# A Standardized Probability Comparison Approach for Evaluating and Combining Pixel-based Classification Procedures

**DongMei Chen**

## Abstract

*In this paper, a standardized probability approach is presented to evaluate the pixel labeling confidence of each pixel and then combine the classification maps generated from different classification procedures for improving classification accuracy. This approach examines the* posterior *probability of the maximum-likelihood classifier or inverse-distance weight for the minimum-distance classifier for each pixel. It recommends that, for every classification, a standardized probability map should be outputted along with the classified map to show the pixel labeling confidence for all pixels. Tests based on different feature combinations and training strategies from Ikonos data show that the proposed approach was effective in improving the labeling confidence, as well as overall classification accuracy when classified maps from different classification procedures were combined. This standardized probability can be used to provide additional spatial information along with the traditional accuracy assessment.*

## Introduction

Improving classification accuracy of remotely sensed images has been the major purpose in most classification research. A computer-assisted classification procedure typically includes five steps: feature extraction, training, classification decision rules, postprocessing, and accuracy assessment (Jensen, 1996). Previous efforts at improving classification accuracy have placed emphasis on the first four steps of classification procedures. Different algorithms and strategies have been developed and tested for various environments by selecting and developing training strategies (Chuviceo and Congalton, 1988; Hixson *et al.*, 1980; Story and Campbell, 1986; Foody *et al.*, 1995; Chen and Stow, 2002), incorporating and selecting appropriate spatial/contextual information and ancillary data (Wharton, 1982; Gong and Howarth, 1990a; Marceau *et al.*, 1990; Gong and Howarth, 1992; Harris and Ventura, 1995; Flygare, 1997; Sharma and Sarker, 1998; Mesev, 1998; Franklin *et al.*, 2000), developing more robust classifiers and knowledge system (Moller-Jensen, 1990; Hepner *et al.*, 1990; Woodcock and Harward, 1992; Civco, 1993; Baraldi and Parmiggiani, 1994; Palubinskas *et al.*, 1995; Rodrigue, 1995; Mathieu-Marni *et al.*, 1996; Maselli, 1996; Kontoes, 2002; Foody, 2002a; Magnussen *et al.*, 2004; Pal *et al.*, 2005); and post-processing (Guo and Moore, 1991; Eyton, 1993; Barnsley and Barr, 1996; Cihlar and Jansen, 2001).

Compared with the growing availability of various algorithms and strategies used in different stages of classification, the development of methods for integrating classified maps from different classification procedures is limited. Several approaches have been proposed from machine learning perspectives to combining the predictions of multiple learned models. These approaches include: the basic ensemble method of taking the unweighted or weighted average from different classifiers, stacked regression methods of combining the membership probabilities estimated for different classifiers (Breiman, 1996; Leblanc and Tibshirani, 1996), the principal components approach of combining regression estimates (Merz and Pazzani, 1999), and correspondence analysis of combining classifiers (Merz, 1999). However, these approaches require substantial computational effort (Steele, 2000) and are difficult for general remote sensing users to implement. Cortijon and La Blanca (1998) demonstrated high accuracy classifications by combinations of spectral-contextual classifiers. Steele (2000) presented two approaches of utilizing spatial information of remotely sensed data for combining multiple classifiers, but his approaches are developed for polygon-based supervised classification. In more recent work of Chen and Stow (2004), three strategies are proposed for improving classification accuracy by integrating information from multiple-resolutions images, in which only one of them was intended to combine classification results for the maximum-likelihood classifier.

The limited development for combining classified maps generated from different classification routines may largely be due to the current practice used in remote sensing classification applications. Usually, a classified map from a classification procedure is the only output provided to map users or map analysts. The accuracy assessment is used to evaluate how good the classified map is through the comparison of sampled reference (or ground truth) data and the classified map (Foody, 2002b). The outcome of the accuracy assessment is often reported in an error matrix from which further accuracy measurements (such as overall accuracy, producer's and user's accuracy, or overall and individual kappa coefficients) can be derived (Congalton and Green, 1999). This kind of accuracy assessment is intended to provide one or several measurements estimating the degree

Department of Geography, Queen's University, Kingston, ON, K7L 3N6, Canada (chendm@post.queensu.ca).

Photogrammetric Engineering & Remote Sensing
Vol. 74, No. 5, May 2008, pp. 000–000.

0099-1112/08/7405–0000/$3.00/0

of the overall agreement of the derived classification with real or ground data. However, it provides no information on the spatial distribution of error and the labeling confidence or accuracy of individual pixels (Jensen, 1996; Pontius, 2000). Using accuracy measurements derived from an error matrix, map analysts or map users can not compare the labeling confidence of individual pixels on different maps generated for the same area from different classification procedures (e.g., classified maps generated by the use of different training data, feature combinations, multiple classifiers, or images with different spectral/spatial resolutions) (Chen and Stow, 2003; Congalton and Green, 1999; Foody *et al.*, 1992; Gong and Howarth, 1990b). Therefore, map users have no way of combining maps from different classification procedures.

The purpose of this paper is to present an objective approach to assess how much confidence should be placed in the classification of each pixel and to combine classification results through comparison of labeling confidences from different classification procedures. This approach recommends that, for every classification, a standardized probability map should be outputted along with the classified map to show the pixel labeling confidence for all pixels. This standardized probability examines the *posterior* probability of the maximum-likelihood classifier or the inverse-distance weight for the minimum-distance classifier. In addition, a strategy is developed to combine classification results from different classification procedures for the same area using the proposed standardized probability. The details of the approach are described in the following section. A test study was conducted using the Ikonos data covering a portion of the western part of the City of Kingston, Ontario, Canada. The results are presented and discussed.

## Methods

The proposed method was developed for traditional pixel-based supervised classification using probability-based algorithms (e.g., the maximum-likelihood classifier) and distance-based algorithms (e.g., the minimum-distance classifier). It is assumed that (a) each pixel can be assigned to one and only one class, (b) the training data can represent the classes for the study area, and (c) the data distribution meets the underlying assumptions that a classifier requires.

To decide into which class a pixel is classified, different algorithms utilize different decision rules. The maximum-likelihood classification is based on an estimated probability function for each class. The class statistics are usually obtained from the training data. Assuming equal prior probability, the probability that a pixel is classified to different classes is calculated according to the Bayesian rule, using the following formula:

$$P(i|X) = \frac{1}{(2\pi)^{n/2}|V_i|^{1/2}} \exp\left(\frac{1}{2}(B_x - M_i)^T V_i^{-1} (B_x - M_i)\right) \quad (1)$$

where, $P(i|X)$ is the probability for pixel $X$ as a member of class $i$, $n$ is the number of the image bands or features, $\boldsymbol{B_X}$ is the feature value vector of pixel X, $\boldsymbol{M_i}$ is the mean vector of class $i$ obtained from the training data, $V_i$ is the variance-covariance matrix for class $i$, $|V_i|$ is the determinant of the $V_i$, and $V_i^{-1}$ is the inverse of the $V_i$.

In the distance-based algorithms such as the minimum-distance classifier, the distance ($D(i|X)$) that a pixel $X$ belongs to class $i$ is calculated according to the distance measurements used. The simplest Euclidean distance measure can be calculated using Equation 2. Other more complicated distance measures can be found in (Schowengerdt, 1997).

$$D(i|X) = \sqrt{\sum_{j=1}^{n} (b_{jX} - m_{ji})^2} \quad (2)$$

where, $b_{jX}$ represents the value of pixel $X$ at band $j$, and $m_{ji}$ represents the mean value of class $i$ at band $j$.

The probabilities or distances that a pixel belongs to different classes are calculated and then ordered. The pixel is assigned to the class with the highest probability or minimum distance in traditional hard classification.

For illustration purposes, we assume that the data to be classified are one-dimensional for the probability-based classifier and two-dimensional for the distance-based classifier. Figure 1a and 2a show how a selected pixel *(X)* is classified in a maximum-likelihood classifier and a mini mum-distance classifier, respectively.

The two normal curves in Figure 1a show the probability distributions for two classes ($C_1$, $C_2$). The probabilities of pixel $X$ being classified into the two classes are $P_1(C_1|X)$ and $P_1(C_2|X)$. In Figure 1a, pixel $X$ will be assigned to class $C_1$ since $P_1(C_1|X)$ is greater than $P_1(C_2|X)$. Figure 2a shows the locations of pixel $X$, and the two classes' means *(M1, M2)* in feature space. The distances of pixel $X$ to the two classes are $D_1(C_1|X)$ and $D_1(C_2|X)$. Based on the rule of minimum-distance classifier, pixel $X$ will be assigned to class $C_1$ for the example in Figure 2a.

As can be seen from Equations 1 and 2, the changes of mean vector of classes ($M_i$) or the variance-covariance matrix of classes can easily shift normal curves of classes or locations of classes in feature space, which in turn changes the probabilities or distances that a pixel belongs to different classes. In practice, the mean vector or variance-covariance matrix of a class is likely to change by using different training samples and feature bands (such as texture), or classifying data at another spatial and/or spectral resolution.

Suppose the probability distributions and mean locations of two classes obtained from another classification procedure for pixel $X$ is the same as shown in Figure 1b and Figure 2b, using the Bayesian rule and distance algorithm, respectively. The probabilities of pixel $X$ being classified into two classes in this procedure are $P_2(C_1|X)$ and $P_2(C_2|X)$ in a maximum-likelihood classifier, while $D_2(C_1|X)$ and $D_2(C_2|X)$ are the distances of pixel $X$ to two classes, respectively. From an analyst's point of view, the closer the two probabilities (or distances), the more ambiguity (or less confident) there is to assign a pixel to a class in a classification decision. Comparing Figure 1a and Figure 2a, it is clear that pixel $X$ has less ambiguity in Figure 1a in terms of being classified into class $C_1$ since $P_1(C_2|X)$ is zero, while the overlap between $P_2(C_1|X)$ and $P_2(C_2|X)$ for pixel $X$ makes it ambiguous to label pixel $X$ to class $C_1$. Similarly, pixel $X$ has less ambiguity to be classified into class $C_1$ in a classification decision of Figure 2a than in Figure 2b since the difference between $D_1(C_1|X)$ and $D_1(C_2|X)$ is greater than the difference between $D_2(C_1|X)$ and $D_2(C_2|X)$.

Reducing ambiguity in pixel labeling is one of the major purposes in most previous classification research. If a pixel has less ambiguity in labeling it to a class, more confidence can be put in its labeling. However, the maximum probabilities or minimum distances obtained from different classification procedures usually cannot be directly used as indicators of confidence of a pixel's classification. As shown in Figure 1, the maximum probability ($P_2(C_1|X)$) in Figure 1b is higher than the maximum probability ($P_1(C_1|X)$) in Figure 1a, but there is more ambiguity of labeling pixel $X$ in Figure 1b than in Figure 1a. Similarly, in Figure 2, there is a less confidence of labeling pixel X in Figure 2b than in Figure 2a, even though the minimum distance ($D_1(C_1|X)$) in Figure 2a is greater than the minimum distance ($D_2(C_1|X)$). Gong and Howarth (1990c) suggest the use of the direct difference

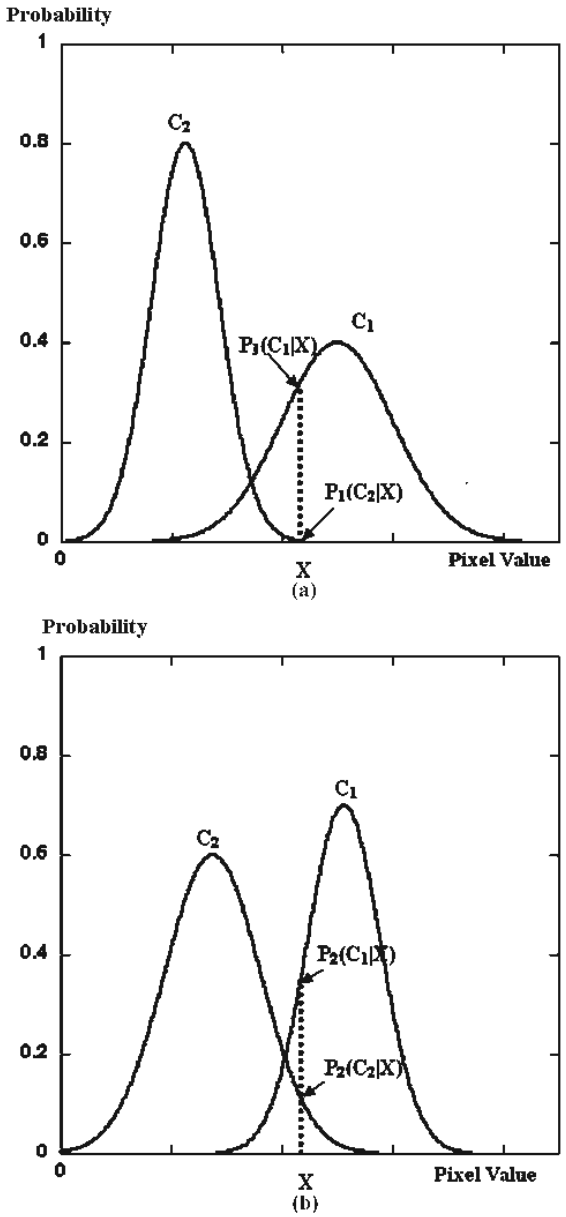


Figure 1. Illustration of labeling confidence for pixel X in two classification procedures using probabilities. $P_1(C_1|X)$ and $P_1(C_2|X)$ represent probabilities of pixel X being classified into classes $C_1$ and $C_2$, respectively, in one maximum-likelihood classification, while $P_2(C_1|X)$ and $P_2(C_2|X)$ represent the probabilities in another classification. Pixel X can be more confidently assigned to class $C_1$ in a) than in b) even though $P_1(C_1|X) < P_2(C_1|X)$.

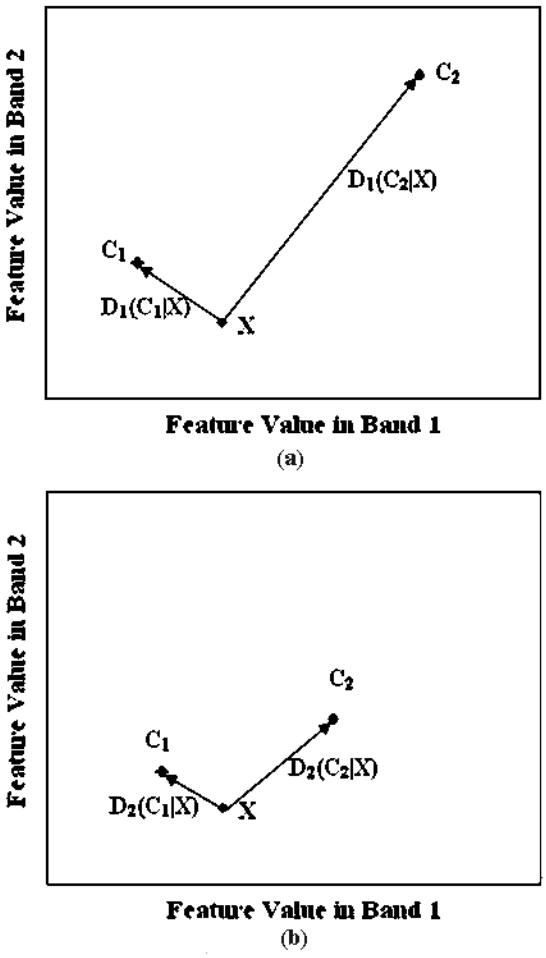


Figure 2. Illustration of labeling confidence for pixel X in two classification procedures using the minimum distance classifier. $D_1(C_1|X)$ and $D_1(C_2|X)$ represent distances of pixel $X$ belonging to classes $C_1$ and $C_2$, respectively, in one classification, while $D_2(C_1|X)$ and $D_2(C_2|X)$ represent distances in another classification. Pixel $X$ can be more confidently assigned to class $C_1$ in a) than in b) even though $D_1(C_1|X) > D_2(C_1|X)$.

between the two largest probabilities as an indicator of ambiguity to select and develop training strategies and feature combinations in a classification procedure. However, the magnitude of direct difference of two probabilities is dependent on the magnitude of probability (or distance) and distance measurements when different classification procedures are used.

In order to compare the ambiguity (or confidence) for a pixel to be assigned to a class from different classification procedures, a standardized measurement independent of the magnitude of probability (or distance) is needed by making all the probabilities (or distance weights) sum to one for each pixel. Let $L(i|X)$ denote this measurement of the standardized probability of pixel $X$ to class $i$.

For the maximum-likelihood classification, $L(i|X)$ is:

$$L(i|X) = \frac{P(i)P(i|X)}{\sum_{i=1}^{N} P(i)P(i|X)} \quad (3)$$

where $P(i|X)$ is the probability for pixel $X$ as a member of class $i$, $P(i)$ is the *a priori* probability of membership of class $i$, and $N$ is the total number of classes. This measurement is the same as the *posterior* probability of class membership for the maximum-likelihood classification, which is used to assess how much confidence should be placed on the

classification of each pixel (Jensen, 1996) and to estimate classification accuracy (Steele, 2005).

For the distance measurement, the standardized probability $L(i|X)$ can be determined by the inverse function of distance to classes with the following Equation 4:

$$L(i|X) = \frac{a_i / D^b_{(i|X)}}{\sum_{i=1}^{N} (a_i / D^b_{(i|X)})} \quad (4)$$

where $D_{(i|X)}$ is the distance of pixel $X$ to class $i$, $a_i$ is the *a priori* weight assigned to class $i$, and $b$ is the power to which distance is raised. For the distanced-based measurement, $L(i|X)$ is same as the inverse-distance weight. The $L(i|X)$ can also be modified, based on distance-based classification rules used. For example, Equation 4 will work for the minimum-distance classification algorithm. However, when the parallelepiped classification algorithm is used, the rule used for assigning $L(i|X)$ may be modified by giving higher or full membership to pixels falling in the threshold of distance specified for each class.

For any classification, when a pixel falls in the area in which two or more probability curves overlap or has similar distances to different classes, the standardized probabilities are reduced. Since the final labeling of a pixel is given to the class with the maximum probability or the minimum distance, the highest of $L(i|X)$ ($i = 1, 2, 3, \ldots N$ possible classes) indicates the level of confidence (or ambiguity) of labeling a pixel. Usually, the closer the maximum probability (or the minimum distance) is to other probabilities (or distance) for a pixel, the lower the highest of $L(i|X)$ ($i = 1, 2, 3, \ldots N$ possible classes) will be. Therefore, the higher the maximum of $L(i|X)$ a pixel has, the more confidence can be placed on the labeling of that pixel. In Figure 1a, the highest standardized probability of $L(i|X)$ is 1 for pixel $X$, while the value of $L(i|X)$ is much lower in Figure 1b, although the maximum probability is higher in Figure 1b than in Figure 1a. It is recommended that the maximum of $L(i|X)$ is displayed along with the final classification map for each classification procedure to indicate to analysts or users the locations of possible problematic pixels with low confidence labeling. The maximum of $L(i|X)$ can also provide additional information on areas that should be focused on and suggest if additional fieldwork is required.

The standardized probabilities suppress the absolute values of posterior probabilities, which are important in determining the confidence. It should be mentioned that the use of the maximum of $L(i|X)$ to represent labeling confidence is under assumptions that the training data and their classes can represent classes in the region, and that each pixel can be classified into only one class, as well as that the data distribution meets the underlying requirement of the classifier. If these assumptions do not hold, the use of the maximum of $L(i|X)$ would lead to a wrong judgment. For example, for images with a low-resolution scene model (Woodcock and Strahler, 1987), in which the majority of pixels are mixed with different classes, the maximum of $L(i|X)$ will not help the confidence of assigning this pixel to a class. If a pixel does not belong to any classes specified in the training data, this would result in very low probabilities of $P(i|X)$ (e.g., 0.02 and 0.08 for two classes) to all classes. Its labeling confidence would be exaggerated when the standardized probability is used (e.g., 20 percent and 80 percent in the above case). Therefore, the maximum probability should be examined before it is converted to the standardized probability.

It should also pointed out that the standardized approach is not recommended or should be used with caution in the situation where many spectral classes are classified separately in the classifier, but then subsequently combined into a smaller number of final map classes. In this situation, standardized posterior probabilities should only be calculated by comparing the probabilities of spectral classes associated with other map classes, and not the same map class.

As pointed out in previous research, different features may be more accurately classified using different training strategies, feature combinations, and high- or low-resolution images (Woodcock and Strahler, 1987; Marceau *et al.*, 1994; and Chen *et al.*, 2004). For each classification procedure $l$, the maximum of $L_l(i|X)$ and its related class can be output for each pixel. By comparing all the maximum of $L_l(i|X)$ derived from different classification procedures, pixel $X$ can be assigned to the class with the highest maximum of $L_l(i|X)$. Thus, in the combination of classification results from different procedures, $X$ in class $i$, if and only if:

$$L(i|X) \geq L_l(k|X), \quad (5)$$

where, $k = 1, 2, 3, \ldots N$ possible classes, and $l$ = all possible classifications.

The rule in Equation 5 can be applied to all pixels in the study area. To obtain a general trend for the maximum of $L_l(i|X)$ for a group of pixels (such as training data) or all pixels in an image, the average of the maximum of $L_l(i|X)$ can be calculated for evaluation.

## Experiment

The data used for this study were Ikonos multispectral imagery covering a portion of the western part of the City of Kingston, Ontario, Canada. The Ikonos image was acquired on 25 April 2000, and has a spatial resolution of 4 m with four spectral bands (R, G, B, and NIR). Ten land-cover classes were used: residential roof, industrial/commercial roof, road surface, irrigated grass, conifer tree, deciduous tree, bare/cleared land, natural grass, water, and crop. Figure 3 shows the BW image of the NIR band of the study area.

Classifications with different feature combinations and spatial resolutions were conducted using both the maximum-likelihood classifier and the minimum-distance classifier to test the effectiveness of the proposed approach in the combination of classification results from different classification procedures to achieve better classification accuracy.

Training data was selected by block (or cluster) training. As a general rule in block training, the length and width of small blocks for each class were close to the range obtained from the semi-variogram so that each block was big enough to represent the spectral and spatial properties of each class (Chen and Stow, 2002).

Two texture bands derived as the variance of 3 by 3 and 5 by 5 moving windows from the near-infrared band were incorporated into the four spectral bands, respectively. Images with only spectral bands as well as two texture combinations were classified with the same training data. In order to test the effectiveness of the proposed approach in combining classifications from different spatial resolutions, the 4 m Ikonos image was also aggregated into 8 m resolution level by an averaging method. The classification results obtained from 4 m and 8 m resolution images using the same training data selected above were compared.

For each classification, the probabilities that each pixel belongs to each class were generated, and then converted to the standardized probabilities of $L_l(i|X)$ ($i = 1, 2, \ldots, n$ possible classes) using Equation 3. The maximum of $L_l(i|X)$ obtained from different feature combinations were compared

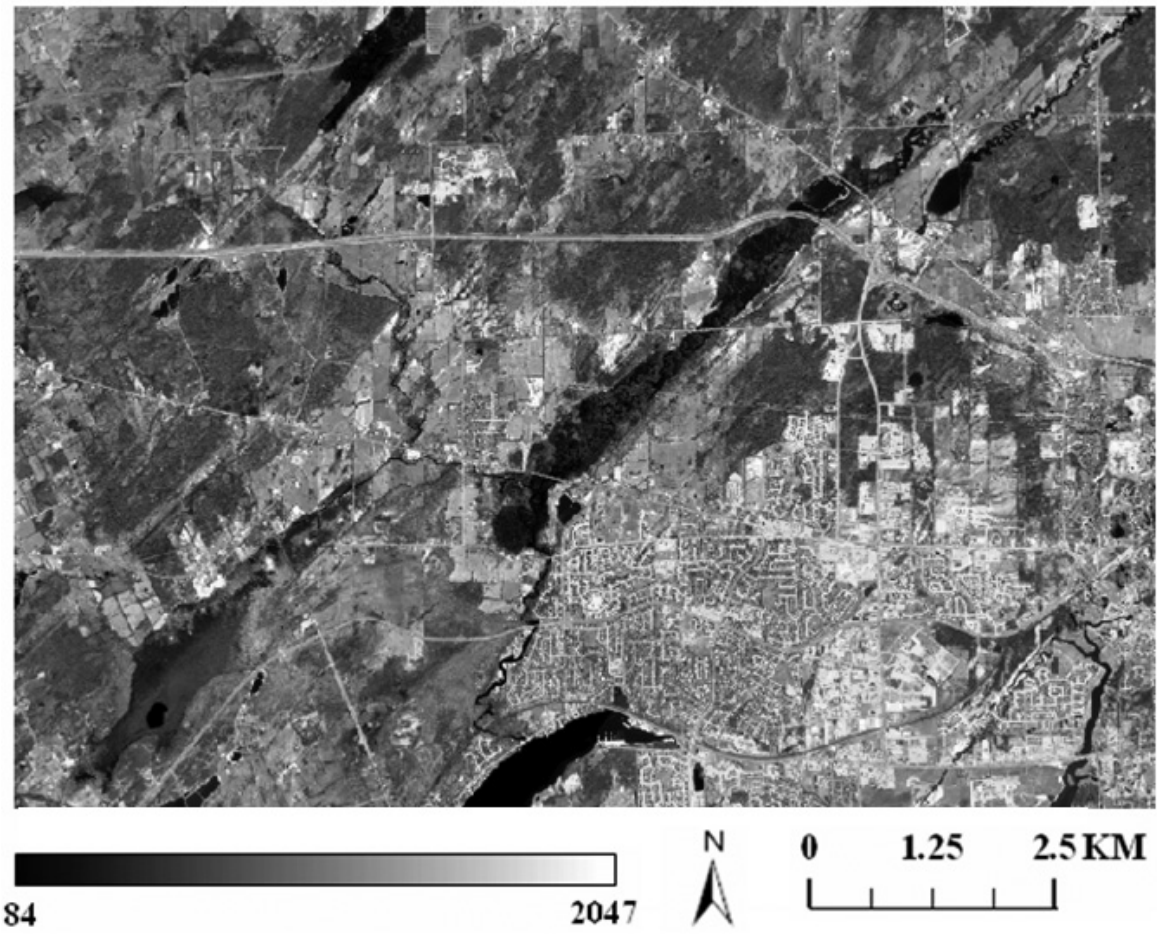


Figure 3. Illustration of NIR band of 4 m Ikonos images for the study area located in the western part of the City of Kingston, Ontario, Canada.

with each other, and then were combined using the rule listed in Equation 5, and each pixel was assigned to the class having the highest maximum standardized probability in all classifications. The same process was also applied to classifications obtained from 4 m and 8 m resolution images. Final combined classification maps were generated along with maps of the maximum standardized probability.

The same land-cover classification scheme and the same reference data were used for accuracy estimation in all classifications. A total of 863 randomly selected samples different from training data were used as reference data for the study area. The size of each class in the reference data is listed in Table 1. The land-cover classes for the reference data were identified by manually interpreting Ikonos images, 1 m digital color aerial photographs and by making ground-level observations. An error matrix was generated for each classified map. A discrete multivariate evaluation of error matrix, Kappa coefficient, was used to evaluate the agreement of classified map and reference data (Jensen, 1996). The Kappa coefficient is similar to the user's and producer's accuracies, which are descriptive evaluation of error matrix. The overall and individual (class specific) kappa coefficients were reported for all classified and combined maps. The average of the maximum $L_l(i|X)$ of reference data and training data were calculated and compared with their accuracy measurements.

## Results

Table 1 summarizes the overall and individual kappa coefficients for all classified and combined maps in discriminating between ten land-cover classes. Since the results generated from the minimum-distance classifier are very close to those obtained from the maximum-likelihood classifier (the correlation coefficient between them is around 0.96 for their individual kappa values and 0.95 for their mean standardized probabilities), only results from the maximum-likelihood classifier are reported. The overall and individual kappa coefficients are not much different at 4 m and 8 m spatial resolutions. As expected, inclusion of texture bands improved classification accuracy. However, adding a $3 \times 3$ window texture band did not result in substantially different classification accuracy than adding a $5 \times 5$ window texture band in this study. It is clear that the highest accuracy of each individual class was not always achieved at the classification with the highest overall accuracy when compared to the individual kappa coefficients among the classified maps generated from different feature combinations and spatial resolution images. This confirms the conclusion from other researchers that there is not a single optimal spatial resolution and window size appropriate for all classes (Marceau *et al.*, 1994; Chen *et al.*, 2004). Since the effectiveness of different feature combinations and spatial resolutions for different classes to achieve high

TABLE 1. THE REFERENCE DATA SIZE AND CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFICATIONS (THE ACCURACY IS MEASURED IN THE INDIVIDUAL KAPPA IN PERCENT FOR EACH CLASS)

| Classes | Reference Size (pixel) | Spectral Bands at 4 m (1) | Spectral Bands + 3 × 3 Texture at 4 m (2) | Spectral Bands + 5 × 5 Texture at 4 m (3) | Spectral Bands at 8 m (4) | Combination of (1) and (2) | Combination of (1), (2), and (3) | Combination of (1) and (4) | Combination of (1), (2), (3), and (4) |
|---|---|---|---|---|---|---|---|---|---|
| RE | 48 | 50 | 66 | 57 | 47 | 67 | 68 | 58 | 73 |
| IN | 28 | 55 | 56 | 51 | 57 | 58 | 59 | 59 | 55 |
| RS | 98 | 62 | 62 | 59 | 62 | 66 | 66 | 68 | 69 |
| IG | 82 | 67 | 73 | 68 | 70 | 74 | 74 | 73 | 75 |
| CT | 158 | 87 | 88 | 86 | 89 | 88 | 89 | 89 | 91 |
| DT | 101 | 72 | 65 | 63 | 74 | 72 | 69 | 75 | 79 |
| CR | 105 | 61 | 73 | 75 | 65 | 78 | 74 | 69 | 75 |
| CL | 31 | 71 | 73 | 73 | 73 | 72 | 73 | 72 | 72 |
| NG | 103 | 67 | 69 | 71 | 67 | 69 | 71 | 69 | 76 |
| WA | 109 | 99 | 99 | 99 | 99 | 99 | 99 | 99 | 99 |
| Overall | 863 | 61 | 70 | 68 | 67 | 72 | 73 | 72 | 77 |

(RE: residential roof, IN: industrial/commercial roof, RS: road surface, LG: irrigated grass, CT: conifer tree, DT: deciduous tree, CR: crops, CL: cleared and bare field, NG: natural grass, WA: water)

accuracy is not the major concern of this paper, the main focus will be on the proposed standardized probability $L(i|X)$ in the reference data and classification accuracy in the following.

The overall kappa coefficients obtained from combined classifications using the standardized probability comparison are greater when compared to those from single classification procedures, as evidenced in Table 1. The lowest kappa value from four combinations is 0.72, while the highest accuracy from all single classification procedures is 0.70. The highest accuracy occurred on the combined map of all four single classifications for the majority of classes. All individual kappa values on combined maps are greater or equal to the highest kappa value obtained from their corresponding single classification procedures for all classes except: the industrial/commercial roof on the combined map of all four single classifications, the deciduous tree on the combined map of all three single classifications at 4 m, and the cleared field on the three combinations. However, in those exceptional cases, the individual kappa on the combined map was still higher or equal to the lowest kappa value obtained from the single classification procedure. It is evidential that the combination of different classification procedures using the proposed approach improved the overall classification accuracy as well as the accuracy for almost all classes.

It is clear from Table 1 that the improvement of accuracy for individual classes varied in different combinations. For classes that already achieved very high individual kappa values from individual classification procedures, combining classification results from those individual procedures led to no or very limited accuracy improvement. For example, the individual kappa value of water surface was 0.99 in all four single classification procedures, which indicated that water surface can be very accurately distinguished from other classes. Combination of single classification procedures did not change its kappa value. For classes having similar kappa values from single procedures, their individual classification accuracy showed very small or no increases after combinations. For example, in all four single classifications, cleared field had kappa values of 0.71 or 0.73. This means that none of the four different feature combinations used in four single classification procedures was more effective in discriminating cleared field than another. In all four cases of combination, its individual kappa values were 0.72 or 0.73. Similar trends can be found for the conifer tree, although conifer tree did show an increase on its kappa value on the combined map of all four single classifications. The substantial increase of individual coefficients occurred when single procedures in which a class achieved obviously different individual accuracy were combined. Among all classes, the residential roof had the highest increase of individual kappa. The average increase of its kappa was 18 percent when the four single procedures were combined. The individual kappa coefficient of residential roof was 0.5 and 0.47 when only spectral bands at 4 m and 8 m were used, respectively. When the 3 × 3 and 5 × 5 texture bands were added as another layer into classification, the individual kappa coefficient of residential roof was increased to 0.66 and 0.57. When they were combined together using the proposed approach, the individual kappa of residential roof increased to 0.73 in the final classified map.

For each class, the mean, minimum, maximum, and the standard deviation values of the $L(i|X)$ were obtained for their respective reference data. Table 2 summarizes the mean of maximum of $L(i|X)$ of each class in different classifications for reference data. It can be seen that combining the classification results from different procedures using the rule defined in Equation 5 increases the mean standardized maximum probability, which means that combined maps using the proposed approach have higher confidence of pixel labeling compared with maps generated from single classification procedures. In fact, the increase of the standardized maximum probability is more obvious than the improvement of the classification accuracy on the combined maps. Comparing the mean standardized maximum probability in the combined classification map to those in the individual classification yielded significant difference at the 0.10 significance level as demonstrated by the t-tests. When the classified result obtained from the use of only spectral bands was combined with the results generated from the inclusion of 3 × 3 and 5 × 5 texture bands, the average increase of the mean standardized maximum probability was 8 percent for all classes, while the average increase was 7 percent when combining the two classification results generated from 4 m and 8 m resolutions. The class with the highest increase was residential roof in both combinations. The increase of the maximum $L_{(i|X)}$ for commercial/industrial roof and road surface was higher in the combination of classified results from 4 m and 8 m resolutions than in the combination of classification results from texture and spectral bands.

Comparing the values in Table 2 to the accuracy estimates in Table 1, it is clear that changes of the standardized

TABLE 2. THE MEAN OF THE STANDARDIZED PROBABILITIES ($L_i(i|X)$)(%) FOR THE REFERENCE DATA IN DIFFERENT CLASSIFICATIONS

| | Single Classification Procedure | | | | Combination of Multiple classification Procedures | | | |
|---|---|---|---|---|---|---|---|---|
| Classes | Spectral Bands at 4 m (1) | Spectral bands + 3 × 3 Texture at 4 m (2) | Spectral Bands + 5 × 5 Texture at 4 m (3) | Spectral Bands at 8 m (4) | Combination of (1) and (2) | Combination of (1), (2), and (3) | Combination of (1) and (4) | Combination of (1), (2), (3), and (4) |
| RE | 78 | 85 | 86 | 85 | 89 | 91 | 90 | 93 |
| IN | 79 | 78 | 74 | 82 | 77 | 85 | 89 | 91 |
| RS | 80 | 81 | 81 | 81 | 84 | 85 | 87 | 89 |
| IG | 86 | 86 | 85 | 85 | 89 | 95 | 92 | 96 |
| CT | 91 | 92 | 93 | 90 | 91 | 97 | 97 | 99 |
| DT | 88 | 91 | 91 | 90 | 92 | 95 | 94 | 97 |
| CR | 82 | 88 | 89 | 84 | 92 | 92 | 89 | 94 |
| CL | 84 | 87 | 86 | 89 | 85 | 92 | 92 | 94 |
| NG | 81 | 82 | 82 | 83 | 83 | 88 | 89 | 92 |
| WA | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

(RE: residential roof, IN: industrial/commercial roof, RS: road surface, LG: irrigated grass, CT: conifer tree, DT: deciduous tree, CR: crops, CL: cleared and bare field, NG: natural grass, WA: water)

maximum probability correspond with changes of individual kappa values. The correlation coefficient between individual kappa coefficients and the mean of standardized maximum probability is 0.82 in all classifications. However, this correlation coefficient varies from 0.21 to 0.93 for individual classes. For some classes the correlation coefficient between individual kappa coefficients and the mean of standardized maximum probability is high (0.74 for residential, 0.93 for road surface, 0.89 for crops, 0.74 for natural grass, 0.76 for irrigated grass), but for several other classes, the correlation coefficients are quite low (0.42 for deciduous tree, 0.47 for industrial/commercial roof, 0.21 for cleared field, and no correlation for water surface). The no correlation for water surface is caused by the total lack in variation in the water values. This indicates that the high mean of the maximum standardized probability does have a tendency to correspond with high classification accuracy, but it is not directly related to the classification accuracy.

As previously discussed, one advantage of the proposed approach is the ability to combine different maps generated from different classification procedures. Another advantage of this approach is that the standardized probability can be calculated for every pixel. Therefore, the confidence level of pixel labeling can be examined spatially. Figure 4 shows the standardized maximum probabilities of several classified maps generated in this study. It is evident that labeling confidence is different for different pixels at different locations for different classification procedures. The map of the maximum standardized probability in classified map generated from using only spectral bands (Figure 4a) appears darker than other maps, indicating that the majority of the pixels have a lower labeling confidence compared with other maps in Figure 4. By adding a 3 × 3 texture, the main areas with improved standardized probability are residential area, cropland, and natural grassland. The two combined maps (Figure 4c and 4d) appear brighter than the two maps above, indicating that the combination of the labeling confidence has been increased for most pixels. By comparing the standardized probability of individual pixels, the changes of different classification procedures on pixel labeling can also be identified. For example, pixels with changed standardized probability can be obtained easily by subtracting the probability map in Figure 4a from the map in Figure 4b.

By examining the maps in Figure 4 it is also clear that pixels in some regions have low labeling confidence in all classified and combined maps. One obvious feature with low standardized probability is the road network in dark color in all maps of Figure 4. This may indicate that neither spectral bands nor variance texture bands are capable of separating linear road features from other features. In order to better identify linear road surfaces, other texture features that can better capture linear objects should be applied. Further investigation on the other dark regions of Figure 4d reveals that most regions with low standardized probability are located in wetlands, where the soil moisture is significantly higher than the rest of the study area. The training data used in this study did not capture the impact of soil moisture on the spectral values of different classes in wetland. For those regions with high soil moisture, separate classification should be conducted by using separate training sets.

## Conclusions

This paper presents an approach that uses standardized probability measurement to compare the labeling confidence of the final labeling of each pixel, and then combines the classification maps generated from different classification procedures to improve classification accuracy. The use of the proposed measurement representing labeling confidence is under assumptions that the training data and their classes can represent classes in the region, and that each pixel can be classified into only one class, as well as that the data distribution meets the underlying requirement of the classifier. This approach examines the *posterior* probability of the maximum-likelihood classifier or inverse-distance weight for the distance-based classifier for each pixel. It is objective and easy to implement. This approach recognizes that pixels of different classes at different locations might require different training strategies, feature combinations, and spatial resolutions in order to be better identified. It recommends that, for every classification, a standardized probability map should be outputted along with the classified map to show the pixel labeling confidence for all pixels. The advantage of using standardized probability is that it is available for all pixels and can be used to identify regions with low labeling confidence. This standardized probability can be used to provide additional spatial information along with the traditional accuracy assessment and probability or distance map.

The results of the experimental study demonstrate that the proposed standardized probability can substantially improve the reliability of final classification labeling when classified maps from different procedures are combined. Although there is a trend that the increase of maximum standardized probability corresponds with the increase of

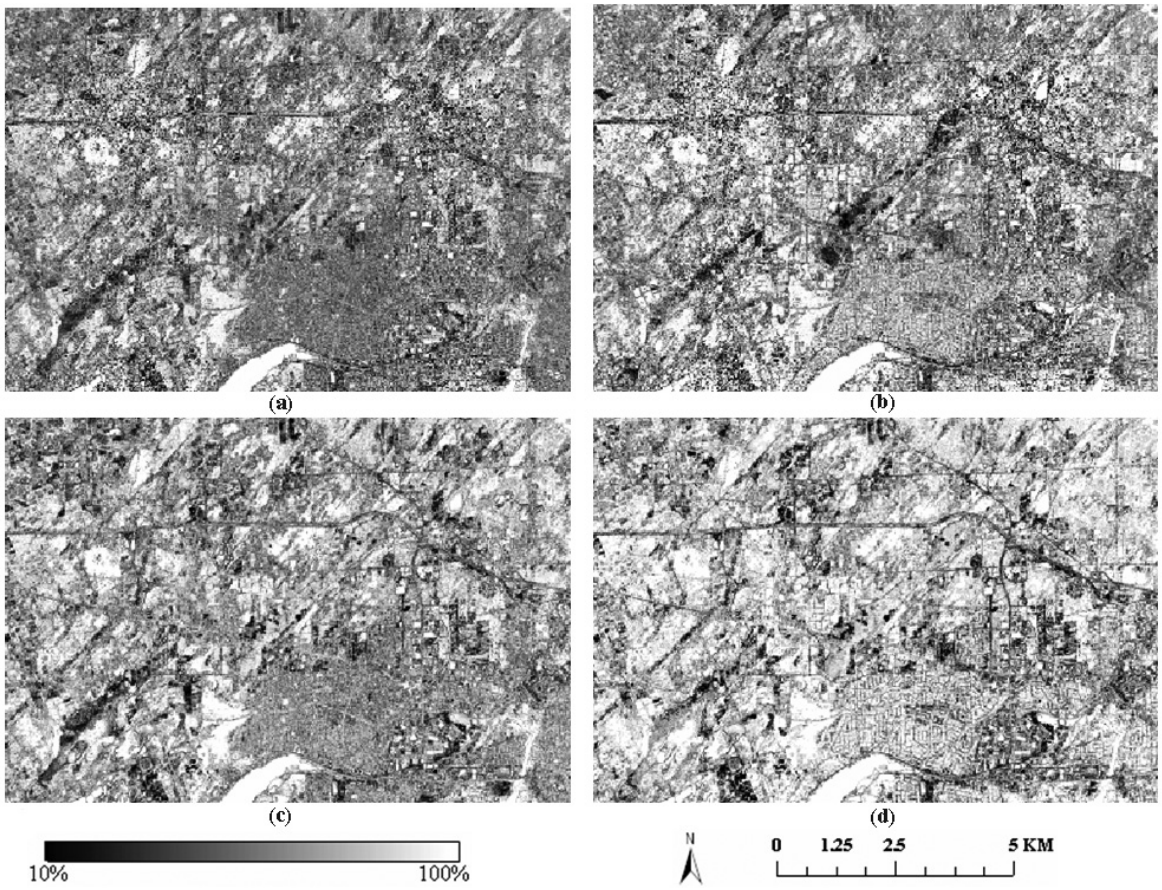


Figure 4. The maximum standardized probabilities of classified maps generated from (a) using only spectral bands at 4 m, (b) adding a 3 × 3 texture band into the spectral bands, (c) combining the classifications from 4 m and 8 m resolutions, and (d) combining all classification results

classification accuracy, it should be noted that the increase of standardized probability is not a direct measurement of map accuracy increase. How much improvement of each class's accuracy on a combined map depends on the characteristic of individual classes on single classification procedure. For classes that already achieved very high accuracy or similar individual accuracy from single classification procedures, combining classification results from those individual procedures would lead to no or very limited accuracy improvement. The substantial increase of classification accuracy of an individual class occurs when classification results are combined from single classification procedures with obviously different effectiveness of discriminating that class.

## Acknowledgments

This research is supported by a discovery grant from National Science and Engineering Research Council of Canada. The author would like to thank Jie Tian's assistance in the collection of reference data. Special thanks go to the four anonymous reviewers for their extensive and thoughtful comments and suggestions that led to substantial improvements in this article.

## References

Baraldi, A., and F. Parmiggiani, 1994. A Nagao-Matsuyama approach to high-resolution satellite image classification, *IEEE Transactions on Geoscience and Remote Sensing*, 4(32):749–758.

Barnsley, M.J., and S.L. Barr, 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification, *Photogrammetric Engineering & Remote Sensing*, 8(62):949–958.

Breiman, L., 1996. Stacked regressions, *Machine Learning*, 24:49–64.

Chen, D., and D. Stow, 2002. The effect of training strategies on supervised classification at different spatial resolutions, *Photogrammetric Engineering & Remote Sensing*, 68(11):1155–1161.

Chen, D., and D. Stow, 2003. Strategies for integrating information from multiple spatial resolutions into land use/cover classification routines, *Photogrammetric Engineering & Remote Sensing*, 69(11):1279–1287.

Chen, D., D.A. Stow, and P. Gong, 2004. Examining the effect of spatial resolution and texture window size on classification accuracy: An urban environment case, *International Journal of Remote Sensing*, 25(11):2177–2192.

Chuvieco, E., and R.G. Congalton, 1988. Using cluster analysis to improve the selection of training statistics in classifying remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 54(9):1275–1281.

Cihlar, J., and L.J.M. Jansen, 2001. From land cover to land use: A methodology for efficient land use mapping over large areas, *Professional Geographer*, 53(2):275–289.

Civco, D.L., 1993. Artificial neural networks for land-cover classification and mapping, *International Journal of Geographic Information System*, 2(7):173–186.

Congalton, R.G., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, New York, 137 p.

Cortijon, F.J., and N.P. La Blanca, 1998. Improving classical contextual classifications, *International Journal of Remote Sensing*, 19(8):1591–1613.

Eyton, J.R., 1993. Urban land use classification and modeling using cover-type frequencies. *Applied Geography*, 13:111–121.

Flygare, A.-M., 1997. A comparison of contextual classification methods using Landsat TM, *International Journal of Remote Sensing*, 18(18):3835–3842.

Foody, G.M., 2002a. Hard and soft classifications by a neural network with a non-exhausively defined set of classes, *International Journal of Remote Sensing*, 23(18):3853–3864.

Foody, G.M., 2002b. Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, 80(1):185–201.

Foody, G.M., M.B. McCulloch, and W.B. Yates, 1995. Classification of remotely sensed data by an artificial neural network: Issues related to training data characteristics, *Photogrammetric Engineering & Remote Sensing*, 61(3):391–401.

Foody, G.M., N.A. Campbell, N.M. Trodd, and T.F. Wood, 1992. Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification, *Photogrammetric Engineering & Remote Sensing*, 58(9):1335–1341.

Franklin, S.E., R.J. Hall, L.M. Moskal, A.J. Maudie, and M.B. Lavigne, 2000. Incorporating texture into classification of forest species composition from airborne multispectral images, *International Journal of Remote Sensing*, 21(1):61–79.

Gong, G., and P.J. Howarth, 1990a. The use of structural information for improving land-cover classification accuracies at the rural-urban fringe, *Photogrammetric Engineering & Remote Sensing*, 56(1):67–73.

Gong, P., and P.J. Howarth, 1990b. A graphical approach for the evaluation of land-cover classification procedures, *International Journal of Remote Sensing*, 11(5):899–905.

Gong, P., and P.J. Howarth, 1992. Frequency-based contextual classification and gray-level vector reduction for land-use Identification, *Photogrammetric Engineering & Remote Sensing*, 58(4):423–437.

Guo, L.J., and J.M. Moore, 1991. Post-classification processing for thematic mapping based on remotely-sensed image data, *Proceedings of the International Conference of IEEE Geoscience and Remote Sensing Society*, IEEE, New York, Espoo, Finland, pp. 2203–2206.

Gurney, C.M., and J.R.G. Townshend, 1983. The use of contextual information in the classification of remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 49(1):55–64.

Harris, P.M., and S.V. Ventura, 1995. The integration of geographic data with remotely sensed imagery to improve classification in an urban area, *Photogrammetric Engineering & Remote Sensing*, 61(8):993–998.

Hepner, G.F., T. Logan, N. Ritter, and N. Bryant, 1990. Artificial neural network classification using a minimal training set: Comparison to conventional supervised classification, *Photogrammetric Engineering & Remote Sensing*, 56(4):469–473.

Hixson, M., D. Scholz, N. Fuhs, and T. Akiyama, 1980. Evaluation of several schemes for classification of remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 46(12):1547–1553.

Jensen, J.R., 1996. *Introductory Digital Image Processing – A Remote Sensing Perspective*, Prentice Hall, Upper Saddle River, New Jersey, 316 p.

Kontoes, C.C., V. Raptis, M. Lautner, and R. Oberstadler, 2000. The potential of kernel classification techniques for land use mapping in urban areas using 5m-spatial resolution IRS-1C imagery, *International Journal of Remote Sensing*, 21(16):3145–3151.

LeBlanc, M., and R. Tibshirani, 1996. Combining estimates in regression and classification, *Journal of the American Statistical Association*, 91:1641–1650.

Magnussen, S., P. Boudewyn, and M. Wulder, 2004. Contextual classification of Landsat TM images to forest inventory cover type, *International Journal of Remote Sensing*, 25(12):2421–2440.

Marceau, D.J., P.J. Howarth, J.M.M. Dubois, and D.J. Gratton, 1990. Evaluation of the grey-level co-occurance matrix method for land-cover classification using SPOT imagery, *IEEE Transactions on Geoscience and Remote Sensing*, 4(28):513–519.

Marceau, D.J., P.J. Howarth, and D.J. Gratton, 1994. Remote sensing and the measurement of geographical entities in a forest environment 1: The scale and spatial aggregation problem, *Remote Sensing of Environment*, 49:93–104.

Maselli, F., A. Rodolfi, and C. Conese, 1996. Fuzzy classification of spatially degraded Thematic Mapper data for the estimation of sub-pixel components, *International Journal of Remote Sensing*, 17(3):537–551.

Mathieu-Marni, S., S. Moisan, and R. Vincent, 1996. A knowledge-based system for the computation of land cover mixing and the classification of multi-spectral satellite imagery, *International Journal of Remote Sensing*, 17(8):1483–1492.

Merz, C.J., 1999. Using correspondence analysis to combine classifiers, *Machine Learning*, 36:33–58.

Merz, C.J., and M.J. Pazzani, 1999. A principal components approach to combining regression estimates, *Machine Learning*, 36:9–32.

Mesev, V., 1998. The use of census data in urban image classification, *Photogrammetric Engineering & Remote Sensing*, 64(5):431–438.

Moller-Jensen, L., 1990. Knowledged-based classification of an urban area using texture and context information in Landsat-TM imagery, *Photogrammetric Engineering & Remote Sensing*, 56(6):899–904.

Pal, N.R., A. Laha, and J. Das, 2005. Designing fuzzy rule based classifier using self-organizing feature map for analysis of multispectral satellite images, *International Journal of Remote Sensing*, 26(10):2219–2240.

Palubinskas, G., R.M. Lucas, G.M. Foody, and P.J. Curran, 1995. An evaluation of fuzzy and texture-based classification approaches for mapping regenerating tropical forest classes from Landsat-TM data, *International Journal of Remote Sensing*, 4(16):747–759.

Pontius, R.G., 2000. Quantification error versus location error in comparison of categorical maps, *Photogrammetric Engineering & Remote Sensing*, 66(8):1011–1016.

Rodrigue, J.-P., 1995. The Heuristic classification of functional land use: A knowledge-based approach, *Geographical Systems*, 2:103–120.

Schowengerdt, R.A., 1997. *Remote Sensing Models and Methods for Image Processing*, Academic Press, San Diego, California, 522 p.

Sharma, K.M.S., and A. Sarker, 1998. A modified contextual classification technique for remote sensing data, *Photogrammetric Engineering & Remote Sensing*, 64(4):273–280.

Steele, B.M., 2000. Combining multiple classifiers: An application using spatial and remotely sensed information for land cover type mapping, *Remote Sensing of Environment*, 74(3):545–556.

Steele, B.M., 2005. Maximum posterior probability estimators of map accuracy, *Remote Sensing of Environment*, 99(3):254–270.

Story, M.H., and J.B. Campbell, 1986. The effect of training data variability on classification accuracy, *Proceedings of the Annual ASPRS Conference*, Washington, D.C., pp. 370–379.

Wharton, S.W., 1982. A context-based land-use classification algorithm for high-resolution remotely sensed data, *Journal of Applied Photographic Engineering*, 1(8):46–50.

Woodcock, C.E., and V.J. Harward, 1992. Nested-hierarchical scene models and image segmentation, *International Journal of Remote Sensing*, 16(13):3167–3187.

Woodcock, C.E., and A.H. Strahler, 1987. The factor of scale in remote sensing, *Remote Sensing of Environment*, 21(3):311–332.

(Received 10 August 2006; accepted 22 September 2006; revised 03 October 2006)