Large-Area Land-Cover Mapping through Scene-Based Classification Compositing

B. Guindon and C.M. Edmonds

Abstract

Over the past decade, a number of initiatives have been undertaken to create definitive national and global data sets consisting of precision corrected Landsat Multispectral Scanner (MSS) and Thematic Mapper (TM) scenes. One important application of these data is the derivation of large area land-cover products spanning multiple satellite scenes. A popular approach to land-cover mapping on this scale involves merging constituent scenes into image mosaics prior to image clustering and cluster labeling, thereby eliminating redundant geographic coverage arising from overlapping imaging swaths of adjacent orbital tracks. In this paper, arguments are presented to support the view that areas of overlapping coverage contain important information that can be used to assess and improve classification performance. A methodology is presented for the creation of large area land-cover products through the compositing of independently classified scenes. Statistical analyses of classification consistency between scenes in overlapping regions are employed both to identify mislabeled clusters and to provide a measure of classification confidence for each scene at the cluster level. During classification compositing, confidence measures are used to rationalize conflicting classifications in overlap regions and to create a relative confidence layer, sampled at the pixel level, which characterizes the spatial variation in classification quality over the final product. The procedure is illustrated with results from a synoptic mapping project of the Great Lakes watershed that involved the classification and compositing of 46 Landsat MSS scenes.

Introduction

Within the last decade, a number of initiatives have been undertaken to assemble databases of precision processed Landsat imagery. These activities have included both MSS, e.g., the North American Landscape Characterization (NALC) program (Lunetta *et al.*, 1998), and TM imagery, for example, the Multi-resolution Land Characteristics (MRLC) consortium (Loveland and Shaw, 1996) and the GEOCover (Dykstra *et al.*, 2000) programs. One of the most important systematic uses of these data is large-area land-cover mapping, for example, the creation of the National Land Cover Data (NLCD) product for the conterminous United States (Vogelmann *et al.*, 2001).

Satellite-based land-cover products gained widespread endorsement within the remote sensing community because of their potential for providing valuable large-area information. On the other hand, the acceptance of such information products by the "outside world," (i.e., decision makers and resource managers), requires that their accuracy characteristics be rigorously quantified and documented. This presents two important challenges for information producers:

- For moderate resolution sensors such as those carried by the Landsat series of satellites, large-area coverage can only be achieved by merging scenes acquired over a relatively broad temporal window. For example, in the case of the NALC program, this window was targeted to span three consecutive growing seasons for each "epoch" data set (Sohl and Dwyer, 1998). As a consequence, the constituent scenes can exhibit a range in information detail related to differences in vegetative development, soil moisture, and atmospheric conditions. This characteristic, in conjunction with intra-product variations in class proportions, can be expected to lead to significant intra-product spatial variations in accuracy.
- In addition, current accuracy assessment methods rely in large part on detailed comparison between derived information and independently acquired "ground truth" (Congalton and Green, 1999). Because of cost and logistics, ground truth tends to be limited and can be viewed as a "spot checking" exercise. In addition, truth information is most easily acquired in regions where thematic class purity exists, whereas image classifications are most likely to be uncertain near inter-class boundaries. We suggest that a preferred accuracy assessment strategy is one that couples limited ground-truth testing (absolute accuracy checking) with a "wall-to-wall," albeit indirect, method of relative classification consistency mapping at the individual pixel level.

A widely-used processing approach to large-area mapping involves merging multiple scenes into regional image mosaics (e.g., Homer *et al.*, 1997; Vogelmann *et al.*, 1998) that are then classified. The apparent advantages of this procedure are that (1) it reduces the processing load by eliminating redundant ground coverage that is present due to overlapping imaging swaths of adjacent satellite tracks and (2) it eases the data management and classification load because only a relatively small number of mosaics have to be dealt with rather than a large number of distinct scenes. For example, in the NLCD land-cover initiative, the conterminous U.S. was mapped with a set of regional mosaics, each consisting of 16 to 20 scenes (Vogelmann *et al.*, 1998; Vogelmann *et al.*, 2001).

The repeat coverage available from overlapping imaging swaths has been a largely unexploited characteristic of scene data sets. It is argued that, because an individual scene may be subject to significant limitations in information content, it is desirable that the full information content of the parent image data set be harnessed, including all data in overlap regions. In this paper, a large-area land-cover mapping methodolgy is

B. Guindon is with the Canada Centre for Remote Sensing, 588 Booth Street, Ottawa, Ontario K1A 0Y7, Canada (bert.guindon@ccrs.nrcan.gc.ca).

C.M. Edmonds is with the U.S. Environmental Protection Agency, P.O. Box 93478, Las Vegas, NV 89193-3478.

Photogrammetric Engineering & Remote Sensing
Vol. 68, No. 6, June 2002, pp. 589–596.

0099-1112/02/6806–589$3.00/0

described involving independent clustering and classification of individual scenes followed by classification "compositing" to create a final large-area product. Within this methodology, overlapping classifications are compared and employed in two ways:

- *Classification Error Checking.* By comparing the consistency of a scene classification with those of its overlapping neighbors, clusters that are likely to be mislabeled can be identified.
- *Accuracy Characterization.* An indirect measure of classification confidence can be derived for each cluster in each scene based upon classification consistency of its member pixels with neighboring scenes. This measure can then be used during the compositing process both to rationalize conflicting classifications in overlap regions as well as to generate a net classification "confidence" for each pixel in the final land-cover product, thereby encapsulating intra-product quality variations. It should be noted that this form of "relative" accuracy assessment is complementary to conventional comparison with "ground truth" (i.e., "absolute" accuracy assessment). The former has the added advantage that overlap regions can constitute a significant portion of a mapped area, especially at high latitudes, while ground-truth sampling tends to be sparse. Others have suggested using overlap regions for accuracy characterization, not for classification but rather for landscape metric estimation (Brown *et al.*, 2000).

The research presented here was undertaken as part of a joint Canada Centre for Remote Sensing (CCRS) United States Environmental Protection Agency (USEPA) effort to generate and interpret multi-temporal synoptic land-cover maps of the Great Lakes watershed derived from combined NALC and Canadian-processed Landsat MSS imagery. The goal is a spatially consistent classification involving six broad classes (water, forest, agriculture, urban/developed, natural grasslands, and barren). A conventional classification approach is being employed involving scene-based unsupervised clustering followed by interactive cluster labeling.

In the next two sections we discuss the broad accuracy issues of cluster-based labeling and introduce the concept of classification consistency analysis in inter-scene overlap regions as a way to characterize accuracy at the cluster level. This will be followed by the development of specific accuracy assessment methodologies to support label error identification, compositing of scene-based classifications to generate large area products, and the creation of a corresponding confidence layer. The approach was applied to the generation of a land-cover product of the Great Lakes watershed from 46 composited scenes. A simple stratification of the product (forest vs. non-forest land) is analyzed within the context of a simple statistical model. Finally, we briefly describe how consistency analysis can be adapted to mosaic-based classification.

Relevant Aspects of Cluster-Based Classification

There are a number of characteristics of cluster-based labeling that have consistency and accuracy implications:

- Unlike "textbook" examples of clustering, pixels rarely aggregate into distinct clusters in spectral space that correspond to the classes of interest. This leads to difficulties in determining the number of relevant clusters of a data set. For this reason, the number of clusters is usually set much higher than the number of classes sought (e.g., Vogelmann *et al.*, 1998).
- Each class typically is represented by a number of clusters. The "classification" qualities of these clusters can be expected to vary. Intuitively, one expects clusters residing far from intra-class transition zones in spectral space to contain far fewer misclassified pixels than those near such zones As a result, an accuracy confidence measure is needed at both the cluster and the class levels.

To better understand the labeling scenarios that can arise in the case of clusters, consider the simple case of two classes, Class A and Class B. Of the clusters labeled as Class A, we identify four major types based upon their pixel contents and labeling results.

Type a. Clusters that consist predominately of "pure" "true" Class A pixels that have been correctly labeled as Class A.
Type b. Clusters that consist predominately of "pure" "true" Class B pixels that have been incorrectly labeled as Class A.
Type c. Clusters that consist primarily of spectrally "mixed" pixels that have been labeled as Class A.
Type d. Clusters that contain significant numbers of both "pure" "true" Class A and "pure" "true" Class B pixels. These clusters are located in a portion of spectral space where classes A and B are spectrally indistinct. For example, the spectral signatures of urban areas and fallow agricultural fields will be similar within the restricted spectral space of Landsat MSS imagery.

Now consider the case where two partially overlapping scenes are separately clustered and labeled. These classifications will exhibit a degree of independence because (1) the scenes may have been acquired at different dates (as with Landsat scenes from adjacent tracks) and (2) the cluster characterizations will be different because they are based upon different overall parent pixel populations. As a result, the pixels constituting a single cluster of one scene are expected to be distributed among a number of clusters in the other scene. Classification consistency analyses should provide insights into the likely labeling type of those clusters with significant representation in the overlap region and hence point to those clusters whose labels require further review.

Label Error Checking

A three-step methodology is used for assigning a classification "category" to each scene-based cluster where the category indicates the confidence in the labeling result. This categorization is then used iteratively to improve classification consistency across scene boundaries.

Step 1

Because consistency checking is a relative process, it will be increasingly effective with increasing and comparable levels of producer accuracies of the two scenes. If this is the case, clusters with suspect labels should be identifiable through an "outlier" analysis process. This prerequisite condition can be assessed by generating a contingency table of label agreement at the pixel level whose $(I,J)^{th}$ element represents the number of pixels in the overlap region that have been assigned label I in one scene (Scene #1) and label J in the other (Scene #2).

Step 2

In this step we analyze the level of classification agreement at the cluster level. For example, consider some cluster α, in Scene #1, that is labeled as Class A. For pixels of cluster α located in the overlap region, we generate a summary of their corresponding assigned labels in Scene #2. We can assign cluster α to one of three possible categories.

Category 1. A high proportion of the pixels of cluster α have the same label (Class A) in Scene #2. We have added confidence that the pixels of cluster α are correctly labeled, i.e., that cluster α is of Type a described in the previous section.

Category 2. A high portion of the pixels of cluster α have been labeled as Class B in Scene #2 where $B \neq A$. A significant inconsistency is present, indicating that cluster α may be of Type b, (i.e., it may be mislabeled and hence requires further scrutiny).

Category 3. Significant portions of the cluster α population are assigned to each of two or more classes in Scene #2. This inconsistent labeling result suggests that cluster α may be of

either Type c or Type d, but the resolution of this ambiguity requires further information.

To employ the above categorization, one needs to define two "proportion-of-agreement" thresholds; an upper one, T_U, to delineate Category 1 clusters and a lower threshold, T_L, to delineate Category 2 clusters. Those clusters with agreement proportions between these thresholds will belong to Category 3.

If we define F_A to be the overall fraction of pixels of Class A in Scene #1 that have the same label in Scene #2, then as a first approximation we could use F_A as the upper threshold to test for all clusters. A drawback of this simple approach is that it does not take into account uncertainties in proportion estimation associated with unequal cluster populations. To account for cluster size variations, we derive cluster-specific threshold values based on F_A and binomial theory confidence estimation (Thomas and Allcock, 1984). For example, consider the case of a cluster, α, labeled A in Scene #1 with a population of N_α pixels in the overlap region of which a proportion, F_α, has the same label in Scene #2. We assign the cluster to Category 1 if

$$F_\alpha > T_U = F_A - \Delta F_A \tag{1}$$

where ΔF_A allows for a statistical uncertainty in an estimate of F_A given a sample size of N_α. We set this parameter to the 99.9 percent confidence interval.

$$\Delta F_A = [3s(1 + 1\sqrt{(N_\alpha)} + 1/\sqrt{(2N_\alpha)})]/N_\alpha \tag{2}$$

where $s = \sqrt{(N_\alpha F_A Q_A)}$ and $Q_A = 1 - F_A$, the proportion of disagreement.

The lower threshold can be based on the overall proportion of disagreement for pixels, Q_A, of Class A (i.e., $Q_A = (1 - F_A)$). In this case, if the observed proportion of agreement satisfies the inequality

$$F_\alpha < T_L = Q_A - \Delta F_A, \tag{3}$$

then the cluster is assigned to Category 2.

All clusters not satisfying conditions expressed by either Equation 1 or Equation 3 are assigned to Category 3.

Step 3

Once each cluster in Scene #1 has been given a provisional categorization, clusters other than those of Category 1 should be reviewed, resulting in possible re-labeling or cluster splitting actions. Note that the same process is applied to Scene #2 clusters. Following review, the categorization process can be repeated and in this way the most consistent, iterative solution can be reached.

In practice, a Landsat scene will exhibit significant overlap with up to four other scenes (i.e., two cross-track and two along-track neighbors). When assigning a category to a cluster, we compute the above statistics based on the aggregate level of agreement over all overlap regions.

Classification Compositing

Once final classifications for each scene have been achieved, their fusion into a final seamless classification mosaic can be undertaken through a compositing process. To proceed, one requires a classification layer and a "confidence" layer for each scene. The confidence layer should provide an estimate of the classification quality of each pixel. As discussed earlier, absolute accuracy estimation is not feasible at this level of detail. Instead, an estimate of confidence is computed based upon inter-scene classification consistency. For a given scene, this involves tabulating the aggregate fractional agreements in classification for each of its clusters with the classifications of its available overlapping neighbors. To each pixel in the scene, a measure of classification confidence is assigned that is proportional to the level of agreement of its parent cluster, thereby creating a scene "confidence" layer based upon consistency. The statistical relationship between consistency and inherent scene classification accuracy will be developed in the next section. In the compositing process, confidence is "accumulated" at the product pixel level as new scenes are added. The final composited product will then also contain two layers, namely, the final classification and the accumulated confidence layer that reflects, at the pixel level, both the number of available scene classifications and their levels of agreement. The following compositing algorithm is proposed:

> Let L(x,y) be the current classification label of the composite at location (x,y), C(x,y) be the current accumulated confidence for the current label, sl(x,y) be the classification label at location (x,y) of a new scene, and sc(x,y) be the confidence value at location (x,y) of the new scene.

Case 1: L(x,y) = 0, i.e., the composite has no current label. This can arise either if the location has had no coverage from previously composited scenes or if all earlier scenes were corrupted by cloud or cloud shadow at location (x,y). Then

$$L(x,y) = sl(x,y) \text{ and} \tag{4}$$

$$C(x,y) = sc(x,y). \tag{5}$$

Case 2: *L(x,y) > 0 and L(x,y) = sl(x,y), i.e., the current composite has a label which is the same as that as the scene being added.* Then

$$L(x,y) \text{ is unchanged and} \tag{6}$$

$$C(x,y) = C(x,y) + sc(x,y),$$

i.e., the confidence is increased by the confidence level of the scene.

Case 3: *L(x,y) > 0 but L(x,y) ≠ sl(x,y), i.e., there is a conflict* between the current composite classification and the classification of the new scene. This leads to three sub-cases.

Case 3a: If C(x,y) > sc(x,y), i.e., the accumulated confidence of the current composite exceeds the confidence level of the new scene, then

$$L(x,y) \text{ is unchanged and} \tag{7}$$

$$C(x,y) = C(x,y) - sc(x,y).$$

Case 3b: If C(x,y) < sc(x,y), i.e., the confidence level of the new scene exceeds that of the accumulated confidence of the current composite classification, then

$$L(x,y) = sl(x,y) \text{ and} \tag{8}$$

$$C(x,y) = sc(x,y) - C(x,y). \tag{9}$$

Case 3c: If C(x,y) = sc(x,y), i.e., there is equal confidence supporting each of the conflicting classifications. In this case, one must utilize additional information to resolve the conflict. In our implementation, a set of heuristics, based on the distribution of classifications in a 3 by 3 window centered on (x,y) in both the composite and the scene, is utilized. First, if either L(x,y) or sl(x,y) differs from all of its eight neighbors, its label is discarded, thereby reducing "salt and pepper" effects in the final composite. If this occurrence is not present, the label is selected that is in best agreement with its eight neighbors. Once a class label is selected, the composite confidence is set to zero.

The compositing methodology has a limitation for those product locations, (x,y) where three or more scenes provide

classification estimations. Because, in our current implementation, scenes are added sequentially to the composite, classification conflict resolution will only be independent of the order of scene entry if it involves a conflict between two classes. This risk has been deemed acceptable given that for Landsat the proportion of area covered by three or more scenes is small and that most realistic classification confusions arise between class pairs. However, it is emphasized that this problem is an implementation issue that can be overcome through the use of additional temporary storage layers that allow for the simultaneous comparison of all classification candidates at a given (x,y).

Great Lakes Example

A rudimentary land-cover product of the Great Lakes watershed, *in which* land has been categorized into two classes (forest and non-forest), is used to illustrate key issues, in particular, accuracy impacts related to spatial variations in class proportions. The Great Lakes product is sampled at 3 arc seconds in longitude by 2 arc seconds in latitude (approximately 70 meters) and has been created through the independent classification and compositing of 46 scenes using the algorithms described in previous sections. During this process, each scene was partitioned into 150 clusters using the K-means algorithm, and the clusters were assigned one of five labels (water, forest, non-forest, cloud, or cloud shadow). Because land-water confusion is low and the cloud-related classes can be considered cases of "no data," we have restricted our analysis to the classification consistency of the two land categories. A total of 76 cross-track overlap regions were used in the analysis. Figure 1 illustrates a portion of the product, centered in northern Michigan and containing contributions from approximately 20 scenes.

Statistical Model for a Two-Class Consistency

Consider the case of a classification scenario involving two classes, A and B, and an overlap region that has been independently classified in two scenes. Furthermore, let the numbers of true Class A and Class B pixels in the region be N_A and N_B, respectively. One can now proceed to generate a contingency table summarizing the consistency of the two classifications. The numbers of pixels classified as A and B in Scene #1 will be given respectively by

$$M_A = N_A p_A + N_B(1 - p_B) \tag{10}$$

and

$$M_B = N_B p_B + N_A(1 - p_A) \tag{11}$$

where p_A and p_B are the probabilities of correction classification of true Class A and Class B pixels, respectively (i.e., producers accuracies). In each equation, the first term indicates the number of correctly classified pixels while the second term represents the number of commission errors.

If these classified pixels are compared to the corresponding classification in Scene #2 that exhibits probabilities of correct classification of q_A and q_B, one can formulate the four elements of the two-way contingency table. For example, of the pixels classified as A in Scene #1, the numbers classified as A and B in Scene #2 will be equal to

$$M_{AA} = N_A p_A q_A + N_B(1 - p_B)(1 - q_B) \tag{12}$$

and

$$M_{AB} = N_A p_A(1 - q_A) + N_B q_B(1 - p_B), \tag{13}$$

respectively. Similarly, for those pixels classified as B by Scene #1, the numbers classified as B and A, i.e., M_{BB} and M_{BA}, can be readily estimated from the above two equations by reversing the subscripts of A and B.

The number of true Class A and Class B pixels (i.e., N_A and N_B) are unknown. However, one can derive a number of statistical measures that can be estimated from the population contingency tables extracted from real scene overlap regions. This comparison between theory and observation is done using only overlap regions between adjacent-track scene pairs in order to ensure maximum independence of the classifications (76 cases). In addition, some further observations lead to a simplified formulation. First, two broad factors will affect the relative sizes of the entries in a contingency table, i.e., (1) the individual scene producer accuracies as indicated by the probabilities of correct classification and (2) the relative proportions of true class pixels. Given the broad classes (forest vs. non-forest) in the Great Lakes example, one would expect consistently high producer accuracies for most scenes. On the other hand, the relative proportion of forest to non-forest land varies dramatically from approximately 1:10 in the south to 10:1 in the north and, hence, variations in user accuracy will be dominated by the proportional factor. The Great Lakes example provides a good opportunity to study this factor, which is typically difficult to assess from conventional confusion matrices.

From the above arguments, the model formulation can be simplified by replacing all probabilities by a single unknown probability p. As a result, the elements of the contingency table become

$$M_{AA} = N_A p^2 + N_B(1 - p)^2, \tag{14}$$

$$M_{AB} = M_{BA} = (N_A + N_B)p(1 - p), \text{ and} \tag{15}$$

$$M_{BB} = N_B p^2 + N_A(1 - p)^2. \tag{16}$$

Observable Statistical Measures

Below we list a number of statistical measures that can be extracted from overlap contingency tables of real scenes.

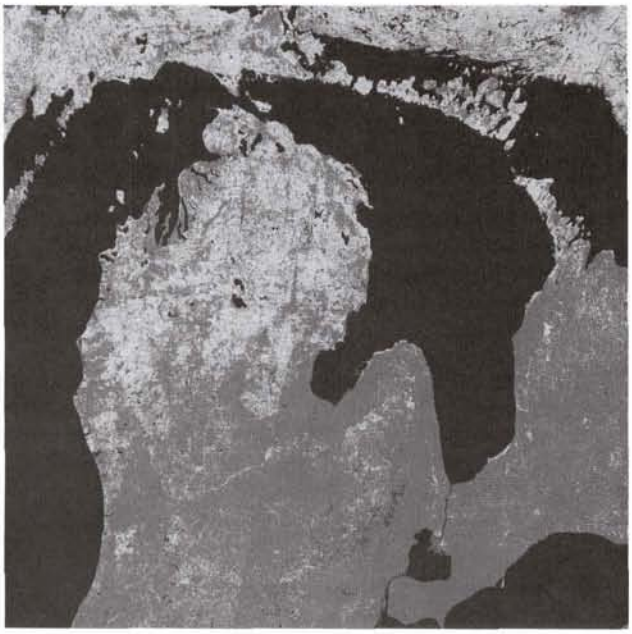

Figure 1. A portion of the Great Lakes land-cover product centered on the region of northern Michigan. The classes include water (dark), non-forest (medium grey), and forest (white). The area includes land-cover information from approximately 20 Landsat scenes.

F_A = fraction of those pixels classed as A in Scene #1 which are also classed as A Scene #2.
$= M_{AA}/(M_{AA} + M_{AB})$. (17)

Similarly, the fractional component of classification agreement for Class B will be

$$F_B = M_{BB}/(M_{BB} + M_{BA}). \quad (18)$$

E_A = estimated proportion of the overlap region that is of Class A based on the Scene #1 classification
$= (M_{AA} + M_{AB})/(M_{AA} + 2M_{AB} + M_{BB})$. (19)

Figure 2a illustrates the relationships of F_A versus F_B for selected producer accuracies, p, where Class A is set to forest (F) and Class B to non-forest (NF). Each curve is derived by varying the ratio of N_A to N_B from 0 to 1. To understand the predicted behavior, consider one example, namely, the curve for $p = 0.9$. If $N_A = 0$, then all pixels classified as A will be commission errors, (i.e., 10 percent of N_B), and, hence, the classification agreement between scenes for this class will be low. On the other hand, 90 percent of the Class B pixels will be correctly classified in each scene, and of these, 90 percent will be in agreement between scenes or 81 percent of the complete N_B sample. When all of the overlap pixels are of true Class A, the roles of the classes are reversed. As the fraction of true Class A pixels in the overlap region gradually increases from 0 toward 1, F_A increases. This occurs because the proportions of pixels classed as A in each scene that are correct increase, thereby increasing the probability

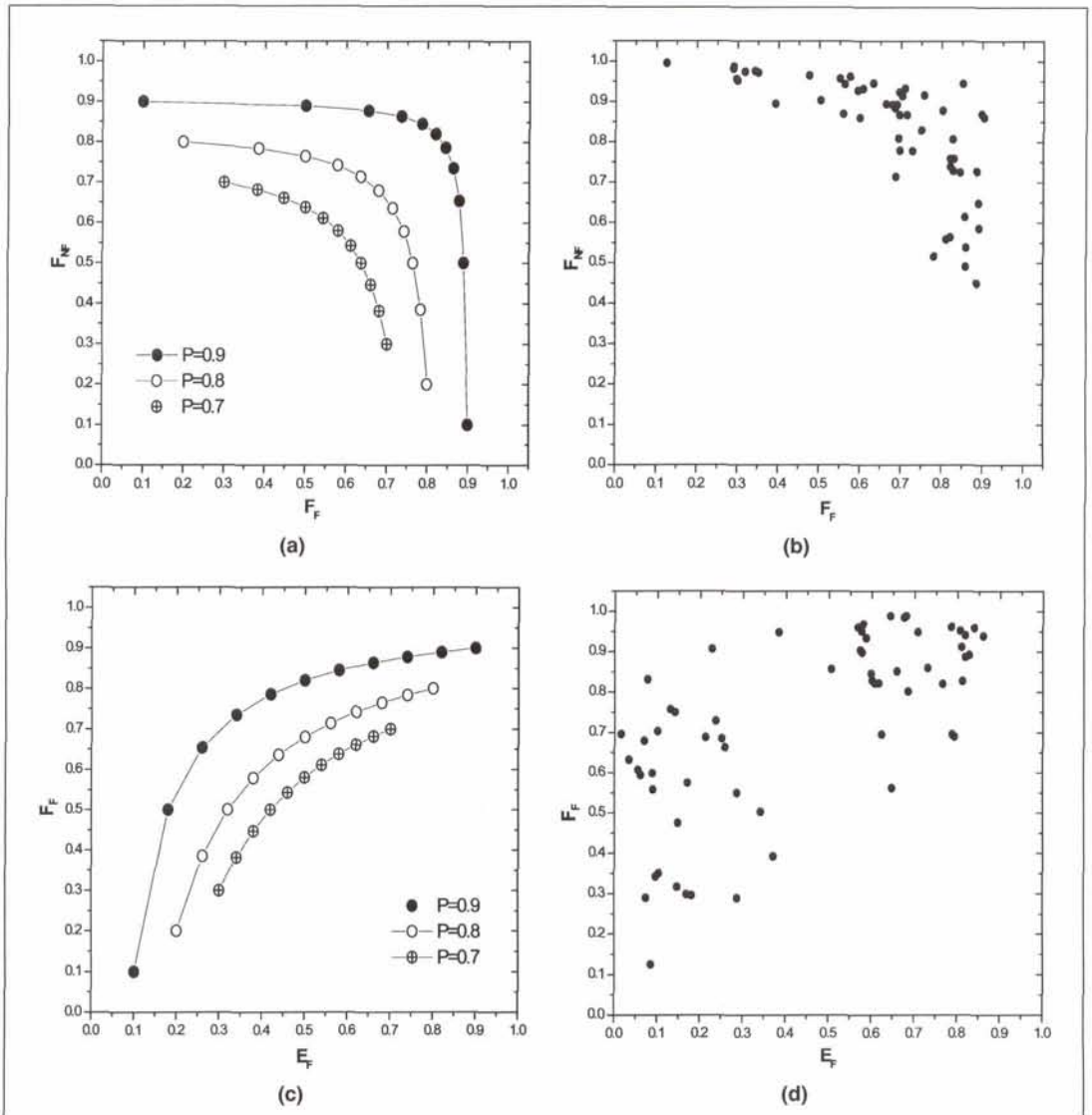

Figure 2. Comparison of predicted classification consistency parameters for a two-class statistical model with observed forest (F) vs non-forest (NF) classification results for 76 overlap regions of the Great Lakes data set. F_F is the fractional agreement in forest classification, F_{NF} is the fractional agreement in non-forest classification, and E_F is the estimated proportion of forest based on the classification of one scene. Model predictions for three producer accuracy levels (0.7, 0.8, and 0.9) are shown in plots (a) and (c) with corresponding observed data in (b) and (d), respectively. The majority of overlap regions in the Great Lakes data set are consistent with a producer accuracy of approximately 0.9.

TABLE 1. SUMMARY OF THE IMPACT OF CLASSIFICATION COMPOSITING ON USER ACCURACY FOR THE TWO-CLASS SCENARIO (ASSUMED PRODUCER ACCURACY OF 0.9) AS A FUNCTION OF THE FRACTIONAL REPRESENTATION OF CLASS A, I.E., $N_A/(N_A + N_B)$. FOR A CLASS WITH LOW TRUE PROPORTIONAL REPRESENTATION, I.E., CLASS A, COMPOSITING CAN SIGNIFICANTLY ENHANCE USER ACCURACY (FC_A) OVER A SINGLE IMAGE CLASSIFICATION (FS_A) THROUGH THE REDUCTION OF RANDOM COMMISSION ERRORS. THIS IS ACHIEVED AT THE EXPENSE OF A CORRESPONDING MODEST REDUCTION IN USER ACCURACY OF THE DOMINANT CLASS B (i.e., $FC_B < FS_B$)

$N_A/(N_A + N_B)$	FC_A	FS_A	FC_B	FS_B
0.1	0.9	0.5	0.98	0.99
0.2	0.95	0.69	0.95	0.97
0.3	0.97	0.79	0.92	0.96
0.4	0.98	0.86	0.89	0.93

of coincident A classification. Similarly, F_B will decrease because of the increasing importance of Class B commission errors. Finally, it can be seen that the ranges of the F values decrease as p decreases. An F value can never be less than $(1 - p)$ nor larger than p. In the extreme situation of random classification, i.e., $p = 0.5$, both F_A and F_B will always be equal to 0.5 no matter what the true class proportions are in the overlap regions.

Figure 2c illustrates the relationships for F_A versus E_A for the same values of p. It can be seen that E_A is restricted by the same bounds as F_A, again due to the limiting effects of commission errors.

Figures 2b and 2d show the scatter plots of F_A vs F_B and F_A vs E_A for the 76 Great Lakes cross-track overlap regions. From a comparison with the theoretical plots, the following conclusions are drawn:

- Comparing Figures 2a and 2b, we observe that the observed Great Lakes population is consistent with a model of relatively high, consistent scene classification rate, of approximately $p = 0.9$. Even with this high producer accuracy, agreement levels between classifications can be low ($\ll 0.5$). These cases arise when fractional class coverage is low (e.g., sparse forest cover in the south), and result from a preponderance of commission errors.
- Comparing Figures 2c and 2d, it can be seen that, while a small number of overlap regions, about 10 percent of the sample, can be explained by values of p below 0.8, most cases of low apparent classification agreement between scenes (i.e., low values of F) are consistent with the effects of dominating commission errors in regions where the land is primarily non-forest.
- For a region as diverse as the Great Lakes watershed, spatially variations in land-cover proportions will result in significant intra-product spatial variations in user classification accuracy, even for rudimentary classes, if land cover is derived from a single classification.

Compositing Model

In this subsection, the impact of classification compositing within the two-class model is discussed. In practice, classification compositing is based upon confidence comparisons at the cluster level derived from multiple overlap regions. Here, however, it is dealt with at a simpler level, i.e., the case of two scenes with a single overlap region in which further subdivision associated with multiple clusters per class is ignored.

In the compositing process, conflicting classifications are rationalized by comparing measures of "confidence" of the two competing interpretations. For the compositing algorithm outlined above, confidence for a class in Scene #1 is defined to be proportional to the level of agreement in classification with Scene #2, i.e., this confidence for Class A in Scene #1 is proportional to F_A. In the two-class case, the relationship of F_A versus F_B is symmetric about its midpoint (i.e., the point where N_A equals N_B). As a result, we need only deal with the case of

$$N_A < N_B, \text{ hence } F_A < F_B. \tag{20}$$

Only those pixels which were classed as A in both scenes will be composited as Class A, i.e.,

$$\text{number of pixels assigned to Class A} = N_A p^2 + N_B (1 - p)^2. \tag{21}$$

Of these pixels, the fraction that are true Class A pixels will be given by

$$FC_A = N_A p^2/(N_A p^2 + N_B(1 - p)^2). \tag{22}$$

The impact of compositing can be assessed by comparing FC_A to the fraction, FS_A, of true Class A pixels identified if only a single Scene #1 were used, where

$$FS_A = N_A p/(N_A p + N_B(1 - p)). \tag{23}$$

Turning to Class B, the number of pixels assigned to this class following compositing will be the total of those pixels classed as B in either both or only one of the scenes, i.e.,

$$\text{Pixels assigned to Class B} = M_{BB} + M_{BA} + M_{AB} = N_B(2p - p^2) + N_A(1 - p^2). \tag{24}$$

As with class A, one can compute the corresponding fractions of correctly classified pixels with and without compositing, i.e.,

$$FC_B = N_B(2p - p^2)/(N_B(2p - p^2) + N_A(1 - p^2)) \tag{25}$$

and

$$FS_B = N_B p/(N_B p + N_A(1 - p)). \tag{26}$$

Table 1 contains a summary of the variations of these fractional parameters as a function of the true fractional proportion of Class A, $N_A/(N_A + N_B)$, again for the example case of $p = 0.9$. The implications of compositing can be summarized as follows:

- In the case of Class A, compositing tends to reduce random commission errors, resulting in a greatly "purified" final Class A population, i.e., an improved user accuracy. This is especially important when $N_A \ll N_B$.
- The purification of Class A occurs at the expense of identifying a smaller portion of the true Class A population, i.e., a reduced producer accuracy. The numbers of true Class A pixels found through compositing versus single scene classification are $N_A p^2$ and $N_A p$, respectively, or 81 percent versus 90 percent for our case of $p = 0.9$.
- In the case of the dominant Class B, the effects are reversed but less dramatic. Compositing results in an improved recovery rate of true Class B pixels, e.g., 99 percent for $p = 0.9$ versus 90 percent without compositing. On the other hand, the final population of pixels classed as B exhibits a marginally higher proportion of commission errors.

Model Extension

While the above statistical model dealt with a simple two-class case, it can provide insights into more complex cases.

More Than Two Classes

Increasing the number of classes to m potentially leads to n-way class interactions where $3 \leq n \leq m$. However, in some practical cases, confusion may still exist at the pair-wise level (e.g., grass and row crop confusion (Vogelmann *et al.*, 1998; Zhu *et al.*,

2000)). In such cases, the above formulations can be directly applied. In situations where a class is confused with a collection of classes (e.g., "mixed" forest with pure deciduous, conifers, and schrubland classes), a first level of understanding can be achieved by treating these confusing classes as a single "aggregate" class, i.e., an aggregate "Class B."

Multiple Clusters per Class

Typically, in cluster-based classification, more than one cluster will be assigned the same class label. In our classification compositing methodology, consistency measures (e.g., F values) are estimated for each cluster. The above theory can therefore be applied at the cluster level as well because the measure of confidence used in compositing is proportional to the F values developed in this section. The cumulative impacts of factors such as commission errors can be assessed at the class level by weighting the effects of constituent clusters by their relative pixel populations.

Application to Mosaic-Based Classification

As a final point, it should be noted that the methodologies discussed here may also be retro-fitted to model the spatial consistency of classifications based on multi-scene image mosaics. Two levels of retro-fitting are shown in Figure 3. In the traditional mosiac-based classification, individual scenes are first radiometrically normalized to a common standard; then the scenes are merged into an image mosaic. The mosaic is then treated as a single image entity, i.e., it is clustered and the clusters are labeled. The labeling process in many cases may be complex. In the case of the NLCD, the initial 100 clusters of each regional mosaic represented only a "first cut" at data partitioning. Extensive use was made of auxiliary data and additional imagery to achieve final a "clustering" and labeling of pixels.

At the first level of modification, consistency checking can be accomplished by applying classifying each of the parent (radiometrically normalized) scenes using the final cluster descriptors and labels derived from the mosaic. Inter-scene overlap analyses can then be undertaken, leading to the estimation of a confidence measure for each cluster and the creation of a confidence layer for the original mosaic. At the second level, the labeled mosaic is replaced with a fully composited product, including a layer of accumulated confidence. This level is attractive because it combines the manual efficiency of mosaic processing (i.e., only one set of clusters, derived from the mosaic, need be labeled), while at the same time exploiting the full available image data set, including all overlaps.

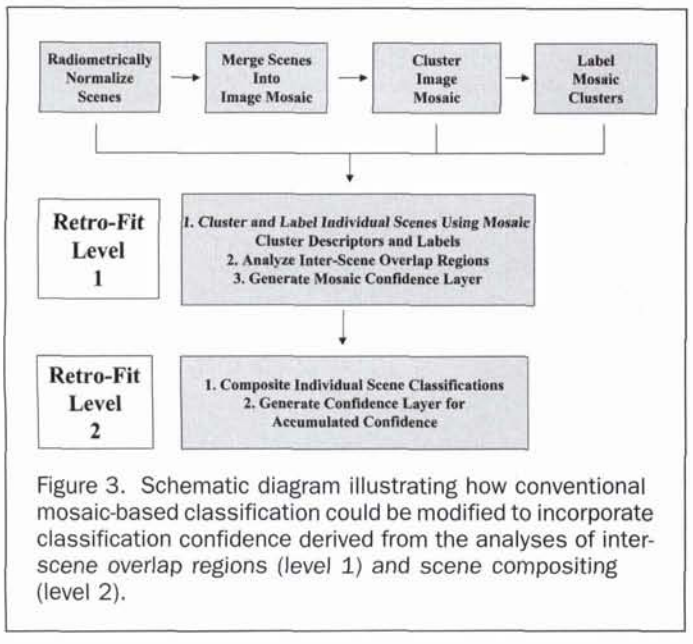

Figure 3. Schematic diagram illustrating how conventional mosaic-based classification could be modified to incorporate classification confidence derived from the analyses of inter-scene overlap regions (level 1) and scene compositing (level 2).

Conclusions

Large-area land-cover mapping based upon Landsat archival imagery involves the integration of derived information from scenes that exhibit diverse seasonal and atmospheric conditions and significant overlap coverage. Traditionally, scenes have been combined into image mosaics prior to classification, thereby eliminating multiple coverage. It is argued that redundant coverage has the potential both to improve classification performance and to characterize spatial variations in quality of the final land-cover product. An alternate land-cover mapping approach has been developed which involves classification at the scene level followed by the integration of these independently classified entities. Within the context of this approach, analyses of the classification consistency in overlap regions can be used to identify mislabeled clusters and to model classification confidence at the cluster level. Finally, the same categorization can be employed to rationalize conflicting scene classifications and to generate an overlay of comparative classification confidence during the step of integrating scene classifications into a final seamless land-cover product.

This methodology has been employed to generate a synoptic land cover product of the Great Lakes watershed. An analysis of a simple two-class case (forest vs non-forest land) provides a number of useful insights into accuracy issues and implications for large-area mapping.

- If the region to be mapped exhibits significant regional variations in thematic class proportions, significant intra-product variations in user accuracy can occur even if consistent producer accuracy is maintained. This has important implications for the subsequent use of such products because many, such as change-detection analysis, are dependent on high user accuracy.
- Intra-product accuracy variations imply that accuracy modeling is desirable at a spatial detail that would be difficult and prohibitively expensive to achieve through conventional ground-truth comparison. It is argued that the level of classification agreement in overlap regions provides an indirect but complementary confidence measure that can be assessed on a cluster basis and independently applied on each pixel.
- Improvement in user accuracy, particularly for classes of low areal coverage, can be achieved by employing multiple classification estimates because random commission errors present in each classification are reduced during a compositing process. Because parent scene data sets typically include extensive overlap coverage, this should be exploited rather than eliminating it through mosaic creation and classification.
- Even if overlap regions are exploited, desired levels of user accuracy may not be achievable without resorting to further redundancy (e.g., the use of complementary "leaf-off" and "leaf-on" image pairs, as in the case of the NLCD). The proposed compositing methodology lends itself to this scenario.
- The compositing process includes the generation of a cumulative confidence layer for each land-cover product. This layer provides an important ancillary information source for post-processing activities that involve comparison of multiple land-cover products, for example, for change detection. Confidence can be used to assess the statistical significance of observed changes, again at the pixel level.
- There are a number of pros and cons of mosaic- versus scene-based classification in the creation of large-area products. Because labeling is a labor-intensive process, the mosaic approach is attractive because cost and timeliness are usually important issues. On the other hand, mosaics have the added complexity of temporal variability, which in turn can result in added cluster mixing (i.e., more Type c and Type d clusters) and class confusion. A detailed performance comparison goes beyond the scope of this paper. Finally, while the compositing

methodology presented here has been illustrated within the context of independent scene-based classification, many of its concepts can be fully integrated within the framework of a mosaic-based mapping strategy, leading to the creation of additional accuracy characterization based on consistency analyses.

Acknowledgments

The authors acknowledge the valuable comments provided by journal reviewers that have led to numerous improvements in the final manuscript. Also, the authors wish to thank Dr. Ying Zhang for support in preparing the figures.

References

Brown, D.G., J.-D. Duh, and S.A. Drzyzga, 2000. Estimating Error in an Analysis of Forest Fragmentation Change Using North American Landscape Charcterization (NALC) Data, *Remote Sensing of Environment*, 71:106–117.

Congalton, R.G., and K. Green, 1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Lewis Publishers, Boca Raton, Florida, Chapters 5 and 6.

Dykstra, J.D., M.C. Place, and R.A. Mitchell, 2000. GEOCOVER-ORTHO: Creation of a Seamless, Geodetically Accurate, Digital Base Map of the Entire Earth's Land Mass Using Landsat Multispectral Data, *Proceedings of the ASPRS 2000 Conference*, 22–26 May, Washington, D.C., 7 pages.

Homer, C.G., R.D. Ramsey, T.C. Edwards, and A. Falconer, 1997. Landscape Cover-Type Modeling Using a Multi-Scene Thematic Mapper Mosaic, *Photogrammetric Engineering & Remote Sensing*, 63:59–67.

Loveland, T.R., and D.M. Shaw, 1996. Multiresolution Land Characterization: Building Collaborative Partnerships, *GAP Analysis: A Landscape Approach to Biodiversity Planning, Proceedings of the ASPRS/GAP Symposium* (J.M. Scott, T. Tear, and F. Davis, editors), Charlotte, North Carolina (National Biological Service, Moscow, Idaho), pp. 83–89.

Lunetta, R., J.G. Lyon, B. Guindon, and C.D. Elvidge, 1998. North American Landscape Characterization Dataset Developments and Data Fusion Issues, *Photogrammetric Engineering & Remote Sensing*, 64:821–829.

Sohl, T.L., and J.L. Dwyer, 1998. North American Landscape Characterization Project: The Production of a Continental Scale Three-Decade Landsat Data Set, *Geocarto International*, 13:43–51.

Thomas, I.L., and C.M. Allcock, 1984. Determining the Confidence Level for a Classification, *Photogrammetric Engineering & Remote Sensing*, 50:1491–1496.

Vogelmann, J.E., T. Sohl, and S.M. Howard, 1998. Regional Characterization of Land Cover Using Multiple Sources of Data, *Photogrammetric Engineering & Remote Sensing*, 64:45–57.

Vogelmann, J.E., S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and N. Van Driel, 2001. Completion of the 1990s National Land Cover Data Set for the Conterminous United States from Landsat Thematic Mapper Data and Ancillary Data Sources, *Photogrammetric Engineering & Remote Sensing*, 67:650–662.

Zhu, Z., L. Yang, S.V. Stehman, and R.L. Czaplewski, 2000. Accuracy Assessment for the U.S. Geological Survey Regional Land-Cover Mapping Program: New York and New Jersey Region, *Photogrammetric Engineering & Remote Sensing*, 66:1425–1435.

(Received 05 September 2001; accepted 24 October 2001; revised 12 November 2001)

Forthcoming Articles

D.H.A. Al-Khudhairy, C. Leemhuis, V. Hoffman, I.M. Shepherd, R. Calaon, J.R. Thompson, H. Gavin, D.L. Gasca-Tucker, G. Zalidis, G. Bilas, and *D. Papadimos,* Monitoring Wetland Ditch Water Levels Using Landsat TM and Ground-Based Measurements.

A.K. Chong, A Rigorous Technique for Forensic Measurement of Surveillance Video Footage.

A.K. Chong and *P. Stratford,* Underwater Digital Stereo-Observation Technique for Red Hydrocoral Study.

Arie Croitoru and *Yerach Doytsher,* Monocular Right-Angle Building Hypothesis Generation in Regularized Urban Areas by Pose Clustering.

Olivir Depeir, Isabelle Van den Steen, Patrice Latinne, Philippe Van Ham, and *Eléonore Wolff,* Textural and Contextual Land-Cover Classification Using Single and Multiple Classifier Systems.

J.R. Eastman and *R.M. Laney,* Bayesian Soft Classification for Sub-Pixel Analysis: A Critical Evaluation.

Jeanne Epstein, Karen Payne, and *Elizabeth Kramer,* Techniques for Mapping Suburban Sprawl.

Giles M. Foody, The Role of Soft Classification Techniques in the Refinement of Estimates of Ground Control Point Location.

Clive S. Fraser and *Harry B. Hanley,* Bias Compensation in Rational Functions for IKONOS Satellite Imagery.

B. Guindon and *C.M. Edmones,* Large-Area Land-Cover Mapping through Scene-Based Classification Compositing.

Jack T. Harvey Population Estimation Models Based on Individual TM Pixels.

Yong Hu and *C. Vincent Tao,* Updating Solutions of the Rational Function Model Using Additional Control Information.

Carl J. Legleiter, W. Andrew Marcus, and *Rick L. Lawrence,* Effects of Sensor Resolution on Mapping In-Stream Habitats.

Yan Li and *Jia-Xiong Peng,* Remote Sensing Texture Analysis Using Multi-Scale and Multi-Parameter Features.

Zhilin Li, Xiuxiao Yuan, and *Kent W.K. Lam,* Effects of JPEG Compression on the Accuracy of Photogrammetric Point Determination.

Hans-Gerd Maas, Methods for Measuring Height and Planimetry Discrepancies in Airborne Laserscanner Data.

Jill Maeder, Sunil Narumalani, Donald C. Rundquist, Richard L. Perk, John Schalles, Kevin Hutchins, and *Jennifer Keck,* Classifying and Mapping General Coral-Reef Structure Using IKONOS Data.

Assefa M. Melesse and *Jonathan D. Jordan,* A Comparison of Fuzzy ve. Augmented-ISODATA Classification Algorithms for Cloud-Shadow Discrimination from Landsat Images.

Boniface O. Oindo, Predicting Mammal Species Richness and Abundance Using Multitemporal NDVI.

Åsa Persson, Johan Holmgren, and *Ulf Söderman,* Detecting and Measuring Individual Trees Using an Airborne Laser Scanner.

Jiann-Yeou Rau and *Liang-Chien Chen,* True Orthophoto Generation of Built-Up Areas Using Multi-View Images.