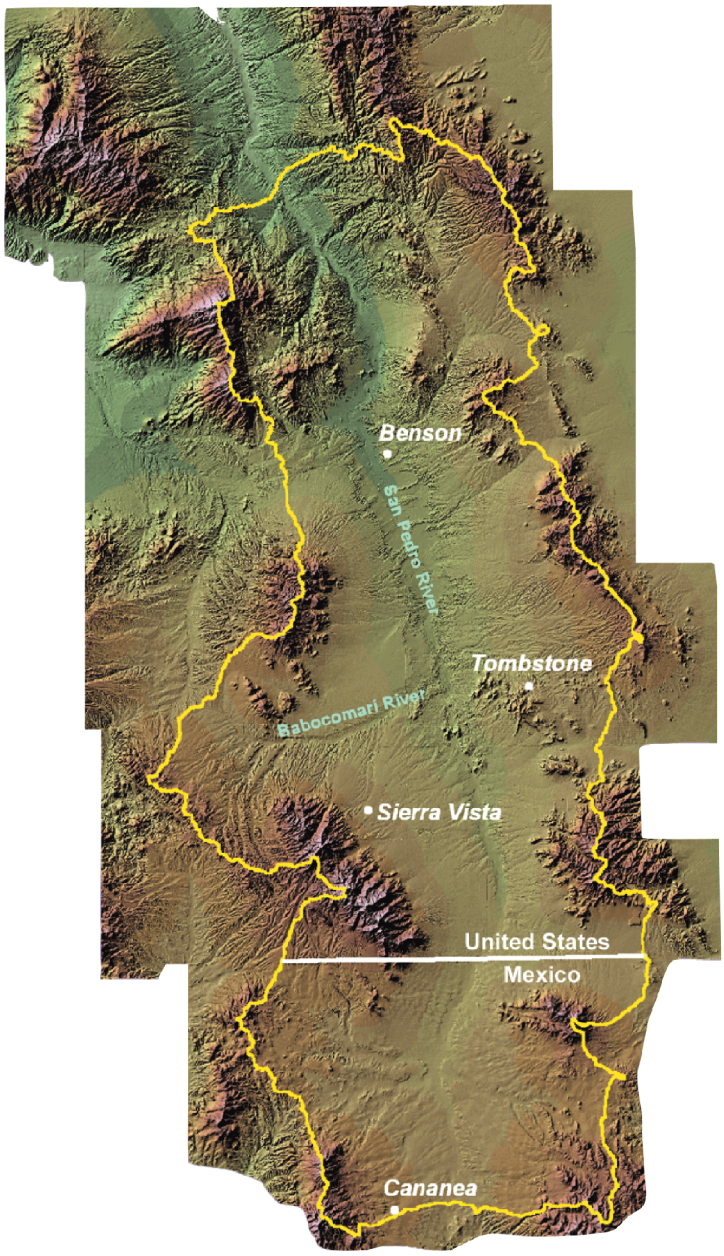
United States
Environmental Protection
Agency

An Accuracy Assessment of 1992 Landsat-MSS Derived Land Cover for the Upper San Pedro Watershed (U.S./Mexico)



EPA/600/R-02/040
June 2002

An Accuracy Assessment of 1992 Landsat-MSS Derived Land Cover for the Upper San Pedro Watershed (U.S./Mexico)

by

John K. Maingi and Stuart E. Marsh
University of Arizona
Arizona Remote Sensing Center
Tucson, AZ

William G. Kepner and Curtis M. Edmonds
U.S. Environmental Protection Agency
National Exposure Research Laboratory
Las Vegas, NV

Acknowledgments

I gratefully acknowledge Dr. L. Dorsey Worthy, U.S. Environmental Protection Agency, National Exposure Research Laboratory, and John Lowry, Department of Geography and Earth Resources, Utah State University, for their helpful suggestions as reviewers for this report.

Notice

This report has been peer reviewed by the U.S. Environmental Protection Agency (EPA), through its Office of Research and Development (ORD) and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by EPA for use.

Executive Summary

The utility of Digital Orthophoto Quads (DOQs) in assessing the classification accuracy of land cover derived from Landsat MSS data was investigated. Initially, the suitability of DOQs in distinguishing between different land cover classes was assessed using high-resolution airborne color video data. A cross-tabulation of the analyst's DOQ labels and the reference video label was produced and had an overall accuracy of 92%. This indicated that the DOQ data could be used to identify and distinguish between the different land cover classes.

A 1992 land cover map for the Upper San Pedro Watershed was available for accuracy assessment. The map was interpreted and generated by Instituto del Medio Ambiente y el Desarrollo Sustentable del Estado de Sonora (IMADES), Hermosillo, Sonora. The Environmental Protection Agency (EPA) supplied Arizona Remote Sensing Center (ARSC) with approximately 60 DOQs for 1992. Most of the land cover classes were fairly well represented in the DOQs and covered between 24% and 41% in eight out of ten land cover classes. Only the Barren and Agriculture classes were poorly represented in the available DOQs covering 5.3% and 14.2% of the map area, respectively.

A total of 457 sample points was used for the accuracy assessment. Allocation of sample points to land cover classes was through stratified (by land cover class area) random sampling, with a 20-sample minimum for the smallest classes. Map labels for the sample points were compared with reference DOQ labels and an error matrix generated. An overall classification accuracy of about 75% was obtained.

Table of Contents

Page

Section 1 A Review – Land Cover Accuracy Assessment 1

Section 2 Sampling in Support of Accuracy Assessment 4

Section 3 Classification Accuracy Assessment Sampling Design for the San Pedro Watershed 7

Section 4 Methods 12

Section 5 Accuracy Assessment of the 1992 Land Cover Map 15

Conclusions 17

Appendix 1 Ground Control Points Used to Georeference the 1992 NALC Subset Image to a Precision Corrected 1997 Landsat TM Image 18

References 19

List of Figures

Figure Page

1 Location of the Upper San Pedro River Basin, Arizona/Sonora (Adapted from Kepner *et al.*, 2000) 8

2 Appearance of some land cover classes on 1992 digital orthophoto quadrangles 13

List of Tables

Table Page

1 Land cover class descriptions for the Upper San Pedro Watershed (Adapted from Kepner *et al*., 2000) 9

2 Minimum number of sample points, N1 and N2, required to achieve an allowable error of 5% (E1) and 2.5% (E2), respectively, at the 95% confidence interval and for accuracies ranging from 60% to 95% 10

3 Minimum number of sample points per land cover class stratified by area 11

4 Error matrix illustrating the analyst's ability to use the 1992 DOQs for land cover classification accuracy assessment. A summary of classification errors is appended below . . . 14

5 Classification accuracy error matrix for the 1992 land cover map using 1992 DOQs 15

Section 1

A Review – Land Cover Accuracy Assessment

Land cover maps derived from remotely sensed data inevitably contain error of various types and degrees. It is therefore very important that the nature of these errors be determined, in order for both users and producers of the maps to be able to gauge their appropriateness for specific uses. In addition, identifying and correcting the sources of errors may increase the quality of map information. Classification accuracy assessment is necessary for comparing the performance of various classification techniques, algorithms, or interpreters (Congalton and Green, 1998). Classification accuracy assessment is now recognized as a critical component of any mapping project.

Development of criteria and techniques for testing map accuracy began in the 1970s (Hord and Brooner, 1976; van Genderen and Lock, 1977; Ginevan, 1979). More in-depth studies and development of new techniques were initiated in the 1980s (Rosenfield *et al.*, 1982; Congalton and Mead, 1983; Aronoff, 1985). Today, the error matrix has become the standard medium for reporting the accuracy of maps derived from remotely sensed data (Congalton and Green, 1993). More recent research into classification accuracy assessment has focused on factors influencing the accuracy of spatial data, such as sampling scheme and sample size, classification scheme, and spatial autocorrelation (Congalton, 1991; Congalton and Green, 1993). Other important considerations in classification accuracy assessment include ground verification techniques, and evaluation of all sources of error in the spatial data set.

The accuracy of a classified image refers to the extent to which it agrees with a set of reference data. Most quantitative methods to assess classification accuracy involve an error matrix built from the two data sets (i.e., remotely sensed map classification and the reference data). An error matrix is a square array of numbers set out in rows and columns which express the number of sample units assigned to a particular category relative to the actual category or as verified on the ground or typically large scale (at least 1:12,000) color aerial photography (Congalton and Green, 1993). The columns normally represent the reference data, while the rows indicate the classification generated from the remotely sensed data. An error matrix is a very effective way to represent accuracy because the accuracy of each category is clearly described, along with both errors of inclusion (commission errors) and errors of exclusion (omission errors), as well as summary statistics for the entire matrix (Congalton *et al.*, 1983; Congalton, 1991, Ma and Redmond, 1995).

Overall map accuracy is computed by dividing the total correct (obtained by summing the major diagonal of the error matrix) by the total number of pixels in the error matrix. Accuracy of individual categories is computed by dividing the number of correct pixels in a category by either the total number of pixels in the corresponding row or the corresponding column (Congalton, 1991). When the number of correct pixels in a category is divided by the total number of pixels in the corresponding row (i.e., total number of pixels that were classified in that category), the result is an accuracy measure called “user’s accuracy,” and is a measure of commission error. “User’s accuracy” or reliability is indicative of the probability that a pixel classified on the map actually represents that category on the ground (Story and Congalton, 1986). On the other hand, when correct number of pixels in a category are divided by the total

number of pixels in the corresponding column (i.e., total number of pixels for that category in the reference data), the result is called "producer's accuracy." "Producer's accuracy" indicates the probability of a reference pixel being correctly classified and is really a measure of omission error.

An error matrix is an appropriate beginning for many analytical statistical techniques, especially discrete multivariate techniques. Discrete multivariate techniques are appropriate because remotely sensed data are discrete rather than continuous. The data are also binomially or multinomially distributed, and therefore, common normal theory statistical techniques do not apply (Jensen, 1996).

KAPPA is a discrete multivariate technique developed by Cohen (1960) and has been utilized for land cover and land use accuracy assessment derived from remotely sensed data (Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Gong and Howarth, 1990). The result of performing a KAPPA analysis is the KHAT statistic (an estimate of KAPPA) which is another measure of accuracy or agreement. Values of KAPPA greater than 0.75 indicate strong agreement beyond chance, values between 0.40 and 0.79 indicate fair to good, and values below 0.40 indicate poor agreement (SPSS, 1998). Overall accuracy uses only the main diagonal elements of the error matrix, and, as such, it is a relatively simple and intuitive measure of agreement. On the other hand, because it does not take into account the proportion of agreement between data sets that is due to chance alone, it tends to overestimate classification accuracy (Congalton and Mead, 1983; Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986; Ma and Redmond, 1995). KHAT accuracy has come into wide use because it attempts to control for chance agreement by incorporating the off-diagonal elements as a product of the row and column marginals of the error matrix (Cohen, 1960).

The KAPPA coefficient is a powerful tool because of its ability to provide information about a single matrix and as a means to statistically compare matrices (Congalton, 1991). The Kappa coefficient serves as a more rigorous estimate of accuracy considering agreement that may be expected to occur by chance. Verbyla (1995) gives a formula for computing KHAT:

$$\hat{K} = \frac{\text{Overall Classification Accuracy} - \text{Expected Classification Accuracy}}{1 - \text{Expected Classification Accuracy}}$$

The expected classification accuracy is the accuracy expected based upon chance alone or the expected accuracy if we randomly assigned class values to each pixel. It can be calculated by first using the error matrix to produce a matrix of the products of row and column totals. The expected classification accuracy is then computed as the sum of the diagonal cell values divided by the sum of all cell values (Verbyla, 1995).

However, Foody (1992) has shown that, without modifications, KAPPA overestimates the proportion of agreement due to chance, and underestimates the overall classification accuracy. For this reason, Foody (1992) proposed the use of a modified KAPPA statistic for use with classifications based on equal probability of group membership that resembles and is derived more properly from the Tau coefficient. Kendall's Tau is a measure of the association between two variables and is limited to the range [-1, +1]. A value near zero indicates that the values of one variable are uncorrelated with values of the other.

In follow-up research to Foody's findings, Ma and Redmond (1995) introduced the Tau coefficient, which measures the improvement of a classification over a random assignment of pixels to groups, and compared its performance to that of KAPPA and percentage agreement (overall accuracy). They found that Tau did better at adjusting percentage agreement than KAPPA, and that it was also easier to calculate and

interpret. They therefore recommended the Tau statistic as a better measure of classification accuracy for use with remote sensing data than either KAPPA or percentage agreement.

Other techniques for assessing the accuracy of remotely sensed data have been suggested. Aronoff (1985) suggested an approach based on the binomial distribution of data, which is very appropriate for remotely sensed data. This approach involves the use of a minimum accuracy value as an index of classification accuracy. The advantage of the index is that it expresses statistically the uncertainty involved in any accuracy assessment. The major disadvantage of the approach is that it is limited to a single overall accuracy value rather than using the entire error matrix.

Analysis of variance is another technique for accuracy assessment suggested by Rosenfield (1981). However, violation of normal theory assumption and independence assumption when applying this technique to remotely sensed data has severely limited its application (Congalton, 1991).

Section 2

Sampling in Support of Accuracy Assessment

The overriding assumption in the entire classification accuracy assessment procedure is that the error matrix is indicative or representative of the entire area mapped from the remotely sensed data. For this reason, a proper sampling approach must be used in generating the error matrix on which all future analyses will be based (Congalton, 1988). Since a total enumeration of mapped areas for verification is impossible, sampling is the only means by which the accuracy of a land cover map can be derived. Using the wrong sampling design can be costly and yield poor results.

Congalton and Green (1998) list four considerations that are critical to designing an accuracy assessment sample that is truly representative of the map, i.e., (1) statistical distribution of map information, (2) appropriateness of sampling unit, (3) number of samples to be collected, and (4) choice of sample units. Most statistics assume that the population to be sampled is continuous and normally distributed, and that samples selected will be independent. However, map information is discrete and frequently not normally distributed. Therefore, normal statistical techniques that assume continuous distribution may be inappropriate for map accuracy assessment. Spatial autocorrelation is also an important consideration in the formulation of a sampling design for map accuracy assessment. Spatial autocorrelation is said to occur when the presence, absence, or degree of a certain characteristic affects the presence, absence, or degree of the same characteristic in neighboring units (Cliff and Ord, 1973), thereby violating the assumption of sample independence. Campbell (1981) found this condition particularly important in map accuracy assessment when an error in a certain location was found to positively or negatively influence errors in surrounding locations.

It is critical that reference data be collected using the same classification scheme as was used to create the land cover classification map. Classification schemes are a means of organizing spatial information in an orderly and logical fashion, and therefore fundamental to any mapping project. The classification scheme makes it possible for the map producer to characterize landscape features and for the user to readily recognize them. A classification scheme has two critical components: (1) a set of labels, and (2) a set of rules for assigning the labels (Congalton and Green, 1998). The number and complexity of the categories in the classification scheme strongly influence the time and effort needed to conduct the accuracy assessment. The classification scheme should be both mutually exclusive (i.e., each mapped area is one and only one category) and totally exhaustive (i.e., no area on the map can be left unlabeled).

In order to obtain unbiased ground reference information to compare with the remote sensing classification map and fill the error matrix values, we need to determine the most appropriate (i.e., minimum) sample size acceptable for a valid statistical testing of accuracy of the land cover map. In addition, an appropriate sampling scheme must be used to locate the sample points. The binomial distribution or the normal approximation to the binomial distribution is recognized as the appropriate mathematical model to use for determining an adequate sample size for accuracy assessment (Hord and Brooner, 1976; Hay, 1979; Rosenfield and Melley, 1980; Fitzpatrick-Lins, 1981; Rosenfield, 1982). These

techniques are statistically sound for computing the overall accuracy of the classification or even the overall accuracy of a single category (Congalton and Green, 1998).

The equation based on binomial probability theory that relates classification accuracy assessment sample size to overall classification accuracy and allowable error can be used to calculate the allowable error on the accuracy of each land cover map (van Genderen and Lock, 1977; Fitzpatrick-Lins, 1981, Marsh *et al.*, 1994). The equation is:

$$N = \frac{Z^2pq}{E^2}$$

where,

N = Number of samples
p = Expected or calculated accuracy (in percentage)
q = 100-p
E = Allowable error
Z = Standard normal deviate for the 95% two-tail confidence level (1.96)

A decision needs to be made on the allowable error, E, in order to determine the minimum number of sample points necessary to achieve the error. Since we do not have an overall accuracy of any of the land cover maps, nor an allowable error, a decision has to be made on each. We begin by assuming an initial allowable error of 5% and an overall map accuracy of between 60% and 95%. We can now use the formula given above to determine the minimum number of sample points required to achieve this allowable error for maps whose accuracy ranges between 60% and 95%, at the 95% confidence level. The minimum number of sample points necessary to achieve an allowable error of 2.5% can also be calculated in a similar manner. Spatial autocorrelation will affect sample size and especially the sampling scheme to be used in map accuracy assessment because it violates the assumption of sample independence. Autocorrelation may be responsible for periodicity in data that could affect the results of any type of systematic sample (Congalton and Green, 1998).

There are three important considerations in the design of a successful sampling scheme: (1) samples must be selected without bias, (2) choice of sampling scheme determines what further analysis can be carried out, and (3) the sampling scheme will determine the distribution of samples across the landscape, and in turn significantly affect the costs of the accuracy assessment (Congalton and Green, 1998). There are five common sampling schemes that have been applied for collecting reference data in map accuracy assessment:

1. Simple random sampling,
2. Systematic sampling,
3. Stratified random sampling,
4. Cluster sampling,
5. Stratified systematic unaligned sampling.

Many researchers have expressed divergent opinions about the proper sampling schemes to use. Berry and Baker (1968) recommended systematic sampling design as the most efficient when used to assess the accuracy of land use data where geographic autocorrelation was known to decline monotonically with increased distance. However, they concluded that stratified systematic unaligned sampling yielded both the greatest relative efficiency and safety to estimation procedures, where the shape of the autocorrelation function was unknown. Stratified systematic unaligned sampling attempts to combine the advantages of randomness and stratification with the ease of a systematic sample without falling into the pitfalls of

periodicity common to systematic sampling. This method is a combined approach that introduces more randomness than just a random start within each stratum (Congalton and Green, 1998). Several other researchers have supported the use of stratified systematic unaligned sampling (Ayeni, 1982; Mailing, 1989). Campbell (1987) recommended stratified systematic unaligned sampling in situations where the analyst knew enough about the region to make a good choice of grid size. Rosenfield and Melley (1980) and Rosenfield *et al.* (1982) recommended stratified systematic unaligned sampling, with augmentation of the sample by addition of randomly selected pixels in rare map categories to bring the sample sizes in these categories up to some minimum number.

Van Genderen *et al.* (1978) concluded that stratified random sampling techniques were readily accepted as the most appropriate method of sampling in resource studies using remote sensing imagery, because important minor categories could be satisfactorily represented. Several studies conducted earlier by Rudd (1971) and Zonneveld, (1974) had also come to the same conclusion. In one of the few empirical studies specifically addressing sampling in remote sensing, Congalton (1988) conducted a simulation study of three populations by comparing five sampling schemes: simple random, stratified random, cluster, systematic and stratified systematic unaligned sampling. He found that simple random sampling without replacement always provided adequate estimates of the population parameters, provided the sample size was sufficient. However, he found that random sampling may under-sample small but possibly very important classes unless the sample size was sufficiently large. For the less spatially complex agriculture and range areas, systematic sampling and stratified systematic unaligned sampling greatly overestimated the population parameters. For this reason, Congalton (1988) recommended that systematic or stratified systematic unaligned sampling be used with great caution as they tend to overestimate the population parameters. In systematic designs, an unbiased estimator of sampling variance is unavailable and so variance has to be estimated by treating the systematic sample as a simple random sample. It is because of the reasons outlined above that most analysts prefer stratified random sampling (Jensen, 1996). However, stratified random sampling can only be carried out after the map has been completed (i.e., when location of strata is known). This rules out the possibility of simultaneously collecting sample data with training data, thereby potentially increasing project cost. Stratified random sampling can also be a problem if carried out long after the classification map was prepared since there may be temporal changes.

Section 3

Classification Accuracy Assessment Sampling Design for the San Pedro Watershed

A 1992 land cover map for the Upper San Pedro Watershed (Figure 1) was available for classification accuracy assessment. The digital map was interpreted and generated from a 2 June 1992 North American Landscape Characterization (NALC) Landsat Multi-Spectral Scanner (MSS) image by Instituto del Medio Ambiente y el Desarrollo Sustentable del Estado de Sonora (IMADES), Hermosillo, Sonora. The NALC project is a component of the National Aeronautics and Space Administration (NASA) Landsat Pathfinder Program (US-EPA, 1993). The goals of the NALC project are to produce standardized data sets for the majority of the North American continent. The purpose of the project was to develop standard data analysis methods to perform inventories of land cover, quantify land cover change analyses, and produce digital data base products for the U.S. and international global change research programs. A specific objective of the NALC project has been the assembly of three-date georeferenced data sets, called triplicates, for the U.S. Global Change Research Program (GCRP) and retrospective evaluations of change. A set of these NALC triplicates has been generated for the San Pedro Watershed and evaluated for change detection (Kepner *et al.*, 2000).

The Environmental Protection Agency (EPA) supplied the Arizona Remote Sensing Center (ARSC, University of Arizona) with 60 Digital Orthophoto Quadrants (DOQs) for 1992. Most of the land cover classes (Table 1) were fairly well represented in the DOQs and covered between 24% and 41% in eight out of ten land cover classes. Only the Barren and Agriculture classes were poorly represented in the available DOQs covering 5.3% and 14.2% of the map area, respectively.

The available DOQs were black and white and at a scale of approximately 1:24,000. The recommended reference data for classification accuracy assessment, when ground truth data is unavailable, is large scale (1:12,000 or larger) color aerial photography (Congalton and Green, 1993). Since the DOQ data did not meet this criterion, we felt there was a need to use other high-resolution data to determine its suitability for conducting the classification accuracy assessment. We had high resolution airborne color video (at a scale of 1:200 when displayed on a 13-inch monitor) for a subset of the Upper San Pedro River Watershed in the U.S. This video data was acquired in November 1995. Full-zoom video frames (n = 557) with a swath width of 50 m were selected systematically from continuously recorded video over a grid of flight lines. These points had been interpreted and each assigned a cover class based on the Brown, Lowe and Pase system (Brown, 1982). In addition, an estimate of the canopy cover and plant density of species or groups of species present was also made for each analyzed frame. These points represent six of the ten land cover classes in the NALC data sets. The only classes not included in the 557-video frames were (1) agriculture, (2) water, (3) barren, and (4) urban. Fortunately, these classes are clearly the easiest to identify in the DOQ data.

We then used the airborne video data to perform an accuracy assessment of our ability to recognize and identify the six vegetation classes (forest, oak woodland, mesquite woodland, grassland, desertscrub, and riparian) on the DOQ data. We also produced a complete report on the performance of these data for

identification of the six vegetation classes. Conversion of the Brown, Lowe and Pase class video labels to the IMADES land cover classification scheme and video registration issues are described in more detail in Section 4. Both producer's and user's accuracy of individual land cover classes was computed, in addition to overall accuracy, Kappa and Tau statistics.

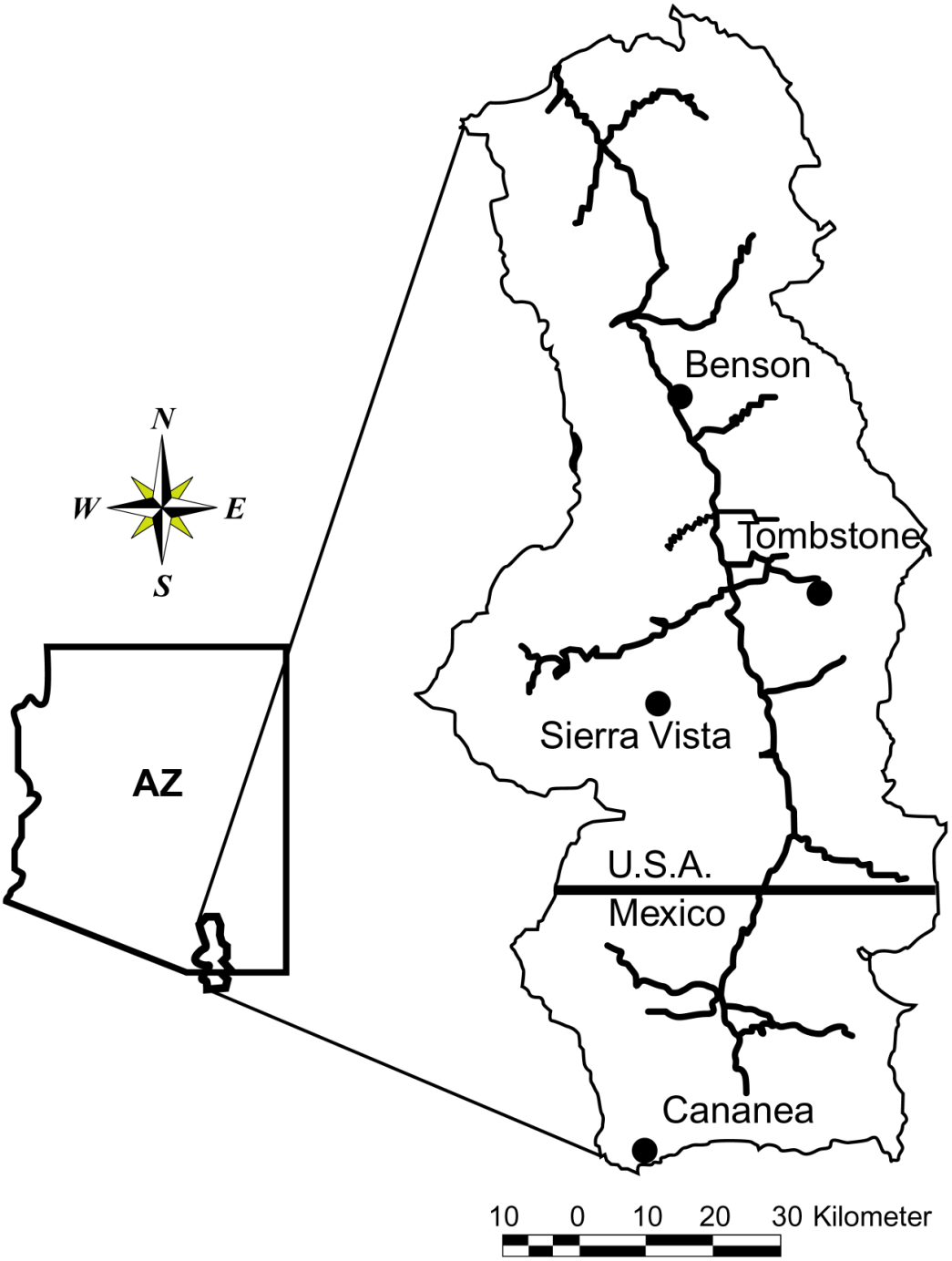


Figure 1. Location of the Upper San Pedro River Basin, Arizona/Sonora (Adapted from Kepner *et al.*, 2000).

Table 1. Land cover class descriptions for the Upper San Pedro Watershed (Adapted from Kepner *et al*., 2000)

Forest	Vegetative communities comprised principally of trees potentially over 10 m in height and frequently characterized by closed or multilayered canopies. Species in this category are evergreen (with the exception of aspen), largely coniferous (e.g., ponderosa pine, pinyon pine), and restricted to the upper elevations of mountains that arise off the desert floor.
Oak Woodland	Vegetative communities dominated by evergreen trees (*Quercus spp.*) with a mean height usually between 6 and 15 m. Tree canopy is usually open or interrupted and singularly layered. This cover type often grades into forests at its upper boundary and into semiarid grassland below.
Mesquite Woodland	Vegetative communities dominated by leguminous trees whose crowns cover 15% or more of the ground often resulting in dense thickets. Historically maintained maximum development on alluvium of old dissected flood plains; now present without proximity to major watercourses. Winter deciduous and generally found at elevations below 1,200 m.
Grassland	Vegetative communities dominated by perennial and annual grasses with occasional herbaceous species present. Generally grass height is under 1 m and they occur at elevations between 1,100 and 1,700 m; sometimes as high as 1,900 m. This is a landscape largely dominated by perennial bunch grasses separated by intervening bare ground or low-growing sod grasses and annual grasses with a less-interrupted canopy. Semiarid grasslands are mostly positioned in elevation between evergreen woodland above and desertscrub below.
Desertscrub	Vegetative communities comprised of short shrubs with sparse foliage and small cacti that occur between 700 and 1,500 m in elevation. Within the San Pedro River basin this community is often dominated by one of at least three species, i.e., creosotebush, tarbush, and whitethorn acacia. Individual plants are often separated by significant areas of barren ground devoid of perennial vegetation. Many desertscrub species are drought-deciduous.
Riparian	Vegetative communities adjacent to perennial and intermittent stream reaches. Trees can potentially exceed an overstory height of 10 m and are frequently characterized by closed or multilayered canopies depending on regeneration. Species within the San Pedro basin are largely dominated by two species, i.e., cottonwood and Goodding willow. Riparian species are largely winter deciduous.
Agriculture	Crops actively cultivated (and irrigated). In the San Pedro River basin these are primarily found along the upper terraces of the riparian corridor and are dominated by hay and alfalfa. They are minimally represented in overall extent (less than 3%) within the basin and are irrigated by ground and pivot-sprinkler systems.
Urban (Low and High Density)	This is a land-use dominated by small ejidos (farming villages or communes), retirement homes, or residential neighborhoods (Sierra Vista). Heavy industry is represented by a single open-pit copper mining district near the headwaters of the San Pedro River near Cananea, Sonora (Mexico).
Water	Sparse freestanding water is available in the watershed. This category would be mostly represented by perennial reaches of the San Pedro and Babocomari rivers with some attached pools or repressos (earthen reservoirs), tailings ponds near Cananea, ponds near recreational sites such as parks and golf courses, and sewage treatment ponds east of the city of Sierra Vista, Arizona.
Barren	A cover class represented by large rock outcropping or active and abandoned mines (including tailings) that are largely absent of aboveground vegetation.

Before we began our accuracy assessment, we needed to determine the minimum number of sample points required so that our calculated classification accuracy would have an allowable error of 5% or less at the 95% confidence interval. Since the overall accuracy of the land cover map was unknown at this time, we needed to assume that it fell within a certain range in order to be able to calculate the minimum number of sample points required to achieve the specified allowable error. For this calculation, we assumed that the overall accuracy of the land cover map was between 60% and 95%. This assumption was based on frequently reported accuracy's of land cover maps derived from satellite data (e.g., Jensen *et al.*, 1993; Marsh *et al.*, 1994; Dimyati *et al.*, 1996; Miguel-Ayanz and Biging, 1997; Ramsey *et al.*, 1997). Using the equation based on binomial probability theory (van Genderen and Lock, 1977; Fitzpatrick-Lins, 1981, Marsh *et al.*, 1994), we calculated the minimum number of sample points for a range of accuracy's, and allowable errors of 5% and 2.5% at the 95% confidence interval. The results are summarized in Table 2.

Table 2. Minimum number of sample points, N1 and N2, required to achieve an allowable error of 5% (E1) and 2.5% (E2), respectively, at the 95% confidence interval and for accuracies ranging from 60% to 95%

N_1	N_2	Z^2	p	Q	E_1^2	E_2^2
369	1475	3.8416	60.00	40.00	25	6.25
360	1441	3.8416	62.50	37.50	25	6.25
350	1398	3.8416	65.00	35.00	25	6.25
337	1348	3.8416	67.50	32.50	25	6.25
323	1291	3.8416	70.00	30.00	25	6.25
306	1225	3.8416	72.50	27.50	25	6.25
288	1152	3.8416	75.00	25.00	25	6.25
268	1072	3.8416	77.50	22.50	25	6.25
246	983	3.8416	80.00	20.00	25	6.25
222	887	3.8416	82.50	17.50	25	6.25
196	784	3.8416	85.00	15.00	25	6.25
168	672	3.8416	87.50	12.50	25	6.25
138	553	3.8416	90.00	10.00	25	6.25
107	426	3.8416	92.50	7.50	25	6.25
73	292	3.8416	95.00	5.00	25	6.25

We then decided to use the lowest expected land cover map accuracy (60%) in determining the minimum number of sample points (~ 370) for the accuracy assessments. If the map accuracy eventually turned out to be higher than this value, this would result in a smaller allowable error (less than 5%) around our estimates at the 95% confidence interval.

Based on the literature review described in Section 2, we concluded that a stratified random sampling design was the most appropriate for the land cover accuracy assessment. Apportionment of the sample points to the different land cover categories is shown in Table 3. However, because the area covered by some of the smaller land cover classes is negligible compared to the rest of the classes, these classes were not apportioned a sufficient number of sample points. If sample size within a stratum is too small, chances

are that even if the classification is poor we could not sample any classification errors (Miguel-Ayanz and Biging, 1997). In such situations, van Genderen and Lock (1977) suggested that the smallest sample in this class should be 20 or 30 for maps in which the admissible percentage errors are 15% and 10%, respectively. For this reason, we set 20 as the minimum number of sample points for any class, therefore increasing our total number of sample points from 370 to 457.

Table 3. Minimum number of sample points per land cover class stratified by area

Land cover	Area (Ha)	Proportion of Area (%)	Estimated Samples	Final Number of Samples
Forest	7385.76	0.98	4	20
Woodland Oak	93663.36	12.42	46	46
Woodland Mesquite	106766.64	14.15	52	52
Grassland	255024.00	33.81	125	125
Desertscrub	232583.40	30.84	114	114
Riparian	5918.04	0.78	3	20
Agriculture	20991.60	2.78	10	20
Urban	24492.24	3.25	12	20
Water	310.68	0.04	0	20
Barren	7139.52	0.95	4	20
Total	**754275.24**	**100.00**	**370**	**457**

For each sample point, the land cover category on the map was noted and entered in the “map” column of a spreadsheet, while the interpreted class on the photo was entered in a “reference” column. This spreadsheet was then used to generate an error matrix in SYSTAT and used to compute the accuracy of each category, along with both commission and omission errors. In addition, summary statistics for the matrix as a whole (overall classification accuracy, Kappa and Tau statistics) were calculated.

Section 4
Methods

DOQs were used to conduct a classification accuracy assessment of the 1992 land cover map. These photos were acquired in 1992 and were therefore current with the land cover map. However, because of the relatively coarse resolution of the DOQs (approximately 1:25,000), it would have been difficult to use them to distinguish between some of the vegetation communities without access to some other higher resolution data. We therefore used high-resolution airborne color video data to help associate subtle changes in shape, texture, or configuration in the DOQs to specific land cover types. This video data was acquired in November 1995 and was at a scale of approximately 1:200. This exercise constituted the training of the analyst in order to be able to distinguish between the different land cover types on the DOQs.

Each video-frame had associated GPS coordinates obtained using a Trimble Basic Receiver at the time of video-frame acquisition. Although the nominal accuracy of the GPS receiver is 100 m, ground sampling revealed that the accuracy was much closer to 20 m (S. Drake, Personal Communication, 1999). There were 557 full-zoom video frames (each frame had a swath width of 50 m) that had been selected systematically from continuously recorded video over a grid of flight lines, but only 105 were coincident with the available DOQs used in the accuracy assessment. Each full-zoom video frame had previously been interpreted (with detailed information on percentage cover by species and by land cover) and a land cover class based on the Brown, Lowe and Pase System (Brown, 1982) assigned. We then re-interpreted the detailed land cover classes based on this system and aggregated them into one of the ten land cover classes used in the San Pedro Watershed mapping. This was one of the most challenging parts of the exercise and required an understanding of the criteria used in the San Pedro Watershed mapping by IMADES. We needed to understand the characteristics and thresholds used by IMADES to assign land cover classes. To help understand these criteria better, a team from both IMADES and ARSC met in the San Pedro Watershed where the former explained their mapping criteria. Typical vegetation classes such as mesquite woodland, desertscrub, oak woodland, and grassland sites were visited and GPS coordinates obtained. In addition, mixed sites that represented various thresholds between land cover classes were visited and GPS coordinates acquired. These sites were later located in available DOQs and compared to similar classes that were both on the DOQs and the video frames.

All land cover classes in the San Pedro Watershed maps were represented in the 105 full-zoom video frames except for agriculture, urban, water and barren classes. The forest and riparian forest classes though present in both the video and DOQs were greatly underrepresented with only one sample each. Fortunately, these absent or underrepresented classes in both the video and DOQs were relatively easy to identify and distinguish in the DOQs.

Since each full-zoom video frame included detailed information on location (GPS coordinates), species composition, and percent land cover, it was possible to locate the corresponding sites on the DOQs. By coupling prior field experience with information derived from the full-zoom video frames (each covering an area approximately 50 m × 50 m) in conjunction with DOQs, the analyst was able to quickly and

confidently develop the knowledge base to recognize each land cover class on the DOQs. Some of the classes were easily identifiable in the DOQs but others such as desertscrub, woodland mesquite, and grassland required considerable training. Figure 2 shows DOQ chips that are 'typical' examples of some of the land cover classes. Each chip has been extracted from a DOQ and represents a 3-pixel × 3-pixel area on a Landsat MSS scene, i.e., 180 m × 180 m area on a DOQ.

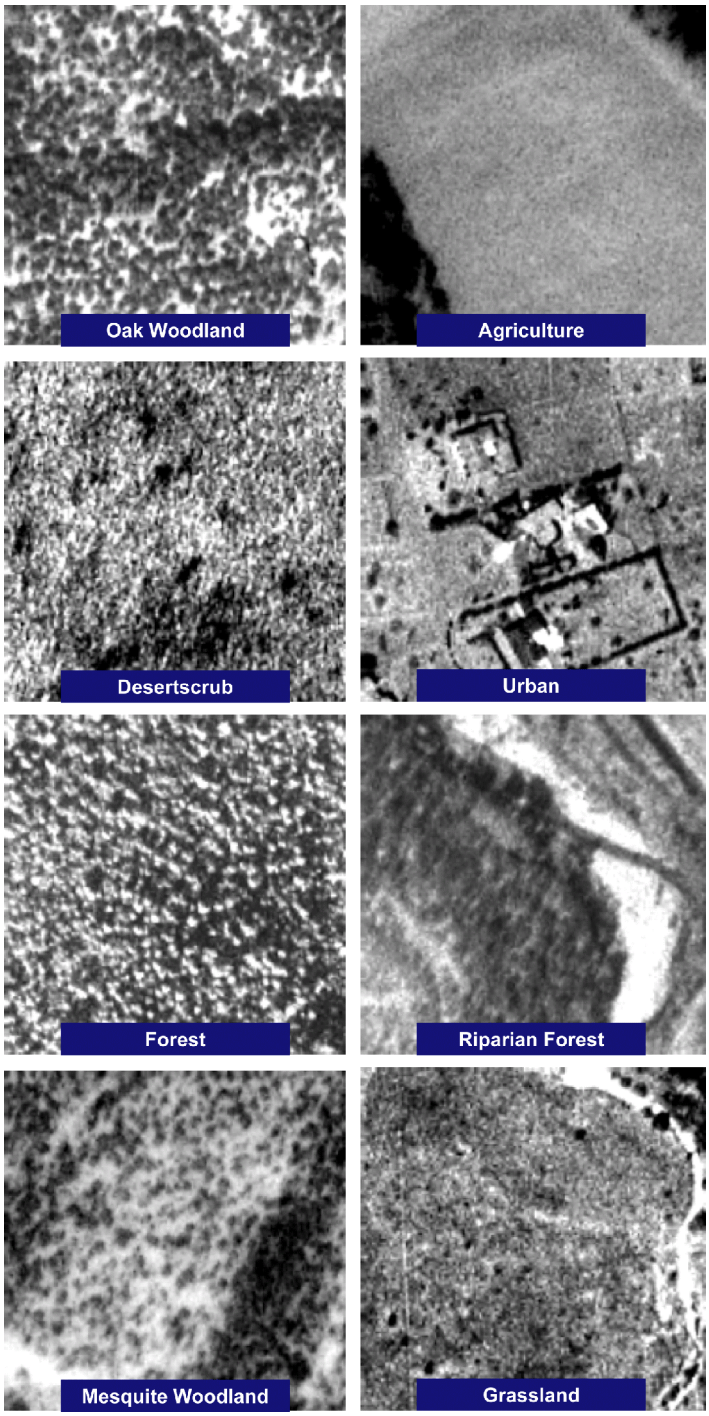


Figure 2. Appearance of some land cover classes on 1992 digital orthophoto quadrangles.

A randomized set of full-zoom video frame locations that correspond to available DOQ coverage was evaluated. The analyst then identified in the DOQs the land cover classes associated with each video frame location. Since the 'revised' class label for each video location was known, it was possible to determine how reliably the DOQ data could be used to identify the vegetation cover classes. A cross-tabulation of the analyst's DOQ labels and the 'reference' video labels were performed in SYSTAT (SPSS, 1998). One limitation in this analysis was the small number of samples for some of the land cover classes but this was not seen as a great problem because the missing classes were those easiest to identify. The result of this assessment (Table 4) is a measure of the analyst's ability to identify the land cover classes.

These results indicated that the DOQ data could be used to identify and distinguish between these vegetation land cover classes.

Table 4. Error matrix illustrating the analyst's ability to use the 1992 DOQs for land cover classification accuracy assessment. A summary of classification errors is appended below

		Airborne Color Infrared Video (1995)						
Digital Orthophoto Quadrangles (1992)		***1***	***2***	***3***	***4***	***5***	***6***	***Grand Total***
	1	**1**	0	0	0	0	0	**1**
	2	0	**20**	0	0	0	0	**20**
	3	0	0	**10**	1	1	0	**12**
	4	0	0	2	**38**	0	0	**40**
	5	0	0	1	2	**28**	0	**31**
	6	0	0	0	0	0	**1**	**1**
Grand Total:		**1**	**20**	**13**	**41**	**29**	**1**	**105**

Land Cover Class	**1992 DOQs**	**1995 Video**	**Number Correct**	**Producer's Accuracy (%)**	**User's Accuracy (%)**
1. Forest	1	1	1	100.00	100.00
2. Woodland Oak	20	20	20	100.00	100.00
3. Woodland Mesquite	12	13	10	76.92	83.33
4. Grassland	40	41	38	92.68	95.00
5. Desertscrub	31	29	28	96.55	90.32
6. Riparian Forest	1	1	1	100.00	100.00
Total:	**105**	**105**	**97**		

Overall Accuracy (%): 92.381

Coefficient	**Value**	**Standard Error**
Kendall's Tau-B	0.912	0.039
Cohen's Kappa	0.907	0.034

Section 5

Accuracy Assessment of the 1992 Land Cover Map

In Section 2 of this report we indicated that at least 370 sample points were required for the classification accuracy assessment. These samples would be sufficient to result in an allowable error that would be within 5% of the estimated accuracy at the 95% confidence level, assuming an overall map accuracy of at least 60%. The sample size was increased to 457 so that after stratification by area, each land cover class would have at least 20 samples.

Generation of sample points was performed in ERDAS IMAGINE (ERDAS, 1998) and relied on a window majority rule. In generating each stratified random sample point, a window kernel of 3 × 3 pixels moved across each land cover class and would result in selection of a sample point only if a clear majority threshold of six pixels out of nine in the window belonged to the same class. If this majority threshold rule was not satisfied, that window would be discarded (ERDAS, 1998) and the kernel would move to a different window. Generation of sample points in this manner ensured that the points were extracted from areas of relatively homogenous land cover class. It is also for this reason that we used a 180 m × 180 m DOQ sample size as it would be equivalent to a 3 × 3 - pixel window on the map.

A total of 457 points was used for the assessment with stratification by land cover area. The error matrix showing producer's and user's, and overall classification accuracy, and including the Kappa and Tau coefficients is shown in Table 5.

Table 5. Classification accuracy error matrix for the 1992 land cover map using 1992 DOQs

		Reference (Digital Orthophoto Quads)										
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	**Grand Total**
Land Cover Classes 1992	***1***	**22**	2	0	0	0	0	0	0	0	0	**24**
	2	0	**44**	0	3	1	0	0	0	0	0	**48**
	3	0	2	**40**	9	10	1	0	0	0	0	**62**
	4	0	6	12	**68**	17	0	0	0	0	0	**103**
	5	0	1	8	11	**89**	0	0	0	0	0	**109**
	6	0	0	0	0	0	**20**	3	0	0	0	**23**
	7	0	0	1	0	0	4	**18**	0	0	0	**23**
	8	0	0	2	1	10	0	1	**11**	0	0	**25**
	9	0	0	1	0	0	0	0	0	**19**	0	**20**
	10	0	0	0	7	2	0	0	0	0	**11**	**20**
Grand Total:		**22**	**55**	**64**	**99**	**129**	**25**	**22**	**11**	**19**	**11**	**457**

Table 5 *(Continued)*

Land Cover Class	92_Map Total	DOQ Total	Number Correct	Producer's Accuracy (%)	User's Accuracy (%)
1. Forest	24	22	22	100.00	91.67
2. Woodland Oak	48	55	44	80.00	91.67
3. Woodland Mesquite	62	64	40	62.50	64.52
4. Grassland	103	99	68	68.69	66.02
5. Desertscrub	109	129	89	68.99	81.65
6. Riparian Forest	23	25	20	80.00	86.96
7. Agriculture	23	22	18	81.82	78.26
8. Urban	25	11	11	100.00	44.00
9. Water	20	19	19	100.00	95.00
10. Barren	20	11	11	100.00	55.00
Total:	457	457	342		

Overall Accuracy (%):	74.836 ± 3.979

Coefficient	Value	Standard Error
Kendall's Tau-B	0.770	0.025
Cohen's Kappa	0.701	0.025

An overall accuracy of about 75% was obtained. Although the producer's accuracy for the urban and barren classes is 100%, the user's accuracy is only 44%, and 55%, respectively. This means that, all the urban and barren class pixels examined in the DOQs were also labeled as urban and barren classes in the 1992 map. However, there were many more pixels in the map labeled 'urban' and 'barren' that turned out to be some other class in the DOQs. Indeed, only 44% of all pixels labeled urban in the 1992 turned out to be urban while 55% of map pixels labeled barren turned out to be 'barren' in the DOQs.

Conclusions

The results discussed in this report indicate that DOQ data when used together with higher resolution data can be successfully used to perform classification accuracy assessment on land cover maps derived from historical satellite data. It is essential that geometric rectification between digital maps being assessed and the DOQs be equal (Appendix 1). It is expected that newer DOQs will be even more effective for accuracy assessment because of their multispectral characteristics. Because DOQs are already georeferenced, they could be used to georeference other historical photography for more valid assessment of land cover maps generated from data before 1992.

The use of DOQ data sets to assess satellite derived classification accuracies appears to be a viable methodology. In addition, this methodology could be applied to assess classification accuracy in other project areas that have used Landsat MSS data obtained from the NALC program.

Appendix 1

Ground Control Points Used to Georeference the 1992 NALC Subset Image to a Precision Corrected 1997 Landsat TM Image

GCP_ID	X source	Y source	X destination	Y destination	X residual	Y residual	RMS Error
GCP #1	576.625	750.875	567530.688	3547876.688	-0.156	-0.424	0.452
GCP #2	534.625	808.625	565001.313	3544399.688	-0.100	-0.080	0.128
GCP #3	359.125	886.625	554470.563	3539739.938	0.309	-0.058	0.314
GCP #4	466.781	1635.031	560794.891	3494862.234	0.062	0.440	0.444
GCP #5	525.531	1662.531	564343.141	3493244.859	0.458	-0.148	0.481
GCP #6	489.875	1054.625	562279.563	3529636.688	0.038	0.392	0.394
GCP #7	587.875	937.125	568177.281	3536692.219	-0.112	-0.010	0.113
GCP #8	625.781	1401.781	570399.391	3508848.609	0.292	0.019	0.293
GCP #9	492.406	1625.656	562319.641	3495464.297	-0.208	-0.270	0.341
GCP #10	735.625	1699.125	576914.313	3491004.938	-0.249	0.351	0.430
GCP #11	749.031	981.781	577841.453	3534006.984	-0.256	-0.073	0.266
GCP #12	623.781	1220.531	570285.391	3519732.047	-0.121	-0.384	0.402
GCP #13	478.750	1441.250	561562.609	3506524.078	0.322	-0.445	0.549
GCP #14	982.031	2413.531	591603.391	3448200.609	-0.091	0.046	0.102
GCP #15	912.375	2564.125	587377.375	3439174.125	-0.361	0.212	0.418
GCP #16	962.125	2015.125	590455.375	3472091.625	-0.427	-0.285	0.514
GCP #17	1112.625	2004.875	599546.875	3472647.375	0.398	0.476	0.621
GCP #18	1133.531	2163.781	600767.031	3463162.219	0.254	-0.142	0.291
GCP #19	471.531	941.031	561179.641	3536450.859	-0.252	0.276	0.374
GCP #20	145.625	1640.375	541488.813	3494588.813	-0.170	0.111	0.203
GCP #21	508.031	722.406	563431.141	3549560.859	0.110	0.030	0.114
GCP #22	991.125	1662.875	592290.063	3493156.688	0.119	0.326	0.347
GCP #23	825.375	1247.625	582414.813	3518065.688	0.290	0.052	0.295
GCP #24	550.531	579.031	565999.703	3558185.672	-0.037	-0.611	0.612
GCP #25	804.531	1950.031	581015.641	3476002.359	-0.168	-0.244	0.296
GCP #26	661.531	1987.031	572437.141	3473807.859	0.124	-0.415	0.433
GCP #27	440.625	357.375	559413.531	3571419.469	-0.402	0.314	0.510
GCP #28	365.625	154.375	554967.531	3583603.219	0.006	0.060	0.060
GCP #29	309.906	591.906	551565.344	3557368.969	0.328	0.483	0.584

X RMS Error	**0.249**
Y RMS Error	**0.301**
Total RMS Error	**0.390**

References

Ayeni, O.O. 1982. Optimum sampling for digital terrain models: a trend towards automation. *Photogrammetric Engineering and Remote Sensing* 48: 1687-1694.

Aronoff, S. 1985. The minimum accuracy value as an index of classification accuracy. *Photogrammetric Engineering and Remote Sensing* 51(1): 99-111.

Berry, B.J.L., and A.M. Baker. 1968. Spatial analysis: a reader in statistical geography, In B.J.L. Berry and D.F. Marble (Eds.), Geographic sampling. Prentice Hall, Englewood Cliffs, NJ, pp. 91-100.

Brown, D.E. (Editor). 1982. Biotic communities of the American southwest-United States and Mexico. *Desert Plants* (special issue) 4(1-4): 1-342.

Campbell, J.B. 1981. Spatial autocorrelation effects upon the accuracy of supervised classification of land cover. *Photogrammetric Engineering and Remote Sensing* 47(3): 355-363.

Campbell, J.B. 1987. Introduction to remote sensing. Guilford Press: New York.

Cliff, A.D. and J.K. Ord. 1973. Spatial autocorrelation. Pion, London.

Cohen. I. 1960. A coefficient of agreement of nominal scales. *Educational and Psychological Measurement* 20(1): 37-46.

Congalton, R.G. and R.A. Mead. 1983. A quantitative method to test for consistency and correctness in photo-interpretation. *Photogrammetric Engineering and Remote Sensing* 49(1): 69-74.

Congalton, R.G. 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogrammetric Engineering and Remote Sensing* 54(5): 593-600.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37: 35-46.

Congalton R.G. and K. Green. 1993. A practical look at the sources of confusion in error matrix generation. *Photogrammetric Engineering and Remote Sensing* 59(5): 641-64.

Congalton, R.G., R. Oderwald, and R. Mead. 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing* 49(12): 1671-1678.

Congalton R.G. and K. Green. 1998. Assessing the accuracy of remotely sensed data: principles and practices. Lewis Publishers, New York.

Dimyati, M., K. Mizuno, S. Kobayashi, and T. Kitamura. 1996. "An Analysis of Land Use / Cover Change Using the Combination of MSS Landsat and Land Use Map – a Case Study in Yogyakarta, Indonesia." *International Journal of Remote Sensing* 17: 931-944.

ERDAS. 1998. *ERDAS IMAGINE*. Version 8.3. Atlanta, Georgia: ERDAS Inc.

ESRI. 1998a. *ARC/INFO*. Vers. 7.1.2. Redlands, California: Environmental Systems Research Institute.

ESRI. 1998b. *ArcView*. Vers. 3.1. Redlands, California: Environmental Systems Research Institute.

Fitzpatrick-Lins, K. 1981. Comparison of sampling procedures and data analysis for a land use and land cover maps. *Photogrammetric Engineering and Remote Sensing* 47(3): 343-351.

Foody, G. 1992. On the compensation of chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing* 58(10): 1459-1460.

Genevan, M.E. 1979. Testing land use map accuracy: another look. *Photogrammetric Engineering and Remote Sensing* 45(10): 1371-1377.

Hay, A.M. 1979. Sampling designs to test land use map accuracy. *Photogrammetric Engineering and Remote Sensing* 45(40): 529-533.

Hord, R.M. and W. Brooner. 1976. Land use map accuracy criteria. *Photogrammetric Engineering and Remote Sensing* 42(5): 671-677.

Gong, P. and P.J. Howarth. 1990. An assessment of some factors influencing multispectral land cover classification. *Photogrammetric Engineering and Remote Sensing* 56(5): 597-603.

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

Jensen, J. R., S. Narumalani, O. Weatherbee, and H. E. Mackey. 1993. "Measurement of Seasonal and Yearly Cattail and Waterlily Changes Using Multidate SPOT Panchromatic Data." *Photogrammetric Engineering and Remote Sensing* 59:519-25.

Kepner, W.G., C.J. Watts, C.M. Edmonds, J.K. Maingi, S.E. Marsh, and G. Luna. 2000. A landscape approach for detecting and evaluating change in a semi-arid environment. Environmental Monitoring and Assessment, 64(1): 179-195.

Ma, Z. and R.L. Redmond. 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing* 61(4): 435-439.

Marsh, S. E., J. L. Walsh, and C. Sobrevila. 1994. Evaluation of airborne video data for land cover classification accuracy assessment in an Isolated Brazilian Forest. *Remote Sensing of Environment* 48: 61-69.

Mailing, D.H. 1989. Measurements for maps: principles and methods of cartometry. Pergamon Press: New York.

Miguel-Ayanz, J.S. and G.S. Biging. 1997. "Comparison of Single-Stage and Multi-Stage Classification Approaches for Cover Type Mapping with TM and SPOT Data." *Remote Sensing of Environment* 59: 92-104.

Ramsey, E.W. and S.C. Laine. 1997. "Comparison of Landsat Thematic Mapper and High Resolution Photography to Identify Change in Complex Coastal Wetlands." *Journal of Coastal Research* 13: 281-292.

Rosenfield, G.H. 1981. Analysis of variance of thematic mapping experiment data. *Photogrammetric Engineering and Remote Sensing* 47(12): 1685-1692.

Rosenfield, G.H. 1982. Sample design for estimating change in land use and land cover. *Photogrammetric Engineering and Remote Sensing* 48(5): 793-801.

Rosenfield, G.H. and M.L. Melley. 1980. Applications of statistics to thematic mapping. *Photogrammetric Engineering and Remote Sensing* 48(5): 1287-1294.

Rosenfield, G.H. and K. Fitzpatrick-Lins. 1986. A coefficient of agreement as a measure of thematic accuracy. *Photogrammetric Engineering and Remote Sensing* 52(2): 223-227.

Rosenfield, G.H., K. Fitzpatrick-Lins, and H.S. Ling. 1982. Sampling for thematic accuracy testing. *Photogrammetric Engineering and Remote Sensing* 48(1): 131-137.

Rudd, R.D. 1971. Macro land use mapping with simulated space photographs. *Photogrammetric Engineering* 37: 365-372.

Story, M. and R.G. Congalton. 1986. Accuracy assessment: a users perspective. *Photogrammetric Engineering and Remote Sensing* 52(3): 397-399.

SPSS Inc. 1998. *SYSTAT*. Version 8.0, Chicago, Illinois: SPSS Inc.

U.S. Environmental Protection Agency. 1993. North American Landscape Characterization (NALC). Research Brief. EPA/600/S-93/0005, Las Vegas, NV.

van Genderen, J. L. and B. F. Lock. 1977. Testing land use map accuracy. *Photogrammetric Engineering and Remote Sensing* 43:1135-37.

van Genderen, J. L., B. F. Lock, and P.A. Vass. 1978. Remote sensing: statistical testing of thematic map accuracy. Proceedings of the 12th International Symposium on Remote Sensing of Environment, ERIM, Ann Arbor, MI, pp. 3-14.

Verbyla, D. L. 1995. Satellite remote sensing of natural resources. New York: CRS Lewis Publishers.

Zonneveld, I.S. 1974. Aerial photography, remote sensing and ecology. *ITC Journal*, 5(4): 553-560.

United States
Environmental Protection
Agency

Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
P.O. Box 93478
Las Vegas, Nevada 89193-3478

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE 9; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35

Official Business
Penalty for Private Use
$300

EPA/600/R-02/040
June 2002