Accuracy Assessment of a Land-Cover Map of the Kuparuk River Basin, Alaska: Considerations for Remote Regions

S.V. Muller, D.A. Walker, F.E. Nelson, N.A. Auerbach, J.G. Bockheim, S. Guyer, and D. Sherba

Abstract

An accuracy assessment of a Landsat MSS-derived land-cover map of the Kuparuk River basin, Alaska was performed. We used a stratified systematic transect-based sampling design with a homogeneous 3- by 3-pixel block sampling unit. The ramifications of the sampling strategy are discussed. Sample sites were located using a helicopter and a Y-Code GPS receiver. Estimates of overall classification accuracy (P), Tau (T_o), producer's accuracy, and user's accuracy were calculated from an error matrix. Assessment methods based on fuzzy sets theory were used to supplement the error matrix. The accuracy estimates indicate a classification with high accuracy. However, they are likely to have a fair degree of optimistic bias and can only be applied reliably to homogeneous 3 by 3 blocks of pixels. The combined use of an error matrix and fuzzy sets allowed for a more precise analysis of errors. Based on this analysis, changes were made to the final map. Several methodological advantages contributed to the high classification accuracy.

Introduction

Purpose

Quantifying and documenting the accuracy of maps and spatial data are important components of any mapping process. However, assessing the accuracy of a map can be a time-consuming and expensive process. This is especially true for maps of remote areas, such as the North Slope of Alaska, which can involve considerable financial, logistical, and technical constraints. This paper presents an accuracy assessment of a satellite-derived land-cover map of the Kuparuk River basin on the North Slope of Alaska and examines how the choice of sampling strategy and analysis methods affects estimates of classification accuracy and their usefulness.

S.V. Muller, D.A. Walker, and N.A. Auerbach are with the Tundra Ecosystem Analysis and Mapping Laboratory, Institute of Arctic and Alpine Research, University of Colorado, CB 450, 1560 30th St., Boulder, CO 80309-0450 (mullers@taimyr.colorado.edu).

F.E. Nelson is with the Department of Geography, University of Delaware, Newark, DE 19716.

J.G. Bockheim is with the Soils Department, University of Wisconsin, 1525 Observatory Dr., Madison, WI 53706.

S. Guyer and D. Sherba are with the Bureau of Land Management, Division of Cadastral Survey (AK-924), 222 West 7th Ave. #13, Anchorage, AK 99513-7599.

The Kuparuk River Basin Land-Cover Map

As part of the of the National Science Foundation's (NSF) Land-Atmosphere-Ice-Interactions (LAII) Flux Study (Weller *et al.*, 1995), a land-cover map of the Kuparuk River basin was derived from Landsat Multi-Spectral Scanner (MSS) satellite image data. The Kuparuk River basin, located on the North Slope of Alaska, covers approximately 9,201 km^2 and extends 216 km from north to south, and 78 km from east to west (Figure 1). Except for the Dalton Highway and oil drilling bases along the Arctic coast, this area of the North Slope remains undeveloped, remote, and difficult to access.

General vegetation land-cover types were derived by classification of the Landsat MSS satellite data (Plate 1; N. Auerbach *et al.*, unpublished data, 1996). A framework for the land-cover type designations was provided by Braun-Blanquet vegetation analysis of a tussock tundra landscape in the Brooks Range Foothills, Alaska (Walker *et al.*, 1994). Derived classes include (1) Barrens, (2) Moist nonacidic tundra (MNT), (3) Moist acidic tundra (MAT), (4) Shrublands, (5) Wet tundra, (6) Water, (7) Clouds and Ice, and (8) Shadows. To expedite image processing, the digital data for a rectangular region encompassing the Kuparuk River watershed were extracted from an existing mosaic of MSS frames covering the Central Arctic Management Area (CAMA) and Arctic National Wildlife Refuge (ANWR), northeast Alaska, produced by the U.S. Geological Survey, EROS Data Center, Sioux Falls, South Dakota. Images for the entire mosaic were acquired during the snow-free growing seasons of 14 August 1976 through 2 August 1985. Due to prevalent cloud cover over the North Slope during most growing seasons, single-time-period (e.g., one week) mosaics of imagery from sun-synchronous satellites are generally not feasible. The mosaic (80 m nominal spatial resolution) was resampled to 50-m pixels, and was geometrically corrected using cubic convolution interpolation by means of a second-order polynomial registration, with a resultant root-mean-square error (RMSE) of 57.4 m. An IsoData unsupervised classification approach was implemented based on input of the green, red, and infrared spectral bands of the MSS image. Forty cluster classes were initially generated and then aggregated into the eight land-cover classes. We used first-hand experience and familiarity with the area, as well as geobotanical maps and earlier Landsat-derived maps of the region, as supplementary information to interpret the spectral classes (Walker *et al.*, 1982; Walker and Acevedo, 1987; Walker *et al.*, 1989; Walker and Walker, 1991;

Photogrammetric Engineering & Remote Sensing, Vol. 64, No. 6, June 1998, pp. 619–628.

0099-1112/98/6406–619$3.00/0

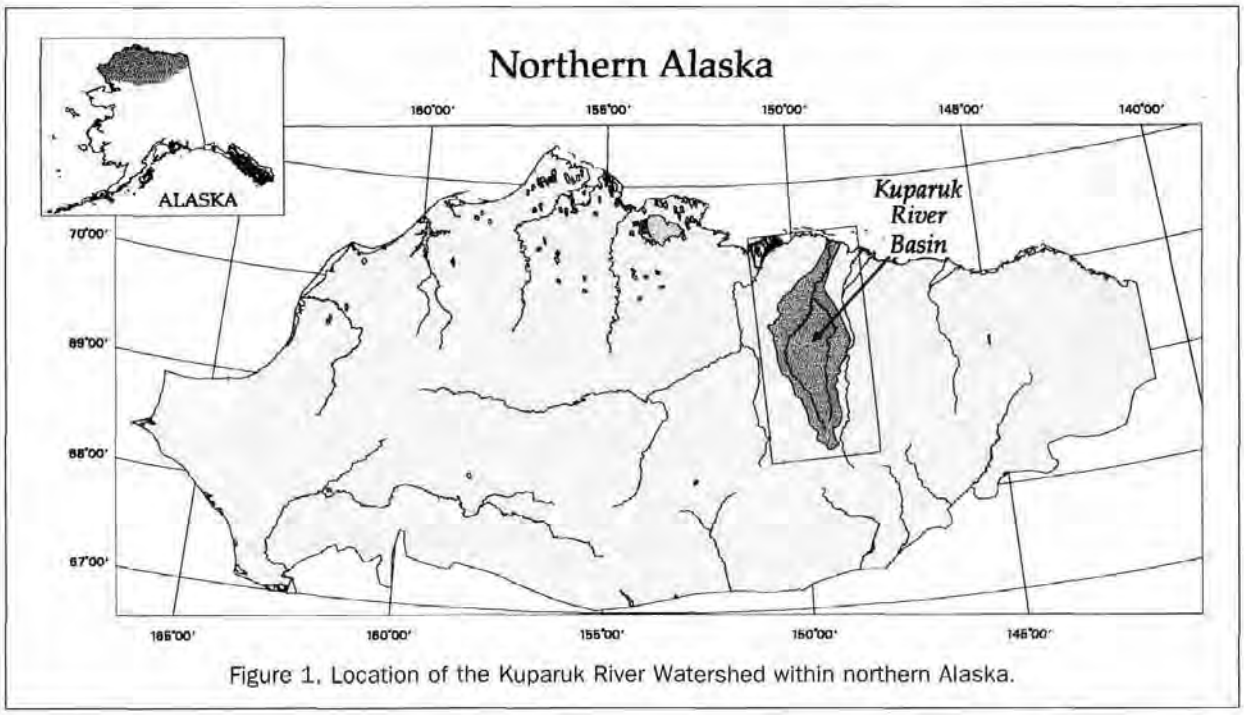

Figure 1. Location of the Kuparuk River Watershed within northern Alaska.

Walker and Walker, 1996; Walker *et al.*, unpublished data, 1996). The MSS classification was refined through post-classification sorting using ancillary data (Hutchinson, 1982).

Background

In general, an accuracy assessment of a map classification is performed by comparing it to a more detailed, independently sampled data set. A considerable amount of research has focused on various methods of assessing the accuracy of maps derived from remotely sensed data; most of these are designed for and applicable to categorical data. Hord and Brooner (1976), van Genderen and Lock (1977), and Hay (1979) presented and discussed methods for determining appropriate sample size. Others have presented discussions, research, and empirical experiments on various sampling designs (Congalton, 1988; Gong and Howarth, 1990; Stehman, 1992). Story and Congalton (1986) discussed the use of error matrices for the analysis of reference data, including calculation of descriptive statistics such as producer's and user's accuracy. Congalton *et al.* (1983), Rosenfield and Fitzpatrick-Lins (1986), Hudson and Ramm (1987), and Foody (1992) discussed various uses and iterations of the Kappa coefficient of agreement, which is derived from an error matrix. Card (1982), Rosenfield (1986), Naesset (1995), and Ma and Redmond (1995) presented alternative statistical methods for measuring accuracy. Gopal and Woodcock (1994) presented methods for using fuzzy sets in accuracy assessments of thematic maps. Recently, some researchers have shifted focus from assessing spectrally caused classification errors to assessing a myriad of spatial and temporal sources of error (e.g., positional error between image and reference data) that can bias estimates of classification accuracy (Congalton and Green, 1993; Verbyla and Hammond, 1995; Hammond and Verbyla, 1996).

Mapping of land cover and vegetation in remote areas of northern Alaska have been performed by Morrissey and Ennis (1981), Walker *et al.* (1982), C. J. Markon (unpublished data, 1986), Walker and Acevedo (1987), Fleming (1988), Stow *et al.* (1989), Markon (1992), Jorgenson *et al.* (1994), Markon and Kirk (1994), and Pacific Meridian Resources (1995). Of these studies, quantitative accuracy assessments were performed by Fleming (1988), Felix and Binney (1989), Stow *et al.* (1989), and Pacific Meridian Resources (1995). Positional accuracy of field-collected training and reference data was either unknown or unstated in these studies. Uncertainty in the positional accuracy of field data (training or reference) has been one of the biggest obstacles to mapping remote, undeveloped areas. Positional error in reference data is a source of classification confusion (Congalton and Green, 1993) and can cause conservative bias in an accuracy assessment due to co-registration problems between reference data and map data (Verbyla and Hammond, 1995). Therefore, it is important that any field-collected data be positionally accurate.

Methods

Sampling Issues and Strategy

The goal of the reference data collection phase of our accuracy assessment was to adopt methods that would result in reference data that were logistically feasible, spatially accurate, and sampled in a statistically sound manner. The undeveloped nature of the Kuparuk River region required the use of a helicopter to collect ground-truth information. Our sampling strategy was planned around our need to pre-select sampling location coordinates so that we could locate each site using a helicopter and a GPS unit.

Because positional errors in the reference data and the map data can cause bias in estimates of classification accuracy, it is important to consider potential and actual positional error when planning ground-truth data collection. The positional accuracy of our reference data was going to depend on the positional accuracy of the GPS unit we would be using. Through a collaborative effort with scientists at the U.S. Department of the Interior Bureau of Land Management (BLM), we were able to use a Y-Code GPS receiver (Y-Code is restricted to government use through encryption). The unit provided by the BLM was a Rockwell Precision Lightweight Global Positioning System Receiver (PLGR). The Y-Code provides an accuracy of 22 metres at 2distance root-mean-

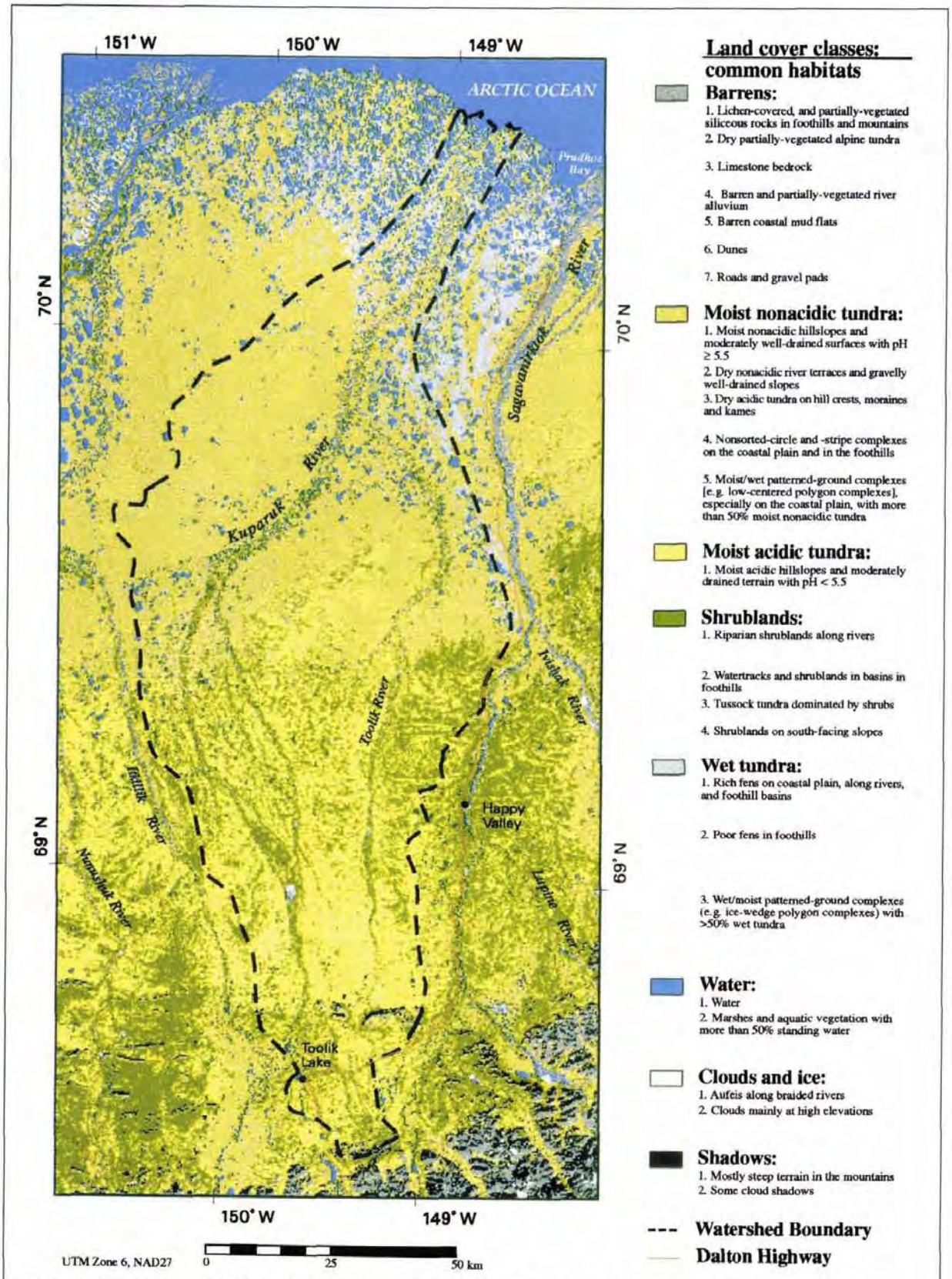

Plate 1. Land-cover map of the Kuparuk River watershed and surrounding area. The map shown here includes changes that were made based on the results of the accuracy assessment.

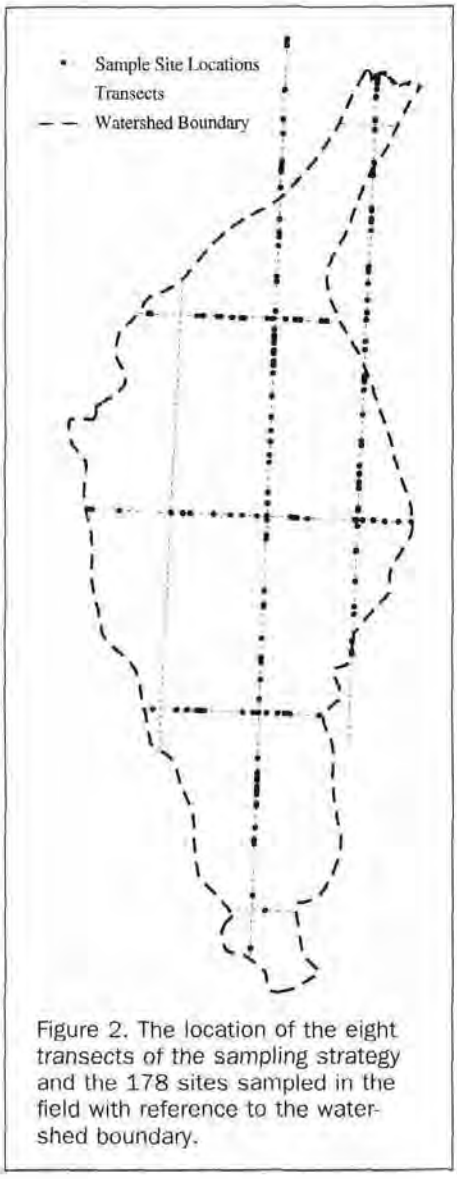

Figure 2. The location of the eight transects of the sampling strategy and the 178 sites sampled in the field with reference to the watershed boundary.

squared [2drms — i.e., 95 percent confidence interval (Federal Radionavigation Plan, 1994)]. This unit also had Wide Area GPS Enhancements (WAGE) which increased its accuracy to 4 metres circular error probable (50 percent confidence level). Collaboration with the BLM was not absolutely confirmed until just prior to our data collection phase. For this reason, we planned our sampling strategy around the possible use of a C/A Code GPS receiver, which has a horizontal accuracy no worse than 100 metres at 2drms. The difference in accuracy between the two GPS technologies means a substantial difference in the accuracy of data and thus can influence the choice of sampling unit used in a sampling strategy. The use of a helicopter hindered the use of other techniques that can substantially improve data collected with a C/A Code receiver.

Choices made in creating a sampling strategy can affect the reliability of accuracy assessments (Congalton, 1991; Janssen and van der Wel, 1994). Understanding such effects is important for interpreting accuracy assessment results. The three major components of a sampling strategy are (1) the sampling unit (e.g., pixels or polygons), (2) the sampling design (e.g., random sampling), and (3) the sample size. Our choice in each of these components was balanced between logistics and the effect that it would have on estimates of classification accuracy (e.g., bias).

The first step in devising the sampling strategy was to determine the sampling unit. For maps derived from satellite data, the ideal sampling unit is the individual pixel (Janssen and van der Wel, 1994). However, the accuracy of a C/A Code GPS receiver combined with the positional accuracy of the map itself could add a potentially high level of uncertainty in locating individual pixels on the ground. Such uncertainty in our reference data set was unacceptable. Therefore, we used a sampling unit of 3 by 3 blocks of pixels with the same land-cover class. This criterion would allow for GPS receiver imprecision when trying to compare reference data with map data. With a sampling unit of this size, we could minimize bias caused by co-registration errors.

The second component to determine was the sampling design. Theoretically, the statistically ideal design would be either a simple random sample or a stratified random sample (Hord and Brooner, 1976; Hay, 1979; Congalton, 1988). However, logistical constraints often prevent the use of ideal sampling designs and many spatially explicit adaptations have been introduced, including systematic sampling (Berry and Baker, 1968; Thompson, 1992), stratified systematic unaligned sampling (Berry and Baker, 1968), cluster sampling (Thompson, 1992), and transect sampling (Thompson, 1992).

To maximize the use of limited helicopter time, we implemented a stratified systematic transect-based sampling design. Sampling was stratified within map categories, and transects were systematically selected throughout the watershed. Flying along transects is far more efficient than flying to random sites. Eight transects were identified for use in the sampling effort (Figure 2): three equally spaced transects running north-south and five equally spaced transects running east-west. The interval between transects one, two, and three was 35 minutes of longitude (approximately 22.5 km). The interval between the remaining five transects was 25 minutes of latitude (approximately 46 km). The transects were spaced to encompass the widest range of physiographic regions, vegetation, and spectral reflectance within the watershed.

Using a randomly selected starting point between 0 and 500 m from the start of each transect, 3 by 3 blocks every 250 metres were analyzed for homogeneity; using this interval prevented overlap between 3 by 3 blocks of 50-m^2 pixels. Locations that met the homogeneity criterion were considered "potential" sites, because the analysis was likely to produce more sites than were needed for a sample. We planned on randomly sampling from this pool of potential sites to obtain our sample set.

The final component of the sampling strategy was to determine the sample size. For estimating overall classification accuracy, a minimum of 50 samples is recommended for maps with less than a dozen categories (Hay, 1979; Congalton, 1991). However, if each category's accuracy is to be estimated, 50 sample sites per category is desirable. Based on the planned uses of the map, we felt that the accuracy of individual categories should be available to map users. However, we estimated that, under optimal field conditions, 225 to 250 sites could be visited. Based on this, we set the minimum sample size per category at 30 and concentrated the remaining samples in the most important categories, MNT (Moist Nonacidic Tundra) and MAT (Moist Acidic Tundra). We determined the sample size of these two categories on the basis of their percentage occurrence on the map relative to the overall sample size. For example, 38.9 percent of the watershed was covered by MNT; therefore, 87 (38.9 percent of 225) was the number of samples in this category. Stratifying the sample in this manner would allow for sample sizes

TABLE 1. RESULTS OF THE SAMPLING STRATEGY. POTENTIAL SITES ARE SITES ALONG THE TRANSECTS THAT MET THE HOMOGENEITY CRITERIA. THE SAMPLE SET WAS DERIVED BY RANDOMLY SAMPLING POTENTIAL SITES OR FINDING MORE POTENTIAL SITES WHEN A CLASS WAS UNDER-REPRESENTED

Land-Cover Category	Potential Sites	Sample Set
Barrens	7	17
MNT	341	87
MAT	128	69
Shrublands	22	30
Wet tundra	10	30
Water	61	30
Total	569	263

large enough to produce meaningful accuracy estimates in all categories and more statistically sound estimates for the MNT and MAT categories.

Based on the sampling strategy, we used Environmental Systems Research Institute's (ESRI) Arc/Info geographic information system (GIS) software package to derive a set of potential sites that met our homogeneity criterion (Table 1). The less common land-cover categories — Barrens, Wet tundra, and Shrublands — had fewer than 30 potential sites. To increase the number of potential sites in these categories, the initial spacing requirement was removed and every pixel along each transect that fell into one of these categories was analyzed for 3- by 3-block homogeneity. This introduced the possibility of spatial autocorrelation between samples in these categories which is a potential source of bias. The remaining three categories were randomly sampled to meet targeted sample sizes. The class counts for the sample set are shown in Table 1. The minimum of 30 samples was not achieved in the Barrens category. However, this was deemed acceptable because it is a spectrally distinct class and unlikely to be confused with other classes.

Data Collection

Waypoint data (X, Y coordinates with unique site identification numbers) for the 263 samples sites were loaded into the PLGR's memory. Using the waypoint information, azimuth and range data were constantly being calculated and were used to guide the helicopter to the exact position of each sample site. For the majority of sample sites, ground-truth evaluations were made from the landing point. The primary land-cover type in a 25-metre radius from the helicopter was recorded. If other land-cover types covered more than 30 percent of the observation area, then secondary and possibly tertiary land-cover was also recorded. It is impossible to truly assess dominance of vegetation types by visual inspection of a 25-m radius circle from a single observation point. Therefore, error in classifying land cover at sample sites is an unknown source of bias when estimating classification accuracy. However, a simple classification scheme and expert knowledge of the vegetation mitigates some of this uncertainty. It is important to note that the analysis of errors was done by comparing the ground-truth data to the center pixel of each 3 by 3 area selected from the map.

Error Matrix Analysis

An error matrix (Story and Congalton, 1986; Congalton, 1991) was used to calculate overall classification accuracy (P), a confidence interval for P, producer's accuracy, and user's accuracy. P is a simple, intuitive measure of the proportion of total sampling units that were correctly classified; it indicates the overall probability that a unit on the ground was correctly classified. User's accuracy is a measure of commission errors, indicating the probability that a unit within an individual category is correctly classified. Producer's accuracy is a measure of omission errors, indicating the probability that a reference data sample is correctly classified. It is useful for the map producer because it measures the degree to which a land-cover type on the earth can be distinguished and mapped using remote sensing data.

Another widely used measure for estimating overall classification accuracy has been the Kappa coefficient of agreement (Congalton *et al.*, 1983; Rosenfield and Fitzpatrick-Lins, 1986). The calculation of Kappa attempts to remove chance agreement from estimates of classification accuracy by incorporating the row and column totals of the error matrix (Cohen, 1960). Recently, Foody (1992) has shown that Kappa tends to over-account the level of chance agreement. Therefore, Kappa will consistently underestimate the overall classification accuracy. Ma and Redmond (1995) expanded upon Foody's (1992) work and presented a Kappa-like statistic, Tau (T), which they argue is a more precise measure of accuracy that removes chance agreement from the estimate of classification accuracy. In our analysis, we have adopted the case of T_e, which is appropriate when an error matrix is derived from a classification that had equal possibilities of group membership, *a priori*. Confidence intervals were also calculated for T_e.

Fuzzy Sets Analysis

Because vegetation often occurs as a mosaic, classifying the vegetation or land cover of an areal unit into only one class is often difficult or erroneous. For this reason, we also adopted accuracy assessment methods based on fuzzy sets theory (Gopal and Woodcock, 1994). Fuzzy sets theory recognizes uncertainty in the mapping process and allows for an areal unit to correctly fall into more than one category. While in the field we used Gopal and Woodcock's linguistic scale to assign a "fuzzy value" to each map class for each site (Table 2). The benefit of using fuzzy sets theory is that it provides additional tools for analyzing map error.

The results of assigning fuzzy values at each sample site were analyzed using three different functions. Following Gopal and Woodcock's (1994) nomenclature, a correct classification is called a match, and an incorrect classification is called a mismatch. The first function measures the frequency of errors and is divided into two sub-functions: the *max* function and the *right* function. Frequency tabulations for the number of matches under each function result in a measure of overall classification accuracy. The *max* function is a more conservative measure of accuracy than the *right* function. The *max* function measures how frequently the mapped class is the best classification — indicated by the highest fuzzy value being the same as that of the classification. The *right* function measures how frequently the mapped class

TABLE 2. LINGUISTIC SCALE OR "FUZZY VALUES" USED IN EXPERT EVALUATION OF LAND COVER AT EACH SAMPLE SITE (GOPAL AND WOODCOCK, 1994)

Value	Description
1	*Absolutely wrong* (*Very Wrong*): This answer is absolutely unacceptable.
2	*Understandable but wrong* (*Not Right*): Not a good answer. There is something about the site that makes the answer understandable, but there is clearly a better answer. This answer would pose a problem to users of the map.
3	*Reasonable or acceptable answer* (*Right*): Maybe not the best possible answer but it is acceptable; this answer does not pose a problem to the user.
4	*Good answer* (*Very Right*): Would be happy to find this answer given on the map.
5	*Absolutely right* (*Perfect*): No doubt about the match.

TABLE 3. ERROR MATRIX COMPARING THE REMOTELY SENSED MAP TO THE GROUND VISITED REFERENCE DATA

Landsat Classification	Reference Data: Barrens	MNT	MAT	Shrublands	Wet Tundra	Water	**Totals**	**User's Accuracy**
Barrens	**11**	·	·	·	·	·	11	**100.0%**
MNT	·	**51**	·	·	6	·	57	**89.5%**
MAT	·	12	**38**	1	·	·	51	**74.5%**
Shrublands	·	·	2	**17**	·	·	19	**89.5%**
Wet Tundra	·	2	·	·	**14**	·	16	**87.5%**
Water	·	·	·	·	·	**24**	24	**100.0%**
Totals	11	65	40	18	20	24	**178**	
Producer's Accuracy	**100.0%**	**78.5%**	**95.0%**	**94.4%**	**70.0%**	**100.0%**		
Classification Accuracy:		**P = 87.08%**	95% confidence interval for P: 82.07 – 91.95%					
		T_e = 84.49%	95% confidence interval for T_e: 78.73 – 90.25%					

was given a *right* or better fuzzy value (i.e., three, four, or five). This function is more optimistic than the *max* function because a match does not have to be the highest fuzzy value.

The usefulness of the *max* and *right* functions lies in the calculation of the differences between the total number of matches under each function. Because the *right* function measures how frequently the mapped class was given an acceptable or higher fuzzy value, the improvement of the *right* over the *max* function identifies the percentage of cases that had an acceptable answer but not the best answer. If the amount of improvement is considerable, then even though the *max* function indicates a category to be a problem for the map, the *right* function indicates that the effect on the map user may not be as large as the *max* function suggests (Gopal and Woodcock, 1994).

The second function is the *difference* function, a measure of the magnitude of correctness or incorrectness of a sample site. This function also indicates the degree of ambiguity or heterogeneity identified in land cover for each class. A frequency count is tabulated for each magnitude level of correctness or incorrectness, both for the entire map and for individual categories. This function is calculated by subtracting the fuzzy value assigned to the mapped class from the highest assigned fuzzy value among all the other classes. The resulting difference value can range between −4 and +4. For example, when a site's mapped class is given a fuzzy ranking of 5 and all other classes are given fuzzy values of 1, then the difference is 4. Positive difference values occur when a site is classified correctly and vice versa for negative difference values. We have slightly modified the analysis of the *difference* function as presented by Gopal and Woodcock (1994). Instead of calculating the arithmetic mean for each category as a whole, we calculated the arithmetic mean for the mismatches and matches within each category separately. The arithmetic mean of mismatches can range from −4 to −1 and the arithmetic mean of matches can range from 0 to +4. This alteration allows for the gleaning of more information on the degree of incorrectness and correctness within each land-cover class.

The third function is the *membership* function, a measure of the number of "sets" to which a sample site belongs. The purpose of this function is to analyze the source of errors. A fuzzy value of three or better means that the site is a member of that class' set. Because more than one class can have a fuzzy value of 3 or better, a site can have multiple set memberships; in theory, a site can be a member of zero through all defined sets (map classes). The *membership* function tabulates the number of cases that fall into a given membership category. For example, if a site has a fuzzy value of 4 for one class, a fuzzy value of 3 for another, and a fuzzy value of 1 for the remaining classes, then that case falls into membership category 2. Within set membership categories, the number of matches and mismatches — according to the *max function* — is indicated. The usefulness of the *membership* function is that it aids in understanding possible sources of errors in a map by indicating the nature of the environment in which errors are occurring. This function can help identify whether map errors occurred in ambiguous (i.e., heterogeneous) areas or if the classification performed poorly in unambiguous (i.e., homogeneous) areas.

Map Homogeneity Analysis

A study by Hammond and Verbyla (1996) showed that choosing a homogeneous, multiple-pixel sampling unit could introduce a considerable amount of "optimistic" bias (i.e., overestimation) in estimates of classification accuracy. However, the degree of bias depends on the degree of homogeneity in the map. Likewise, the degree of optimistic bias for accuracy estimates for individual land-cover classes will vary with the degree of homogeneity within each class. To gain a sense of the amount of optimistic bias caused by our sampling unit, we analyzed the homogeneity of the map using two different definitions of map homogeneity. First we analyzed the map for the amount of area covered by homogeneous 3- by 3-cell blocks. Second, we analyzed the map for the amount of area covered by continuous polygons of any shape that were greater than or equal to the area covered by nine pixels. This less-restrictive definition would allow identification of non-symmetrical homogeneous areas. We also analyzed the homogeneity of individual classes.

Results

Due to helicopter difficulties and poor weather conditions, transects one and four were not visited. To maintain the desired ratio of samples to land-cover classes, the field team randomly eliminated sites from the remaining transects. This resulted in a sample size of 178 sites out of the 263 targeted (Figure 2). The number of sites sampled in each land-cover class is listed in the Totals column of the error matrix (Table 3). Elimination of all samples along transect one and four potentially introduced major systematic bias into the sampling design, especially the lack of samples in the western portion of the watershed. However, this is unlikely to be a large problem for two reasons. First, the major trend in landscape variation within the watershed is north-south. Samples collected along transects two and three adequately cover this gradient. Second, there were no large, unique landscape types in the western area of the watershed that were not sampled elsewhere.

Based on the error matrix, the probability that any given location on the map is correct (P) is 87.1 percent with a 95 percent confidence interval of 82.1 to 92.0 percent. The central diagonal of the matrix highlights correctly classified cases. Eliminating chance agreement, the probability that any

TABLE 4. RESULTS OF THE *MAX* AND *RIGHT* FUNCTIONS. IMPROVEMENT COLUMN INDICATES THE DIFFERENCE IN ACCURACY BETWEEN THE *RIGHT* FUNCTION AND THE *MAX* FUNCTION

Map Class	Sample Sites	Max (M) - Best Answer				Right (R) - Correct				Improvement (R−M)	
		Matches		Mismatches		Matches		Mismatches			
Barrens	11	11	100%	0	0%	11	100%	0	0%	0	0%
MNT	57	51	89%	6	11%	52	91%	5	9%	1	2%
MAT	51	38	75%	13	25%	39	76%	12	24%	1	2%
Shrublands	19	17	89%	2	11%	18	95%	1	5%	1	5%
Wet Tundra	16	14	88%	2	13%	15	94%	1	6%	1	6%
Water	24	24	100%	0	0%	24	100%	0	0%	0	0%
Total	178	155	87%	23	13%	159	89%	19	11%	4	2%

given point on the map is correct (T_p) is 84.5 percent with a 95 percent confidence interval of 78.7 to 90.3 percent. All measures of producer's and user's accuracy were greater than 70 percent; only MAT had a user's accuracy below 80 percent and only MNT and Wet tundra had a producer's accuracy below 80 percent.

The categories of greatest concern, MNT and MAT, had user's accuracy of 89.5 percent and 74.5 percent, respectively, and producer's accuracy of 78.5 percent and 95 percent, respectively. All misclassifications in the MNT class were identified as Wet tundra on the ground, and all but one of the sites that were misclassified in the MAT class were identified as MNT on the ground. This accounts for 18 out of the 23 sample sites that were misclassified, indicating a pattern of non-random classification errors within these categories.

The results of the *max* and *right* functions are shown in Table 4. The *max* function indicates an overall accuracy of 87 percent, while the *right* function indicates an accuracy of 89 percent. The improvement column shows a 2 percent difference between the *max* and *right* functions. The range of improvement for individual categories was between 2 percent and 6 percent.

The results of the *difference* function can be found in Table 5. Sixty percent of the sites had a difference value of +4, which indicates that these sites were classified perfectly. Another 19 percent of the sites had difference values of either +2 or +3, which indicates that the classification of these sites was highly correct. Nine percent of the sites were marginally correct, with a difference value of 0 or +1 more than one best answer. Another 6 percent of the sites were incorrect, with difference values of either −1 or −2. The remaining 6 percent of the sites were very incorrect or perfectly incorrect, with values of either −3 or −4, respectively.

Also found in Table 5 are the arithmetic means of the mismatches and the matches. These values were calculated for the map as a whole and for each map class. For individual map classes, only one class (MNT) had an arithmetic mean of mismatches below −3, with the remaining five classes ranging between 0 and approximately −2. The arithmetic mean of matches for the six map classes ranged from between +4 and approximately +3. Wet tundra had the lowest arithmetic mean of matches at 2.79.

The results of the *membership* function can be found in Table 6. There were no sites that were members in zero sets or in more than two sets. Eighty-seven percent of all samples were members of only one set, and the remaining 13 percent were members of two sets. Of the 155 sites with single set membership, 138 (89 percent) were correctly classified. Of the 23 sites with multiple set memberships, 17 (73 percent) were classified correctly. Of the 23 samples classified incorrectly, 17 had single set membership and six had multiple set memberships.

Results of the analysis of map homogeneity are listed in Table 7. Homogeneity analysis using the first definition determined that approximately 22 percent of the watershed's pixels were at the center of a 3 by 3 homogeneous block of cells. Analysis using the second definition revealed that approximately 82 percent of the watershed was covered by continuous land-cover areas that were greater than or equal to the area of nine pixels. Analyzing the homogeneity of individual categories revealed that the Wet tundra and Shrublands categories were most heterogeneous in nature, while the MNT, Water, and MAT categories were most homogeneous in nature.

Discussion

Overall Classification Accuracy

For a map derived from satellite imagery, the measures of overall classification accuracy (87.1 percent) indicates that, through our classification method, we were generally able to correctly distinguish map classes. The results of our accuracy assessment compare favorably with the few accuracy assess-

TABLE 5. RESULTS OF THE *DIFFERENCE* FUNCTION, SHOWING THE FREQUENCY AND MAGNITUDE OF MISMATCHES AND MATCHES BASED ON THE *MAX* FUNCTION

Map Class	Sites	Mismatches				Matches					Arithmetic Mean of Mismatches	Arithmetic Mean of Matches
		−4	−3	−2	−1	0	1	2	3	4		
Barrens	11	0	0	0	0	0	0	0	0	11	0.00	4.00
MNT	57	4	1	0	1	7	1	9	5	29	−3.33	2.94
MAT	51	0	6	3	4	2	0	5	4	27	−2.15	3.42
Shrublands	19	0	1	0	1	0	3	3	1	10	−2.00	3.06
Wet Tundra	16	0	0	1	1	1	1	4	2	6	−1.50	2.79
Water	24	0	0	0	0	0	0	0	0	24	0.00	4.00
Total	178	4	8	4	7	10	5	21	12	107	−2.39	3.30
% of Total		2%	4%	2%	4%	6%	3%	12%	7%	60%		

TABLE 6. RESULTS OF THE *MEMBERSHIP* FUNCTION, SHOWING THE DISTRIBUTION OF ALL SAMPLE SITES (T), INCLUDING A BREAKDOWN OF MATCHES (M) AND MISMATCHES (N), WITHIN MEMBERSHIP CATEGORIES (USING THE *MAX* FUNCTION)

		Membership					
		1			2		
Map Class	Sites	T	M	N	T	M	N
Barrens	11	11	11	0	0	0	0
MNT	57	47	42	5	10	9	1
MAT	51	44	34	10	7	4	3
Shrublands	19	15	14	1	4	3	1
Wet tundra	16	14	13	1	2	1	1
Water	24	24	24	0	0	0	0
Total	178	155	138	17	23	17	6
% of Total Sites		87.08%	77.53%	9.55%	12.92%	9.55%	3.37%

ments that have been performed for Landsat-derived maps of remote Alaskan landscapes. Fleming (1988) used training set samples as reference data and estimated a *P* of 78.2 percent for a land-cover classification of Kanuti National Wildlife Refuge. Felix and Binney (1989) calculated the accuracy of a vegetation map of the Arctic National Wildlife Refuge (ANWR) to be 37 percent. More recently, Jorgenson *et al.* (1994) found their land-cover map of the coastal plain of ANWR to have a *P* of 63 percent, and Pacific Meridian Resources (1995) found their land-cover map of the western portion of the National Petroleum Reserve to have a *P* of 84 percent.

Using the results of the fuzzy sets analysis, the overall classification accuracy was analyzed in more detail. The improvement of the *right* function over the *max* function was small. This shows that most errors were of significance to the user. Antithetically, 2 percent of the map's error is not likely to have a strong impact on map users. The results of the *difference* function indicate that 79 percent of the map was highly or perfectly correct (*difference* values of +2 to +5). This implies that the classification scheme performed very well for a large portion of the mapped area, and that in general the map classes were well-defined and easy to distinguish from one another. This implication is reinforced by the overall arithmetic mean of matches (3.3), which indicates that matches were on average very right or perfect. Of the total set of sample sites, 22 (13 percent) appear to have considerable levels of land-cover heterogeneity (i.e., *difference* values between +1 and −1); of these, 15 were mapped correctly. Because there was more than one right answer for these sites, the classification given at these sites does not pose a serious problem for map users. Based on these numbers, sample sites with mixed land cover appear to be classified relatively well. Conversely, 6 percent of the sample sites had a high degree of classification error (i.e., difference values of −3 or −4), despite relatively homogeneous land cover.

TABLE 7. RESULTS OF MAP HOMOGENEITY ANALYSIS FOR INDIVIDUAL CATEGORIES AND THE OVERALL MAP (I.E., WATERSHED)

	Percent of area covered by homogeneous units	
	3 by 3 Blocks*	Continuous Areas**
Barrens	16.5%	77.8%
MNT	35.5%	87.2%
MAT	17.6%	89.1%
Shrublands	6.4%	68.1%
Wet tundra	2.7%	53.6%
Water	33.5%	85.1%
Watershed	22.4%	82.1%

*First definition of map homogeneity - 3 by 3 square blocks.
**Second definition of map homogeneity - continuous area, nine cells or more, any shape.

The *membership* function contributes additional information about the performance of the classification. First, 87 percent of the sample sites had single set membership which indicates that the area being mapped is fairly homogeneous with respect to map classes. Second, the results also suggest that the classification performed relatively well at sample sites with both homogeneous and heterogeneous land cover. Third, when errors did occur, they were most often at sites with relative homogeneity (17 of 23). This shows once again that, when errors did occur, they could be of considerable concern to the map user.

Category Accuracy

The more specific measures of category accuracy indicate that, generally, the classification performed well in distinguishing the differences between all land-cover classes. Barrens, Shrublands, and Water categories all had very high producer's and user's accuracy. This is likely to be in part due to their spectral distinctiveness. The producer's and user's accuracy values also indicate that the classification method had the most difficulty in separating the MNT, MAT, and Wet tundra categories.

The errors in the MNT and MAT categories were of greatest concern to us and other map users. Ten of 13 mismatches in the MNT category have only single set membership. This indicates that classification errors in this category tended to occur in homogeneous sites. Based on the spatial occurrence of these errors and our knowledge of the North Slope region, we determined that these errors occurred primarily in a moist nonacidic and moist acidic tundra ecotone in the northern foothills. Although dominated by plant taxa characteristic of moist nonacidic plant communities (Walker *et al.*, 1994), these misclassified sites also had generally well-developed tussocks (*Eriophorum vaginatum*) and high cover of willows, making them structurally and spectrally similar to MAT. In the MNT category, five of six errors had only single set membership, which indicates that most errors occurred in homogeneous sites. Although all of these sites had large amounts of standing water and species characteristic of Wet tundra, they also had a dense sedge cover that made them spectrally similar to MNT.

Performing spectral analysis based on the results of the accuracy assessment, we attempted to fix classification errors through refinement of the spectral definitions of each problem class. Areas that were mapped as MNT but were ground-truthed as Wet tundra did not appear to have spectral distinction, and no adjustments were made. However, we were able to better delineate the spectral boundary between MAT

and MNT. A map-wide spectral redefinition of the spectral boundary between these two classes resulted in the reclassification of half of the samples that were MAT on the ground but misclassified as MNT. This redefinition did not alter the classification of sample sites that were correctly classified in the MAT and MNT categories. It is likely, but uncertain, that this adjustment improved the accuracy of the MAT and MNT categories and the overall classification.

Confounding Issues of the Sampling Strategy

Several facets of the sampling strategy affected the reliability of our results. The facet with the largest effect on our results was the sampling unit. Due to optimistic bias caused by the sampling unit, our estimates of P and T_o can only be applied reliably to the 22 percent of the watershed that is covered by 3 by 3 blocks with homogeneous land cover. As with P and T_o, user's and producer's accuracy can only be applied reliably to the amount of area in each category that was covered by 3 by 3 homogeneous blocks (e.g., the accuracy of MNT applies to 35.5 percent). The results of the less-restrictive analysis of map homogeneity (82 percent homogeneous by definition) appears to indicate that the optimistic bias in the accuracy estimates is not as large as one would expect based on the initial analysis of map homogeneity. However, this should not be over-emphasized because this analysis is based on a loose definition of homogeneity, which includes linear features and allows for island polygons within homogeneous areas.

Another facet of the sampling methods which has an effect on our estimates of accuracy is the sampling design. Some researchers such as Congalton (1988) and Stehman (1992) have discussed the fact that the estimated variance, and thus confidence intervals, will be biased when based on systematic sampling designs. Also, Congalton (1988) discussed the potential of bias due to systematically sampling a data set with periodicity. The use of transects with irregular spacing of samples avoids regularities in the data, hidden or otherwise. However, the confidence intervals calculated for the accuracy estimates are likely to be based on a biased estimate of variance.

Final sample sizes of individual categories are the last part of the sampling strategy affecting the usefulness of our accuracy estimates. The final sample size in all but two categories was below 30. Due to this, estimates of user's and producer's accuracy have a greater chance of considerably over- or under-estimating the true values. The MNT and MAT categories had sample sizes large enough to have statistically reliable estimates of user's and producer's accuracy. Although the remaining four categories do not have statistically reliable estimates of user's and producer's accuracy, they can still be considered general indicators of their true accuracy.

Methodological Advantages

The methods used in the accuracy assessment contributed to obtaining meaningful results. The use of a PLGR and effectively planned sampling strategy produced a spatially accurate reference data set. The sampling strategy also minimized bias while maximizing resources. The use of fuzzy methods allowed for a more realistic field observation method, and the combination of fuzzy sets with the error matrix lent itself to a more precise analysis of the nature and source of classification errors.

The results of the accuracy assessment indicate a fairly accurate image classification. Several factors contributed to this, including few map categories, the generally flat or gently rolling landscape, the simple vegetation canopy, prior knowledge, and previous experience with two Landsat MSS-derived classifications within the basin (Walker and Acevedo, 1987; D.A. Walker, unpublished data, 1985). Probably the most important factor was the development of a classification that represented spectrally distinct and ecologically meaningful land-cover classes based on detailed phytosociological information (Walker *et al.*, 1994).

Conclusions

- The accuracy assessment indicates that the classification of the Kuparuk River basin was highly accurate and performed relatively well in both homogeneous and heterogeneous areas. However, the estimates are likely to be optimistically biased and can only be applied reliably to the 22 percent of the watershed covered by 3 by 3 homogeneous blocks.
- All of the accuracy estimates for the individual categories indicate mid to high levels of accuracy. However, the statistical reliability of accuracy estimates for all but the MNT and MAT categories are affected by small sample sizes. Also, all of these estimates are affected by optimistic bias. Despite statistical unreliability and potential bias, the estimates of individual category accuracy are still useful indicators of true accuracy.
- The majority of classification errors occurred in the MAT and MNT. Based on our analysis of errors, we determined that we could not correct the errors between Wet tundra and MNT. However, we did redefine the spectral boundary between MAT and the MNT classes, which changed the classification of half of these sample sites to MAT. This most likely improved the classification.
- The high accuracy of the classification and the meaningfulness of the accuracy assessment's results are attributable to the following factors: a well-planned sampling strategy, spatially accurate reference data, a simple and spectrally distinct classification based on detailed phytosociological information, and prior knowledge and experience in creating land-cover maps for the region.
- The combined use of fuzzy sets theory and an error matrix allowed for a more precise analysis of the classification accuracy and its errors. This gave greater insight into classification reliability and usefulness than either method would have if used alone.

Acknowledgments

The work in this paper, performed by members of the Tundra Ecosystem Analysis and Mapping Lab, was supported by NSF grants OPP-9318530 and OPP-9415554. The contribution of Dr. F.E. Nelson was supported by NSF grant OPP-9612647. The authors would like to thank Dr. Marilyn Walker for crucial comments that greatly improved the focus of the paper, the Alaska Office of the BLM for providing use of the PLGR GPS unit, and to the anonymous reviewers whose comments helped significantly improve this work.

References

Berry, B.J.L., and A.M. Baker, 1968. Geographic sampling, *Spatial Analysis: A Reader in Statistical Geography* (B.J.L. Berry and D.F. Marble, editors), Prentice Hall, Inc., Englewood Cliffs, N.J., pp. 91–100.

Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy, *Photogrammetric Engineering & Remote Sensing*, 48(3):431–439.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20(1):37–40.

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 & Remote Sensing*, 54(5):593–600.

———, 1991. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of the Environment*, 37: 37–46.

Congalton, R.G., and K. Green, 1993. A practical look at the sources of confusion in error matrix generation, *Photogrammetric Engineering & Remote Sensing*, 59(5):641–644.

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

Federal Radionavigation Plan, 1994. Department of Defense and Department of Transportation, Springfield, Virginia, 229 p.

Felix, N.A., and D.L. Binney, 1989. Accuracy assessment of a Landsat-assisted vegetation map of the coastal plain of the Arctic National Wildlife Refuge, *Photogrammetric Engineering & Remote Sensing*, 55(4):475–478.

Fleming, M.D., 1988. An integrated approach for automated covertype mapping of large inaccessible areas in Alaska, *Photogrammetric Engineering & Remote Sensing*, 54(3):357–362.

Foody, G.M., 1992. On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering & Remote Sensing*, 58(10):1459–1460.

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

Gopal, S., and C. Woodcock, 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets, *Photogrammetric Engineering & Remote Sensing*, 60(2):181–188.

Hammond, T.O., and D.L. Verbyla, 1996. Optimistic bias in classification accuracy assessment, *International Journal of Remote Sensing*, 17(6):1261–1266.

Hay, A.M., 1979. Sampling designs to test land-use map accuracy, *Photogrammetric Engineering & Remote Sensing*, 45(4):529–533.

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

Hudson, W., and C. Ramm, 1987. Correct formulation of the Kappa coefficient of agreement, *Photogrammetric Engineering & Remote Sensing*, 53(4):421–422.

Hutchinson, C., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement, *Photogrammetric Engineering & Remote Sensing*, 48(1):123–130.

Janssen, L.L.F., and F.J.M. van der Wel, 1994. Accuracy assessment of satellite derived land-cover data: A review, *Photogrammetric Engineering & Remote Sensing*, 60(4):419–426.

Jorgenson, J.C., P.E. Joria, T.R. McCabe, B.E. Reitz, M.K. Raynolds, M. Emers, and M.A. Williams, 1994. *User's Guide for the Land-Cover Map of the Coastal Plain of the Arctic National Wildlife Refuge*, U.S. Dept. of the Interior, U.S. Fish and Wildlife Service Region 7, Anchorage, Alaska, 46 p.

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

Markon, C., 1992. Land cover mapping of the Upper Kuskokwim Resource Management Area, Alaska, using Landsat and a digital database approach, *Canadian Journal of Remote Sensing*, 18(2): 62–70.

Markon, C., and W. Kirk, 1994. *Development of a Digital Land Cover Data Base for the Selawik National Wildlife Refuge*, U.S. Geological Survey, Open-File Report 94-627, 14 p. plus appendix.

Morrissey, L.A., and R.A. Ennis, 1981. *Vegetation Mapping of the National Petroleum Reserve in Alaska Using LANDSAT Digital Data*, U.S. Geological Survey, Open File Report 81-315, 25 p.

Naesset, E., 1995. A method to test for systematic differences between maps and reality using error matrices, *International Journal of Remote Sensing*, 16(16):3147–3156.

Pacific Meridian Resources, 1995. *National Petroleum Reserve Alaska Land-Cover Inventory: Phase 1 Western NPR-A*, Final Report, Pacific Meridian Resources Sacramento, California, 30 p.

Rosenfield, G.H., 1986. Analysis of thematic map classification error matrices, *Photogrammetric Engineering & Remote Sensing*, 52(5): 681–686.

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

Stehman, S.V., 1992. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 58(9):1343–1350.

Story, M., and R.G. Congalton, 1986. Accuracy assessment: A user's perspective, *Photogrammetric Engineering & Remote Sensing*, 52:397–399.

Stow, D., B. Burns, and A. Hope, 1989. Mapping Arctic tundra vegetation types using digital SPOT/HRV-XS data: A preliminary assessment, *International Journal of Remote Sensing*, 10(8):1451–1457.

Thompson, S.K., 1992. *Sampling*, John Wiley & Sons, New York, New York, 343 p.

van Genderen, J.L., and B.F. Lock, 1977. Testing land-use map accuracy, *Photogrammetric Engineering & Remote Sensing*, 43(9): 1135–1137.

Verbyla, D.L., and T.O. Hammond, 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids, *International Journal of Remote Sensing*, 16(3):581–587.

Walker, D.A., and W. Acevedo, 1987. *Vegetation and a Landsat-Derived Land Cover Map of the Beechey Point Quadrangle, Arctic Coastal Plain, Alaska*, CRREL Report 87-5, U.S. Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire, 63 p. (plus map).

Walker, D.A., K.R. Everett, W. Acevedo, L. Gaydos, J. Brown, and P.J. Webber, 1982. *Landsat-Assisted Environmental Mapping in the Arctic National Wildlife Refuge, Alaska*, CRREL Report 82-27, U.S. Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire, 59 p.

Walker, D.A., and M.D. Walker, 1991. History and pattern of disturbance in Alaskan arctic terrestrial ecosystems: A hierarchical approach to analyzing landscape change, *Journal of Applied Ecology*, 28:244–276.

———, 1996. Terrain and vegetation of the Imnavait Creek Watershed, *Landscape Function: Implications for Ecosystem Disturbance, a Case Study in Arctic Tundra* (J.F. Reynolds and J.D. Tenhunen, editors), Springer-Verlag, New York, New York, pp. 73–108.

Walker, M.D., D.A. Walker, and N.A. Auerbach, 1994. Plant communities of a tussock tundra landscape in the Brooks Range Foothills, Alaska, *Journal of Vegetation Science*, 5:843–866.

Walker, M.D., D.A. Walker, and K.R. Everett, 1989. *Wetland Soils and Vegetation, Arctic Foothills, Alaska*, Report 89(7), U.S. Fish and Wildlife Service.

Weller, G., F.S. Chapin, K.R. Everett, J.E. Hobbie, D. Kane, W.C. Oechel, C.L. Ping, W.S. Reeburg, D. Walker, and J. Walsh, 1995. The Arctic Flux Study: a regional view of trace gas release, *Journal of Biogeography*, 22:365–374.

(Received 7 May 1997; accepted 25 September 1997; revised 14 November 1997)