

**CEO Technical Report 218**

March 29, 2001

**Final Report**

Submitted to  
Lockheed Martin  
for the MRLC Accuracy Assessment Service Agreement Entitled

**Accuracy Assessment of the Region 5 Dataset of the MRLC Consortium's  
National Land Cover Data**

Submitted by



**Center for Earth Observation  
North Carolina State University**

U.S. Congressional District 4

**Final Report**

Submitted to  
Environmental Protection Agency (EPA)  
for the MRLC Accuracy Assessment Service Agreement Entitled

**Accuracy Assessment of the Region 5 Dataset of the MRLC Consortium's  
National Land Cover Data**

Submitted by

Siamak Khorram, Center for Earth Observation, NCSU, Principal Investigator  
Joseph Knight, Center for Earth Observation, NCSU, Project Manager  
Halil Cakir, Center for Earth Observation, NCSU, Photo Interpreter  
Zhiyan Mao, Center for Earth Observation, NCSU, Photo Interpreter  
Okan Pala, Center for Earth Observation, NCSU, Photo Interpreter  
Hui Yuan, Center for Earth Observation, NCSU, Photo Interpreter



Center for Earth Observation  
North Carolina State University

U.S. Congressional District 4

Contact:  
[Khorram@ncsu.edu](mailto:Khorram@ncsu.edu)

Box 7106, North Carolina State University  
Raleigh, North Carolina 27695-7106  
Phone: (919) 515-3430  
FAX: (919) 515-3439

# Accuracy Assessment of the Region 5 Dataset of the MRLC Consortium's National Land Cover Data

## Summary

The Center for Earth Observation (CEO) at North Carolina State University (NCSU) was funded by the Lockheed Martin, Inc. to evaluate and analyze the accuracy of the Region 5 dataset of the MRLC National Land Cover Data (NLCD). This dataset provides a consistent and conterminous land cover map of the lower 48 States at approximately an Anderson Level II thematic detail. To conduct the accuracy assessment for the Region 5 (Midwest States), the US EPA and USGS EROS Data Center (EDC) in Sioux Falls, SD provided the Center for Earth Observation (CEO) of North Carolina State University (NCSU) with the following data and information: the geographic locations of 1800 reference sites in a digital file of X, Y map coordinates; and NAPP photos containing these reference sites; Landsat Thematic Mapper bands 3, 4, and 5 image data for the area surrounding each reference site. The accuracy assessment approach included formal training in photo interpretation, on-the-job training of photo interpreters, photo interpretation and image analysis, frequent group meetings and discussion, interpretation agreement analysis using overlapping reference sites, hierarchical QA/QC procedures, and generation and analysis of error matrices. Results include the reference database of 1643 sample sites collected by photo interpretation and image analysis, and accuracy assessment as compared to the classified data. The overall classification accuracy of the MRLC Region 5 data for the primary interpreted class was 36%. When alternate classes were included, the accuracy was 48%. When the primary or alternate matched any pixel in a 3 by 3 pixel area, the overall accuracy was 71%. A single factor ANOVA test ( $\alpha=0.05$ ) was used to compare the interpretations of 270 overlap sites. The test failed to reject the null hypothesis that there is no difference between the four interpreters ( $P=0.246$ ). Therefore, we conclude with confidence that inter-interpreter variation did not have a statistically significant affect on the accuracy estimates. The base overall accuracy (36%) of the Region 5 NLCD seems quite low. However, we believe that this number, derived from a simple site for site comparison, is based on a strict interpretation of overall accuracy, and so may underestimate the usability accuracy of the classification. This estimate does not take into account complicating factors such as errors in locating sites on the photos, seasonal and annual differences in photo and image acquisition dates, LU/LC change between photo and image acquisition dates, difficulty in separating the row crops vs. small grains classes, difficulty in identifying the shrubland and natural grassland classes, ambiguity of class definitions, heterogeneity of land cover, and the inclusion of black and white photos. Recommendations for future studies include: field sampling, revision of Region 5 classes, decreasing the time difference between the acquisition dates of the photos and imagery, use of only stereo photographs, and use of only CIR photographs.

## **Acknowledgments**

The results reported here were generated through a contract funded by Lockheed Martin, Inc. The views expressed in this report are those of the authors and do not necessarily reflect the views of Lockheed Martin, Inc. or any of its sub-agencies. The authors would like to thank Larry Tinney and Dave Williams of Lockheed Martin, Inc. and Jim Wickham of the U.S. Environmental Protection Agency for their support on this project. The authors would also like to thank Dr. Heather Cheshire of the Center for Earth Observation for contributing her expertise in photo interpretation and Ms. Linda Babcock of the Center for Earth Observation for organizational support throughout the duration of this project.

## 1. Introduction

The Multi-Resolution Land Characteristics (MRLC) Consortium, including the U.S. Geological Survey (USGS), the U.S. Environmental Protection Agency (EPA), and other Federal agencies, has completed the National Land Cover Data (NLCD) program. This dataset provides a consistent and conterminous land cover map of the lower 48 States at approximately an Anderson Level II thematic detail. The program used Landsat Thematic Mapper (TM) 30-meter resolution imagery as the baseline data. The central goal of the program is to provide a regionally consistent land cover product for use in a broad range of applications (Lunetta *et al.*, 1998). Each of the ten Federal regions was mapped independently. This project presents the results of the accuracy assessment for Region 5 of the NLCD.

Accuracy assessment must be an integral component of any remote sensing-based mapping project. There are two primary motivations for this accuracy assessment:

- ?? To provide an overall assessment of the reliability of a land cover map from TM data
- ?? To estimate the class-specific and overall accuracy of the classification

Quantitative accuracy assessment of large-area land cover maps, produced from remotely sensed data, involves comparing thematic maps with reference data (Congalton, 1991). Since there is no suitable existing ground reference data that can be used for all federal regions, a practical and statistically sound sampling plan was designed to characterize the accuracy of common and rare classes for the NLCD. The MRLC program has adopted a two-stage cluster sampling scheme and selected National Aerial Photography Program (NAPP) photography as the primary source of reference data.

To conduct the accuracy assessment for the Region 5 (Midwest States), the US EPA and USGS EROS Data Center (EDC) in Sioux Falls, SD provided the Center for Earth Observation (CEO) of North Carolina State University (NCSU) with the following data and information:

- ?? The geographic locations of 1800 reference sites in a digital file of X, Y map coordinates;
- ?? NAPP photos containing these reference sites;
- ?? Landsat Thematic Mapper bands 3, 4, and 5 image data for the area surrounding each reference site;

Using these data and following the MRLC protocol for reference data collection and evaluation for accuracy assessment, CEO at NCSU provided computer resources and staff to conduct reference data collection. This was accomplished by interpreting on NAPP photography the appropriate land use and land cover (LU/LC) class(es) for the pre-selected 1800 sites, while maintaining high standards of quality assurance (QA) and quality control (QC).

## 2. Objectives

The objectives of the project were (1) to produce reference data for accuracy assessment by interpreting NAPP photography, and (2) to subsequently analyze the accuracy of the MRLC Region 5 NLCD product created by the MRLC Consortium.

## 3. Study Area and Sample Determination

The study area for this project was EPA Region 5, which is composed of the six states of Ohio, Indiana, Illinois, Michigan, Wisconsin, and Minnesota.

A stratified random sampling scheme was used to select 1800 sample sites – 100 for each of the 18 classes present in Region 5. This was done as follows: First, the Region was divided into frames measuring 60km by 30km. Next, each frame was further subdivided into 50 Primary Sampling Units (PSUs). Next, one PSU was selected at random from each frame. Finally, the 1800 sample sites were selected by stratifying based on NLCD land cover class within the selected PSUs.

The sampling design was based on the following criteria:

- ?? Ensuring the objectivity of sample selection and validity of statistical inferences drawn from the sample data;
- ?? Distributing the sites spatially across the region to ensure adequate coverage of the entire region;
- ?? Reducing the variance for estimated accuracy parameters;
- ?? Ensuring a low cost approach in terms of budget and time;
- ?? Being feasible to implement and analyze.

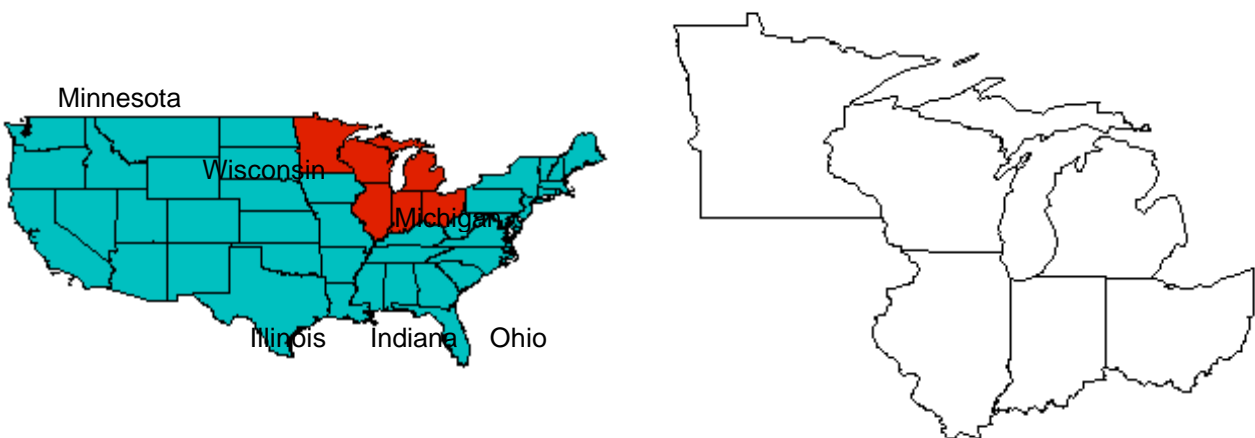


Figure 1. The study area: MRLC Region 5

## 4. Training

To provide equal footing for and consistency among the interpreters, a comprehensive training program was devised. The program consisted of a several full-day training sessions and "on the job" training. The formal classroom training sessions were led by experienced airphoto interpretation and photogrammetry instructors. The training sessions included the following:

- ?? Discussion of color theory and photo interpretation principles and techniques;
- ?? Understanding of the class definitions;
- ?? Interpretation of over 100 sample sites of different classes during the training sessions, followed by interactive discussions about potential discrepancies;
- ?? Creation of sample sites for later reference; and
- ?? Repetition of photo interpretation practice after the sessions.

The focus was on situations that the interpreters would encounter during the project. Each participant was presented with approximately 100 pre-selected sites and was asked to provide their interpretation of the land cover for these sites. Their calls were analyzed and subsequently discussed to minimize any misconceptions. During the "on the job" portion of the training, the interpreters were randomly assigned 199 sites from the 1800 sample that they interpreted as a group. This approach was used to "calibrate" the interpreters so that there would be as little inter-interpreter variation as possible. Their progress was monitored daily for accuracy and proper methodology. The interpreters kept a log of their calls and the sites for which they were uncertain about the land cover classes. Problem sites were discussed until the group reached a consensus on the proper class for each site. This calibration procedure provided a solid foundation for further independent photo interpretation work.

## 5. Preprocessing of Imagery

The TM imagery for the study area provided by EPA was composed of image "chips" for each of the PSUs. These chips provided image data for the PSUs plus a 10 pixel buffer. Each chip was composed of three Arc/GRID format files containing Landsat TM bands 5 (Infrared), 4 (Infrared), and 3 (Red). Data preprocessing included the following steps:

- ?? Data decompression. The image chips were provided in Unix tar format. The chips were decompressed and organized by PSU number.
- ?? Data import. The Arc/GRID format chips were imported into Erdas Imagine format.
- ?? Layer stack. The individual band images were stacked to create image files with all three bands.
- ?? Backup. After all the preprocessing procedures were complete, the images were backed up to CD and tape for data integrity.

## 6. Photo Interpretation

### 6.1. Photo Interpretation Protocol

The photo interpreters followed a common interpretation protocol for all sites.

- ?? Photo interpretation was based on the NAPP photographs.
- ?? To locate the sites on the photos, the reference site locations were plotted on the TM data. The corresponding locations on the photos were found by examining the context on TM color composite image.
- ?? When a photo interpreter was interpreting the sites, land cover class and related information such as homogeneity, confidence of location, and confidence of interpretation was recorded for later analysis.
- ?? When there was a difference in acquisition times between image and photo, the interpreted results were based on the photo.
- ?? The interpreters examined an approximately 3 by 3 TM pixel area on the photos to determine the correct class for each reference site.
- ?? For heterogeneous or confused sample sites, both primary and alternate land cover classes were recorded. The two classes are considered to be equally correct.

### 6.2. Sample Site Preparation and Photo Interpretation Team

The PSUs were assigned randomly to one of the four photo interpreters. Due to manpower and time limitations, the interpreters did not complete equal numbers of sites. The number of sites completed by each of the interpreters is summarized in the following table:

Interpreter #1	270
Interpreter #2	417
Interpreter #3	365
Interpreter #4	392
Training (all 4)	199
Total	1643

Only 1643 of the 1800 random points were completed because of three factors: missing or unavailable photos for parts of the study area, incorrect photos supplied, and unavailable image data for parts of the study area. The four interpreters were assigned additional sites from each of the other three interpreters' sites for use in consistency checking. The objective was to have a 15% overlap for the total 1800 site sample. The overlapping 270 sites were interpreted by all four interpreters and were used for quality assurance purposes.

### 6.3. Interpretation Procedures

The provided reference site locations were plotted on the Landsat TM image chips using Erdas Imagine software. The photo interpreters could then precisely locate each site on the image. Then, based on the context from the Landsat TM False Color Composite (FCC) image, the interpreters located the reference site on the corresponding photo. The interpreters examined a 3 by 3 TM pixel window around the site on the aerial photo and determined the LU/LC class label for each sample site according to the NLCD classification scheme. In the event that the site was heterogeneous or there was confusion as to the correct class, the interpreters were given the option of specifying an alternate LU/LC class. The interpreters also indicated whether the site changed significantly between the acquisition dates of the photo and image (Boolean 0 or 1 value), whether the site could be considered very heterogeneous (Boolean 0 or 1 value), their confidence in their interpretation (1 – 4), and their confidence that they correctly located the site on the photo (1 – 4). Please see the Appendix for the linguistic scales used to determine interpretation and location confidence and for an enumeration of the NLCD classes for Region 5. ***It should be noted that the photo interpreters did not have access to the MRLC classification values during the interpretation process.***

Following interpretation, each interpreter was responsible for entering each site's data into a computer database. This database was designed to store and organize the data collected in this project. This database was created using Microsoft Access and consists of one table to store data and one two-page form to enter, edit, and display the data. The form is composed of two parts: the first part shows the reference data for each site, and the second part is for the fuzzy analysis (not part of this contract and data; to be conducted in the future subsequent to completion of this project).

The first part of the form contains the following information: site ID number, coordinates, photo acquisition date, PSU ID, photo ID, primary class, alternative class (if any), general site description, likelihood of site change between photo and image acquisition dates, site heterogeneity, interpretation confidence, site location confidence, and comments.

The second part of the form, the fuzzy part, allowed the interpreter to interpret the membership of each site in the different classes. Membership probabilities include: *Absolutely Correct*: there is no doubt that this is the correct class for the site; *Probably Correct*: it is likely that this is the correct class; *Acceptable*: maybe not the best possible answer but it is reasonable and acceptable; *Probably Incorrect*: not a good answer and there is clearly a better class; *Absolutely Incorrect*: totally unacceptable. The fuzzy results were not included in the results reported in this paper. Only the primary and alternate photo interpretation classes were used. Except where noted below, only the primary class was used for accuracy assessment as per the project objectives. The additional fuzzy information could be helpful in the future research on this subject.

The database was designed to have built-in error-checking and data-validation rules. For example, coordinates entered that are not within the study area would be rejected.

Also, typographical errors are minimized by having the interpreters select items from menus whenever possible. In addition, comprehensive training was provided on database use to all project personnel.

Following the photo interpretation analysis, error matrices were created using a Visual Basic macro within Microsoft Excel. This automated approach reduced the possibility of errors associated with the data entry and manual generation of error matrix tables.

### **7. Quality Assurance/Quality Control (QA/QC)**

During the interpretation process, QA/QC procedures were enforced as described in the Project Organization Chart, shown in Table 1. In addition to the interactions between the CEO group and the scientists from Lockheed Martin, a large number of meetings of the CEO project team, hosted by the Principal investigator and Project Supervisor, were held to review project progress and discuss problems encountered. In addition, impromptu meetings among members of the CEO group on an informal basis provided an opportunity to discuss problems that occurred and, in most cases, provide a solution on the spot. Information shared in these meetings helped the interpretation quality and consistency. The following QA/QC procedures were used:

- 1) At the completion of each day’s work, the photo interpreters reviewed the points they finished that day for completeness, consistency, and to correct any typographical errors.
- 2) In addition to resolving problems for certain sites, the Project Supervisor performed random QA/QC checks during the photo interpretation process to ensure that the interpreters’ performance were acceptable and corresponded with the established guidelines.
- 3) The Project Supervisor and Principal Investigator randomly checked 10% of the data sheets against the photos to ensure that the interpretation was correct.

#### **Project Organization**

<b>Photo Interpreters</b>	Interpretation of photos, Data entry, Log keeping
<b>Project Supervisor</b>	Random consistency checks, Examining problem sites, Random checking of all sites, Overall QA/QC
<b>Principal Investigator</b>	Procedure establishment, Discussions of interpretation issues, Random checking, Overall QA/QC, Reports and Presentations

Table 1. Project team organization.

## 8. Results

The interpreted results of the 1643 sample sites were used to produce various analyses of the accuracy of the Region 5 NLCD data.

### 8.1 Standard Accuracy Analysis

The following matrix shows the results of a site-for-site comparison of the NLCD data vs. the reference data. The numbers across the top and sides of the matrices are the 18 NLCD land cover classes present in Region 5. (Please see the Appendix for class definitions).

		Interpreted Class																	Tot	%		
		11	21	22	23	31	32	33	41	42	43	51	71	83	82	81	85	91			92	
NLCD Class	11	32							1			1								34	94	
	21		66	1	5				2		3	2	1	1	3	1	4	1	1	91	73	
	22			70	11	11			1		1				2		3			99	11	
	23	1	9	1	59	2			2		1	2	2	2	5		6	1	1	94	63	
	31	12	2		1	40	4		3		3				2	17	1	2		2	89	45
	32	15			1	9	1	23	4	2		5	11	4	1	3	1	1	13	4	98	23
	33	1				4	3	4	30	14	1	6	8	2	3	4	4	1	1	6	92	33
	41	1	2			2			5	44	2	11	5	2	1	3	3	1	10	4	96	46
	42	2						1	5	4	35	31	3	1					11		95	37
	43	2	1			1			6	17	7	47	1			1	1		8	1	93	51
	51	1	1			2	3	1	30	11	1	6	19	2	1	2			7	9	96	20
	71			3		1	1	2	3	20	3	3	10	3	1	22	1	4	10	7	94	3
	83			1		3				1			1	7	10	62	8	1	1	2	97	10
	82			3		1	1			3		1	2	1	1	82	1	1	2		99	83
	81			3		5			2	4		4		7	6	46	6	7	4		94	6
	85	1	31			4	1	2	1	2	1	1	2	5		8		34	1	1	95	36
	91	3	1						3	11	10	16	11						32	7	94	34
	92	8				2	1		2	6	3	10	4	4	3	15	1		11	23	93	25
Tot	79	193	14	110	53	37	91	148	63	149	82	41	32	277	28	65	113	68	1643			
%	41	34	79	54	75	62	33	30	56	32	23	7	31	30	21	52	28	34			36.3	

Table 2. This matrix shows the interpreted results for the 1643 completed sites versus the NLCD classification for those sites. The overall percent accuracy is 36.3%.

### 8.2 Extended Analysis

To account for positional and correspondence errors (Congalton and Green, 1993), additional analysis was completed for this project. We examined the relationships of the primary and alternate interpreted classes with: 1) the center NLCD pixel (based on the provided sample site coordinates), 2) the most common value in a 3x3 window around the center pixel in the NLCD classification, and 3) Any value in a 3x3 window around the center pixel. The **overall accuracies** under these scenarios are as follows:

	# Samples	% Acc.
Primary interpreted matches classified pixel	596 / 1643	36%
Primary is most common in classified 3x3 area	645	39%
Primary matches any pixel in classified 3x3 area	955	58%
Primary or alternate matches classified pixel	787	48%
Primary or alternate is most common in 3x3 area	852	52%
Primary or alternate matches any pixel in 3x3	1174	71%

Table 3. Overall accuracies when taking into account primary and alternate interpreted classes and a 3x3 window around the NLCD center pixel. "Interpreted" refers to the classes chosen during the airphoto interpretation process, "primary class" and "alternate class" are the most likely land cover classes for a particular site, and "classified" refers to the MRLC classification result for that site.

We extended the previous results by summarizing the values by NLCD land cover class. This is presented in Tables 5 and 6.

Class	Num Sites	Primary PI		Primary PI is Mode of 3x3		Primary PI Matches any 3x3		Prim or Alt PI Matches NLCD		Prim or Alt PI is Mode of 3x3		Prim or Alt PI Matches any 3x3	
		#	%	#	%	#	%	#	%	#	%	#	%
		<b>11</b>	79	32	40.5	36	45.6	49	62.0	40	50.6	42	53.2
<b>21</b>	193	66	34.2	84	43.5	134	69.4	88	45.6	101	52.3	148	76.7
<b>22</b>	14	11	78.6	10	71.4	11	78.6	12	85.7	11	78.6	12	85.7
<b>23</b>	110	59	53.6	48	43.6	75	68.2	64	58.2	58	52.7	88	80.0
<b>31</b>	53	40	75.5	31	58.5	40	75.5	44	83.0	37	69.8	46	86.8
<b>32</b>	37	23	62.2	23	62.2	23	62.2	24	64.9	23	62.2	25	67.6
<b>33</b>	91	30	33.0	29	31.9	41	45.1	47	51.6	46	50.5	59	64.8
<b>41</b>	148	44	29.7	83	56.1	123	83.1	52	35.1	92	62.2	128	86.5
<b>42</b>	63	35	55.6	32	50.8	43	68.3	39	61.9	36	57.1	47	74.6
<b>43</b>	149	47	31.5	37	24.8	73	49.0	66	44.3	64	43.0	104	69.8
<b>51</b>	82	19	23.2	14	17.1	24	29.3	28	34.1	27	32.9	35	42.7
<b>71</b>	41	3	7.3	1	2.4	3	7.3	11	26.8	10	24.4	16	39.0
<b>83</b>	32	10	31.3	10	31.3	10	31.3	12	37.5	16	50.0	25	78.1
<b>82</b>	277	82	29.6	111	40.1	173	62.5	137	49.5	163	58.8	217	78.3
<b>81</b>	26	6	23.1	9	34.6	13	50.0	9	34.6	11	42.3	16	61.5
<b>85</b>	65	34	52.3	31	47.7	40	61.5	41	63.1	41	63.1	50	76.9
<b>91</b>	113	32	28.3	35	31.0	55	48.7	45	39.8	48	42.5	71	62.8
<b>92</b>	68	23	33.8	21	30.9	25	36.8	28	41.2	26	38.2	32	47.1
<b>Total</b>	1643	596	36.3%	645	39.3%	955	58.1%	787	47.9%	852	51.9%	1174	71.5%

Table 5. Summary of extended analysis represented by the number of sites and percent in each class.

The following table (Table 6) breaks down these same six scenarios by interpreter. We list the number of points and the percent accuracy vs. the NLCD for each of the four interpreters plus the 199 Training points.

### Interpreter

	#1		#2		#3		#4		Training	
	#	%	#	%	#	%	#	%	#	%
Prim. interpreted matches class. pixel	114	42	152	36	124	34	142	36	74	37
Prim. is most common in class. 3x3 area	120	44	166	40	142	39	138	35	79	40
Prim. matches any pixel in class. 3x3 area	185	69	235	56	202	55	214	55	119	60
Prim. or alt. matches class. pixel	133	49	189	45	170	47	188	48	107	54
Prim. or alt. is most common in 3x3 area	151	56	210	50	188	52	191	49	112	56
Prim. or alt. matches any pixel in 3x3	222	82	279	67	255	70	265	68	153	77

Table 6. Scenario breakdown by interpreter. The # column is the number of the interpreter's points that satisfy each condition. The % column gives the percent accuracy vs. the NLCD for each condition.

### 8.3 Overlap Analysis

Upon completion of the interpretation, the 270 overlap sites were assessed for inter-interpreter variation using standard statistical procedures. A single factor ANOVA test ( $\alpha=0.05$ ) was used to compare the interpretations of the overlap sites. The test failed to reject the null hypothesis that there is no difference between the four interpreters ( $P=0.246$ ). Therefore, we conclude with confidence that *inter-interpreter variation did not have a statistically significant affect on the accuracy estimates*. The results of the test are summarized below:

#### Anova: Single Factor

##### ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	117.8	3	39.26667	1.383577	0.24626	2.613177
Within Groups	30537.47	1076	28.38055			
Total	30655.27	1079				

Table 7. Overlap ANOVA.

### 8.4 High Confidence Sites Analysis

We attempted to account for the affect that problematic sites had on the error assessment by constructing another error matrix in which only sites with high location and interpretation confidence (on the linguistic scales) were included. High confidence was defined as a value of 3 or 4 on the four-point confidence scale. The results are summarized in the following table.

#### Interpreted Class

	11	21	22	23	31	32	33	41	42	43	51	71	83	82	81	85	91	92	Tot	%
11	32							1			1								34	94
21		63	1	4				2		2	2	1	2	1	4	1	1		84	75
22			68	11	10			1		1				2	3				96	11

NLCD Class	23	1	9	1	56	2			2	1	2	1	1	4		5	1	1	87	64		
	31	12	2		1	37	1		3	3			2	16		2		2	81	46		
	32	15			9	1	22	4	2	4	11	3		3	1	1	13	4	93	24		
	33	1			4	3	3	30	13	1	5	7	1	3	3	3	1	6	84	36		
	41	1	2		2			3	43	2	10	5	2	1	3	2	1	10	2	89	48	
	42	2						4	4	34	30	3	1		2			9	89	38		
	43	1	1		1			5	17	7	41	1			1	1		6	82	50		
	51	1	1		2	3	1	30	11	1	5	19	2	1				7	9	93	20	
	71		2		1	1	1	2	19	2	3	8	1	1	22	1	3	10	4	81	1	
	83		1		3				1			1	6	10	60	8		1	2	93	11	
	82		3		1				3		1	2	1	1	80		1	2		95	84	
	81		2		5			2	3		4		7	6	43	6	6	4		88	7	
	85	1	30		4	1	1		2	1	1	2	3		5			33	1	1	86	38
	91	2	1					3	11	10	14	10							29	7	87	33
	92	8			2	1			5	3	8	3	3	3	15	1		8	22		82	27
	Tot	77	185	13	105	49	29	83	143	61	133	77	31	30	261	24	60	102	61	1524		
	%	42	34	85	53	76	76	36	30	56	31	25	3	33	31	25	55	28	36			37.3

Table 8. Accuracy analysis of only the 1524 sites with high location and interpretation confidence

### 8.5 Merged Categories Analysis

Many of the most significant sources of error in the classification were present within the urban and agriculture classes. Presented below is an error matrix with the urban and agricultural classes merged so that there is one urban class (20) and one agricultural class (80).

	11	20	31	32	33	41	42	43	51	71	80	91	92	Tot	%
11	32					1			1					34	
20	1	233	2			5		5	4	3	27	2	2	284	
31	12	3	40	4		3		3			22		2	89	
32	15	10	1	23	4	2		5	11	4	6	13	4	98	
33	1	4	3	4	30	14	1	6	8	2	12	1	6	92	
41	1	4			5	44	2	11	5	2	8	10	4	96	

42	2			1	5	4	35	31	3	1	2	11		95
43	2	2			6	17	7	47	1		2	8	1	93
51	1	3	3	1	30	11	1	6	19	2	3	7	9	96
71		4	1	2	3	20	3	3	10	3	28	10	7	94
80	1	51	2	2	3	10	1	6	5	20	273	8	3	385
91	3	1			3	11	10	16	11			32	7	94
92	8	2	1		2	6	3	10	4	4	19	11	23	93
Tot %	79	317	53	37	91	148	63	149	82	41	402	113	68	1643
														50.8

Table 9. Accuracy analysis with merged urban (20) and agriculture (80) classes

### 8.6 Random Quality Assurance Checks

The Project Manager and Principal investigator examined a 10% simple random sample of the interpreted sites. These random checks occurred throughout the project as the previously selected points were completed by the interpreters. In the event that the Project Manager did not agree with, or had a question about, the interpretation, the issue was discussed with the interpreter. The site was discussed until either interpreter justified their interpretation to the project manager or the interpreter learned what the correct interpretation should have been. These corrections were then disseminated to the group on an informal basis to ensure consistent results. Extensive training prior to beginning the project, and the fact that three of the four interpreters for this project also worked on the similar MRLC Region 4 accuracy assessment, resulted in very few corrections being required.

## 9. Discussion

At first glance, the overall accuracy (36%) of the Region 5 NLCD seems quite low. However, we believe that this number, derived from a simple site for site comparison, is based on a strict interpretation of overall accuracy, and so may underestimate the usability accuracy of the classification. This estimate does not take into account complicating factors such as errors in locating sites on the photos, seasonal and annual differences in photo and image acquisition dates, LU/LC change between photo and image acquisition dates, difficulty in separating the row crops vs. small grains classes, difficulty in identifying the shrubland and natural grassland classes, ambiguity of class definitions, heterogeneity of land cover, and the inclusion of black and white photos. We will examine each of these issues in detail. The issues discussed below (9.1 to 9.8) are not necessarily unique to the classification under evaluation in this study. Other investigators have reported similar issues (Congalton 1991 and 1999; Gopal and Woodcock 1994; Lunetta, et al 1998; Zhu, et al 2000).

### 9.1 Location Errors

The interpreters reported that it was, at times, difficult for them to exactly locate the reference site on the photo. This was most often the case 1) when the land cover had changed between the image and photo acquisition dates, and 2) when there were few

clearly identifiable features for reference. The interpreters feel confident that the vast majority of the sites were correctly located, but there are many that were troublesome. The inclusion of alternate photo interpretation classes and the use of a 3x3 pixel window in the accuracy analysis may help to mitigate the effect of this problem. We have included the primary/alternate and 3x3 window analyses in the Results section. See Plates 1, 2, 3, 4, and 6 for examples.

### *9.2 Seasonal and Annual Differences in Acquisition Dates*

This issue has a large effect in Region 5 because of the possibility that there is snow cover in the photos, the images, or both. Portions of the Region are snow and ice covered in the provided winter-acquired photos. This made it difficult, if not impossible, to determine the underlying ground cover. In these situations, a “best guess” was made and the interpretation confidence measure was, therefore, lowered.

### *9.3 LU/LC Change Between Acquisition Dates*

Always an issue when using reference data acquired at a different time than the classification, this factor likely resulted in errors when the photos were older than the images and development or logging had occurred between the acquisition dates. In clearly discernible situations such as these, the image was given more weight in choosing the correct NLCD class for the site. See Plate 3 for an example of this change.

### *9.4 Row Crops vs. Small Grains*

This problem is one of the largest single error sources in the accuracy assessment. We found during training that it is virtually impossible to separate row crops and small grains with winter NAPP photography. The only situation we found in which it is feasible to do so is when a crop such as winter wheat clearly shows up as live vegetation on the photos. However, fields in which row crops and small grains were planted during the normal summer growing season are indistinguishable with the provided winter photography. Table 2 shows that 62 of the 100 randomly selected small grains sites were identified from the reference data as row crops. Only 10 small grains sites were actually identified as small grains. We suggest combining the small grains and row crops classes into a broader crops class to reduce this significant source of confusion.

The interpreters also encountered problems when interpreting cropland and pasture/hay. This is attributed to the fact these classes look very similar on the photos and typically coexist geographically. In addition, cropland could have been changed to pasture/hay or vice versa during the interval of the two acquisition dates.

### *9.5 The Shrubland and Natural Grassland Classes*

The shrubland class caused problems similar to the small grains class. The interpreters found that it was very difficult to differentiate shrubland from the natural grassland and forest classes. In addition, the definition of the shrubland class describes a land cover

type that is very rare in Region 5. The shrubland class seems more appropriate for inclusion in the western United States where there are large expanses of mountain shrub species. See Plate 3 for an example.

The natural grassland class also caused many problems. Differentiation of natural grassland from pastures and the other grassland class proved problematic. The natural grassland class, like the shrubland class, occurs very infrequently in the study area. It is our opinion that most of the areas assigned to the natural grassland class in the Region 5 NLCD are mis-classified. Of the 94 randomly selected natural grassland sites, only approximately 3% were actually identified on the photos as natural grassland. The most significant sources of confusion were with deciduous forest, shrubland, and rowcrops.

### *9.6 Ambiguity of Class Definitions*

Table 2 indicates that there was significant confusion between low intensity residential, high intensity residential, and commercial/transportation. It is our opinion that the definitions of these categories contribute the most to this confusion. Categories that cannot be directly determined from remotely sensed data not only compromise the accuracy of classification, but also the accuracy of reference data collection. In this case, all three categories belong to urban and are distinguished from each other by the amount of vegetation. Generally speaking, it is difficult to identify the vegetation amount within a pixel, i.e. at sub-pixel level. Therefore, the definitions themselves contain ambiguity. Similar situations occurred when trying to differentiate shrublands, forests, and natural grasslands.

### *9.7 Heterogeneity of Land Cover*

The heterogeneity of cover types in much of the study area caused significant confusion in assigning exact class labels. Since the spatial resolution of the Landsat TM data is 30 meters by 30 meters, in many cases one pixel could consist of more than one land cover class. For example, a site on the image is often composed of some trees, some grassland and one or more structures, so the reflectance values for a given pixel is actually a combination of reflectance from several cover types within that pixel. This factor contributed to confusion between evergreen forest and mixed forest, deciduous forest and mixed forest, barren ground and other grassland, low intensity residential and mixed forest, and transitional and several classes. The inclusion of alternate photo interpretation classes and the use of a 3x3 pixel window in the accuracy analysis (as presented in the results section) may help to mitigate the effect of this problem.

### *9.8 Black and White Photos*

The vast majority of the photos provided for use as reference data in this project were color infrared. However, few, due to unavailability of CIR, were black and white. The interpreters reported problems selecting land-cover classes from the black and white photos. Particularly, differentiation between forest types and the shrubland class was difficult. A comparison of the accuracy of CIR vs. black and white sites may be helpful.

## 9.9 Summary of Discussion

It is our opinion that, from a usability standpoint, the true accuracy of the Region 5 dataset may be greater than 36%. Factors such as those discussed above significantly impacted the accuracy assessment of the NLCD Region 5 dataset. Where possible, we have recommended alternate analyses, which may help to provide more representative accuracy estimates. In addition, we have presented the results of the accuracy analysis with the inclusion of primary and alternate photo interpretation classes and with the use of a 3x3 window in the classified data, which have produced the overall accuracy estimates shown in Table 5. This method, while not a standard procedure for accuracy assessment, serves to highlight the effects of using strict point for point assessment methods. We leave to the users the decision as to which accuracy estimate is most consistent with their intended use of the NLCD data.

## 10. Summary and Conclusions

The 1643 photo interpretation sample sites for the Region 5 NLCD resulted in an overall accuracy estimate of 36%. Evaluation of only high confidence sites increased this estimate to 37%. When alternate classes were included, the accuracy estimate was 48%. When the primary or alternate matched any pixel in a 3 by 3 pixel area, the overall accuracy was 71%. The accuracy analysis was repeated for combined urban and agriculture classes and yielded an accuracy estimate of 51%. An Analysis of Variance (ANOVA) test indicated that inter-interpreter variation was not a significant source of error in the accuracy assessment.

## 11. Recommendations for Future Studies

- ?? Field sampling – We recommend that a small, statistically valid, ground sample be incorporated into the accuracy assessment. This sample would serve as further validation of both the NLCD data and the photo interpretation quality. Budget limitations in this project, based on contractual agreement, did not allow us to complete a field sample.
- ?? Revision of Region 5 classes – For reasons discussed in detail in this report, we recommend that the shrubland, small grains, and natural grassland be removed from consideration for inclusion in future Region 5 classifications.
- ?? Stereo coverage– Given the extreme difficulty of determining vegetation height with the provided non-stereo photographs, we recommend that future accuracy assessment of the NLCD data be conducted only with stereo photos.
- ?? CIR photos – We further recommend that sites for which there are no CIR photos be removed from the sample. Identification of many land cover classes is unreasonably difficult with bland and white photos.

?? Acquisition dates – Reducing the gap between the acquisition dates of the NLCD data and the NAPP or other photographs for accuracy assessment will reduce the uncertainty in the accuracy estimates.

## References

- R. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of the Environment*, vol. 37, pp. 35-46, 1991.
- R. Congalton and K. Green, "A practical look at the sources of confusion in error matrix generation," *Photogrammetric Engineering and Remote Sensing*, vol. 59. No. 5, pp. 641-644, 1993.
- X. L. Dai and S. Khorram, "The effects of image misregistration on the accuracy of remotely sensed change detection." *IEEE Transactions on Geoscience and Remote Sensing*, vol.36, no.5, pp. 1566-1577, 1998.
- S. Gopal and C. Woodcock, "Theory and methods for accuracy assessment of thematic maps using fuzzy sets," *Photogrammetric Engineering and Remote Sensing*, vol. 60, No. 2, pp. 181-188, 1994.
- S. Khorram, G. S. Biging, N. R. Chrisman, D. R. Colby, R. G. Congalton, J. E. Dobson, R. L. Ferguson, M. F. Goodchild, J. R. Jensen, and T. H. Mace, "Accuracy assessment of remote sensing-derived change detection," *Monograph, American Society of Photogrammetry and Remote Sensing (ASPRS)*: Bethesda: Maryland, 64p. 1999.
- R. S. Lunetta, J. G. Lyon, B. Guidon, and C. D. Elvidge, "North American Landscape Characterization Dataset Development and Data Fusion Issues," *Photogrammetric Engineering & Remote Sensing*, vol. 64, no. 8, pp. 821-829, August 1998.
- Z. Zhu, L. Yang, S. V. Stehman, and R. L. Czaplewski, "Accuracy Assessment for the U.S. Geological Survey Regional Land Cover Mapping Program: New York and New Jersey Region," *Photogrammetric Engineering & Remote Sensing*. 2000.

## Appendix

### A1: Classification Scheme and Class Definitions

The MRLC program utilizes a consistent classification scheme for all EPA Regions at approximately an Anderson Level II thematic detail. While there are 21 classes in the MRLC system, only 15 were mapped in EPA Region 5. The following classification scheme was applied to EPA Region 5 data set:

- 11 Water
- 21 Low Intensity Residential
- 22 High Intensity Residential
- 23 Commercial / transportation / industrial
- 31 Bare Rock / Sand
- 32 Quarries / Strip Mines / Gravel
- 33 Transitional
- 41 Deciduous Forest
- 42 Evergreen Forest
- 43 Mixed Forest
- 51 Shrubland
- 71 Natural Grassland
- 81 Pasture/Hay
- 82 Row Crops
- 83 Small Grains
- 85 Other Grassland (maintained)
- 91 Woody Wetlands
- 92 Emergent Herbaceous Wetlands

*The class definitions are as follows:*

*Water - All areas of open water or permanent ice/snow cover.*

- 11. Open Water - All areas of open water; typically  $\geq 25$  % cover of water (per pixel).

*Developed - Areas with by a high percentage ( $\geq 30$  %) of constructed materials (e.g. asphalt, concrete, buildings, etc).*

- 21. Low Intensity Residential - Includes areas with a mixture of constructed materials and vegetation. Constructed materials are 30-80% of cover. Vegetation may are 20-

70% of cover. These areas most commonly include single-family housing units. Population densities will be lower than in high intensity residential areas.

22. High Intensity Residential - Includes highly developed areas where people reside in high numbers. Examples include apartment complexes and row houses. Vegetation is  $\leq 20\%$  cover. Constructed materials are 80-100% of cover.

23. Commercial/Industrial/Transportation - Includes infrastructure (e.g. roads, railroads, etc.) and all highly developed areas not classified as High Intensity Residential.

*Barren - Areas characterized by bare rock, gravel, sand, silt, clay, or other earthen material, with little or no "green" vegetation present regardless of its inherent ability to support life. Vegetation, if present, is more widely spaced and scrubby than that in the "green" vegetated categories; lichen cover may be extensive.*

31. Bare Rock/Sand/Clay - Perennially barren areas of bedrock, desert pavement, scarps, talus, slides, volcanic material, glacial debris, beaches, and other accumulations of earthen material.

32. Quarries/Strip Mines/Gravel Pits - Areas of extractive mining activities with significant surface expression.

33. Transitional - Areas of sparse vegetative cover ( $\leq 25\%$  cover) that are dynamically changing from one land cover to another, often because of land use activities. Examples include forest clear cuts, transition between forest and agriculture, the temporary clearing of vegetation, and changes due to natural causes (e.g. fire, flood, etc.).

*Forested Upland - Areas with tree cover (woody vegetation generally  $\geq 6$  m); tree canopy is 25-100% of cover.*

41. Deciduous Forest - Areas dominated by trees where  $\geq 75\%$  of the tree species shed foliage due to seasonal change.

42. Evergreen Forest - Areas dominated by trees where  $\geq 75\%$  of the tree species maintain their leaves all year. Canopy is never without green foliage.

43. Mixed Forest - Areas dominated by trees where neither deciduous nor evergreen species represent  $\geq 75\%$  of the cover.

*Shrubland - Areas characterized by natural or semi-natural woody vegetation with aerial stems generally  $\leq 6$  m, with individuals or clumps not touching to interlocking. Both evergreen and deciduous species of true shrubs, young trees, and trees or shrubs that are small or stunted because of environmental conditions are included.*

51. Shrubland - Areas dominated by shrubs; shrub canopy accounts for 25-100% of the cover. Shrub cover is generally  $\geq 25\%$  when tree cover is  $\leq 25\%$ . Shrub cover may be  $\leq 25\%$  in cases when cover of other life forms  $\leq 25\%$  and shrubs cover exceeds the cover of the other life forms.

*Herbaceous Upland - Upland areas characterized by natural or semi-natural herbaceous vegetation; herbaceous vegetation accounts for 75-100% of the cover.*

71. Grasslands/Herbaceous - Areas dominated by upland grasses and forbs. In rare cases, herbaceous cover  $\leq 25\%$ , but exceeds the combined cover of the woody species present. These areas are not subject to intensive management, but they are often utilized for grazing.

*Planted/Cultivated - Areas characterized by herbaceous vegetation that has been planted or is intensively managed for the production of food, feed, or fiber; or is maintained in developed settings for specific purposes. Herbaceous vegetation accounts for 75-100% of the cover.*

81. Pasture/Hay - Areas of grasses, legumes, or grass-legume mixtures planted for livestock grazing or the production of seed or hay crops.

82. Row Crops - Areas used for the production of crops, such as corn, soybeans, vegetables, tobacco, and cotton.

83. Small Grains - Areas used for the production of graminoid crops such as wheat, barley, oats, and rice.

85. Urban/Recreational Grasses - Vegetation (primarily grasses) planted in developed settings for recreation, erosion control, or aesthetic purposes. Examples include parks, lawns, golf courses, airport grasses, and industrial site grasses.

*Wetlands - Areas where the soil or substrate is periodically saturated with or covered with water as defined by Cowardin et al.*

91. Woody Wetlands - Areas where forest or shrubland vegetation accounts for 25-100% of the cover and the soil or substrate is periodically saturated with or covered with water.

92. Emergent Herbaceous Wetlands - Areas where perennial herbaceous vegetation accounts for 75-100% of the cover and the soil or substrate is periodically saturated with or covered with water.

## **A2: Location and Interpretation Confidence Linguistic Scales**

Location:

- ?? Can't locate the point
- ?? Location is doubtful
- ?? Location is probably correct
- ?? Location is absolutely correct

Interpretation:

- ?? Can't choose a class
- ?? Class chosen is doubtful
- ?? Class is probably correct
- ?? Class is absolutely correct