

# Issues Involved in the Accuracy Assessment of Large Scale Land Use / Land Cover Mapping from Remotely Sensed Data

Siamak Khorram, Joseph Knight, X. Long Dai, Hui Yuan, Halil Cakir, Zhiyan Mao  
Center for Earth Observation, Box 7106  
North Carolina State University  
Raleigh, NC 27695-7106  
(919) 515-3430, Khorram@ncsu.edu

## INTRODUCTION

The Multi-Resolution Land Characteristic (MRLC) Consortium, composed of several U.S. government agencies, sponsored the creation of the National Land Cover Data (NLCD) [3,4,5,6]. This dataset provides a consistent land cover classification system for the lower forty-eight states. It is based on thirty meter spatial resolution Landsat Thematic Mapper (TM) satellite data. The NLCD, as a national land cover classification, is unprecedented in its coverage and resolution.

The objective of this project was to assess the accuracy of a portion of the NLCD. An implementation scheme for the project was designed by the Center for Earth Observation group to assure the quality of the reference data collection when applied to such a large scale project. A number of issues were encountered in dealing with a project of this magnitude. These issues are of importance to researchers involved in large scale land cover mapping and change detection [1,2]. Issues addressed include: Examining the impact of including alternate classes in the accuracy assessment, compensating for positional errors in the dataset, and examining reference data collection issues.

## STUDY AREA AND SAMPLING SCHEME

The study area for this project was Federal Region 4, which is composed of the states of North Carolina, South Carolina, Georgia, Florida, Kentucky, Tennessee, Mississippi, and Alabama. This Region contains a wide variety of land cover types, including grassy plains, mountains, saltwater marshes, coastal plains, piedmont, agricultural areas, and urban centers.

The sampling scheme was a two-stage design. The first stage, the Primary Sampling Units (PSU), were the size of a National Aerial Photography Program (NAPP) photo of the scale of 1:40,000. Each PSU was randomly selected from a cluster of 128 photos. These clusters of photos were formed using a geographic unit frame of 30 minutes by 30 minutes. The second stage was a stratified random sample, within the PSUs, of 100 points per land-cover class. The selected points were referred to as Secondary Sampling Units (SSU). The number of points per photo ranged from one to approximately 70. The total number of sample points in the study was 1500 (100 for each of the fifteen MRLC land-

cover classes present in Region 4). This approach was chosen by the EROS Data Center over a standard random sampling over the entire study area to reduce the cost of purchasing the NAPP photography [7].

## METHODS

### Training

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

The focus was on likely situations that the interpreters would encounter during the project. Each participant was presented with over 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, each interpreter was assigned approximately 400-500 points to examine. Their progress was monitored daily for accuracy and proper methodology. The interpreters kept a log of their calls and the points for which they were uncertain about the land cover classes. On a weekly basis, their questions were addressed by the project supervisors. The problematic sites were discussed until each team member felt comfortable with the class definitions and their consistency in interpretation.

### Photo Interpretation Procedure:

NAPP 1:40,000 scale aerial photos were used as the reference data source for this project. The majority of the photos were color infra-red. However, in some parts of the study area color infra-red photos were not available. In these areas, black and white photos were substituted. The EROS Data Center provided the locations of the 1500 sample points to be used in the accuracy assessment. The sample points were each assigned randomly to one of the three photo interpreters. The interpreters were assigned 500 points plus 75 points drawn randomly from each of the other two interpreters' points. This gave each interpreter a total of 650 points and provided 225 overlapping points (15%)

which were interpreted by all three interpreters. The overlapping points were used for consistency analysis and quality assurance.

The interpreters examined the sample point's characteristics on the aerial photo and determined the proper MRLC class label for each sample point. This was referred to as the **primary** land cover class for the sample. In situations where, due to heterogeneity of the sample or uncertainty in the correct class, a single land cover class could not be chosen, the interpreters also selected an **alternate** land cover class for the sample. This was very common in residential areas where points often fell in mixed areas of houses, lawns, trees, and other cover types. The primary classes were considered to be the correct classes for the samples and were used in the standard point for point accuracy assessment of the MRLC classification. For the purposes of the additional neighborhood accuracy analysis presented in this paper, the alternate classes were considered to be as correct as the primary classes. In addition to the primary and alternate classes, the interpreters collected fuzzy set data for each reference point. This data will be used in future accuracy assessment research.

#### Accuracy Assessment Procedure

The accuracy assessment was divided into two stages. First, a standard point for point accuracy assessment, and then additional analysis which took into account both the primary and alternate (if present) interpreted classes and a three by three pixel neighborhood around the provided sample points. The additional accuracy analysis was broken down into six groups:

1. Primary interpreted class matches the classified value for the center pixel only (standard point-for-point method);
2. Primary interpreted class matches the most common classified value in a 3x3 area around the sample point;
3. Primary interpreted class matches any classified value in a 3x3 area around the sample point;
4. Primary or alternate interpreted class matches the classified value for the center pixel only;
5. Primary or alternate interpreted class matches the most common classified value in the 3x3 area;
6. Primary or alternate interpreted class matches any classified value in the 3x3 area.

#### QUALITY ASSURANCE PROCEDURES

Strong Quality Assurance and Quality Control (QA/QC) procedures were enforced in this project. Upon completion of the training portion of the project, a test was given to determine how similarly the interpreters would interpret the same sites. The 225 overlap sites were chosen (75 per interpreter) for which all three interpreters would provide a

land cover class. These overlap points were then analyzed for consistency. Initially, there was confusion as to the proper interpretation of some of the classes. This necessitated a re-training phase in which the interpreters collaboratively identified a series of training sites. Upon completion of this additional training, a consistency test was re-administered. The average overall agreement between the photo interpreters was 84%.

In addition to the training procedures and consistency checks, weekly meetings of the project team were held to review project progress and to discuss problems encountered. Apart from these formal meetings, discussions among the research group on an ad hoc basis provided an opportunity to discuss problems that occurred and, in most cases, provide a solution on the spot. Information gathered from these meetings greatly increased the interpretation quality and consistency.

#### RESULTS

The results of the accuracy analysis are presented in the following table. Fig. 1 lists the overall accuracy estimates derived from incorporating the different combinations of primary and alternate classes and classified 3x3 neighborhood values.

Prim. matches classified:	640/1474 = 43 %
Prim. is most common in 3x3:	674/1474 = 46 %
Prim. matches any in 3x3:	859/1474 = 58 %
Prim. or alt. matches classified:	824/1474 = 56 %
Prim. or alt. is most common in 3x3:	985/1474 = 67 %
Prim. or alt. matches any in 3x3:	1170/1474 = 79 %

Figure 1: Summary of accuracy analysis

#### DISCUSSION

From Fig. 1, we see that the standard method of accuracy assessment produced an overall accuracy estimate for this data set of 43%. The main factors combining to lower the overall estimate of classification accuracy included: heterogeneity of the study area, similarity between certain land classes, and positional errors in sample point location. To decrease the effects of these factors on the overall accuracy, extended accuracy analysis was performed which examined, not only the primary interpreted class, but also an alternate class (if specified), as well as the classified values in a three-by-three pixel neighborhood around the center point. The addition of an alternate interpreted class helped to compensate for heterogeneity, and class similarity. Inclusion of the three-by-three classified neighborhood values was used to help compensate for positional errors. The greatest increases in the estimate of overall accuracy occur when the alternate classes are taken into consideration. The estimate increases from 43%, with the point-for-point method, to 56%, with the inclusion of the alternate class

values. The impact of the neighborhood values are not as great. After including the "most common" value in the classified three-by-three area, the overall estimate shows small increases over using the alternate classes alone. Only when the interpreted values are matched with "any pixel" in the classified three-by-three area are large increases in the overall accuracy estimate evident.

There were some issues with the reference data collection process. Some of the classes were particularly problematic. There was much confusion between low intensity residential, high intensity residential, and commercial/transportation. It is likely that the definitions of these categories [5] contributed 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 a general urban category 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. The interpreters assigned many high intensity residential areas in the classified images to the low intensity residential and commercial/transportation class. Similarly, the interpreters had difficulty differentiating cropland and pasture/hay. This was attributed to the fact that both of these classes have very similar spatial patterns and commonly co-exist within the same general areas. Confusion also existed between classes of evergreen forest vs. mixed forest; and deciduous forest vs. mixed forest. The similarity of these classes on the photos made differentiation a very complex task for the interpreters. In addition, there are, no doubt, errors associated with incorrect location of the sample points on the photos. It is difficult to quantify such errors, but the interpreters reported that, for some points, they were not certain that they had accurately located the sample point on the photo. This was frequently due to spatial resolution differences between the image data and the photos and land cover change between the acquisition dates of the imagery and photos. The accuracy assessment protocol would clearly have benefited from the inclusion of reference data derived from actual ground sampling, but due to the financial constraints of the project this was not possible.

### CONCLUSIONS

An accuracy assessment project of this magnitude introduces many complexities. These issues, heterogeneity of cover types, similarity between certain land cover classes, and positional errors in sample point location, land cover change between the acquisition dates of the raw data and the reference data make implementing the accuracy assessment more difficult. We have proposed solutions to some of these problems by including primary and alternate classes in the

accuracy assessment, compensating for positional errors in the dataset through the use of a three by three pixel window, fuzzy set analyses, and ensuring proper training of photo interpreters. These solutions provide a much more representative measure of the accuracy of a thematic classification than current methods.

### ACKNOWLEDGMENTS

This research was funded by the United States Environmental Protection Agency as a part of the MRLC-NLCD Region IV accuracy assessment. Sample point coordinate locations and NAPP photography were provided by the EROS Data Center in Sioux Falls, S.D. The authors would like to thank Jim Wickham, Limin Yang, and Steve Stehman for their valuable guidance in the completion of this project.

### REFERENCES

- [1] Congalton, R.G., K. Green. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC Press, Inc., Boca Raton, FL., pp. 29-30.
- [2] Khorram, S., 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, 1999. Monograph, Accuracy Assessment of Remote Sensing-Derived Change Detection. American Society of Photogrammetry and Remote Sensing (ASPRS), 58p.
- [3] Loveland, T.R., and D.M. Shaw. (1996). Multiresolution land characterization: building collaborative partnerships, Gap Analysis: A Landscape Approach to Biodiversity Planning (J.M. Scott, T. Tear, and F. Davis, editors), Proceedings of the ASPRS/GAP Symposium, Charlotte, North Carolina, National Biological Service, Moscow, Idaho, pp. 83-89.
- [4] Van Driel, N., T. Loveland, and D. Lauer. (1999). The MultiResolution Land Characteristics Consortium, a government success story. The 50th International Astronautical Congress, October, 1999, Amsterdam, The Netherlands.
- [5] Vogelmann, J.E., T.L., Sohl, P.V. Campbell, and D.M. Shaw. (1998a). Regional land cover characterization using Landsat Thematic Mapper data and ancillary data sources, Environmental Monitoring and Assessment 51: 415-428.
- [6] Vogelmann, J.E., T. Sohl, and S.M. Howard. (1998b). Regional characterization of land cover using multiple sources of data, Photogrammetric Engineering and Remote Sensing 64: 45-57.
- [7] Zhu, Z, L. Yang, S.V. Stehman, and R.L. Czaplewski. In Press. Accuracy assessment for the U.S. Geological Survey regional land cover mapping program: New York and New Jersey Region. Photogrammetric Engineering and Remote Sensing.