GROUND TRUTH SAMPLING AND LANDSAT ACCURACY ASSESSMENT

by

Jon W. Robinson

Computer Sciences Corporation

Fred J. Gunther

Computer Sciences Corporation

William J. Campbell

Goddard Space Flight Center

INTRODUCTION

The work reported in this paper was supported by a contract (The Power Plant Siting Study) from the Nuclear Regulatory Commission to the National Aeronautics and Space Administration, Goddard Space Flight Center. The work was carried out by government and contractor personnel at Goddard Space Flight Center in cooperation with the Nuclear Regulatory Commission and Pennsylvania Power and Light Company.

The purpose of the study was to compare the cost and accuracy of various remote sensing data types and processing procedures for updating Geographic Information Systems (GIS). This paper reports a portion of the work carried out under that contract. A complete report of the work carried out under the contract will be submitted to the Nuclear Regulatory Commission at the end of the contract period and will be available to the public from the Nuclear Regulatory Commission.

The key factor in any accuracy assessment of remote sensing data is the method used for determining the ground truth, independent of the remote sensing data itself. This paper will describe the sampling and accuracy procedures developed for the Power Plant Siting Study.

The purpose of the sampling procedure was to provide data for developing supervised classifications for the two study sites and for assessing the accuracy of that and the other procedures used. The purpose of the accuracy assessment was to allow the comparison of the cost and accuracy of various classification procedures as applied to various data types.

There were two study sites, one centered on the city of Lancaster, Pennsylvania and the other centered on the Susquehanna Steam (nuclear) Generating Plant near Berwick, Pennsylvania. The methods described here were used at both sites, but only the results from the Berwick site will be presented here. The final report to the Nuclear Regulatory Commission will contain the results from both sites.

Each site contained 400 square miles, 20 miles on a side. Both sites were within the Pennsylvania Power and Light Company's service area and were covered by that company's Environmental Land Use Data System (ELUDS) data base (a geographic information system). The data base includes a variety of data types, including land cover, geology, slope, infrastructure, and historic sites.

METHODS

In this section, the materials used and the methods employed for both the sampling procedure and the accuracy assessment procedure will be presented. The sampling and accuracy procedures involved the use and merging of several data types. These included Landsat Multispectral Scanner (MSS) data, Thematic Mapper Simulator (TMS) and low altitude aerial photography which was digitized for further manipulation by computer. All of these data were registered to United States Geological Survey 7.5- minute maps so they would be congruent with each other. The results of a ground survey were then combined with the previous data to provide estimates of the accuracy of the two types of classifiers used on the MSS and TMS data. Since the study area was too large to be completely surveyed, a sampling procedure was developed.

Sampling Methods

The goal of the sampling procedure was to generate as many ground truth pixels per given amount of effort as possible, yet maintain a statistically valid procedure. The sampling procedure chosen was cluster sampling (Cochran, 1977). This allowed areas to be chosen at random and a large number of pixels to be identified in each chosen area.

The areas were chosen by taking United States Geological Survey (USGS) 7.5-minute quadrangle maps of the study site and picking points at random from selected quadrangles. Because of time constraints, a contiguous group of maps within the study area was selected. That group of maps included the Susquahanna Steam Generating Plant.

The vertical and horizontal borders of each map were marked at one inch intervals. Pairs of two-digit random numbers were then taken from a random number table (Rohlf and Sokal, 1969) to select pairs of horizontal and vertical tick marks from the edges of the maps. If a two-digit number was beyond the range of the tick marks, another two-digit number would be chosen until one within the range was selected. Each pair of tick marks identified a centroid of a one-inch-by-one inch square on the map. Due to the dense road network, each square selected on the map was crossed by or closely approached by at least one road. Each site so selected was then visited with a survey crew provided by Pennsylvania Power and Light Company. Table 1 lists the name of each quadrangle selected and the approximate latitude and longitude of each site visited within that quadrangle.

TABLE 1

Latitude and Longitude of Ground Truth Sample Areas

Quadrangle	#	Latitude	Longitude
Shickshinny	6	41 9.3 N	76 9.0 W
"	4	41 10.0 N	76 10.9 W
"	1	41 10.7 N	76 12.2 W
"	2	41 13.3 N	76 10.0 W
"	3	41 13.3 N	76 11.0 W
"	5	41 13.0 N	76 14.1 W
"	7	41 12.6 N	76 14.8 W
Stillwater	14	41 9.4 N	76 18.2 W
"	16	41 2.9 N	76 19.4 W
"	15	41 8.2 N	76 18.0 W

On arriving at a site, landmarks that would show up on low altitude aerial photography were identified. Then the location of field boundaries and the boundaries between landcover types were measured relative to the landmarks. Detailed notes on the crop types and landcover types surveyed were taken along with 35mm. photographs on Kodachrome and Infrared Aero Ektachrome. The Infrared Ektachrome pictures were taken so that the observations obtained on the ground could be compared with low altitude color infrared photography and infrared photography taken by the Thematic Mapper Simulator flight.

The original plan was to have the low altitude aerial photography performed on or close to the date of the field work which was during the last week of August 1981 and to have this coincide with the flight of the Thematic Mapper Simulator (TMS). The low altitude photography was being provided by a subcontractor for Edgerton Gearson & Greer Corporation (EG&G) for the Nuclear Regulatory Commission on a separate contract. Because of contracting delays, the flight was not made until

the 25th of September 1981. The Thematic Mapper Simulator (TMS) flight was being flown by National Aeronautics & Space Administration National Space Technology Laboratories (NSTL) in Mississippi. Although the field work was undertaken with the understanding that NSTL would make the TMS flight during the ground-truth field work, it was in actuality not flown until the 12th of October.

The low altitude aerial photography was digitized by the University of California Santa Barbara on a (subcontract from EG&G) into three digital images for each frame. Each digital image was filtered by the appropriate red, green or blue filter so that the color information content of the original color infrared photograph would be retained. Each frame of digitized photography was entered into the Interactive Digital Image Manipulation System (IDIMS) on a HP3000 computer. Each frame that covered one of the ground-truth study sites was then registered to the 7.5-minute quadrangle map in which it occurred. The registration was to within 15 meters, which is the accuracy limit of the 7.5-minute quadrangle maps.

The registered images were then displayed on a color raster display using the IDIMS programs; and the boundaries of the landcover types were drawn in and the polygons thus generated labeled using the data collected during the ground-truth collection field trip. Because all of the remote-sensing images were registered to the same 7.5-minute maps, the identity of any pixel falling within one of the ground-truth polygons could be determined. Thus, the accuracy of the classifications generated by the various processing methods could be determined for each type of data used by counting the number of pixels of known ground cover that were correctly labeled by a classification.

Accuracy Methods

For the accuracy assessment, the identity of pixels falling within the ground truth polygons and urban-area polygons (which were photointerpreted) were compared with the classification labels produced by a particular classification method. The two primary methods of classification used were maximum likelihood and cluster analysis with the ISOCLS routine in the IDIMS system.

The maximum likelihood classifier required that statistics, sample mean vectors and sample variance-covariance matrices be generated for each landcover type. Half of the ground truth sites were used to generate these statistics and the other half were used to estimate the accuracy of the method.

Theoretically, one could use the pixels used to generate the maximum likelihood decision rule to estimate its accuracy. This estimate of the accuracy would only be unbiased if the sample used to generate the classification was unbiased. Therefore it is best to use an independent sample of pixels, if that is possible, to test the accuracy of a maximum likelihood classifier. The practice of using the classifier to classify the pixels which generated it and then using the accuracy of that classification to estimate accuracy of the classifier is called back classification. A close agreement between accuracy estimates from back classification and from a classification of an independent sample of pixels of known identity indicates that the two samples are less likely to have been drawn in a biased manner from the population of pixels and that more faith can be placed in the estimates so derived.

Thus to check for bias in selecting which sites would be used for generating the classification and which sites would be used for accuracy determination, the back classification accuracy was determined for the training site pixels as well as for an independent sample of pixels.

Because the ground-truth sites had been broken into two groups for testing the accuracy of the maximum likelihood classification, the accuracy of the ISOCLS classifications were estimated by comparing the accuracy for each group of ground-truth sites separately. This provided two independent estimates of the accuracy for each ISOCLS classification.

A table like table 2, was generated from a CONTABLE (an IDIMS program) run on each classification. The values in these tables were then used to calculate the following estimates: the probability that a pixel is correctly classified; the probability that a pixel belonging to class i is classified into class i, and the probability that a pixel classified as class i is in fact a member of class i.

Table 2 shows the unweighted procedure for calculating accuracy figures. This means that the number of pixels in each category are in proportion to their frequency in the ground truth polygons. Because urban areas were photointerpreted, the relative frequency of those pixels in the accuracy assessment procedure were greater than their relative frequency in the image being classified. If the accuracy figures were adjusted to the relative frequency of each category of pixel in the image being classified, then they would be weighted (or a weighted accuracy assessment).

It has been been pointed out (Chrisman, 1980) that simple accuracy figures, by themselves, may be missleading. A better measure of how well a classifier is performing would be the percentage improvement over a random classifier based on the

relative frequencies of the classes. The kappa statistic (Everitt, 1968) provides such a measure. Using the frequency of pixels in each class in the ground-truth polygons to calculate the expected frequencies for a random classifier, the kappa statistic was calculated for each data type and classification procedure.

TABLE 2

Accuracy Calculations

	Classifier Label					Number Belonging To Each Class
	1	2	3	...	M	
True Label						
1	m_{11}	m_{12}	m_{13}	...	m_{1M}	$m_{1\cdot}$
2	m_{21}	m_{22}	m_{23}	...	m_{2M}	$m_{2\cdot}$
3	m_{31}	m_{32}	m_{33}	...	m_{3M}	$m_{3\cdot}$
.	.	.	.	...	.	.
.	.	.	.	...	.	.
.	.	.	.	...	.	.
M	m_{M1}	m_{M2}	m_{M3}	...	m_{MM}	$m_{M\cdot}$
Number Classified As	$m_{\cdot 1}$	$m_{\cdot 2}$	$m_{\cdot 3}$	...	$m_{\cdot M}$	$m_{\cdot\cdot}$

$$\text{Total Pixels Checked} = TP = \sum_{i=1}^{M}\sum_{j=1}^{M} m_{ij} = m_{\cdot\cdot}$$

$$\text{Total Pixels Correct} = TC = \sum_{i=1}^{M} m_{ii}$$

Probability that a pixel in the sample is correctly classified. $= P_{cc} = TC/TP$

Probability that a pixel classified as class i is a member of class i. $= P_{ci} = m_{ii} / \sum_{j=1}^{M} m_{ji} = m_{ii}/m_{\cdot i}$

Probability that a pixel that is a member of class i is classified as class i. $= P_{ic} = m_{ii} / \sum_{j=1}^{M} m_{ij} = m_{ii}/m_{i\cdot}$

RESULTS

The results of the sampling can only be presented in terms of an analysis of the accuracy figures. Table 3a gives the results of the unweighted accuracy calculations based on the maximum likelihood classification of the independent sample of pixels of known identity for both the MSS and TMS images. Table 3b gives the results of the unweighted accuracy calculations based on the maximum likelihood classification of the pixels used to generate the classification functions (back classification).

Tables 4a and 4b present similar results for the unsupervised method (cluster analysis) of classification. Because the classes are not predefined as in the supervised method (maximum likelihood) the analyst must assign names to the classes generated by the clustering algorithm. This led to the merging of several ELUDS landcover classes into more general categories. The merged ELUDS classes are identified by the numbers associated with each landcover name in tables 4a and 4b.

It should be noted that those categories that have small samples for the training sets, i.e. the N columns in table 3b, have low accuracies. Beyond this, the results for the accuracy assessment based on the back classification are not very different from those based on the independent sample. The small pixel counts for the landcover class "barren land" in the unsupervised classification do not provide an accurate estimate of the probabilities for that class.

There is little difference between the probabilities of correct classification for the different classification methods. The primary difference is in the number of classes that can be differentiated. The kappa statistic also reflects this situation.

The overall quality of the classifications based on the TMS data are better for all of the classification procedures and assessment data sets. Since the quality of the TMS data was very bad it contained a large amount of noise, the quality of classifications based on real Thematic Mapper (TM) data should be better.

TABLE 3A

BERWICK
ACCURACY ASSESSMENT MAXIMUM LIKELIHOOD CLASSIFIER

(INDEPENDENT SAMPLE)

ELUDS CODE	FREQ.*	MSS			TMS		
		N	P_{ci}	P_{ic}	N	P_{ci}	P_{ic}
1 URBAN	.0537	1572	.882	.803	6091	.930	.953
2 BARREN LAND	.0375	68	.186	.353	116	.019	.017
3 AGRICULTURAL	.3225	568	.418	.563	2110	.696	.667
5 TREE PLANTAT.	.0018	7	.0	.0	27	.039	.111
7 CONIF. FOREST	.0084	18	.0	.0	77	.055	.182
9 DECID. FOREST	.3852	1490	.780	.590	2694	.675	.591
11 MIXED FOREST	.1603	138	.071	.094	512	.073	.060
13 SCRUB LAND	.0048		NONE			NONE	
14 MEADOW	.0009	15	.0	.0	50	.0	.0
15 FORESTED WETL	.0099	26	.023	.115		NONE	
16 UNFOREST WETL	.0000		NONE			NONE	
99 WATER	.0148	212	.864	.962	843	.840	.890
		P_{cc} = .6578			P_{cc} = .7669		
		KAPPA = .5108			KAPPA = .6590		

*FREQ. - The frequency of each ELUDS data type in the entire 400 square mile Berwick study site.
N - The counts of pixels of each ELUDS landcover type in the ground truth polygons used for the independent accuracy assessment.

TABLE 3B

BERWICK
ACCURACY ASSESSMENT MAXIMUM LIKELIHOOD CLASSIFIER

(BACK-CLASSIFICATION)

ELUDS CODE	FREQ.*	MSS			TMS		
		N	P_{ci}	P_{ic}	N	P_{ci}	P_{ic}
1 URBAN	.0537	1567	.951	.678	6305	.977	.908
2 BARREN LAND	.0375	104	.121	.106	61	.750	.443
3 AGRICULTURAL	.3225	285	.281	.537	1070	.688	.827
5 TREE PLANTAT.	.0018	6	.231	1.000	23	.188	.826
7 CONIF. FOREST	.0084	83	.619	.157	347	.441	.478
9 DECID. FOREST	.3852	1125	.732	.762	2439	.832	.801
11 MIXED FOREST	.1603	24	.066	.333	97	.179	.433
13 SCRUB LAND	.0048		NONE			NONE	
14 MEADOW	.0009	12	.240	.500	40	.440	.825
15 FORESTED WETL	.0099	10	.063	.600		NONE	
16 UNFOREST WETL	.0000		NONE			NONE	
99 WATER	.0148	96	.929	.958	349	.795	.943
		P_{cc} = .6685			P_{cc} = .8553		
		KAPPA = .4906			KAPPA = .7726		

*FREQ. - The frequency of each ELUDS data type in the entire 400 square mile Berwick study site.
N - The counts of pixels of each ELUDS landcover type in the ground truth polygons. This is also the sample size for each class's training set.

TABLE 4A

BERWICK
ACCURACY ASSESSMENT UNSUPERVISED CLASSIFICATION

(INDEPENDENT SAMPLE)[1]

LAND COVER	FREQ.*	MSS			TMS		
		N	P_{ci}	P_{ic}	N	P_{ci}	P_{ic}
1 URBAN	.0537	1572	.953	.502	6076	.988	.706
2 BARREN LAND	.0375	68	.289	.191	114	.073	.491
3 + 14 AGRICUL.	.3234	583	.369	.877	2160	.583	.787
5 + 7 + 9 + 11 + 15 FOREST	.5656	1679	.855	.846	3313	.836	.912
99 WATER	.0148	212	.964	.892	838	.925	.952
		P_{cc} = .7103			P_{cc} = .7892		
		KAPPA = .5639			KAPPA = .6802		

[1]The subtitle Independent sample is used for identification purposes only. The unsupervised classification procedure does not use training sites.
*FREQ. - The frequency of each ELUDS data type in the entire 400 square mile Berwick study site.
N - The counts of pixels of each landcover type in the ground truth polygons used for the independent accuracy assessment.

TABLE 4B

BERWICK
ACCURACY ASSESSMENT UNSUPERVISED CLASSIFICATION

(BACK-CLASSIFICATION)[1]

LAND COVER	FREQ.*	MSS			TMS		
		N	P_{ci}	P_{ic}	N	P_{ci}	P_{ic}
1 URBAN	.0537	1567	.967	.318	6292	.998	.512
2 BARREN LAND	.0375	104	.018	.019	58	.028	.448
3 + 14 AGRICUL.	.3234	297	.190	.761	1107	.281	.703
5 + 7 + 9 + 11 + 15 FOREST	.5656	1248	.750	.851	2873	.738	.869
99 WATER	.0148	96	1.000	.865	336	.898	.917
		P_{cc} = .5652			P_{cc} = .6407		
		KAPPA = .3036			KAPPA = .3671		

[1]The subtitle back-classification is used for identification purposes only. The unsupervised classification procedure does not use training sites.
*FREQ. - The frequency of each ELUDS data type in the entire 400 square mile Berwick study site.
N - The counts of pixels of each landcover type in the ground truth polygons used for the training of the maximum likelihood.

DISCUSSION

The results of the accuracy assessment of the supervised classification indicate that there was no strong bias in the sampling procedure. The low accuracies for certain categories may be due to either the similarities in their spectral reflectivities or the small samples used to characterize their spectral reflectivities. At an intuitive level, it is easy to understand how the various forest landcover types would be spectrally confusing. The causes of confusion between the other classes are not so obvious.

One remedy for the small sample sizes of certain categories would be to use a stratified sampling procedure (Cochran, 1977), where the strata would be the landcover categories. This would allow for adequate sample sizes for all but the rarest categories. There is one requirement for this procedure that makes it more difficult to carry out. That is, a landcover map of the area must already be available. It does not have to be perfect, but it must be sufficiently accurate so that the majority of the field checks are made in the correct categories.

A further problem with cluster sampling is that neighboring pixels are used for the training set pixels and for the accuracy assessment pixels. Studies by a variety of authors have shown that the spectral characters of the pixels are spatially autocorrelated. It is also clear that other characteristics may be spatially autocorrelated. Since one of the basic assumptions behind the estimation procedures used is that the observations are statistically independent, the confidence bounds of the quantities presented here can not be reliably determined. Further, because of theoretical considerations it may be that the classifications themselves would be quite different if the autocorrelation in the spectral values of neighboring pixels were removed.

The overall accuracies of the two classification procedures do not differ much between themselves when compared with the variation within a procedure. The prime differences are that in the supervised classification, the classes are defined in advance and that in the unsupervised classification, the classes are assigned names on an adhoc basis. The success of the adhoc assignment of class identities by the skilled analyst are vindicated by the small differences between the supervised classification and the unsupervised classification accuracies.

A major consideration in choosing which classification procedure will be used in a study will be cost. The cost to properly execute a supervised classification is considerably greater than the cost to properly execute an unsupervised classification. In many situations, where the classes of

landcover that are to be distinguished are coarse, the unsupervised methods are the most efficient. In those situations where a statistically rigorous procedure is required and where many categories must be distinguished, the extra cost of the supervised procedure may be justified.

The accuracies achieved by both classification methods were consistently better with the TMS data than with the MSS data. This was inspite of the fact that the TMS data was very noisy and required both geometric and spectral correction for the bow tie effect. This indicates that the increased spectral and spatial resolution provide for a consistently more accurate classification. The results with real Thematic Mapper data should be much better than the results presented here.

A more detailed analysis of the data developed in this study should provide a better understanding of the results presented here. Such analysis could look at the trade off between noise in individual sensor channels and greater spectral and spatial resolution. Such analysis could also examine the effects of autocorrelation on all aspects of a classification procedure: the classification, and the accuracy estimates.

CONCLUSION

The sampling design and the associated accuracy assessment presented above indicate that Thematic Mapper data should provide consistently better classification results than the old Multispectral Scanner data of Landsat 1, 2 and 3. In addition it appears that the choice of a classification procedure will depend on the purposes to which the classification will be put and the resources available to execute it. In a supervised classification the sampling procedure by which ground truth is obtained will be dictated by the requirements of the particular study. If the accurate classification of rare classes is not of great importance, than cluster sampling may prove quite efficient. However, other sampling procedures should be considered when rare classes are important and the necessary ancillary information is available.

LITERATURE CITED

Anonymous. 1981. IDIMS Functional Guide Volume I. Technical Manual ESL-TM705. ESL Incorporated. Sunnyvale, California. viii + 716p

Chrisman, Nicholas R. 1980. Assessing Landsat accuracy and correcting for missclassification errors. Unpublished Manuscript. 18p.

Cochran, William G. 1977. Sampling Techniques, third edition. John Wiley & Sons, xvi + 428.

Everitt, B. S. 1968. Moments of the statistics kappa and weighted kappa. The British Journal of Mathematical and Statistical Psychology. 21:97-103.

Rohlf, F. James and Robert R. Sokal. 1969. Statistical Tables. W. H. Freeman and Company. xi + 253.