Engineering Soils Mapping in Indiana by Computer from Remote Sensing Data[1]

T. R. West

Department of Geosciences

Purdue University, Lafayette, Indiana 47907

Abstract

Multispectral imagery of the surface terrain may be obtained from airborne scanning devices that measure reflectance in the visible through infrared portion of the electromagnetic spectrum. At the Laboratory for Applications of Remote Sensing (LARS), Purdue University, multispectral imagery is analyzed using a high-speed digital computer. The final result is a detailed map or classification of the surface material.

Pattern recognition (classification) is accomplished by separating materials according to their spectral response statistics in a known area and then applying these criteria to unknown areas. As an example of the LARS mapping technique, a detailed analysis of a soils area in north-central Indiana is presented.

In this study the correct classification of sandy floodplain soils versus till plain soils was obtained for 98+% of the ground area elements (remote sensing units) for the training fields. Similarly a 76+% correct discrimination between the two materials was made for the test fields in the flightline. This degree of accuracy indicates that the analysis of these data using the LARS techniques was successful in differentiating the materials sufficiently well to be of considerable use in reconnaissance surveys.

Introduction

Aerial photography has been used for several decades in reconnaissance mapping of soil, rock, vegetation and man-made features in a variety of scientific specialties, including geology, geography, agronomy, forestry, hydrology, civil engineering and a number of others. With the development of improved instrumentation systems for measuring electromagnetic radiation of the terrain and with the advent of high-speed computers, it is now possible to obtain more precise data which can be analyzed in great detail by digital computers.

Remote sensing involves the identification and classification of physical objects through analysis of data obtained from sensing devices which do not come in direct contact with those objects. In most cases the instruments measure radiation intensities from selected portions of the electromagnetic spectrum. For remote sensing studies that incorporate analysis by computer, the data-gathering phase is followed by some data reduction and reformatting which expedites subsequent data processing. The final phase, analysis, involves classification of the target materials through the use of pattern recognition techniques.

This paper presents a discussion of remote sensing of soil and rock materials with specific emphasis on the techniques developed at the Laboratory for Applications of Remote Sensing, Purdue University.

[1]Research was supported in part by NASA Grant NRG 15-005-112.

Methods: LARS Automatic Classification Technique

Perhaps the first question to answer is why use an automatic classification system at all? Why not rely on the current method of mapping soil and rock for geological and other purposes. An obvious answer is that computers are faster and more accurate. Computer use does not remove the scientist from the operation; it makes him more important. As before, accurate field work is necessary; however, the same amount of field effort can now be used to map a larger area. Reliable data are needed to train the computer for proper classification but this detail is not required for the entire mapping area.

Currently, multispectral imagery of the terrain is obtained for LARS by the Institute of Science and Technology, University of Michigan, using a multispectral optical-mechanical scanner mounted on their DC-3 aircraft. As the plane proceeds along the flight path, the energy radiated by a specific ground resolution element passes through the scanner optics, is divided according to its spectral wavelength, and is directed to the appropriate detector. The output of all such detectors is simultaneously recorded by a multiband recorder. The transverse motion provided by the rotating mirror and the forward motion of the aircraft cause a continuous signal to be recorded for each spectral band of the scanner output. The information is stored in analog form by the data recording system.

The multispectral bands recorded on the aircraft storage tapes vary according to the needs of the researcher. In this study, 12 channels of data were obtained from the visible and reflective infrared portion of the spectrum, between 0.4 and 2.6 micrometers[1] (μm).

Following the multispectral remote sensing flights, the aircraft scanner analog data tapes are forwarded to LARS for conversion to digital form. Each analog scan is sampled, normally at a sampling rate which yields 220 data points for an 80° field of view across the flightline. For later reference, each digitized data point (remote sensing unit or RSU) is assigned a unique address in a two-dimensional coordinate system based on scan line numbers ("line numbers") and samples within the line ("column numbers").

Typically the first step in the analysis is to obtain gray scale printouts of the data for several channels. Gray scale printouts are digital displays of the spectral response of the terrain but limited to one band or channel of the scanner data per display. They resemble low-resolution photographs. Alpha-numeric symbols (letters of the alphabet, typing symbols and single digit numbers) are assigned to the radiance levels so that high-intensity or light areas receive symbols which cover a low percentage of the paper (Example ---), and dark areas receive symbols with a high percentage of cover (Example MMM). Because the data are not equally distributed between the highest and lowest radiance values, the symbols are not simply assigned so that each represents an equal range of radiance. Instead, to achieve a maximum contrast display, the data

---

[1] One micrometer (μm) $= 10^{-6}$ meters.

are first histogrammed and the symbols are assigned so that each will occur with approximately the same frequency.

Areas of known materials are located on the gray scale printouts by the researcher and their addresses are recorded. This information, referred to as "ground truth," is used to train the computer to recognize similar kinds of material. In agricultural studies, this may consist of fields containing corn, wheat, oats, soybeans, etc.; in forestry studies—conifers and deciduous trees or individual tree species; in geology and highway engineering—specific bedrock and soil types. While some of the ground truth areas are used to train the computer to recognize the classes of interest, the remaining areas are reserved for testing the accuracy of the computer classification after it is completed. Addresses of training areas are provided to the computer on punched cards; these addresses are boundary corners indicated by column and line number, and the fields are limited to rectangular shapes (commonly referred to as "fields").

After January 1, 1971, an alternative to using gray scale printouts for field selection became available at LARS. This tool, known as the digital display unit, provides a television-like image of the scanner data with each digital value represented by a different brightness level. This yields an image on the television screen having greater detail than is possible with gray scale symbols on computer paper. Images for each channel can be displayed, and fields of interest can be outlined on the screen using a light pen with their addresses automatically punched on cards.

Another method for obtaining training field sites is sometimes used in conjunction with the above procedure of manually selecting these sites. Accomplished by using the LARS non-specified classifier routine (NSCLAS), this relatively new technique divides the scanner data into groups or clusters based on similarity of spectral response within clusters. Typically 4 to 6 channels of imagery are analyzed simultaneously and the program is requested to obtain 10 clusters. A map of the results is printed for each area analyzed in this way and the researcher can observe the patterns of spectrally differentiable material occurring in the data. The training fields may then be selected from these maps. The resulting clusters may not in fact be spectrally distinct, but the researcher must decide this to his own satisfaction based on separability information for the clusters which is printed by the program. He then has the option of repeating the analysis using a different number of clusters to increase their separability. An important point is that this method of selecting training fields takes into detailed account the multispectral response data as well as the ground truth information.

In the clustering approach, the actual differences in spectral response are displayed, but a major problem exists in determining what each of the clusters represents. Some may be vegetation types, some water bodies, man-made features or tonal aspects of bare soil and rock. Aerial photography is helpful in affixing names to the patterns observed. Despite this difficulty, clustering is a powerful tool in obtaining workable training fields.

The next step in the analysis is to obtain histograms of each class of material identified in the training fields. An example of classes for a geologic study area might be alluvium, limestone soil, shale soil, trees, mixed crops, and water. The histograms for each class show the distribution of reflectance intensity for each spectral channel.

Unimodal or single-peaked distributions in the histograms for a class suggest that the proposed class is spectrally an individual group. Bimodal or trimodal distributions must be subdivided manually into unimodal classes by the researcher; histograms for individual fields can be obtained to help in locating the multimodal contribution within a class.

The next operation involves the application of a divergence (statistical separability) analysis (known as $DIVERG) to determine the best channels to use for classification. Only the best four to six channels are used for classification to save computer time; in general, the accuracy is not meaningfully increased when more channels are added. In addition to indicating the preferred channels, the divergence analysis indicates the separability of the designated classes. If separability between significant materials is poor, some of the preceding steps are repeated in an attempt to improve this separability and hence classification accuracy.

Next, training field statistics are used to classify the designated portion of the flightline. The computer classifies each data point (RSU) based on the maximum likelihood criterion. On request, the computer calculates how accurately it classified the areas used for training by comparing the classification of each point in the training fields with the initial ground truth designation. A high level of agreement means there is minimal confusion within the training field statistics for the various materials and that the classes are being separated properly.

The final steps are to print out a computer classification for the whole area and to determine how well the test fields were classified. The test fields are those areas of known material from ground truth studies that were not previously used for training purposes. If test fields and training fields show a high degree of accuracy, and the test fields are representative of the entire area, the classification is a good one. If the accuracy is low, some reworking of the classes should be done. The researcher may also have to conclude that the classes involved are not spectrally separable.

Results: Tippecanoe County, Indiana, Study Area

To illustrate the automatic classification technique for engineering soils mapping, a flightline in Tippecanoe County, Indiana, was selected for study. The test site known as Tippecanoe County Flightline 23 is a north-south line running approximately through the center of the county, including a central strip through the Lafayette-West Lafayette metropolitan area along the Wabash River (Fig. 1).

Scanner data were obtained for Flightline 23 during various seasons of the year in 1969 and 1970 for the on-going LARS program of study of the midwest corn belt area. Seasonal changes of soil patterns and crop

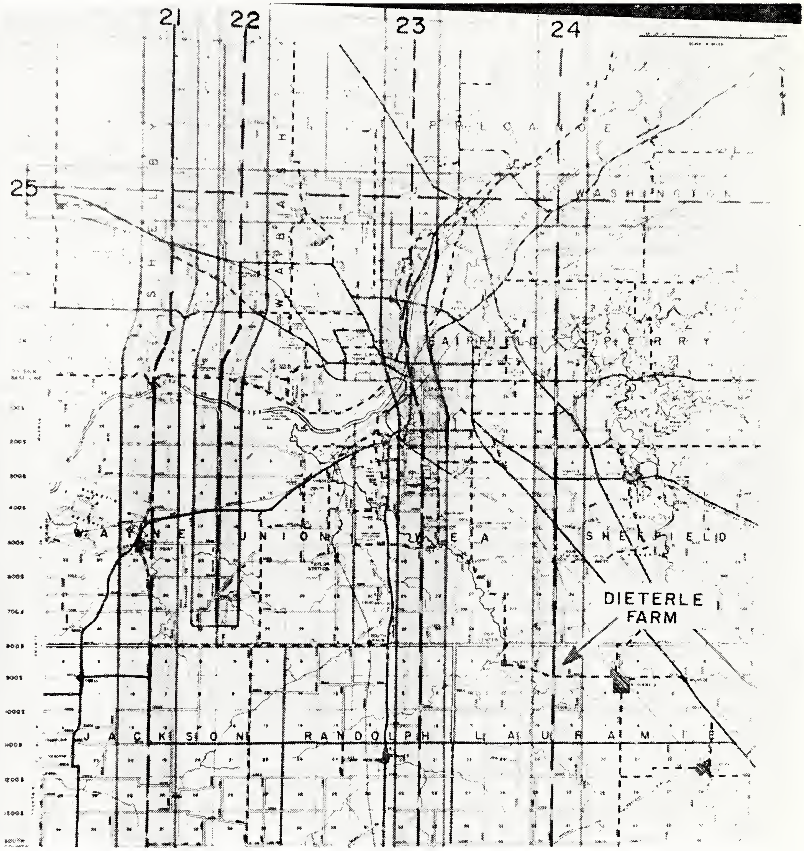

FIGURE 1. *Tippecanoe County, Indiana, flightlines.*

cover were observed by these periodic flights over the established flightline. In the current study, the scanner flight dated May 6, 1970, was selected for analysis. In general, spring flights are best for soil classification studies as the maximum percentage of bare soil is exposed at this time of year.

Analysis was limited to the northern third of the flightline which contains examples of the soil materials of greatest interest in the study. The northern 5 miles of the flightline cover a portion of the Tipton Till Plain. South of this segment is a 2-mile section which runs across the Wabash River floodplain. These two sections comprise contrasting parent materials and form the basis for the soils study of the 7-mile area. Flown at an elevation of 3,000 feet, the flightline has a width of about 5,000 feet.

The initial step in the analysis technique was to compare a gray scale printout of the flightline to black and white photography taken at the time of flight. The Wabash River was outlined, as were the floodplain and till plain areas and several significant man-made features: the By-Pass 52 bridge and Interstate 65, which was under construction at the time of flight.

To determine the material types present in the study area, 13 east-west strips of data were taken at various points of interest along the flightline. An NSCLAS was run on these combined transverse strips which collectively comprised nearly the maximum number of data points that the routine can handle.

In the 13 clustered strips, vegetation, water, man-made features and bare soil were identified with the aid of the aerial photography. Rectangular fields were selected in the bare soil areas and a second NSCLAS was run on these combined areas using ten clusters for the analysis. This yielded a numerous collection of rectangular fields with soil patterns designated in each.

By selecting those fields (or portions of fields) which had predominantly one pattern, a collection of bare soil fields, each of essentially one soil type, was assembled. Two distinct groups were maintained, however: those fields from the floodplain area and those from the till plain.

Next, statistics were obtained for each field and the histograms examined to insure that each was unimodal. Three divergence analyses ($DIVERG) were made on the fields to determine how to combine them into separable classes. Two runs were required initially because the programs accept a maximum of 30 fields in the divergence analysis and the 50 fields were divided into two groups to accomplish this. Fields deemed similar by the divergence analysis were combined, whereas, others were left by themselves and a few confusing fields discarded. After this was achieved for the two divergence runs, the remaining groups (numbering less than 30) were collected in the final divergence analysis and the combining technique repeated. The net effect is that 18 classes remained after this compilation, 8 from the floodplain area and 10 from the till plain.

Several classifications were made of the 7-mile segment using the training statistics from these classes. Adjustments were made in successive classifications to improve the training field performance. In the final classification this performance was sandy floodplain 99.1%, and till plain 98.5%. These values are somewhat misleading as a considerable number of floodplain symbols were scattered within the till area although the till symbols are rather minimal in the floodplain area.

To obtain a more meaningful indication of the classification's accuracy, a large number of bare soil fields were selected as test fields. In the sandy floodplain, 28 test fields containing a total of approximately 1,900 RSU's and 59 test fields in the till plain (approximately 8,700 RSU's) were included. Test results showed 92.7% accuracy for the sandy floodplain and 76.8% accuracy for the till plain with an overall accuracy of 76.8%.

Because of limitations on the size of figures and their number in this paper, the computer printout, which is quite large and loses much detail in photo reduction, is not included here.

An intriguing situation was noted during the classification revisions. The till plain areas most often confused with floodplain soils were in recent road cuts where reflectance of the till was the greatest. This was observed near the By-Pass 52 bridge cut on the west side of the Wabash River and for borrow and fill areas along Interstate 65.

An interesting substudy was made on the flightline as a final effort. Interstate 65 runs diagonally NW-SE across the flightline. Experience at LARS has indicated that straight line, man-made features not oriented essentially parallel or perpendicular to the flightline are difficult to map in detail. Diagonal features of this sort run both away from the center of the flightline yielding changes in viewing angle from the aircraft, and up or down the flightline simultaneously. Also, as remote sensing units measured on the ground are square, the units along the pavement boundary have an averaged value of reflectance for both pavement and soil. If training samples are taken from this boundary portion, erroneous pavement symbols will appear in soil areas in other locations of the classification.

To master this problem, we clustered (NSCLAS) rectangular areas across the highway from the southeastern edge to the northwestern edge of the flightline. Ten cluster groups were designated. The pavement, median strip and shoulders were outlined in the clustered map results. As might be expected, the same cluster group was not associated with the pavement in the center as at the edges of the flightline, but the contrast of pavement and soil was apparent at each location. To insure proper classification, pavement classes from the middle, quarter point and edge of the flightline would be included for accurate designation of the highway in an overall classification.