ENGINEERING SOIL MAPPING FROM MULTISPECTRAL IMAGERY USING AUTOMATIC CLASSIFICATION TECHNIQUES

Terry R. West, Purdue University

Multispectral imagery collected over southeastern Pennsylvania by Willow Run Laboratories, University of Michigan, was analyzed at the Laboratory for Applications of Remote Sensing using its current capabilities for automatic classification. The project involved the evaluation of imagery in 13 discrete bands of the spectrum as a source of data on engineering soils. Detailed computer classifications of two 4-mile segments of the Pennsylvania flight line, each containing predominantly 1 parent soil material, were obtained. Classification accuracy as measured by training field and test field performance was more than 90 percent. These 2 segments were used as a basis to obtain a computer-implemented map for 12 miles of the flight line.

•IN JUNE 1970, the research project to analyze by digital computer multisensor data collected over various test sites was initiated by the Laboratory for Applications of Remote Sensing (LARS), Purdue University. This research involves using the LARS digital computer and automatic multispectral data analysis techniques for analyzing remote-sensing data and for performing engineering soil mapping. The work done for the Pennsylvania test site is discussed in this paper.

Soil pattern recognition using the LARS approach predates this study. Agricultural soil studies have involved detailed mapping of relatively small areas (about 100 acres) located in the glacial areas of the midwestern corn belt where topographic relief is low. Previous work applying the LARS techniques to engineering soils has been reported by Rib and Miles (3) and by Tanguay and Miles (4) and involved glaciated terrain and bedrock areas. The Pennsylvania flight line, over a region of bedrock-derived residual soils having moderate relief, presented a new challenge to the continually evolving LARS techniques.

DESCRIPTION OF TEST SITE

The test site location in southeastern Pennsylvania was selected by the Pennsylvania Department of Transportation and the Federal Highway Administration because of its accessibility and because of the favorable geologic features of the area. It is approximately 47 miles long. The flight line lies at right angles to the prominent structural trend of the Appalachian Mountain system. This orientation maximizes the number of different rock units intercepted by the flight line. Also, the extensive cultivation in the area maximized the percentage of bare soil present in the spring when the flight was made. The southern end is located in the Piedmont physiographic province, but the flight line soon enters the Valley and Ridge province and ends there about 40 miles to the north.

The geology of the Pennsylvania test area can be described as consisting of simply to complexly folded, steeply dipping sedimentary rock strata cut by several fault systems and igneous intrusions of Triassic age. The sedimentary rocks are primarily sandstone, shale, limestone, and dolomite sequences of Paleozoic age, but they also contain some Triassic sandstone, shale, and conglomerate. Southeastern Pennsylvania has not been glaciated, and the soils consist of residual material formed by surface weathering of the parent bedrock. Local relief in the area ranges from 100 to 500 ft and the elevation from 350 to 1,000 ft above sea level. Annual precipitation is approximately 41 in./year, and dense forests develop in areas not under active cultivation.

Because of the complex structure of the sedimentary rocks, portions of the flight line were selected that contained predominantly 1 parent bedrock material. These specific "pure soil" areas were selected with the intention that analysis would first be accomplished for these and subsequently extended to the more diverse areas between them. Table 1 gives information on the pure soil areas.

REMOTE-SENSING DATA COLLECTED

The multispectral imagery used in this study was collected by the Willow Run Laboratories, University of Michigan, under a separate contract with the Federal Highway Administration. The data were collected with a multispectral optical-mechanical scanner mounted in a DC-3 aircraft and were recorded on magnetic tape. Duplicate copies of the tape were furnished to LARS for this study.

The specific multispectral bands recorded on aircraft storage tapes vary according to the needs of the researcher. In this study, a total of 15 channels of data were obtained, 13 in the visible and reflective infrared portion of the spectrum (between 0.4 and 2.6 μm) and 2 in the thermal infrared (4.5 to 5.5 μm and 8.0 to 14.0 μm).

The flight line was flown May 15, 1969, at 3,000 ft above the average terrain elevation. The data were recorded on 2 tapes: tape 1 for the southern half and tape 2 for the northern half of the flight line.

Each segment was flown during both the day and the night. For the daytime flights, visible, reflective infrared, and thermal infrared data were obtained. At night, measurements were limited to the thermal infrared.

In addition to the aircraft data tapes, other information made available to LARS included black-and-white 9- by 9-in. aerial photographs plus an aerial photomosaic for the flight line; topographic maps for the area and agricultural soil maps for Berks and Lancaster Counties; cathode-ray tube imagery of 3 wavelength bands (one each in the visible, reflective infrared, and thermal infrared); 70-mm color photography; and 9- by 9-in. color and color infrared photography.

LARS TECHNIQUES FOR MULTISPECTRAL DATA ANALYSIS

This section is intended to familiarize the reader with the current LARS analysis techniques for multispectral remote-sensing data. Because of space limitations, this discussion has been abbreviated. Detailed descriptions of the techniques are given by Tanguay and Miles (4), Hoffer and Goodrick (1), and West (5).

Remote sensing involves the identification and classification of surfaces through analysis of data from sensing devices not in direct contact with those surfaces. At LARS these data are analyzed on a digital computer using multivariate-pattern recognition techniques. Currently, imagery collected in the visible through thermal infrared portions of the spectrum are included in pattern recognition studies.

The data collected during the flight are in an analog format. By means of the LARS data-handling system, these analog data tapes are converted to digital form. Each analog scan is sampled, normally at a sampling rate that yields 220 data points for an 80-deg field of view across the flight line. For later reference, each digitized data point is assigned a unique address in a 2-dimensional coordinate system based on scan line numbers (line numbers) and samples within the line (column numbers).

A final preprocessing function, data overlaying, must precede analysis of imagery from multiple-aperture scanners. In 1969, when these data were obtained, the University of Michigan scanner system collected data from 3 apertures (visible, reflective infrared, and thermal infrared). In such cases, data overlaying is performed to align the data so that pattern recognition can be based on all channels. Checkpoints, such as field corners and highway intersections, are located on gray scale printouts (described below) for channels obtained through different apertures. After an array of such checkpoints is established throughout the flight line, the computer aligns these points and forces the data between them to line up as well as possible. Typically, the discrepancy remaining after data overlaying is at most 2 or 3 resolution elements. A resolution element or remote-sensing unit (RSU) is the area on the ground represented by a single reflectance symbol or letter on a gray scale printout. At the 3,000-ft altitude of the Pennsylvania flight, this represents about a 20-ft square on the ground.

The computer programs used at LARS to analyze multispectral scanner data are shown in Figure 1. Indicated are the sequential and alternative steps involved in the analysis procedures.

Typically the initial step in the analysis is to obtain gray scale printouts of the data for several channels ($PIC, Fig. 1). Gray scale printouts are digital displays of the spectral response of the terrain but limited to 1 band or channel of the scanner data per display. They resemble low resolution photographs.

The researcher locates areas of known materials on the gray scale printouts and records their addresses. This information, referred to as ground truth, is used to train the computer to recognize similar material. In agricultural studies, this training information may consist of fields containing corn, wheat, oats, or soybeans; in forestry studies, it may consist of conifers and deciduous trees or individual tree species; and in geology and highway engineering, it may consist of specific bedrock and soil types. Although 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.

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

Another method for obtaining training field sites is sometimes used in conjunction with the procedure given above. This involves the use of the clustering program NSCLAS (nonsupervised classifier, Fig. 1). 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 computer display 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. 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 that are printed by the program. He then has the option of repeating the analysis and using a different number of clusters to increase their separability. An important point is that this method of selecting training fields takes into 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. They may be vegetation types, 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 clustering technique, which was under development during the analysis phase of the work reported here, was not applied to these data. It was subsequently used for detailed soil mapping of the Pennsylvania data (2) and in a brief study in Indiana (6).

The next step in the analysis is to obtain histograms for each class of material identified in the training fields ($STAT). 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 4 to 6 channels are used for classification in order to save computer time; in general, the accuracy of classification 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 may be repeated in an attempt to improve this separability and hence classification accuracy.

Next, the training field statistics are used to classify the designated portion of the flight line ($CLASSIFY). The computer classifies each data point (RSU) based on a 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 of the various materials and that the classes are being separated properly.

The final steps are to print out a computer classification display map for the whole area ($DISPLAY) 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 OF STUDY FOR THE BLUE BALL AND REAMSTOWN-DENVER AREAS

Initial work during the contract period was performed on data from the Blue Ball area. Prior to this, a brief study had been made on a 12-mile section immediately to the south of Blue Ball near Welsh Mountain, a prominent quartzite, tree-covered ridge. This preliminary study, set up to determine the potential of the LARS system for engineering soils, was most rewarding in that it suggested an improved approach for analyzing the total flight line.

The Welsh Mountain study was complicated by the diverse geologic and topographic nature of the area. It is a complex region of the Piedmont province displaying as much as 500 ft of relief due to differential erosion of rocks with contrasting resistance to weathering. Soils derived from limestone, schist, quartzite, and a mixture of these, talus or colluvium, mantle the area.

Because of these complexities, very accurate ground truth information was necessary for training the computer to distinguish between the various materials. Such detailed information, as later determined, was unavailable. In areas of high relief composed of resistant rocks, it is common for pieces of durable rock to be moved downslope by gravity and to accumulate as a mixture of materials at a lower level. This colluvium displays some of the spectral characteristics of all the parent materials involved. As the extent of colluvium exposed at the surface was not known, it was omitted from the training samples.

In the computer-generated classification of the Welsh Mountain area, vegetation and water were accurately delineated; however, classification of the soil types did not agree well with the known geological conditions of the area for the reasons previously stated. Because of these inaccuracies in this complex area, it was decided to concentrate on pure soil areas where the complications are minimized. This supplied the impetus to use the pure soils area approach for the Pennsylvania flight line.

Blue Ball Area

The first pure soil area studied, located near Blue Ball, is 3.6 miles long and approximately 5,000 ft wide. It lies near the southern end of the flight line in Lancaster County and is designated as pure soil areas 1 and 2 (Table 1). The following discussion, a step-by-step description of the manner in which the final Blue Ball classification was accomplished, is presented as an example of the details involved in obtaining an accurate classification.

This flight line segment consists primarily of residual limestone soil with some river alluvium located along a small creek. It was eventually determined that 3 conditions exist in this residual limestone soil: eroded areas where the subsoil is exposed, noneroded areas, and places where local alluvium derived from erosion has accumulated. These combined conditions (which can be observed on the color photos and to a lesser extent on the gray scale printouts of the multispectral scanner data) made analysis of the Blue Ball site difficult. Visible and reflective infrared channels were included in the analysis.

The first classification of the Blue Ball area was based on 10 fields outlined on air photos that were suggested for training by the FHWA photo interpreter. Training fields in the river alluvium along the creek, which had not been represented previously, were added, and the Lancaster County soils map provided ground truth information. Early results suggested that more spectral classes of limestone soil were present than were represented by the training fields. In the subsequent refinement, many additional limestone training fields were added from the entire area in an attempt to represent all possibilities. Test fields were selected at the same time.

For this next classification, results still indicated difficulty in differentiating river alluvium, local alluvium, and limestone soil. Field checks were made by staff from the Department of Agronomy, Pennsylvania State University, at LARS' request. This field study indicated that the agricultural soil map was not entirely accurate, and this new information was used to discard some training fields, regroup others, and add new ones, specifically categories for eroded limestone and local alluvial accumulation. A series of classifications was made, fields that proved troublesome were deleted, and several new factors were added such as training fields from a limestone quarry. Approximately 75 limestone training fields were used in the final classification. The training fields were evaluated after the classification was completed, and these results are given in Table 2. The overall performance was 97.7 percent correct, and the average performance by class was 97.8 percent correct.

After the classification was made, 58 test fields in the limestone soil were used to determine the accuracy of the classification. For these fields, which comprised 3,340 data points or RSU's, an accuracy of 90.6 percent was obtained.

In the process of improving the classification, several class divergence analyses were made. As a result, the 75 limestone fields were eventually combined into 14 separable classes. Several of the final classes were relatively similar but collectively were sufficiently different so that, if combined, they would yield a combined class too broad for useful classification. Hence, the 14 classes were maintained as separate entities.

The divergence analysis identified the 4 best channels for classification. As vegetation types are not an important aspect in this study, the vegetation classes were omitted in the divergence analysis. This meant that only separability between different soil groups would be a factor in determining which channels to use. Maximum significance was assigned to distinguishing between limestone, local alluvium, river alluvium, and eroded soil. This resulted in the selection of channels 2, 7, 12, and 13 as the 4 best channels for classification (0.44 to 0.46, 0.58 to 0.62, 1.00 to 1.40, and 2.00 to 2.60 μm respectively).

A third divergence analysis, identical to the preceding except for the deletion of channels 12 and 13, was then performed in order to determine whether channels 12 and 13 contributed enough toward class separability to warrant their inclusion. Because these channels are obtained from a scanner aperture different from channels 1 through 11, these upper channels are significantly involved in the overlay problem, with the misalignment ranging up to 3 resolution units despite overlay adjustments. Results indicated a marked reduction in separability of some key classes when channels 12 and 13 were omitted, signifying that they were needed for proper separation. Thus, channels 2, 7, 12, and 13 appeared best for the classification.

Reamstown-Denver Area

The Reamstown-Denver area, designated as pure soil area 4 (Table 1), was the second pure-soil area studied. The Reamstown-Denver pure soil study area is about

4½ miles long and 5,000 ft wide. This study area contains several soils that are derived from very different parent materials (bedrock). Much of the area is underlain by Ordovician age limestone, which yields a limestone soil similar to that found in the Blue Ball area. The limestone soil conditions in Reamstown-Denver are simpler, however, because the severely eroded areas are absent.

Present also in the Reamstown-Denver area are soils derived from dark-gray Ordovician age shales and soils derived from Triassic age red-colored shales, sandstones, and conglomerates. In addition, river alluvium and river terrace materials are present. The water and vegetation classes are similar to those in the Blue Ball area.

Training samples were selected from the different soil types in the Reamstown-Denver site, combined into classes, and used to classify the existing surface materials. Twenty-three classes consisting of 76 fields were used. The same channels used in the Blue Ball classification, 2, 7, 12, and 13 (0.44 to 0.46, 0.58 to 0.62, 1.00 to 1.40, and 2.00 to 2.60 μm respectively), were used. Several classifications were required because adjustments for the troublesome training fields were needed before an acceptable classification was obtained for the area. The training fields were evaluated for the Reamstown-Denver classification, and these results are given in detail in Table 3. The overall performance was 98.2 percent accurate, and the average performance by class was 97.8 percent.

After the completion of the Reamstown-Denver classification, the training classes for both the Blue Ball and the Reamstown-Denver areas were combined to yield 43 classes and approximately 150 training fields. The area from Blue Ball continuously through to Denver, a distance of 11.6 miles, was classified by the use of the resulting statistics. Included in this 11.6-mile segment is pure soil area 3, Independence School. This area of Triassic age shale, sandstone, and conglomerate was mapped in this manner. The results obtained for area 3 are realistic and quite good considering that no training samples were taken directly from that area.

For the 11.6-mile segment, the training field performance averaged 95.6 percent correct. This high level is somewhat misleading in regard to the actual accuracy, for no new test fields were evaluated. A problem that occurs is the confusion between shale-derived soils in Reamstown-Denver and the limestone soils in Blue Ball. This misclassification is related to the presence of transitional local alluvium and limestone soil. Despite these difficulties, the extended classification is a significant step toward the objective of using the LARS computer techniques to map engineering soils for large areas of a flight line.

SUMMARY AND FUTURE WORK

Results from the 2 pure soil areas, Blue Ball and Reamstown-Denver, indicate that engineering soil materials can be mapped at the Pennsylvania test site in considerable detail when adequate ground truth is available. Vegetation, water, roads, rooftops, and quarries can also be discerned. Vast amounts of detail are available from multispectral analyses, much of which is due directly to soil properties and conditions. For example, locations of wet soil and areas where soil cover over bedrock is thin may be outlined.

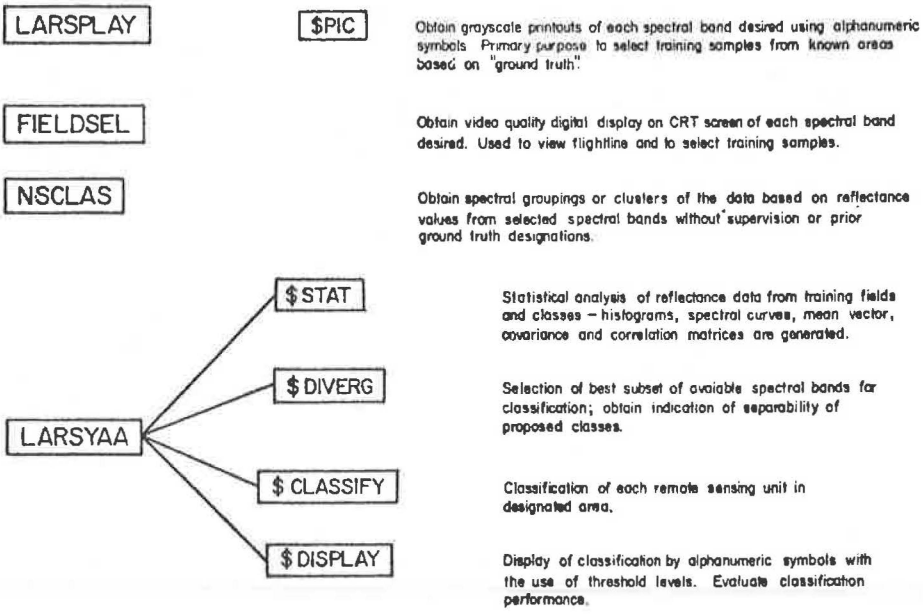
Classification of the 11.6-mile segment from Blue Ball to Reamstown-Denver marks a step toward classifying longer portions of a flight line by extending classification from smaller segments. This could possibly be extended to achieve a classification of the southern half of the flight line if additional training fields from the unrepresented soil types were obtained. The achievement of a combined flight-line map was an objective of the study, and the pure soils area concept was initiated in hopes of accomplishing it.

Two problems arise when mapping is done for a large area that contains numerous spectral classes of material. First, there is a physical limit on the number of classes that can be processed by the computer programs; the number is primarily a function of computer memory limitations. The current limit at LARS is 60 classes, and the 11.6-mile classification with 43 classes approached this limit. Second, accurate classification becomes more difficult when numerous material types are considered. Simply stated, the probability of having overlapping spectral characteristics of materials is increased when greater numbers of similar materials are included.

Table 1. Pure soil areas.

Pure Soil Area	Designation	Predominant Parent Materials	Geologic Age
1	Blue Ball	Limestone	Cambrian
2	Blue Ball	Limestone	Cambrian
3	Independence School	Sandstone and shale	Triassic
4	Reamstown to Denver	Limestone	Ordovician
5	Union House	Sandstone and shale	Triassic
6	Sheridan	Limestone	Ordovician
7	Mt. Aetna	Sandstone and shale	Ordovician
8	Bethel	Sandstone and shale	Ordovician

Figure 1. Computer programs for classification of multispectral scanner data.



Obtain grayscale printouts of each spectral band desired using alphanumeric symbols. Primary purpose to select training samples from known areas based on "ground truth".

Obtain video quality digital display on CRT screen of each spectral band desired. Used to view flightline and to select training samples.

Obtain spectral groupings or clusters of the data based on reflectance values from selected spectral bands without supervision or prior ground truth designations.

Statistical analysis of reflectance data from training fields and classes – histograms, spectral curves, mean vector, covariance and correlation matrices are generated.

Selection of best subset of avaiable spectral bands for classification; obtain indication of separability of proposed classes.

Classification of each remote sensing unit in designated area.

Display of classification by alphanumeric symbols with the use of threshold levels. Evaluate classification performance.

Table 2. Training field performance—Blue Ball.

Group	Number of Samples	Correct (percent)	Samples Classified: Veg	Lime	Alv	Loclv	Rival	Erod	Threshold[a]
Vegetation	519	100.0	519	0	0	0	0	0	0
Limestone soil	3,586	97.4	16	3,491	3	62	1	0	13
Alluvium	43	97.7	0	0	42	0	1	0	0
Local alluvium	86	96.5	0	3	0	83	0	0	0
River alluvium	149	98.7	0	1	1	0	147	0	0
Eroded limestone soil	32	96.9	1	0	0	0	0	31	0
Total	4,415		536	3,495	46	145	149	31	13

[a]No classification made (data very unlike any training class).

Table 3. Training field performance—Reamstown-Denver.

Group	Number of Samples	Correct (percent)	Samples Classified: Qua	Tpk	Roof	Water	Alv	Ter	Lime	Shale	Tri
Quarry	142	97.2	138	1	3	0	0	0	0	0	0
Turnpike	27	100.0	0	27	0	0	0	0	0	0	0
Roof	53	98.1	0	0	52	0	0	0	0	1	0
Water	104	100.0	0	0	0	104	0	0	0	0	0
Alluvium	70	92.9	0	0	0	0	65	0	0	1	4
Terrace	194	98.5	0	0	0	0	2	191	1	0	0
Duffield (limestone soil)	534	99.3	0	0	0	0	1	1	530	0	2
Berks (shale soil)	224	99.6	0	0	1	0	0	0	0	223	0
Penn (from Triassic red beds)	165	94.5	0	0	0	0	5	0	4	0	156
Total	1,513		138	28	56	104	73	192	535	225	162

A new approach now being explored is to include in the first classification attempt only the soil types that are anticipated in the extended area. For example, soils derived from metamorphic rocks such as schist, quartzite, and slate would not be included as training classes for an area derived from sedimentary bedrock. An initial subdivision into broad soil groups along the flight line would be made first. The cluster analysis would be applied to delineate these general groups and to remove the human bias in training field designation. Detailed analysis of each general group using typical classes of material for that area would follow.

ACKNOWLEDGMENTS

This research project was sponsored by the Federal Highway Administration with some support by the National Aeronautics and Space Administration, the primary funding agency of the LARS facility. The ongoing project is a portion of the Cooperative Remote-Sensing Research Program of FHWA monitored by H. T. Rib, whose efforts in coordinating the research are hereby acknowledged. Special thanks are afforded to R. L. Cunningham and H. L. Mathews, Department of Agronomy, Pennsylvania State University, who aided in the field identification of surface materials, and to P. H. Swain, of LARS, who critically reviewed the manuscript.

REFERENCES

1. Hoffer, R. M., and Goodrick, F. E. Geographic Considerations in Automatic Cover Type Identification. Proc., Indiana Academy of Science, Vol. 80, 1971.
2. Mathews, H. L., Cunningham, R. L., West, T. R., and Cipra, J. E. Application of Multispectral Remote Sensing to Soil Survey Research in Southeastern Pennsylvania. Paper presented at American Society of Agronomy, New York, 1971.
3. Rib, H. T., and Miles, R. D. Multisensor Analysis for Soils Mapping. HRB Spec. Rept. 102, 1968, pp. 22-37.
4. Tanguay, M. G., and Miles, R. D. Multispectral Data Interpretation for Engineering Soils Mapping. Highway Research Record 319, 1970, pp. 58-77.
5. West, T. R. Applications of Remote Sensing to Highway Locations. Proc., 21st Highway Geology Symposium, Lawrence, Kan., 1970.
6. West, T. R. Engineering Soils Mapping in Indiana by Computer From Remote Sensing Data. Proc., Indiana Academy of Science, Vol. 81, 1972.