1-1-1976

Stratification of Landsat Data by Clustering

Marvin E. Bauer

Barbara J. Davis

LARS Information Note 051576

Stratification of Landsat Data by Clustering

M. E. Bauer
B. J. Davis

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1976

STRATIFICATION OF LANDSAT DATA BY CLUSTERING

Marvin E. Bauer and Barbara J. Davis

Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana

I. ABSTRACT

Full realization of the potential advantages of the synoptic coverage provided by Landsat will require the development and use of data analysis techniques which take into account the large variation and diversity of patterns found over many Landsat scenes. Stratification of the scene into units which are internally homogeneous is recommended as a first step in the analysis of data for whole or multiple frames of Landsat data. The use of clustering as an objective and efficient method of dividing scenes into areas which are spectrally similar (strata) is discussed and initial results, including classification performances and comparisons of spectral strata with major physical factors, are presented.

II. INTRODUCTION

The capability for acquiring and utilizing multispectral remote sensing data was greatly increased when Landsat-1 was launched in 1972. Two of the most significant characteristics of the Landsat data are its wide area and repetitive coverage. These attributes together with machine-assisted data analysis and classification methods provide the basis for global crop production surveys in which Landsat data is used to identify and estimate the areal extent of crops.[1]

Full realization of the potential advantages provided by the synoptic Landsat coverage, however, will require the development and use of data analysis techniques which take into account the large amount of variation found in many scenes of Landsat data. Analysis techniques which are satisfactory for data acquired by airborne sensor systems or for limited areas of Landsat data cannot be effectively used to classify an entire Landsat frame of data. The diversity of landscape patterns found in Landsat data is readily seen in Figure 1. Fortunately, however, the variation found in Landsat scenes is not random, but occurs in very definite patterns. These landscape patterns are associated with the different topographic features, soils, crops, farming practices, and climatic zones found in a 10,000 square mile area.

This suggests that one of the first steps in the analysis and classification of Landsat data covering one or more Landsat scenes is to divide the scene into areas that have similar characteristics. Division of a heterogeneous population (or area) into subpopulations (or subareas), each of which is internally homogeneous is known as stratification.[2] This is suggested by the term *strata* with its implication of division into layers. Stratification is frequently used by statisticians performing surveys to increase the precision of estimates. If each stratum is homogeneous in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. Estimates from several strata can then be combined into a precise estimate for the whole population. Use of stratification in the sampling designs used for remote sensing applications is therefore advantageous. The use of Landsat data for construction of an area sampling design or frame is being developed by Wigton.[3]

A second use of stratification directly related to remote sensing applications is to permit training statistics developed for one segment or portion of the scene to be successfully used to classify other segments which are spatially and/or temporally removed from the training segment. In this context the term *spectral stratification* is useful in that it connotes the division

of the scene into areas which are internally spectrally similar. A spectral stratum may be defined as an area within which the scene and atmospheric effects are sufficiently similar that training statistics from one segment can be used to classify other segments of the stratum without significant change in classification performance. Conversely, if the same training statistics are applied to segments outside the stratum in which they were developed, classification performance will decrease.

Computer-implemented clustering techniques provide an objective and efficient method for determining the similarity of units within Landsat scenes. The objectives of our research are: (1) develop multivariate pattern recognition procedures for determining and delineating spectral strata in Landsat data and (2) determine quantitatively the physical factors which account for the spectral strata. We will discuss alternate methods to quantitatively determine and delineate spectral strata, some experimental results, and an outlook on the potential of this technique.

III. STATIC AND DYNAMIC STRATIFICATION

Stratification may be of two forms: static and dynamic. Static stratification or partitioning is the division of the geographic area of interest into subareas whose boundaries are fixed over time. Static partitioning will generally result in boundaries between major soil associations and climatic zones with different crops and cropping practices. This type of stratification can best be performed using soil, climatic, and land use maps in conjunction with Landsat imagery from appropriate seasons. Landsat imagery may be used to good advantage as a base map because many boundaries of interest will be apparent on it. However, static partitions can only use the information present in constant or slowly changing characteristics of a scene. Static stratification cannot take into account the dynamic factors of day-to-day atmospheric changes, current crop year weather patterns, and scanner system variations.

Dynamic stratification is the division of the geographic area of interest into subareas whose boundaries are dependent on changing variables and therefore not fixed over time. Examples of such dynamic variables are: a difference in crop maturity between two areas with similar crops and soils, the change in reflectance caused by a rain storm a few days before a Landsat overpass, or the division of an otherwise homogeneous area due to differences in atmospheric haze. Spectral stratification or stratification based on the spectral characteristics of Landsat data will include dynamic as well as static effects.

Both static and dynamic stratification should be beneficial in remote sensing applications. Static stratification is most applicable as the basis for constructing sampling designs and allocating sample segments. The use of Landsat imagery in stratifying land uses for this purpose has been demonstrated by Hay.[4] On the other hand, dynamic stratification based on the spectral characteristics of the scene will be useful for determining areas to which training statistics can be satisfactorily extended.

IV. USE OF CLUSTERING FOR STRATIFICATION

The technique of clustering has been adopted to define the spectral strata present in Landsat scenes. Clustering has been used extensively in remote sensing to group together units which are similar, based on observation vectors and a measure of similarity. Most remote sensing data analysts are familiar with the process of clustering pixels into spectral classes to be used later in classification. The observation vector in that type of clustering is the spectral response of the pixel in each waveband, and a commonly used measure of similarity is the Euclidean distance in the observation space.[5]

In spectral stratification, the sample unit is much larger than a single pixel and the objectives of the clustering technique are slightly different from the familiar process mentioned above. Instead of grouping together vectors of spectral responses for single pixels, we wish to group distributions of the spectral responses of sample units. Two units are spectrally similar if the distribution of spectral response in one unit is close to the distribution of the spectral response in the second area.

We can state the generalized procedure for clustering to define spectral strata in five steps.

1. Select sample units in the scene.

2. Characterize the distribution of the spectral response of each unit.

3. Choose a measure of similarity.

4. Apply a clustering algorithm to the units to determine groups of spectrally similar units.

5. Delineate the strata boundaries.

Each step and its application to stratification will explained further.

A. Selection of Sample Units

The sample units to be used in this procedure may either be segments whose geographic position has been fixed by a sampling scheme before the Landsat data is acquired or rectangular areas chosen from the Landsat data itself without regard for their geographic position. The size of the sampling unit affects the kind of strata that can be found as it is the effective lower limit on the size of strata that can be observed. For example, if the sampling unit is larger than the largest city in the scene, then urban areas cannot be separated as distinct strata. The smaller the sampling unit chosen, the smaller the geographic extent of the strata and the finer the division, or levels, that can be observed. For example, if a pixel is chosen as the sampling unit, the strata essentially are the spectral subclasses of cover types present in the scene.

B. Characterization of Spectral Response

The distribution of the spectral response within a sampling unit may be characterized in several ways. Two methods are being pursued in our research. In the first method the distribution of the spectral response in an area is represented by its first and second moments, that is, by its mean vector and covariance matrix. These parameters are easy to calculate and to use with similarity measures. However, they do not contain complete information on the skewness, multimodality, and non-normality of the distribution, all of which may be important in applying a statistical measure of distance between distributions.

A second method is essentially nonparametric. The distribution of the spectral response is characterized by the marginal density functions of the distribution. The marginal density functions rather than the joint density function are used to meet computer space limitations when dealing with large numbers of sample units. The characterization of distributions of the sample units is accomplished by first tabulating a base histogram for each feature (wavelength band) for the entire scene which is to be stratified. Equally probable bins are established from these histograms. Then a vector is constructed for each sampling unit in which each entry in the vector is the number of pixels in the sampling unit which fall in the corresponding bin in the base histograms. Thus the histograms or marginal densities of each sampling unit are characterized relative to the base histograms. The "histogram vectors" formed in this manner can then be used as data by a clustering routine.

C. Similarity Measures

In addition to the choice of characterization of the distribution of each unit's spectral response, a choice must be made of how to measure the similarity of two or more sample units. Sample units will be spectrally similar if the distance between their distributions or density functions is small. For the first method, that of representation of a distribution by its mean vector and covariance matrix, several statistical measures are possible.[5]

The transformed divergence has been the primary similarity measure used in this research as its properties are closer to the Jeffreys-Matusita distance than are the properties of divergence, yet it is computationally less complex than the Jeffreys-Matusita distance. The desirable properties of the Jeffreys-Matusita distance are that it is a metric among multivariate normal densities and it is related to the probability of error (amount of overlap) between two densities.

The implementation of these distance measures assumes that the distributions involved are multivariate normal. The assumption of normality may be violated when the sampling unit contains bad data or clouds which saturate the dynamic range of the data or when the sampling unit is divided into two distinct spectral classes, leading to bimodality. Use of large sample units has tended to alleviate the second problem, and we have tried to avoid bad data lines. Examinations of histograms have indicated that the normality assumption is not unreasonable for the data we have been using.

For the second method, that of "histogram vectors", the Euclidean distance between the vectors was chosen as a similarity measure for two reasons. First, it is a familiar measure whose properties are well known, and secondly, it has been previously implemented and extensively used in clustering analysis.

D. Clustering of Sample Units

Once a characterization of the spectral response and a distance or similarity measure have been selected, groups of spectrally similar units must be determined. If the analyst were to manually examine all possible pairs of units, the process would quickly become unwieldly and the results difficult to interpret for a large number of units. For example, if 150 units are to be stratified, over 10,000 pairwise comparisons are necessary. A

machine-implemented clustering algorithm calculates the many pairwise distances and combines the information before presenting the analyst with the natural groups of the sample units.

Two clustering algorithms have been applied in this research. The first is an iterative algorithm which has been available for both observation space and parameter space clustering.[5] The algorithm can be simply stated in its general form.

Step 1. Determine initial group centers.

Step 2. Assign each unit to the nearest group center.

Step 3. If no unit has changed allegiance, go to step 4. Otherwise, determine new group centers and return to step 2.

Step 4. If groups are distinct, stop. Otherwise modify the number of groups, determine new group centers, and return to step 2.

In our research this algorithm has been applied to cluster units characterized by their mean vectors and covariance matrices in the parameter space, and to cluster the histogram vectors in the observation space manner.

The second clustering algorithm is a systematic procedure for grouping spectrally similar units in such a way as to minimize the total number of groups while avoiding the grouping of non-similar units.[6] This procedure is slightly more complex than the first, as is seen in the following description.

Step 1. Assign each unit to its own, G_1, G_2,..., G_n.

Step 2. Order the pairwise distances $\{d_{ij}\}$, by magnitude. The algorithm considers $\{d_{ij}\}$ in ascending order. Let d_{xy} equal the smallest d_{ij}.

Step 3. If $d_{xy}>T$, a threshold of non-similarity, grouping is completed. Otherwise, proceed to step 4.

Step 4. If the units x and y belong to the same group, go to step 7. Otherwise proceed to step 5.

Step 5. Construct the average distance $\bar{d}_{xu}$ between G_x and each other group $G_u \neq G_x$ for which $d_{ab} \leq T$ for all a in G_x and b in G_u. The average distance between groups is defined as the average of all pairwise distances between units in the different groups.

Step 6. If $\bar{d}_{xy}$ is the minimum of the set of inter-group distances constructed step 5, then combine G_x and G_y into one group.

Step 7. Set d_{xy} to the next d_{ij} and return to step 3.

We have used this algorithm to group the spectrally similar units characterized by their mean vectors and covariance matrices.

E. Delineation of Strata Boundaries

After clustering is completed, the strata boundaries are delineated. Presently, this process is done manually when full Landsat frames or portions of frames have been stratified, although in the future we intend to adapt the "Extraction and Classification of Homogeneous Objects" (ECHO) approach to establish the boundaries of strata determined on the basis of fixed segments or a small sample of a Landsat frame.[7] When fixed segments based on a sampling scheme are stratified, a list of the segments in each stratum is produced rather than a map since this is the knowledge desired in this case and since the geographic location of strata boundaries between the segments is uncertain. That is, even though it is known that the boundary is between certain segments, the exact location is unknown.

V. EVALUATION OF STRATIFICATION RESULTS

The success of stratification of Landsat data by clustering is being measured in two ways, classification performance and correlation with physical factors. The criteria for success are first that classification accuracies for all segments within a stratum classified using training statistics developed within a given stratum should be similar and secondly the strata should correspond with major agronomic and other physical factors.

A. Classification performance.

To statistically evaluate a stratification, two or more areas with known crop identification data must be available within each stratum. These test areas should fall entirely within the stratum, and should be large enough to conduct a reasonable classification analysis. Such a data set will give an adequate test of the stratification of the test areas, but can not be used to determine the accuracy of the strata boundaries.

Classification results from one stratification of segments in Landsat scenes acquired June 12, 1974 for central Kansas are presented in Table 1. Each segment is either a 5x6 or 3x3 square mile area for which the crop types and Landsat

data coordinates of the agricultural fields are known. The stratification procedure treated the segments as the sample units and characterized each segment by its first and second moments. The procedure placed the segments from Stafford, Ellis, Ellsworth, and Rice Counties in one stratum along with one of the segments from Barton County. The other segment from Barton County was placed in a different stratum. Both of the procedures described in section IV.D gave the same result when transformed divergence was used as the similarity measure.

The classification results show that the stratification technique was successful in identifying segments which are indeed different. In no case was a high classification performance achieved when using training statistics from segments outside the stratum. For segments identified as members of the same stratum, similar high (approximately 90 percent correct) classification performances were obtained for both local and non-local classifications of several combinations of segments. This indicates that these segments are from the same stratum. But, in several other instances the non-local classification result was lower than the local classification performance, indicating these segments may be from different strata. This would mean that the clustering procedure is grouping the segments into groups or strata which are too broad.

Similar results have been obtained with two other data sets. With the available data, however, we cannot state with certainty whether the stratification procedure should be modified or whether the inconsistencies in results are due to limitations of the available data sets. Lack of a more adequate test data set is a major problem at this time; greater emphasis will need to be placed on this requirement of stratification evaluation before additional progress can be made.

B. Correlation with Physical Factors

The accuracy of the stratification can be assessed indirectly by comparing the strata found by clustering with maps of physical factors which are known to influence spectral response. Presently the Landsat imagery, strata maps, and physical factor maps are being compared manually. Later, when the physical factor data are digitized, we plan to conduct a regression analysis which will quantify the degree of correlation between the strata and various physical factors. Such an analysis will not only provide a measure of the accuracy of stratification, but also provide quantitative information on the influence of major agronomic and meteorological factors on spectral reflectance. The physical factors being investigated include crop maturity stage, soil association, land use, precipitation, temperature, and grain yield.

The illustrations in Figures 1-4 permit a qualitative comparison of Landsat imagery, spectral stratification, and soil and land use maps for the same area of southwestern Kansas. The spectral stratification shown in Figure 2 was produced by the "histogram vectors" method described in section IV.B. Only the marginal density function from band five was used so that the information would correspond to that present in the Landsat image shown in Figure 1. The sample units in this example are 50 pixels x 50 pixels or roughly 2½ miles x 2 miles.

The soil association map shown in Figure 3 exhibits several features easily seen in both the Landsat imagery and the spectral stratification. The areas of the Udic Ustolls (12) are easily visible, as are the patterns of the Typic Ustolls (9, 10, and 11).[8] The land use map, Figure 4, was developed from Landsat imagery acquired during June and July 1973.[9] Almost two years later, the same land use patterns appear again in the May 21, 1975 image shown in Figure 1.

VI. OUTLOOK ON THE USE OF SPECTRAL STRATIFICATION

Large scale surveys using satellite-acquired multispectral data require classifications to be made over areas at least the size of individual Landsat scenes. The diversity of landscape patterns found over many areas of this size indicates that a logical first step in the analysis and classification of Landsat data is to stratify or divide the scene into units which are internally similar. Such a stratification will be helpful in constructing sampling frames which minimize the variance among sample units and in determining the boundaries of areas over which training statistics can be satisfactorily extended.

Stratification for sampling purposes can be based on static factors whose boundaries are either static or change only very slowly. For classification, however, the stratification should be based on the Landsat spectral data and will include the effects of dynamic as well as static factors.

The use of computer-implemented clustering procedure for dynamic stratification has been developed and tested over several Landsat scenes of Kansas. Initial results indicate that the technique can be used to determine the similarity of sample

units and that the strata produced agree with major physical factors. The use of such a procedure should enable scenes to be more efficiently and objectively stratified than would be possible using manual methods.

We recommend that stratification be considered a prerequisite of signature extension or signature adjustment algorithms such as the multiplicative and additive signature correction (MASC) technique described by Henderson.[10] Our observation of results from such algorithms is that the results are highly variable and are data dependent. This shortcoming may be largely overcome by applying such signature adjustment algorthms only within a stratum, thus taking advantage of the knowledge gained from spectral stratification.

VII. REFERENCES

1. MacDonald, R.B., F.G. Hall, and R.B. Erb. 1975. The Use of Landsat Data in a Large Area Crop Inventory Experiment (LACIE). Proceedings, Symposium on Machine Processing of Remotely Sensed Data, June 3-5, 1975, Purdue University, West Lafayette, Indiana (IEEE Catalog No. 75 CH 1009-0-C), pp. 1B, 1-23.

2. Cochran, W.G. 1963. Stratified Random Sampling. pp. 87-113. *In* Sampling Techniques. John Wiley and Sons, Inc., New York.

3. Wigton, W.H. 1976. Applications of Landsat Data to Crop Acreage Estimation by SRS. Proceedings, Symposium on Machine Processing of Remotely Sensed Data, June 29-July 1, 1976, Purdue University, West Lafayette, Indiana.

4. Hay, C.M. 1974. Agricultural Inventory Techniques with Orbital and High-Altitude Imagery. Photogrammetric Engineering 40: 1283-1294.

5. Wacker, A.G. and D.A. Landgrebe. 1972. The Minimum Distance Approach to Classification. Information Note 100771, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

6. Davis, B.J. and P.H. Swain. 1974. An Automated and Repeatable Data Analysis Procedure for Remote Sensing Applications. Proceedings, Ninth International Symposium on Remote Sensing of Environment, April 15-19, 1974, Ann Arbor, Michigan, pp. 771-774.

7. Kettig, R.L. and D.A. Landgrebe. 1975. Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects. IEEE Transactions of Geoscience Electronics, GE-14: 19-26.

8. Bidwell, O.W., and C.W. McBee. 1973. Soils of Kansas. Kansas Agricultural Experiment Station, Department of Agronomy Contribution No. 1359.

9. Williams, D.L., and B.L. Barker. 1974. Kansas Land-Use Patterns. Space Technology Laboratories, University of Kansas, Lawrence, Kansas.

10. Henderson, R.G. 1976. Signature Extension Using the MASC Algorithm. IEEE Transactions on Geoscience Electronics, GE-14: 34-37.

VIII. ACKNOWLEDGEMENTS

The research reported in this paper was funded by Contract NAS9-14016 from the National Aeronautics and Space Administration. The assistance provided by Dr. Philip Swain and Ms. Tina Cary of LARS is gratefully acknowledged.

Table 1. Classification Performances (Wheat vs. Other) for Segments Within and Outside of Strata Determined by Clustering

Strata No.	Source of Training Statistics	Areas Classified*					
		Barton-1	Barton-2	Rice	Ellsworth[1]	Ellis	Stafford
		Overall Percent Correct					
1	Barton-1	83.7	42.9	15.1	69.4	54.1	61.5
2	Barton-2	27.1	96.0	93.8	90.0	56.2	52.5
2	Rice	34.1	92.0	93.4	85.7	47.4	69.1
2	Ellis	63.4	43.4	26.4	60.4	64.8	51.4
2	Stafford	58.2	55.4	42.0	59.9	61.7	89.9

* Landsat scenes 1689-16392 and 1689-16385 acquired June 12, 1974 over Central Kansas.

[1] Ellsworth was not used as a source of training statistics because only wheat field coordinates were available.

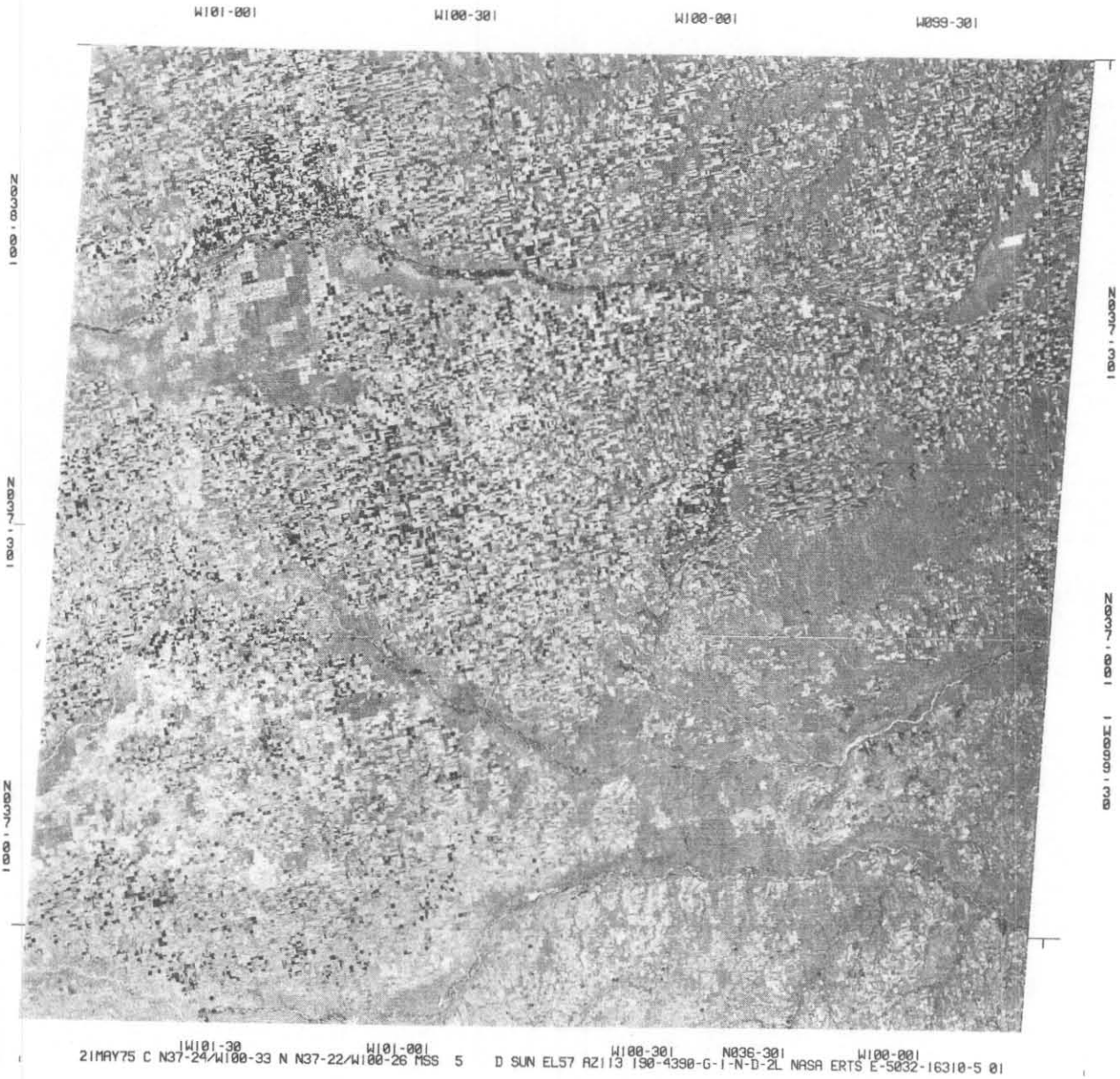

Figure 1. Landsat Scene 5032-16310, Band 5 (0.6-0.7μm), Acquired May 21, 1975 over Southwestern Kansas. Several landscape units or strata corresponding to different soils and land uses are present in the scene.

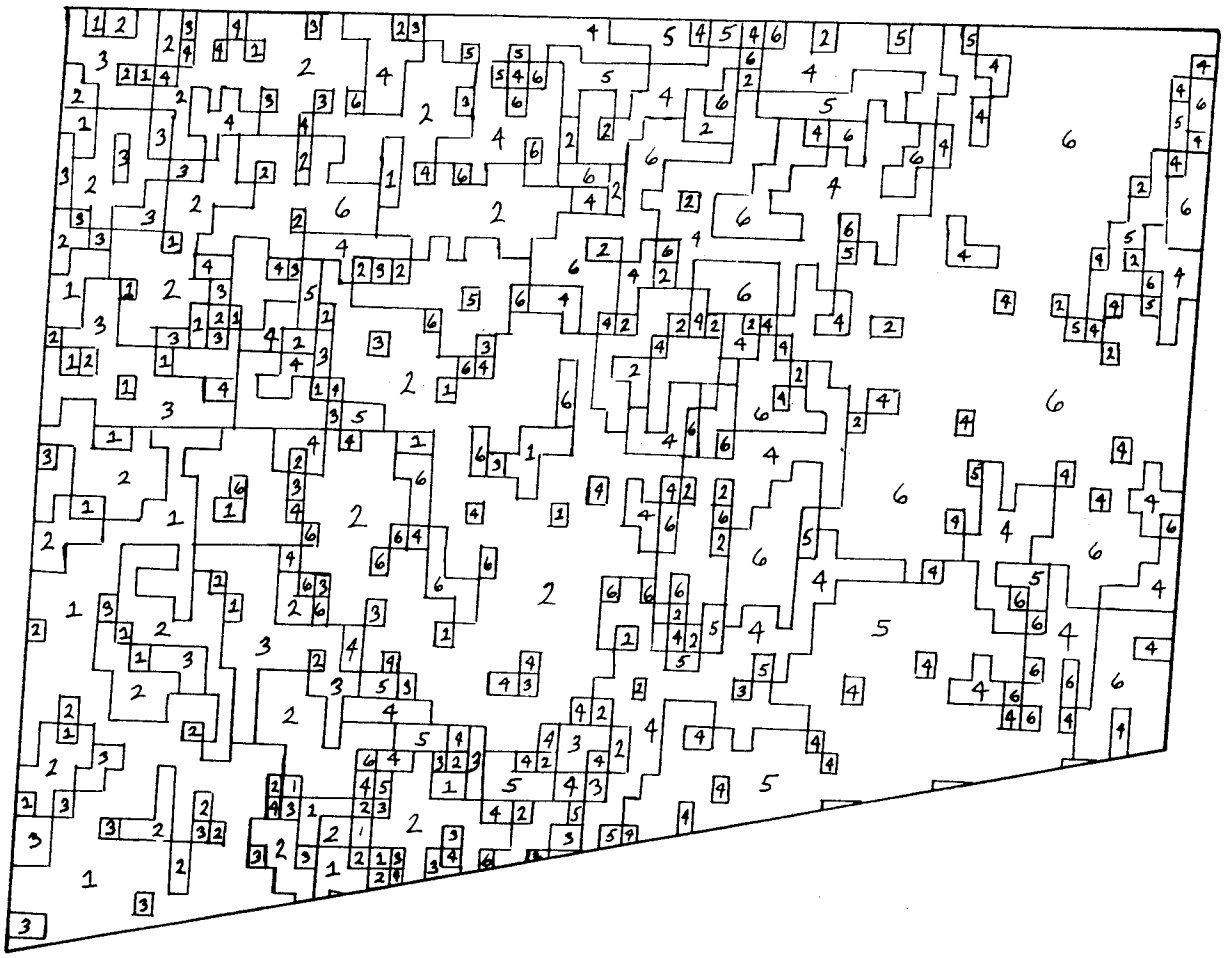

Figure 2. Machine-implemented Stratification of the Kansas Portion of Landsat Scene 5032-16310. Each number represents a different stratum.

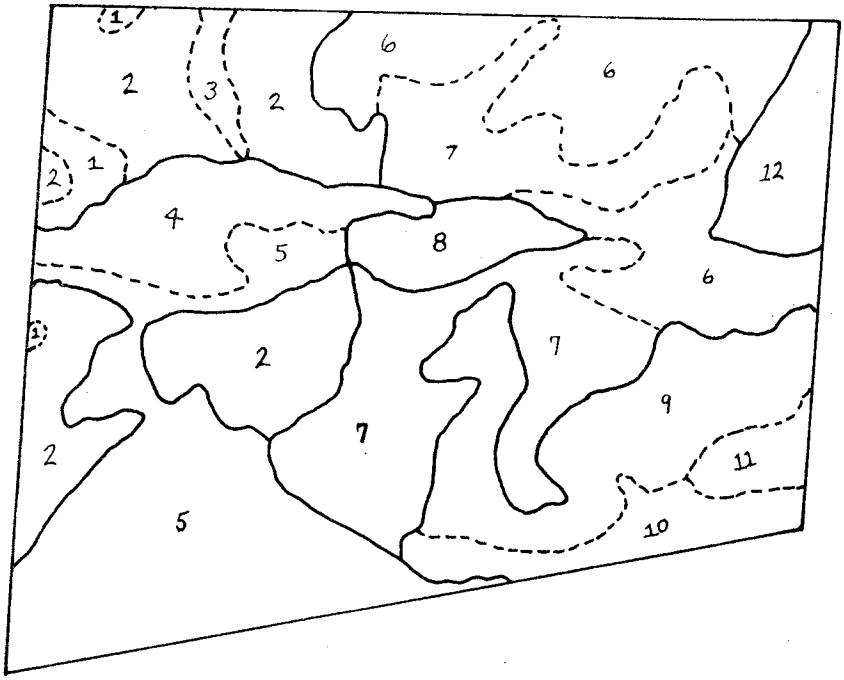

Figure 3. Soil Association Map of the Kansas Portion of the Landsat Scene shown in Figure 1.

SOILS ASSOCIATIONS

ARIDIC USTOLLS

Ustolls, Orthents, and Ustalfs
Deep, grayish-brown and dark grayish-brown silt loams
1. Ulysses, Colby
2. Richfield, Ulysses
3. Ulysses, Drummond

Ustalfs, Psamments, Ustolls, and Argids
Deep, grayish-brown silt loams and sandy loams, and pale-brown loamy fine sands and fine sands
4. Tivoli, Vona
5. Dalhart, Richfield, Vona

TYPIC USTOLLS

Ustolls and Usterts
Deep and moderately deep, dark grayish-brown silt loams and moderately deep, gray clays
6. Harney, Uly, Wakeen
7. Harney, Spearville

Ochrepts, Ustolls, Ustalfs, and Psamments
Moderately deep and shallow, reddish-brown loams and clays, and deep, grayish-brown silt loams and clay loams and pale-brown loamy fine sands and fine sands
8. Manter, Pratt
9. Mansic, Mansker
10. Tivoli, Pratt
11. Woodward, Carey

UDIC USTOLLS

Ustalfs, Ustolls, and Aquolls
Deep, dark grayish-brown loams and fine sandy loams and pale-brown loamy fine sands
12. Pratt, Carwile

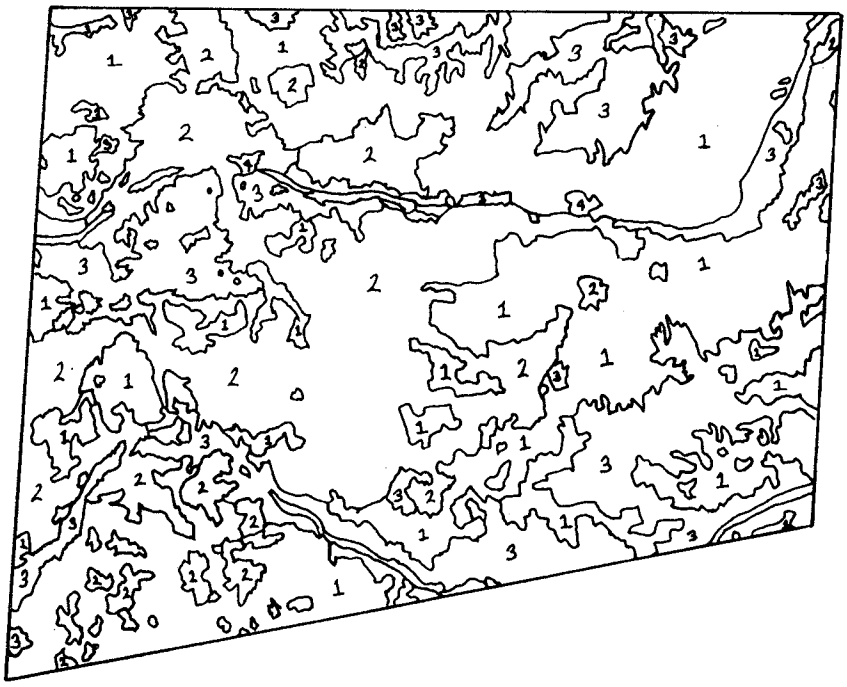

Figure 4. Map showing Major Land Use Categories for the Kansas Portion of the Landsat Scene shown in Figure 1.

LAND USE CATEGORIES

1. Unirrigated - areas with greater than 50% unirrigated cropland
2. Irrigated - areas with greater than 50% irrigated cropland
3. Rangeland - areas with greater than 50% rangeland
4. Urban and built-up land
5. Water and wetlands