E82-10038
TM - 82198

NASA

Technical Memorandum 82098

A Comparison of Unsupervised Classification Procedures on Landsat MSS Data for an Area of Complex Surface Conditions in Basilicata, Southern Italy

Chris Justice and John Townshend

(E82-10038) A COMPARISON OF UNSUPERVISED CLASSIFICATION PROCEDURES ON LANDSAT MSS DATA FOR AN AREA OF COMPLEX SURFACE CONDITIONS IN BASILICATA, SOUTHERN ITALY (NASA) 40 p HC A03/MF A01 CSCL 05B G3/43

N82-19613

Unclas
00038

MARCH 1981

National Aeronautics and
Space Administration

Goddard Space Flight Center
Greenbelt, Maryland 20771

TM 82098

A COMPARISON OF UNSUPERVISED CLASSIFICATION PROCEDURES ON LANDSAT MSS DATA FOR AN AREA OF COMPLEX SURFACE CONDITIONS IN BASILICATA, SOUTHERN ITALY

Chris Justice
and John Townshend[1]
Earth Resources Branch
Code 923
Goddard Space Flight Center

March 1981

1. Both authors hold National Research Council resident research post-doctoral fellowships in the Earth Resources Branch (Code 923) at GSFC.

A COMPARISON OF UNSUPERVISED CLASSIFICATION PROCEDURES ON LANDSAT MSS DATA FOR AN AREA OF COMPLEX SURFACE CONDITIONS IN BASILICATA, SOUTHERN ITALY

ABSTRACT

In this study, two unsupervised classification procedures are applied to ratioed and unratioed Landsat MSS data of an area of spatially complex vegetation and terrain. An objective accuracy assessment is undertaken on each classification and a comparison is made of the classification accuracies. The two unsupervised procedures use the same clustering algorithm. By one procedure the entire area is clustered and by the other, a representative sample of the area is clustered and the resulting statistics are extrapolated to the remaining area using a maximum likelihood classifier. Explanation is given of the major steps in the classification procedures including image preprocessing; classification; interpretation of cluster classes; and accuracy assessment. Of the four classifications undertaken, the monocluster block approach on the unratioed data gave the highest accuracy of 80% for five coarse cover classes. This accuracy was increased to 84% by applying a 3 X 3 contextual filter to the classified image. A detailed description and partial explanation is provided for the major misclassifications. In outline, classification of the unratioed data produced higher percentage accuracies than for the ratioed data and the monocluster block approach gave higher accuracies than clustering the entire area. The monocluster block approach was additionally the most economical in terms of computing time.

PRECEDING PAGE BLANK NOT FILMED

OUTLINE

I. INTRODUCTION

II. DESCRIPTION OF THE STUDY AREA

III. DEFINITION OF COVER CLASSES

IV. METHODOLOGY AND TECHNIQUES

- i) Image preprocessing
- ii) Classification procedure and accuracy assessment techniques

V. DESCRIPTION OF CLASSIFICATION RESULTS

- i) Ratioed data – clustering the entire area
  - – monocluster block approach
- ii) Unratioed data – clustering the entire area
  - – monocluster block approach
- iii) Classification of agricultural sites

VI. COMPARISON OF RESULTS AND EXPLANATION OF MISCLASSIFICATIONS

- i) Comparison of clustering methods
- ii) Comparison of results from ratioed and unratioed data
- iii) Misclassifications for the monocluster block approach with unratioed data

VII. CONTEXTUAL CONSIDERATIONS

VIII. DISCUSSION OF RESULTS AND CONCLUSION

A COMPARISON OF UNSUPERVISED CLASSIFICATION PROCEDURES ON LANDSAT MSS DATA FOR AN AREA OF COMPLEX SURFACE CONDITIONS IN BASILICATA, SOUTHERN ITALY

I. INTRODUCTION

In the last seven years the NASA Landsat series of satellites have provided remotely sensed data for many parts of the world. During this time studies have been undertaken to demonstrate the utility of the data for a wide variety of Earth resources applications. One of the most fruitful applications areas, has been the use of multispectral scanner (MSS) data for surface cover mapping (e.g. NAS 1978). Surface cover mapping involves the identification and discrimination of vegetation or surface materials followed by classification into surface cover types. The most successful results have been obtained for large areas of contrasting cover types and units, with little or no topography, suitable for discrimination at the spectral and spatial resolutions of the MSS system. Several recent studies have examined the problems of cover type identification in more complex areas of small and mixed cover units with rugged terrain (e.g. Hoffer and Staff 1975, Fleming 1977).

This paper is part of a series of studies to examine methods for interpreting Landsat data of such an area of complex surface conditions in southern Italy (Justice et al. 1976, Justice 1978, Townshend and Justice 1980). The objective of this particular study is to examine the success with which two unsupervised classification procedures can be applied to ratioed and unratioed Landsat MSS data for areas of spatially complex vegetation and terrain. The two unsupervised procedures compared in this study both use the same clustering algorithm. By one procedure the entire study area is clustered and by the other a representative sample of the area is clustered and the cluster statistics are then extrapolated to the remaining area using a maximum likelihood classifier. The latter procedure is known as the monocluster block approach and has been used by Fleming (1977) and Townshend and Justice (1980). As part of this study a thorough and objective accuracy assessment is undertaken of each of the classifications and the accuracy results are compared for both ratioed and unratioed data. A further analysis was undertaken to examine possible improvements to the classification by applying a contextual filter to the classified data.

The following seven sections of this paper provide: a description of the study area; a definition of the cover classes; a description of the methodology and classification procedure; the classification results; a comparison of the results from the four classifications; a description of the results from the contextual filter and finally a summary and conclusion.

II. DESCRIPTION OF THE STUDY AREA

The study area covers approximately 743 square kilometers (512 X 350 Landsat pixels) and is located in Basilicata Region, Southern Italy (Figure 1). The geological, morphological, pedological and botanical phenomena found in this area are representative of many areas within the Mediterranean region and as such provide a useful test site from which to extrapolate results.

The geological structure of the area is dominated by the Sant Arcangelo Basin, which was infilled in the Pliocene and Pleistocene and shows a continuous depositional sequence from coarse unconsolidated conglomerates in the west, to fine clays in the east. The Quaternary Basin is bordered to the west and east by Eocene nappe formations of sandstones and flysch which create a hilly terrain of moderate ruggedness. Within the study area the nappe formations rise to 864 m at Mount St. Arcangelo. The conglomerates form a deeply incised tableland at approximately 500 m. The sand deposits are heavily dissected and have undergone considerable faulting and subsidence. The marine clays are characterised by a series of cuestas in the north of the study area and rolling convexo-concave terrain to the south.

The area is traversed by two major river systems, the Agri and Sinni, which drain, predominantly west-to-east from the southern Appennines to the Gulf of Taranto, Gulley networks occur throughout the study area where rapid Quaternary uplift combined with poor land use management has contributed to a severe soil erosion problem (Williams 1981).

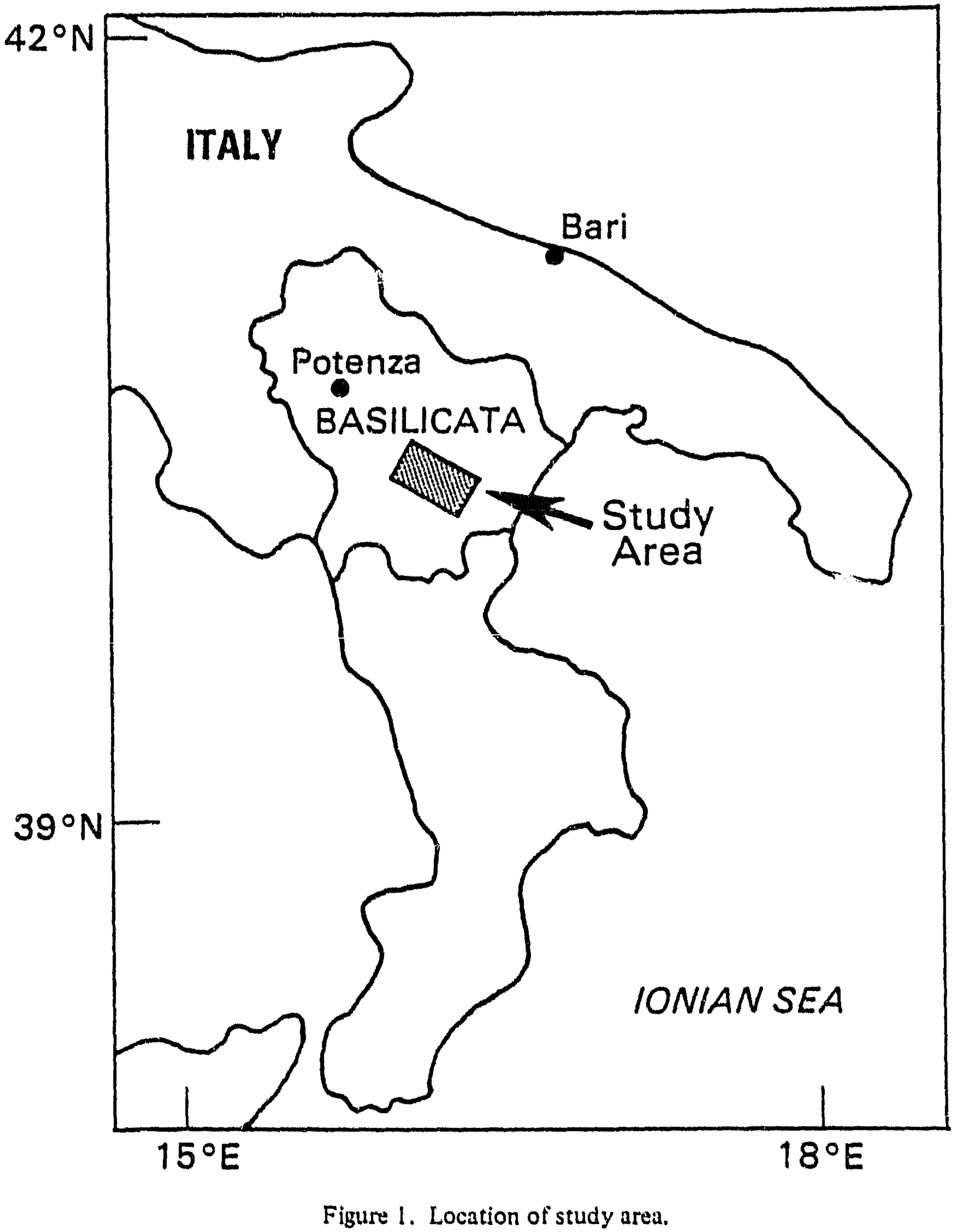

Figure 1. Location of study area.

Intensive use of the land throughout history at least since 500 B.C. has led to substantial alteration of the natural vegetation communities. Four major altered vegetation communities occur within the area, deciduous oak woodland, evergreen oak woodland, open macchia and riparian scrub. The deciduous woodland community occurs on the conglomerate deposits and at the summit of Mt. Sant Arcangelo both in open and closed stands. The lower more sheltered areas on the conglomerates host a degenerate evergreen oak community. Open degraded macchia is the dominant altered vegetation community in the central and eastern parts of the study area. This consists of low sclerophyll evergreen shrubs surrounded by rough grass. The community exists in a wide variety of densities and maturity on open and rugged hillslopes and on the margins of and in the bottoms of the gulley systems.

Agricultural and managed pasture land make up the major remaining parts of the study area. Farming with a large subsistence component is predominant except for mechanized wheat farming in the rolling claylands. The dependence on subsistence farming has led to cultivation wherever possible, a complex land tenure system and an interculture of tree and grain crops. A two-year rotation scheme of wheat and fallow is implemented by the larger holdings but the majority of the smaller land tenure units have no regular rotation scheme. Olives are still an important crop for the subsistence farmer and olive groves are scattered throughout the study area. Where subsistence farming occurs it gives rise to the following vegetated landscape: small arable plots of wheat or vetch with scattered fruit trees and vines; small olive groves, irregularly spaced with underlying arable or grazing land; clumps of thinned deciduous oak trees; heavily grazed macchia in the valley bottoms and bordering the areas of accelerated erosion. Market farming is undertaken along the more fertile valley floodplains. The tenure units are often very small, c. 1/8th hectare and cultivation is intensive. Most farming families own a small herd of sheep and goats which are kept as mixed flocks and graze on virtually all the uncultivated land. The open grazing land which consists generally of rough grass with scattered evergreen shrubs and occasional deciduous trees, occurs extensively on the footslopes of Mt. Sant Arcangelo.

III. DEFINITION OF COVER CLASSES

For the purposes of land cover classification it is necessary to assign the surface conditions occurring within the study area, to specific cover types or classes. The term 'cover type' is used loosely within remote sensing circles, to refer to both vegetation and exposed surface materials including soils. The classification scheme should be designed to satisfy two requirements which are often conflicting i.e., suitability for spectral discrimination and utility for subsequent applications. For studies that consider such requirements, a compromise is often attempted. To be suitable for spectral discrimination the surface cover classes should be defined by ground variables which have been shown to be highly correlated to spectral response such as vegetation type and density, chlorophyll content, physiognomy, soil color, and moisture content. These often differ from the parameters that would enable inferences to be made about land-use types such as subsistence farming.

To take account of the mixture of cover types typical of this study area, classes were defined to give an indication of the dominant and secondary cover types at a site, with names such as "herbaceous cover with trees and shrubs." A physiognomic subdivision which facilitated field description of cover types, was selected as the basis for the classification. To be more quantitative the percentage cover of each physiognomic type was estimated for each site. Broad physiognomic classes and bare surface types were defined to include the major cover types found within the study area. Although describing the degree of mixture at a site, a purely physiognomic classification did not adequately separate agricultural and non-agricultural cover types. Three agricultural classes were added to the physiognomic subdivision, namely arable rotation, olive and fruit trees and market gardens. The major cover classes occurring in this study are listed in Table 1.

Two other conflicting demands which need to be considered when defining the cover classes are precision and accuracy of discrimination. The user usually requires both high precision and high accuracy which in practice have a negative relationship and thus result in an inevitable trade-off. The refinement of the classes will be dependent on the discriminating ability of the Landsat scanner

Table 1
The major cover types occurring within the study area.

Water (Reservoir)	Herbaceous with shrubs and/or trees
Water (River channels)	Evergreen shrubs
Bare river gravels	Mixed evergreen and deciduous shrubs
Bare eroded slopes	Orchards and Olive groves
Eroded slopes with shrubs	Open woodland (deciduous)
Bare ground with herbaceous	Closed woodland (evergreen)
Herbaceous (permanent pasture)	Closed woodland (deciduous)

system and the sophistication of the classification techniques used. Often it may well be that the categories discriminable using Landsat data are less precise than those ideally required. Users of Landsat data should be aware both of the type of categories obtainable from such data and of the way in which they can relate to more detailed land cover classes derived from other data sources. The most common cover classification system used for Landsat analysis is the hierarchical scheme developed for land use mapping by Anderson et al. (1972) for the U.S. Geological Survey, but in several respects it is poorly suited to conditions found in our study area. In particular, where several cover types exist at an individual site, classification into a single cover type, such as the dominant one, is unlikely to be satisfactory.

IV. METHODOLOGY AND TECHNIQUES

The description of the methodology adopted in this study is subdivided into two sections, firstly image preprocessing and classification procedures, and secondly accuracy testing. All the image processing and analysis was undertaken using the Electromagnetic Systems Laboratory (ESL) Interactive Digital Image Manipulation System (IDIMS) at the ERRSAC (Eastern Region Remote Sensing Applications Center) Facility, NASA/Goddard Space Flight Center.

The IDIMS system includes a Hewlett Packard (HP) 3000 series III minicomputer, an ESL Advanced Scientific Array Processor (ASAP) with an HP 21 MX minicomputer, three disc drives with 450 megabytes of storage, a Comtal color image display, Versatec printer plotters and an Optronics color film recorder. A more comprehensive description of the system is given by Campbell (1980).

The imagery used in this analysis was recorded by Landsat 1 on August 8, 1972. The high sun angle imagery was selected to avoid misclassification caused through possible topographic effects on the data (Holben and Justice, 1979). Panchromatic aerial photography (1:20,000) flown at the same season, two years after imaging, was used to aid the interpretation of the Landsat data.

i) Image preprocessing

Selected image processing techniques were applied to the Landsat data to provide the optimum image for classification.

Destriping was undertaken using the IDIMS Histnorm function (ESL 1976) to reduce the six line banding, which was particularly noticeable on MSS Channel 4. This function uses a relatively simple procedure of normalizing the mean values for the six detectors in each channel of the MSS system to either the maximum or minimum value or the middle four values averaged. The user is prompted for multiplication factors to adjust the mean and standard deviation for any detector and create a destriped image.

A simple atmospheric correction was undertaken to remove the different diffuse light components in the four MSS channels. This method is known as dark area subtraction and is demonstrated in Bentley et al. (1976): it involves subtracting a constant value equal to the darkest response on the image, from all the pixel values.

Linear contrast stretching was undertaken for each MSS channel to provide an improved visual product. A stretch function was also applied to the pixels, thereby rendering the image more comparable geometrically with maps and aerial photography. This was achieved by multiplying Landsat lines and samples by factors of 7 and 5, respectively. MSS channel 5 of the final processed image used is shown in Figure 2.

Figure 2. The final processed image of Landsat MSS 5 for the study area.

Spectral band ratioing of the form (channel i/channel j) was undertaken, since a previous study (Justice 1978) indicated that ratioed data may lead to improved classification of cover types. The theory behind ratioing is that multiplicative environmental factors affecting the spectral response can be reduced by dividing one channel by another (Vincent 1973, 1977) though Holben and Justice (1980) demonstrated that for some areas there may be serious limits to the degree of reduction of environmental factors that can be expected from such ratioing. Four ratio combinations were produced for this study namely 7/5, 4/7, 5/6, 7/6, and these were combined into one multiband image. These ratios were chosen based on preliminary analysis of ratioing for cover discrimination by Justice (1978).

ii) Classification Procedure and Accuracy Assessment Techniques

Unsupervised classification procedures are characterized by the use of image properties to produce an initial definition of the cover types, which are subsequently interpreted after classification of the data, as distinct from supervised classification techniques, which define the cover classes to be discriminated prior to classification. The term 'unsupervised' can be misleading since extensive user interaction is usually required to implement the technique. Implementation of the unsupervised technique normally requires the definition of parameters to control the size and number of classes prior to the classification, followed by interpretation and regrouping of the cluster classes after classification.

The unsupervised classification procedure used in this study was the ESL IDIMS 'ISOCLS' function (ESL 1976, Townshend and Justice 1980). ISOCLS is a clustering algorithm which requires input by the analyst of maximum standard deviation and minimum distance parameters to control splitting and combining of the clusters. Parameters are used to control the number and minimum size of cluster classes derived from the data. Ideally the user requires a knowledge of the approximate number of cluster classes finally needed. Approximately twice the number of cluster classes ultimately required were created in each cluster analysis, to allow for the same cover type being represented by different spectral responses in different locations. Extensive field work in the study area revealed approximately 10 major cover types for discrimination, excluding water surfaces which covered only a small part of the study area.

Two methods of applying the clustering algorithm were used in this study: firstly, applying the clustering algorithm to the entire study area and secondly, applying the clustering algorithm to a representative sample of the area and then extrapolating the cluster statistics to the remaining study area using a maximum likelihood technique.

The principal advantage in performing the latter monocluster block approach is that it reduces the computer time involved in clustering the whole area (Fleming 1977). The monocluster block approach used in this study, involved selection of four sample areas representative of the cover conditions occurring in the area. The four sample areas, which amounted to 6.3 percent of the total area, were merged to create one image, which was subsequently clustered. The sample areas contained the major cover classes occurring in the study area and were selected based on field experience.

After clustering the image into a suitable number of classes it is necessary to identify the cluster classes in terms of the ground conditions they represent. Accurate identification of the cluster classes requires detailed knowledge of ground conditions. For this study, information from previous field visits and aerial photography were used to interpret the cluster classes. An earlier study to assess the interpretability of the aerial photographs revealed that the major physiognomic composition could be identified consistently with 95% accuracy (Townshend and Justice 1980). A two-level interpretation of the cluster classes was undertaken, firstly by identifying homogeneous areas of each class occurring within the study area and locating and interpreting these areas on the aerial photographs and secondly by examining each of the cluster classes occurring in the monocluster block sample areas and identifying the classes on the aerial photographs. The two-level interpretation was undertaken both interactively on the color Cromtal display and by using hard-copy color products created using the Optronix system, at approximately the same scale as the aerial photographs. Interpretation of the cluster classes was facilitated by ranking the clusters in order of the 7/5 ratio prior to displaying the classes. This gave an indication of the amount of green biomass in the class (Richardson and Wiegand 1977, Tucker 1979) and provided more homogeneous areas of similar classes for interpretation.

When the final classified image had been produced by clustering the entire area or by extrapolating the monocluster block statistics, interpretations were undertaken for each of the cluster classes, and percentage cover estimates of the four major physiognomic classes were made, i.e., bare, herbaceous, shrubs, and trees. Interpretations of the cluster classes for the different parts of the study area were then compared and limits for the interpretation classes formulated, which included the range of ground conditions within each cluster class. Similar classes were then grouped together to produce the final number of interpreted classes required. This stage resulted in a reduction in the precision of discrimination of the cover classes representing differing proportions of cover types in different parts of the study area. One way to preserve precision may be to stratify the area prior to classification, but this was not undertaken as part of this study.

When the cluster classes had been identified and where necessary grouped together to provide the final number of interpreted classes on the classified image, the accuracy testing phase was executed. Objective accuracy assessment is critical in evaluating the utility of the classification. To facilitate accuracy assessment a set of random testing sites was created which was then used to evaluate the several different classifications. The testing set included 6 sites, of 3 X 3 pixels for each of the 10 major cover types. The percentage cover of the four major physiognomic classes was then estimated for each randomly selected test site. These were then located visually on the Landsat scene using the Cromtal color video display. The locations of these test sites were stored in terms of line and sample coordinates and subsequently transferred to each classified image. The individual test sites were then assigned to their respective interpreted cluster classes using percentage cover criteria. The cluster class of each pixel within the training sites was listed and a confusion matrix created to show the errors of commission and omission and to provide the final percentage accuracies. Once the final accuracies were calculated a final stage of regrouping the classes was undertaken to provide the optimum classification accuracy for a given range of cover classes.

V. DESCRIPTION OF CLASSIFICATION RESULTS

This section describes and compares the accuracy results for the four classifications undertaken using the two classification procedures on both multiband ratioed and non-ratioed images (Figure 3). The first sub-section presents the results for the ratioed data by clustering the entire study area and subsequently by using the monocluster block approach. The second sub-section presents results

2. RATIOED DATA MONOCLUSTER-BLOCK APPROACH	4. NON-RATIOED DATA MONOCLUSTER-BLOCK APPROACH
1. RATIOED DATA CLUSTERING THE ENTIRE STUDY AREA	3. NON-RATIOED DATA CLUSTERING THE ENTIRE STUDY AREA

Figure 3. The relationship between the four classification schemes examined in this study.

for the unratioed data in the same order. The third sub-section describes the results for the agricultural test sites.

i) Classification results from the ratioed data

The 25 cluster classes derived by clustering the entire study area (Figure 3, Box 1), were interpreted and regrouped to form the 10 classes described in Table 2. The 10 classes were classified with an overall accuracy of only 36.4% (Table 3). Low accuracies (<20%) were found for the open woodland, herbaceous with trees and shrubs and one herbaceous class. Regrouping the classes into five major cover classes gave an improved accuracy of 67%. Misclassification of deciduous and evergreen woodland, with herbaceous and herbaceous with trees and shrubs (Table 3, classes 6 and 7) accounted for the particularly poor accuracy figures for this classification.

Table 2.
Table showing the cover classes derived from the ratioed data by clustering the entire area.

Final class number	Original cluster number	% trees and shrubs	% herbaceous	% bare ground	Class description
1	1	< 3	< 3	>97	Bare ground
2	21, 14, 9	<10	<25	90-96	Bare ground
3	18	<10	10-35	65-89	Bare with herbaceous
4	24, 5, 17	<10	36-94	10-64	Herbaceous with bare
5	15, 12, 20, 3, 22	< 5	>95	<10	Herbaceous
6	10, 7, 2, 6	5-15	>85	<10	Herbaceous
7	4, 23, 16, 11	15-30	65-84	<10	Herbaceous with trees and shrubs
8	25, 19	35-70	<65	<20	Open woodland
9	8, 13	>70	<30	<20	Closed woodland
10		<10	0-100	0-100	Agriculture Rotation

Table 3
Unsupervised classification of ratioed Landsat data by clustering the entire study area.

Actual Classes: Final Classes	Predicted Classes: 1	2	3	4	5	6	7	8	9	n	Percent Correct Classification	Percent Correct Classification for the Regrouped Classes
1	17	9	1							27	63.0	63.0
2	3	13	2	5	4					27	48.1	
3	2	13	9	17	2	2				45	20.0	81.4
4		1	2	4	2					9	44.4	
5			1	10	21	13				45	46.7	
6			1	4	3	1				9	11.1	82.0
7					10	26				36	0	
8						23	41	17	18	99	17.2	31.0
9					2	28	29	21	55	135	40.7	
10		23	3	14	7	7				54	74.1	74.1

Overall Accuracy

10 classes including arable rotation sites = 36.4%
5 classes including arable rotation sites = 66.9%
4 classes excluding arable rotation sites = 62.0%

Twenty preliminary clusters were obtained by applying the monocluster block approach to the ratioed data (Figure 3, Box 2). These cluster classes were interpreted and regrouped into the 10 interpretation classes described in Table 4. The 10 classes were discriminated with an overall accuracy of 44.3% (Table 5). Major misclassification occurred between open woodland and the two herbaceous classes with trees and shrubs. Regrouping the classes into five coarse cover classes gave an accuracy of 71%. Remaining misclassification was highest between herbaceous cover and herbaceous with shrubs and trees.

Table 4.
Table showing the cover classes derived from the ratioed data by applying the monocluster block classification

Final class number	Original cluster number	% trees and shrubs	% herbaceous	% bare ground	Class description
1	1	≤10	<20	>80	Bare ground (River Gravels)
2	2	≤20	<33	66-80	Bare ground (Eroded)
3	3	≤20	<66	33-80	Bare ground with herbaceous
4	4, 5, 6	≤ 5	66-84	<33	Herbaceous, with some bare ground
5	7, 8, 9, 10	<15	>85	< 5	Herbaceous
6	11, 12, 13, 14	15-40	60-84	<10	Herbaceous with shrubs and trees
7	15, 16, 17	41-60	<60	≤20	Herbaceous with trees and shrubs
8	18, 19	61-90	≤40	≤20	Open woodland with herbaceous and bare ground
9	20	>90	≤10	≤10	Closed woodland
10		<10	0-100	0-100	Agricultural Rotation

Table 5

Unsupervised classification of ratioed Landsat data using the monocluster block approach.

Actual Classes: Final Classes	Predicted Classes: 1	2	3	4	5	6	7	8	9	n	Percent Correct Classification	Percent Correct Classification for the Regrouped Classes
1	15	19	11	7	2					54	27.8	
2	2	10	1	2						15	66.7	63.2
3		1	13	28	2	1				45	28.9	
4										0	0	71.1
5			1	14	18	12				45	40.0	
6				5	14	33	2			54	61.1	61.1
7						10	25	14	5	54	46.3	
8						34	17	21	7	79	26.6	78.5
9						4	10	31	45	90	50.0	
10		10	11	20	6	8				54	68.5	68.5

Overall Accuracy

10 classes including arable rotation sites = 44.3%
5 classes including arable rotation sites = 71.2%
4 classes excluding arable rotation sites = 71.6%

ii) Classification results from the unratioed data

The 23 cluster classes derived by clustering the entire multiband unratioed image (Figure 3, Box 3) were interpreted and regrouped to produce the 10 interpretation classes described in Table 6. Overall classification accuracy for the 10 classes was 43% (Table 7). Misclassification was high between the herbaceous and herbaceous with shrubs and trees class (Table 7, classes 4 and 5) and between herbaceous with shrubs and trees, and shrubs and trees with herbaceous (Table 7, classes 5 and 6). Regrouping the classes into five coarse cover classes gave an accuracy of 71%, but misclassification remained high between the herbaceous and bare ground classes and between deciduous woodland and the evergreen trees and shrub class.

Table 6
Table showing the cover classes derived from the unratioed data by clustering the entire area.

Final class number	Original cluster number	% trees and shrubs	% herbaceous	% bare ground	Class description
1	1, 2, 3	<10	<20	>80	Bare ground
2	4, 5	<10	≤50	50-80	Bare ground with herbaceous
3	6	< 3	50-65	35-49	Herbaceous with bare ground
4	7, 8, 9	<15	>65	<35	Herbaceous
5	10, 11, 12	15-40	60-80	<10	Herbaceous with shrubs and fruit trees
6	13, 14, 15	40-60	40-60	<10	Evergreen shrubs and trees with herbaceous
7	16, 17, 18, 19	>60	<40	<10	Evergreen trees and shrubs with herbaceous
8	20, 21	60-90	<40	<10	Open Woodland (deciduous)
9	22, 23	>90	<10	<10	Closed Woodland (deciduous)
10		<10	0-100	0-100	Agricultural Rotation

Table 7
Unsupervised classification of ratioed Landsat data, by clustering the entire study area.

Final Classes (Predicted Classes)	Predicted Classes 1	2	3	4	5	6	7	8	9	n	Percent Correct Classification	Percent Correct Classification for the Regrouped Classes
1	33	16	2	3						54	61.0	
2	16	23	12	1	2					54	42.6	81.5
3										0	0	
4		20		12	18	4				54	22.2	40.4
5		7		8	10	20				45	22.2	
6						14	25	14	1	54	25.9	
7						15	63	1		90	70.0	81.2
8						5	14	8		27	29.6	
9							16	13	25	54	46.2	56.8
10	24	13	4	5	7	1				54	85.1	85.1

Overall Accuracy

10 classes including arable rotation sites = 43.5%
5 classes including arable rotation sites = 71.0%
4 classes excluding arable rotation sites = 69.2%

Sixteen clusters were derived using the monocluster block approach on the unratioed data (Figure 3, Box 4). These were interpreted and regrouped to form the 10 classes described in Table 8. The overall accuracy for these 10 classes was 59% (Table 9). Misclassification was highest between the herbaceous with trees and shrubs and the herbaceous with trees class (Table 9, classes 5 and 4). When regrouped into the five coarse cover classes, the overall accuracy increased to 80.2%, which was the highest percentage accuracy of all the four classification schemes examined. The remaining misclassification was highest between the deciduous and evergreen woodland classes.

Table 8
Table showing the cover classes derived from the unratioed data derived by applying the monocluster block approach.

Final class number	Original cluster number	% trees and shrubs	% herbaceous	% bare ground	Class description
1	1, 13, 15	<10	<20	>80	Bare ground
2	3, 5, 9	<10	20-75	25-80	Bare ground with herbaceous
3	7, 2	10-19	>75	10-25	Herbaceous with bare ground and shrubs
4	4, 11	10-19	>80	<10	Herbaceous
5	2	20-33	66-80	<10	Herbaceous with ever-green trees
6	10	34-50	40-65	<10	Evergreen shrubs with herbaceous
7	8	>50	<50	<20	Dense evergreen shrubs with herbaceous
8	16	>33	<66	<20	Evergreen woodland
9	6, 14	>33	<66	<20	Deciduous woodland
10		<10	0-100	0-100	Agricultural Rotation

Table 9
Unsupervised classification of unratioed Landsat data using the monocluster block approach.

Actual Classes: Final Classes	Predicted Classes: 1	2	3	4	5	6	7	8	9	n	Percent Correct Classification	Percent Correct Classification for the Regrou[illegible] Classes
1	22	38	3							63	34.9	
2		42		3						45	93.3	94.4
3										0	0	
4	1	2	28	15	1	7	1			55	27.3	84.9
5			7	10	1					18	5.6	
6			7		10	15	2			34	44.2	
7				1	16	3	25	7	2	54	46.3	72.8
8						1	13	44	5	63	69.8	
9					10	2	11	7	78	108	72.2	72.2
10	6	32	7	8		1				54	98.1	98.1

Overall Accuracy

10 classes including arable rotation sites = 59.9%
5 classes including arable rotation sites = 80.2%
4 classes excluding arable rotation sites = 78.4%

iii) Classification accuracies for the arable rotation sites

As no ground data concerning the physiognomic conditions of the agricultural areas was available for the time of imaging, it was necessary to isolate the arable rotation sites to form a separate cover class. The agricultural test sites were classified with a relatively high degree of accuracy (76.8%) for all data sets (Tables 3, 5, 7 and 9). The highest accuracies were obtained for the unratioed data, independent of the clustering procedure used. Inclusion of the agricultural class with the four semi-natural cover classes, increased the overall accuracy in all but one case. Misclassifications frequently occurred between the agricultural sites and the herbaceous (permanent pasture) cover classes, which is certainly understandable in terms of their spectral similarity. It should be made clear that the accuracies quoted refer to distinguishing the rotational arable classes from cover types with contrasted physiognomic properties, and not to distinguishing them from cover types with similar physiognomic properties. For the latter situation accuracies would inevitably be much poorer.

VI. COMPARISON OF RESULTS AND EXPLANATION OF MISCLASSIFICATIONS

This section is divided into three sub-sections, the first of which provides a comparison of the accuracy results for the two clustering methods used in the study. The second sub-section compares the accuracy results derived using the ratioed and unratioed data. The third sub-section provides an explanation for some of the major misclassifications for the monocluster block approach for the unratioed data.

i) A comparison of classification accuracies obtained by clustering the entire area and those obtained using the monocluster block approach.

Results obtained using the monocluster-block approach gave higher classification accuracies than those obtained by clustering the entire image (Figure 4). A completely satisfactory explanation has not been found for this, but we hypothesise that selection of the monocluster block sample sites provides a bias to the type and variability of the final classes required. Selection of these sample areas to include typical cover types, will inevitably reduce the 'noise' from atypical cover

	RATIOED DATA	UNRATIOED DATA
MONOCLUSTER BLOCK APPROACH	2 10 classes (inc. agric.) = 44% 5 classes (inc. agric.) = 71% 4 classes (exc. agric.) = 71%	4 10 classes (inc. agric.) = 59% 5 classes (inc. agric.) = 80% 4 classes (exc. agric.) = 78%
CLUSTER ENTIRE AREA	1 10 classes (inc. agric.) = 36% 5 classes (inc. agric.) = 67% 4 classes (exc. agric.) = 62%	3. 10 classes (inc. agric.) = 43% 5 classes (inc. agric.) = 71% 4 classes (exc. agric.) = 69%

Figure 4. Summary of classification results for the four classification schemes.

types which would be incorporated when forming clusters for the entire area. Two observations from the analysis support this hypothesis. Firstly, correct classification of woodland classes was consistently higher for the monocluster block approach (Tables 3, 5, 7 and 9). The choice of sample areas for clustering included some of the most uniform and homogeneous woodland areas which were sufficiently large and spectrally distinctive to form a separable cluster class. Secondly, classification of the first bare ground class, which in all cases represented river gravels, was higher by clustering the entire image than by using the monocluster block approach. Visual examination of a standard color composite of the area, after the analysis, showed the river gravels to have a higher degree of spectral diversity than was represented by the sample areas.

(ii) A comparison of classification accuracies derived from the ratioed and unratioed data. Classification accuracies for the unratioed data were consistently higher than for the ratioed data (Figure 4). The result appears to contradict the preliminary findings reported by Justice (1978) but it should be noted that different criteria were used to define the classes in the discriminant analysis performed in this previous study. Although a topographic effect can be expected in August Landsat data (Justice et al. 1980), it does not appear either to have affected the unratioed data sufficiently to limit classification accuracy or, alternatively, to have been removed by band-ratioing. Two distinct groups of cover classes were derived using ratioed and unratioed data, which may have affected the resulting accuracies. Both class descriptions for the unratioed data (Tables 6 and 8) included evergreen shrub classes which were absent from the classes derived from the ratioed data (Tables 2 and 4). The classes represented more dominant herbaceous cover classes. Woodland classes were classified with a higher accuracy using the unratioed data than with the ratioed data, though individual herbaceous classes were classified with higher accuracies using the ratioed data. Evergreen woodland was discriminated from other deciduous woodland classes most successfully using the unratioed data.

iii) Description and explanation of the major misclassifications for the monocluster block analysis of the unratioed data.

A detailed examination of the major misclassifications will provide a partial explanation for the general levels of accuracy achieved during this study. Although only results from the optimum scheme are discussed, the more general explanations apply to all the classifications. The distribution of the original 16 clusters derived from clustering the sample areas for the unratioed data are presented for the MSS 5 and MSS 7 feature space in Figure 5. There is good separability with no overlap for all the clusters for MSS 5 and MSS 7 but poorer separation in the MSS 4 and MSS 6 feature space (Figure 6). It is likely that misclassification could well arise between classes 1 and 2 and classes 9 and 5, when the data from the remaining parts of the study area are assigned to these classes. The distribution of the clusters in the MSS 4 and MSS 5 feature space (Figure 7) shows a strong correlation between the channels and their high degree of redundancy for discrimination. Cluster classes for cloud, cloud shadow and water are not shown in these figures, but these classes were sufficiently spectrally distinct for all classification schemes to warrant no further analysis.

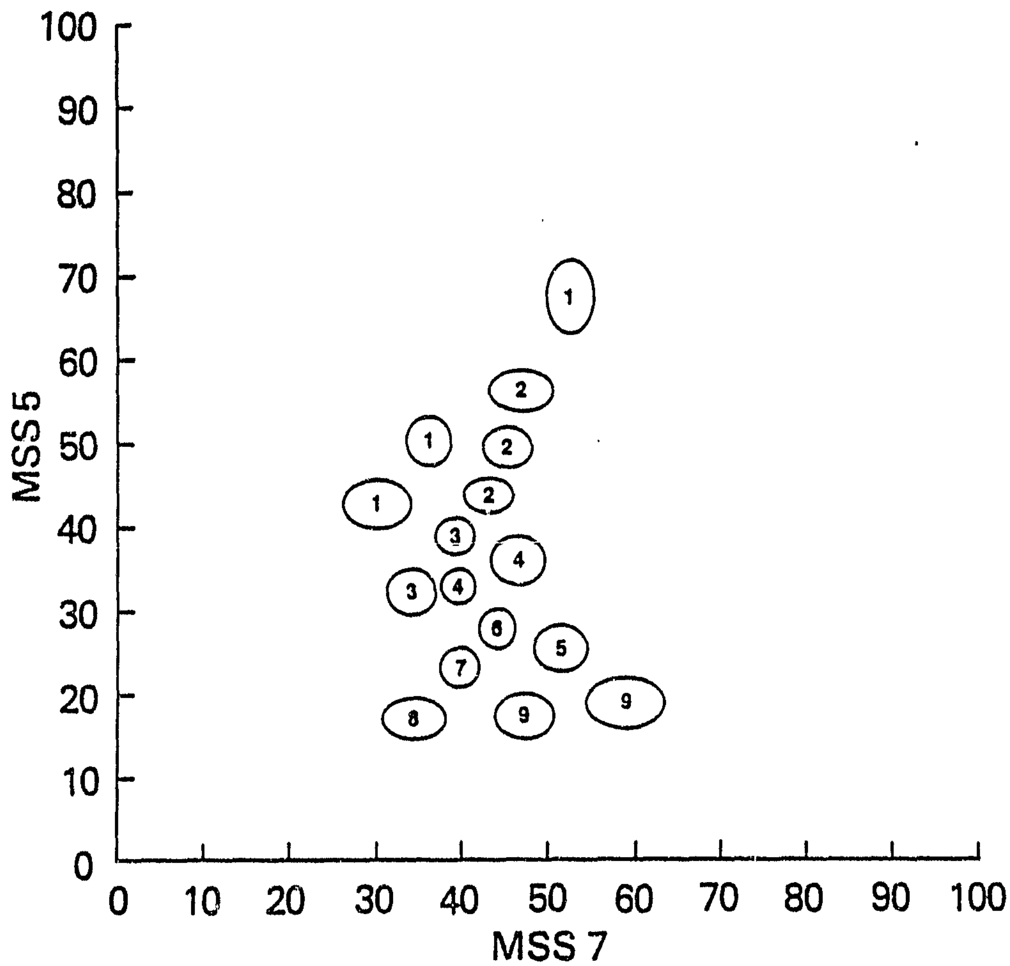

Figure 5. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 5 and 7 feature space.

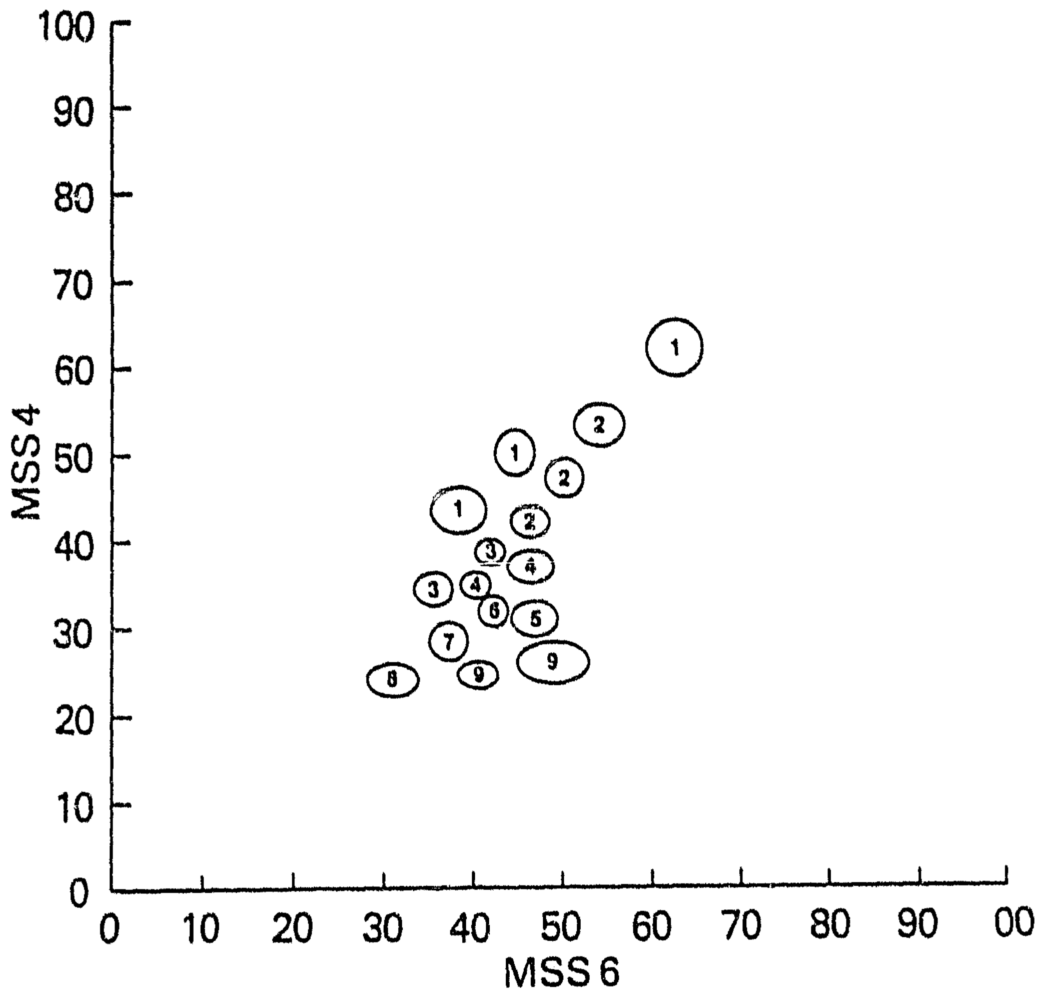

Figure 6. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 4 and 6 feature space.

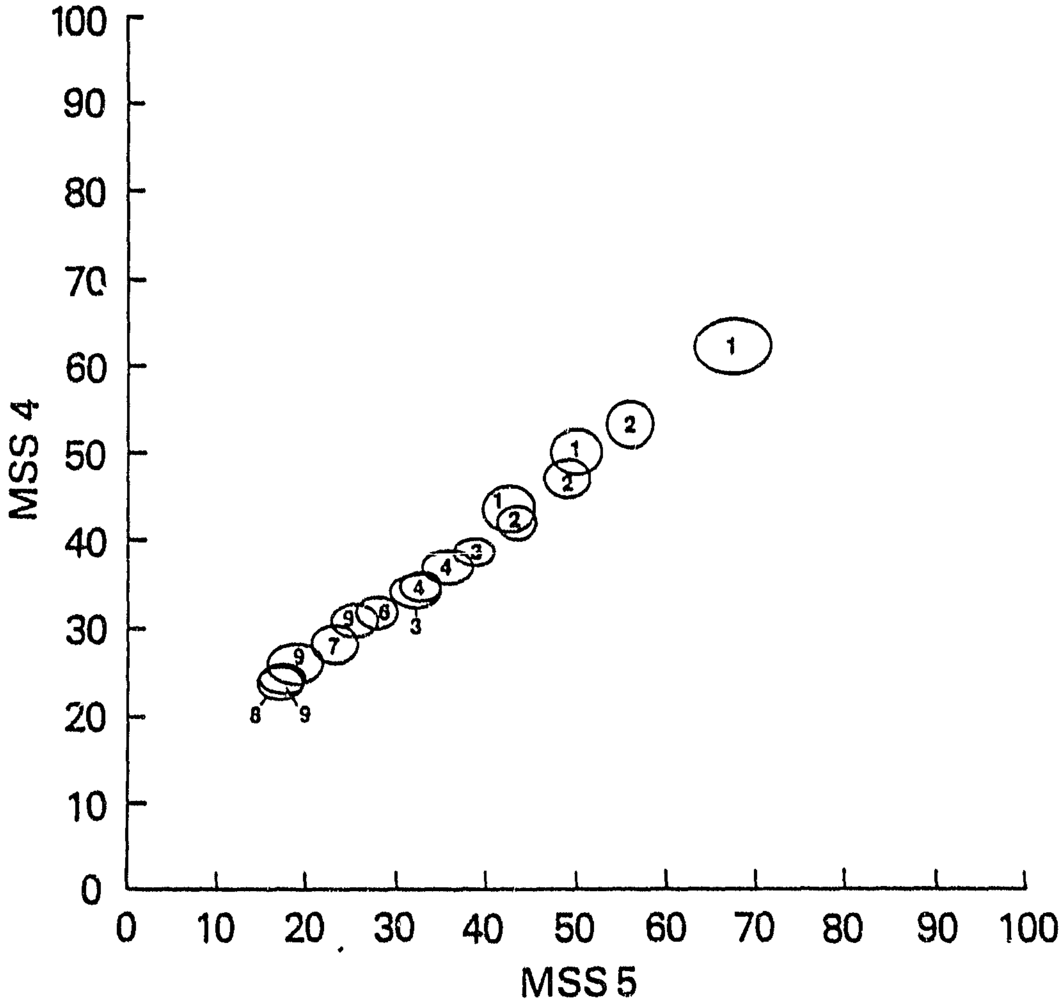

Figure 7. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 4 and 5 feature space.

Examination of the confusion matrix for the above classification (Table 9) and the results for the individual testing sites, showed the herbaceous class to be assigned to the largest number of cover classes. Fifty-one percent of the herbaceous class (Table 9, class 4) was misclassed as herbaceous with bare ground. There is no immediate explanation for this, apart from the wide variety of ground conditions which fall under the herbaceous physiognomic category.

The largest single percentage misclassification was between classes 1 and 2 (Table 9), where bare ground was misclassed as bare ground with herbaceous. The only test site with 100% bare ground was classified correctly; all other sites showed some confusion with class 2. In the final classification, no consideration was made of material types such as the distinction between river gravels and eroded clays; discrimination between sites was based purely on percentage cover criteria. The inherent spectral diversity between the bare areas, which is indicated by three clusters for class 1 in Figure 5, may account in part for the high degree of misclassification. Some of the misclassification between the evergreen shrubs and trees with herbaceous class and the herbaceous class (Table 9, classes 6 and 5) occurred for olive grove sites, the understory of which may have considerably altered the spectral response. A similar confusion may have arisen between the market garden sites in class 7 and herbaceous with evergreen shrubs (class 5) and may be explicable by the minimum percentage cover of trees (> 33%) used to define the woodland class. A general observation from several of the classes is that misclassification often occurred for those testing sites that came close to the class limits of the cover class. Some of the misclassifications were overcome by regrouping adjacent classes, but this led to a reduction in the precision of classification.

VII. CONTEXTUAL CONSIDERATIONS

So far only spectral information has been used in the classification of the pixels. Additionally, we can use contextual information concerning the classes of surrounding pixels to modify the classification of a pixel. This potentially has the benefit of improving classification accuracies by removal of isolated inliers within homogeneous areas. The procedure used involved the execution of the

Reclass function of IDIMS (ESL 1976). Specifically, each pixel was reassigned to the most common class of its eight immediate neighboring pixels. Although large homogeneous areas are not typical of the study area, comparison of Tables 9 and 10 shows that a modest improvement in the overall accuracy of 4% was achieved. The regrouped bare class and herbaceous class both showed improvement and a substantial improvement occurred in the evergreen shrub with herbaceous understory category.

VIII. DISCUSSION OF RESULTS AND CONCLUSION

The results and experience of this study have indicated certain methods that may lead to improvement of the present class accuracies. Division of the classification feature space into approximately 20 clusters is achieved by applying the same statistical thresholds, i.e. maximum standard deviation and minimum distance criteria to all the data. It is likely that a more subtle division of the feature space could be achieved equally successfully by stratifying the data prior to final clustering. For example, on the basis of visual examination of the data, finer discrimination of deciduous woodland cover types and bare surface types than was achieved by this study, appears feasible. Similarly, stratification of the study area into areas with a similar range of cover types and ground conditions, prior to classification, would reduce the loss in precision experienced when the same cluster class represents different physiognomic characteristics in different parts of the study area. Furthermore, improvements in classification may be achieved by adjusting the cover classification scheme to incorporate more than just physiognomic criteria. The scheme used in this study relies heavily on a strong relationship between physiognomic composition and spectral response, which probably does not always exist.

Of the four unsupervised schemes examined in this study (Figure 3), the monocluster block approach on the unratioed data gave the highest classification accuracies. When 'reclassed' using a 3 X 3 pixel grid, the accuracy results were 61% for 10 cover classes and 84% for 5 cover classes. The monocluster block approach on the unratioed data was also the most economical in terms of

Table 10

Unsupervised classification of unratioed Landsat data using the monocluster block approach. Reclassed using a contextual filter.

Actual Classes: Final Class Number	Predicted Class: 1	2	3	4	5	6	7	8	9	n	Percent Correct Classification	Percent Correct Classification for the Regrouped Class
1	19	43	1							63	30.2	
2		42		3						45	93.3	96.2
3										0	0	
4		1	36	10		7				52	19.2	90.0
5			8	10						18	0	
6			9			18				27	66.7	
7				2	18		32	2		54	59.2	72.2
8					9	1	14	37	2	63	58.7	
9				1	10		9	6	82	108	75.9	75.9
10	1	35	9	9						54	100.0	100.0

Overall Accuracy

10 classes including arable rotation sites = 60.7%
5 classes including arable rotation sites = 84.3%
4 classes without arable rotation sites = 82.3%

computing time. Classification of the unratioed data produced higher percentage accuracies than for the ratioed data and the monocluster block approach gave higher accuracies than clustering the entire study area. The results from the different classifications were on the whole disappointing, the majority of classes being discriminated with less than 80% accuracy.

The results from this study can be compared with those presented by Townshend and Justice (1980). In the latter study, unsupervised classification was undertaken on ratioed data for the same four sample areas used in this present study. Instead of creating an objective testing set of random sites, Townshend and Justice (1980) selected two test areas which gave conflicting accuracy figures. Accuracies of 84.7% (4 classes, excluding agricultural sites) and 65.5% (7 classes, excluding agricultural sites) were achieved for the two sites. These results were a little higher than those shown in Box 2 Figure 4, and show the importance of developing a representative testing set to derive a realistic statement of the success of classification. It would appear that the accuracy figures quoted are not an underestimate as indicated in the previous paper and provide a fair indication of the classification accuracies obtainable for the time of imaging using this approach.

Selection of the sample areas used for monocluster block classification, plays an important part in determining the success of the classification. Although the four selected areas were known to possess examples of all the major cover classes within the study area, the choice of these areas in terms of size and cover variability was largely arbitrary. One consideration arising from applying the monocluster block method is that extrapolation of the cluster statistics by the maximum likelihood rule may lead to inclusion of numerically insignificant cover classes unless the minimum size and number of clusters is carefully controlled during clustering of the sample blocks.

However rigorous the design of the testing phase, the reliability of the testing will ultimately depend on the success with which the test sites are located on the satellite data. With the present resolution of Landsat MSS, location of random test sites will remain difficult in areas with a high degree of mixture and may possibly lead to a spurious increase in errors of misclassification. With the advent of higher resolution satellite systems ground location will undoubtedly become easier. This as much as the improved spatial resolution *per se* may help improve classification accuracies in areas with terrain as complex as the area described in this paper.

ACKNOWLEDGEMENTS

The authors wish to thank the personnel of the Eastern Regional Remote Sensing Applications Center (NASA/GSFC) for their assistance in facilitating our use of the ESL IDIMS facility and to thank the Earth Resources Branch for providing support to enable us to use this image processing system. Helpful suggestions on an earlier draft were provided by Ross Nelson and Steve Wharton.

This work was performed while the authors were National Research Council Resident Research Associates in the Earth Resources Branch at NASA/GSFC.

BIBLIOGRAPHY

Anderson, J. R., E. E. Hardy and J. T. Roach 1972. A land use classification scheme for use with remotely sensed data. U.S. Geological Survey Circular, 671, Washington. pp 16.

Bentley, R. G., C. Bette, D. Salmon *et al.* 1976. A Landsat study of ephemeral and perennial rangeland vegetation and soils. Final Report, Bureau of Land Management.

Campbell, W. 1980. ERRSAC interactive computer systems for image processing. Eastern Regional Remote Sensing Application Center, Goddard Space Flight Center: NASA, pp 9.

ESL (Electromagnetic Systems Laboratory) Inc. 1976. IDIMS User's Guide. Technical Memorandum ESL-TM 705, Sunnyvale, California.

Fleming, M. D. 1977. Computer-aided analysis techniques for an operational system to map forest lands utilizing Landsat MSS data. LARS Technical Report 11227. Laboratory for Applications of Remote Sensing, W. Lafayette, Indiana.

Hoffer, R. M. and Staff 1975. Computer aided analysis of Skylab multispectral scanner data in mountainous terrain for land use, forestry, water resource and geologic applications. LARS Information Note 121275. Laboratory for Application of Remote Sensing, W. Lafayette, Indiana.

Holben, B. N. and C. O. Justice 1979. Evaluation of modeling of the topographic effect on the spectral response from nadir pointing sensors. NASA/GSFC TM 80305.

Holben, B. N. and C. O. Justice 1980. An examination of spectral band ratioing to reduce the topographic effect on remotely sensed data. NASA/GSFC TM 80604.

Justice, C. O. 1978. An examination of the relationships between selected ground properties and Landsat MSS data in an area of complex terrain in southern Italy. Proc. Am. Soc. Photogramm. Fall Convention, 303-329.

Justice, C. O., D. F. Williams, J. R. G. Townshend, R. A. G. Savigear 1976. The evaluation of space imagery for obtaining environmental information in Basilicata Province, Southern Italy. In: Land Use Studies by Remote Sensing (eds. W. G. Collins and J. L. van Genderen) Sheffield: Remote Sensing Society, 1-39.

Justice, C. O., S. W. Wharton, B. N. Holben 1980. Application of digital terrain data to quantify and reduce the topographic effect on Landsat data. NASA/GSFC TM 80760.

NAS (National Academy of Sciences) 1978. Resource Sensing from Space: Prospects for Developing Countries, Washington, D. C.

Richardson, A. J. and X. L. Wiegand 1977. Distinguishing vegetation from soil background information. Photogramm. Engng. and Remote Sensing, Vol. 43, No. 12, pp 1541-1442.

Townshend, J. R. G. and C. O. Justice 1980. Unsupervised classification of MSS Landsat data for mapping spatially complex vegetation. International Journal of Remote Sensing, 1, 105-120.

Tucker, C. J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment 8: 127-150.

Vincent, R. K. 1973. Spectral ratio imaging methods for geological remote sensing from aircraft and satellites. Proc. Am. Soc. Photogramm. Management Utilization of Remote Sensing Data Conference, Sioux Falls, South Dakota, 377-397.

Vincent, R. K. 1977. Geochemical mapping by spectral ratioing methods. In: Remote Sensing Applications for Mineral Exploration (ed. W. L. Smith) 252-276. Dowden, Hutchinson, and Ross Inc., Pennsylvania.

Williams, D. F. 1981. Integrated land survey methods for the prediction of gulley erosion. In: Terrain Analysis and Remote Sensing (ed. J. R. G. Townshend), George Allen and Unwin: London, 154-68.

LIST OF FIGURES

Figure 1. Location of the study area.

Figure 2. The final processed image of Landsat MSS 5 for the study area.

Figure 3. The relationship between the four classification schemes examined in this study.

Figure 4. Summary of classification results for the four classification schemes.

Figure 5. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 5 and 7 feature space.

Figure 6. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 4 and 6 feature space.

Figure 7. Distribution of the original 16 clusters for the unratioed data, monocluster block approach in MSS 4 and 5 feature space.

LIST OF TABLES

1. Major Cover Types occurring in the study area.

2. Cover classes derived from the ratioed data by clustering the entire area.

3. Classification accuracies for the ratioed data derived by clustering the entire area.

4. Cover classes derived from the ratioed data by applying the monocluster block classification.

5. Classification accuracies for the ratioed data derived by applying the monocluster block approach.

6. Cover classes derived for the unratioed data by clustering the entire area.

7. Classification accuracies for the unratioed data derived by clustering the entire area.

8. Cover classes derived for the unratioed data by applying the monocluster block classification.

9. Classification accuracies for unratioed data derived by the monocluster block approach.

10. Classification accuracies for the reclassed monocluster block approach on the unratioed data.