108

STIF

II

THE UNIVERSITY OF KANSAS SPACE TECHNOLOGY CENTER
Raymond Nichols Hall
2291 Irving Hill Drive—Campus West Lawrence, Kansas 66045

Telephone:

7.6 - 1 0. 4 8 3
CR - 148790

A Comprehensive Data Processing Plan for Crop Calendar MSS Signature Development from Satellite Imagery.

Third Progress Report

Technical Report 286-3

11 July 1976

(E76-10483) A COMPREHENSIVE DATA PROCESSING PLAN FOR CROP CALENDAR MSS SIGNATURE DEVELOPMENT FROM SATELLITE IMAGERY Progress Report (Kansas Univ.) 59 p HC $4.50 CSCL 05B G3/43 N76-31629 Unclas 00483

R. M. Haralick

G. Minden

A. Singh

Funded by:

Goddard Space Flight Center
Contract #NAS 5-20943
Greenbelt, Maryland 20771

23/2B

STIF

/08

II

THE UNIVERSITY OF KANSAS SPACE TECHNOLOGY CENTER
Raymond Nichols Hall

2291 Irving Hill Drive—Campus West Lawrence, Kansas 66045

Telephone:

7.6-10.483
CR-148790

A Comprehensive Data Processing Plan for Crop Calendar MSS Signature Development from Satellite Imagery.

Third Progress Report

Technical Report 286-3

11 July 1976

(E76-10483) A COMPREHENSIVE DATA PROCESSING PLAN FOR CROP CALENDAR MSS SIGNATURE DEVELOPMENT FROM SATELLITE IMAGERY Progress Report (Kansas Univ.) 59 p HC $4.50 CSCL 05B G3/43 N76-31629 Unclas 00483

R. M. Haralick

G. Minden

A. Singh

Funded by:

Goddard Space Flight Center
Contract #NAS 5-20943
Greenbelt, Maryland 20771

23/2B

CRES —— REMOTE SENSING LABORATORY

TEMPORAL CROP SIGNATURES

The Kansas intensive LACIE test site data consists of three to five LANDSAT images of each site taken in the fall, early spring, and early summer. We have treated this multi-temporal data set in a classical way with non-parametric supervised decision rules and have determined poor results from crop discriminations. Some of these results were reported last quarter.

During this quarter, we have been discovering the reasons for poor performance. We have done careful comparisons of the category statistics and cross compared these statistics with those obtained from our unsupervised clustering approach.

1. Signature Comparison

Temporal data were collected and graphed by the ground truth categories for Finney and Morton Counties. The mean gray tone values were computed by band and category and were plotted. These plots were presented in Figures 1.1 and 1.2. The plots for ground truth categories are easily discriminated for the Finney County test site. However, in Morton County the plots are very similar except for wheat and grass. One would expect that corn, summer fallow, grain sorghum, and rye would be easily confused. Looking at the same category in both counties, we see that the plots are not similar at all.

For example, wheat in Finney County tends to have a wider range of values. Band 6 for wheat increases, decreases, then increases again in Finney County, whereas it decreases, then increases in Morton. Band 7 decreases in brightness after April in Finney County, but in Morton County band 7 is increasing after May 9.

Clearly it would be difficult to use data from one test site to discriminate categories in another.

To see what regions the spatial clustering procedure was finding, the temporal plots for 6 clustered regions were produced. These are presented in Figures 1.3 and 1.4. The plots for the first clusters in Figure 1.3 match very closely with the ground truth category plots for Finney County in Figure 1.1. Cluster 1 matches with wheat in Figure 1.1; Cluster 2 matches with corn, etc. As the cluster number increases, it is

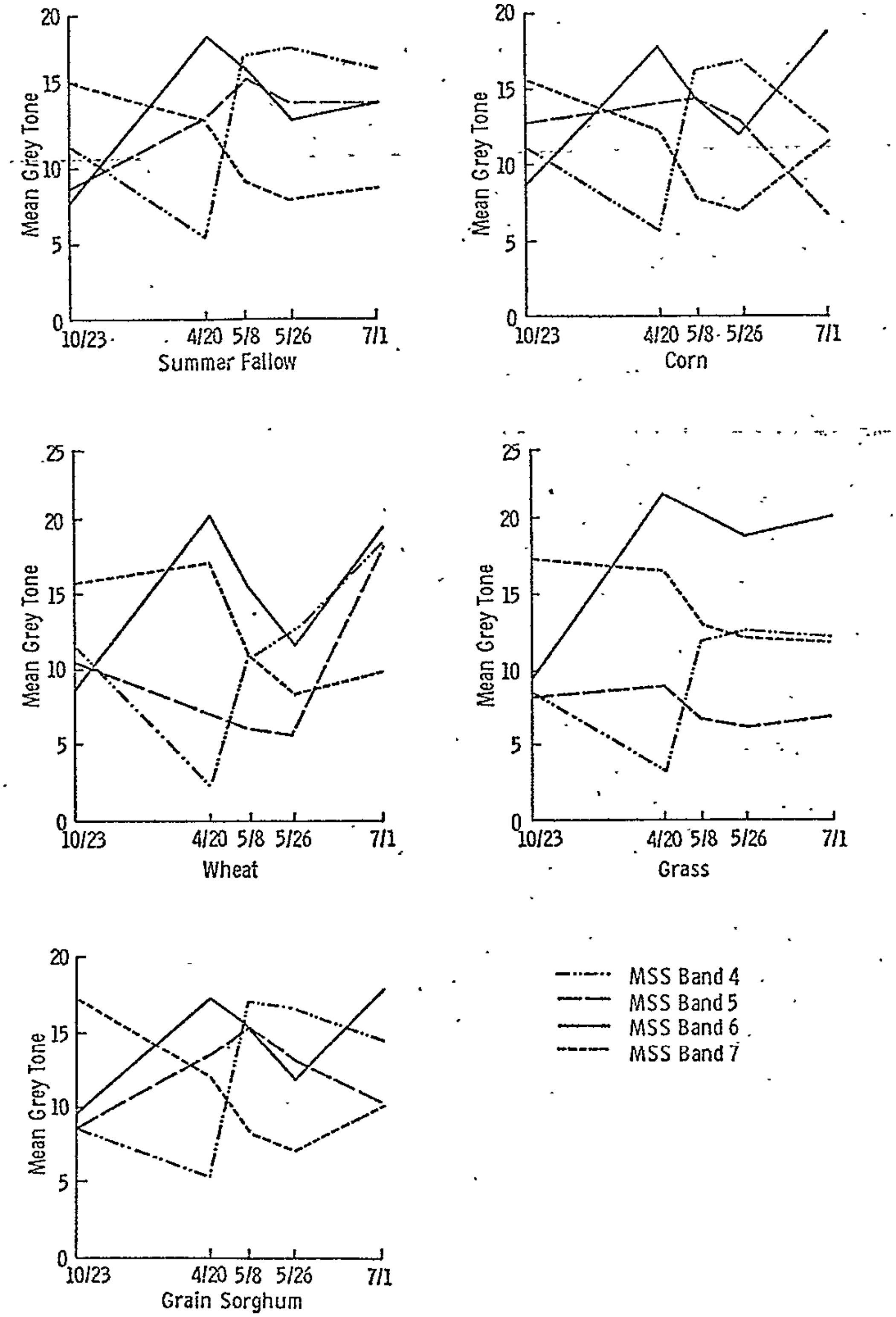


Figure 1.1 Graphs of mean grey tone of ground truth data for Finney County.

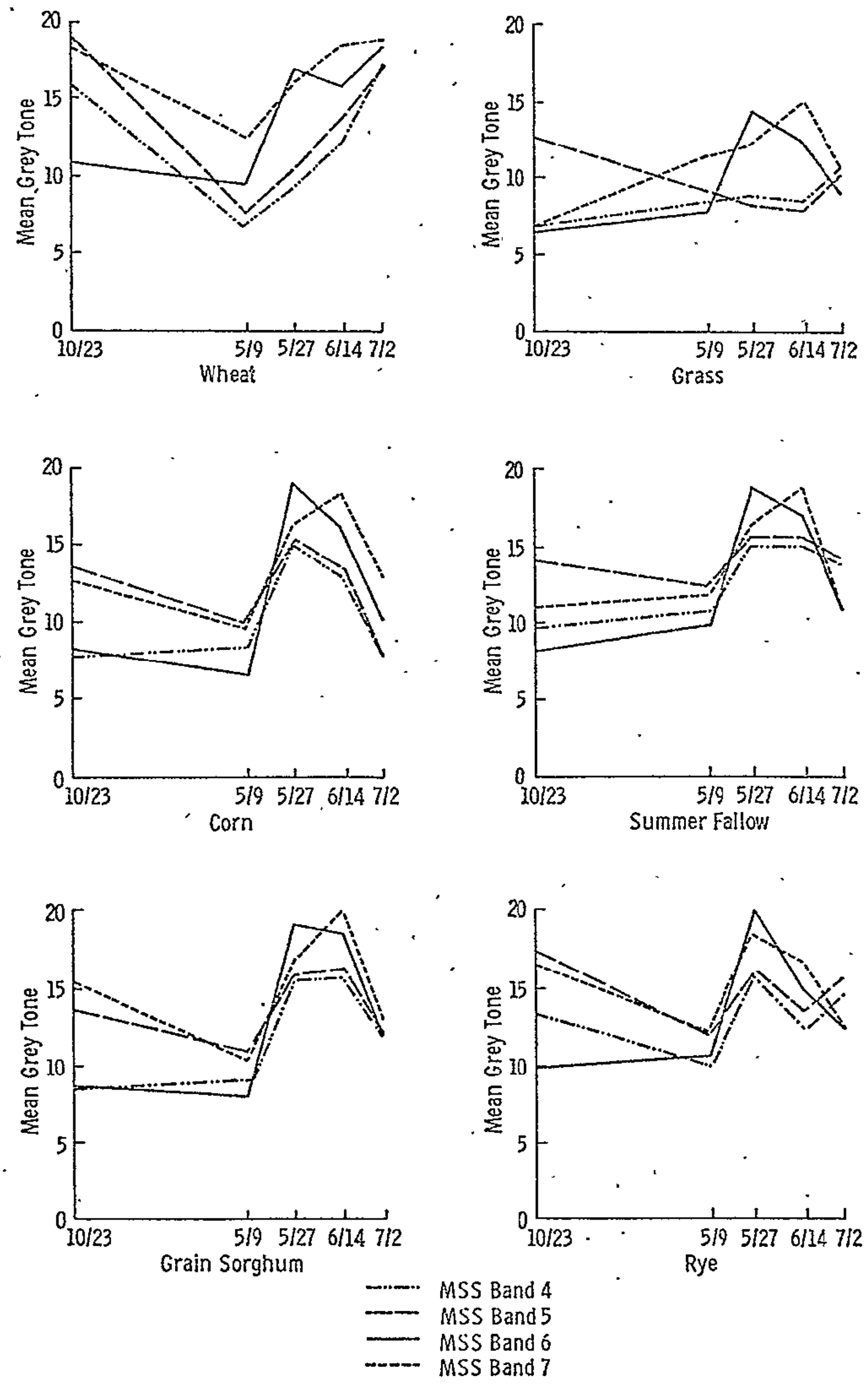


Figure 1.2 Graphs of mean grey tone of ground truth data for Morton County.

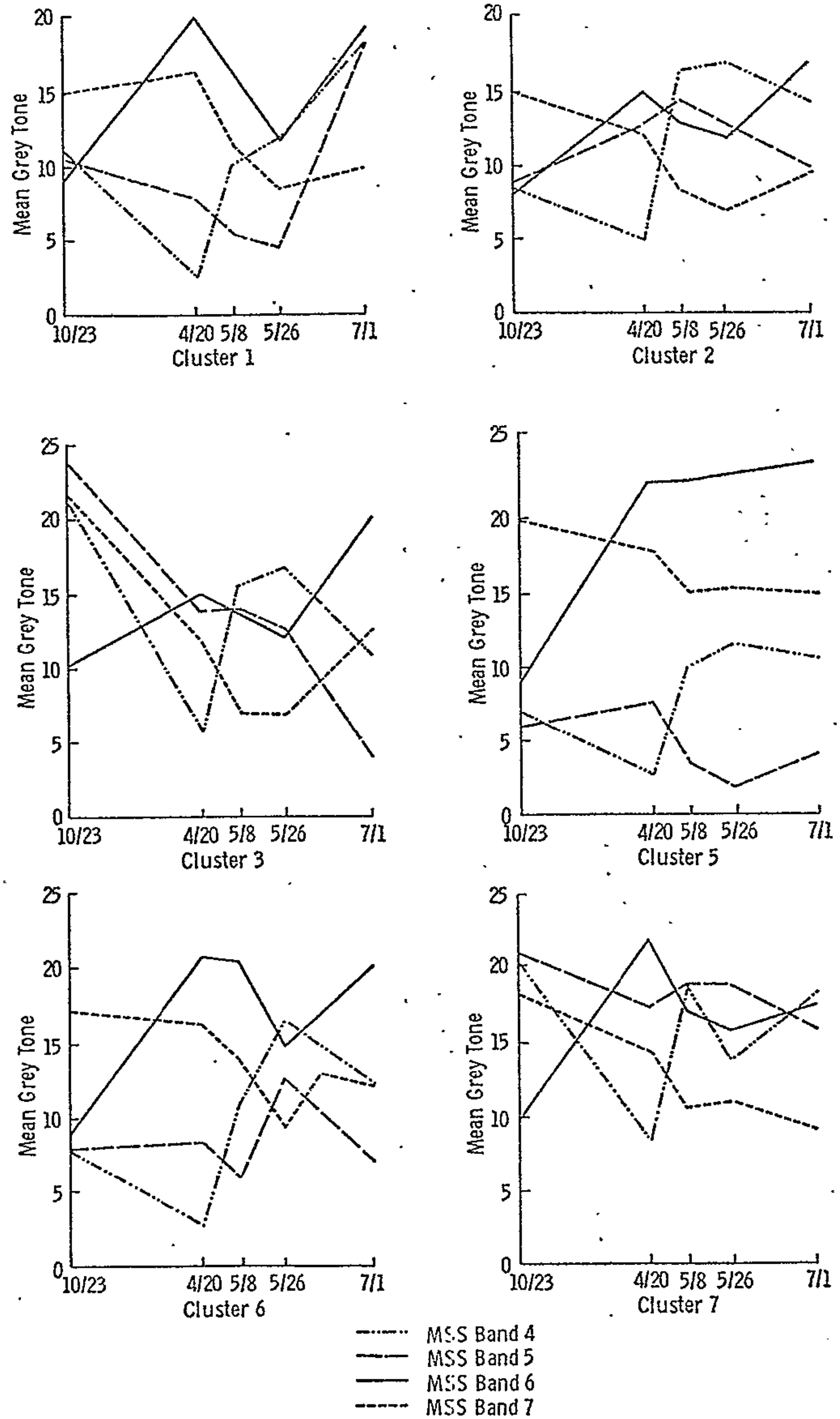


Figure 1.3 Graphs of mean grey tone of 6 spatial clusters for Finney County.

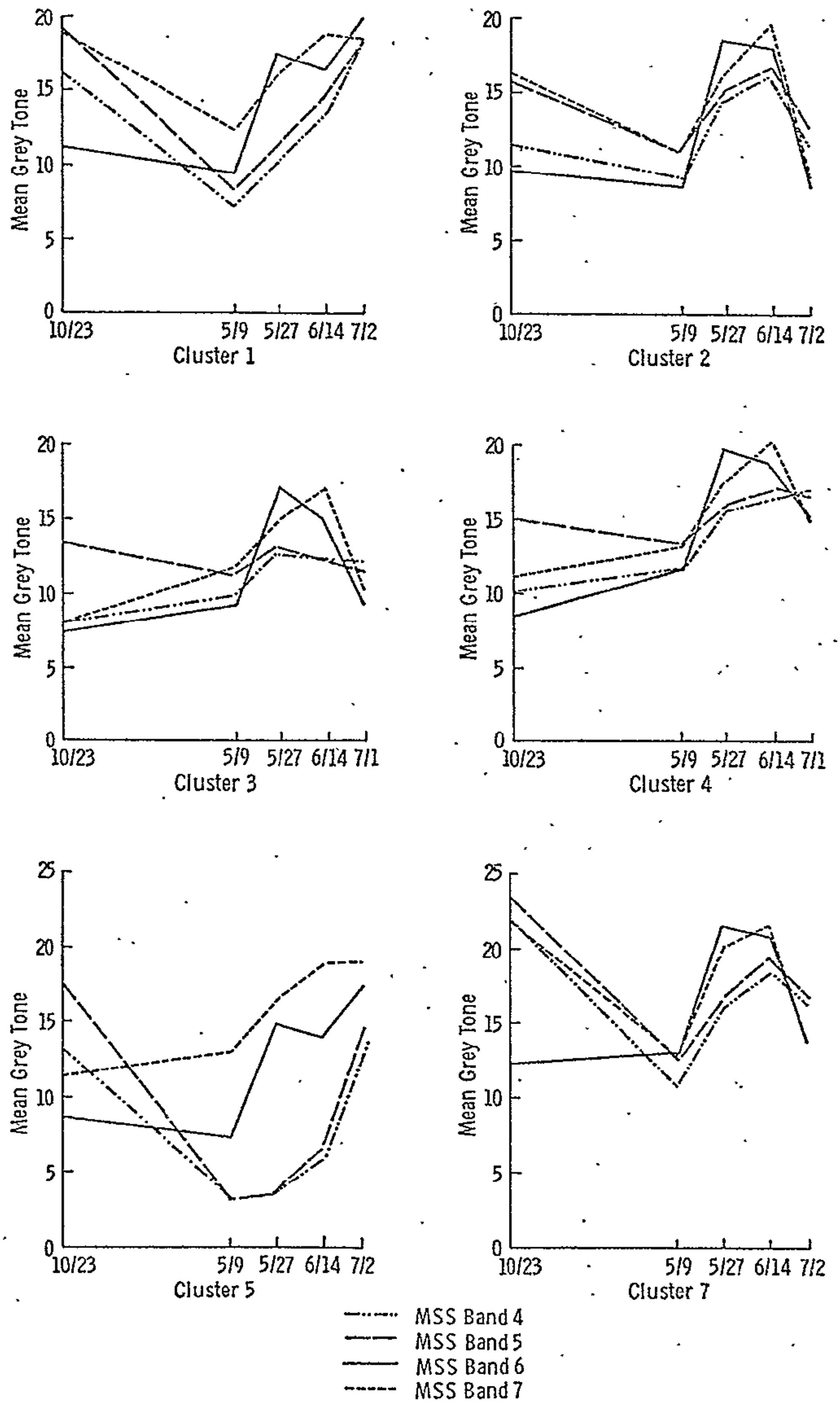


Figure 1.4 Graphs of mean grey tone of 6 spatial clusters for Morton County.

more difficult to identify the ground truth category. Cluster 7 from Figure 1.3 is probably corn, but one is not absolutely sure.

The clusters for Morton County are harder to identify. Clusters 1 and 5 in Figure 1.4 match with wheat in Morton County Figure 1.2. The remaining clusters could be assigned to any of corn, summer fallow, grain sorghum or rye.

These huge differences for data between counties and the generally large variance for the data for any one category prompted the idea that some ground truth labels were in fact wrong and/or that there was a large difference in the reflectance of crops due to different fields being in different growth stages. To test the idea that differences in growth state can increase the variance for a ground truth sample, the following experiment was performed. Each wheatfield in the ground truth was given a unique label. The mean and standard deviation for each field were computed and the mean plotted. The graphs of the mean gray tone for 6 fields seem to fall into 2 categories: fields 2, 3, and 10 are similar, as are 1, 18, and 20. Table 1 presents the standard deviation by band for all wheatfields and six selected wheatfields. In all cases, the standard deviation for the individual fields is less than (by factors of 3 to 10) the standard deviation computed over all fields of the same category. Difference in growth stage is a major contribution to the category confusion problem.

Throughout these experiments we have been concerned with the validity of the ground truth data. One method for testing the ground truth data is to compare the clustered regions with the ground truth. Figure 2.2.5 shows the original LACIE ground truth for Morton County. The clustered data is shown in Figure 2.2.2. Clustering was performed on all dates/bands 5 and 7 and using dates/bands: OCT23/5 OCT23/7 MAY9/5 MAY9/7 for gradient information. To see the confusion shown by the cluster, we can associate the cluster label with the ground truth label it covers. We can represent this association by the graph shown in Figure 1.8. It can easily be seen that almost every cluster covers more than one ground truth category. To see what causes this, we can look at the fields that produce each of the edges from a cluster and the edges to a ground truth category. For example, a field labelled wheat and cluster 1 has a mean gray tone temporal plot shown in Figure 1.9. Another wheatfield, labelled cluster 2, has the

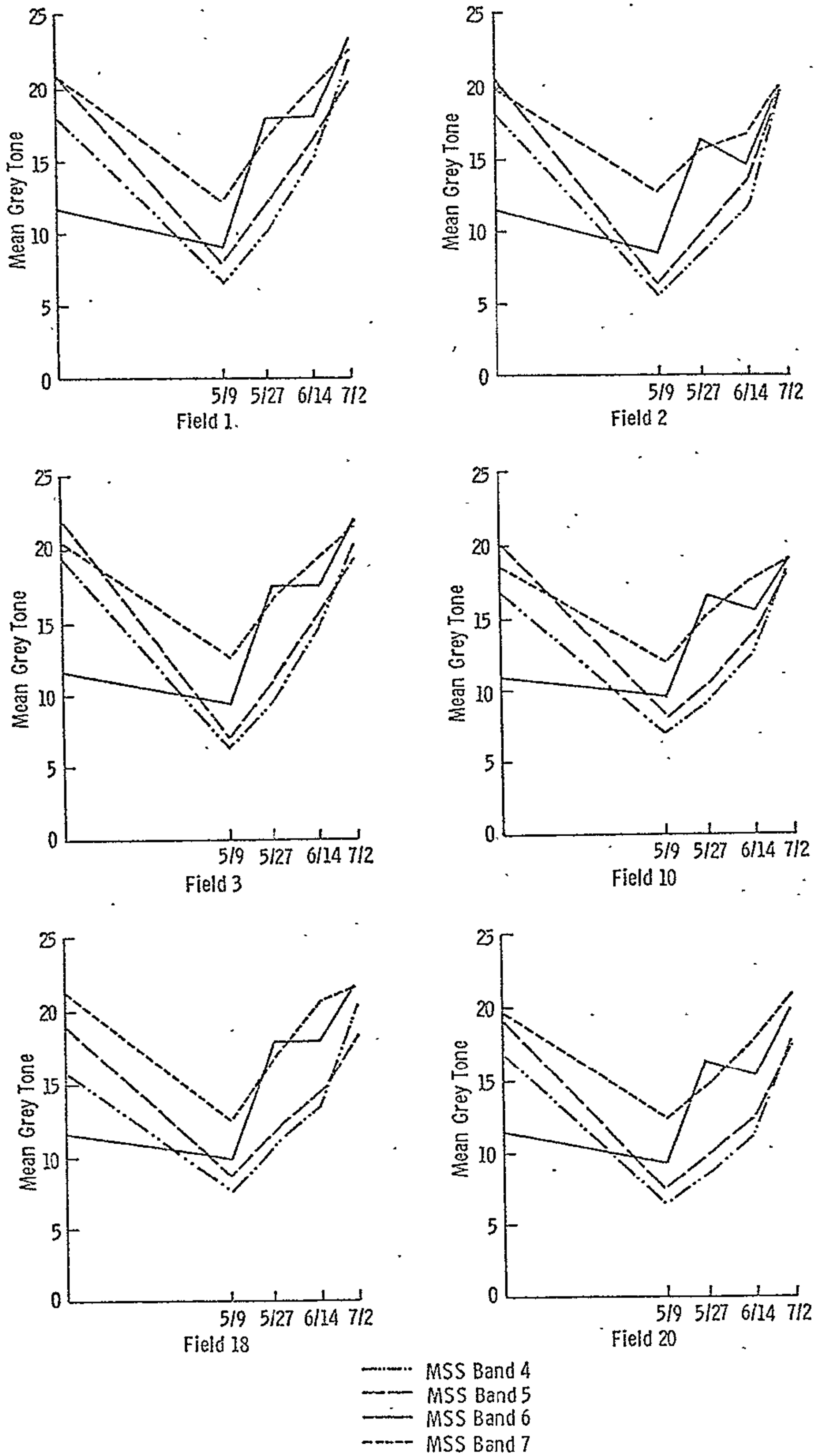
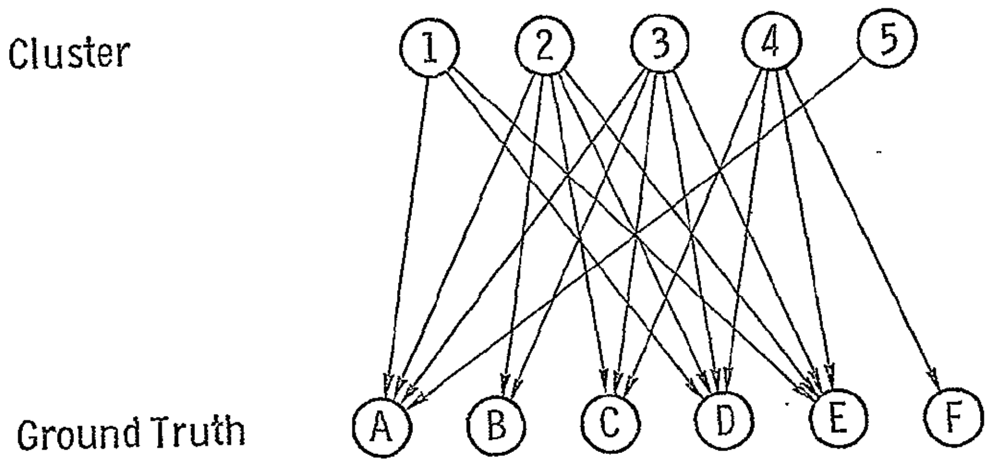


Figure 1.5 Graphs of mean grey tone of 6 wheat fields in Morton County.

STANDARD DEVIATION

BAND	ALL WHEATFIELDS	FIELD 1	FIELD 2	FIELD 3	FIELD 10	FIELD 18	FIELD 20
OCT23/ 4	81.1	11.6	0.0	20.3	25.2	25.0	4.4
OCT23/ 5	90.7	11.1	0.0	25.9	22.8	21.9	9.9
OCT23/ 6	112.0	24.1	10.1	29.8	37.1	37.5	17.4
OCT23/ 7	98.1	16.3	0.0	28.8	26.4	28.9	0.0
MAY 9/ 4	61.4	17.3	4.3	18.5	19.3	25.7	9.7
MAY 9/ 5	56.8	14.6	2.4	16.9	19.2	21.4	8.6
MAY 9/ 6	85.6	15.6	6.7	21.5	30.1	23.3	10.3
MAY 9/ 7	120.4	21.1	10.0	26.4	35.4	31.7	19.5
MAY27/ 4	63.1	17.6	4.5	20.5	19.8	19.9	5.9
MAY27/ 5	64.0	16.3	6.3	21.3	21.3	20.2	6.9
MAY27/ 6	105.8	25.1	0.0	26.0	32.3	24.9	4.9
MAY27/ 7	95.3	15.0	0.0	22.1	26.9	19.1	0.8
JUN14/ 4	70.0	17.3	0.0	19.2	22.7	23.0	11.8
JUN14/ 5	75.0	20.0	0.0	24.2	28.4	22.7	11.5
JUN14/ 6	93.3	16.7	0.0	21.1	31.0	30.8	11.3
JUN14/ 7	94.8	14.7	0.0	14.6	27.0	29.1	6.7
JUN 2/ 4	80.1	0.0	0.0	19.1	17.0	25.5	2.0
JUN 2/ 5	83.4	0.0	0.0	22.4	16.9	25.8	9.8
JUN 2/ 6	71.5	3.8	0.0	12.9	13.7	23.3	0.0
JUN 2/ 7	79.8	0.0	0.0	14.0	14.0	24.9	0.0

Table 1. Standard deviation of all wheatfields and six individual wheat fields in Morton County.



A - Wheat
B - Grass
C - Corn
D - Summer Fallow
E - Grain Sorghum
F - Rye

Figure 1.8 Graph showing association of 5 clusters with ground truth categories for Morton County.

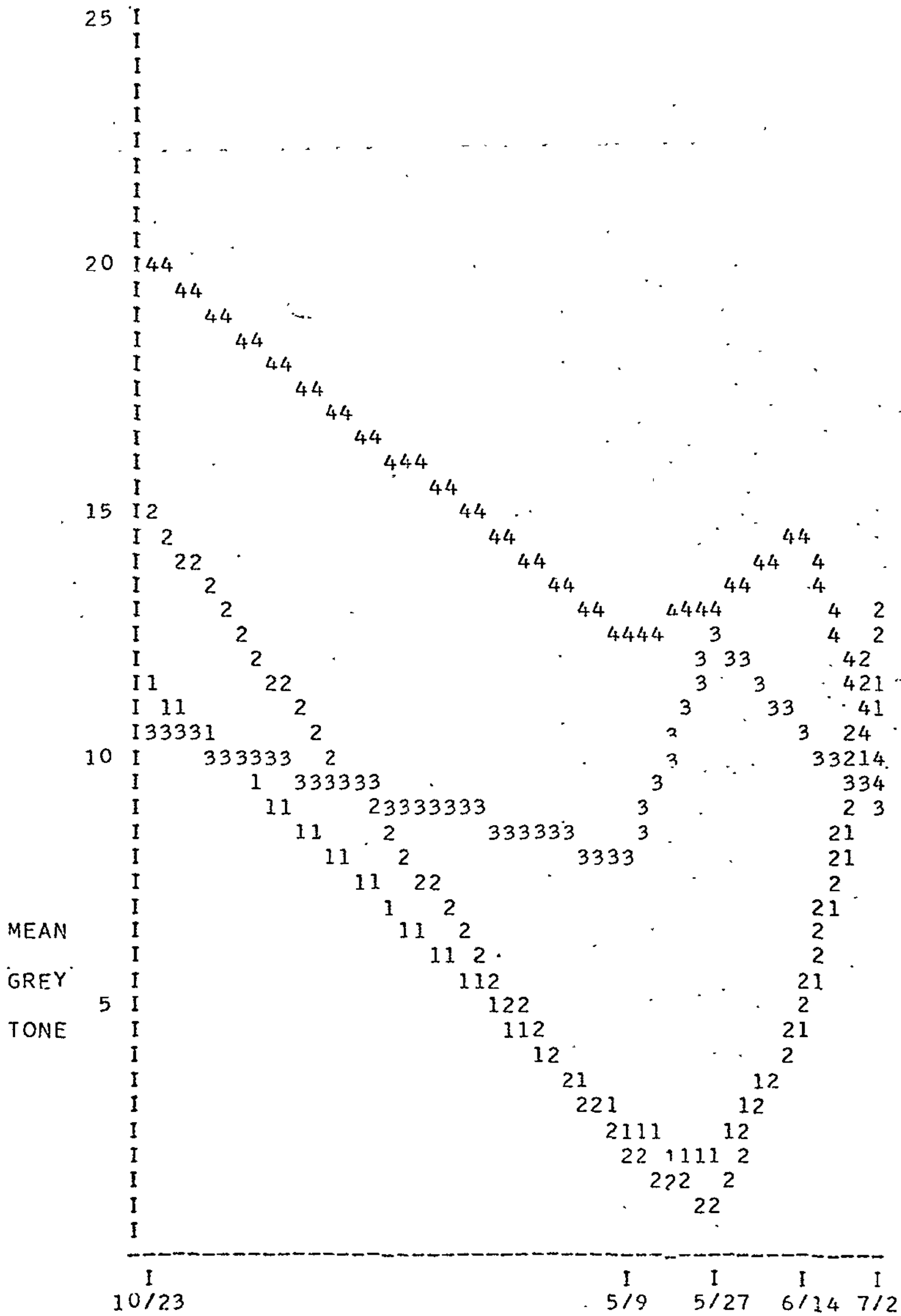


Figure 1.9 Mean grey tone for wheat and cluster 1 in Morton County.

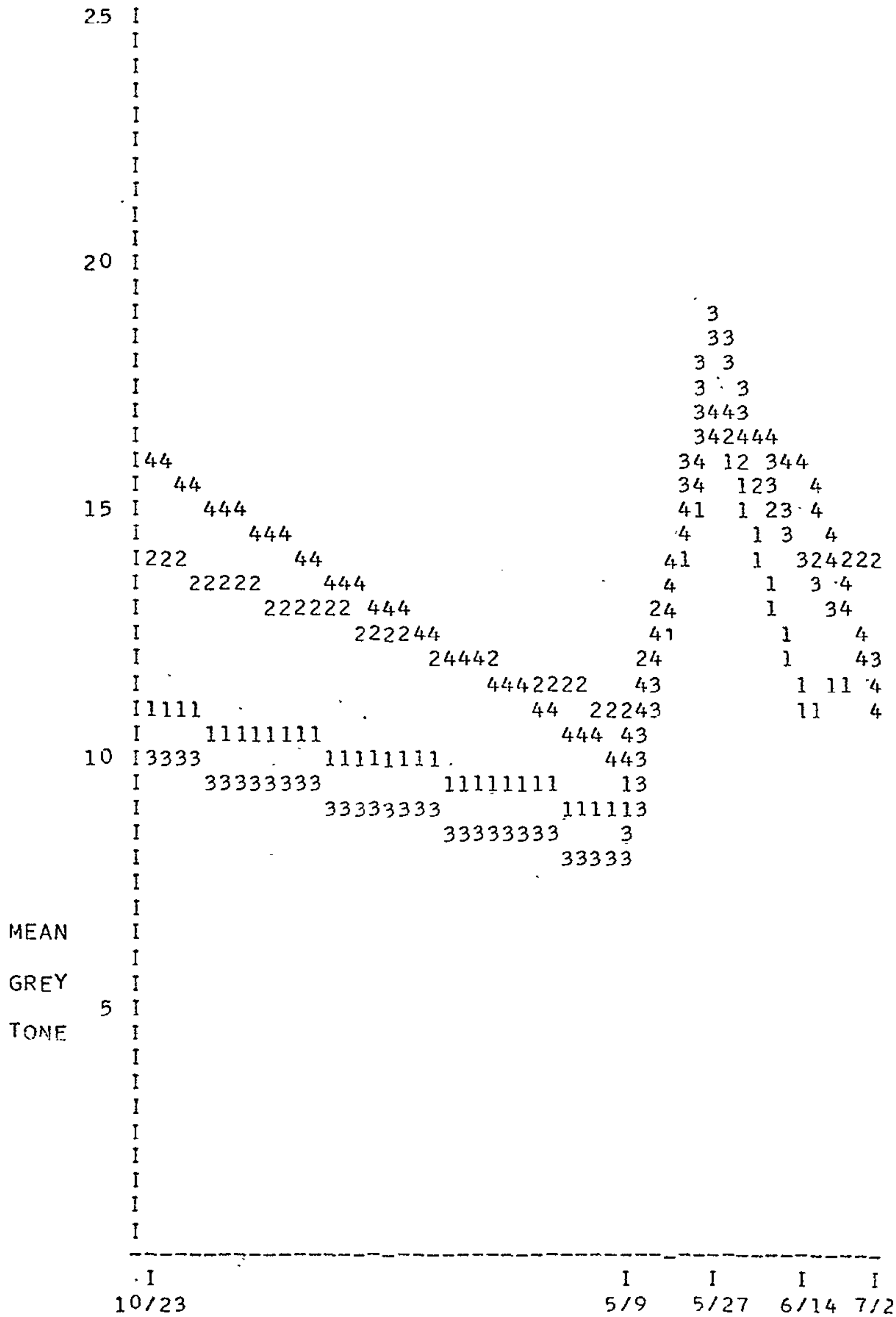


Figure 1.10 Mean grey tone for wheat and cluster 2
in Morton County.

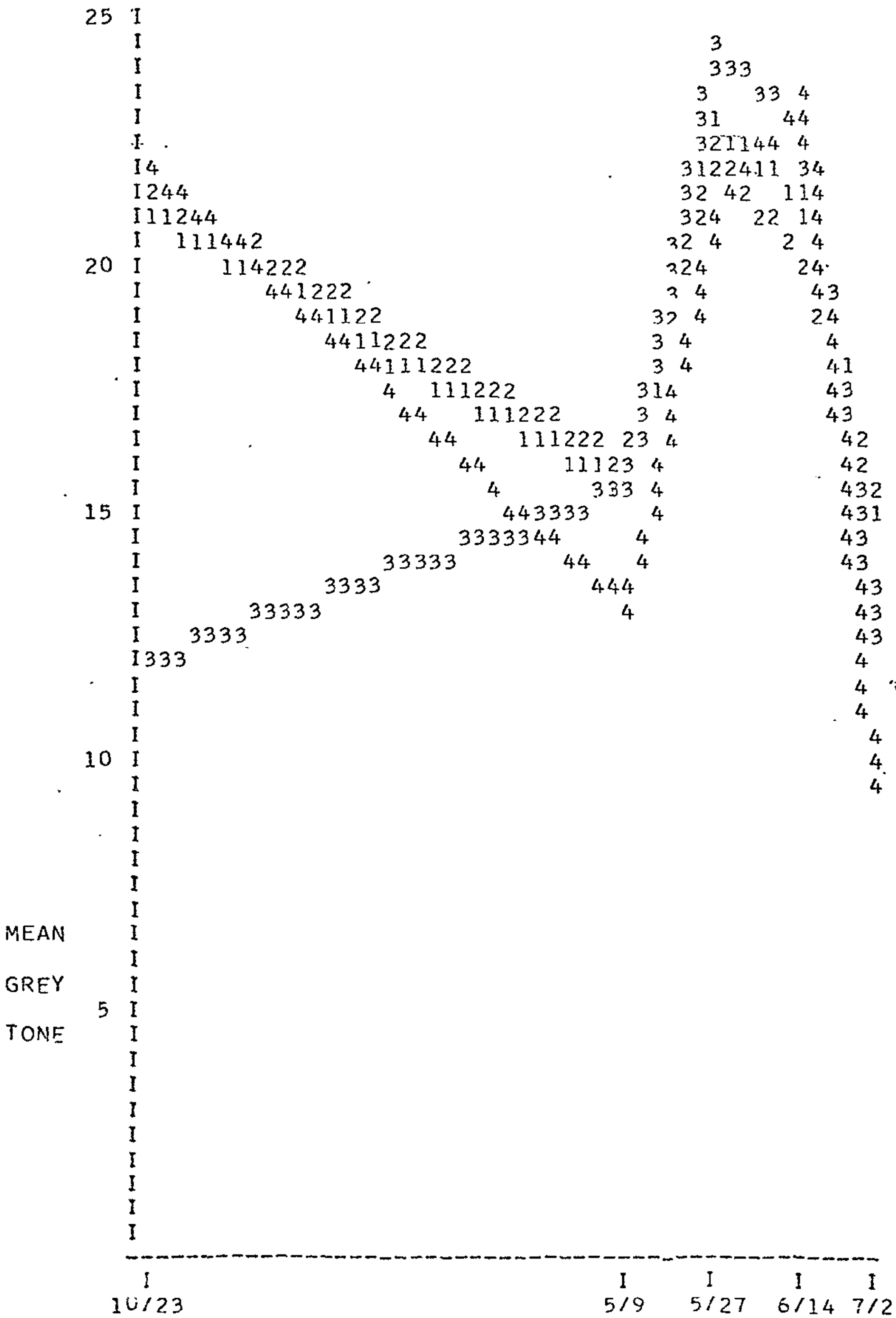


Figure 1.11 Mean grey tone for wheat and cluster 1 in Morton County.

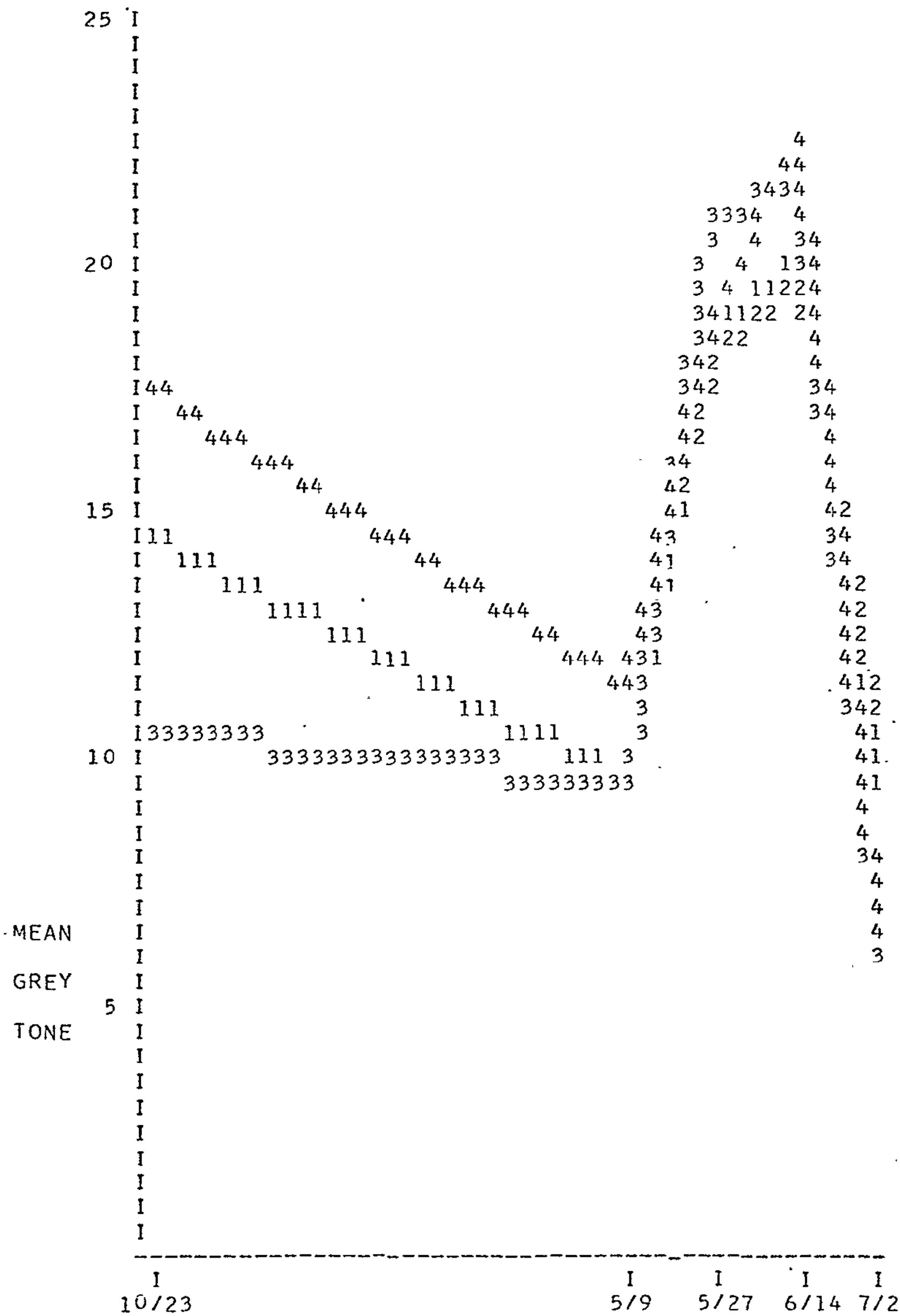


Figure 1.12 Mean grey tone for summer fallow and cluster 2 in Morton County.

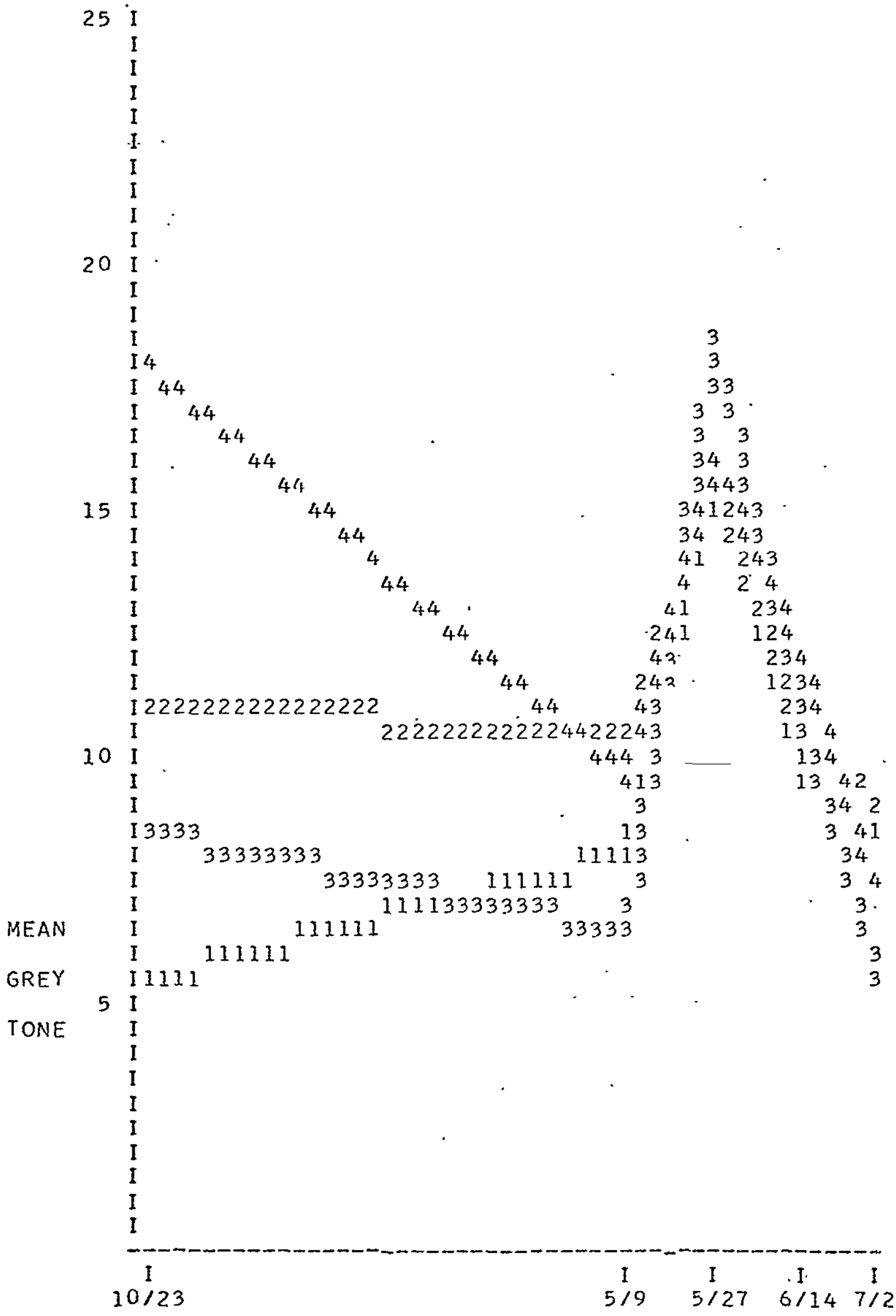


Figure 1.13 Mean grey tone for corn and cluster 2
in Morton County.

plot shown in Figure 1.10, and a third wheatfield with cluster label 1 has the plot in Figure 1.11. These three graphs are marked by different characters. This difference can be explained by either the ground truth being incorrect or the fields being at different points in the "normal" growth pattern when the imagery was acquired. If we assume that we are seeing different points on a "normal" growth curve, then this data can be used to construct such a growth curve. On the other hand, if it seems that the ground truth is incorrect, we can correct the ground truth and perhaps obtain better discrimination results.

We can also fix our attention on a single cluster label. Figures 1.10, 1.12, 1.13 and from fields labelled cluster 2 and are ground truth categories wheat, summer fallow, and corn, respectively. These plots, from the same cluster, are similar. The plots for a given cluster label are not always so similar and in some cases vary as much as those shown above for wheat. This indicates that the clustering process has perhaps gone too far in its grouping together.

2.0 Unsupervised Clustering

The last report included a qualitative description of the unsupervised clustering approach. It was mentioned that because of the high error rate in the classification of some crops, the ground truth information on some fields was suspect. In the last quarter, an attempt was made to overcome this problem, using a couple of techniques. Unsupervised clustering was one of them. In this section, we will describe some of the results obtained by this process.

The procedure, strictly speaking, is not completely unsupervised. There are a few parameters (described below) which the user enters. These serve to control the degree or level of clustering. In that sense, the user does supervise the process. However, when compared to the Bayes or discriminant analysis classification, the meaning of unsupervised becomes clearer. The latter two procedures require a prior knowledge of some ground truth data. That is, in order to do any classification, they need some training sites to work on. The classification is supervised by these training sites, along with some control parameters specified by the user.

Unsupervised clustering, on the other hand, groups like things together, without any regard to what they are. This makes the unsupervised clustering very easy to perform, as all that is needed is the original multiband image. Only at the end of the processing do we need any ground truth information. This is not for training sites, but for identification and measuring classification accuracy. Thus, if parts of an image are identified to correspond to certain crops or land use, the clustered image can be used to check the ground truth data that seemed faulty.

In order to understand some of the results of this procedure, the various steps of the process are described below. Basically, it is a two-part procedure. In the first part, usually referred to as spatial clustering, an image is generated, which makes and delineates the different homogenous areas of the input image. The process uses the spatial information surrounding a cell to determine these areas. In the second part, or measurement space clustering, these homogenous areas are grouped together according to their spectral signatures. Let us examine the operations involved in these two stages.

Usually before any spatial clustering, some preprocessing of the image is done. This involves quantization of the image, followed by contrast enhancing. The preprocessing options are performed to increase the effect of the first spatial clustering operation--generating a gradient image. The image gradient operation is defined on a multiband image and serves to distinguish between the spectrally homogenous and boundary areas in the image. The preprocessing options tend to make the edges between regions sharp.

It is also important here that the bands of the original image be registered correctly. The better the registration, the sharper the boundaries on the gradient image, and the better the definition of the homogenous regions. Unfortunately, the registration between bands from different dates was not very good. Thus, in the first run-through, the results from this operation were poor. The problem was partially overcome by not using all the bands, but only some selected ones, either the ones showing the least misregistration between dates, or bands from one date only.

The gradient operation, like its calculas counterpart, generates an image in which boundaries or points of rapid change show up as high values. The interior of the fields in which there is little change in the gray tones, show up as low gradient values. Thus, by thresholding the image, it is possible to separate boundary cells.

There are various kinds of gradient operations available. We have used a Roberts gradient throughout.

The thresholding of the image is done using a fraction of its running mean as the cutoff. The running mean of a row is defined as the mean of 10 rows above and below it. By changing the fraction, one can raise or lower the cutoff, and control the degree of thresholding. For most cases, this fraction was set to 1.0. Thus the running mean of the row became the cutoff. There are other methods available for thresholding. This scheme was used because it is fast and works well enough on images of the sizes involved.

After thresholding, the cells which fall below the cut-off are considered as homogenous cells. This resulting image is usually very noisy, as it is hard to get very sharp gradient images, despite all the preprocessing. A "cleaning" operation is then carried out. It removes isolated cells and the "salt and pepper" effect which constitutes the noise. The result is an image with the different homogenous areas separated out and the boundary cells set to zero.

These areas can be expected to correspond to fields in an image. However, in many cases, a field may come out subdivided into more than one homogenous region. This happens if the process does not find the field spectrally homogenous. For example, if half the field has been watered, the wet and the dry areas may come out separated.

In the next step, for purposes of identification and clustering, each separate homogenous area is marked with a unique label. The last image usually constitutes the end of the spatial clustering section. However, if it is felt that some of the regions were not separated out sufficiently, a "splitting" operation is available. This operation takes a region marked homogenous and splits it into two or more sections, if it determined the region was not homogenous enough. In this process, the operation examines both the labelled file and the original multiband image file.

The idea behind separating each homogenous region is that we want to cluster only on the interior of the fields. The signature from the core of a field is more uniform, and a better representation of the field, than a signature obtained by including boundary or edge cells. The latter reflect a mixture of different classes and is prone to give rise to error in the classification.

The first step in the second stage is to generate the spectral signatures for the homogenous regions. For this we go back to the multiband image. The signature for a region is defined as the average gray tone for that region, for each of the bands of the original image. Thus, for a multiband image we have a spectral signature vector for each region. In addition, an area count is generated for each region. This is used to weight the signatures when regions or clusters are grouped together.

The measurement space clustering is a pure clustering procedure. While it is not limited to image clustering, it has been implemented to cluster spectral signatures. The procedure is an iterative one. The distance used for measuring closeness between clusters is the Euclidean distance defined on the measurement space dimensions. Within an iteration, the procedure finds the closest neighbor for each signature. These pairs of closest neighbors are linked together to form groups. For this, the user enters two parameters to insure that closest neighbors, which may be too far apart, are not clustered together. After clustering, a new table is generated, which consists of the signatures of the new clusters obtained in the last iteration.

The iterations are repeated until the number of clusters is reduced to what is felt adequate. For example, for the Rice County image, 6 clustering iterations were carried out. In these, 409 homogenous regions were reduced to 37 clusters.

Usually the number of clusters is left at a number larger than the number of categories expected. It is done to prevent overclustering. In almost all images, there are a few isolated regions with outlying signatures. The way the process is set up, these come out as isolated clusters. They don't normally define a well-known land use class and can be ignored. The discarding of clusters is made after a visual examination or a check on the relative size of the clusters. If one did not allow for these isolated cases to occur in the final number of clusters, it might cause different classes to merge. Forcing a smaller number of clusters may put together different classes whose signatures are closer than those for the outlying bunch.

The final step in the classification process is another spatial operation. In this, we grow the clusters into the area which had been marked as boundary. The operation assigns each boundary cell to the cluster it is closest to. This way we include all the edge cells which were stripped off during the initial phase of the spatial clustering. The last process is analogous to the generalization a cartographer makes when creating a land use classification map.

One other fact remains to be mentioned before discussing each county image. The spatial sample is the area or the number of resolution cells that the image covers. By increasing the sample size by a factor of four, doubling the image vertically and horizontally, the results of the spatial clustering are considerably enhanced. This is because each individual field has more cells, which gives it a better definition. The increase in sample size is done by expanding the image vertically and horizontally by a factor of two. While the results are better, the price paid for this is in terms of processing time and memory space. However, for images of this small size, this increase was not significant.

For the figures included in this report, the images were compressed down and scaled. This was done to accomodate the size of an ERTS cell as well as the uneven printer cell ratio. The compression for display purposes is done by omitting selected lines or columns. The selection is determined by the compression ratio. All this may result in the omission or drastic reduction of some small regions on the printout, even though they were included in the processing. The figures are included here for qualitative purposes generally. The character and shape of all the major fields is maintained. Any quantitative assessment is, of course, done at the expanded scale.

2.1 Rice County Image

In the last report, the processing of the Rice County image was done on the principal component image. This last quarter, the processing was repeated on the original ERTS images. The Rice County site consisted of four images taken on October 21, 1973, April 18, 1974, June 12, 1974, and July 18, 1974. Thus, there were sixteen bands of information. All these bands were expanded vertically and horizontally to increase the sample size. For the spatial clustering, MSS bands 5 & 7 of the April and June dates were selected. Contrast enhancing was performed on each of these bands. The spatial clustering procedure was then carried out. It resulted in

386 homogenous regions which are depicted in Figure 2.1.1.

As mentioned above, some of the smaller regions may not be seen because of the compression. Also, though each region has a unique label, they are shown here with repeating labels. This is because the printer only had 44 symbols available. Without resorting to overprinting or color, it is not possible to display each region uniquely. The figure, however, does illustrate the different regions of the image. The blank areas between the fields constitute the boundary section which separates the regions.

It was felt that the spatial clustering was not strong enough, and some fields came out connected, when they should not have been. A splitting operation was then performed. In this, the different parts of the homogenous regions of Figure 2.1.1, were examined to see if their means differed enough to separate them. This checking was done using the four MSS bands 5 & 7 of the April and June dates. Under the threshold given, the 286 regions were split into 409 regions. These can be seen in Figure 2.1.2.

The 409 regions were clustered on the basis of their spectral signatures defined by all 16 bands. The series of reduction of the number of clusters was 409-202-139-106-64-47-37, in 6 iterations. In order to see some pattern in the clustering, the results of the last four iterations are shown in Figures 2.1.3 to 2.1.6. These figures correspond to 106, 64, 47 and 37 clusters respectively.

The result of spatially generalizing Figure 2.1.6 is shown in Figure 2.1.7. Here the boundary or unclassified cells are assigned category labels using the nearest neighbor rule.

No quantitative analysis could be performed on the classification results for Rice County, as was done for the others. It shall be done this coming quarter. For completeness sake, we have included some soil and ground truth maps for this test site. Figure 2.1.8 shows the crop ground truth map. This will be used for quantitative verification of the clustering. The four crop classes shown on this map are wheat, grain sorghum, corn and summer fallow. These are denoted by letters A through D respectively.

It should be noted that out of the 37 categories in Figure 2.1.6, 95% of the image consists of the six classes, A, C, G, J, K, and X. The

FILE NAME - RICCMPACT

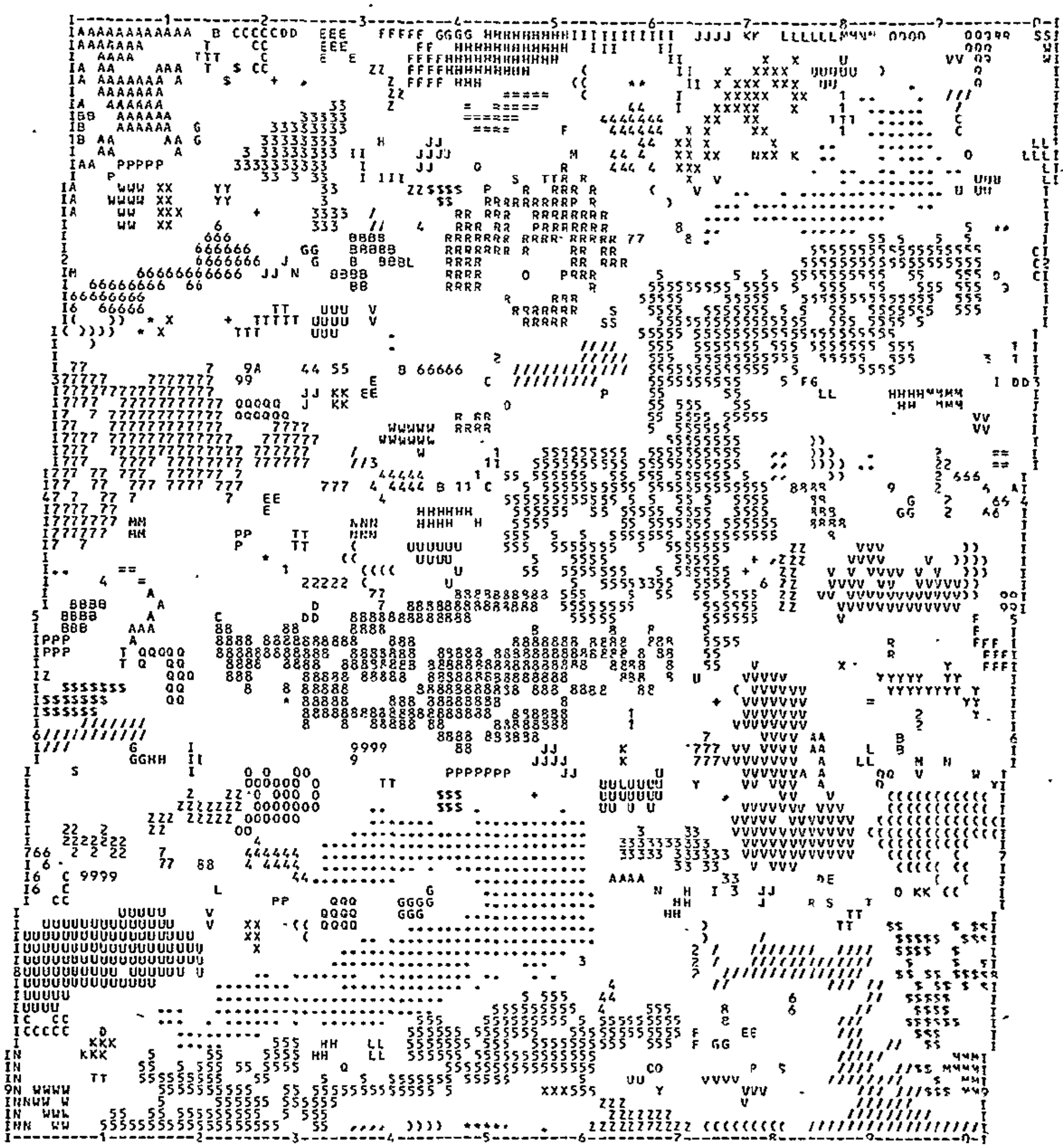


Figure 2.1.1 Rice County - Homogenous regions

FILE NAME - RICCRPSPL

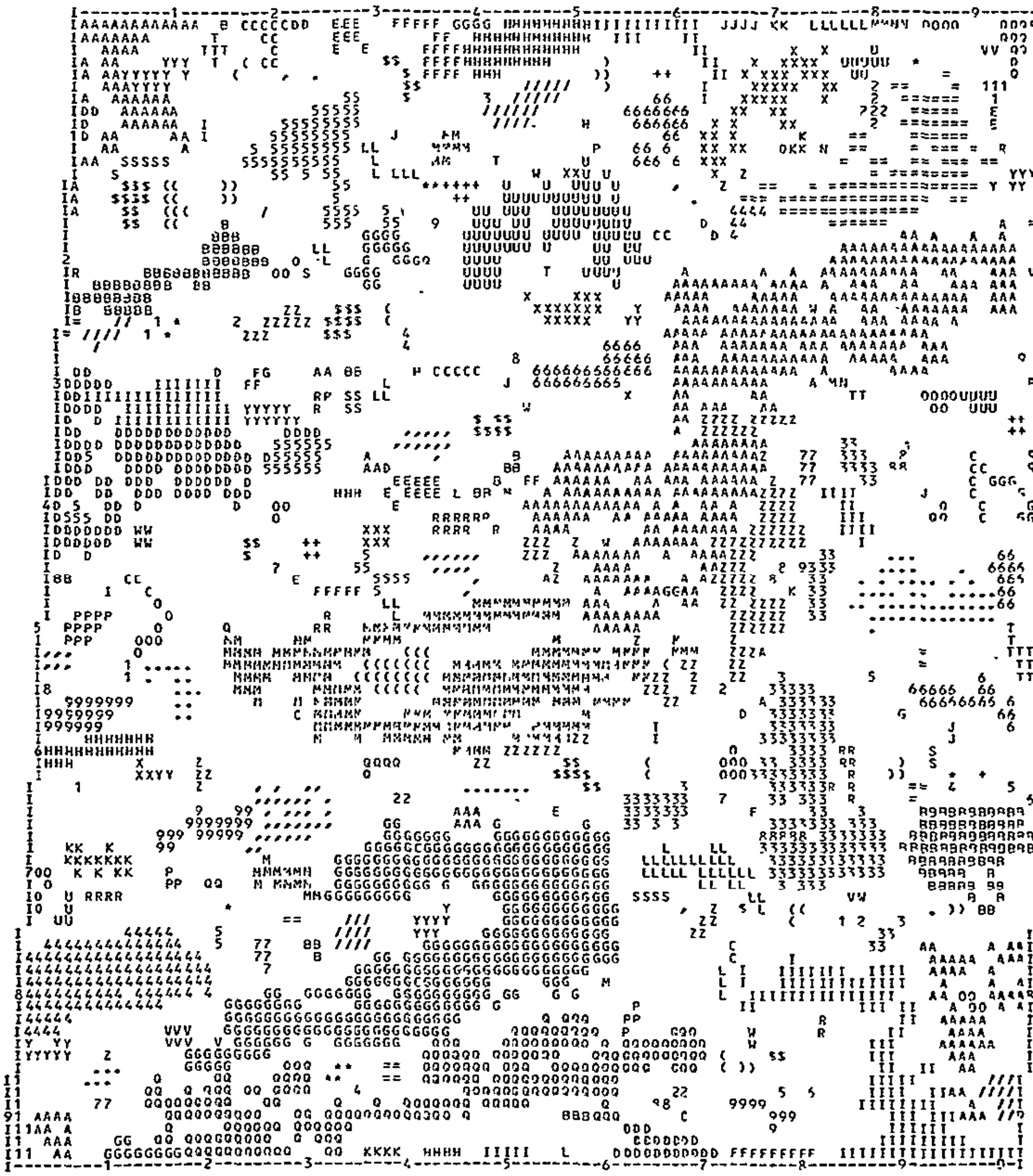


Figure 2.1.2 Rice County - "Split" homogenous regions

FILE NAME - RICCMPCL3

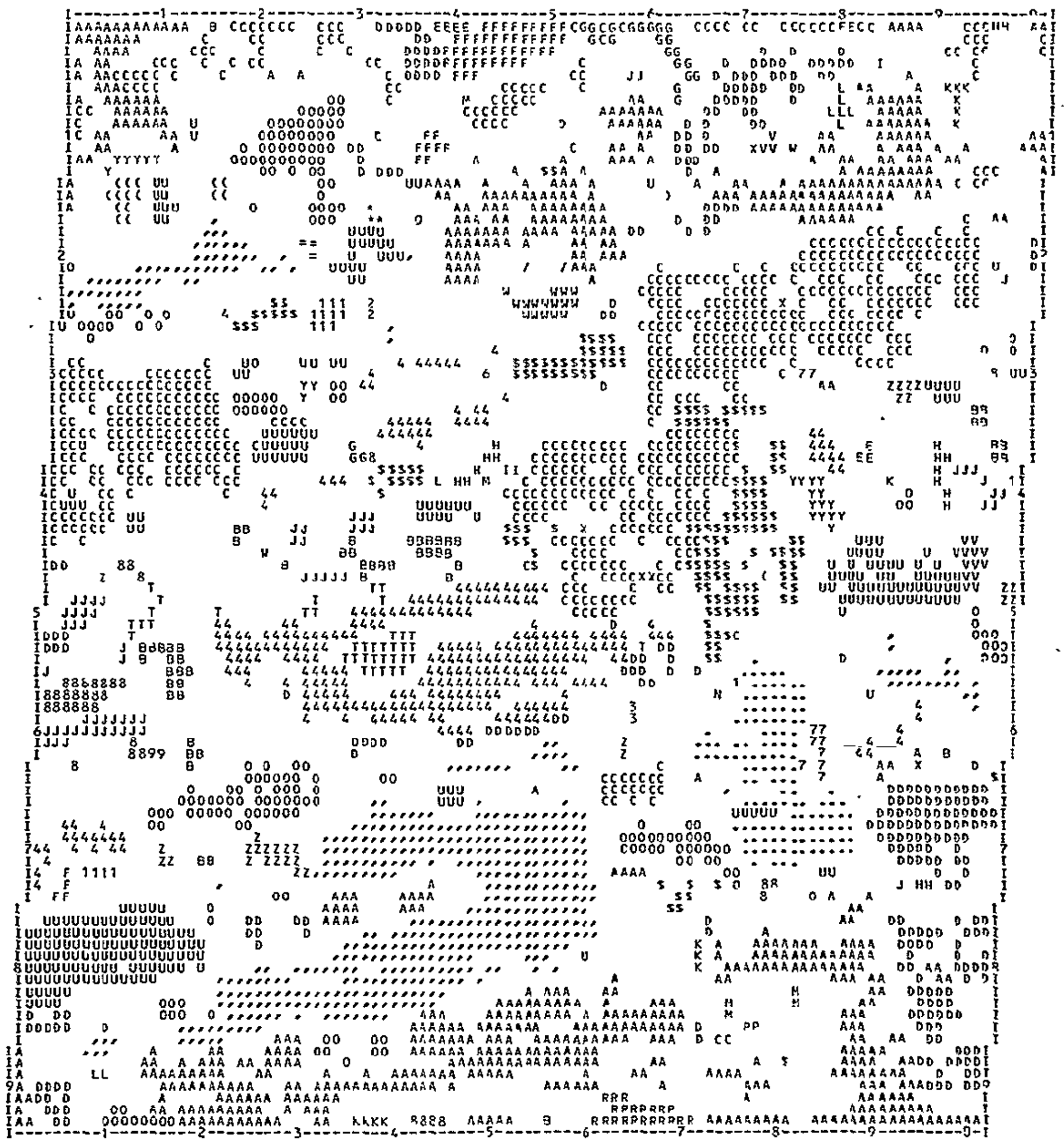


Figure 2.1.3 Rice County - Third Clustering iteration

106 regions

FILE NAME - RICCMPCL4

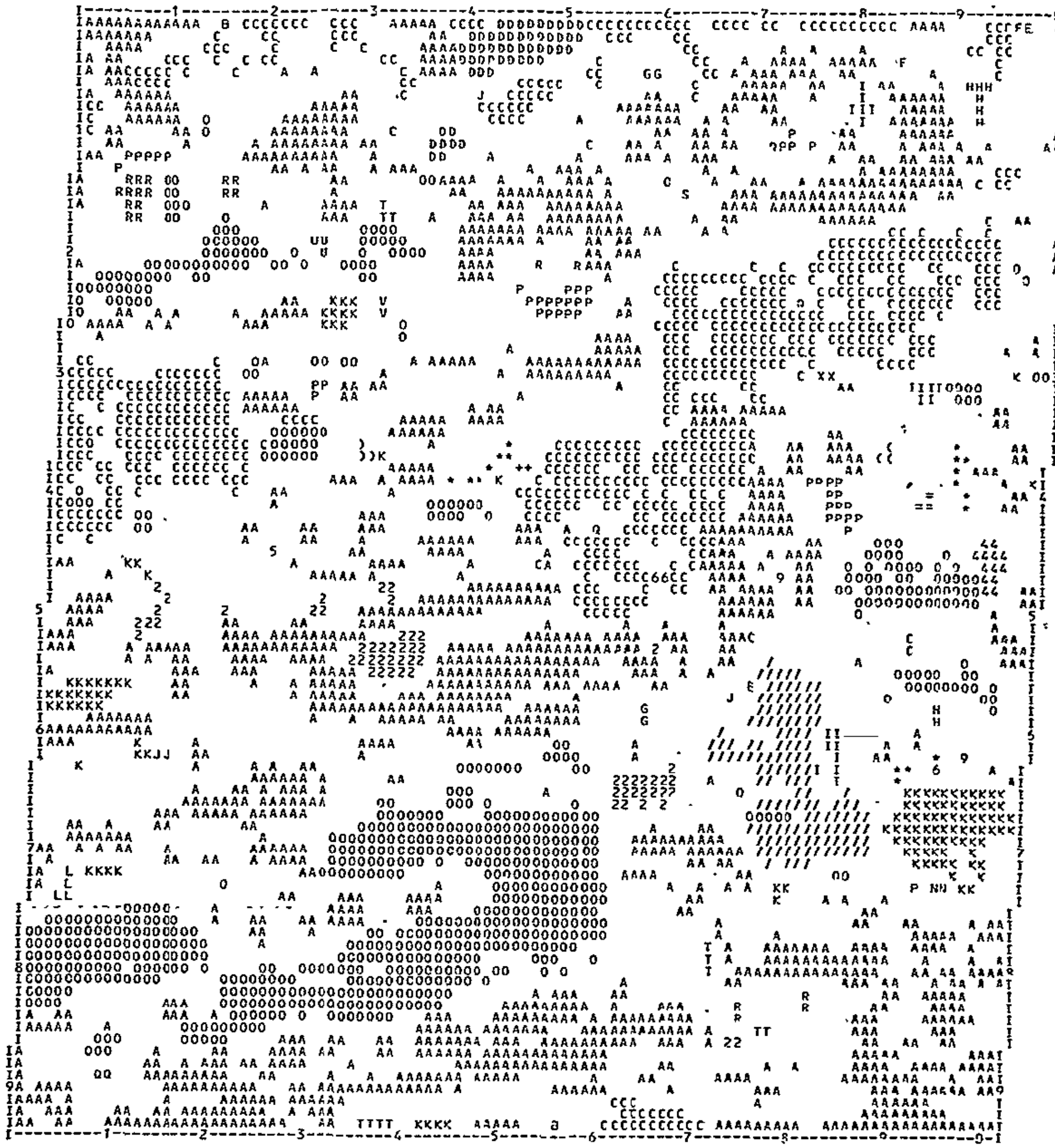


Figure 2.1.4 Rice County - Fourth Clustering iteration

64 regions

FILE NAME - RICCHPCL5

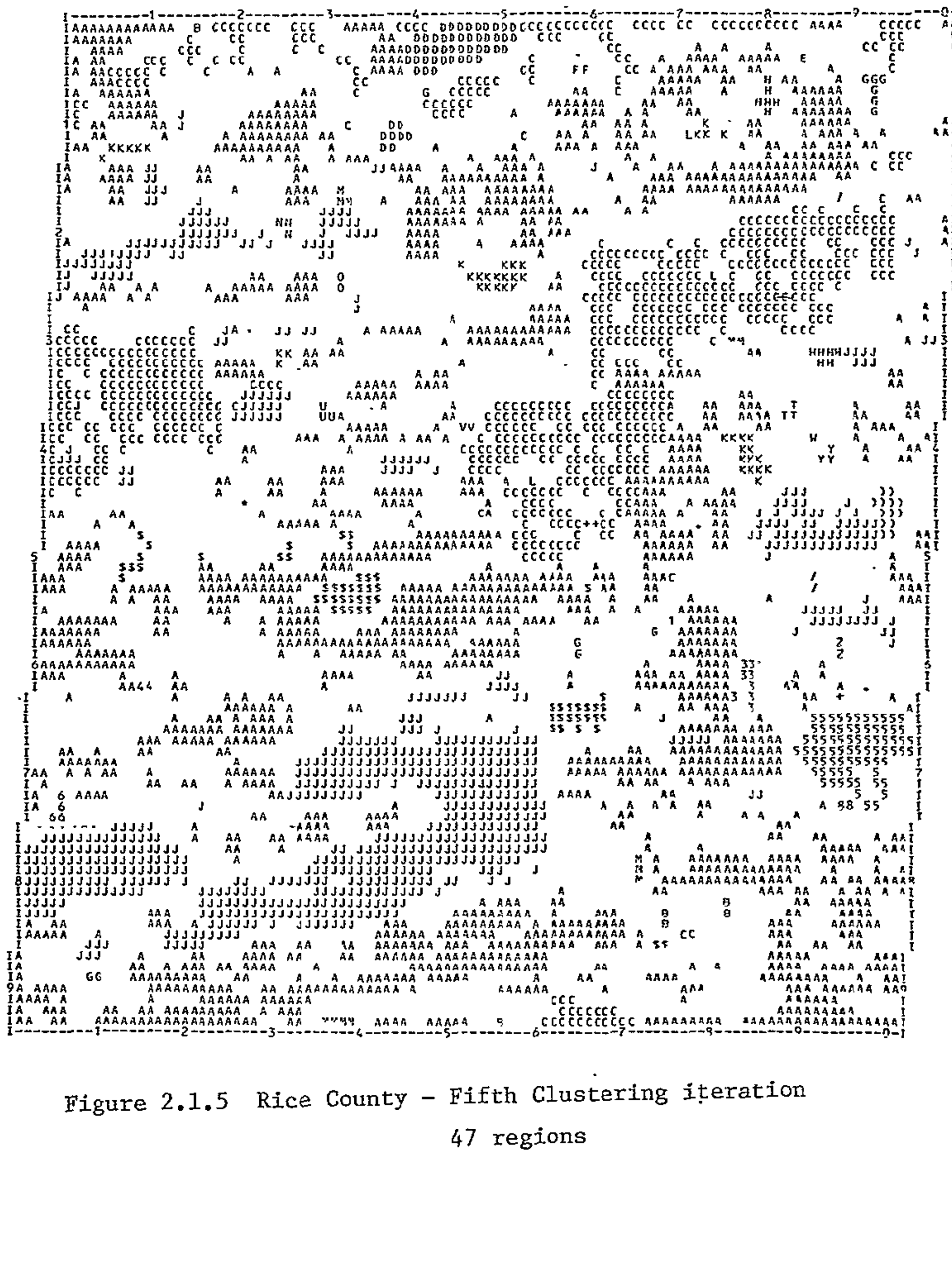


Figure 2.1.5 Rice County - Fifth Clustering iteration
47 regions

FILE NAME - RICCMPCL6

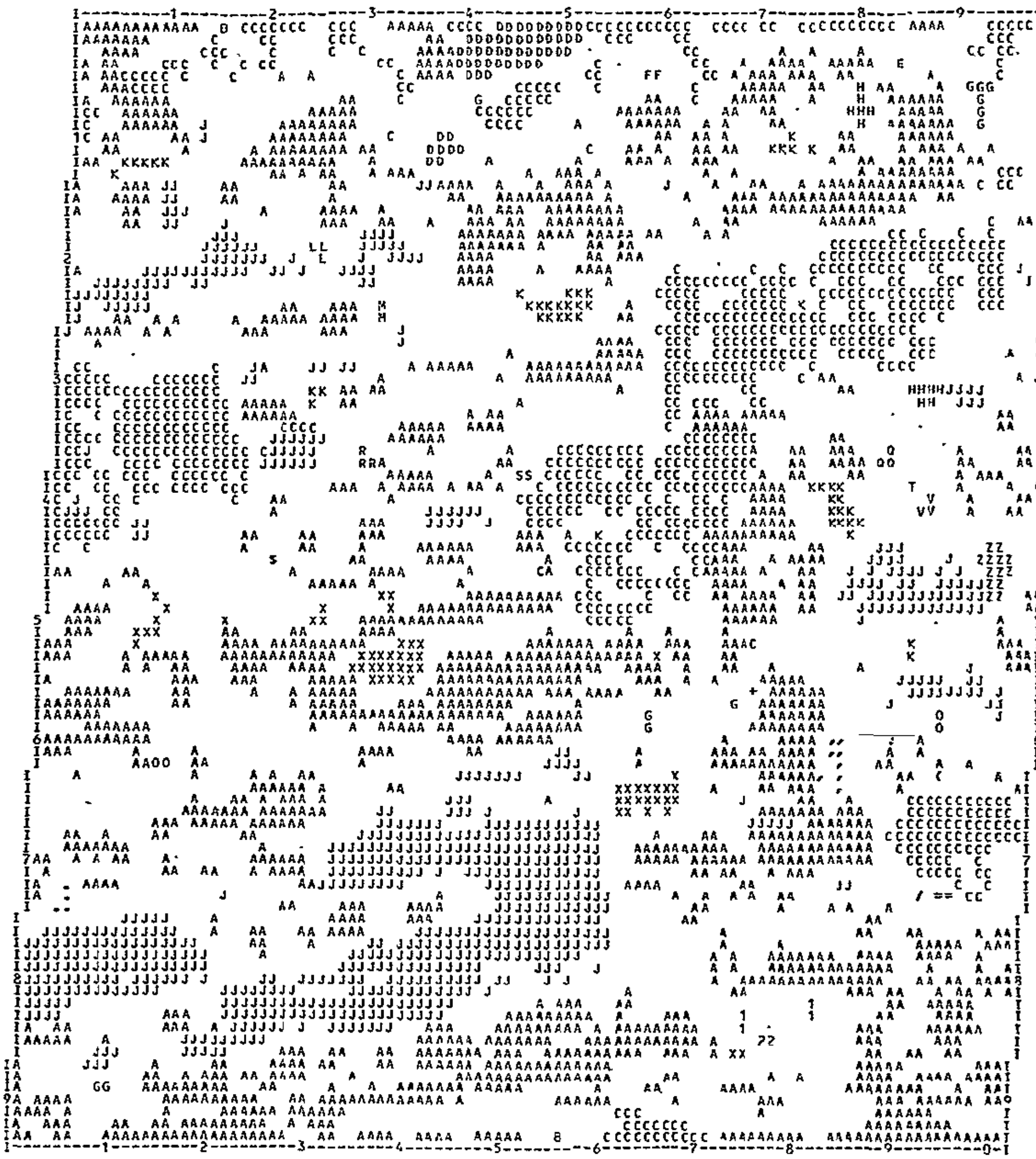


Figure 2.1.6 Rice County - Final Clustering

37 regions

FILE NAME - RICCMPFIL

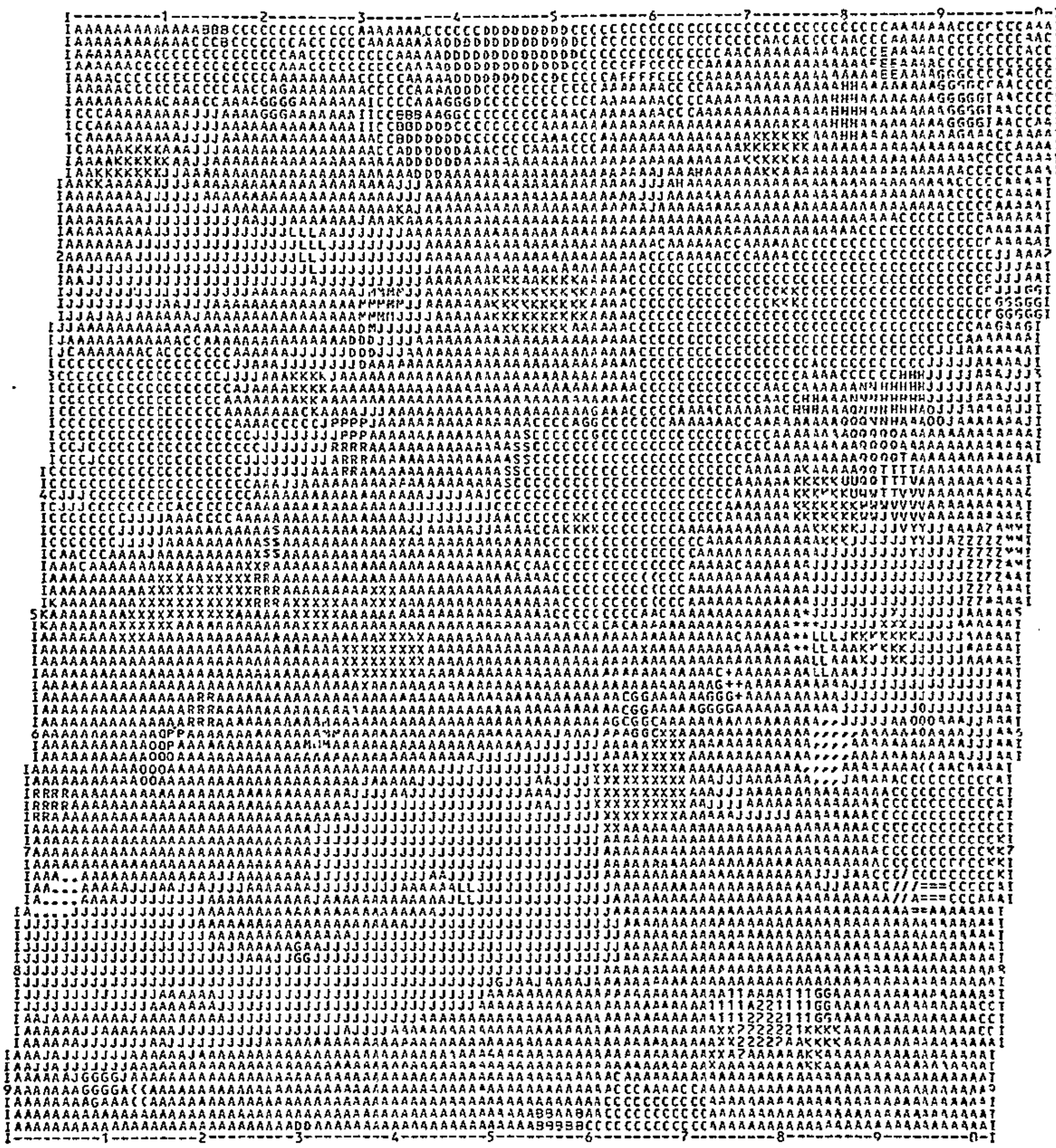


Figure 2.1.7 Rice County - Spatially generalised clustered image

FILE NAME - RICCMPSYM BAND 1

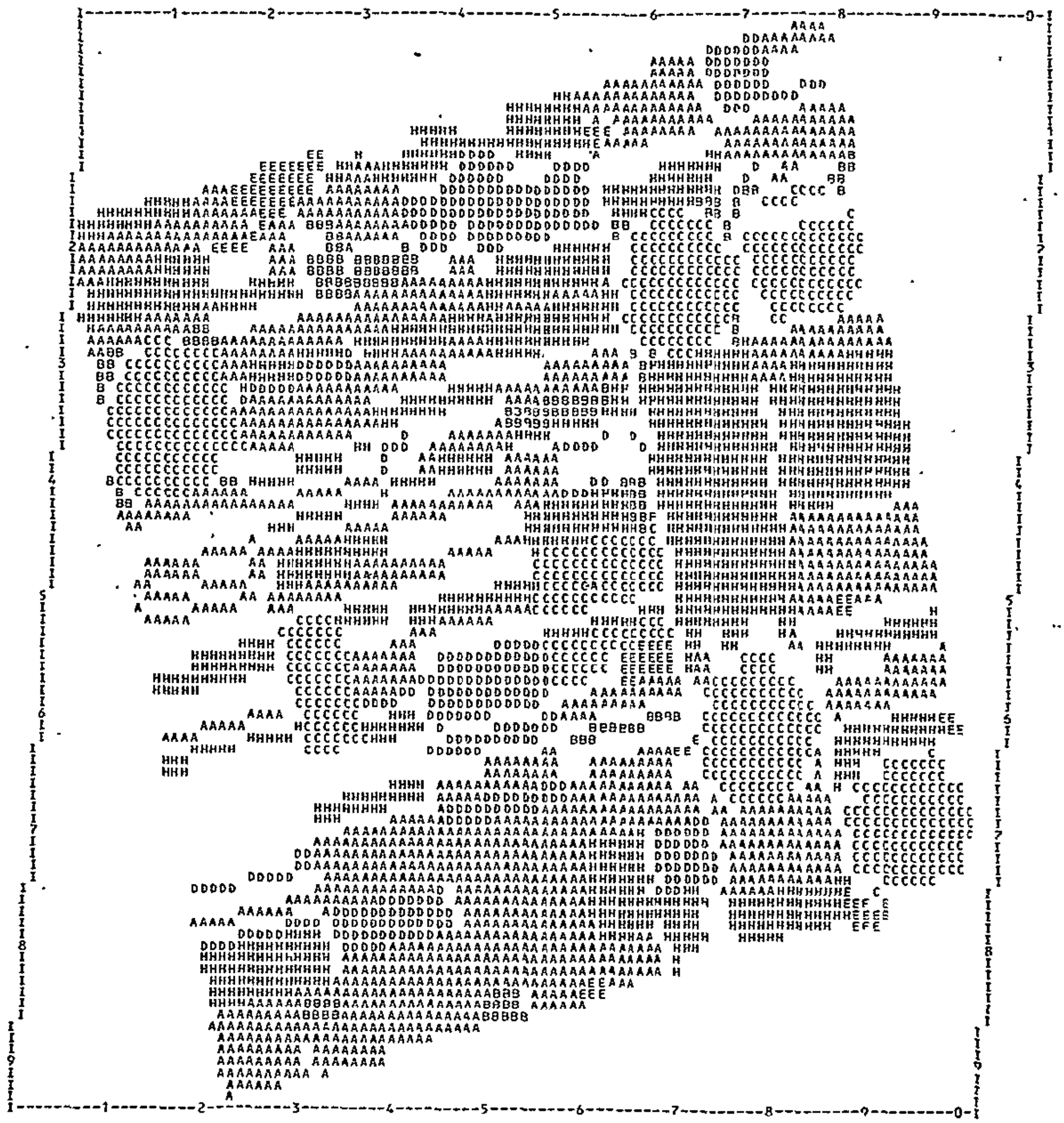


Figure 2.1.8 Rice County - Crop ground truth

FILE NAME - RICCMPSYM BAND 3

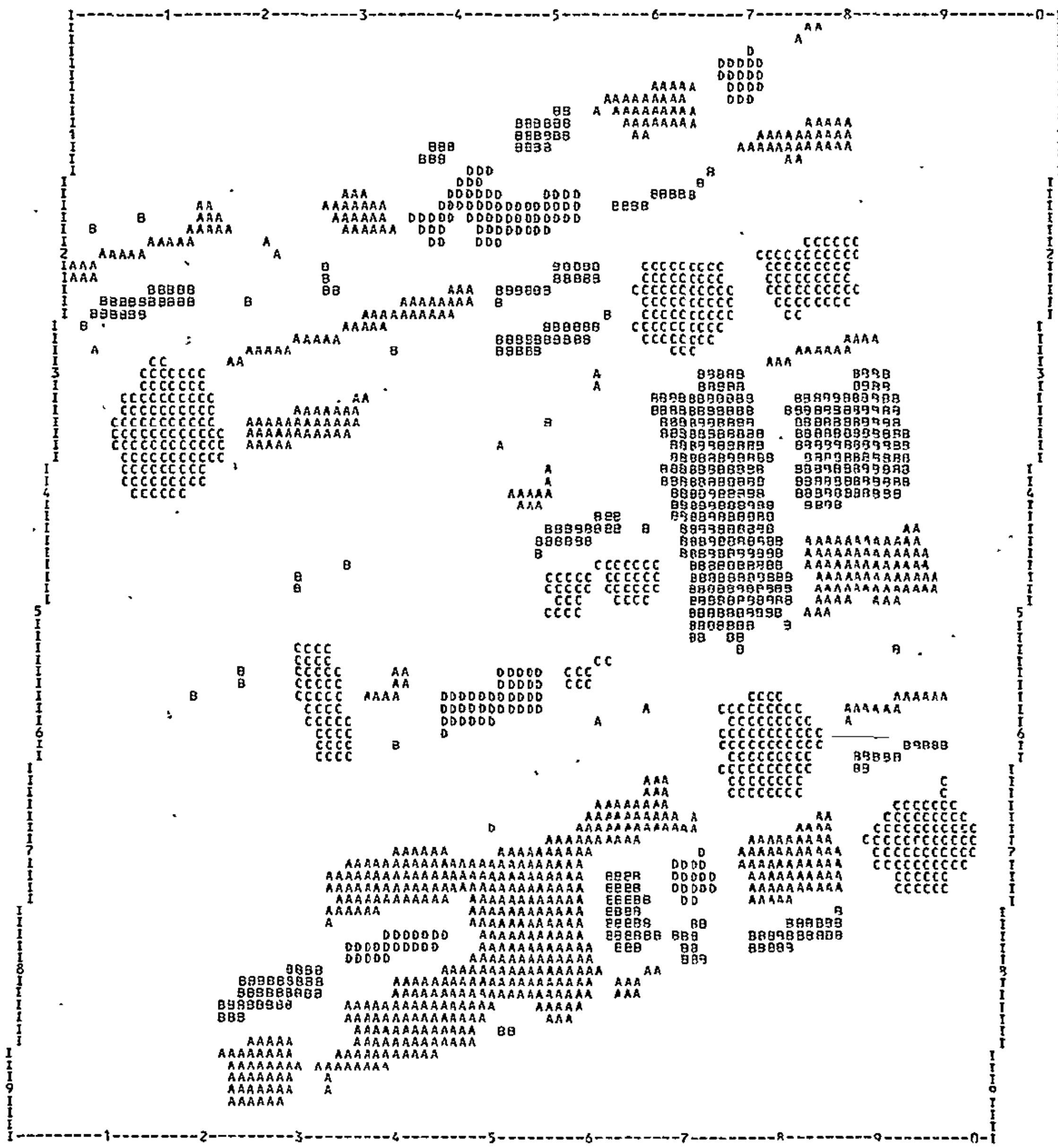


Figure 2.1.9 Rice County - "Shrunken" crop ground truth

FILE NAME - RICCMPSYM BAND 2

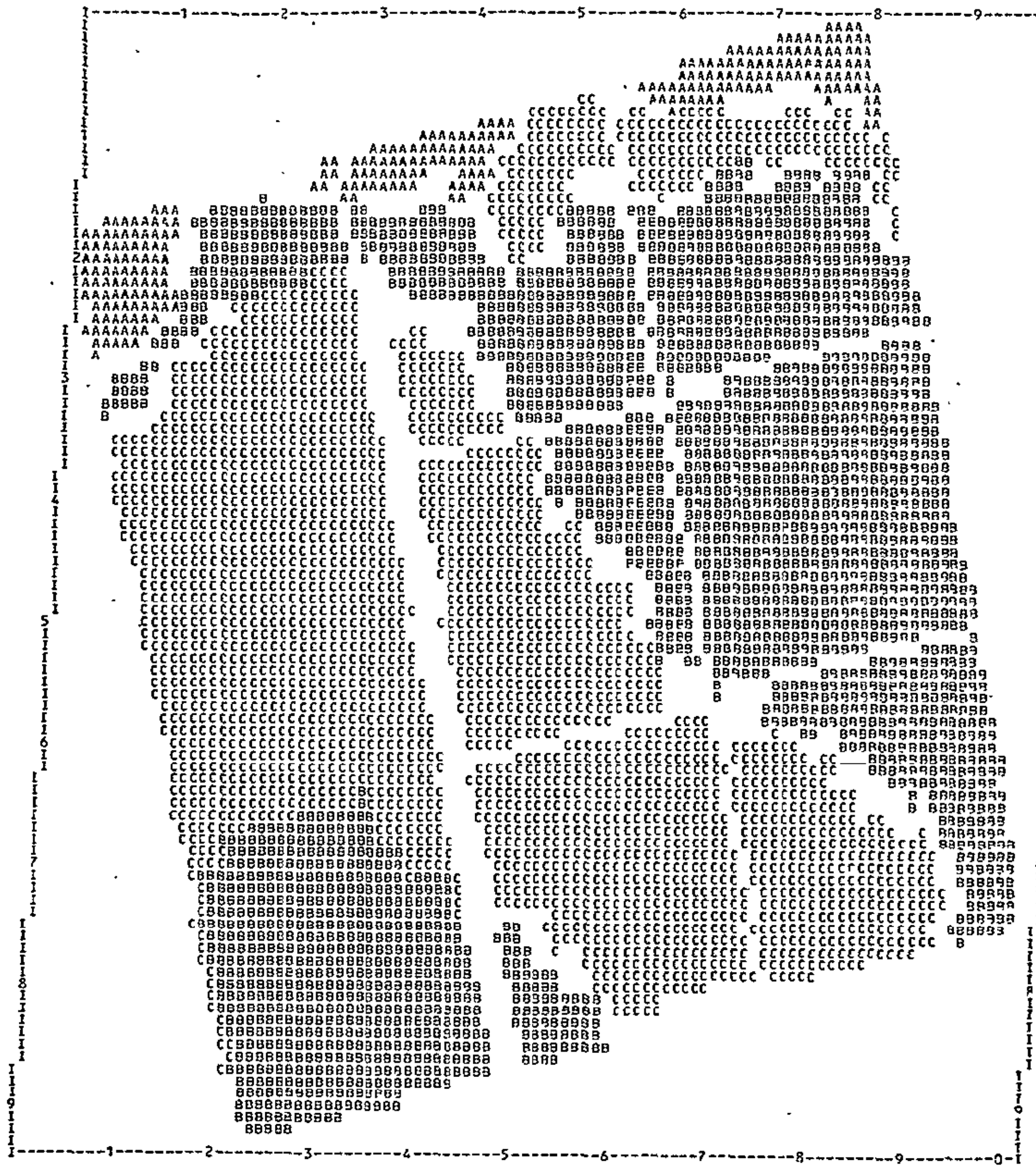


Figure 2.1.10 Rice County - Soil map

rest of the classes can either be merged into these six or discarded, as they are too small.

In Figure 2.1.9, the shrunken ground truth of the area is provided. It is the result of spatially shrinking the ground truth map of Figure 2.1.8. This gets rid of the cells on the edges of the field. It is used when the signatures of some fields are not very reliable.

Figure 2.1.10 shows the soil map for this test site.

2.2 Morton County Image

The Morton County image consisted of twenty bands of information. There are five ERTS images taken on October 23, 1973, May 9, 1974, May 21, 1974, June 14, 1974, and July 2, 1974. Again, for the spatial clustering, only a few bands were selected to minimize the misregistration error. Each of these bands was spatially expanded by two and run through a contrast enhancing operation. This was followed by Roberts Distance 1 gradient option. After thresholding, cleaning and labelling, the resulting image contained 607 regions. Figure 2.2.1 shows us a scaled version of this image. The fields and boundaries were brought out well by this series of operations. Thus, it was not felt necessary to perform a splitting function.

The measurement space clustering was performed using the 10 MSS bands 5 & 7 of the five dates. All twenty bands could have been used, for spatial misregistration is not as critical for this as it is for the gradient operation. Only 10 bands were chosen, however to keep processing time small. Figure 2.2.2 shows the clustered Morton County image. It has 23 classes, which were the result of clustering the original 607 regions of Figure 2.2.1 in 10 iterations. In Figure 2.2.3 we see the final result. It is the spatially generalized clustered image.

While 23 classes seems too many for a small image like this, it should be noted that most of these can be discarded. They constitute a very small percentage of the image. As mentioned before, these small classes have outlying spectral signatures and do not usually correspond to any useful land use class. While it would be interesting to find out exactly what they correspond to, it is difficult to do so, because of their small size.

In Figure 2.2.3, 91% of the image consists of classes A, B, C, and D.

Figure 2.2.1 Morton County - Homogenous regions

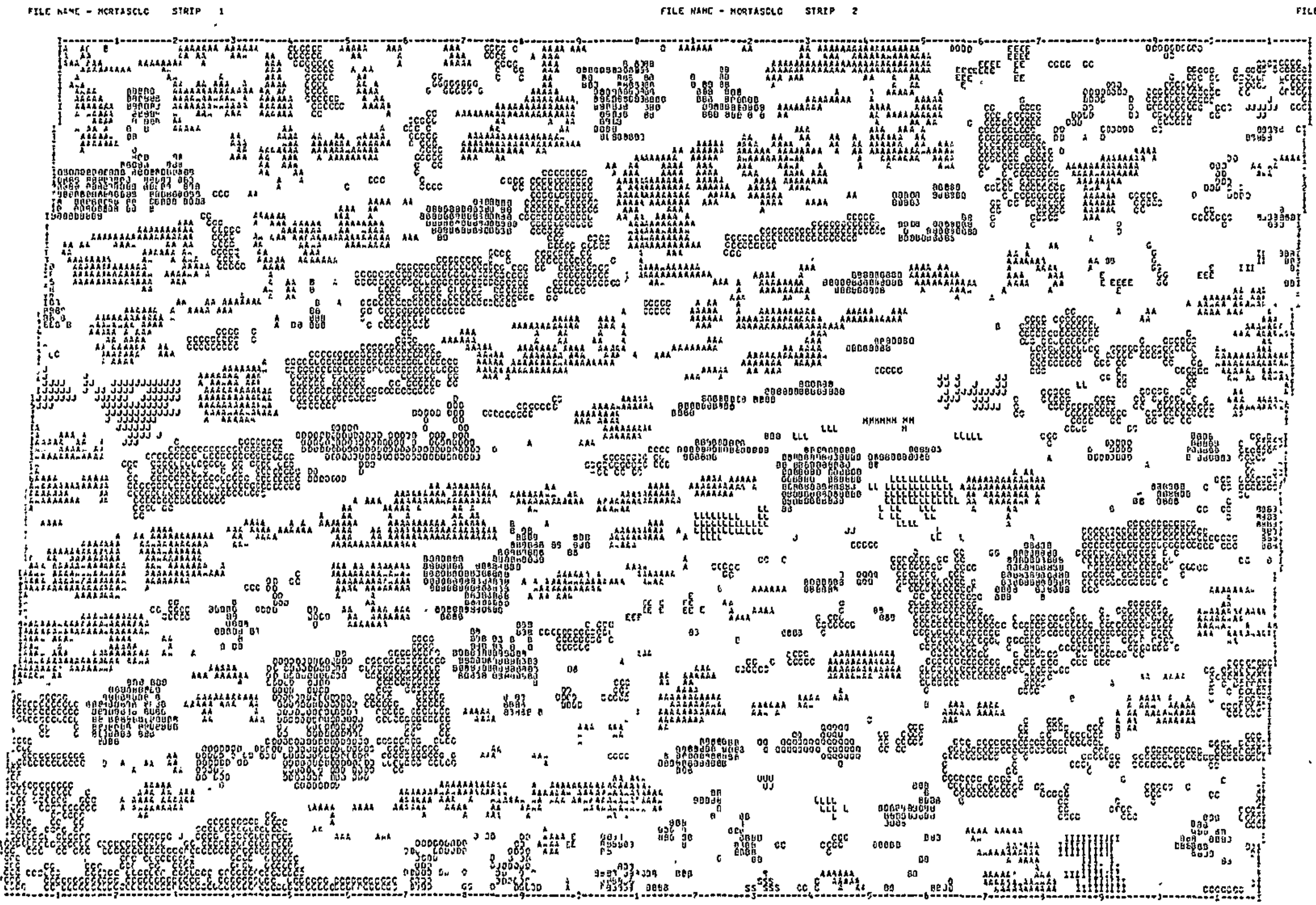


Figure 2.2.2 Morton County - Clustered image - 23 clusters

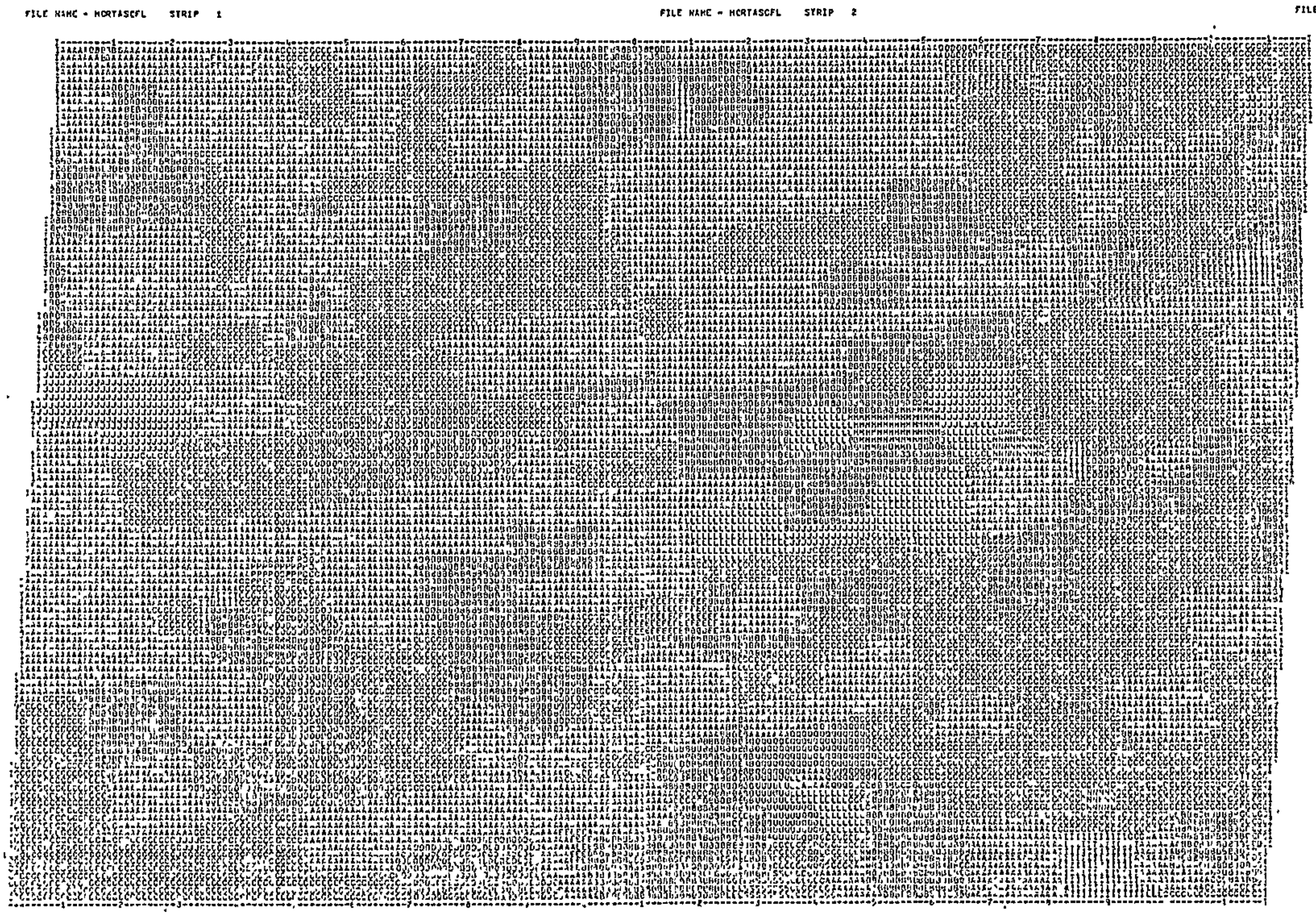


Figure 2.2.3 Morton County - Spatial generalisation of Clusters image

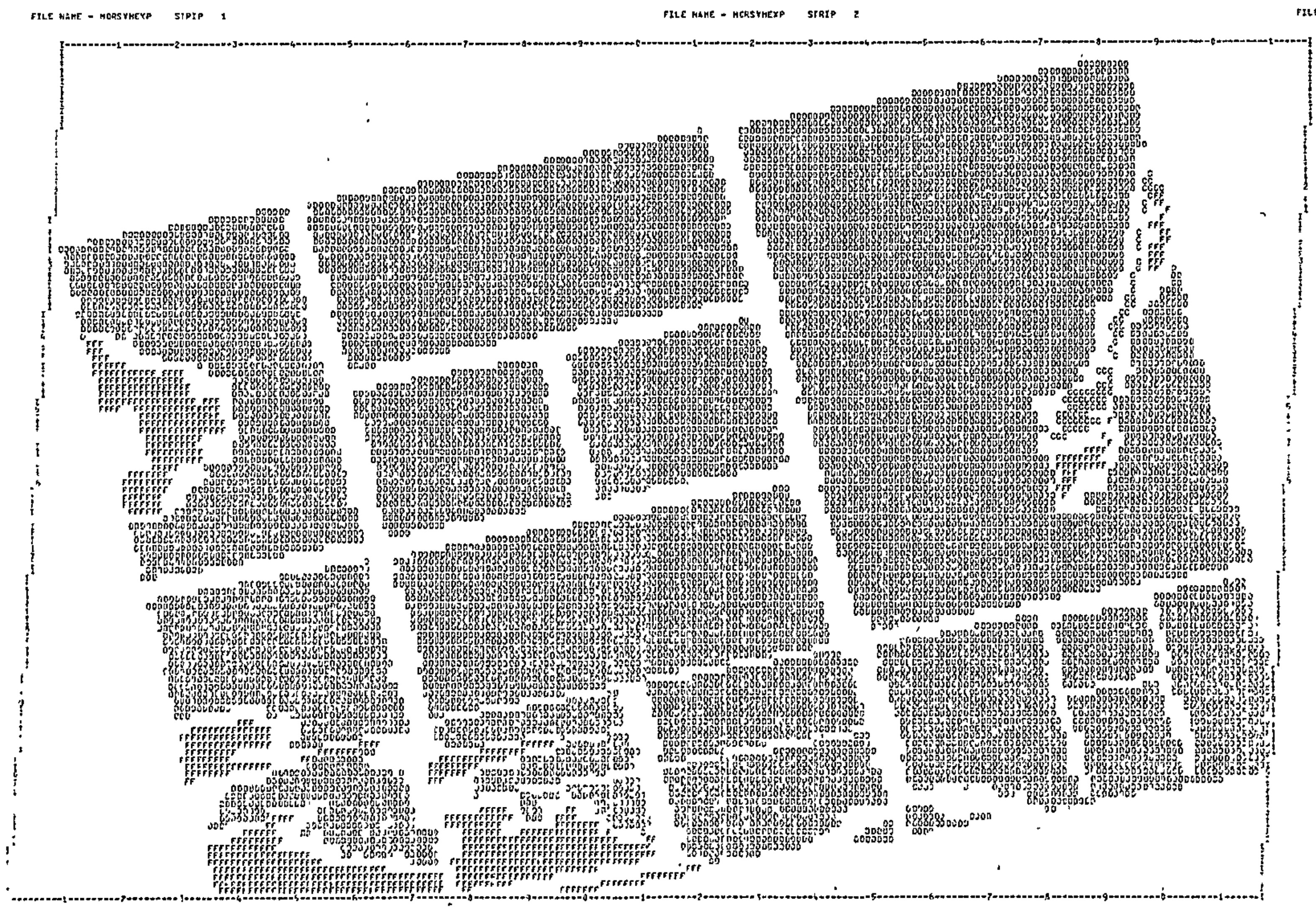


Figure 2.2.4 Morton County - Soil map

Figure 2.2.5 Morton County - Crop ground truth

FILE NAME = MORSTHEXP STRIP 1 FILE NAME = MORSTHEXP STRIP 2 FILE

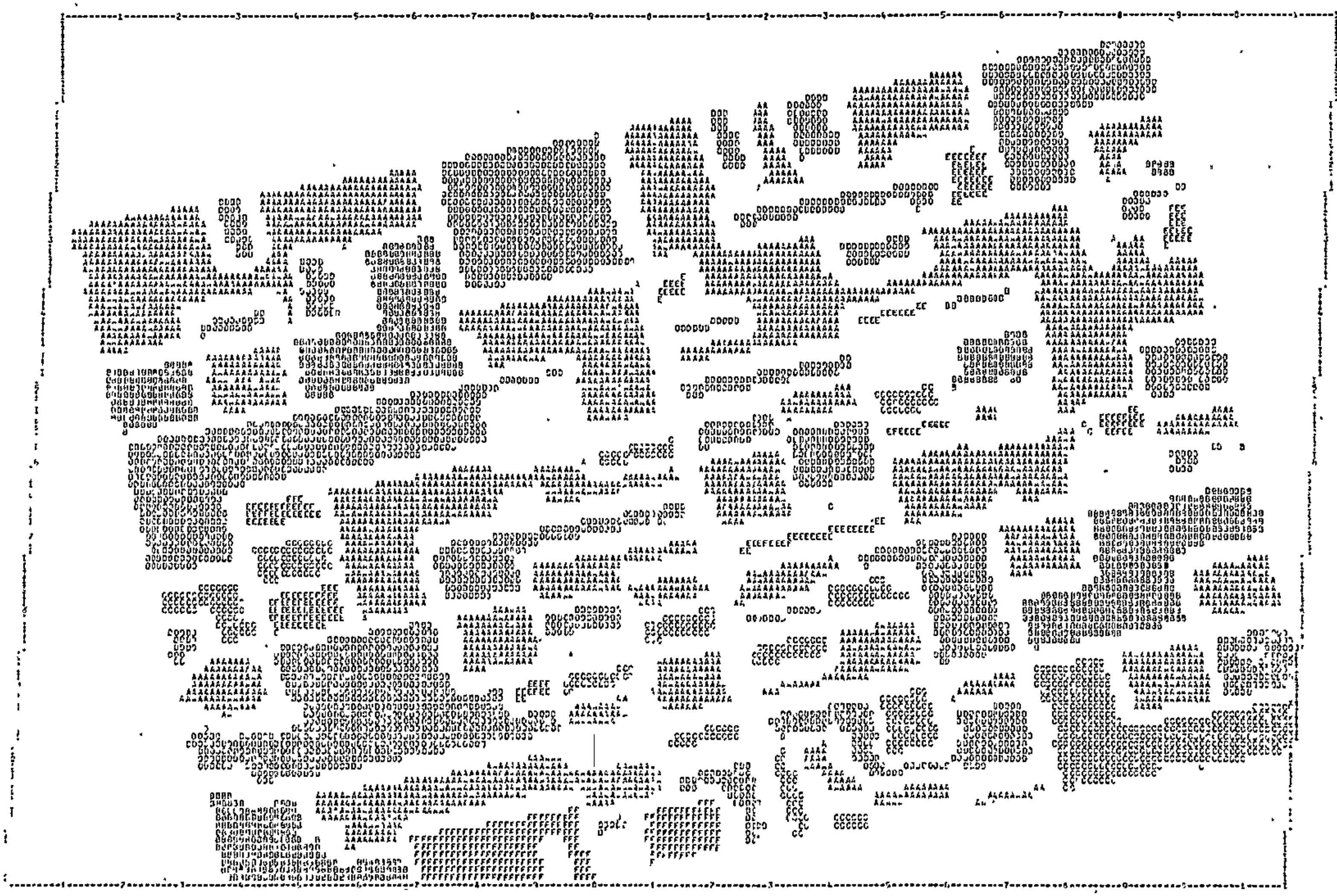


Figure 2.2.6 Morton County - "Shrunken" crop ground truth

Figure 2.2.7 Morton County - Wheat (a) and non-wheat (b) ground truth

By adding the next three largest clusters, E, J, and K, 96% of the image is included. If we go up to the 10 largest classes, it leaves 1.2% of the image divided up between the 13 small classes. These 13 classes may be ignored.

The quantitative analysis and an attempt at semi-automatic interpretation of these classes is discussed in a later section. It is the hardest and the most time-consuming part of this process. Included here are some figues giving the ground truth information which was used for interpretation. Figure 2.2.4 shows the soil map of the area. Figure 2.2.5 is the crop ground truth with 6 categories. These are wheat, grass, corn, summer fallow, grain sorghum and rye. They are denoted by the letters A through F, respectively.

Figure 2.2.6 is the shrunken ground truth map for Norton County. Figure 2.2.7 gives us a wheat versus the non-wheat map. Since the determination of the wheat yeld is important, the wheat class was isolated to be checked as a special case.

2.3 Finney County Clustering

Most of the processing on Finney County was similar to that done on Morton County. This image also consisted of 20 bands of information, and only some were selected for the spatial processing. The dates for the five images were October 23, 1973, April 20, 1974, May 8, 1974, May 26, 1974, and July 1, 1974. Two sets of MSS bands 5 & 7 of pre- and post-harvest dates were chosen. The operations of contrast enhancing, gradient, thresholding, cleaning and labelling were performed on these four bands. Again, all processing was done on a spatially expanded image. The spatial clustering process resulted in 1148 homogenous regions, which are depicted in Figure 2.3.1. This is a compressed and scaled version of the original file. Like for Morton County, it was felt that a splitting operation was not necessary. The fields seemed to be separated nicely from each other.

Measurement space clustering was then carried out using these regions and the 10 MSS bands 5 & 7 of the five original dates. It took 10 iterations to reduce the 1148 regions down to 29 clusters, which are shown in Figure 2.3.2. In the next figure, we have the result from spatial generalization of this image.

Out of the 29 classes, the bulk of the image falls under 6 clusters.

Figure 2.3.1 Finney County - Homogenous regions

Figure 2.3.2 Finney County - Clustered image

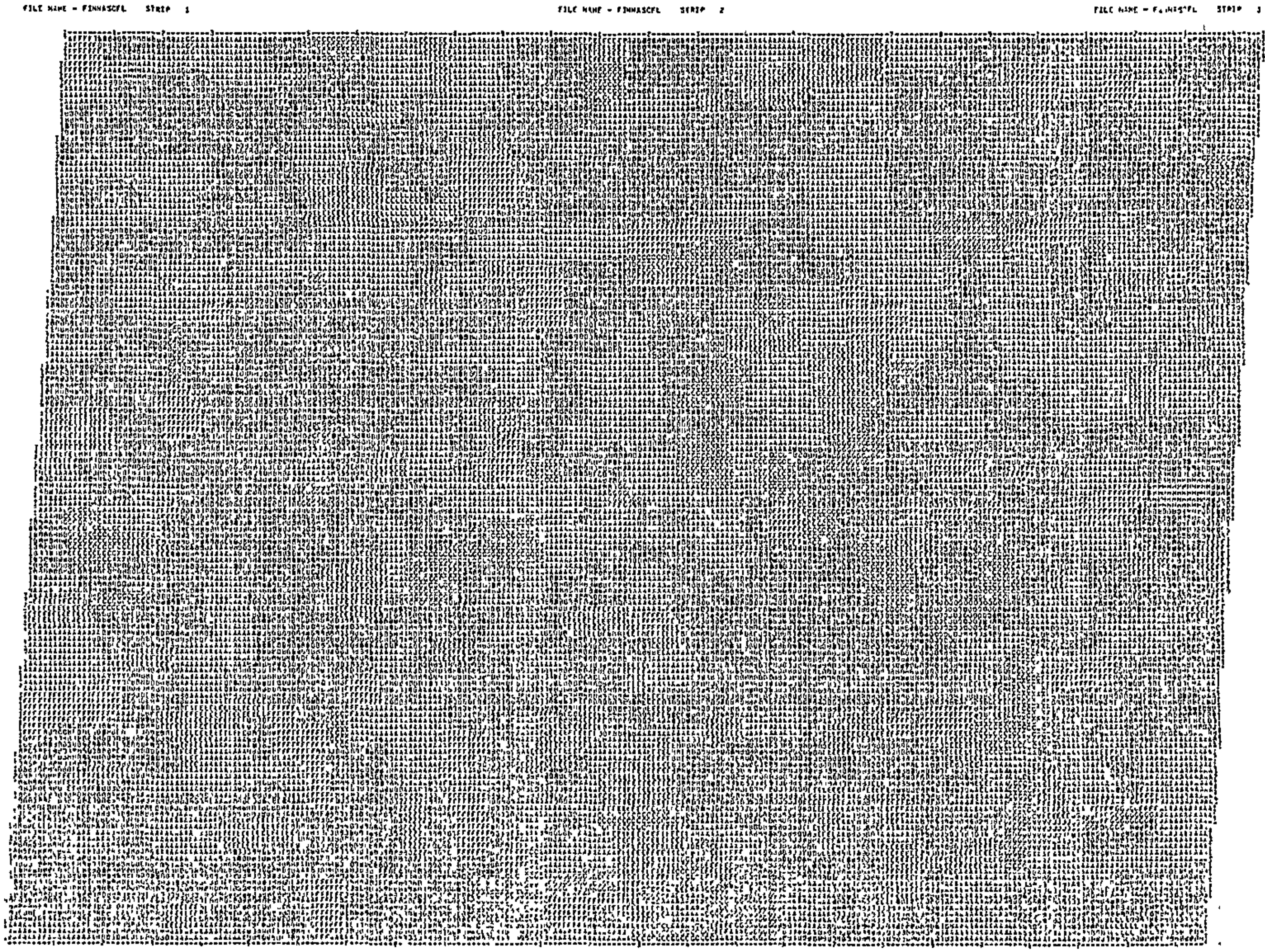


Figure 2.3.3 Finney County - Spatial generalisation of clustered image

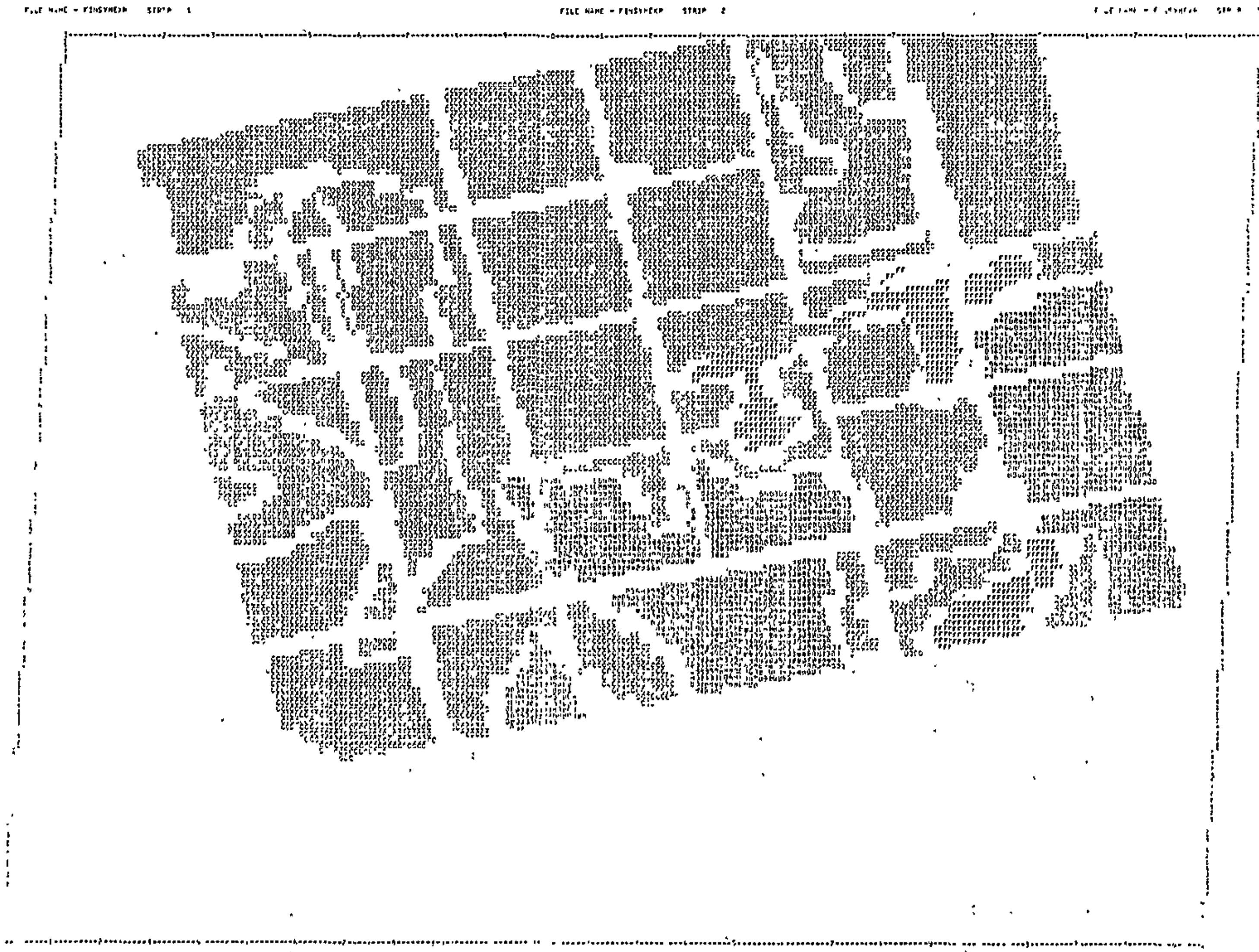


Figure 2.3.4 Finney County - Soil map

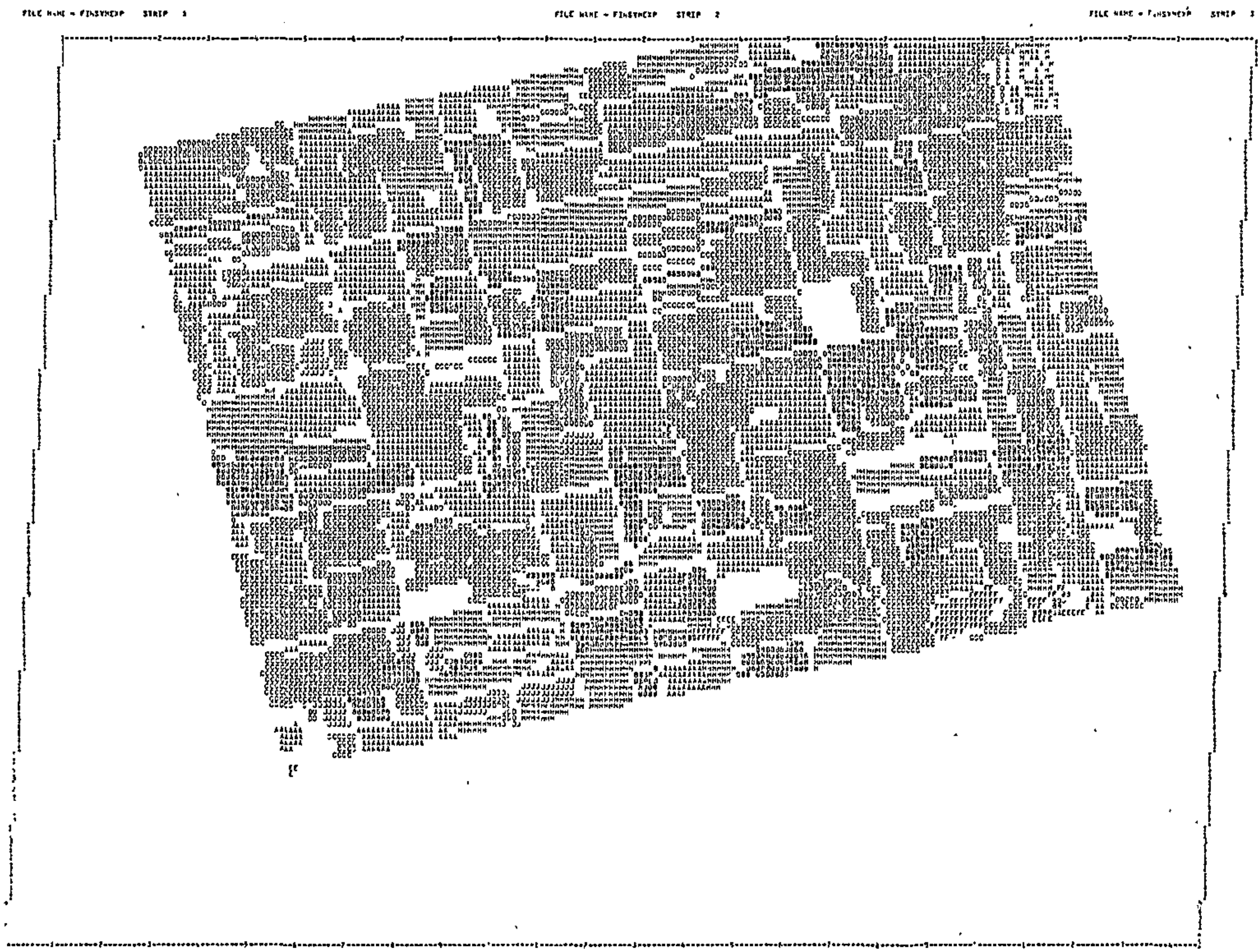


Figure 2.3.5 Finney County - Crop ground truth

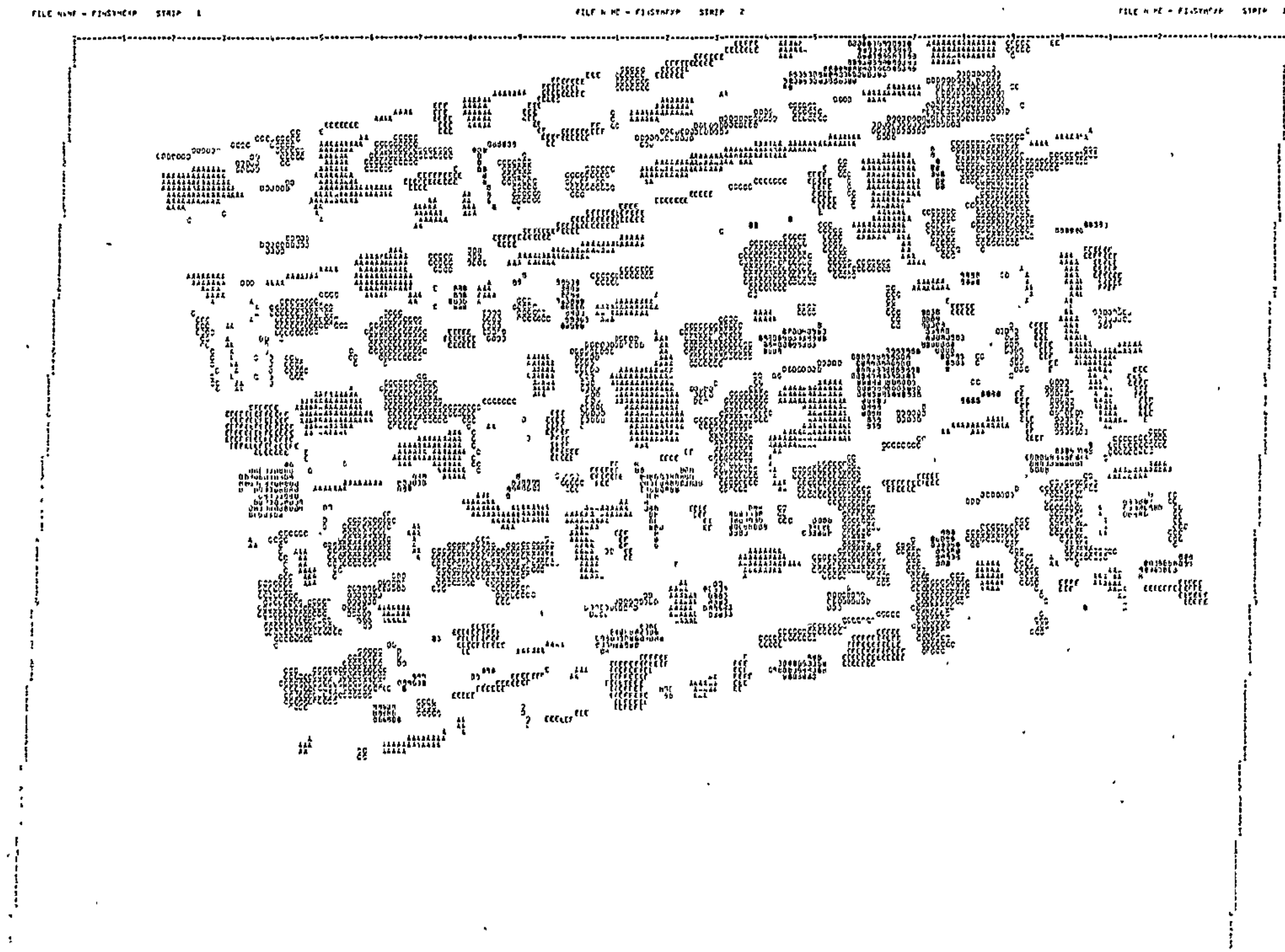


Figure 2.3.6 Finney County - "Shrunken" crop ground truth

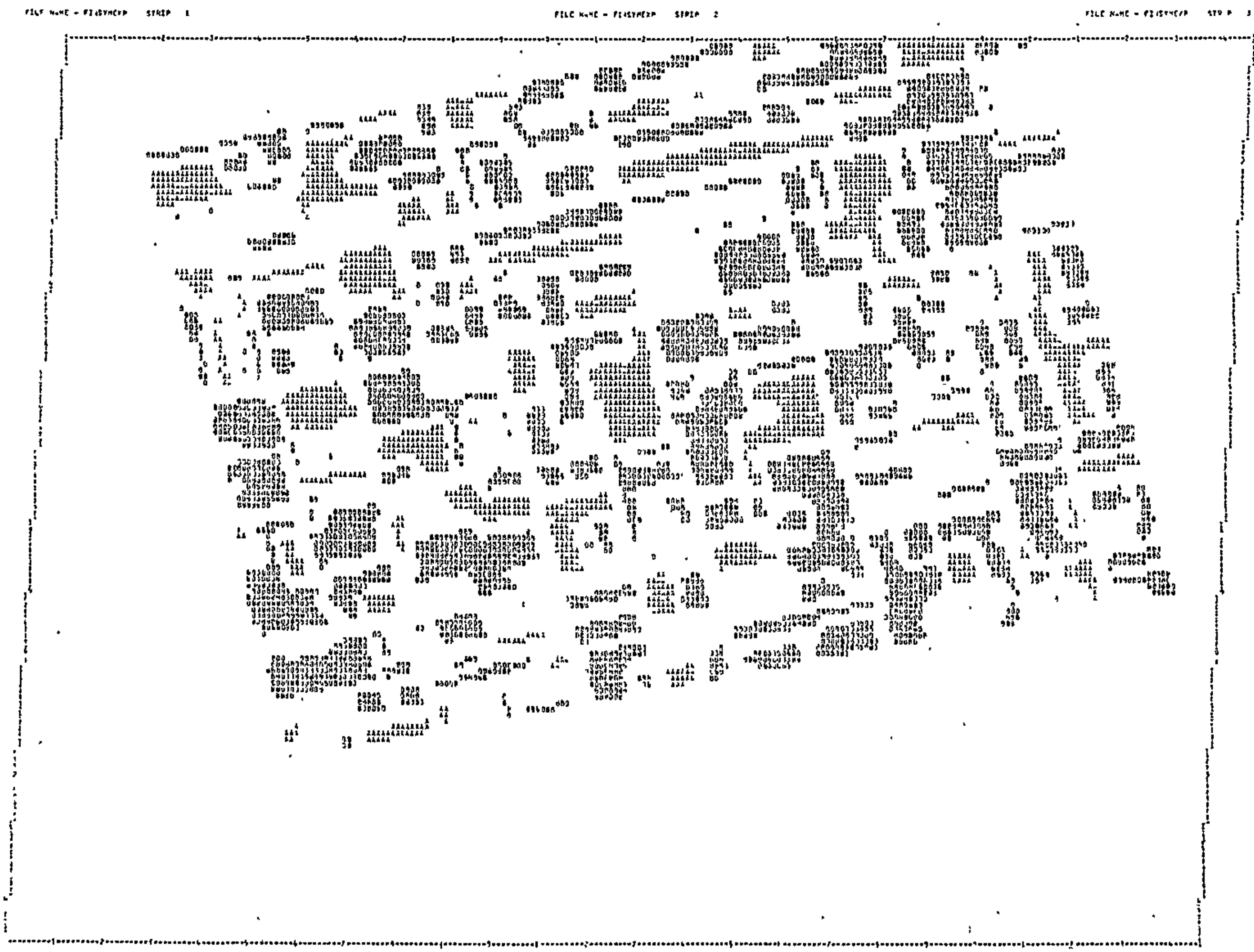


Figure 2.3.7 Finney County - Wheat (a) and non-wheat (b) ground truth

These occupy 98.2% of the image.

The analysis and interpretation of processing the Finney County image is done later on. Included here are the corresponding ground truth maps for this area. Figure 2.3.4 is the soil map of the test site. In Figure 2.3.5 we have the crop information. There are five classes, shown by labels A through E, corresponding to wheat, grass, corn, summer fallow, and grain sorghum. Figure 2.3.6 is the shrunken crop ground truth, while Figure 2.3.7 shows the wheat versus non-wheat layout for this area.

2.4 Saline County Image

It was mentioned in the last report that the ground truth of the images seemed incorrect in some places. The Saline County test site image gave us an opportunity to check both the clustering process and the ground truth. The geography applications laboratory at KU had some aerial photography over part of Saline County taken in April of 1974. Mr. Jim Merchant of the above lab consented to do some manual interpretation of the unsupervised clustering product, using this photography.

Unfortunately, the aerial run did not cover the LACIE test site, but was to the north of it. It did, however, overlap with the top half of the Saline County image provided on tape by NASA. The LACIE test site occupies the bottom right corner of this image.

The image on the tape was 419 rows by 290 columns. The bottom right 331 rows and 150 columns of this were extracted for processing. This gave a vertical strip image with the aerial photograph covering the top half of it.

The analysis of the image was carried out in two ways. First, a computer printout of MSS band 7 of the April date was generated. This is shown in Figure 2.4.1. Using the aerial photograph and county maps, fields and regions were marked out on the printout for ground truth classification. These areas were used for both the Bayes classification rules and interpretating the unsupervised clustering result.

First time through, the unsupervised clustering was carried out on all twelve bands. However, this result was difficult to match to the printout with the ground truths marked on it. To get some idea on the accuracy of clustering, the unsupervised clustering was redone, using the 4 bands of the April date. These were first expanded, which gave an image 662 rows

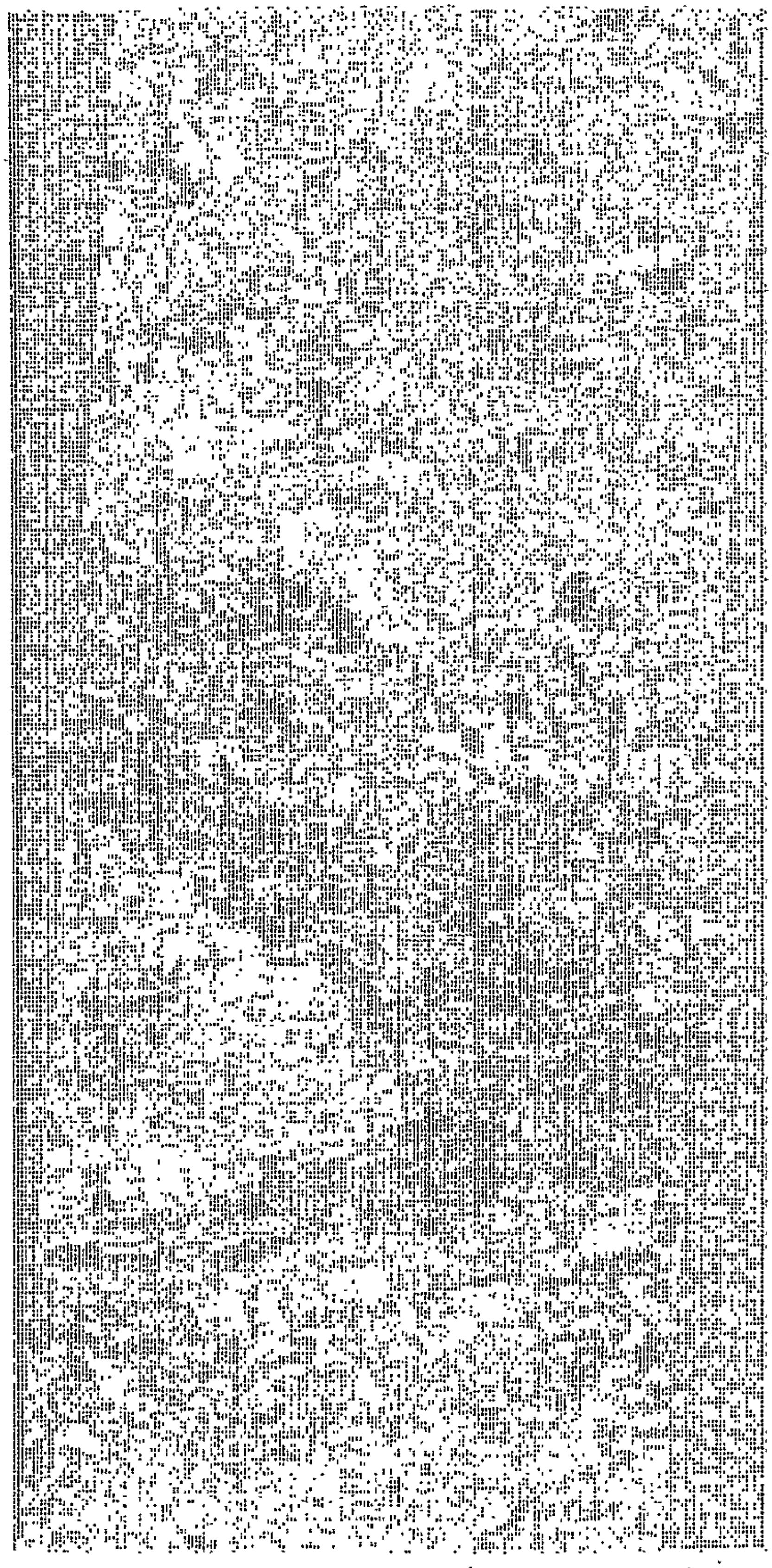


Figure 2.4.1 Saline County - MSS Band 7 - April 18, 1974

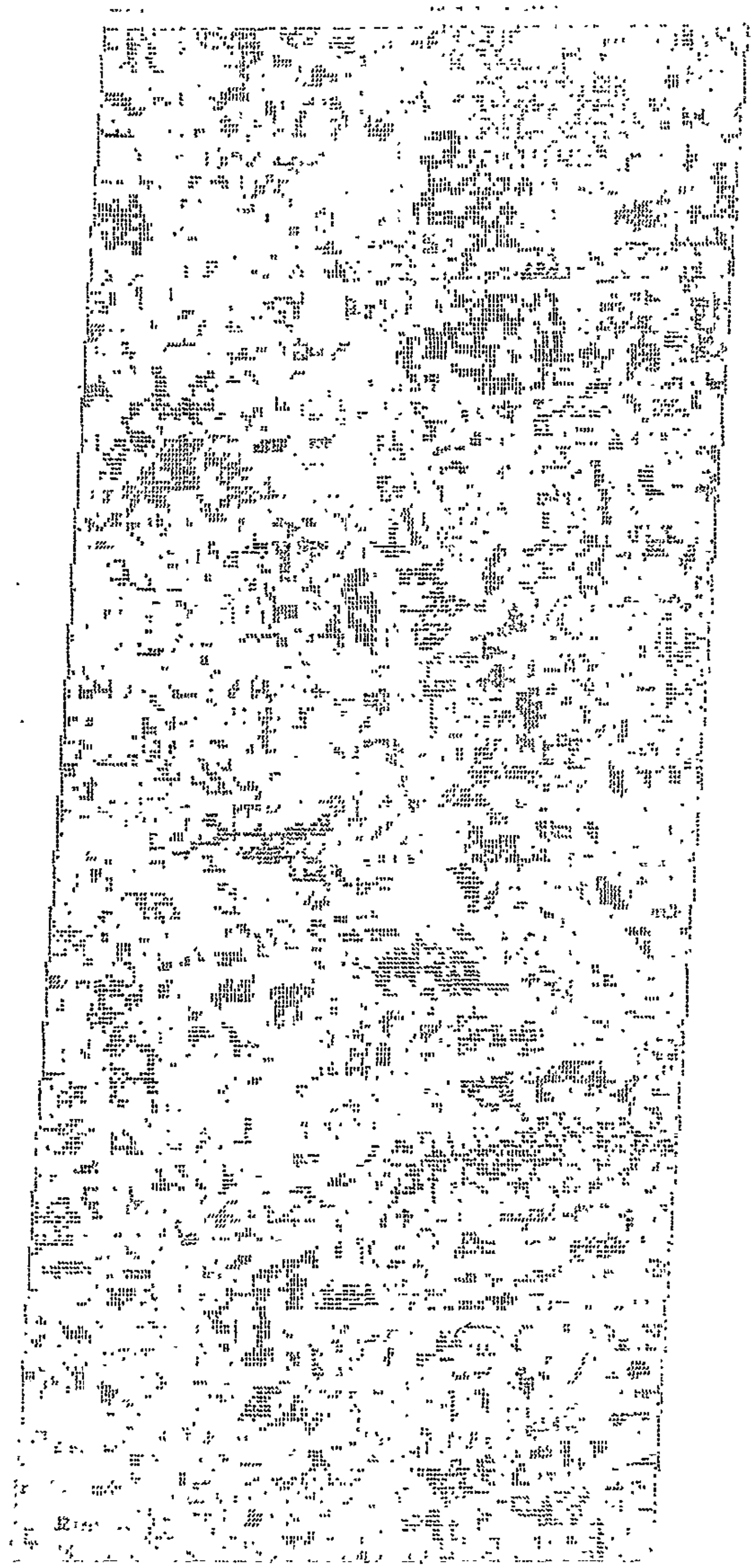


Figure 2.4.2 - Saline County - Homogenous regions

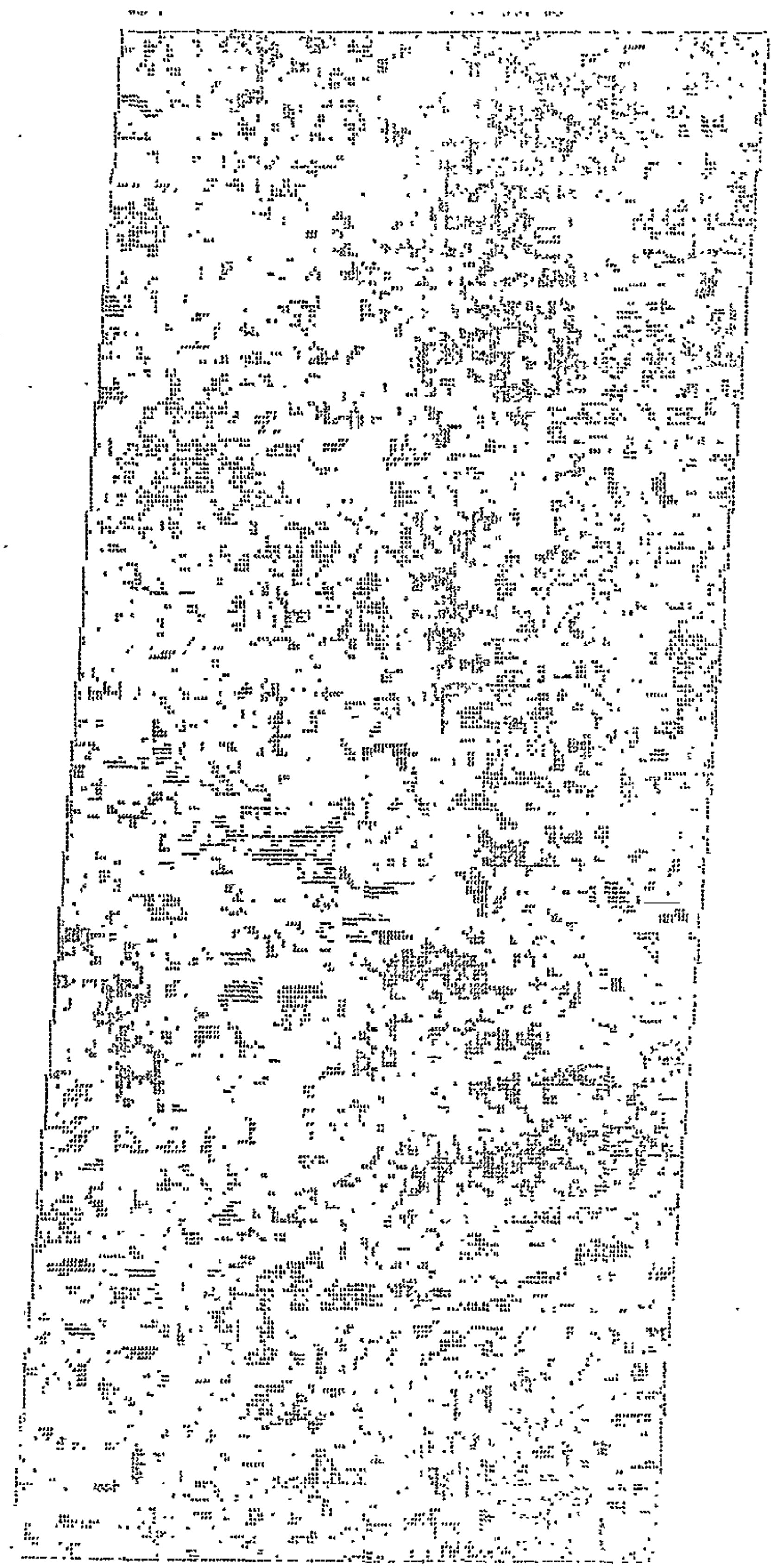


Figure 2.4.3 Saline County - Clustered image

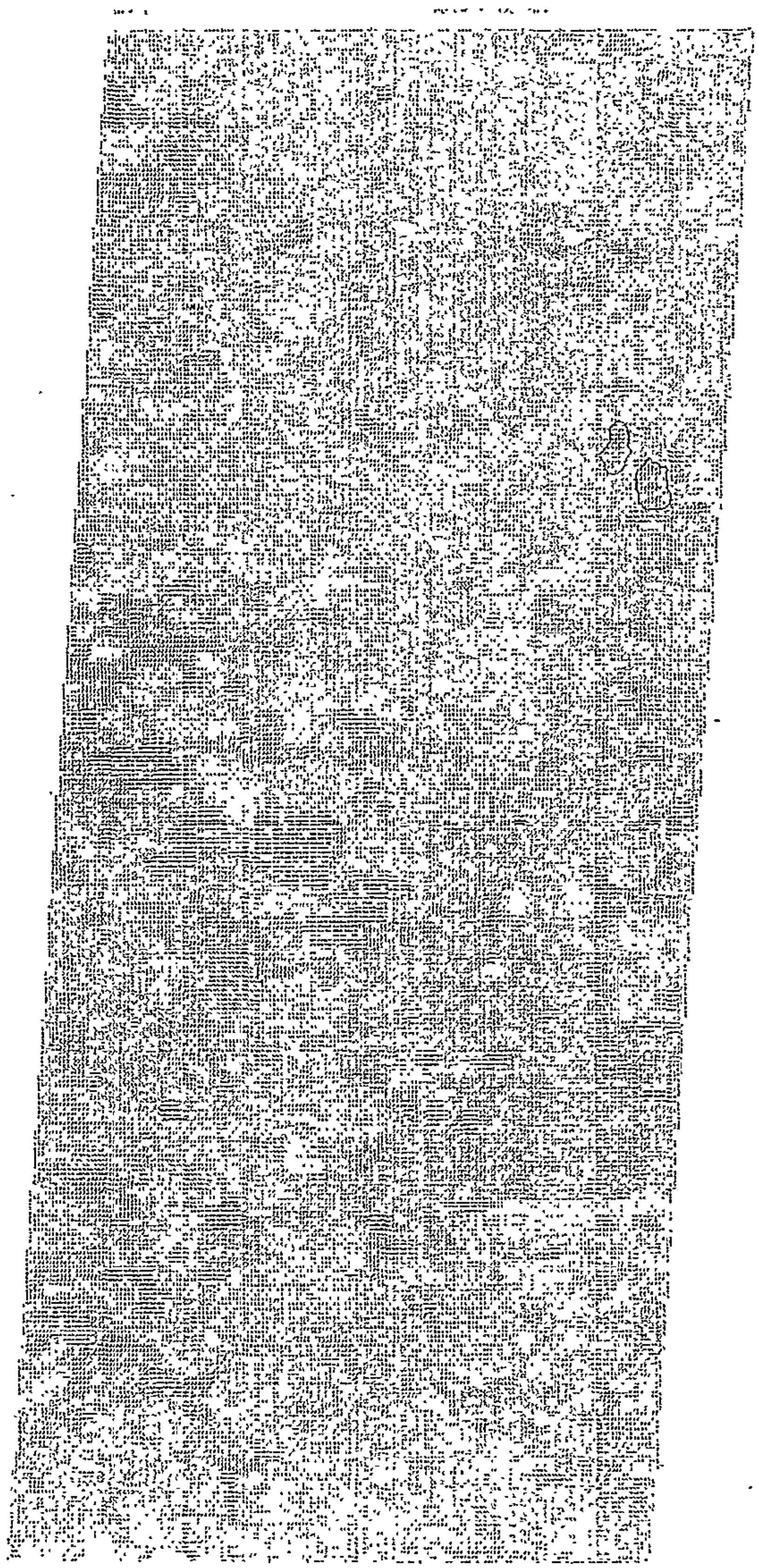


Figure 2.4.4 Saline County - Spatial generalisation of clustered image

by 300 columns. The spatial clustering was carried out on MSS bands 5 & 7 only. The resulting image had 2359 regions, and a compressed version of this is shown in Figure 2.4.2. The clustered image is shown in the next figure. It was obtained using all four bands of the April date. The 2359 regions were reduced to 23 classes in 8 iterations. The image was spatially generalized and this result is shown in Figure 2.4.4.

An interpretation of the cluster labels was carried out using the aerial photograph and the ground truth computer printout. The result of this is given by Table 2.4.1.

LABEL	FREQUENCY	CATEGORY
A	18.5%	Wheat
B	1.7%	Wheat
C	3.5%	Rangeland
D	18.6%	Rangeland
E	3.1%	Woodland
F	7.0%	Bare ground
G	17.5%	Bare ground
H	4.5%	Rangeland-grass-pasture
I	0.3%	Rangeland
J	2.9%	Bare ground
K	17.1%	Wheat
N	2.7%	Wheat
P	1.0%	Woodland
L,M,Q-W	1.6%	Not classifiable

Table 2.4.1

A few points emerge upon examing the table. The clustering came up with no less than 4 labels for wheat, along with multiple labels for rangeland and bare ground. At first glance, it would suggest that the clustering may have been stopped prematurely. Another iteration or two would have merged them together. However, this may not be true. The different shapes of wheat or different types of rangeland, may constitute the same nominal class, but have different spectral signatures on the satellite image. Thus, the computer treats them as separate spectral categories. Imposing further, clustering may cause the merging of some spectral categories across nominal classes. For example, if one sub-class of wheat has a signature closer to a rangeland sub-class than to the other wheat sub-classes, in an additional iteration, it may be put together with rangeland.

As a matter of fact, this may already have happened in two instances. As may be seen from the table, there was no water class detected. However, on the image printout, there are occurrences of water. One example is the meandering river flowing in the center left of the picture. It was not expected that the spatial process would pick this river, as its width is too thin for the gradient and cleaning operators to resolve. That holds true also for some scattered ponds, which consist of a few cells each. However, there is one large lake about a hundred cells in size which lies one-third way down from the top and about 40 cells from the right border (see Figure 2.4.4). It was formed to be classified with a label "G", which corresponds mostly to bare ground. This indeed is unfortunate, as water is an important class and should have come out separately. The second instance for misclassification is for an occurrence for label "E". The aerial photograph tells us that the area lying just below and to the right of the lake is bare ground. The clustering process, however, gave it the label for woodland category.

Other than these cases, the clustering seemed to be quite good. It, however, needs to be determined why this misclassification occurred. If these misassignments occurred in the last iteration, then perhaps the clustering went too far, or the user entered parameters that were too liberal. This calls for an examination of the clustering process interation by iteration for these particular areas. Unfortunately, at this time the intermediate signatures of the clusters are not available. It would let us see whether the clustering was overdone or if the original signatures of the areas were misleading to begin with. Currently, the process is being modified to print these intermediate results for such verification.

Once the labels are associated with classes, some relabelling may be carried out. This would allow all the sub-classes of the same nominal category to be assigned the same label. That then would be the final clustering result.

In the table above, the current labels cover about 98.4% of the image.

Out of the remaining 1.6%, the label "0" constitutes 1% of the image. It seems to be the only unclassified area large enough to be important. However, as it occurs in quantity only in the lower half of the image, it was difficult to find what category it represented. The supporting photograph of the ground truth only covered the top half of the strip.

The relabelling will be done once the misclassifications in the image have been rectified.

The human interpretation yielded only one crop class--wheat. There are a few reasons for this. Wheat was predominant over the entire strip. The other crops occurred only in small areas. Also, the interpretation was made using two bands of the April date, for which only wheat stood out.

For the April date alone, the classification seemed pretty good. It is hoped that similar analysis can be done using other dates, if additional aerial photography can be obtained.

To check the ground truth after relabelling, we can compare the clustered result with the ground truth of the test site. The fields which are of the same class should have the same label in both images. Discrepancies can be singled out and their signatures checked. The corresponding areas in the ground truth can then be corrected.

2.5 Conclusion

Although we have finished the clustering, the comparison between the clusters and the ground truth has not been completed. During the next quarter we will fix any incorrectly given ground truth and report on the detailed comparison. It is hoped also that some of the county images will be reclustered, using an "Isodata" option, in the measurement space clustering. This is at present being implemented. It should reduce the occurrencesof mis-assignments, that came up for the Saline County image.

3.0 Spectral-Tempral Classification Using Vegetation Phenology

The usual model for classification implicitly assumes that the phenological growth stage for each vegetation category is the same for all observations made at a single time. It is well known, however, that even in a geomorphologically homogeneous area, the phenological growth stages for each vegetation type is not the same. This slop in phenological growth stage is then reflected in probability distributions of larger variance than they should be. The larger variance causes a lower classification accuracy for an optimal decision rule. The solution to this problem is to focus on what phenological growth stages are appropriate for any spectral reflectance for each category.

One classification algorithm which makes use of vegetation phenology has a direct and simple description. For example, if a 2-band spectral observation (α_1, α_2) is made using wavelengths (λ_1, λ_2) at time t_1, classification can be done by determining for each category c all those phenological growth stages of vegetation of category c which can yield spectral return α_1 at wavelength λ_1 and spectral returns α_2 at wavelength λ_2. If there is not a phenological growth stage of category c which yields spectral returns α_1 and α_2 at wavelengths λ_1 and λ_2, then category c is not a possible choice. Classification may then be done by eliminating inappropriate category choices. Spectral observation taken at a later calendar time can be naturally constrained to be associated with later phenological growth stages in order to keep an earlier accepted possibility of category c remaining open at the later observation time.

3.1 Bayesian Perspective

The usual model for the classification of remotely sensed data has been one relying on simple statistical structure. For example, to discriminate corn from wheat, it is assumed that data vectors coming from the corn category are distributed according to one probability distribution and data vectors coming from wheat are distributed according to another probability distribution. A Bayes or maximum likelihood criteria then determines an optimal decision rule.

Unfortunately, the world is not as simple as the model assumes. Besides atmospheric haze and geomorphologic soil and moisture variations, which undoubtedly affect spectral returns, it is not the case that data vectors coming from corn are distributed according to a simple probability distribution. Rather, for each phenological growth stage for corn, spectral data vectors coming from corn are distributed according to a probability distribution. The statistical model, therefore, requires the probability function of spectral reflectance vector x coming from the category c in phenological growth stage g at calendar time t. We denote this probability by

$$P_t(x, c, g).$$

For multi-temporal multi-spectral data, the probability function of spectral reflectance vectors $x_1,\ldots x_N$ coming from categories $c_1,\ldots c_N$ in phenological growth stages $g_1,\ldots,g_N$ at calendar times $t_1,\ldots,t_N$, respectively is denoted by

$$P_{t_1,\ldots,t_N}(x_1,\ldots,x_N, c_1,\ldots,c_N, g_1,\ldots,g_N).$$

To determine a Bayes rule, the probability

$$P_{t_1,\ldots,t_N}(c_1,\ldots,c_N, x_1,\ldots,x_N)$$

must be determined. Now,

$$P_{t_1,\ldots,t_N}(c_1,\ldots,c_N, x_1,\ldots,x_N) = \sum_{g_1}\ldots\sum_{g_N} P_{t_1,\ldots,t_n}(x_1,\ldots,x_N, c_1,\ldots,c_N, g_1,\ldots,g_N)$$

$$= \sum_{g_1}\ldots\sum_{g_N} P_{t_1,\ldots,t_N}(x_1,\ldots,x_N|c_1,g_1,\ldots,c_N,g_N) P_{t_1,\ldots,t_N}(c_1,g_1,\ldots,c_N,g_N)$$

The causal mechanism which produces a reflectance x given a category c at phenologic growth stage g guarantees that

$$p_{t_1,\ldots,t_N}(x_1,\ldots x_N|c_1,g_1,\ldots c_N,g_N) = \prod_{n=1}^{N} P\ (x_n|c_n,g_n)$$

since a vegetative category and its phenological growth stage at time t_n are the only determinants of the spectral reflectance x_n. Likewise, the vegetation growth mechanism guarantees that

$$p_{t_1,\ldots,t_N}(g_1,\ldots g_N|c_1,\ldots,c_N) = \prod_{n=1}^{N} P_{t_n}(g_n|c_n).$$

Hence,

$$\begin{aligned} P_{t_1,\ldots t_N}(c_1,\ldots,c_N,x_1,\ldots,x_N) = & \sum_{g_1} p(x_1|c_1,g_1)p_{t_1}(g_1|c_1) \\ & \cdot \sum_{g_2} p(x_2|c_2,g_2)p_{t_2}(g_2|c_2)\cdot\ldots\cdot \\ & \cdot \sum_{g_N} p(x_n|c_N,g_N)p_{t_N}(g_N|c_N) \\ & \cdot\ p_{t_1,\ldots t_N}(c_1,\ldots,c_N) \end{aligned}$$

In theory, the formula just derived could be used to determine a Bayes rule in the usual way. In practice, there are too many distributions to estimate and too many calculations to do to calculate the required probabilities. However, because the required probability has the form of a product, then if any probability in the product is zero, then the product must be zero. And a Bayes rule would never make an assignment to a category with a zero probability. This fact can be capitalized on to make an efficient table look-up rule.

3.2 Summary of Work to be Done for Final Quarter

Using the corrected ground truth for each test site and the spectral time plots of each crop category by field, we will construct spectral-phenological growth state curves for each category. Then using the theory developed in 3.1, we will implement a table look-up rule which classifies using the derived growth state curve. The results of this table look-up scheme will be compared with the classical approach. We are expecting the classification error to decrease by a factor of three with this procedure.