Reprinted from

Ninth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Natural Resources Evaluation

June 21-23, 1983

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

A FLEXIBLE CLUSTERING PROCEDURE FOR USE IN AN UNSUPERVISED CLASSIFICATION OF LANDSAT DATA

W.Y. CHEN

University of Shandong
Jinan, China

W.G. COLLINS

University of Aston
Birmingham, United Kingdom

ABSTRACT

A new procedure for finding the data structure is proposed which can be used for unsupervised classification of MSS data. It can, under certain conditions, find and adjust the clusters. These conditions are that:

1) The number of members in each cluster must be big enough (>NMIN) to overcome noise or abnormality of any pixels.
2) The members of each cluster must be concentrated around the corresponding cluster centre. The difference between the member and the centre, and the variance in each cluster, must be less than a certain value (<DEVMAX), otherwise the cluster will be divided into two or more
3) The distance, in the feature space, between any two cluster centres must be long enough (>DSL), otherwise they will be combined into one single cluster.

In order to get more flexibility and find the data structure more objectively, the parameters (NMIN, DEVMAX, DSL) can be found in the histogram and the scatter diagram, and can be changed according to the user's need.

The procedure can produce one classification map and a group of curves. The correctness of the first one can be checked by the second.

I. INTRODUCTION

It is well known that in the measurement space [1] the MSS data distribute in clusters. The main task of any unsupervised classification algorithm is to identify these clusters, how many there are, the distance between them, the variation in each of them, or, in summary, to find the data structure. The algorithm proposed here serves this purpose, but this algorithm has some features of its own which are as follows:

a) Clarity
At the beginning a general picture of the processed data can be obtained from the histograms in each MSS band and the scatter diagram [2] in each pair of MSS bands.

b) Self-adjustability
The number of clusters is not fixed and can be self-adjusted automatically to suit the processed data set while the program is running.

c) Choice
It provides the optimum choice of the cluster centres. A special "DO LOOP" is employed in the processing, using different groups of cluster parameters in each iteration, and producing a series of results, from which the best choice of the cluster centres can be made.

d) Self-checking
At the end of the procedure, there are two products - a classification map and a group of curves - which are used for showing the reflectances of the pixels in the related spectral band along an arbitrarily chosen line. Comparing these curves with the classification map, an assessment of the procedure can be made.

The whole clustering algorithm includes three programs: "STATESW", "CLASC", and "CLASB", which will be discussed separately and in more detail.

II. EXAMINING THE DATA STRUCTURE

At the start of data processing, it is important to know something about the data structure, and the program "STATESW" is used for this purpose.

Supposing that a general purpose computer is in use and the output channel is through a line printer, the data set to be processed must not be too big and due attention must also be paid to estimating the range of the data distribution in the output. For example, in the program "STATESW", "LLL" and "MMM" are used for representing the minimum and the maximum digital value respectively in the four bands, and all the pixels processed correspondingly.

The program consists of the following steps:

A) Start the program, read in the image data and initiate the parameters.
B) Establish the statistics for each MSS band, find the range and present the results in the form of a histogram. This is done by calling a

pack of subroutines: "MULTISTATIST", "LEVEL", "PERCENT", "BARPIC".

C) Establish the 2-D statistics for each pair of MSS bands and show the results by a scatter diagram. The calling of subroutines: "STATIST2" and "CLSTSW" is for this purpose.

The flow chart is shown in Figure 1. and some results are as follows:

Figure 2. is the histogram, showing the relation between the digital number of the reflectance in band 6 and the percentage of the pixels which have the corresponding reflectance.

Figure 3. is the scatter diagram. The reflectance digital numbers of the two different bands (band 4 and 6) are used as co-ordinates for constituting a two dimensional measurement space. At each point in the space there is a character, (including the blank as a special one), representing the number of a special kind of pixels, which possess the reflectance values corresponding to the co-ordinates of the point. There are ten different characters used in the scatter diagram, viz. "blank", ".", ":", "-", "!", "T", "F", "E", "G" and "W", which are program numbered from one to ten. Different characters represent different numbers of pixels, which are made to increase accordingly to a definite pattern by the statement:

$$N = C * (I - 1) ** 2$$

where N is the number of pixels
C is a constant, depending on the maximum value of the number
of pixels corresponding to one point in the measurement space
I is the cluster number

From Figures 2. and 3. the clustering tendency can be seen clearly; Figure 4. shows that a strong correlation exists between the reflectances in band 6 and in band 7.

III. CLUSTER FINDING

The information obtained from the histograms and the scatter diagrams can be used for estimating the cluster parameters, such as the number of clusters, the minimum distance between the clusters, the maximum variation of the reflectance digital number in a cluster. The estimation of the parameters is very essential, and this has to be done before running the program "CLASC", which is used for finding the cluster centres.

The flow of data is illustrated by Figure 5. Most of the steps in the flow chart are self-explanatory, but there are two which need to be discussed in more detail.

A) Sampling the image data set. It is noted that clustering algorithm involves iteration, and this might incur undesirably high demands on computational time and costs if the number of the pixels to be clustered were to be very large. For these reasons sampling is necessary. While reducing the number of pixels to be clustered, the sample of the image data must be taken over the whole data set, otherwise it is very likely to miss some of the clusters.

B) Call subroutine "AUTOCLASS". The subroutine "AUTOCLASS" is the nucleus of the program, and before it is to be called the following parameters and the transitory cluster centre must be known:

ITM: the maximum number of iterations
ITMAX: an integer number connected with the maximum number of iterations by the statement
$$ITMAX \leqslant ITM \leqslant 2*ITMAX$$
NHOPE: the number of clusters required
DEVMAX: the maximum variation, in any cluster, of the reflectance digital number
DSL: the minimum distance between the cluster centres
NMIN: the minimum number of pixels in any cluster
C: the cluster centres
NC: the number of clusters to be adjusted while the program is running
IT: the count of iteration: the initial value of it must be zero.

The sequence of the procedure is:

Step 1. Count and check the number of iterations. If IT < ITM then go to step 2, otherwise print out the result and return.

Step 2. Assign each sampled pixel to the nearest temporary cluster centre.

Step 3. Compute the mean value of the co-ordinates of the measurement space in each cluster, and make it the new temporary cluster centre.

Step 4. Check the number of pixels in each cluster, cancel those clusters whose number of pixels is less than NMIN.

Step 5. Check the number of clusters. If it is too small (NC $\leqslant$ NHOPE/2), go to step 7 for cluster splitting; if it is too large (NC $\geqslant$ NHOPE), go to step 8 for cluster combining.

Step 6. Check the count of iteration. If it is an odd number, go to step 7, otherwise go to step 8.

Step 7. Calculate the standard deviation for I-th cluster at spectral band K - DEV(I,K), where
I = 1, 2, ...NC
K = 1, 2, 3, 4 (corresponding to the Landsat MSS band 4, 5, 6, 7) and check it. If DEV(I,K) $\geqslant$ DEVMAX, then split the I-th cluster. After that go to step 1.

Step 8. Calculate the distance between each pair of cluster centres - DST(I,J), where
I,J = 1, 2, 3, ...NC, I $\neq$ J
and check it. If DST(I,J) < DSL, then combine the two clusters (I-th to J-th clusters). After that go to step 1.

Figure 6. shows schematically the procedure.

After each call the output from the printer is produced: (Shown in Figure 7.).

The subroutine "AUTOCLASS" can be called as many times as needed, but it is necessary to change the parameters slightly each time in order to obtain a different group of values for the clusters. After that several groups of cluster centres will be available for comparing with one another. The best one will be used for further classification.

IV. OVERALL CLASSIFICATION

Having found the cluster centres, we can proceed with the classification of the whole data set. The program "CLASB" is used for this purpose. The flow chart is illustrated by Figure 8.

The output of the program is a classification map (see Figure 9.), using different colours (or characters) to indicate different classes of ground cover. Accompanying the map is a collection of four curves (see Figure 10.), one for each MSS band and plotted with different marks: ".", "-", "+", "*", representing bands 4, 5, 6 and 7 respectively. These curves show the variation of the reflectances of the pixels along the line (marked in Figure 9.) which is chosen arbitrarily before running the program. Pixels having similar reflectance are considered belonging to the same cluster.

In Figure 10. the reflectance of the first 4 pixels are very similar, and in the classification map of Figure 9. it can be found that the 4 pixels belong to the cluster marked by the character "F". Of course this kind of comparison could be done the other way: first look at the classification map, find the pixels marked by the same character, and then check them against the curves for the similarity in the reflectance. By doing so, we either accept the results as consistent or give them up because of unjustifiable discrepancies.

V. CONCLUSIONS

In a clustering procedure, flexibility and self-checking ability are very important. The former requirement is necessary because of the fact that the ground cover is so varied. The latter is indispensable as it is not always possible to obtain reliable reference data [3]. While the structure of the data set changes with time, season, weather, and location, the parameters used to describe them must also be changed accordingly. The problem of how to change them and how to select values for the parameters still remain to be solved. Due to the high computing cost it is not feasible to adopt the trial and error method, what can be done is to use the information obtained from the program "STATESW" to seek some initial approximate values of the parameters at the outset, and then make slight and systematic changes, and use the modified parameters to find the cluster centres.

VI. ACKNOWLEDGEMENT

The authors wish to record their thanks for the help and guidance they received from staff of the University of Aston. From Dr T Chidley and Miss Gwyneth Thomas, Department of Civil Engineering; and Dr P D Mallinson and Mr H Beswetherwick of the Computing Centre.

VII. REFERENCES

1 SWAIN, P.H., 1978. Fundamentals of pattern recognition in remote sensing. "Remote Sensing: The Quantitative Approach", McGraw-Hill, New York.

2 LILLESAND, T.M., KIEFER, R.W., 1979. Remote Sensing and image interpretation, John Wiley & Sons, New York.

3 HOFFER, R.M., 1978. Biological and physical considerations in applying computer aided analysis techniques to remote sensor data. "Remote Sensing: The Quantitative Approach", McGraw-Hill, New York.

AUTHOR BIOGRAPHICAL DATA

Dr. W.G. Collins is director of a remote sensing research unit at the University of Aston in Birmingham, United Kingdom which is working worldwide on the application of remote sensing. Author of over 60 papers and several journals, he serves on a number of national committees. Founder member of the Remote Sensing Society and its first Secretary and Journal Co-Author.

Mr. Yen Yi Chen is lecturer in the Department of Physics at the University of Shandong. His main interest is in data processing with particular emphasis on the use of satellite imagery for third world problems. He is keen to develop data processing methods which can be used in countries which do not yet have access to interactive image processing facilities.

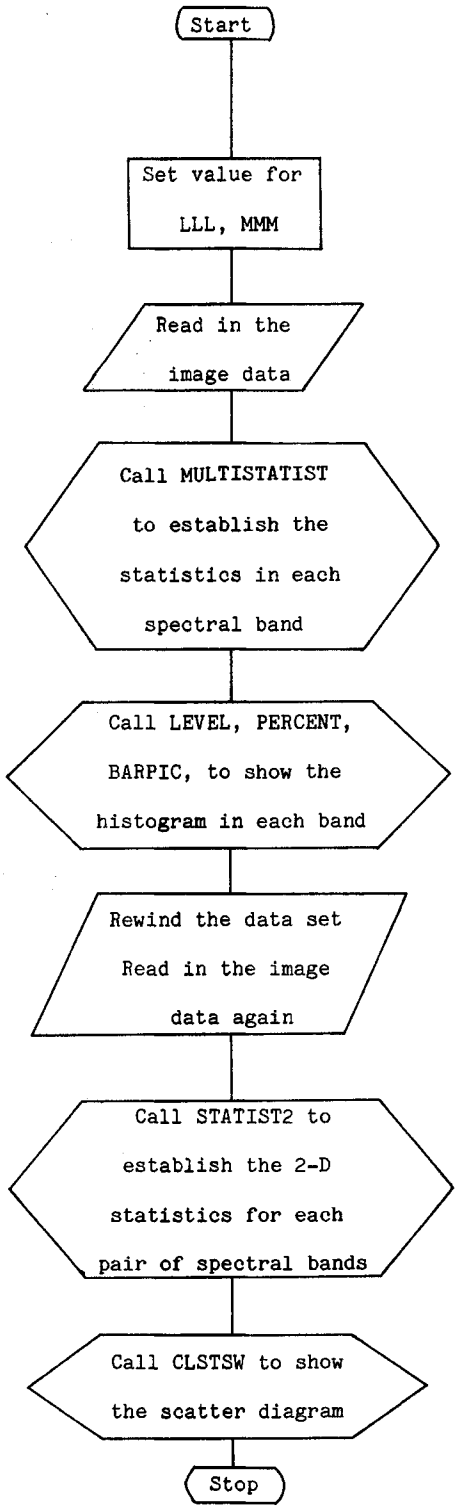


Figure 1. The flow chart of the program STATESW

THE NUMBER OF ALL PIXELS CONCERNED = 33408

RATIO=200.000

percentage of pixels

```
I=  1,A()=  0.000|*
I=  2,A()=  0.000|*
I=  3,A()=  0.000|*
I=  4,A()=  0.000|*
I=  5,A()=  0.000|*
I=  6,A()=  0.000|*
I=  7,A()=  0.000|*
I=  8,A()=  0.000|*
I=  9,A()=  0.001|*
I= 10,A()=  0.010|***
I= 11,A()=  0.012|***
I= 12,A()=  0.008|**
I= 13,A()=  0.004|*
I= 14,A()=  0.003|*
I= 15,A()=  0.006|**
I= 16,A()=  0.018|****
I= 17,A()=  0.036|********
I= 18,A()=  0.061|*************
I= 19,A()=  0.104|***********************
I= 20,A()=  0.144|*******************************
I= 21,A()=  0.123|*************************
I= 22,A()=  0.098|*********************
I= 23,A()=  0.077|****************
I= 24,A()=  0.073|***************
I= 25,A()=  0.057|************
I= 26,A()=  0.042|*********
I= 27,A()=  0.037|********
I= 28,A()=  0.030|******
I= 29,A()=  0.020|****
I= 30,A()=  0.013|***
I= 31,A()=  0.010|**
I= 32,A()=  0.007|**
I= 33,A()=  0.004|*
I= 34,A()=  0.002|*
I= 35,A()=  0.001|*
I= 36,A()=  0.000|*
I= 37,A()=  0.000|*
I= 38,A()=  0.000|*
I= 39,A()=  0.000|*
I= 40,A()=  0.000|*
I= 41,A()=  0.000|*
I= 42,A()=  0.000|*
I= 43,A()=  0.000|*
I= 44,A()=  0.000|*
```

BIG= 0.500 SMALL= 0.000

Digital number in band 6

Figure 2. The histogram of a small part of a Landsat image (band 6)

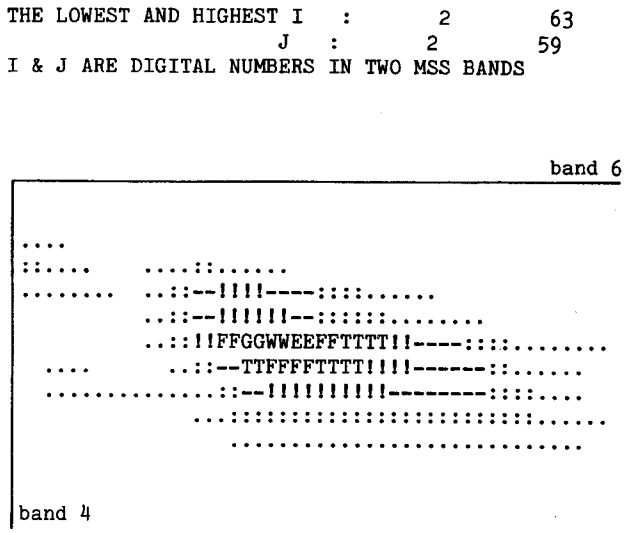


Figure 3. The scatter diagram of a small part of a Landsat image (band 4 and 6)

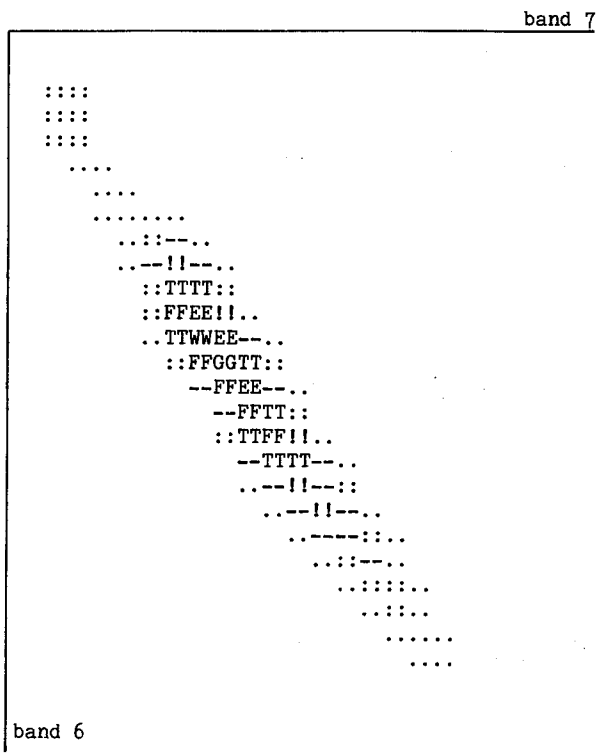


Figure 4. The strong correlation in Landsat band 6 and 7

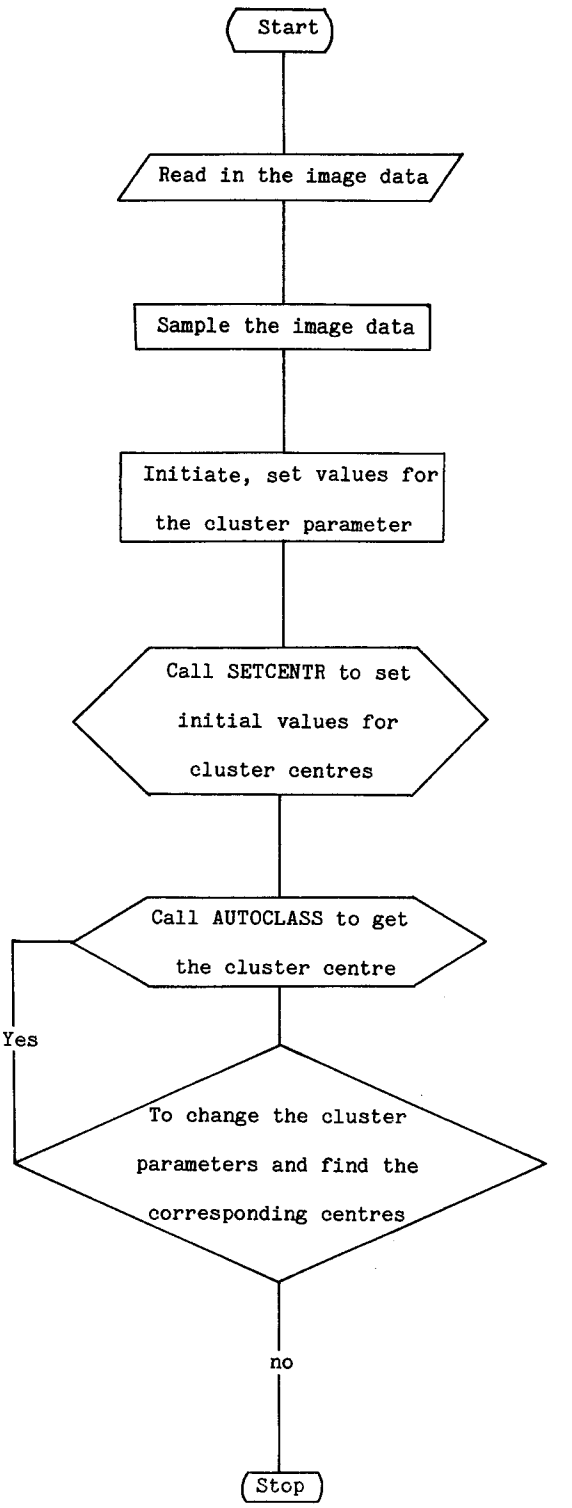


Figure 5. The flow chart to find the cluster centres

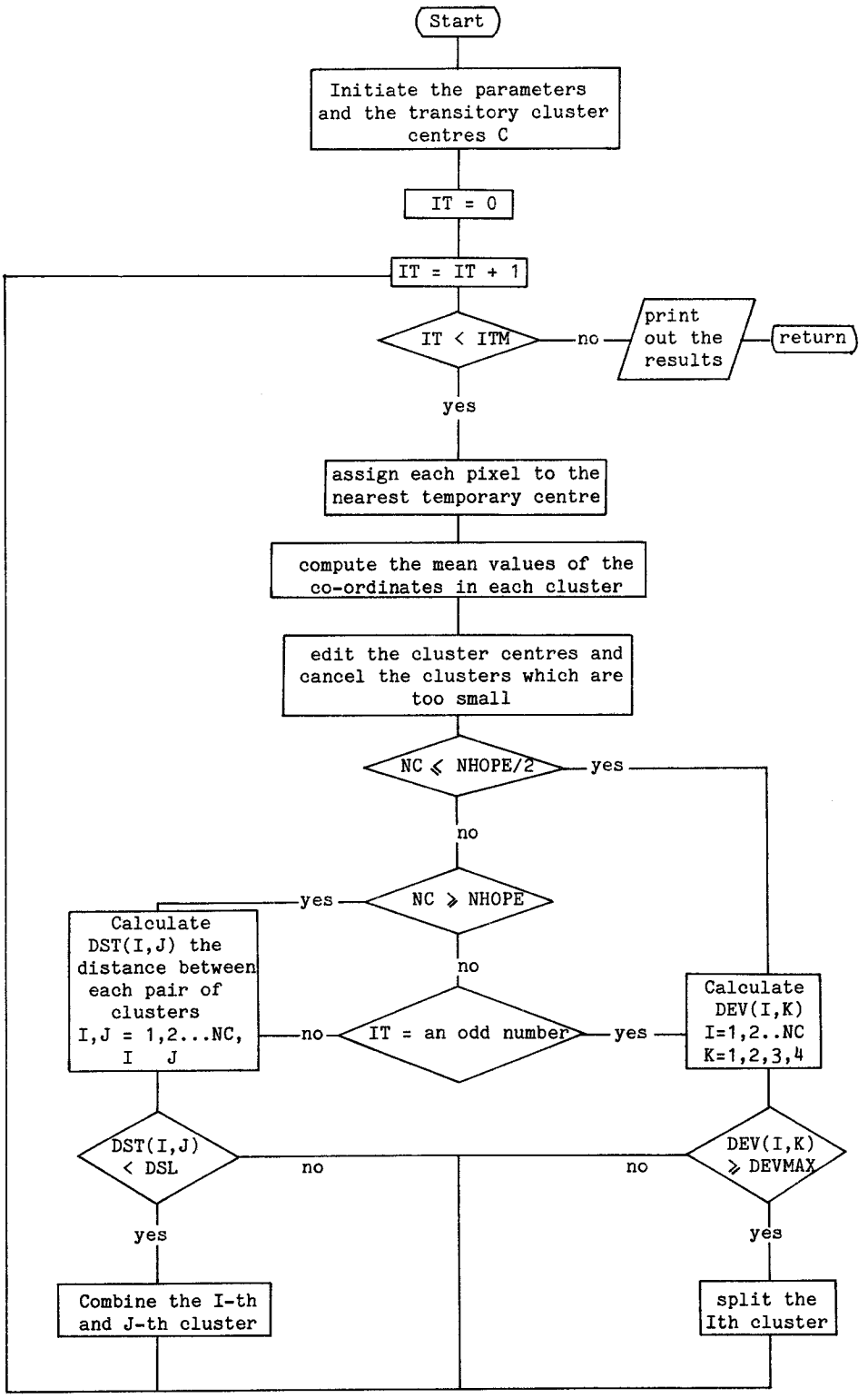


Figure 6. The iteration for adjusting the cluster centres

```
THE CENTRES OF ALL CLASSES
28.6   30.9   32.6   29.8   29.3   29.4   31.9   30.1   29.7   33.8
22.7   28.2   34.1   32.7   30.3   30.0   36.7   31.8   29.1   41.0
13.6   20.7   30.4   43.8   49.9   55.9   49.5   38.2   65.1   56.8
 2.2    4.3    8.7   18.6   23.1   26.9   21.3   14.4   31.7   25.1
         ---------10--------20--------30--------40--------50---------
BAND  4
        I                              19873A
        I                              19873A
        I                              19873A
        I                              19873A
BAND  5
        I                         1    29584 3 7    A
        I                         1    29584 3 7    A
        I                         1    29584 3 7    A
        I                         1    29584 3 7    A
BAND 6
        I              1     2     3     8    4     7    6A
        I              1     2     3     8    4     7    6A
        I              1     2     3     8    4     7    6A
        I              1     2     3     8    4     7    6A
BAND 7
        I 1 2   3    8    4  7 5 A6    9
        I 1 2   3    8    4  7 5 A6    9
        I 1 2   3    8    4  7 5 A6    9
        I 1 2   3    8    4  7 5 A6    9

          THE PIXEL NUMBER OF ALL CLASSES
 124    118    397    418    300    111    139    208     16     12

          THE DISTANCE BETWEEN GROUP CENTERS
 C1     C2     C3     C4     C5     C6     C7     C8     C9     C10    C11
1 0.    10.    22.    36.    43.    50.    43.    29.    60.    52.
2 0.     0.    12.    28.    35.    42.    34.    21.    52.    44.
3 0.     0.     0.    17.    25.    32.    23.    10.    42.    32.
4 0.     0.     0.     0.     8.    15.     8.     7.    25.    17.
5 0.     0.     0.     0.     0.     7.     7.    15.    18.    14.
6 0.     0.     0.     0.     0.     0.    11.    22.    10.    12.
7 0.     0.     0.     0.     0.     0.     0.    14.    20.    10.
8 0.     0.     0.     0.     0.     0.     0.     0.    32.    24.
9 0.     0.     0.     0.     0.     0.     0.     0.     0.    16.
  THE SMALLEST DISTANCE BETWEEN CENTRES DSL=   7.0
  THE MAXIMUM DEVIATION OF A CLUSTER DEVMAX=   4.0
  THE OVERALL AVERAGE DISTANCE TO  CENTRES DAL=   3.2

          THE AVERAGE DISTANCE OF GROUP
2.3    3.6    2.9    3.3    3.1    3.3    3.7    3.8    3.3    4.0

          THE STANDARD DEVIATION
1.2    1.4    1.3    1.5    1.1    1.1    1.4    1.6    1.0    1.0
1.7    1.9    1.6    2.4    1.9    2.0    2.2    2.5    1.4    2.9
1.5    2.8    2.1    1.7    1.9    2.3    2.4    2.2    2.3    2.4
0.5    1.3    1.4    1.4    1.5    1.4    1.8    1.7    2.0    1.8

  THE UNCLASSIFIED PIXEL NUMBER=  13     0.7PERCENT OF THE WHOLE
```

Figure 7. The cluster centres and the distances between them.

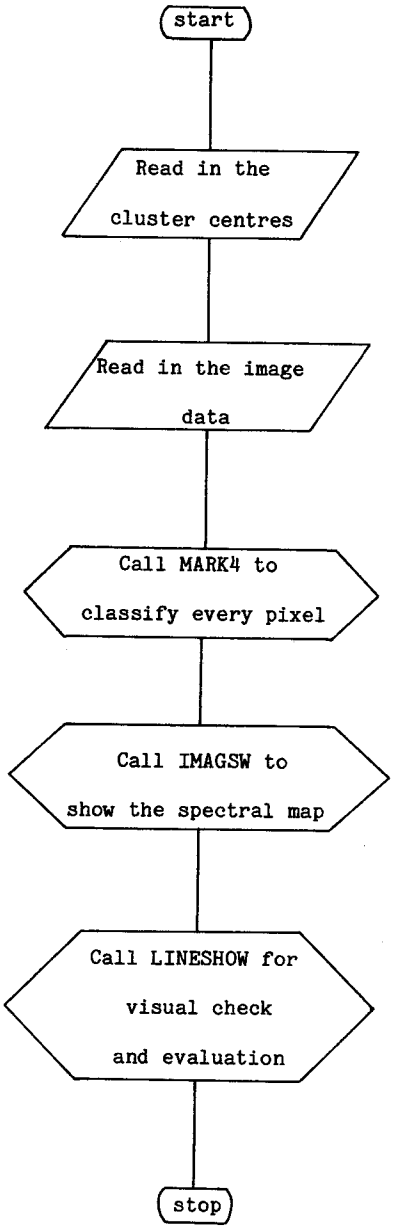
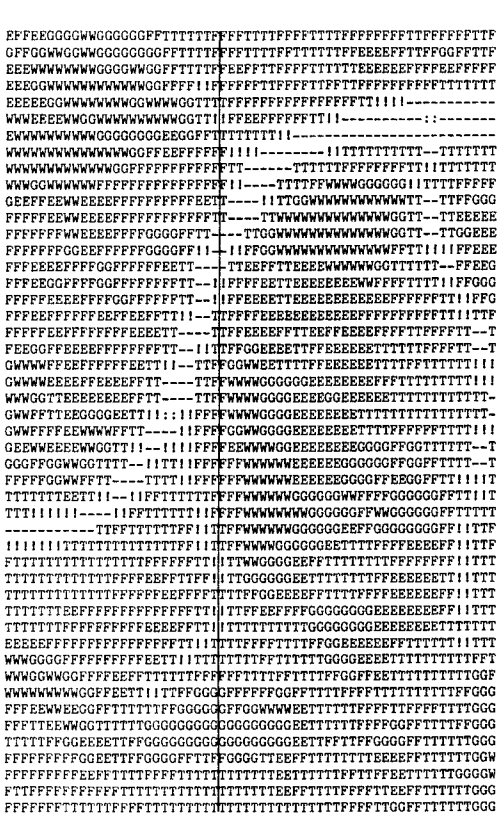


Figure 8. The flow chart for classification.



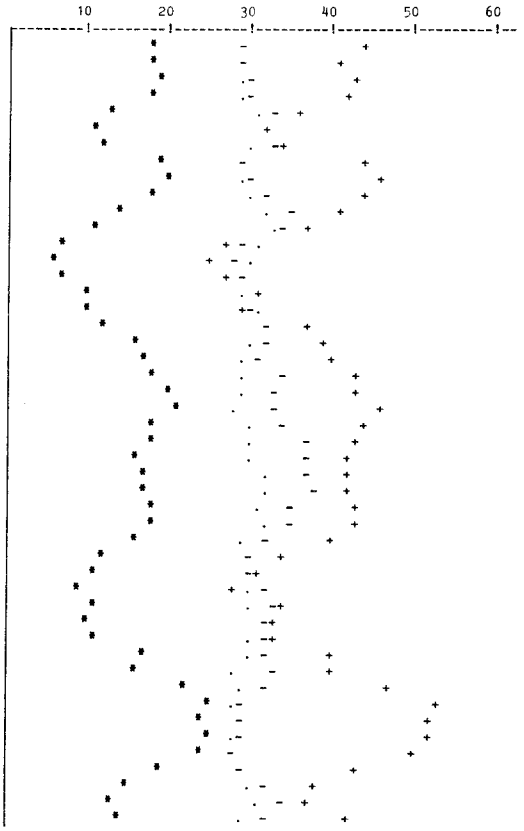


Figure 9. An unsupervised classification map containing a marked line.

Figure 10. The variations along the marked line of the reflectance in the four Landsat bands.