CLUSTER ANALYSIS OF LANDSAT TM DATA

Masatoshi Mori and Keinosuke Gotoh[†]

Department of Management Engineering, Kinki University, Iizuka 820, Japan

[†]Department of Civil Engineering, Nagasaki University, Nagasaki 852, Japan

PURPOSE:

Cluster analysis system by using Statistical Analysis System (SAS) is constructed. The analysis system is applied to Landsat TM data. The six spectral bands except the thermal band of TM data are used for analysis. Three hierarchical algorithms of cluster analysis, Ward method, average linkage, and centroid method, are prepared in the framework of SAS cluster. The results by three algorithms are examined in relation to classification. The classified map by Ward method is the most accurate among them, and is practical. On the other hand, the others are almost same and incorrect.

KEYWORDS: Cluster analysis, Image processing, Landsat TM data, SAS

1. INTRODUCTION

In use of Landsat Thematic Mapper (TM) Data, cluster analysis as an unsupervised classification scheme is one of the most useful methods. Specially, in the case of being not clear of text area in objective image, cluster analysis is effective and powerful (Koontz, 1976). A difficulty in using of cluster analysis is, however, existing in extending spectral band capacity. There are seven spectral bands in Landsat TM sensors. Each of these sensors has a dynamic range with 256 levels (8 bits) nominally. In the first stage of cluster analysis we construct a multi-dimensional feature space (Goldberg, 1978). If using all the seven spectral data of TM in analysis, a seven-dimensional feature space is necessary, which needs 256^7 bits of computer memory (Wharton, 1983). It is, however, impossible to construct this feature space in actuality. Then, as conventional method of clustering, preprocess of sampling has been adopted. In this method, before analyzing data, several ~ several hundred points are selected from original image data. Then, these selected points are directly analyzed by using clustering scheme, and classified to land cover map. Finally every point of original data is assigned to the nearest classified point. As every point of original data is not analyzed directly by this procedure, incorrect classification in assigning points to the resultant class may occur. The reason is that the Euclid metric, not isotropic, is used in assigning, category area is not always an ellipsoid in the multi-dimensional feature space, so erroneous assigning may occur still.

In the present paper we consider a direct analysis of every point of original data of Landsat TM, without preprocessing for sampling. We use Statistical Analysis System (SAS) (Joyner, 1985) so as to complete this analysis system. By using SAS system, we can process binary data of multi-dimensional image. However, as the direct result from SAS is a set of records, it is difficult to reform the record format to binary data of classified image. Moreover, because of direct analysis system, a number of pixels of original image are limited. We try to examine three algorithms of cluster analysis, Ward method, average linkage, and centroid method, compared with resultant classified image. The result using TM data is shown precisely.

2. CLUSTER ANALYSIS SYSTEM

2.1 Image data

In the present study, we used Landsat-4 TM data (the scene: 113-37, obtained on May 22, 1984). The square image of 120×120 pixels was cut off from the scene, which is shown in Fig.1. This image is a part of Fukuoka City in Japan, and contains several typical features of land such as the sea, rivers, ponds, residential, commercial, bare soil, roads, grass, etc. These typical categories are very important for checking cluster algorithms. We adopted six spectral bands except the thermal (the 6th) band for the analysis system, because the spacial resolution of the thermal band is 120m, which is different from that of the others, so we excluded the thermal band. Then we construct the six-dimensional feature space of TM data.

2.2 SAS/cluster procedure

The three algorithms of SAS/cluster procedures are prepared and compared with one another, which are Ward method, average linkage, and centroid method. They are all hierarchical methods. The jobs of clustering are carried out on IBM 3081 computer and FACOM M-1800 computer systems. The CPU time of SAS/cluster procedure is about 11 minutes for three algorithms, but the jobs need more than 9 mega bytes in computer memory.

(a) Ward method
By Ward method clusters are joined so as to maximize the likelihood at each level of the hierarchy. Ward method tends to join clusters with a small munber of observation and is strongly biased toward producing clusters with roughly the same number of observations.

(b) average linkage
In average linkage the distance between two clusters is the average distance between pairs of observations, one on each cluster. Average linkage tends to join clusters with small variances and is slightly biased toward producing clusters with the same variance.

(c) centroid method
In centroid method the distance between two clusters is defined as Euclidean distance between their centroid or means. The centroid method is more robust to outliers than most other hierarchical methods but in other aspects may not perform as well as Ward method and average linkage.

2.3 Reconstruction to image data

For direct analysis of image data by SAS, the six image files of TM data are allocated to device names, respectively. After the process of clustering procedure, the sample results by Ward method as shown in Table I are obtained. In the case of Table I, initially it starts with 1000 pixels, which are equal to 1000 clusters. Each record of the result means the one process of joining two clusters to one. In the right column the frequency means a number of pixels belonging to a new cluster. Finally one cluster is obtained as known

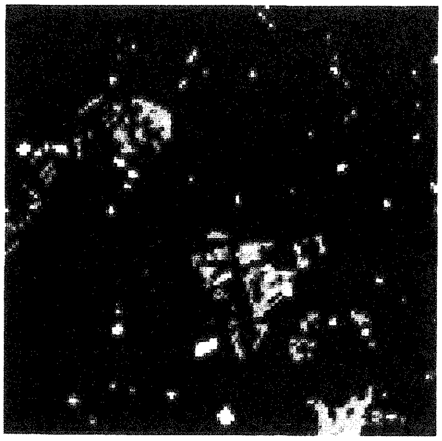


Fig.1. Color composite image of Landsat TM (the scene: 113-37, obtained on May 22, 1984).

in Table I.

Then, we can obtain an expected number of clusters by cluster procedure. In the present study we examined the case of eight clusters for three algorithms. In order to reconstruct classified image from these records, we must know which one among eight clusters esch pixel of image data belongs to. Analysis of this tree network gives the relation between pixels and cluster numbers. Thus we can obtain a final classified image. This reconstruction process is coded in FORTRAN77.

3. RESULTS OF CLASSIFICATION

3.1 Classification land cover map

Results of classification by three algorithms are shown in Fig.2 (a)~(c), respectively. Also the rates of categories are shown. These map were obtained by reconstruction procedure described in Section 2. Each color used in maps is assigned to the most suitable category. The field survay has been performed for checking accuracy of classification.

3.2 Ward method

The result by Ward method is shown in Fig.2 (a). As known from Fig.2 (a)~(c), this classification is most accurate among three algorithms. Eight categories classified in the procedure are water, river and shallows, grass, wood, bare soil, residential, commercial, and asphalted area. Water area contains the sea and a pond with salt water. Wood occupies verdart parks. Bare soil contains playgrounds of school and open area being developed. Asphalted area contains roads, yards, and squares. Although these area are comparatively accurate, residential area contains unclassified area partly.

3.3 average linkage

The result by average linkage is shown in Fig.2 (b). The classification in land area is incorrect. The main land cover gives commercial, residential, grass, and wood, which are joined to one cluster. Only bare soil is comparatively correct. Bare soil 1~4 are essentially the same category. However, these bare soil 1~4 have been separated due to the characteristics of this algorithm.

Table I. A sample of the result of SAS cluster procedure by Ward method.

Number of clusters	Cluster	jointed	Frequency of new cluster
:	:	:	:
21	CL48	CL58	55
20	CL45	CL37	50
19	CL34	CL33	189
18	CL23	CL46	64
17	CL36	CL30	48
16	CL25	CL41	377
15	CL39	CL26	25
14	CL24	CL54	10
13	CL20	CL31	73
12	CL28	CL29	102
11	CL16	CL19	566
10	CL18	CL17	112
9	CL27	CL21	90
8	CL15	CL35	35
7	CL13	CL22	85
6	CL7	CL10	197
5	CL9	CL12	192
4	CL8	CL14	45
3	CL6	CL4	242
2	CL11	CL5	758
1	CL2	CL3	1000

3.4 centroid method

The land cover map by centroid method is shown in Fig.2 (c). The result is almost same as that by average linkage. The main cluster contains almost all categories in land, which is more remarkable than the case by average linkage.

4. CONCLUSION

The cluster analysis system by using SAS has been constructed, and was applied to Landsat TM data. Three algorithms of cluster analysis, Ward method, average linkage, and centroid method, were examined and compared with one another. The classification result by Ward method is most accurate among them. The results by average linkage and centroid method are almost same, but average linkage is somewhat better. However, the separation between water region and land area is rather accurate in classifications by three algorithms naturally.

Although we can analyze six-dimensional image data at one procedure, the image size is limited to around 120×120 pixels due to the framework of SAS/cluster. This size is, however, not sufficientlly practical. One more problem is that three algorithms are all hierarchical. Non-hierarchical algorithms such as the k-means and iso data method are not included in SAS cluster system.

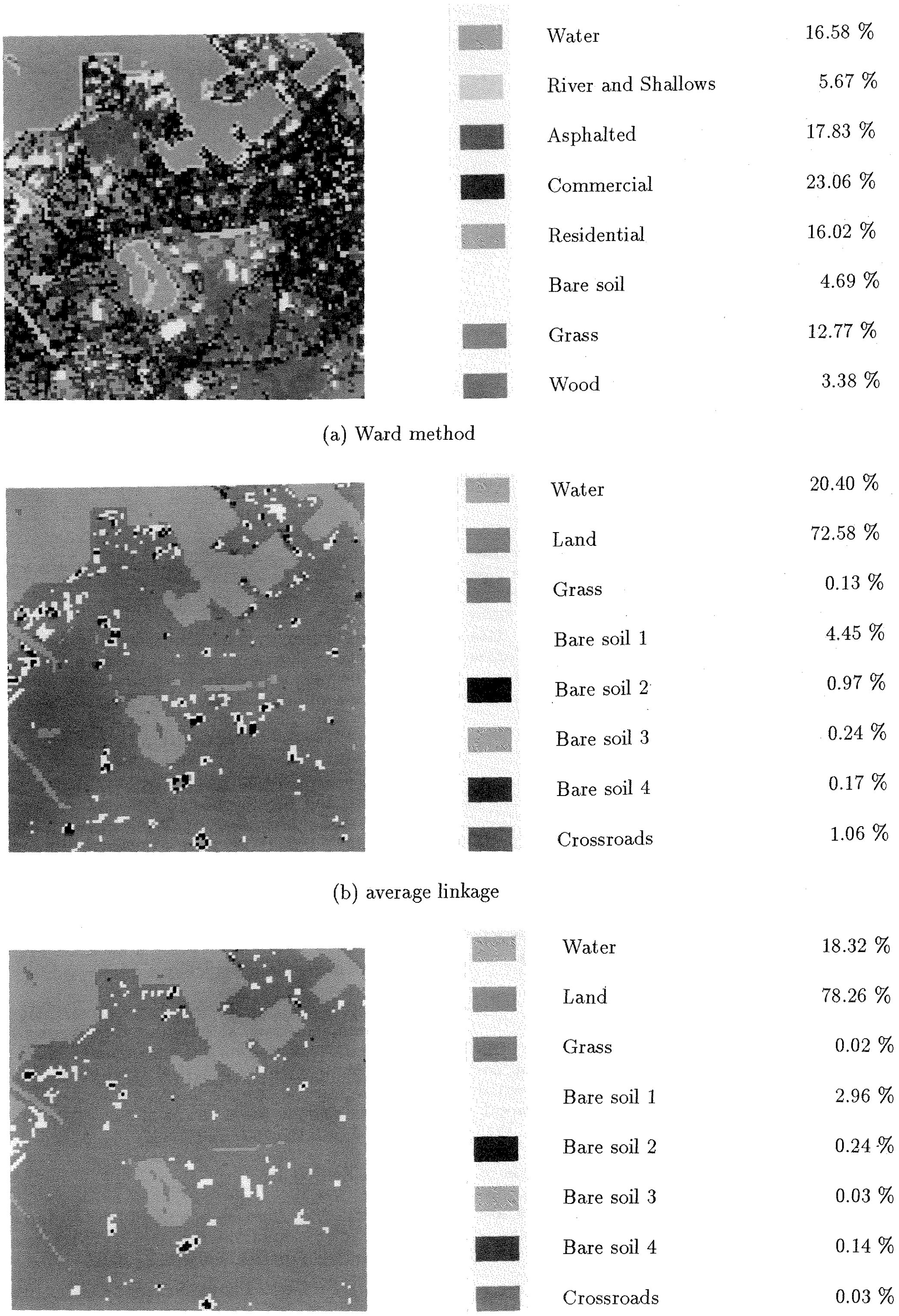
REREFERENCES

Goldberg, S.W., and Shilien, S., 1978. A clustering scheme for multidimensional images. IEEE Transaction on system, man, and cybernetics, SMC-8(2):86-92.

Joyner, S.P., 1985. The SAS User's Guide: Stasistics, Version 5 Edition, SAS Institute Inc.

Koontz, W.L.G., Narendra, P., and Fukunaga, K., 1976. A graph-theoretic approach to nonparametric cluster analysis. IEEE Transaction on computers, C-25(9):936-944.

Wharton, S.W., 1983. A generalized histogram clustering scheme for multidimensional image data. Pattern Recognition, 16(2):193-199.



(a) Ward method

(b) average linkage

(c) centroid method

Fig.2. Results of cluster analysis of Fig.1 by three algorithms, which show classification maps, categories, and rates of occupation.