MEC

Report LARS 060679 Report (1979)

CLUSTERING OF LANDSAT MSS DATA: CERTAIN LIMITATIONS

Luis A. Bartolucci*

and

Ramon Bermudez de Castro**

Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana 47906

Abstract

During this investigation the performance of the LARSYS clustering algorithm has been studied under different conditions. Certain limitations of this algorithm when applied to Landsat MSS data have become apparent, including: 1) the dependence of the clustering performance on the degree of data correlation, 2) the undesired variability of the clustering output when the clustered area is shifted slightly, and 3) the anomalous performance of the clustering algorithm when differential scaling factors are applied to the different axes of the feature space. A number of recommendations are suggested to overcome these limitations.

* Technical Director, Technology Transfer Programs, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana 47906.

** Research Engineer at the Instituto Geográfico Nacional, Madrid, Spain.

Introduction

Since the first digital analyses of the Landsat MSS data were conducted soon after the launch of the Landsat-1 satellite in July 1972, a great deal of research has been carried out to develop, test, and utilize numerical analysis techniques that could be effectively applied to this particular type of data (NASA, 1972). It was soon recognized that the supervised method of developing the statistics for training a classifier (Swain, 1978) was not an adequate means of defining the natural multispectral groupings present in a Landsat data set. The supervised approach did not allow a satisfactory definition of an important type of spectral training classes that represents a large percentage of the total Landsat scene, i.e. the "spectral mixture classes" (Bartolucci, 1979). Therefore, to overcome these limitations of the supervised approach, researchers at LARS have been regularly using a non-supervised procedure to determine the inherent structure of the Landsat data by defining the training statistical parameters through use of a clustering algorithm. However, certain limitations in the LARSYS clustering function have been found when this processor is applied to the Landsat MSS data. This paper illustrates some of these limitations, and the authors propose a possible solution.

The LARSYS Clustering Algorithm

A detailed description of the LARSYS clustering function (*CLUSTER) is given elsewhere by Wacker (1969), Wacker and Landgrebe (1970) and (1971), and Phillips (1973). A summary of the most important features of this algorithm will be described here.

Since the spectral response of every Landsat MSS spatial resolution element (data point) could be represented by a vector in a four-dimensional space, a set of Landsat MSS data could be visualized as a distribution of points in this hyperspace. A clustering algorithm can be used to find a natural grouping of these vectors which possesses strong internal similarities, thus describing the intrinsic structure of a data set.

The ISODATA-type clustering function (Duda and Hart, 1973) implemented in the LARS data processing system (LARSYS) may be described briefly in terms of the following essential steps:

1. Initialization. Given a set of Landsat four-dimensional vectors from a subarea of the image to be analyzed, [1] and a maximum number of classes (N) specified by the analyst, the clustering algorithm defines the N initial cluster centers as follows:

- It computes the mean (μ_i) and variance (σ_i) of the total number of vectors for each one of the four dimensions.

- It constructs a rectangular parallelepiped in the four-dimensional space with vertices defined by $\mu_i \pm \sigma_i$ for each one of the four dimensions.

- It selects the N initial cluster centers in such a way that they are equidistant along the diagonal of the parallelepiped formed by a line connecting the $\mu_i - \sigma_i (i=1,...,4)$ and $\mu_i + \sigma_i \ (i=1,...,4)$ of the parallelepiped.

2. Migration of Cluster Centers. In this step each vector is assigned to its nearest cluster center using a Euclidean distance measure, then the means and covariances of the resulting clusters are computed. If the new means are not equal to the previously computed cluster centers, then the new means are used as cluster centers and this step is repeated. Otherwise, the processing stops.

Limitations of Clustering Landsat Data

Although the LARSYS clustering processor is an indispensable tool to be used by an analyst to adequately define the spectral training classes of a Landsat MSS data set, certain critical factors regarding the application of this algorithm to analyze Landsat MSS data must be considered to insure optimal partitioning of the feature space. The analyst must be aware of the following three limitations of the clustering algorithm:

1. Because of the manner in which this cluster function defines the diagonal of the n-dimensional parallelepiped (as described in the preceding section), the performance of this function will not be as desired in cases when the data features are negatively correlated.

2. It is critical for the analyst to select not only an appropriate number of initial cluster centers but also to bear in mind that the location relative to a particular ground cover

[1] To date, there is not a rigorous method of determining *a priori* the number of initial cluster centers to be input to the clustering function. It is usually left up to the analyst's experience.

feature of an area to be clustered may have a strong influence on the characteristics of the cluster class representing the ground cover feature of interest.

3. It is extremely important that the analyst be aware of the properties regarding invariance of cluster algorithms that utilize the Euclidean distance as a measure of similarity, i.e., that such clustering algorithms are not invariant to rescaling of the feature space axes.

The first of these limitations occurs when the features are negatively correlated in which case the diagonal of the parallelepiped is not in the direction of the axis of maximum variability. Consequently, a number of the initial cluster centers are likely to fall outside the distribution of the data as illustrated in Figure 1. Note in this figure that two of the initial cluster centers indicated by the symbols K and L do not have any data points in their neighborhoods and therefore the clustering algorithm will be forced to reduce the number of initially given cluster centers.

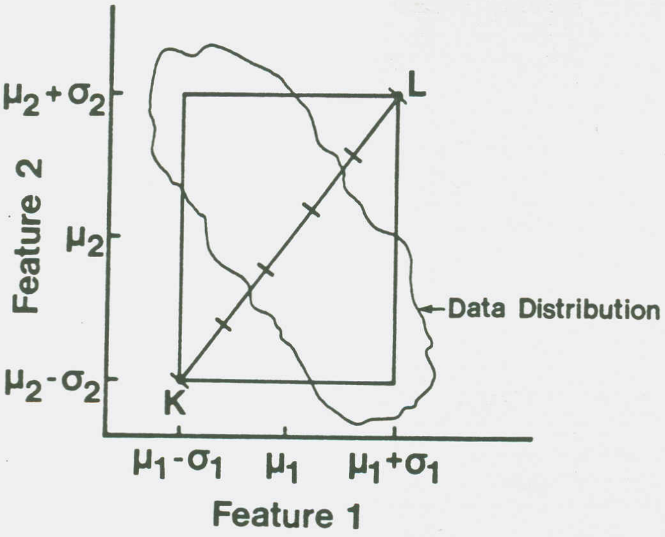

Figure 1. Hypothetical data distribution in a two-dimensional space showing the procedure by which the LARSYS clustering algorithm selects the initial cluster centers for a negatively correlated data set.

Should the features be positively correlated, as shown in Figure 2, the diagonal of the parallelepiped will be in the direction of the axis of maximum variability and therefore it is more likely that all the initial cluster centers will be contained within the distribution of the data.

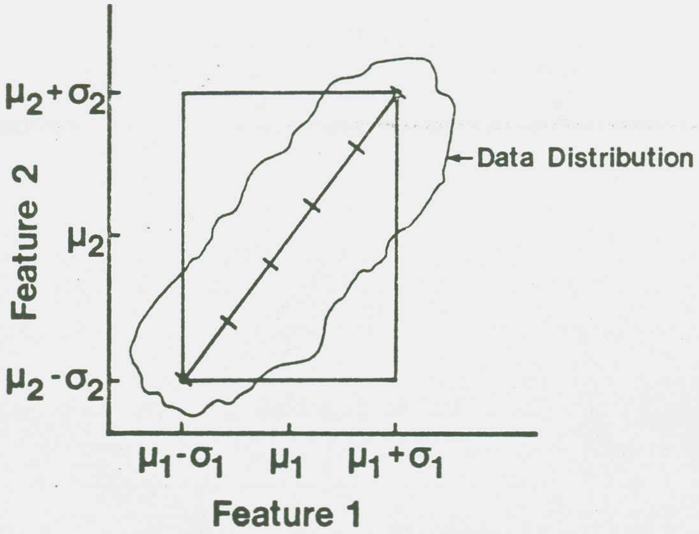

Figure 2. Hypothetical data distribution in a two-dimensional space showing the procedure by which the LARSYS clustering algorithm selects the initial cluster centers for a positively correlated data set.

The second limitation is the possible omission of certain ground cover classes of interest which might occur because of the change in relative distribution of data values in the measurement space when the location of the clustering area is shifted by a number of lines and columns, even though the total number of points and number of classes are left unchanged as shown in Figures 3 and 4. This limitation can be clearly observed when comparing the classes (within the rectangular subarea) obtained as a result of a slight shift in the location of the clustered area. Figure 3 shows two important classes of water defined by the symbols O and Q which represent accurately the state of nature in the scene, i.e., different turbidity levels, as is evident from the reference data (aerial photograph) shown in Figure 5. These two classes were not defined on the clustering output shown in Figure 4.

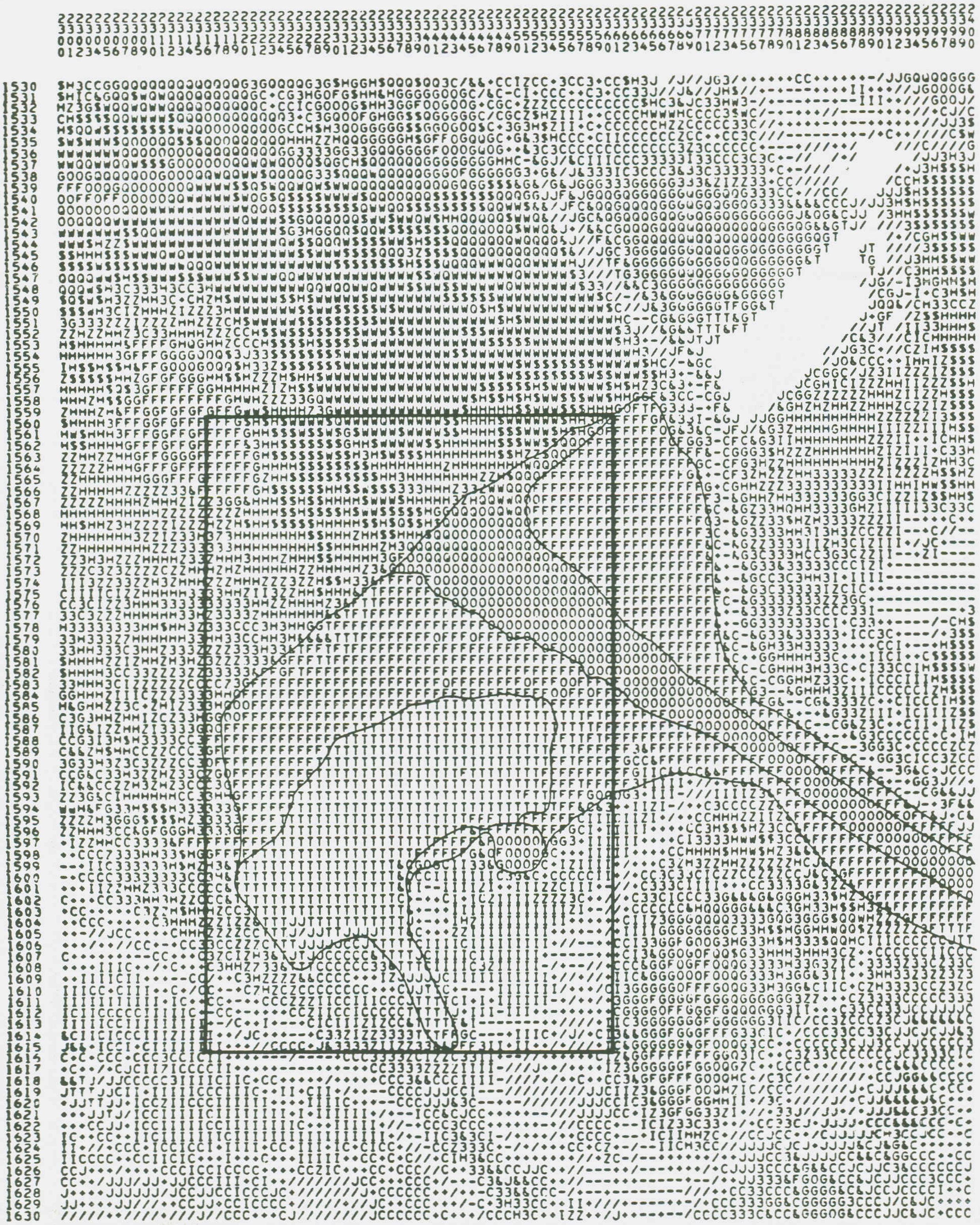

Figure 3. Cluster map of 10,201 Landsat MSS data points and 18 classes. The rectangular area is for comparison with the corresponding area in Figures 4 and 5.

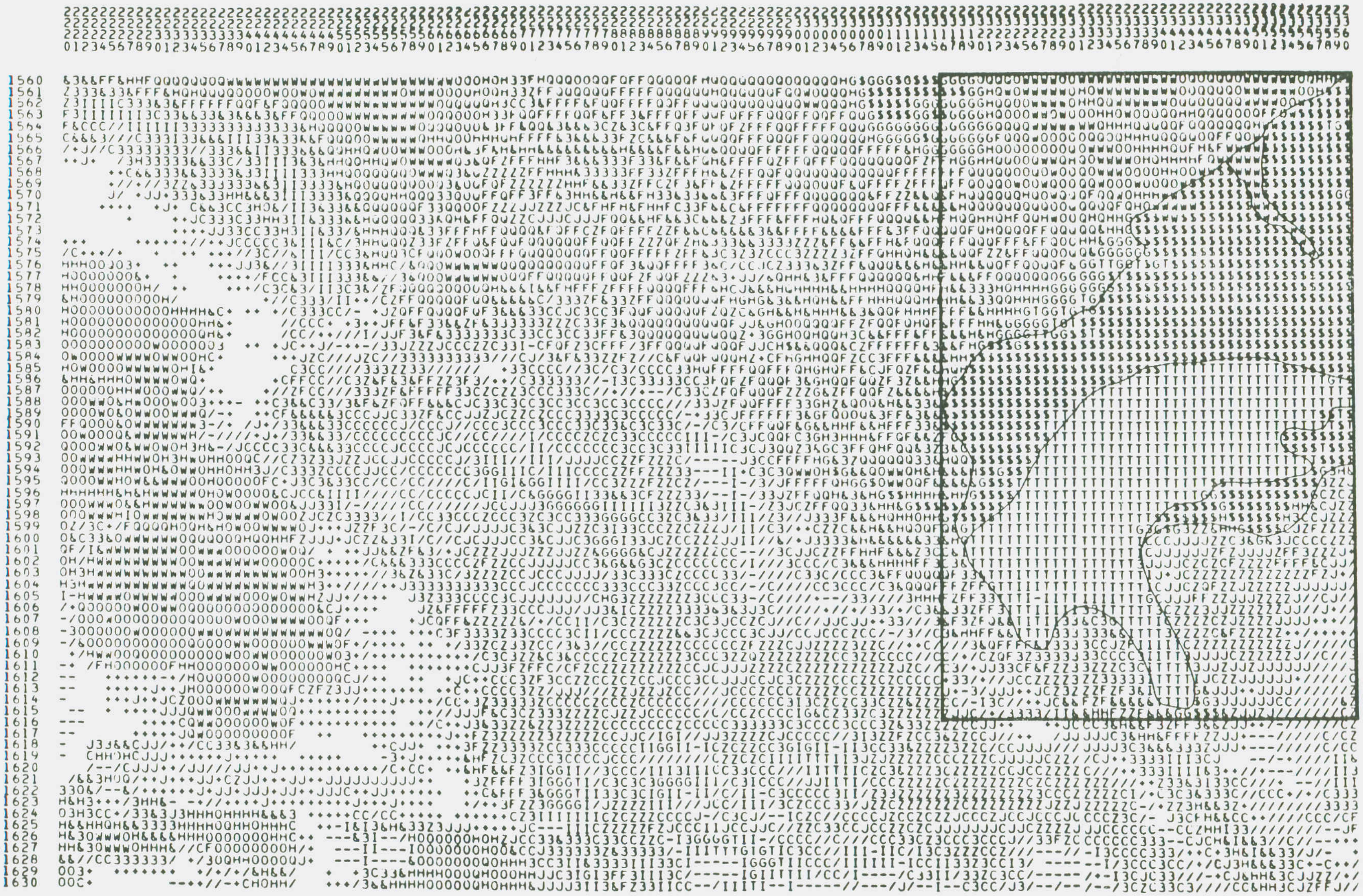

Figure 4. Cluster map of 10,201 Landsat MSS data points and 18 classes. The rectangular area is for comparison with the corresponding area in Figures 3 and 5.

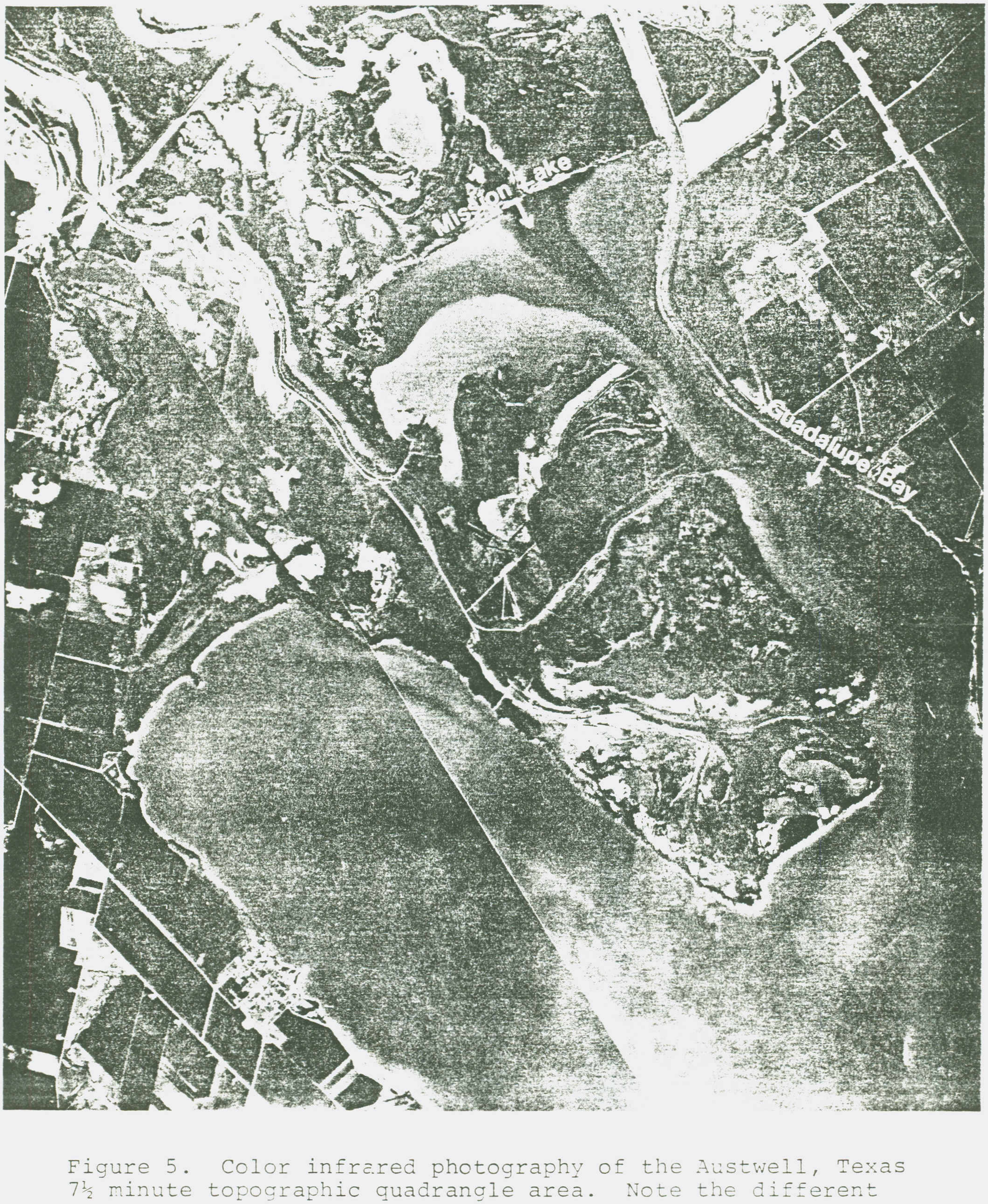

Figure 5. Color infrared photography of the Austwell, Texas 7½ minute topographic quadrangle area. Note the different water classes in the Mission Lake and Guadalupe Bay. The rectangular area is for comparison with the corresponding area in Figures 3 and 4. Features 1 (factory) and 3 (turbid water pond) are outlined for comparison with Figures 7 and 8.

The third and strongest limitation of this clustering algorithm is encountered when differential changes of scale are applied to the axes of the Landsat four-dimensional data. For example, this situation occurs when the Landsat MSS data is expressed in units of "inband radiance", i.e., calibrated Landsat MSS data. Bartolucci (1978) has shown the importance of calibrating the Landsat MSS data as a valuable analysis aid in relating the unsupervised spectral training classes generated by a clustering processor to the physical characteristics of the ground cover types present in the scene.

In an attempt to carry out a complete analysis sequence using calibrated Landsat MSS data, the authors found that the differential rescaling of the axes of the four-dimensional space due to the calibration did produce different results (Figure 6) than the clustering of the uncalibrated original data set which is shown in Figure 7. Note in the calibrated clustering map (Figure 6) that only one class of water was distinguished in the elongated water body. On the other hand, the uncalibrated clustering results show three spectral classes within the water body. Another interesting result that can be observed in the cluster class statistics of the calibrated data is that some of the cluster classes have a distribution in band 7 (LARSYS Channel 4) with a very small (almost zero) variance. Finally, a comparison of the classification results from both uncalibrated and calibrated data indicates that the classification of the calibrated data does not represent the scene as accurately as the classification of the uncalibrated data set. These results are no surprise considering that this clustering algorithm which uses the Euclidean distance as a measure of similarity implies that the feature space is isotropic and thus the performance of the clustering algorithm is invariant to the translations and rotations of the feature space axes, but it is not invariant to general linear transformations, such as axes rescaling (Duda and Hart, 1973). In calibrating the Landsat MSS data, one is essentially applying a different linear transformation to each one of the four axes which produces a distortion of the feature space, and hence significant changes in the distance relationship among all data points are introduced. (The authors have verified that rescaling of the four Landsat feature space axes by applying the same linear transformation to each of the four axes does not distort the feature space as perceived by the clustering algorithm, and therefore the resulting cluster classes correspond exactly to the cluster classes obtained from clustering the original scaled data.)

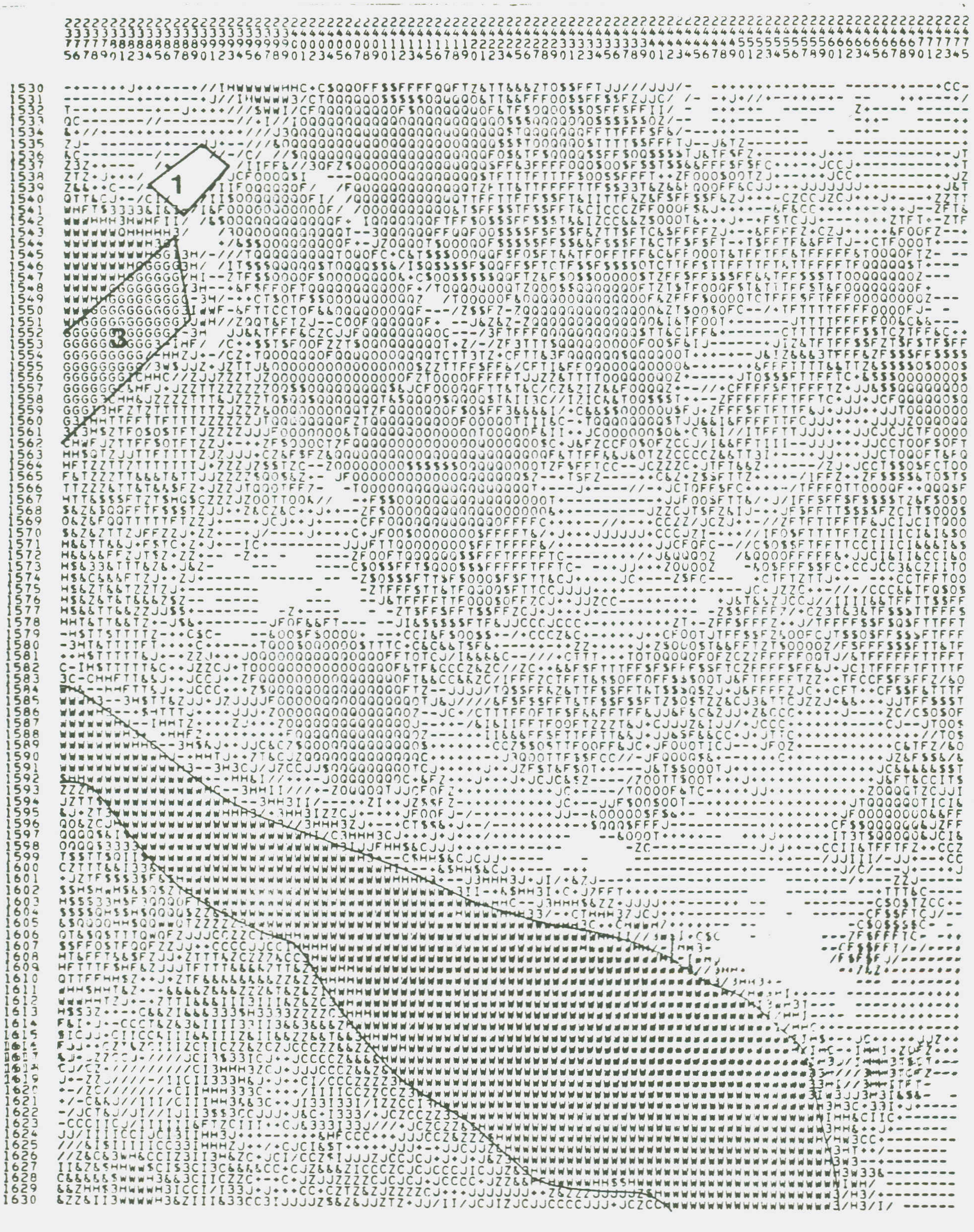

Figure 6. Cluster map of 10,201 Landsat calibrated data points and 18 classes.

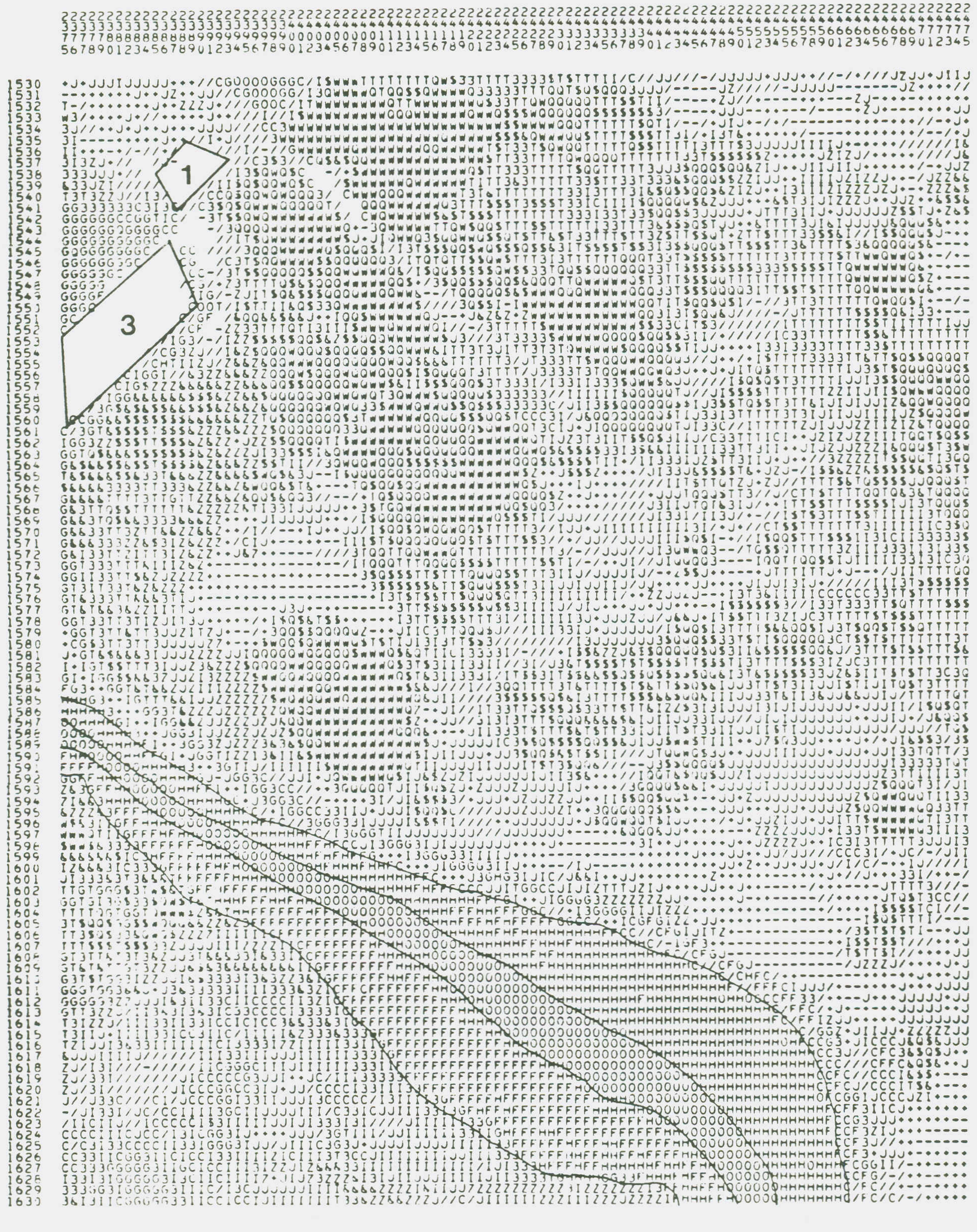

Figure 7. Cluster map of 10,201 Landsat uncalibrated data points and 18 classes.

From these results one might expect that the performance of this clustering algorithm when applied to Landsat MSS data should improve as the scales of the four axes become more similar, approaching an optimum situation in which all the axes of the feature space have an equal dynamic range. In order to equalize the dynamic range of the four Landsat MSS dimensions, one would have to either expand (multiply by a factor of two) the digital values of band 7 (LARSYS Channel 4) to simulate a full scale dynamic range of 7 bits (0-127 digital counts) equal to the dynamic range of bands 4, 5 and 6, or to compress the range of bands 4, 5 and 6 by a factor of two, leaving the original band 7 unaltered.

As part of this investigation, several linear transformations were applied to the Landsat MSS data, including:

1) expansion of band 7 (original dynamic range of 0-63) by a factor of two and leaving bands 4,5 and 6 unaltered,[2]

2) compression of bands 4, 5 and 6 by a factor of two and leaving band 7 unaltered,

3) expansion of all four bands by the same factor (multiplied by 2, 3, 4...),

4) compression of all four bands by the same factor (dividing by two).

Of these four transformations, the third one did not produce cluster classes different from those obtained from clustering the origianl (untransformed) data. Transformations 1, 2 and 4 changed the internal structure of the data distribution considerably. Consequently, the spectral classes resulting from clustering these transformed data sets were different from the spectral classes obtained from clustering the original Landsat MSS untransformed data. The results of applying the clustering algorithm to data sets that have undergone a linear compression in bands 4, 5 and 6 show that by compressing these bands linearly a great deal of the information content in the data set has been lost, as is to be expected if one considers that the compression of the data in the spacecraft and its subsequent decompression at the ground processing facilities (Goddard Space Flight Center) is accomplished through a quasilogarithmic transformation (NASA, 1972).

Out of these four transformations, only the first one, i.e., the expansion of the dynamic range of band 7 by a factor of two and leaving bands 4, 5,and 6 unaltered, allowed the clustering

[2] In order to perform this expansion using LARSYS, one has to utilize the following CHANNELS control card:

```
CHANNELS 1,2,3,4 (4/0,128/)
```

algorithm to define spectral classes that more accurately represent the ground cover types in the scene. Figure 8 shows the output of clustering a Landsat MSS data set which has undergone a linear expansion of the dynamic range of band 7 and left bands 4, 5 and 6 unaltered. This clustered area is the same as the area shown in Figure 7 which is the output of clustering an original (untransformed) Landsat MSS data set. A comparison of the cluster outputs shown in Figures 7 and 8 with the reference photography (Figure 5) clearly indicates that the spectral cluster classes obtained from the expanded data set (Figure 8) represent more accurately the ground cover types in the scene. It is extremely important to note that features "3" and "1" in Figure 7 have been clustered into the same spectral class, although the reference infrared photography (Figure 5) shows that these two features are definitely different cover types. Feature 3 is a turbid water pond, whereas feature 1 is a large factory.[3] These results obviously show the limitations of the clustering algorithm when applied to the original (unexpanded) Landsat MSS data set. On the other hand, note features 3 and 1 in Figure 8 that have been clustered into two different spectral classes which accurately represent the two different ground cover types present in the scene (Figure 5). The spectral separability of these two distinct cluster classes as measured by the Transformed Divergence (Swain and Wacker, 1971; and Swain and King, 1973) indicates that these two classes are completely separable since they have a pairwise transformed divergence value of 2000. These results clearly show that the clustering algorithm was unable to distinguish these two very spectrally different classes when applied to the unexpanded data set, whereas these two features were accurately differentiated by the same clustering algorithm applied to the expanded data. The authors have also observed that when using expanded data, the clustering algorithm is capable of differentiating a larger number of spectral classes without decreasing the overall spectral separability than when using the unexpanded data. In fact, it has also been observed that the minimum pairwise separability of the classes (the same number of classes) obtained from clustering the expanded data is always larger than that obtained from clustering the unexpanded data. In addition, the variation of the cluster output due to slight area shifts when using the original

[3] The turbid water and the factory with associated parking lots have similar spectral responses in bands 4, 5 and 6, and a very different spectral response in band 7. However, in the original (untransformed) data set, band 7 has a range of 0-63 gray levels, i.e., one half the range (scale) of bands 4, 5 and 6, and consequently spectral differences in band 7 contribute (weigh) only half as much as those in bands 4, 5 and 6. If the analyst desires to apply differential weighting factors to each spectral band of the data to be clustered, this can be done by expanding or compressing the dynamic range of the different spectral bands.

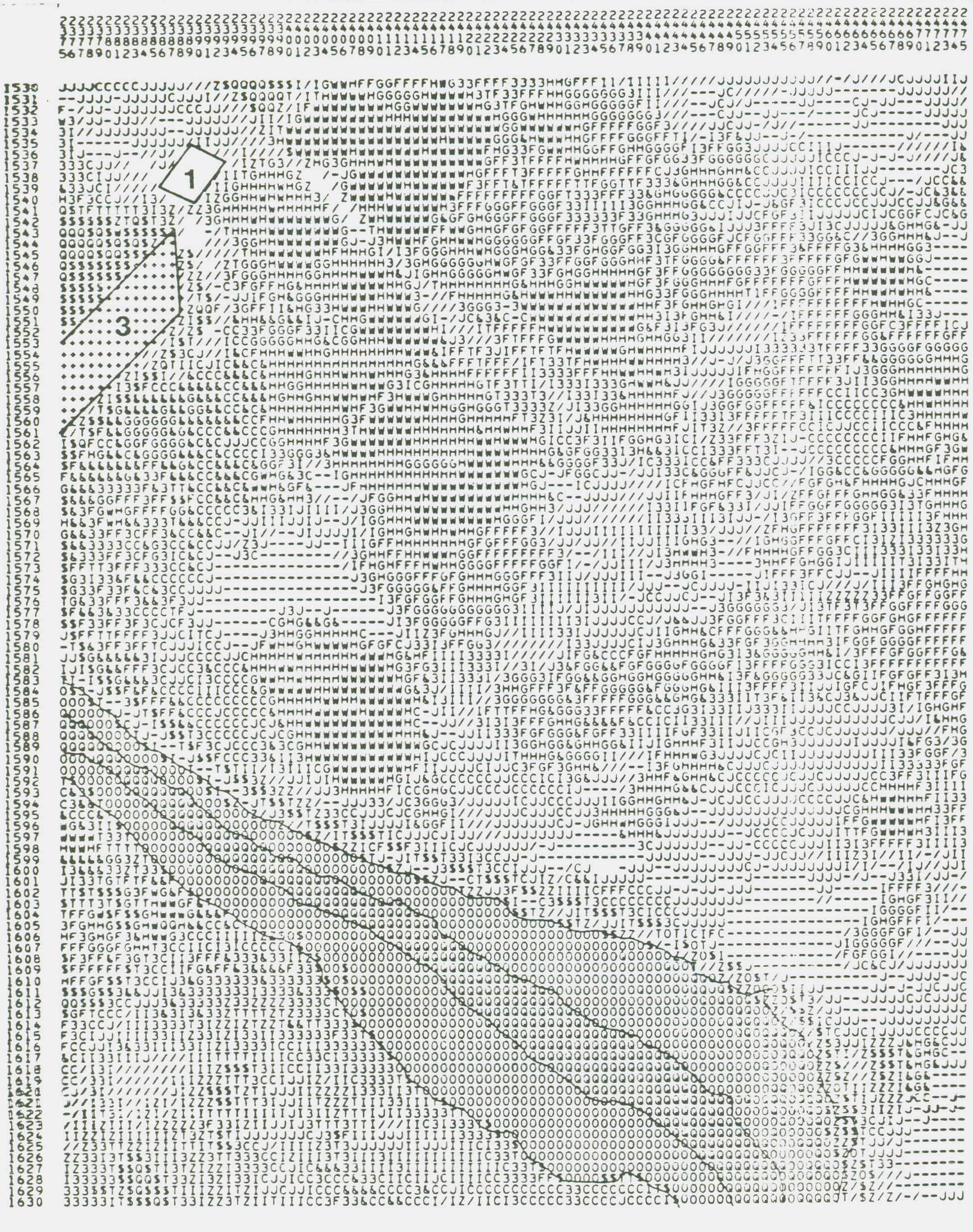

Figure 8. Cluster map of 10,201 Landsat data points with band 7 expanded by a factor of two and with 18 classes

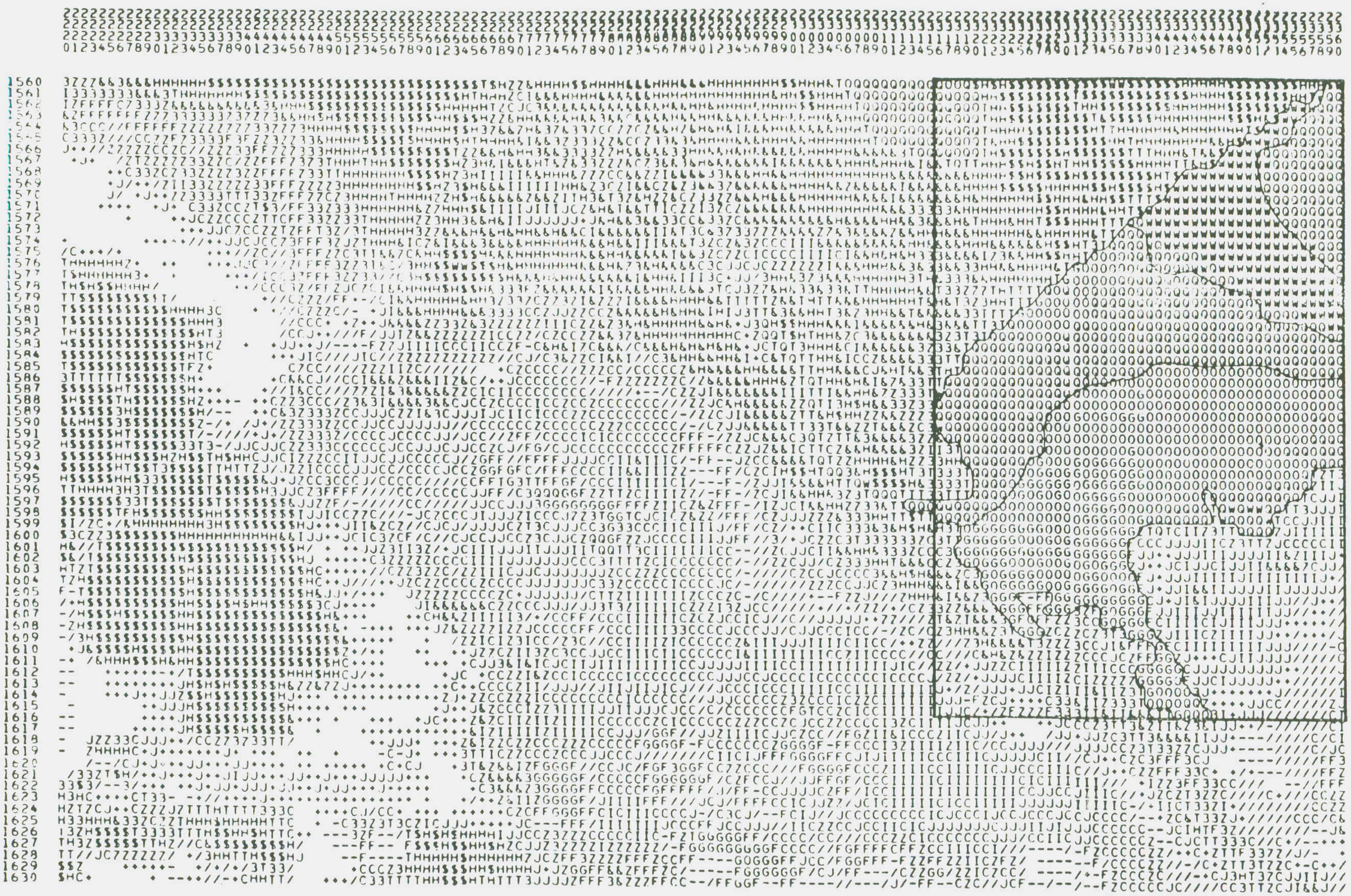

Figure 9. Cluster map of 10,201 Landsat MSS data points with band 7 expanded by a factor of two and with 18 classes.

Landsat MSS data as shown in Figures 3 and 4 may be greatly diminished if one uses expanded data, as illustrated in Figure 9. A comparison of the different water classes shown within the rectangular area of Figure 9 and those within the same area in Figure 4, clearly indicates that by expanding band 7 even after slight area shifts, the resulting cluster classes are still representative of the cover types in the scene. Consequently, the relative location of the areas to be clustered is no longer as critical if expanded data is utilized.

Conclusions and Recommendations

Although it is widely recognized by the remote sensing research and user community that clustering algorithms are a very useful analysis tool for defining the spectral training classes needed to classify Landsat MSS data, the analyst should be aware of the limitations of some clustering algorithms, such as:

1) impact of correlation between channels in the data,

2) sensitivity of the clustering algorithm to shifts of location of the clustered area, and

3) sensitivity of the clustering results to axis scale changes. This investigation has demonstrated these limitations empirically. In order to overcome these limitations, solutions are recommended:

- After the selection of the candidate clustering areas, the analyst should determine the degree of correlation of the data. If the data is not positively correlated, then the initial cluster centers should be located along the first principal component of the data distribution.

- Expansion of the Landsat band 7 dynamic range by a factor of two will reduce the sensitivity of the cluster processor to slight shifts of the area to be clustered and will produce results which better portray the spectral classes actually present in the ground scene.

References

Bartolucci, L.A., 1978, "Calibration of Landsat MSS Data", LARS Technical Report 121278, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

Bartolucci, L.A., 1979, "Digital Processing of Remotely Sensed Multispectral Data", Proceedings of the Latin American Technology Exchange Week Conference, DMA/IAGS, Panama City, Panama, May 14-19, 1979, 17pp.

Duda, R.O. and P.E. Hart, 1973, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

NASA, 1973, "Earth Resources Technology Satellite-1", Symposium Proceedings, Goddard Space Flight Center, Greenbelt, Maryland, September 29, 1972, NASA publication X-650-73-10, 165pp.

NASA, 1972, "Landsat Data Users Handbook", Goddard Space Flight Center, Greenbelt, Maryland.

Phillips, T.L., 1973, "LARSYS Version 3.1 Users' Manual", Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

Swain, P.H., 1978, "Fundamentals of Pattern Recognition in Remote Sensing", Chapter 3, P.H. Swain and S.M. Davis, editors, Remote Sensing: The Quantitative Approach, McGraw-Hill International Book Company, 1978.

Wacker, A., 1969, "A Cluster Approach to Finding Spatial Boundaries in Multispectral Imagery", LARS Information Note 122969, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 25pp.

Wacker, A. and D.A. Landgrebe, 1970, "Boundaries in Multispectral Imagery by Clustering", Proceedings of the IEEE Symposium on Adaptive Processes XI, December 7-9, 1970, pp: 4.1-4.8.

Wacker, A., 1971, "The Minimum Distance Approach to Classification", Ph.D. Thesis, School of Engineering, Purdue University, West Lafayette, Indiana, January, 1972, 361pp.