062383

THE EFFECT OF FEATURE SCALING ON THE CLUSTERING OF LANDSAT MSS DATA

L.A. BARTOLUCCI, S.M. DAVIS

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

P.H. SWAIN

Purdue University/School of Electrical Engineering/Laboratory for Applications of Remote Sensing
West Lafayette, Indiana

ABSTRACT

Nonsupervised classification by clustering has been shown to be a very important tool in the analysis of satellite remote sensing data. However, clustering algorithms which use Euclidean distance as a measure of similarity are highly sensitive to scaling differences among the variables which participate in the clustering process. Since the Landsat MSS spectral bands have different ranges and different calibration functions, this scaling sensitivity is likely to have a significant impact on the results of clustering Landsat MSS data, as is demonstrated by the experiments described in this paper. A rescaling strategy for Landsat MSS data is recommended which seems to give appropriate relative weights to the four spectral bands.

I. 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 applied effectively to this type of multispectral scanner data. It was soon recognized that the supervised method of developing the statistics for training a classifier was not an adequate means of defining the natural multispectral groupings present in a Landsat MSS data set. The supervised approach did not allow a satisfactory definition of an important type of spectral training classes which represent a large percentage of the total Landsat scene, i.e., the "spectral mixture classes."[2] Therefore, to overcome this limitation of the supervised approach, data analysts have been using regularly a nonsupervised procedure to determine the inherent structure of the Landsat MSS data by defining the training statistical parameters through the use of clustering algorithms. However, analysts need to be aware that those clustering algorithms which use Euclidean distance as a measure of similarity may not yield meaningful results when the feature space is not isotropic. Because the Landsat MSS digital data for bands 4, 5 and 6 range from 0 to 127 (7 bits) while those for band 7 range from 0 to 63 (6 bits), the resulting four-dimensional feature space is not isotropic.

II. CLUSTERING ALGORITHMS

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

There are several types of clustering algorithms, and according to Blashfield et al.,[4] there may exist as many clustering software packages as there are users.

The clustering function most commonly used at LARS (*CLUSTER) is a variant of the ISODATA algorithm.[7] *CLUSTER has been described in detail elsewhere.[9,12,13,14] The measure of similarity used in the *CLUSTER function is Euclidean distance, which implies that the cluster classes defined by this function are invariant to rigid-body motions of the data points, i.e., translations or rotations, but the function is highly sensitive to transformations that distort the distance relationships among data points, such as

differential rescaling of the feature space axes.[6]

III. CALIBRATING LANDSAT MSS DATA

The importance of calibrating the Landsat MSS data to aid in labeling spectral training classes generated by a clustering function has been demonstrated and reported elsewhere.[1,3] Calibration of the Landsat MSS data involves changing the scaling of the four-dimensional feature space axes from the original range of 0-127 for bands 4, 5 and 6 and 0-63 for band 7 to "in-band radiance" values which are expressed in terms of mWatts/cm^2-sr. This rescaling procedure alters the distance relationships among the data points and thus affects the performance of a clustering function which uses Euclidean distance as a measure of similarity. Figures 1 and 2 illustrate graphically the effect of calibrating the Landsat MSS data. Figure 1 shows four data points (A, B, C, and D) plotted for an uncalibrated, two-dimensional (bands 6 and 7) feature space. A clustering function which uses the Euclidean distance measure would group points A and B into Cluster 1 and points C and D into Cluster 2. Calibration of the Landsat MSS data would change the scaling of the two-dimensional feature space axes as illustrated in Figure 2.[3] As a consequence, the relative inter-point distances are considerably changed, and a clustering function which uses the Euclidean distance measure would group points A and C into Cluster 1 and points B and D into Cluster 2. It is evident from these illustrations that such a simple change of scale in the feature space can yield completely different cluster classes.

To illustrate the effect that a simple change of scale can have on the performance of clustering real data, a set of Landsat MSS data (Scene ID: 2034-16200) collected over Matagorda Bay, Texas, on February 25, 1975, was clustered using the *CLUSTER function. A total of 10,201 data points (pixels) calibrated into in-band radiance values were clustered into 18 classes based on all four bands. The resulting cluster map is shown in Figure 3.

The uncalibrated data from the same area were clustered into 18 classes, yielding the cluster map shown in Figure 4.

For reference a black-and-white reproduction of a color infrared photograph covering the Austwell, Texas, 7-1/2 minute topographic quadrangle area is shown in Figure 5.

A comparison of the cluster maps presented in Figures 3 and 4 shows clearly the different cluster results obtained from the calibrated (Figure 3) and uncalibrated (Figure 4) data sets. Note that clustering the calibrated data yielded only one spectral class of water along Guadalupe Bay, in contrast to the three spectral classes of water obtained from the uncalibrated data set. On the other hand, clustering the calibrated data produced the differentiation of two man-made features labeled by the numbers 1 (concrete parking lot) and 2 (settling pond), whereas the uncalibrated data did not permit the separation of these two different spectral classes.

These differences in results are not unexpected since, in calibrating the Landsat MSS data, one is essentially applying a different linear transformation to each one of the four axes and, in the process, changing the (Euclidean) distance relationships among all data points. The authors have verified that rescaling the four Landsat feature space axes by applying the same linear transformation (multiplying all values by a constant greater than unity) to each of the four axes does not distort the feature space as perceived by the *CLUSTER algorithm, and therefore the resulting cluster classes correspond exactly to the cluster classes obtained from clustering the original non-rescaled data. This is discussed further in the next section.

IV. THE EFFECTS OF OTHER RESCALING ALTERNATIVES

Rescaling the Landsat MSS data based on the internal calibration reference values given above has roughly the effect of multiplying the band 7 data by a factor

---

* The data were calibrated using the following internal calibration reference values (in mWatts/cm^2-sr.):

	Band 4	Band 5	Band 6	Band 7
Minimum Radiance	0.10	0.07	0.07	0.14
Maximum Radiance	2.10	1.56	1.40	4.15

of four and the other three bands by factors of one and a half or two. Under the Euclidean distance measure, this could cause the clustering program to emphasize information in band 7 to the detriment of information in the other three bands. This effect can be seen by comparing the results of the experiments described below with the results of clustering the calibrated data.

The following experiments were motivated by the notion that the data in the four Landsat MSS bands could be made more commensurate simply by equalizing their ranges. To do this, one would either have to expand (multiply by a factor of two) the digital values of band 7 or compress the range of bands 4, 5 and 6 by a factor of two, leaving the original band 7 unaltered.

Several rescaling transformations were applied to the Landsat MSS data, including:

1) expansion of band 7 (original scaling of 0-63) by a factor of two, leaving bands 4, 5 and 6 unaltered.

2) compression of bands 4, 5 and 6 by a factor of two, 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, only the third did not produce cluster classes different from those obtained from clustering the original (untransformed) data. Transformations 1, 2 and 4 changed the internal structure of the data distribution considerably. The results of applying the clustering algorithm to data sets that have undergone a linear compression in bands 4, 5 and 6 show that compressing these bands linearly causes a great deal of the information content (spectral separability) in the data set to be lost.

Only the first transformation, i.e., expanding the range of band 7 by a factor of two and leaving bands 4, 5 and 6 unaltered, caused the clustering algorithm to define spectral classes that more accurately represented the ground cover types in the scene. Figure 6 shows the result of clustering a Landsat MSS data set which has undergone a linear expansion of band 7 with bands 4, 5 and 6 left unaltered.

A comparison of the cluster maps shown in Figures 4 and 6 with the reference photography (Figure 5) indicates that the spectral cluster classes obtained from the "expanded" data set (Figure 6) represent more accurately the ground cover types in the scene. Note that features "1" and "2" in Figure 4 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 2 is a turbid pond, whereas feature 1 is a large factory.* These results show the limitations of the clustering algorithm when applied to original Landsat MSS data. On the other hand, note that features 1 and 2 in Figure 6 have been clustered into two different spectral classes which accurately represent the two different ground cover types present in the scene.

The spectral separability of these two distinct cluster classes (features 1 and 2) as measured by the Transformed Divergence[10,11] indicates that the classes are completely separable; they have a pairwise transformed divergence value of 2000. These results show that the clustering algorithm was unable to distinguish these two spectrally very different classes when applied to the original (unexpanded) data set, whereas these two features were accurately differentiated by the same clustering algorithm applied to the expanded data.

---

* The turbid water and the factory with associated parking lots have similar spectral responses in bands 4, 5 and 6, and very different spectral responses 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 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 appropriately expanding or compressing the ranges of the different spectral bands.[5]

V. CONCLUSIONS AND RECOMMENDATIONS

Although it is widely recognized by the remote sensing community that clustering is a very useful analysis tool for defining the spectral training classes needed to classify Landsat MSS data, the analyst needs to be aware of the sensitivity to scaling inherent in those clustering algorithms which use the Euclidean distance as a measure of similarity. The authors recommend that, when using such algorithms, the range of the Landsat MSS band 7 be scaled (expanded) by a factor of two (from 0-63 to 0-127) before clustering is performed.

REFERENCES

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

2. 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.

3. Bartolucci, L.A. and S.M. Davis, 1983, "The Calibration of Landsat MSS Data as an Analysis Tool," Proceedings of the 9th International Symposium on Machine Processing of Remotely Sensed Data, LARS/Purdue University, June 21-23, 1983.

4. Blashfield, R.K., M.S. Aldenderfer and L.C. Morey, 1982, "Cluster Analysis Software," Chapter 11, in Classification, Pattern Recognition and Reduction of Dimensionality, P.R. Krishnaiah and L.N. Kanal, eds., North-Holland Publishing Company, Amsterdam, pp. 245-266.

5. Chu, N.Y. and P.E. Anuta, 1979, "Multidimensional Scaling for Clustering of Dissimilar Data Types," LARS Information Note 050279, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

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

7. Hall, D.J. and G.H. Ball, 1965, "ISODATA: A Novel Method of Data Analysis and Pattern Classification," Technical Report SRI Project 5533, Stanford Research Institute, Menlo Park, California.

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

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

10. Swain, P.H. and A.G. Wacker, 1971, "Comparison of the Divergence and B-Distance in Feature Selection," LARS Information Note 020871, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

11. Swain, P.H. and R.C. King, 1973, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing," LARS Information Note 042673, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

12. Wacker, A.G., 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.

13. Wacker, A.G. 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.

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

AUTHOR BIOGRAPHICAL DATA

Luis A. Bartolucci received his B.S., M.S., and Ph.D. in Geophysics from Purdue University. Dr. Bartolucci has been involved in Remote Sensing Research since 1969. He has played an active role in the development of remote sensing technology for applications in the area of water resources and has also made outstanding contributions in the field of thermal infrared radiation for remote sensing applications. In addition, Dr. Bartolucci has served as consultant to the U.S. Information Agency, the U.S. Agency for International Development, the InterAmerican Development Bank and to several Latin American development agencies. He has been Principal Investigator and Project Director of several domestic and international research and training programs involving computer-aided processing and analysis of remotely sensed data for earth resources inventories.

Shirley M. Davis is Senior Education and Training Specialist at Purdue University's Laboratory for Applications of Remote Sensing and Director of Independent Study, Continuing Education Administration. Mrs. Davis received the A.B. degree with honors in English in 1958 from Sweet Briar college and the M.A. degree in English from Case-Western Reserve University in 1962. Her major contributions to remote sensing education have been as co-author and editor of the LARSYS Educational Package; co-editor and contributing author of the textbook Remote Sensing: The Quantitative Approach; Chairman of the 1981 Conference on Remote Sensing Education; and creator/coordinator of the videotape series Introduction to Quantitative Analysis of Remote Sensing Data. Her recent work has involved the development of educational materials for digital image processing.

Philip H. Swain is Associate Professor of Electrical Engineering, Purdue University, and Program Leader for Data Processing and Analysis Research at Purdue's Laboratory for Applications of Remote Sensing (LARS).

Affiliated with LARS since its inception in 1966, Dr. Swain has developed methods and systems for the management and analysis of remote sensing data. He has been employed by the Philco-Ford Corporation and Burroughs Corporation and served as consultant to the National Aeronautics and Space Administration (NASA) and the Universities Space Research Association. His research interests include theoretical and applied pattern recognition, methods of artificial intelligence, and the application of advanced computer architectures to image processing. He is co-editor and contributing author of the textbook Remote Sensing: The Quantitative Approach (New York: McGraw-Hill, 1978).

Dr. Swain received the B.S. degree in electrical engineering from Lehigh University in 1963 and the M.S. and Ph.D. degrees from Purdue University in 1964 and 1970, respectively. He is a member of the Pattern Recognition Society, the IEEE Computer Society and the IEEE Geoscience and Remote Sensing Society.

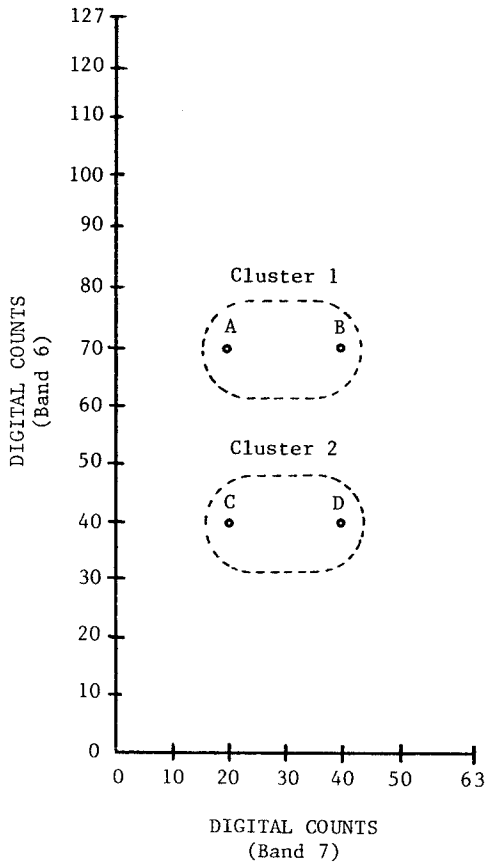

Figure 1. Grouping of Digital Counts. A clustering algorithm that measures similarity with Euclidiance distance would group these hypothetical, two-dimensional data points as shown.

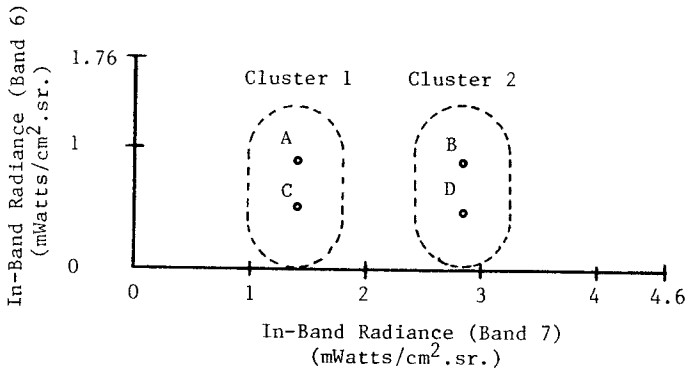

Figure 2. Grouping of Calibrated In-Band Radiance Values.

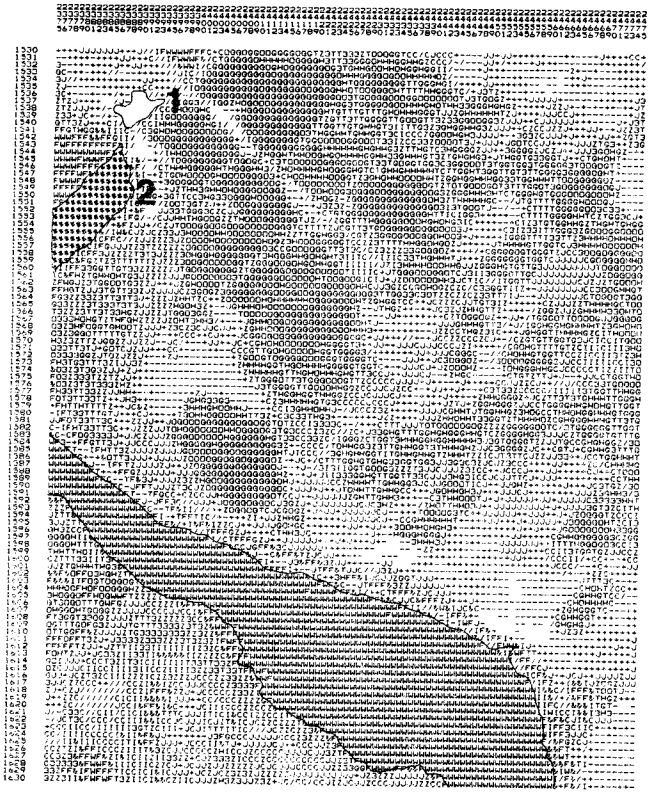

Figure 3. Map of 18 Cluster Classes Derived from Calibrated Data (Matagorda Bay Study Area).

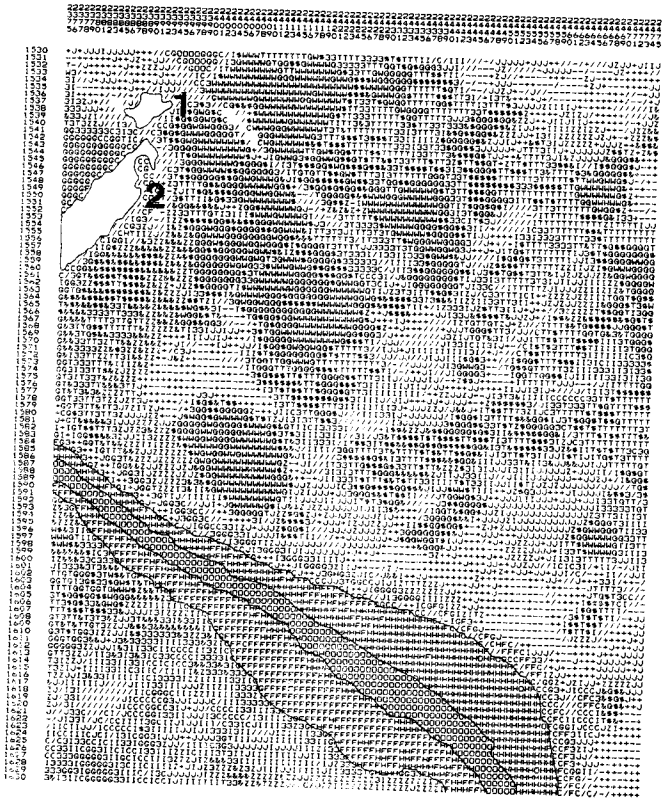

Figure 4. Map of 18 Cluster Classes Derived from Uncalibrated, In-Band Radiance Values (Matagorda Bay Study Area).

Figure 5. Reproduction of a Portion of Color Infrared Photograph of the Matagorda Bay Study Area.

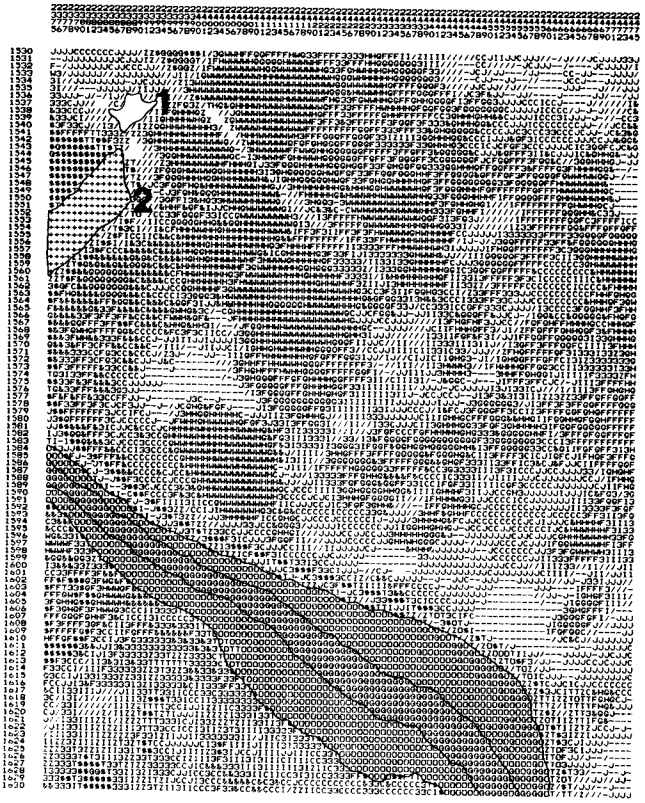

Figure 6. Map of 18 Cluster Classes Derived from Transformed Data. Band 7 data was expanded by a factor of 2; bands 4, 5, and 6 are unaltered.