Paper I 18

CLASSIFICATION OF ERTS-1 MSS DATA BY CANONICAL ANALYSIS

H. M. Lachowski and F. Y. Borden, *Office for Remote Sensing of Earth Resources, The Pennsylvania State University, University Park, Pennsylvania*

ABSTRACT

The objective of canonical analysis is to obtain the maximum separability among a number of categories. The application of canonical analysis was investigated using the merged MSS ERTS-1 data for one area viewed on two dates. The effect of threshold values on classification regions and confusion regions was investigated.

INTRODUCTION

Canonical analysis, which is also known as multiple discriminant analysis, is a multivariate statistical method that has application to classification of multivariate observations into statistically defined categories. Linear combinations of the observed measurable characteristics are found that yield one or more indices that emphasize the differences among the classes. The linear combinations are then applied to data for unknowns and supply the means for classification. Canonical analysis has been used satisfactorily many times in classification problems in the biological and natural sciences. It has obvious potential advantages for a similar application in the field of remote sensing. In this research, the method was tailored for classification of remote sensor data collected by multispectral scanners (MSS) carried by airplanes and satellites.

COMPUTATIONAL METHOD

MSS digital data, having a response for each of p spectral bands or channels, are organized according to scan lines across the flight path and by elements within scan lines. The data can be considered to be multivariate of dimension p with, say, $X_{i,j}$ being the observational vector for scan line i and element j. In the use of canonical analysis for classification of unknown data into k+1 categories (i.e., k defined categories and an "other" category), an estimate for the mean vector of each of the k classes is needed, as well as for the corresponding covariance matrix. For MSS data, these statistics can be obtained by the use of training areas, each of which is considered to be spectrally

homogeneous. At least one training area must be defined for each class and the training areas must be representative of the classes to be investigated. In a geometric sense, in p dimensions, each class is defined by its mean vector and its covariance matrix. The covariance matrix defines the elipsoidal dispersion pattern of points clustering around the class mean vector.

Canonical analysis accomplishes four things in a geometrical sense. First, the origin is translated to the point of the overall mean vector. Second, the original axes are rotated to new orthogonal positions. In this, the first axis is placed according to the maximum possible separability among the class mean vectors. The second axis is positioned orthogonally to the first and according to the maximum remaining class separability and so on for the other axes. The third accomplishment is scaling the axes so that the elipsoidal dispersion patterns are transformed into spherical patterns. The fourth feature is that the space dimension required for satisfactory classification is, in general, substantially reduced. This means that if there are p original variables, the canonical transformation will generally yield substantially less than p transformed new variables that still retain essentially all of the useful information for satisfactory classification. Because it will be necessary to refer to the test results in discussing the method, such discussion will be deferred until after the test data have been described.

THE TEST AREA AND PRELIMINARY COMPUTATIONS

The test area resides in the vicinity of Harrisburg, Pennsylvania, just northwest of the central metropolitan part of the city and stretches across the Susquehanna River encompassing substantial land back from both shorelines (Figure 4B). It includes a part of the river, a railroad marshalling yard, suburban areas, and areas of vegetation. ERTS-1 bulk MSS data from two dates were mapped after being brought into registration by translation and merged. Each element in the test area as a result of merging was composed of eight values, four from each of the two dates. The two dates were August 1, 1972 (scene 1009-15241), and October 11, 1972 (scene 1080-15185). The October 11th overpass was one day prior to the fourth eighteen-day cycle. These two dates were chosen because they were the only ones for which MSS tapes were in hand and for which the test area was cloud free.

The test area was chosen because it was familiar and had a variety of natural and cultural targets. The specific targets that were selected for study were river water, the railroad yard, two suburban targets, and two vegetation targets. The categories are referred to by numbers in figures and tables, therefore, category 1 stands for river, 2 for railroad yards, 3 and 4 for suburbs, and 5 and 6 for vegetation. The area was dominated by these categories.

The mean (spectral signature) and the covariance matrix were computed for selected training areas using various algorithms from the digital processing system for MSS data as described by Borden (1972). These statistics were then input to the canonical analysis program (Lachowski, 1973) where the transformation matrix was computed and transformations of the mean vectors were performed. Following this, each unknown MSS observation was transformed using the transformation matrix and classified into one of the known categories or into the "other" category.

RESULTS AND DISCUSSION

The canonical analysis of the test data showed that essentially all of the separability (99.9%) among the six categories could be recovered in four canonical axes. The four axes accounted for 88.5, 11.1, 0.2, and .06 percent, respectively, of the separability. In Figure 1 the transformed means are plotted for the first two axes as the center points of the circles. The circles represent the 94.5% contour for the transformed dispersion about each mean. The radius of each circle is 1.9, the threshold value, which corresponds to the 94.5% contour. Considering the two axes together, confusion in classification can occur only where circles overlap and, for the 1.9 threshold value, no confusion occurs. The circles have been projected onto each axis in Figure 1, thus producing classification intervals. Overlap occurs for some of these intervals on the first axis and for many on the second axis, thus indicating two axes are needed for separability using a 1.9 threshold value. In Table 1 the indicators of overlap or nonoverlap in classification intervals are presented for each axis. The map of the classification using the 1.9 threshold value is shown in Figure 3, where the vegetation and suburb classes were assigned the same mapping symbol for map clarity in this report.

To show the situation that occurs when classification regions overlap, the test data were run with a threshold value of 3.0 that corresponds to a 99.7% contour for transformed dispersion about the means. Although this would be an unrealistically high value in practice, the influence of the threshold value on confusion regions is amply demonstrated. Plots of the transformed means are presented in Figure 2. In this case, confusion regions exist for the two axes considered together for classes 4 and 5 and 5 and 6. Even with the four axes, confusion could not be entirely resolved although the confusion regions were decreased in size. The mapping results show this in Figure 4A, where the classification confusion is mapped as a separate category. If each of the two axes is considered separately, as seen by the projection of the circles onto the axes in Figure 2, there are four overlapping intervals for axis one and eight overlapping intervals for axis two. The corresponding indicators are given in Table 2.

If an observation falls in a confusion region, what can be done to resolve the confusion? The way in which this was handled was to assign the observation to the class to which it was nearest to the mean; i.e., using a minimum euclidean distance classification scheme based on the transformed data. The mapping results for this are shown in Figure 4B, where it appears that this was a reasonable path to follow. An assignment was made to the "other" category if a transformed observation fell outside every classification region. The "other" category is mapped as the blank areas in all of the figures. The threshold value not only has an influence on the sizes of the confusion regions, but also on the number of "other" classifications. As the threshold value increases, the sizes of the confusion regions increase, but the number of "other" classifications decreases and vice versa. This effect can be seen in the comparison of Figure 4 with Figure 3 where the unclassified area is greater for the smaller threshold value.

The transformation vectors for each of the four axes are presented in Table 3. The values in each vector are analogous to partial regression coefficients of multiple regression. Some interpretation can be made of these for MSS data since the data for all channels have the same order of magnitude. For the first vector, for example, the transformation is dominated by channel seven for each date. The interpretation here is that the emphasis is on the spearation of water from nonwater signatures because of the generally low reflectance of water in channel seven. The plot in Figure 1 bears this out, where the greatest separation is shown to be between the water class and the others for axis one.

FUTURE WORK

Although classification advantages have not been discussed in the report, it appears that strategies can be employed that would reduce computation time in classification by use of canonical analysis. One possibility can be seen with reference to Figure 1. Suppose the classification intervals for the classes are stored for all of the axes, then, by transformation with the first vector, followed by a table look-up, one can determine if the observation falls in any class interval or intervals. If it does not fall in any interval, it is "other." If it falls in one interval only, it is either in that class or "other." If it falls in two or more intervals, it is in a confusion region and the transformation for the next axis has to be made. Work is presently in progress to implement such a procedure, as well as to investigate other similar ones.

LITERATURE CITED

Borden, F. Y. 1972. A Digital Processing and Analysis System for Multispectral Scanner and Similar Data. Remote Sensing of Earth Resources, Vol. 1. Edited by F. Shahrokhi. University of Tennessee Space Institute, Tullahoma, Tenn. pp. 481-507.

Lachowski, H. M. 1973. Canonical Analysis Applied to the Interpretation of Multispectral Scanner Data. M.S. thesis, The Pennsylvania State University, University Park, Pa.

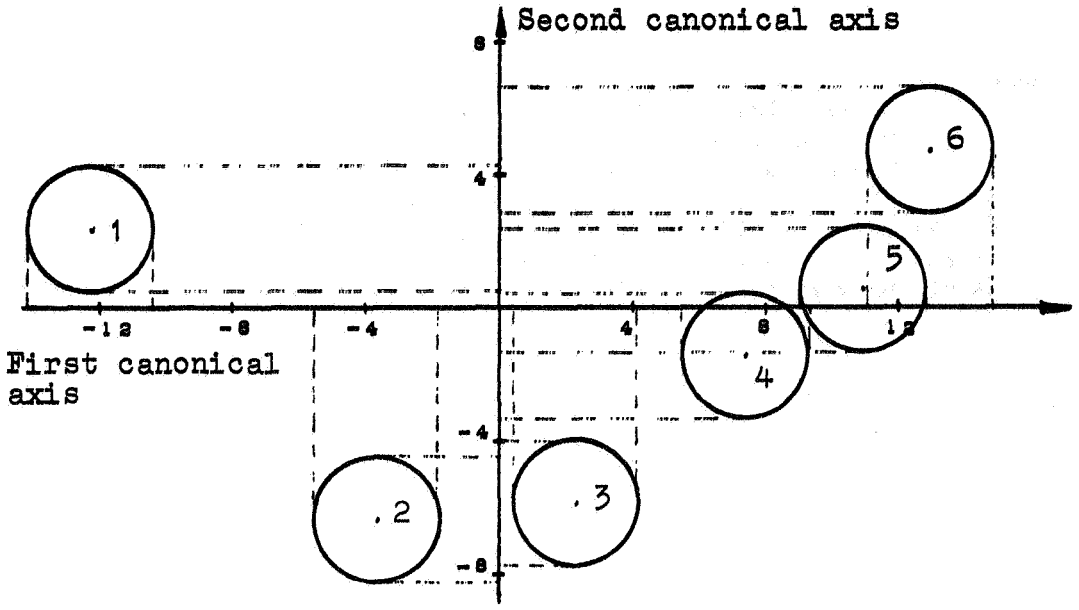


Figure 1. Geometric representation of the separation of the transformed mean estimates for all the categories and their 94.5% classification regions using the first two canonical axes.

Table 1. Matrices showing category overlaps (1's) and nonoverlaps (0's) for the 94.5% classification intervals using the first two canonical axes.

	First Canonical Axis						Second Canonical Axis					
	Category						Category					
Category	1	2	3	4	5	6	1	2	3	4	5	6
1	1						1					
2	0	1					0	1				
3	0	0	1				0	1	1			
4	0	0	0	1			0	0	0	1		
5	0	0	0	1	1		1	0	0	1	1	
6	0	0	0	0	1	1	1	0	0	0	0	1

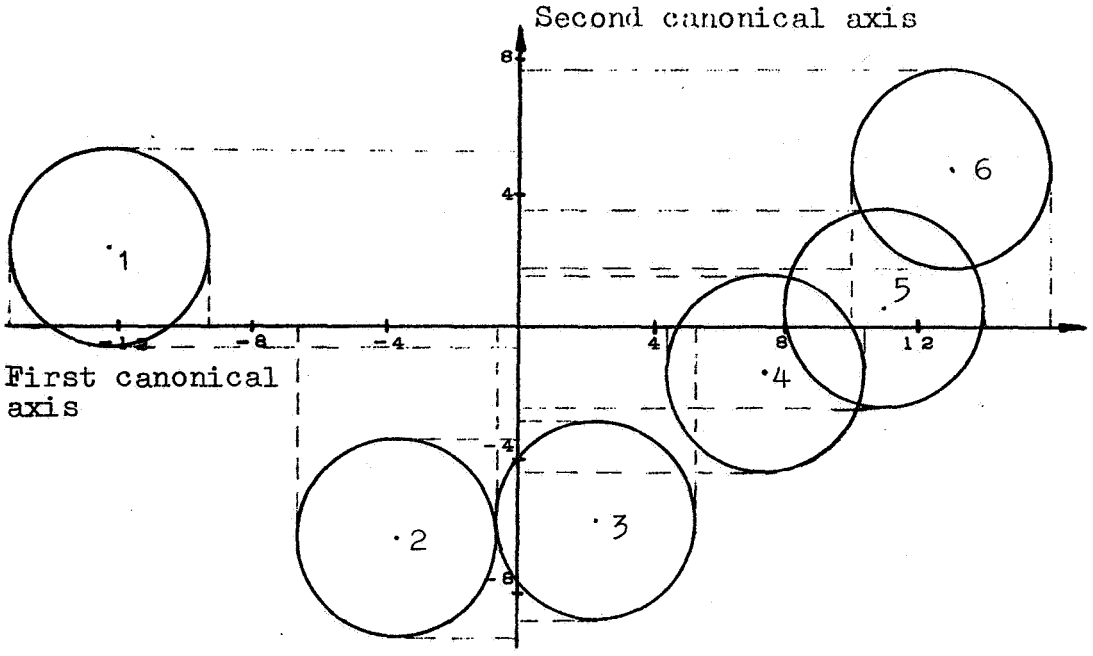


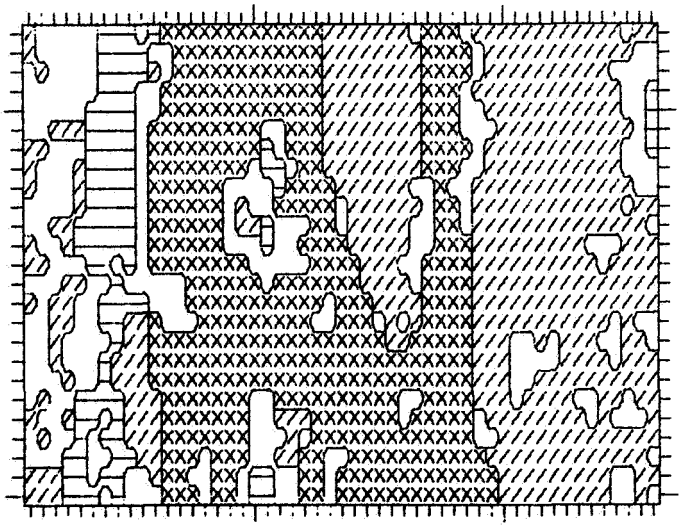
Figure 2. Geometric representation of the separation of the transformed mean estimates for all the categories and their 99.7% classification regions using the first two canonical axes.

Table 2. Matrices showing category overlaps (1's) and nonoverlaps (0's) for the 99.7% classification intervals using the first two canonical axes.

	First Canonical Axis						Second Canonical Axis					
	Category						Category					
Category	1	2	3	4	5	6	1	2	3	4	5	6
1	1						1					
2	0	1					0	1				
3	0	0	1				0	1	1			
4	0	0	1	1			1	1	1	1		
5	0	0	0	1	1		1	0	0	1	1	
6	0	0	0	1	1	1	1	0	0	0	1	1

Table 3. Transformation matrix used for transforming mean estimates (signatures) for the known categories and for transforming each unknown observation.

Channel	Canonical Axes			
	1	2	3	4
August 1				
4	0.045	-0.313	0.303	-0.112
5	0.016	-0.561	-0.686	0.227
6	0.122	0.008	0.476	-0.434
7	0.417	0.009	0.305	0.705
October 11				
4	-0.073	-0.177	0.505	0.565
5	-0.101	-0.332	0.054	-0.096
6	0.205	-0.158	-0.215	-0.370
7	0.546	0.254	-0.382	0.427



Railroad yards ——
Vegetation and suburbs /////
River xxxx
Other blank

Figure 3. Computer generated map for the test area with the threshold value set at 1.9.

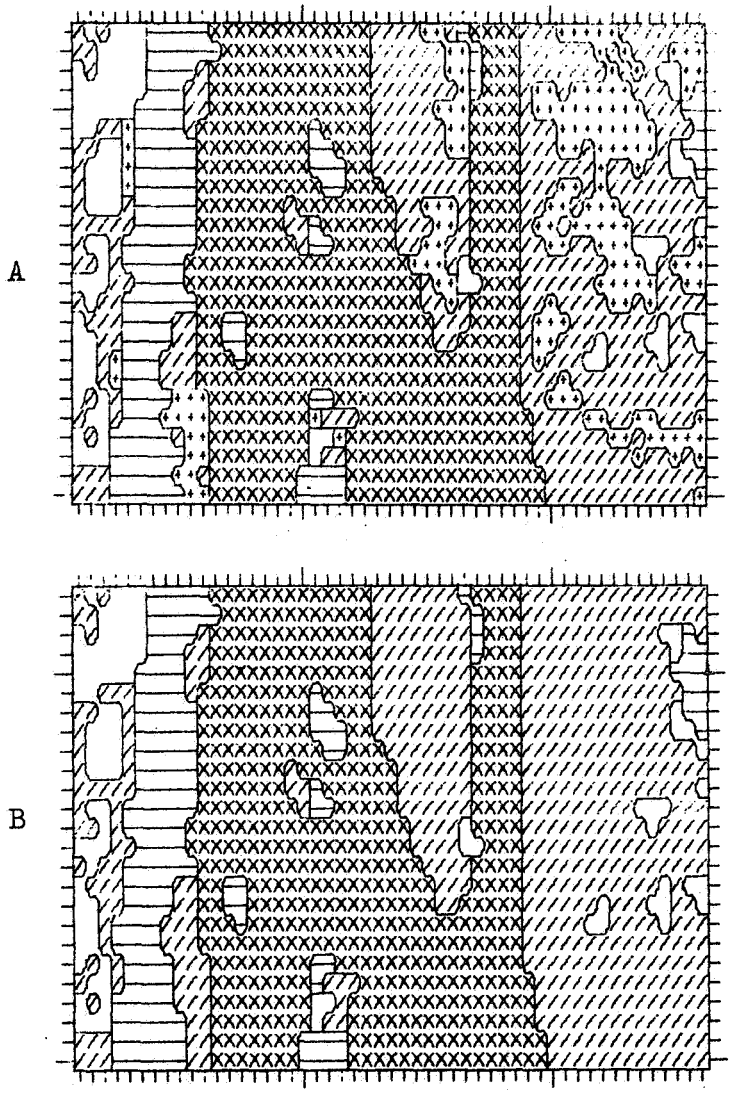


Figure 4. Computer generated maps of the test area with the threshold value set at 3.0. Map A indicates where confusion occurred. Map B shows the confused areas classified into one of the known categories.