X-923-77-182
PREPRINT

Tmx 71387

A REDUCTION IN AG./RESIDENTIAL SIGNATURE CONFLICT USING PRINCIPAL COMPONENTS ANALYSIS OF LANDSAT TEMPORAL DATA

DARREL L. WILLIAMS
F. YATES BORDEN

(NASA-TM-X-71387) A REDUCTION IN AG/RESIDENTIAL SIGNATURE CONFLICT USING PRINCIPAL COMPONENTS ANALYSIS OF LANDSAT TEMPORAL DATA (NASA) 14 p HC A02/MF A01 CSCL 08B G3/43
N77-32566
Unclas
48573

JULY 1977

GSFC — GODDARD SPACE FLIGHT CENTER —
GREENBELT, MARYLAND

X-923-77-182

A REDUCTION IN AG./RESIDENTIAL SIGNATURE CONFLICT USING PRINCIPAL COMPONENTS ANALYSIS OF LANDSAT TEMPORAL DATA

Darrel L. Williams
F. Yates Borden

July 1977

GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland

CONTENTS

A REDUCTION IN AG./RESIDENTIAL SIGNATURE CONFLICT USING PRINCIPAL COMPONENTS ANALYSIS OF LANDSAT TEMPORAL DATA

Darrel L. Williams
Earth Resources Branch – Code 923
NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

BIOGRAPHICAL SKETCH

Darrel L. Williams is an Environmental Scientist with the Earth Resources Branch at NASA's Goddard Space Flight Center, and a forestry discipline scientist for GSFC's Information Transfer Laboratory (Intralab). He received his B.S. and M.S. in Forest Resources from the Pennsylvania State University. Since 1972 his research has involved digital image processing of remotely sensed data, as applied to the Agriculture/Forestry/Range discipline. He is a member of the American Society of Photogrammetry, Society of American Foresters, and the American Forestry Association.

F. Yates Borden
Office for Remote Sensing of Earth Resources
The Pennsylvania State University
University Park, Pennsylvania 16802

BIOGRAPHICAL SKETCH

F. Yates Borden is a Professor of Forest Resources in the School of Forest Resources. He received a B.S. and M.S. in Agronomy from Cornell University and a Ph.D. in Agronomy from The Pennsylvania State University. His research in remote sensing, conducted through the Office for Remote Sensing of Earth Resources (ORSER), has emphasized the development and use of digital processing in analysis of MSS data since 1971.

ABSTRACT

One important objective of a cooperative project between the U.S. Bureau of Census and NASA is to develop the ability to accurately delineate the types of land cover in the urban-rural transition zone of metropolitan areas. The application of principal components analysis to multidate Landsat imagery is being investigated as a method of reducing the overlap between residential and agricultural spectral signatures. The statistical concepts of principal components analysis are discussed, as well as the results of this analysis when applied to multidate Landsat imagery of the Washington, D.C. metropolitan area.

A REDUCTION IN AG./RESIDENTIAL SIGNATURE CONFLICT USING PRINCIPAL COMPONENTS ANALYSIS OF LANDSAT TEMPORAL DATA

INTRODUCTION

Research pertaining to the use of Landsat imagery has indicated that the four band multispectral scanner (MSS) data can be utilized for the delineation of the major types of land cover typically found in urbanized areas (Christenson and Lachowski, 1976). The incentive for exploring this particular application of the Landsat imagery is that it yields repetitive, synoptic views of major metropolitan areas, providing a means of monitoring urban growth and change on a regular basis. In this respect, the Geography Division of the U.S. Bureau of the Census is currently involved in a cooperative project with the Earth Resources Branch at NASA's Goddard Space Flight Center.

One important objective within the context of this project is to develop the ability to accurately delineate the types of land cover in the urban-rural transition zone in the fringe of metropolitan areas. However, given Landsat's spectral and spatial resolution, and the diversity of land cover in the urban-rural transition zone, results have shown that the spectral signatures for certain residential developments are quite often similar to those in areas of agricultural land use. Digital classifications based upon such overlapping signatures often result in false alarms of residentially classified pixels appearing in agricultural areas. This effect is largely due to the heterogeneity over short distances found in these areas relative to Landsat's spectral and spatial resolution. Consider, for example, the intermingling of roof tops, lawns, streets and wood lots in residential areas, surrounded by agricultural lands consisting of row crops, strip cropping patterns, pastures, fence rows, access roads and wood lots. The resulting signature confusion between these two different types of land cover must be overcome prior to the routine use of Landsat MSS data by agencies such as the U.S. Bureau of the Census.

Several methods of digital processing, including image enhancement, are currently being investigated to reduce the problem of ag./residential signature confusion. The approach which will be discussed within this report is the application of principal components analysis. This statistical processing method has been successfully used to discriminate certain rock and soil types (Podwysocki et al., 1977), and therefore warranted investigation for urban area delineation.

DESCRIPTION OF THE STUDY AREA

A portion of Prince Georges County, Maryland was chosen as the study area. The entire county falls within the Washington, D.C. Standard Metropolitan Statistical Area and its selection was governed by: (a) rapid suburban growth in recent years; (b) diversity in land use and good representation of a transition zone from urban to rural; and (c) the availability of supporting aircraft photography, maps, statistical documentation, and accessibility for ground truth surveys.

DATA SOURCES

Previous investigations have indicated that if seasonal Landsat image sets are geometrically registered and merged, the ability to differentiate between certain land use categories increases (Williams, 1976; Kan and Dillman, 1975). Fall (11 Oct. '72) and spring (9 Apr. '73) Landsat-1 scenes of the Washington, D.C. area were being utilized in other phases of the Census Bureau project, and thus were chosen for analysis. These two image sets were geometrically registered and merged prior to analysis so that each pixel represented the same ground area for the two dates of coverage. Thus, a total of eight spectral values were available for each pixel, and this data served as the basis for obtaining the training area statistics for each type of land cover. This image set provided a basis for evaluating differences in classification results between the two dates, between the separate dates and combined dates, or between any of these and the principal components classification. Any differences in these various classifications had to be the direct result of analytic differences since common training areas (i.e., groups of pixels) were used.

PROCESSING OF THE EIGHT CHANNEL MSS IMAGERY

Computer processing of the geometrically corrected, multidate Landsat imagery was accomplished using The Pennsylvania State University ORSER System (McMurty et al., 1974). Preliminary processing efforts followed the conventional supervised method of analysis for each individual image date by: (a) selecting training areas; (b) computing spectral signatures and related statistics for these areas; and (c) classifying the study area using a euclidean distance classifier. A comparison of the spectral signatures and their euclidean distance of separation for each of the two dates revealed a rather close similarity between residential and agricultural signatures. This resulted in considerable misclassification (i.e., residential pixels occurring in agricultural areas).

Eight-channel statistics for combined dates were computed for exactly the same training areas and the study area was again classified. Some improvement in signature separability, and thus in classification results was realized due to the fall/spring seasonal differences between the residential woodland and lawns and the surrounding agricultural crops. However, unacceptable levels of ag./residential confusion remained.

At this point, it was decided to investigate the potential attributes of principal components analysis, applied to the 8-channel multidate image set. The following discussion of principal components analysis is presented so that the reader may become familiar with the basic concepts of this statistical approach prior to the discussion of the results obtained using this method.

PRINCIPAL COMPONENTS ANALYSIS

Background

Principal components analysis has a long history of theoretical evaluation and application in statistics and biometrics dating to the early 1900's (Spearman, 1904). The emphasis has been on phenomenological interpretations based on relationships among the variables. The same analysis, known as the Karhunen-Loeve Expansion (Fu, 1968) in the engineering and pattern recognition disciplines, has been used predominantly for its dimensionality reduction transformation. The two approaches to the same analysis have come together in the field of remote sensing where interpretation and dimensionality reduction are simultaneously important.

Principal components analysis of p original, say X, variables determines a linear transformation which condenses essentially all of the information in the original data into q new, say Y, variables where q is less than p. No distinction is made between meaningful variability (information) and random, undesirable variability (noise). For this reason, recovery of all variability in order to preserve the information is the ideal objective of the transformation into the q Y-variables. However, in most real cases some variability is lost when q is less than p. The variability measure is the total variance e.g., the sum of the variances for the X-variables, and is determined on the basis of the sums of observed deviations from the grand sample or population means. Therefore, in the ideal case the total variance for the q Y-variables would equal the total variance of the p X-variables.

The transformation is found so that the Y-variables are uncorrelated (orthogonal), whereas the X-variables are not. Generally, the greater the correlations among the X-variables, the smaller q will have to be relative to p. This is certainly the case for typical MSS imagery where a high degree of correlation often exists between selected pairs of channels. If the X-variables were nearly uncorrelated to begin with, there would be essentially no advantage in the analysis. Orthogonality results from (i.e., is a condition of) the analysis. However, the unique characteristic of principal components analysis is that the first Y-variable has the greatest possible variance associated with it; the second Y-variable has the greatest possible variance of the remaining variance and so on. Geometrically, the transformation corresponds to a rotation of the original axes to new ones which are orthogonal to each other under the conditions given above. Pattern recognition and multivariate statistical texts typically show the geometric concepts as well as the algebraic relations for principal components analysis (Fu, 1968; Seal, 1964).

Interpretation

The eight channels (i.e., p equal 8) in the multidate image set yield a total variance of 1127. This value is obtained by summing the variances for each of the eight channels (Table 1). The total variance is fairly evenly divided between the two image dates, with 48.5% in the four channels (1-4) of the October '72 image and 51.5% in the four channels (5-8) of the April '73 image. According to the correlations among channels, it would be expected that a substantial reduction in dimensionality could be affected by a principal components analysis (Table 2). Notice the number of strong correlations which are about .8 and greater.

Table 1
Statistics for Eight Channel Multidate Landsat Data
(NOTE: 72,373 pixels in the data set)

	Channels							
	1	2	3	4	5	6	7	8
Mean	47.6	36.5	64.5	73.3	62.1	54.9	74.5	77.6
Variance (Σ = 1127)	114.1	226.8	90.7	116.4	80.9	132.2	164.0	201.5
Percentage Total Variance	10.1	20.1	8.0	10.3	7.2	11.7	14.6	17.9

Table 2
Correlations Between MSS Channels

Channel	Channels							
	1	2	3	4	5	6	7	8
1	1.00							
2	0.96	1.00						
3	0.59	0.61	1.00					
4	-0.03	-0.03	0.71	1.00				
5	0.84	0.81	0.44	-0.10	1.00			
6	0.79	0.80	0.47	-0.05	0.93	1.00		
7	0.59	0.57	0.58	0.28	0.65	0.53	1.00	
8	0.29	0.26	0.49	0.42	0.32	0.18	0.90	1.00

Principal components analysis did confirm that a substantial dimensionality reduction could be made. Because of the estimation precision brought about by the 72,373 pixels in the data set, all eight Y-variables were significant, but a q of four or five would be an acceptable subset. The principal components variance and percentage values show that the percentages of the total variance recovered when q equals four or five components are 98.2 and 99.0 respectively (Table 3). In comparison, the four MSS channels which account for the greatest percentage of the total variance are 2, 6, 7, and 8, and that percentage is only 64.3 (Table 1).

The correlations between X-variables and Y-variables were computed (Table 4) and their interpretation follows the same reasoning as for any simple correlations. In this case, statistical evaluation is uninformative since the estimation of the correlations is so precise that even those less than .1 are statistically significant. For the first principal component, which accounts for 59.4% of the total variance, all channels are strongly positively correlated, with the exception of channel 4 (.8 to 1.1 μm of the Oct. '72 image). This result is typical of the first principal component in biometrics applications and citations of a number of such examples are available (Seal, 1964, p. 122). The interpretation of similar results in the biometrics context is that the first component is a measure of over-all size or magnitude.

Table 3
Statistics Resulting from Principal Components Analysis of the Eight Channel Multidate Landsat Data

Principal Component	1	2	3	4	5	6	7	8
Variance	669.6	273.4	118.3	45.4	8.3	5.5	3.7	2.5
Percentage of Total Variance	59.4	24.3	10.5	4.0	0.8	0.5	0.3	0.2
Cumulative Percentage of Total Variance	59.4	83.7	94.2	98.2	99.0	99.5	99.8	100.0

Table 4
Correlations of MSS Channels with Principal Components

Channel	Principal Component							
	1	2	3	4	5	6	7	8
1	0.89	0.36	0.06	0.19	-0.17	0.01	0.04	-0.08
2	0.89	0.39	0.10	0.21	0.07	0.05	-0.02	0.04
3	0.73	-0.27	0.59	0.05	0.06	-0.18	0.05	-0.00
4	0.21	-0.68	0.68	-0.13	-0.06	0.10	-0.04	0.01
5	0.86	0.34	-0.14	-0.29	-0.17	-0.06	-0.01	0.11
6	0.81	0.42	0.02	-0.40	0.09	0.04	0.04	-0.04
7	0.86	-0.41	-0.27	-0.04	0.01	-0.04	-0.10	-0.04
8	0.63	-0.72	-0.29	0.05	0.01	0.04	0.07	0.02

The correlations of MSS channels with principal components decrease from one component to the next, with fewer and fewer standing out against the others. The two MSS channels least correlated with the first principal component are most strongly correlated with the second component. These are MSS channels 4 and 8, the .8 to 1.1 μ m band for each date. Principal component three is composed mainly of the contrast between the two dates for the two infrared bands (MSS channels 3, 4, 7, and 8). Similarly but less evident for the fourth principal component is the contrast between dates for the pairs of visible bands (MSS channels 1, 2, 5, and 6). Such interpretations can be applied to develop an understanding of which characteristics in the data contribute to each component.

Data Transformation Using the Principal Components

Transformation of the eight X-variables into five Y-variables accounting for 99.0% of the total variance was done by using transformation coefficients (Table 5). The general form of the equation is the same as a multiple linear regression equation, with each of the five Y-variables being computed using the coefficients in the corresponding column in the table. For example:

$$Y_2 = .236X_1 + .352X_2 - .155X_3 - .443X_4 + .186X_5 + .293X_6 - .317K_7 - .616X_8$$

Table 5
Transformation Coefficients for First Five Principal Components

Channel	Principal Component				
	1	2	3	4	5
1	0.3673	0.2356	0.0623	0.2988	-0.6192
2	0.5162	0.3524	0.1367	0.4761	0.3453
3	0.2705	-0.1545	0.5160	0.0697	0.2140
4	0.0896	-0.4432	0.6746	-0.2011	-0.2193
5	0.2984	0.1860	-0.1125	-0.3873	-0.5239
6	0.3578	0.2928	0.0172	-0.6849	0.3510
7	0.4261	-0.3170	-0.3234	-0.0820	0.0658
8	0.3446	-0.6158	-0.3722	0.1137	0.0396

The application of the principal component transformation coefficients reduced the dimensionality of each of the 72,373 pixels in the multidate data set from eight channels of MSS data, to five orthogonal "transformed" axes. There are two major justifications for applying such a transformation to the data. These are the resulting reduction in dimensionality, and the generation of orthogonal coordinate axes. The advantages of this reduction in dimensionality can be summarized as: (a) a reduction in all subsequent computer processing costs; (b) a reduction in the number of black and white and color composite images required to visually interpret the information; and (c) there is some evidence in the literature that a reduction in dimensionality improves classifier performance (Marks and Dunn, 1974). The attributes of orthogonal coordinate axes are related to the assumption of orthogonality implicit in the euclidean distance classifier which was used in this study.

Classification Results Using the Transformed Data

Using the same training areas defined during supervised classification of the MSS data, new "spectral signatures" were obtained for each category. These transformed signatures were then utilized to reclassify the study area, using a common euclidean distance (i.e., threshold) for all categories. The resulting thematic map yielded a noticeable reduction in the number of residential false alarms in agricultural areas, while residential areas continued to be properly represented. These preliminary observations were documented and verified by choosing three known agricultural areas where considerable ag./residential confusion existed using the MSS data. The results of these comparisons substantiated a 3 to 1 reduction in the number residential false alarms in these agricultural test sites (Table 6).

Table 6
Summary of Misclassified Residential Pixels in Agricultural Sites

	Oct. '72	April '73	Merged	Transformed
Site 1	13	6	13	6
Site 2	7	10	10	1
Site 3	50	27	29	10
Summary Totals	70	43*	52	17

*Best MSS False Alarms = 43 vs. 17, or ~ 3 to 1.

Using another, more time consuming approach, a detailed comparison of the signatures and euclidean distances of separation was conducted for each residential and agricultural category of each data set (i.e., Oct '72, April '73, merged, and transformed). Individual thresholds were then calculated to minimize the possibility of "confusion" during a reclassification. In this instance, there was an 8 to 1 reduction in the number of residential false alarms, with the transformed data having a total of only 5 false residential pixels in the three test sites. These results appear to confirm the expected improvement in the performance of the euclidean distance classifier when operating on a reduced number of orthogonal axes.

SUMMARY

Based on the results of this study, the following generalizations can be made. Principal components analysis of geometrically corrected, multidate Landsat imagery seems to be a useful technique for discriminating agricultural and residential land cover in the urban-rural transition zone. Eight MSS bands of temporal data were utilized to accentuate seasonal variations, and the principal components transformation dimensionally reduced the information into orthogonal axes which are highly suited for certain classification algorithms such as euclidean distance. Plans are under way to test this technique in other U.S. cities, as ag./residential confusion seems to be a common problem associated with Landsat data regardless of geographical locale.

REFERENCES

Christenson, J. W. and H. M. Lachowski, 1975. Urban Area Delineation and Detection of Change Along the Urban-Rural Boundary as Derived from Landsat Digital Data. IN: Proceedings of the American Society of Photogrammetry (Seattle, Sept. 1976). Pp. 28-33.

Fu, K. S., 1968. Sequential Methods in Pattern Recognition and Machine Learning. Academic Press, New York.

Kan, E. P. and R. D. Dillman. 1975. Timber Type Separability in Southeastern United States on Landsat-1 MSS Data. IN: Proceedings of the NASA Earth Resources Survey Symposium (Houston, June 1975), Vol. I-A: Technical Session Presentations (Houston: Johnson Space Center). Pp. 135-157.

Marks, S. and O. J. Dunn. 1974. Discriminant Functions when Covariance Matrices are Unequal. JASA, June 1974, Vol. 69, n. 346. Pp. 555-559.

McMurtry, G. J., F. Y. Borden, H. A. Weeden, and G. W. Peterson. 1974. The Penn State ORSER System for Processing and Analyzing ERTS and other MSS Data. IN: Remote Sensing of Earth Resources, Vol. III. F. Shahrokhi, Editor. Univ. of Tennessee Space Inst., Tullahoma, Tenn. Pp. 721-740.

Podwysocki, M. H., F. J. Gunther, and H. W. Blodget. 1977. Discrimination of Rock and Soil Types By Digital Analysis of Landsat Data. NASA/GSFC, X-923-77-17. Greenbelt, Md. 37 pp.

Seal, Hilary L. 1964. Multivariate Statistical Analysis for Biologists. Methven and Co., Ltd., London.

Spearman, C. 1904. General Intelligence Objectively Determined and Measured. Am. J. Psych. 15:201-293.

Williams, D. L. 1976. A Canopy-Related Stratification of a Southern Pine Forest Using Landsat Digital Data. IN: Proceedings of the American Society of Photogrammetry (Seattle, Sept. 1976). Pp. 231-239.