PRINCIPAL COMPONENTS TECHNIQUE ANALYSIS FOR VEGETATION AND LAND USE DISCRIMINATION

Antonio Roberto Formaggio
João Roberto dos Santos
Luis Alberto Vieira Dias

CNPq - Instituto de Pesquisas Espaciais
12200 - São José dos Campos - SP - Brasil

W.G. nº 1/61

ABSTRACT

The main objective of this work was to evaluate the automatic pre-processing technique called Principal Components (PRINCO) in analysing LANDSAT digitized data, for land use and vegetation cover, on the Brazilian "cerrados". The chosen pilot area, 9°55'S, 49°50'W, scene WRS 223/67 of MSS/LANDSAT III, was classified on a GE Image-100 System, through a maximum-likehood algorithm (MAXVER). The same procedure was applied to the PRINCO treated image. PRINCO consists of a linear transformation performed on the original bands, in order to eliminate the information redundancy of the four LANDSAT channels.After PRINCO only two channels were used thus reducing the computer effort. The original channels and the PRINCO channels grey levels for the five identified classes (grassland, "cerrado", burned areas, anthropic areas, and gallery forest) were obtained through the MAXVER algorithm. This algorithm also presented the average performance for both cases. In order to evaluate the results, the Jeffreys-Matusita distance (JM-distance) between classes was computed. The classification matrix, obtained through MAXVER, after a PRINCO pré-processing, showed approximately the same average performance in the classes separability.

1. INTRODUCTION

There is a high correlation degree among the bands of the same scene obtained by the LANDSAT Multispectral Scanner (MSS). The correlation is larger between the visible channels (4, 5) and infrared channels (6, 7). This fact may be considered as information redundancy, and results in an unnecessary high computer processing cost.

According to Santisteban and Muñoz (1977), a cost reduction is possible by using the Karhunen-Loéve transformation, also known as Principal Components Analysis. This technique consists in a linear transformation on the original data set through rotation and translation in the spectral attributes space, defined by four orthogonal axis, that correspond to the four MSS channels. After this transformation, the information is related to another set of axis, non correlated and orthogonal. The redundancy is suppressed, and it is possible to use only the first and second new channels, first and second components, that bear most of the information from the original four channels.

Labrandero and Palou (1980) tested this method on spanish soils data, and concluded that the use of two channels (A, B), instead of the original four, reduced the computer processing time, and the classes discrimination degredation was minimal.

Imhoff and Petersen (n.d.), in soil studies, and Paradella and Vitorello (1981), in geological surveys, among others, also concluded that the Principal Components technique was adequate for the study of natural resources through digitized remote sensing data, like LANDSAT's MSS.

The main objective of this work is to evaluate an automatic classification technique, using the Principal Components (PRINCO), for LANDSAT data, in order to study the vegetal cover and land use of the Brazilian Central Plateau (*"Cerrado"*s region).

2. STUDY AREA

The pilot area for this study was selected in the state of Goiás (Figure 1), to the East of Bananal Island, having as central geographical coordinates 49°50'W and 09°55'S.

This 900 km² area is located at the core of the *"cerrado"* region. The relief is a dissected low-altitude plateau, cut by several flood plains from rivulets, surrounded by gallery forest (Brasil, 1981). The dominant soil types are (Brasil, 1981): concretionary allic soils, with argillic B horizon and "Latossolo Vermelho Amarelo" concretionary (Oxisol).

3. METHODOLOGY

The LANDSAT digitized data automatic analysis, orbit WRS 223/67, pass May, 14, 1983, covering the pilot area, was performed on the I-100 GE Multispectral Data Analyser.

The procedure consisted of the following steps:

- data storage on the I-100 memory;
- pre-processing of data including radiometric correction to eliminate stripping;
- zoom to the 1:100.000 scale for the pilot area;
- acquisition of trainning areas for supervised classification;
- use of the maximum-likelihood algorithm (MAXVER), developed in INPE by Velasco et al (1978), in order to obtain the spectral classes, covariance matrix and classification matrix parameters.

In addition to the above parameters, the Jeffreys-Matusita distance (JM-distance) between classes was obtained (Swain and King, 1973), and an ancillary program to present the hyperellipsoids projection in a bidimensional axis system was performed to visualize the classes definition into the attributes space (Rocha and Minamoto, 1981).

The above described procedure was applied to the original image and to a PRINCO processed image, which consisted of a rotation of the spectral axis, in order to reduce the redundancy, thus saving computer time.

A comparison of the performance of both cases is analysed, after the data were put on tabular or graphic form, in next section.

4. RESULTS AND DISCUSSION

From LANDSAT-4 automatic data analysis, five vegetation and land use classes were identified on the pilot area (Figura 2): grassland, "cerrado", burned areas, anthropic areas and gallery forest.

Table 1 presents the grey levels average values for each class found on the original and enhanced images.

TABLE 1

SPECTRAL PARAMETERS OF THE FIVE CLASSES, ORIGINAL AND PRINCO

CHANNEL / CLASS	WITHOUT PRINCO (*)				WITH PRINCO (**)			
	4	5	6	7	A	B	C	D
1. GRASSLAND	45.45	57.24	61.26	56.26	110.59	131.35	126.50	126.14
2. "CERRADO"	36.84	33.79	64.40	63.15	136.13	128.01	127.40	126.67
3. BURNED AREAS	41.01	47.92	38.97	34.19	115.85	164.51	128.99	125.89
4. ANTHROPIC AREAS	62.72	89.23	85.47	69.30	80.62	101.30	127.38	124.32
5. GALLERY FOREST	38.38	36.01	81.38	78.75	137.75	104.89	126.40	126.27

(*) Average grey levels, original MSS/LANDSAT channels.
(**) Average grey levels, transformed PRINCO channels.

Figure 3 shows the values of Table 1 in graphical form, in order to provide a better visualization. It is noted from Figure 3 that the spectral behaviour for each class is different on the original channels (4, 5, 6, 7), and on the modified channels (A, B, C, D). On the transformed image, the principal components (channels A, B) present most of the information for class discrimination. The eigenvalues obtained by PRINCO procedure show the percentage of information variance on each modified channel (A, B, C, D):

CHANNELS	EIGENVALUES	% OF INFORMATION
A	114.95	50.2
B	92.67	40.5
C	15.13	6.6
D	6.26	2.7

It can be seen that 90.7% of the information is concentrated on channels A and B.

Different spectral behaviour due to PRINCO use is present on the scene under study. Most of the information on channels 4, 5, 6, 7 are transferred to channels A, B, thus an analysis of these two channels, instead of the four original ones, leads to almost the same results with reduced computer effort for "cerrado" type areas.

On Figure 4 the sampling space for "cerrado" and burned areas classes, for MSS channels 5,7 and for PRINCO channels A, B and C, D is presented. The bidimensional plots are generated by the algorithm "Hyperellipsoids Projection" due to Rocha and Minamoto (1981), developed for INPE's I-100 System.

The two classes, "cerrado" and burned areas, were selected as a sample. Figure 4(a) shows the original space variation for channels 5,7; 4(b) for PRINCO channels A, B; and 4(c) for PRINCO channels C, D.

The classification matrix (Tables 2 and 3), obtained through the MAXVER algorithm (Velasco et al, 1978), shows that in spite of using only 90% of the information the overall classes discrimination performance was the same after the use of PRINCO. This difference is not significant, however, being the major advantage the fact that the computer effort is reduced. In case of large volumes of data, as verified by Labrandero and Palou (1980), the advantage is substancial. According to the above cited authors, the savings in computer time due to the reduction in dimensionality is proportional to $N(N+3)/[N'(N'+3)]$, where N is the number of original channels, and N' the number of used transformed channels. In the present work this reduction was different because of the use of the MAXVER algorithm. This algorithm, while reducing considerably the computer effort, has no formula to measure this time advantage. It is necessary to point out that this better overall performance for the transformed channels in the classes discrimination, using PRINCO, is not dependent on the used threshold. In this work it was used threshold 5, with tests being run with threshold 4, 5 and 6.

The JM-distance to measure the classes separability was calculated for the pilot area. It was computed with and without PRINCO. Table 4 presents the average JM-distance for all the considered classes, for all MSS channels (4, 5, 6, 7), for PRINCO (A, B), for the least correlated MSS channels (5, 6), and for the most correlated MSS channels (4, 5) and (6, 7).

The average JM-distance among the five classes (original channels) was 1.86 and 1.65 on the transformed channels. These values according to the probability of correct classification of Swain and King (1973) are of the order of 98% and 96%, respectively, which in other words is a nonsignificant difference, taking into account that using PRINCO, the computer effort was reduced.

5. CONCLUSION

The use of PRINCO on the original channels changed the spectral response of the several classes involved, in some cases producing a better spectral separability.

The overall average performance was slightly better after the use of PRINCO for vegetation and land use, according to the MAXVER algorithm for the "cerrado" region under study.

TABLE 2

CLASSIFICATION MATRIX (THRESHOLD) (MAXVER)

(WITHOUT PRINCO)

C L A S S E S	N	1	2	3	4	5
1. GRASSLAND	0.0	93.9	1.1	2.8	2.2	0.5
2. "CERRADO"	0,9	0.9	87.5	0.0	0.0	10.6
3. BURNED AREAS	0.0	0.7	0.0	99.3	0,0	0.0
4. ANTHROPIC AREAS	0.0	0,0	0.0	0,0	100.0	0.0
5. GALLERY FOREST	0.4	0.0	11,6	0.0	0.4	87.7

AVERAGE PERFORMANCE = 93,1%
AVERAGE ABSTENTTION = 0,2%
AVERAGE CONFUSION = 6,7%

TABLE 3

CLASSIFICATION MATRIX (THRESHOLD) (MAXVER)

(WITH PRINCO)

C L A S S E S	N	1	2	3	4	5
1. GRASSLAND	0,0	95,6	0,8	2,5	1.1	0.0
2. "CERRADO"	0.0	0,5	88.4	0.0	0.0	11.1
3. BURNED AREAS	0,7	0,7	0.0	98,6	0.0	0.0
4. ANTHROPIC AREAS	0,0	0,0	0.0	0.0	100.0	0.0
5. GALLERY FOREST	0.4	0.0	11,2	0.0	0.4	88.0

AVERAGE PERFORMANCE = 93.5%
AVERAGE ABSTENTION = 0.2%
AVERAGE CONFUSION = 6,2%

Regarding the JM-distance, it is better for the original four channels as expected (see Table 4). However, for any pair of channels, depending on the individual classes under study, the results vary. For instance, the original uncorrelated channels (5, 6) presented in general a very good result (see Table 4), the average been almost as good as with the four original channels, but for classes 2-5 the correlated channels on the infrared (6, 7) presented a better result.

The modified channels (A, B) presented, consistently, a better result than the correlated channels (4, 5 and 6, 7). Another advantage is that the modified channels (A, B) combine information of all the four original channels.

Depending on the target it is not possible to say "a priori" which two channels will yield a better classification. If the modified PRINCO channels are used, the chances of a good choice increase with substancial savings in computer time.

When Thematic Mapper (TM) data will be available, the computational effort could be also reduced. It is shortly planned a test of PRINCO with TM data for vegetation and land use studies.

The use of PRINCO channels (A, B) is useful for the visualization, by the interpreter, since it is possible to see clearly in two dimensions the spectral cluster of the classes under study.

This work was performed for one scene of a specific area (*"Cerrado"*), so for different areas the conclusions may differ. It is suggested, for the future, more studies to settle this matter.

TABLE 4

JM DISTANCE, WITH AND WITHOUT PRINCO

CLASSES	AVERAGE JM DISTANCE				
	MSS,4,5,6 e 7	PRINCO A,B	MSS 5 e 6	MSS 4 e 5	MSS 6 e 7
1 - 2	1.96	1.73	1.93	1.87	0.76
1 - 3	1.86	1.32	1.71	0.53	1.60
1 - 4	1.68	1.56	1.67	1.52	1.62
1 - 5	1.99	1.66	1.98	1.69	1.55
2 - 3	2.00	1.90	2.00	0.95	1.91
2 - 4	1.96	1.88	1.94	1.94	1.71
2 - 5	1.29	0.95	1.19	0.09	1.25
3 - 4	1.95	1.82	1.94	1.55	1.92
3 - 5	2.00	1.85	2.00	0.75	1.94
4 - 5	1.99	1.83	1.97	1.89	1.61
AVERAGE	1.86	1.65	1.83	1.28	1.59
STANDARD DEVIATION	0.22	0.30	0.25	0.61	0.35

REFERENCES

BRASIL. Ministério das Minas e Energia. Projeto RADAMBRASIL - Vol. 22; Folha SC.22 Tocantins; 1981.

IMHOFF, M.L.; PETERSEN, G.W. The role of LANDSAT data products in soil surveys. Institute for Research on Land & Water Resources; The Pennsylvania State University, Research Publication 105/OR (n.d.).

LABRANDERO, J.L.; PALOU, F. Application of principal component Analysis to Soil Survey in Central Spain. 14th Congress of the International Society of Photogrammetry. Hamburg, 1980.

PARADELLA, W.R.; VITORELLO, I. Application of Computerized Techniques using LANDSAT Images for Geological Studies. COGEODATA IAMG Meeting for South America, Rio de Janeiro, 1981.

ROCHA, H.L.V.; MINAMOTO, M. Projeção Gráfica de um Espaço de Atributos Quadrimensionais. São José dos Campos, INPE, maio, 1981. (INPE-2069-RPE/309).

SANTISTEBAN, A.; MUÑOZ, L. Application of image principal component Technique to the Geological Study of a Structural Basin in Central Spain. Machine Processing of Remotely Sensed Data Symposium, 1977, pp. 228-236.

SWAIN, P.H.; KING, R.C. Two Effective Feature Selection Criteria for Multispectral Remote Sensing. West Lafayette, IN, Purdue University, LARS, 1973. (LARS Information Note 042673).

VELASCO, F.R.D.; PRADO, L.O.C.; SOUZA, R.C.M. Sistema MAXVER: Manual do Usuário. São José dos Campos, 1978. (INPE-NTI/110).

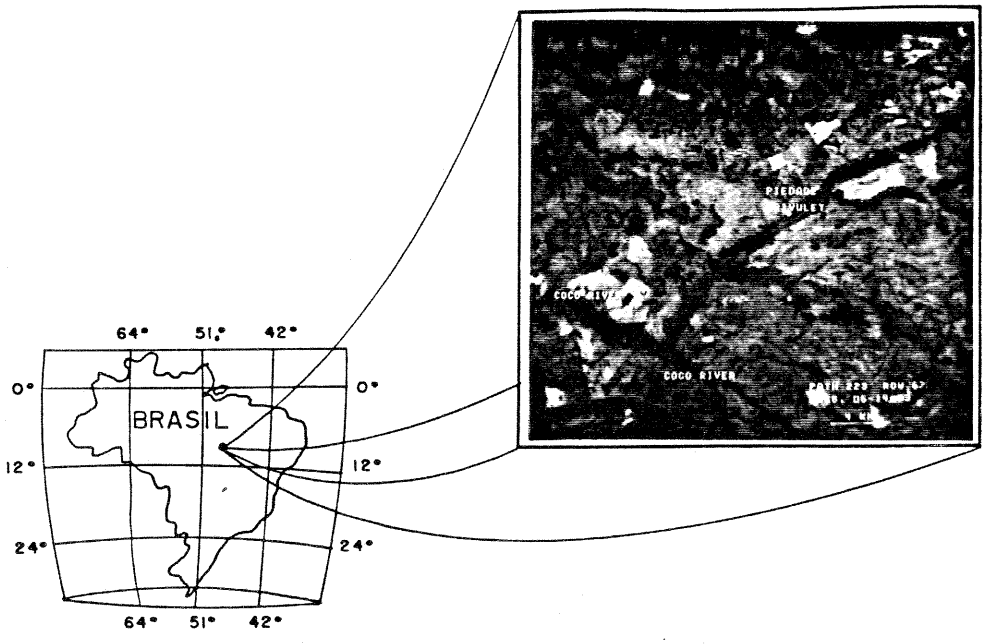

Figure 1 - Geographic location of the area under study.

Figure 2 - Distribution of the five vegetation and land use classes.

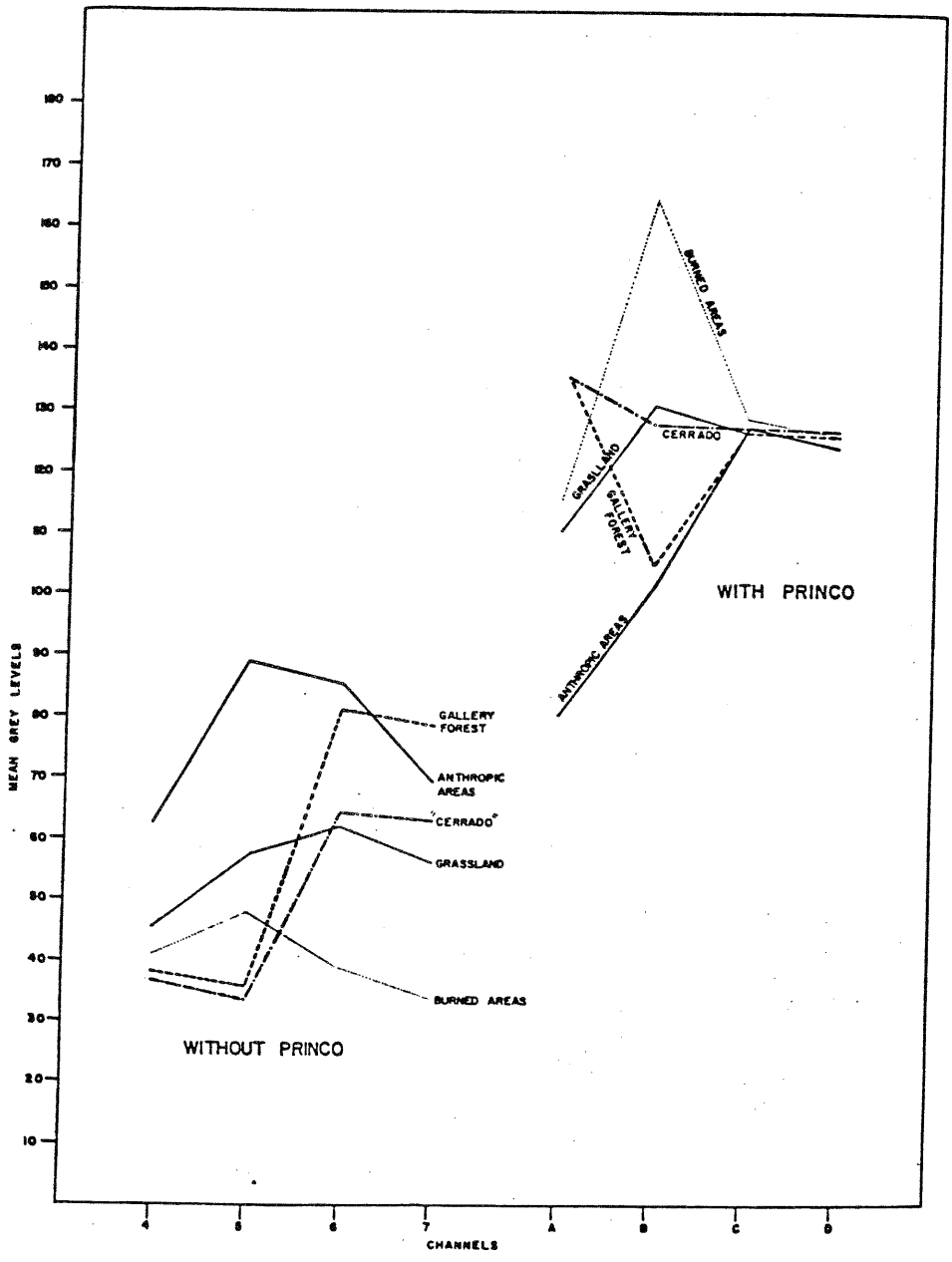

Figure 3 - Spectral distribution of the classes identified on the pilot area.

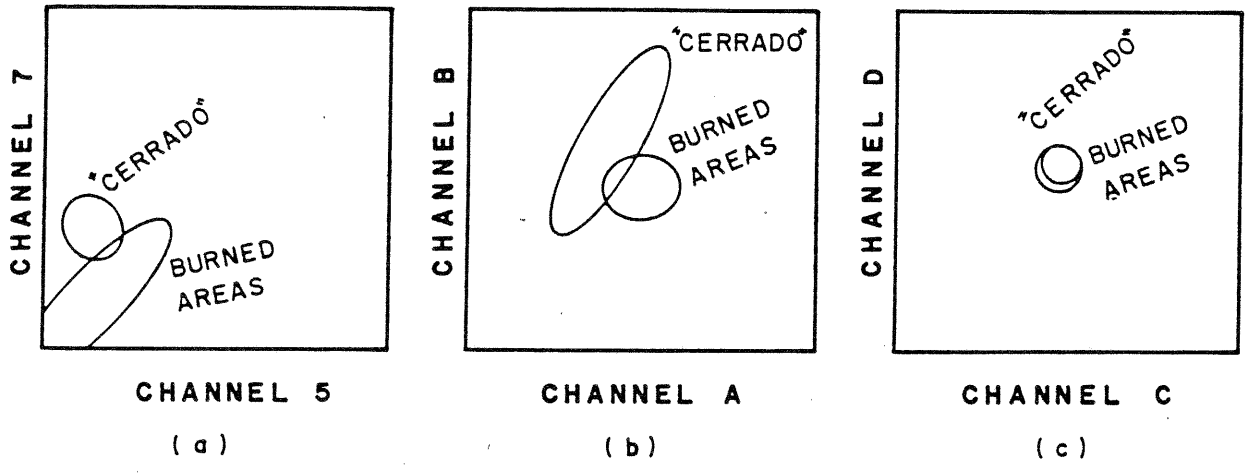

Figure 4 - Graphic representation of the bidimensional spectral space attributes: a) original data; b) components A, B (PRINCO); c) components C, D (PRINCO).