14th CONGRESS OF THE INTERNATIONAL SOCIETY OF PHOTOGRAMMETRY

Hamburg 1980

Commission no. VII

Working group no. 8

Presented paper

LABRANDERO, J.L. Instituto de Edafología y Biología Vegetal.
Serrano 115 dplo. Madrid-6. Spain.

PALOU, F. Centro de Investigación IBM-UAM. Paso de la Castellana, 4. Madrid-1. Spain.

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOIL SURVEY IN CENTRAL SPAIN.

Abstract

This work analyzes the result of applying a principal component transformation to a subimage of LANDSAT frame number 2224-10135. This subimage corresponds to a geographic area in the geologic basin of Madrid, which is formed by a variety of tertiary and quaternary sementary deposits. The analysis aims at the identification and discrimination of soil features, and is carried out with the help of pattern recognition techniques provided by the interactive system ERMAN-II.

COMMISSION NO. VII. WORKING GROUP NO. 8

LABRANDERO, J.L. Instituto de Edafología y Biología Vegetal.
Serrano, 115 dplo. Madrid-6. Spain.

PALOU, F. Centro de Investigación UAM-IBM. Paseo de la Castellana, 4. Madrid-1. Spain.

APPLICATION OF PRINCIPAL COMPONENT ANALYSIS TO SOIL SURVEY IN CENTRAL SPAIN.

Abstract. This work analyzes the result of applying a principal component transformation to a subimage of LANDSAT frame number 2224-10135. This subimage corresponds to a geographic area in the geologic basin of Madrid, which is formed by a variety of tertiary and quaternary sedimentary deposits. The analysis aims at the identification and discrimination of soil features, and is carried out with the help of pattern recognition techniques provided by the interactive system ERMAN-II.

I. INTRODUCTION.

The digital images, which are obtained by the multispectral scanners of the LANDSAT satellites, show a high degree of correlation between spectral bands. This can be interpreted as a redundancy in the data and leads us to consider that the cost of computer processing such data could be reduced by the elimination of its redundancy. The best way to do it is to use the Karhunen-Lòeve transformation to compute the "principal components" of the multiband image, which are linear conbinations of the original bands, uncorrelated to each other. The redundancy is thus suppressed and we can keep only hte first and second components without much loss of information. The third and fourth components show very little variance and part of it is due to noise.

In this paper we apply the "principal component analysis" technique-hereinafter referred as PCA- to specific LANDSAT data corresponding to an area of 2,250 sq. Km. in the provinces of Madrid and Toledo. This area must be analyzed in order to produce a digital map showing the physiographic units and the dominant soils associated to them. The results are compared with those obtained using the four original bands of the image instead of the principal components, but applying the same scheme of training and classification.

II. GEOLOGICAL FEATURES OF THE AREA.

The digital analysis with pedological objectives has been carried out in the tertiary sub-basin of Madrid (Fig. 1). Geologically, in this arid sub-basin we can distinguish chemical facies and detritical facies, both formed by gypsum, marl, clay, gravel, sand and limestone. On these lithologic materials works a typical mediterranean climate (dry warm summers and moderately cold winter). The moisture regime of the soils developing on those lithologic materials is xeric while its temperature regime is mesic.

The most characteristic physiographic units of the area investigated correspond to landforms origined by erosion and sedimentation processes in neogene lands. The existence of calcareous structural levels are very important in the development of the considered landscape. The main physiographic units are: 1) Recent deposits and lower terraces in alluvial valleys, mainly along the Tajo, Henares, Tajuña and tributary rivers, 2) Low hills developed on marl and gypsum, 3) Hillside scarpments developed on limestone, 4) Hillside scarpments origined in gypsum, 5) Mesas on marl, clay and gypsum associated with eolian deposits, 6) Mesas on limestone with different degrees of erosion.

III. LANDSAT IMAGE USED.

The LANDSAT image used was taken on September 2, 1975 and was identified as scene number 2223-10135. On this scene we have selected a subimage between lines 909 to 1,910 and pixels 1,121 to 1,630 that includes part of the provinces of Madrid and Toledo. This sub-scene was referred to UTM coordinates by means of seven ground control points and a second order mapping polynomial, obtaining a geometrically corrected image with pixels of 80 x 80 m. The geographical coordinates of ground control points were selected from topographic maps at scale 1:50,000.

IV. PRINCIPAL COMPONENT ANALYSIS

The computer program used to perform principal component analysis was developed by investigators of the Centro de Investigación UAM-IBM, including one of the authors.

The first step of the analysis consists in computing the band to band covariances of the image. From that we obtain the covariance matrix and the correlation matrix. Both matrices are then diagonalized to obtain their corresponding eigenvalues and eigen-vectors.

The second step uses the transformation that diagonalizes the covariance matrix to generate the principal components, which are new bands uncorrelated to each other. The results of the transformation must be adjusted to values between 0 and 255 in order to be stored in a byte. The program offers four options to do this adjustment. All of them fix the mean for the new bands at the value 127.5. The option we selected adjusts the variance of the first principal component in such a way that the value 127.5 is equal to 2.65 times the standard deviation. This means for a normal distribution that 1/256 of the total population will be beyond the limits 0 and 255 and hence cut off. The remaining principal components will be normalized using the same scale factor. Therefore the spread of the distribution will be much smaller. To be more precise, the higher the order of the component, the narrower will be the distribution, and the less the information contents of the band.

We apply the first step of PCA to a subimage of the original LANDSAT data, before the geometric correction.

The correlation matrix obtained gives the following coefficients

Bands 1 & 2	0.912	2 & 3	0.917
1 & 3	0.843	2 & 4	0.833
1 & 4	0.770	3 & 4	0.961

The eigenvalues of this matrix are 3.62, 0.28, 0.08, 0.02. We can interpret this by saying that 90% of the information contents of the image will be in band A (the first principal component), 7.5% of the information will be in band B and the remaining 2.5% will be in bands C and D, which we will neglect from now on.

V. CLASS ANALYSIS

The aim this analysis is to extract the relevant information contained in the LANDSAT data and put it into a form which makes it easily understandable, like tables, maps, diagrams. We have proceeded along the steps which are summarized in Fig. 2.

The first step is a selection which takes into account both the previous knowledge of the area and the quality of the LANDSAT data available. In this respect it is important that the picture corresponds to a period of the year when both the number of types of cover as well as its extent are kept to a minimum. The training field selection was difficult and complicated. We used conventional soil maps of different scales, aerial photography, geologic maps and interpretation of the LANDSAT image, based on the experience acquired on the field. We selected 56 training fields, which cover 23.5% of the whole subimage and represent all the possible variations of the interesting classes. The next step is a reconsideration of the training fields, which reduces its number from 56 to 47. To do that, we have computed the means of all training fields and made two-dimensional plots: band A versus B, and band 5 versus 7. The fields, whose centers were significantly far from those of other fields belonging to the same group, were eliminated. Next comes the grouping of similar fields into spectral classes. We take into account the two-dimensional plots, sum of the bands, histrograms, and ground truth. The number of spectral classes defined is 14 and we compute the statistics of these classes as the result of merging the pixels of the fields assigned to each one of them. We can now use the statistics to perfom classification a) of the training fields; b) of the whole subimage. The results of such classifications will be discussed in the following sections. To summarize the remaining steps we can say that after interpretation of the previous results we can group the 14 spectral classes into 7 information classes, on which the final results will be based. These classes can be given a meaning as physiographic units and soil associations.

VI. CLASSIFICATION RESULTS

To perform classification we use the maximum likelihood rule, implemented as a part of the system ERMAN-II, in the Centro de Investigación UAM-IBM. The rule consists in assigning a pixel to the class to which the probability of belonging is the highest, and the probability function for each class is a multivariate normal function, whose parameters are those of the training pixel distribution. This is, therefore, a pixel by pixel classification, which can be run on any part of the image that we define previously as a field, or a set of fields. In particular, we can estimate the significance of the results by the consideration of the performance obtained in the classification of the training fields, or, even better, other fields belonging to the class, but not used in the training process, which are called test fields.

We must give an interpretation to the results of the classification in terms of soils or soil associations, taking into account that the dominant soils present in our study area are classified as follows:

ORDERS	SUBORDERS	GREAT GROUPS
Alfisols	Xeralfs	Haploxeralfs Rhodoxeralfs
Inceptisols	Ochrepts	Xerochrepts
Entisols	Orthents	Xerorthents
	Fluvents	Xerofluvents

The 14 classes which have been used for the classification were defined partly on the basis of their spectral properties. With them we have produced a classification map of the area with 14 colors. The map was inspected on a color display monitor, in order to give a definitive interpretation to each of the classes. This interpretation is given in table I. It is apparent that some of the classes have detected vegetation instead of soils, but the natural correlation between one and the others makes the information useful.

Table I also shows how the 14 spectral classes, have been grouped into 7 information classes, in terms of which the results will be given. The reason to do this is the need to give meaningful results. For that it is necesary to reduce the probability of misclasification, what we accomplish by means of mergin statistical classes into information classes, on the basis of its statistical neighbourhood and its interpretation. The 7 information classes are explained in table II, where we also give the estimate of their relative importance inside the study area, basedon the classification of the full area. These results are given in two forms: a percent of the total surface and a number of hectareas. There is also two sets of results, one corresponding to the classification done on the bands A and B of PCA, the other corresponding to the classification done on the four original bands. They do not disagree significantly in comparison.

Regarding the comparison between the 7-class digital map and conventional soil maps, we observe that the results are very good in classes AA, EB and PA. On the opposite side the classes MA, MF and DA tend to show a certain degree of mixture between them.

VII. EVALUATION AND CONCLUDING REMARKS.

Our purpose is to evaluate the interest of using the principal component bands A and B in classification, against using the original spectral bands of the LANDSAT image. The data to be compared for this, is included in tables III and IV. The data is the same in each table, each one pertaining to a different classification. Table III corresponds to the classification with bands A and B and table IV to classification with bands 4,5,6,7. In each case we have classified the training fields in order to obtain the class confusion matrix or class performance matrix. The entries to the left of the table represent the set of training fields associated with a given class. The pixels contained in one of these sets of fields have been classified and assigned to one class or another, as specified along horizontal lines in the matrix. When the majority of elements are assigned to the same class of the training fields, the performance is high. On the opposite side, if the results are distributed more or less ramdomly among several classes the confusion is high.

The comparison between both tables shows that the performance is slightly but consistently better in the case of the four spectral bands. The overall performance rises from 53% to 61% and the average performance by class from 59.7% to 65.4%. The highest improvement corresponds to class PA which increases performance 61% to 75.6%.

The loss of performance with the use of bands A and B, must be weighted against the advantages of the method. The first advantage is the saving of computer time in the classification process, which is important when the amount of data is large. The CPU time consumed by the classification algorithm discussed, is proportional to N x (N + 3), N being the number of bands used by the classifier. In our case the classification time is reduced by a factor of 2.8, and the reduction will be more significant in the case of images with more spectral bands. Another advantage of the method is the bidimensionality of the data after the PCA transformation. This is important because it makes it easy to visualize how the data group into clusters, making the process of classification easier to understand by the analyst. It is also important because it allows to use bidimensional classification tables, the use of which can reduce the CPU time consumed in classification by two orders of magnitude.

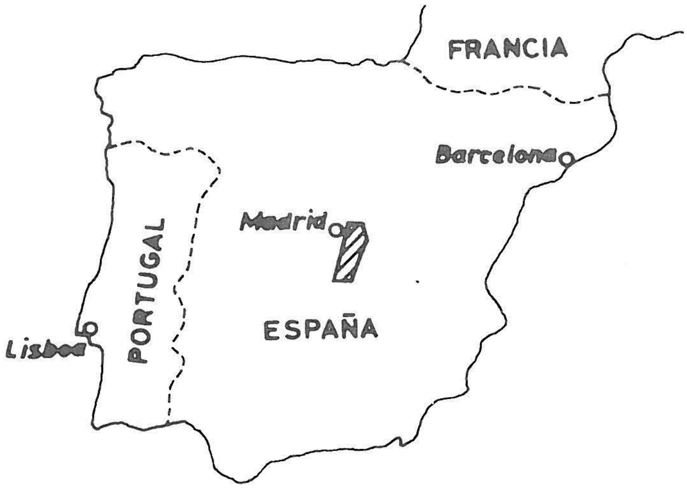

Fig. 1.- Location of study area

CHOICE OF STUDY AREA

SELECTION OF TRAINING FIELDS (56)

REFINEMENT OF TRAINING DATA (47)

DEFINITION OF SPECTRAL CLASSES (14)

CLASSIFICATION

INTERPRETATION OF RESULTS

INFORMATION CLASSES (7)

DIGITAL MAP

FINAL EVALUATION

Fig. 2.- Steps of the analysis

TABLE I. MEANING OF SPECTRAL CLASSES

Spectral Classes	Information Classes	Land Use	Dominant Great Group	Ratios A/B	Ratios V/IR
AA	AA	Horticulture Orchards	Xerofluvents	0.5	0.7
EB	EB	Natural vegetation	Xerorthents	0.3	0.8
EA	EA	Bare soil shadow	Xerorthents	0.4	1.0
EC	EA	Bare soil shadow	Xerorthents	0.7	1.0
MA	MA	Bare soil	Xerochrepts Xerorthents	1.1	1.1
MB	MA	Bare soil	"	1.1	1.1
MC	MA	"	"	1.2	1.0
MD	MA	"	"	1.3	1.1
ME	MF	Bare soil	Xerochrepts	1.4	1.0
MF	MF	"	"	1.5	1.0
DA	DA	Bare soil	Xerochrepts Haploxeralfs	1.4	1.0
DB	DA	"	"	1.3	1.0
PA	PA	Bare soil	Haploxeralfs Rhodoxeralfs	0.8	0.9
PB	PA	"	"	1.0	0.9

TABLE II. CLASSIFICATION SUMMARY

Information Classes	Physiografic Units	Subgroupes of Dominant Soils	Using Bands A,B		Using Bands 4,5,6,7	
			%	Has.	%	Has.
AA	Recent deposits and lower terraces in Alluvial valleys	Typic Xerofluvents	4.7	10560	5.2	11736
EB	Hillside scarpments on limestone	Typic Xerorthents	4.2	9454	4.5	10067
EA	Hillside scarpment on gypsum	Typic Xerorthents	17.9	40193	19.9	44611
MA	Low hills on marl, clay and gypsum	Typic Xerochrepts Typic Xerorthents	13.1	29418	15.6	35127
MF	Mesas on marl, clay and gypsum	Calcic Xerochrepts	8.0	18017	8.1	18086
DA	Mesas on eroded limestone	Typic Xerochrepts Typic Haploxeralfs	20.5	46026	18.3	40762
PA	Mesas on limestone	Typic Haploxeralfs Lithic Rhodoxeralfs	31.2	69929	27.6	62267
THR.			0.4	929	0.8	1868
TOTAL			100.0	224524	100.0	224524

TABLE III. TEST CLASS PERFORMANCE MATRIX. (USING BANDS A, B)

INFOR.	NO OF	NO OF	PCT.	INFORMATION CLASSES							
CLASS	TEST	SAMPS	CRCT	AA	EB	EA	MA	MF	DA	PA	THR
AA	5	2288	77.5	1773	148	76	11	1	10	257	12
EB	3	1097	73.8	38	810	165	1	0	2	81	0
EA	5	3413	71.4	37	239	2437	363	31	49	257	0
MA	9	18939	35.9	178	86	2723	6806	3303	4158	1669	16
MF	4	10469	43.1	99	14	370	1764	4512	3162	532	16
DA	7	13096	54.5	107	7	583	1194	1421	7126	2628	30
PA	14	33246	61.0	1280	343	3866	1576	98	5713	20327	43
TOTAL	47	82548		3512	1647	10220	11715	9366	20220	25751	117

OVERALL PERFORMANCE (1773 + 810 + 2437 + 6806 + 4512 + 7126 + 20327)/82548 = 53.0%

AVERAGE PERFORMANCE BY CLASS (77.5 + 73.8 + 71.4 + 35.9 + 43.1 + 54.5 + 61.0)/7 = 59.7%

TABLE IV. TEST. CLASS PERFORMANCE MATRIX. (USING BANDS 4,5,6,7)

INFOR.	NO OF	NO OF	PCT.	INFORMATION CLASSES							
CLASS	TEST	SAMPS	CRCT	AA	EB	EA	MA	MF	DA	PA	THR
AA	5	2288	82.8	1894	117	137	62	1	14	44	19
EB	3	1097	76.1	38	835	153	6	1	2	61	1
EA	5	3413	80.8	61	162	2758	284	52	24	69	3
MA	9	18939	38.8	296	72	4110	7351	3194	3359	503	54
MF	4	10469	46.4	153	7	906	2278	4861	2117	45	102
DA	7	13096	57.2	74	9	322	1353	647	7492	3136	63
PA	14	33246	75.6	257	345	1104	890	28	5325	25122	175
TOTAL	47	82548		2773	1547	9490	12224	8784	18333	28980	417

OVERALL PERFORMANCE (1894 + 835 + 2758 + 7351 + 4861 + 7492 + 25122)/ 82548 = 61.0%

AVERAGE PERFORMANCE BY CLASS (82.8 + 76.1 + 80.8 + 38.8 + 46.4 + 57.2 + 75.6)/7 = 65.4%

BIBLIOGRAPHY

1. Donker, NHW and Mulder, N.J. "Analysis of MSS digital imagery with the aid of principal component transform". ITC-Journal 1977-3, p. 434-465. (1977).

2. Guerra Delgado et al. "Mapa de suelos de España, Península y Baleares". Escala 1:1.000.000. Inst. Nac. Edaf. y Agro., Madrid (1968).

3. Instituto Geográfico y Catastral. "Thematic Mapping, Land Use, Geological Structure and Water Resources in Central Spain". Project no. 28760. Final Report to NASA. (1976).

4. Labrandero, J.L. y Monturiol, F. "Fisiografía y Suelos de la región de Alcalá". Trabajos sobre Neógeno-Cuaternario. pp. 129-135. Madrid (1977).

5. Labrandero, J.L. and Palou, F. "Soil Inventory in Central Spain by Digital Analysis of Landsat Data". Proceedings of the International Symposium on Remote Sensing for Observation and Inventory of Earth Reosurces and the Endangered Environment. pp. 2309-2318. International Archives of Photogrammetry. Vol. XXII-7. (1978).

6. Palou, F. and Ramirez, F. "IMACNPCA. Computer program to transform a multiespectral image into its principal components" IBM-UAM image processing library.Umpublished.

7. Rebollo, M., Orti, M., and Caramarasa, J.M. "Supervised and Unsupervised Classification of the Delta of the Ebro River: Land use study using LANDSAT Data". Centro de Investigación UAM-IBM, SCR-01.77 (1977).

8. Santisteban, A. and Muñoz, L. "Principal Components of a Multispectral Image: Application to a Geological Problem". IBM J. Res. Develop. Vol. 22. No 5. (1978).

9. User's Guide. "ERMAN-II System". Earth Resources Laboratory. IBM Federal Systems División. Houston, Texas. (1976).