The application of selective principal components analysis (SPCA) to a Thematic Mapper (TM) image for the recognition of geomorphologic features configuration

P. A. SILJESTRÖM†, A. MORENO†, K. VIKGREN‡ and L. M. CÁCERES†

†Instituto de Recursos Naturales y Agrobiología (CSIC), apartado 1052, 41080 Sevilla, Spain

‡Chalmers Centre for Marine Research & Technology, S 41296 Göteborg, Sweden

Abstract. Selective principal components analysis (SPCA) has been applied to highly- and/or little-correlated subgroups of bands. Its usefulness was demonstrated in two ways. First, the final result is a false colour composition based on the first order principal component of each highly correlated subgroup of bands, the resulting image containing more than 95 per cent of the total variance of the six TM bands used. Secondly, the second order principal component of pairs of little-correlated bands will show the information that is unique for each band. Both types of analysis have been applied to characterize the geomorphological units at a site in SW Spain. Both methodologies have demonstrated to be very useful in a difficult to access area, with high vegetation diversity covering quite different geomorphic features.

1. Introduction

The Thematic Mapper (TM) sensor on board Landsat 5 has been demonstrated to be a powerful tool in geological, geomorphological and vegetation remote sensing studies (Drury 1986, Chuvieco 1990). It comprises seven bands along the spectrum, six of which have an improved spatial resolution compared with the MSS, and also an improved spectral resolution compared with both MSS and SPOT. The high spatial resolution (30 m × 30 m for bands 1, 2, 3, 4, 5, and 7) permits the mapping of topographic and other structural features. On the other hand, spectral resolution is important for discerning the different geology, geomorphology, and vegetation of the area, as well as its temporal and status variability.

In general, image processing concentrates on securing a better visual interpretation and information extraction of the images, so that the methodology used will depend on the outcomes to be achieved. Frequently the aim is to map as much information as possible into a reduced subset of images for their analysis or colour composing. In this case, standard principal components analysis (PCA) is often used. However, this results in an information loss (of unused components) and the additional problem of dealing with false colour image interpretation. In other cases, objectives focus on information that is unique to each spectral band. Mapping this

spectral difference or contrast (Chavez and Kwarteng 1989) and interpreting it has been shown to be very useful in geological studies.

The objectives of our Technical Note are to show the visual interpretation and information extraction obtained using both methodologies (i.e., PC transformation for data dimensionality reduction, versus PC for enhancing information unique to each band) in a TM data set, using bands 1, 2, 3, 4, 5, and 7.

2. The study area

The study area is located in Huelva province (in southwestern Spain), and comprises the mouth of the River Piedras and surrounding area (figure 1), holding two big villages inside, Cartaya and Lepe (C and L in figure 1, respectively). The TM image was taken in October 1991 and the scene studied occupies 350 × 500 pixel (10·5 km × 16·5 km). Several geomorphological units can be differentiated: the external part of the basin is characterized by old (Lower Pleistocene) river terraces belonging to the general Tinto–Odiel river system (Cáceres 1995). These terraces are constituted by quartzitic conglomerates and coarse sands rich in iron, even forming iron crusts in certain places. The vegetation covering this soil is a dense pine (*Pinus pinea* L.) forest with patches of Eucalyptus sp. Nearer to the river and in relationship with it, younger (Middle–Upper Pleistocene) terraces appear, in a patchwork with outcrops of Mio–Pliocene marls and silts. This unit is characterized by having a nutrient-rich, sandy soil, covered by intensive crops (strawberry, citrus, etc.). Filling the river estuary, there is a tidal marsh protected by a sand spit advancing parallel to the coast, generated by the strong littoral dynamics of this area (Dabrio *et al.*

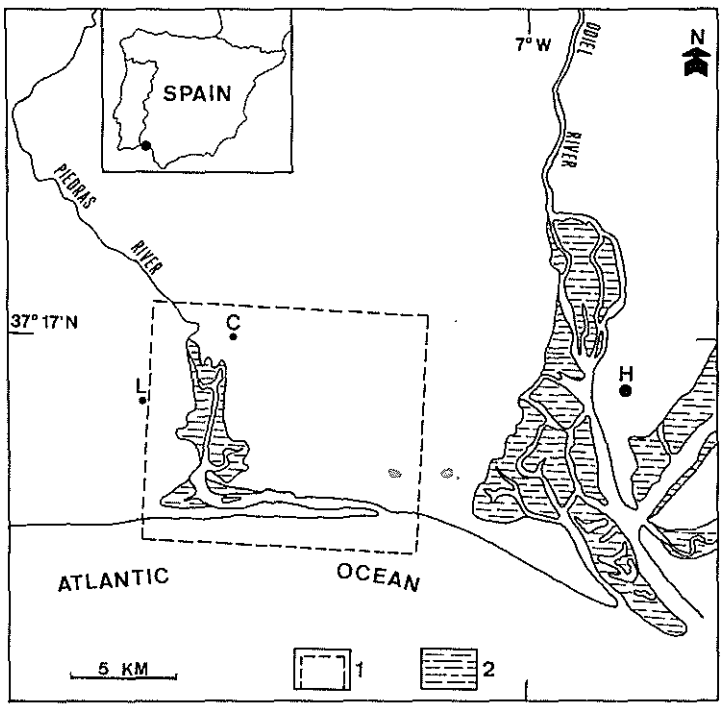


Figure 1. Geographical location of the study area. H = Huelva (city), A = Ayamonte (village), C = Cartaya (village), L = Lepe (village), 1 = study site, 2 = marsh areas.

1980, Cáceres 1995). The marsh can be well defined from its fine-textured soil, high moisture content, and typical patterns of the drainage, as well as from its characteristic halophytic and hydrophytic vegetation.

3. Methodology

Standard principal components analysis (also called Karhunen–Loeve) is used for reducing redundant information. It permits the description of multidimensional data in such way that the first axis (or principal component) contains the maximum variance, diminishing in successive components (Jenson and Waltz 1979, Gonzalez and Woods 1992, Byrne *et al.* 1980). Therefore, the sum of the variance of all the components will be equal to the total variance contained in the input images.

Normally, the first three principal components, containing more than 95 per cent of the total six bands' variance, are used to produce a false colour image (RGB). The high variance values, as dispersion measurements, allow a better differentiation, of geomorphic and landform features in the studied area, which is very rich in different geomorphological units as previously said. The resulting image will be difficult to interpret because of the uncertainty in the data origin (Williams 1983).

Therefore, the biggest benefit achieved in this case would be a great reduction in data dimensionality. This data reduction is particularly interesting when working with a large amount of data (or high number of bands) in regular (tending to small) computer systems.

On the other hand, information belonging to higher order components—of great interest in certain cases (Byrne *et al.* 1980, Richards 1984, Siljeström and Moreno 1995)—is obviated. Sometimes, instead of operating with as much variance as possible, the interest is in the information inherent to each band. In such cases, selective principal components analysis (SPCA) has to be used. The basic difference between the standard method (PCA) and SPCA is that, while the latter uses all the bands, in the SPCA only a subgroup of them is selected. These subgroups (normally of two bands each) vary depending on the aims to be achieved.

The application of SPCA to a pair of bands will give a visual interpretable result similar to a ratio between them. Other researchers suggest that subtle differences will show up, demonstrating the method's high potential, even for multitemporal studies (Byrne *et al.* 1980, Fung and Le Drew 1987). Field work was carried out (geological, geomorphological and vegetation studies) in order to check the reliability of the results.

3. Results and Discussion

3.1. *Data dimensionality reduction*

One of the principal objectives of PCA is data dimensionality reduction, maintaining the variance as high as possible. Data dimensionality reduction will help when working with many bands at a time (with a large amount of data), limited computing memory or time, or if there is any limitation in the process to be applied (such as in a false colour composition, where only three bands can be used). This goal can also be achieved with SPCA, using subgroups of highly correlated bands.

A correlation analysis between the six studied bands was done, shown in table 1. Based on these data, two different (highly correlated) subgroups were chosen: bands 1, 2 and 3 (in the visible range of the spectrum) with a correlation coefficient over 0·937 in all cases, and bands 5 and 7 (in the infrared), with a correlation of 0·974. Band 4 is left by itself because its spectral characteristics.

Table 1. Correlation matrix for the six Landsat TM bands used.

TM band	1	2	3	4	5	7
1	1·00	—	—	—	—	—
2	0·97	1·00	—	—	—	—
3	0·94	0·98	1·00	—	—	—
4	0·71	0·80	0·85	1·00	—	—
5	0·79	0·85	0·91	0·93	1·00	—
7	0·84	0·89	0·92	0·88	0·97	1·00

The principal components transformation was performed on each of both subgroups. Afterwards, a false colour composition (RGB) was applied based on the first component of each group (PC1 of visible bands was assigned to blue, PC1 of mid-infrared to red), and band 4 (assigned to green). This appeared to be quite similar for a visual interpretation to another composition based on a band chosen in each spectral range, which did not contain as much information as the first one (98 per cent of the six bands' variance). As this image has been done for demonstrating the good results for visual interpretation that can be obtained applying PCs to a RGB composition, the simple RGB image (based on the assignment of band 1 to blue, band 4 to green, and band 7 to red), is obviated in the present paper. It is to be noticed that the PC transformation is particularly advised in the case of a high correlation between bands, which is not the present case. Therefore, working with a study site like the present one, where different spectral areas show acceptable correlation values, this type of methodology should be used just in case of doing a multiband data analysis with limited computing capabilities.

The resulting colour image (figure 2) shows the three main geomorphic units composing the landscape: cultivated Mio–Pliocene silts and marls in a patchwork with recent fluvial terraces (in medium-blue tones) with non-vegetation-covered areas (in light-blue tending to white), old (Lower Pleistocene) terraces covered with pines (in green to blue), and tidal marsh (dark-green) and coastal area very well outlined

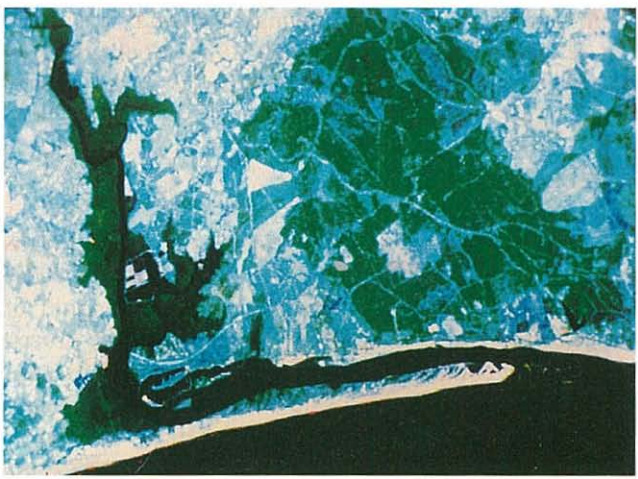


Figure 2. False colour composite (RGB) of the site studied, based on the assignment of red to the first principal component of bands TM 5 and TM 7, green to band TM 4, and blue to the first principal component of the three visible bands (TM 1, 2 and 3).

because the water appears very dark. The good definition of the submerged structures of shoals at the tip of the spit is noteworthy.

It can be seen how SPCA will help in data dimensionality reduction with little information loss. The higher order components (which have not been used) have much noise. The information loss is very small due to the high correlation of the input bands. The RGB image will be easier to interpret as the colours can be associated to certain parts of the spectrum (in a similar way as can be done on a 'simple' RGB image based on three input bands).

3.2. *Information inherent to each band (spectral contrast)*

In this type of analysis, the main objective is the characterization of the spectral contrast along the spectrum, so that the information unique to each band can be obtained. For this purpose, the selected subgroups for applying SPCA contain just one pair of bands. Only the second order principal component (PC2) is to be interpreted as a monochrome image. The PC1 will contain the common information between them (related mainly to albedo and topography). In general, the correlation of two bands is related with the spectral contrast between them. This means that two highly correlated bands will look very similar and will have a low spectral contrast. In general (and for this site), a high correlation has been defined as when the value is higher than 0·90; it is considered medium if it varies between 0·90 and 0·80, and low if less than 0·8.

Based on the correlation values between bands, several pairs were chosen:

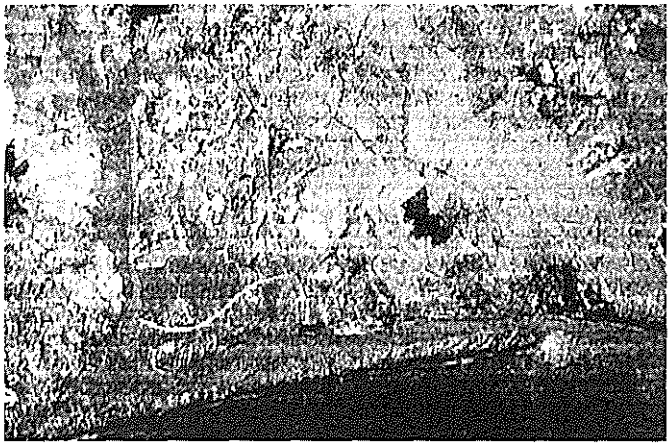
- TM 1 and TM 2 (blue against green, both in the visible);
- TM 5 and TM 7 (mid IR against mid IR);
- TM 4 and TM 7 (near IR against mid IR);
- TM 1 and TM 7 (visible against mid IR);
- TM 1 and TM 4 (visible against near IR).

From the visible part of the spectrum, band 1 was chosen because it shows the lowest correlation values. The correlation coefficient, variance and loading factors in both principal components are shown in table 2. From the corresponding images, only the PC2 has been displayed, as shown in figures 3(*a*), (*b*), and (*c*); and figures 4(*a*) and (*b*).

In some highly correlated images the results are very poor, as can be seen (as a good example) in the PC2 of TM 1 and TM 2 (figure 3(*a*). Both bands present a 0·97 correlation coefficient and the PC2 contains only 1·55 per cent of the total variance. The resulting image contains too much noise and is considered insignificant. This can also be seen at the values of the loading factors in the eigenvectors, which

Table 2. Correlation coefficients, variance corresponding to PC1 and PC2, and loading factors for each pair of TM bands used in the SPCA subgroups.

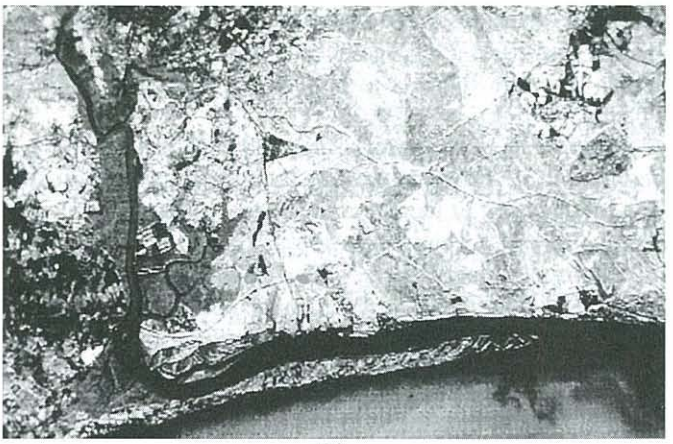
TM pair	Correlation coefficient	PC1% variance	PC2% variance	PC1 loadings	PC2 loadings
TM 1–TM 4	0·71	95·07	4·92	(0·34, 0·94)	(−0·94, 0·34)
TM 1–TM 7	0·84	97·44	2·56	(0·34, 0·93)	(−0·93, 0·34)
TM 4–TM 7	0·88	93·85	6·15	(0·68, 0·72)	(−0·72, 0·68)
TM 1–TM 2	0·97	98·45	1·55	(0·75, 0·65)	(−0·65, 0·75)
TM 5–TM 7	0·97	99·10	0·97	(0·87, 0·47)	(−0·47, 0·87)



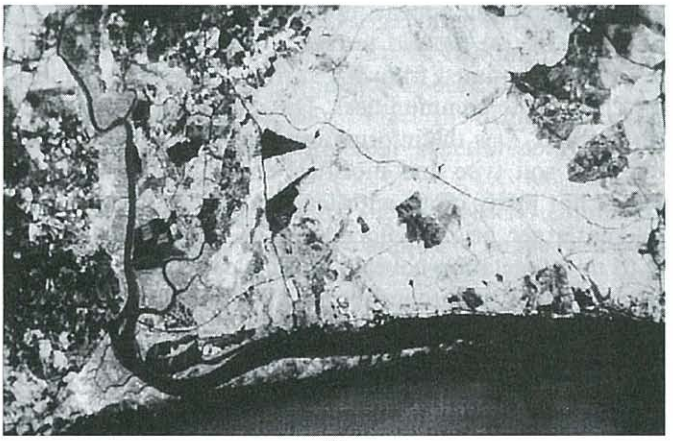
(*a*)

(*b*)

Figure 3. (*a*) Second order principal component of bands TM 1 and TM 2; (*b*) second order principal component of bands TM 5 and TM 7; and (*c*) second order principal component of bands TM 4 and TM 7.



(*a*)



(*b*)

Figure 4. (*a*) Second order principal component of bands TM 1 and TM 7; and (*b*) second order principal component of bands TM 1 and TM 4.

represent the correlation of each band with the principal components (Gonzalez and Woods 1992). These values are quite similar (−0·65, 0·75), meaning that both of them will appear in mid-grey tones.

The resulting PC2 image from processing the other highly correlated pair of bands (TM 5 and TM 7, figure 3(*b*)) offers a much better visual result, although the total variance contained is very low (0·97 per cent). The better contrast in this image is directly related with the loading factors of the eigenvectors (−0·47, 0·87), permitting a higher degree of separation. The coastal features (water border line, dunes, shoals, etc.) can be delineated, as well as the cultivated river terraces (small black and white patches) and the pine forest (mid-dark grey) growing on the old river terraces system. The urban areas (in the middle of the crops) appear as bright spots, as does the bare soil, due to its high response in band 7 (which has a loading factor of 0·87). In contrast, the pine forest and the marsh appear in a similar grey tone since band 5 (loading factor of −0·47) does not distinguish the different plant communities.

The fact that there is some quality difference between the said PC2 and that for

TM 1 and TM 2 lies in the closeness of the bands, as also stated by other authors (Chavez 1989). TM 1 and TM 2 lie together along the spectrum (0·45–0·52 μm TM 1, 0·52–0·60 μm TM 2), while there is a gap between bands 5 and 7 (1·55–1·75 μm TM 5, 2·08–2·35 μm TM 7).

The second component of the analysis applied to TM 4 and TM 7 (0·88 correlation coefficient, figure 3(*c*), shows very nicely the growing sand spit structure because of the vegetation information in band 4, and the moisture and geological information in band 7. Even the new sand bar formation at the tip of the spit can be observed, as band 4 can penetrate water slightly while the radiation at band 7 is completely absorbed (Drury 1986). The resulting PC2 image contains 6·42 per cent of the total variance, which is the highest eigenvalue of all combinations, directly related to the lowest correlation coefficient of all band-pairs tested.

As can be deduced studying the loading factors, one has always a positive value, the other a negative one. In this particular case (−0·72, 0·68), the first corresponds to the correlation between the PC2 image and band 4 and, therefore, all the information inherent to this band will appear in dark-grey colours. This can be seen in the dark tones that characterize the old terrace system, covered by a pine forest. The marsh can also be differentiated from the surrounding area, as it shows a different vegetation type (halophyte communities). The crops do not show up as well as in the PC2 of bands 5 and 7, as the information both images contain is due mainly to band 7, related with soil type and moisture content. As the image was taken in October, there are not many crops with leaves that can appear in band 4. For the same reason, some larger bright patches can be seen inside the pine forest area, corresponding to non-vegetation-covered areas (clear cuttings, new reforested areas, etc.) which will show up as light tones through band 7 (with a positive loading factor of 0·68), as does the dune system along the coast.

The PC2 of bands TM 1 and TM 7 (figure 4(*a*)), with a correlation coefficient of 0·84 and 2·56 per cent of variance, offers a good land–water discrimination. The loading factor for TM 1 (−0·93) appears with a very dark response, showing the coastal shallow waters, where their strong dynamics (currents, eddies) can be easily recognized. The dark tones correspond to the different behaviour in both bands that characterizes the water. The marsh flooding limit can be differentiated because of the soil moisture gradient present. For the same reason, the present river channel can be easily outlined. The intensive greenhouse crops (typical in this area) surrounding the principal villages (C—Cartaya and L—Lepe in figure 1) also appear as dark, well defined patches due to their water irrigation. On the other hand TM 7 with a loading factor of 0·34 offers very little information about the landscape (soil and vegetation characteristics), which appears in very light tones. Differences in vegetation, used for characterizing types of substrate, are not noticeable at all.

The PC2 corresponding to TM 1 and TM 4 (figure 4(*b*), with a 0·71 correlation coefficient and containing 4·92 per cent of the total variance, offers only a little information about the shallow waters and coastline—due to the fact that band 1 can penetrate the water to some extent—but separates very well the geomorphological units depending on their vegetation (as commented earlier, each geomorphological unit is intimately related with the vegetation communities growing on it). The pine forest area shows up in very light tones (band 4 has a loading factor of 0·34), as do a few cultivated zones. Non-vegetation-covered areas, with higher soil moisture content, appear in mid-grey tones (band 1 loads the eigenvector with −0·94). The marsh can be outlined to some extent (the outer limits are not too

clear), due to the light grey tone assigned to its dry part covered with some halophyte vegetation. Finally, urban areas show a very dark response because of their very distinctive spectral behaviour in band 1.

4. Conclusion

In this Technical Note, we have demonstrated that selective principal components analysis is useful in two aspects. First, it is helpful reducing the data dimensionality (data volume), and minimizing the loss of information. In this case, the first order principal component of highly correlated images is the one to be used. Results obtained making false colour combinations with these images are easy to interpret visually, as the origin of the information is relatively controlled. In the present case, the four geomorphological units described could be reliably differentiated by their distinct vegetation and soil moisture content and texture.

Second, it can enhance spectral contrast (or spectral difference) between two different regions in the spectrum. This enables identification of the information unique to each band on the second order principal component obtained by processing pairs of low- to medium-correlated bands. The correlation of both bands with the PC2 will show up in mid-grey with a tendency to dark or light tones depending on their negative or positive sign, respectively. In fact, the degree of correlation between two images can be directly related with their spectral contrast. The higher the correlation, the lower the contrast.

For separating geomorphological units, the most useful combinations were the TM 4 and TM 7, and TM 1 and TM 4, which happen to be the least-correlated bands.

Acknowledgments

The authors thank Juan Cara for his great help in working on the manuscript, and interesting discussions about it, as well as Roger Churchill for his always useful grammatical and vocabulary advice.

References

Byrne, G. R., Crapper, P. F., and Mayo, K. K., 1980, Monitoring land cover changes by principal components analysis of multitemporal Landsat data. *Remote Sensing of Environment*, **10,** 175–184.

Cáceres, L. M., 1995, Geomorfologia del sector occidental de la depresion del Guadalquivir. Doctoral dissertation, University of Huelva.

Chavez, P. S., Jr., 1989, Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis. *Photogrammetric Engineering and Remote Sensing*, **55,** 339–348.

Chavez, P. S., Jr., and Kwarteng, A., 1989, Extracting spectral contrast in Landsat thematic mapper and SPOT for three different sites in the Phoenix, Arizona Region. *Photogrammetric Engineering and Remote Sensing*, **54,** 339–348.

Chuvieco, E., 1990, *Fundamentos de Teledeteccion Espacial*, (Ediciones Rialp S A).

Dabrio, C. J., Boersma, J. R., Fernandez, J., Martin, J. M., and Polo, M. D., 1980, Dinamica costera en el Golfo de Cadiz: sus implicaciones en el desarrollo socio-economico de la region. *Proceedings I Reunion Geologia Ambiental y Ordenacion del Territorio*, pp. 1–19.

Drury, S. A., 1986, *Image Interpretation in Geology* (London: Allen & Unwin).

Gonzalez, R. C., and Woods, R. E., 1992, *Digital Image Processing* (Reading: Addison Wesley).

Jenson, S. K., and Waltz, F. A., 1979, Principal components analysis and canonical analysis in remote sensing. *Proceedings of American Society of Photogrammetry, 45th Annual Meeting*, (Falls Church, Va.: ASPRS), pp. 337–348.

RICHARDS, J. A., 1984, Thematic mapping from multitemporal image data using the principal components transformation. *Remote Sensing of Environment*, **16**, 35–46.

SILJESTRÖM, P. A., and MORENO, A., 1995. Monitoring burnt areas by principal components analysis of multi-temporal TM data. *International Journal of Remote Sensing*, **16**, 1577–1587.

WILLIAMS, R. S., JR., 1983, *American Society of Photogrammetry Manual of Remote Sensing*, vol. 2 (Falls Church, Va.: ASPRS), vol. 2, pp. 1667–2100.