MSEL
Modelling in Science Education and Learning

Modelling in Science Education and Learning
Volumen 6(2), No. 7, 2013.
Instituto Universitario de Matemática Pura y Aplicada

Principal component analysis applied to remote sensing

Javier Estornell, Jesus M. Martí-Gavilá, M. Teresa Sebastiá, Jesus Mengual
Universitat Politècnica de València
jaescre@cgf.upv.es, jemarga@cgf.upv.es, mtsebastia@hma.upv.es, jemencu@hma.upv.es

Abstract

El objetivo principal de este artículo es mostrar una aplicación del análisis de componentes principales (PCA) que se utiliza en dos grados de la ciencia. En particular, se utilizó el análisis de PCA para obtener información de la cobertura del suelo a partir de imágenes de satélite. Tres imágenes Landsat fueron seleccionadas a partir de dos áreas que se encuentran en los municipios de Gandia y Vallat, ambos en la provincia de Valencia (España). En la primera área de estudio, se utilizó una sola imagen Landsat del año 2005. En la segunda área de estudio, se utilizaron dos imágenes Landsat tomadas en los años 1994 y 2000 para analizar los cambios más significativos en la cobertura de la tierra. Según los resultados, el segundo componente principal de la imagen de área Gandia permitió la detección de la presencia de vegetación. El mismo componente en el área de Vallat permitió detectar un área forestal afectada por un incendio forestal. En consecuencia, en este estudio se confirmó la viabilidad del uso de PCA en teledetección para extraer la información territorial.

The main objective of this article was to show an application of principal component analysis (PCA) which is used in two science degrees. Particularly, PCA analysis was used to obtain information of the land cover from satellite images. Three Landsat images were selected from two areas which were located in the municipalities of Gandia and Vallat, both in the Valencia province (Spain). In the first study area, just one Landsat image of the 2005 year was used. In the second study area, two Landsat images were used taken in the 1994 and 2000 years to analyse the most significant changes in the land cover. According to the results, the second principal component of the Gandia area image allowed detecting the presence of vegetation. The same component in the Vallat area allowed detecting a forestry area affected by a forest fire. Consequently in this study we confirmed the feasibility of using PCA in remote sensing to extract land use information.

Keywords: ACP, información territorial, teledetección, Landsat.
PCA, land use, remote sensing, Landsat

1 Introduction

The framework of this study is related to the contents of the optative subject "Applied Remote Sensing". This subject is taught since the 2004/05 academic year in the fourth year of the Degree in Environmental Sciences and since the 2011/12 academic year in the Master's Degree in Assessment and Environmental Monitoring of Marine and Coastal Systems, in the Escola Politècnica Superior de Gandia. The use of Information and Communication Technology contributes to adapt to the student in a technological society which is one of the challenges of the European Higher Education Area [1]. In this subject, satellite images are used to analyse some environmental variables such as land uses, water quality, cover vegetation. In addition, these tools allow to detect and characterize changes in these variables. To perform these applications it is necessary to use mathematical tools in image processing, like the PCA analysis applied to Landsat images.

Landsat images can register the energy reflected by the terrestrial surface at different intervals of the electromagnetic spectrum with wavelengths ranging from the blue region to the mid-infrared ($1:0.45-0.52\mu m; 2:0.52-0.60\mu m; 3:0.63-0.69\mu m; 4:0.76-0.90\mu m; 5:1.55-1.75\mu m; 7:2.08-2.35\mu m$). The information from these wavelength ranges is stored in independent bands. Each band is handled as a matrix structured image where their pixels contain a digital number (DN) which is related with the electromagnetic energy reflected or emitted from a target. In remote sensing two bands located very close in the electromagnetic spectrum will have a high correlation.

The purpose of using a principal component analysis is to reduce the dimensionality of the data, in this case the number of original bands, to maximize the amount of information from the original bands into the least number of principal components. A set of correlated variables (original bands) is transformed in other uncorrelated variables (principal components) which contain the maximum original information with a physical meaning that needs to be explored. This analysis has been successfully applied in Landsat images showing that the first three principal components may contain over 90 percentage of the information in the original seven bands [2]. These calculations have been widely used in remote sensing to classify the land surface [3] and detect changes [4].

2 Methodology

A Landsat image can be expressed in matrix format in the following way:

$$X_{n,b} = \begin{pmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{6,1} & \dots & x_{6,n} \end{pmatrix},$$

where n represents the number of the pixels and b the number of bands. Considering each band as a vector, the above matrix can be simplified as follows:

$$X_k = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_6 \end{pmatrix},$$

where k is the number of bands.

To reduce the dimensionality of the original bands the eigenvalues of the covariance matrix must be calculated. This matrix can be calculated as follows:

$$C_{b,b} = \begin{pmatrix} \sigma_{1,1} & \dots & \sigma_{1,6} \\ \vdots & \ddots & \vdots \\ \sigma_{6,1} & \dots & \sigma_{6,6} \end{pmatrix},$$

where $\sigma_{i,j}$ is the covariance of each pair of different bands.

$$\sigma_{i,j} = \frac{1}{N-1} \sum_{p=1}^{N} (DN_{p,i} - \mu_i)(DN_{p,j} - \mu_j),$$

where $DN_{p,i}$ is a digital number of a pixel p in the band i, $DN_{p,j}$ is a digital number of a pixel p in the band j, μ_i and μ_j are the averages of the DN for the bands i and j, respectively.

From the variance-covariance matrix, the eigenvalues (λ) are calculated as the roots of the characteristic equation,

$$\det(C - \lambda I) = 0,$$

where C is the covariance matrix of the bands and I is the diagonal identity matrix.

The eigenvalues indicate the original information that they retain. From these values the percentage of original variance explained by each principal component can be obtained calculating the ratio of each eigenvalue in relation to the sum of all them [5]. Those components which contain minimum variance and thus minimum information can be discarded.

The principal components can be expressed in matrix form:

$$Y_6 = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{pmatrix} = \begin{pmatrix} w_{1,1} & \dots & w_{1,6} \\ \vdots & \ddots & \vdots \\ w_{6,1} & \dots & w_{6,6} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_6 \end{pmatrix},$$

where Y is the vector of the principal components, W the transformation matrix, and X the vector of the original data. The coefficients of the transformation matrix W are the eigenvectors that diagonalizes the covariance matrix of the original bands. These values provide information on the relationship of the bands with each principal component. From these values it is possible to link a main component with a real variable. The eigenvectors can be calculated from the vector - matrix equation for each eigenvalue λ_k,

$$(C - \lambda_k I) w_k = 0,$$

where C is the covariance matrix, λ_k is the k eigenvalues (six in our example), I is the diagonal identity matrix, and w_k is the k eigenvectors.

The PCA calculation was applied to satellite images from two areas which were located in the municipalities of Gandia and Vallat, both in the Valencia province. In Gandia a Landsat image of 2005 was used; the principal components of the 6 original bands were calculated analysing just those that avoid the loss of information and allowed to explain the presence of vegetation in study area. In Vallat two Landsat images for the years 1994 and 2000 were used and the principal components of the 12 bands (6 for the year 1994 and the other 6 to 2000) were calculated. The components which contain information about the changes in the study area were analysed.

3 Results and discussion

In the first study area, the first three components accounted for 99.3 % (Table 2) of the variance in the original data. The rest of the components were discarded. In the eigenvector matrix, a contrast was observed in the signs and values of the coefficients of the principal component 2 (Table 2). The coefficient of the original band 4 had a high positive value (0.6433) while the remaining bands had negative values or near zero ($-0.4445 - 0.0725$). Since Landsat band 4 contains information of the energy reflected in the near-infrared region and in vegetation zones it presents a very high reflectance value, this could indicate that this component can detect the presence of vegetation.

Component	Eigenvalue	Percentage
1	4295295	88.82
2	368381	7.62
3	142297	2.94
4	18486	0.38
5	8916	0.18
6	2359	0.05

Table 1: Eigenvalues of the variance-covariance matrix in the 2005 image.

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Band 1	0.13113	-0.44457	0.37446	0.66230	-0.09101	-0.44499
Band 2	0.21978	-0.36930	0.38117	0.05236	0.11714	0.80843
Band 3	0.29848	-0.42249	0.22179	-0.68170	0.27908	-0.37500
Band 4	0.53754	0.64332	0.51132	-0.03199	-0.17423	-0.06602
Band 5	0.58073	0.07259	-0.48540	0.28889	0.58173	0.00113
Band 6	0.46820	-0.26143	-0.41018	-0.09707	-0.72893	0.05859

Table 2: Eigenvectors of the variance-covariance matrix in the 2005 image.

The most remarkable results are observed in the component 2 image. As seen in Figure 1 there is correspondence between areas with high values (white colours) in the component 2 and the presence of citrus orchards observed in the orthophotographs. On the other hand, urban areas can be visualized in dark colour. While the other components (1 and 3) are often related with the brightness of the original data (component 1) and the water content (components 3) which are less relevant in this context.

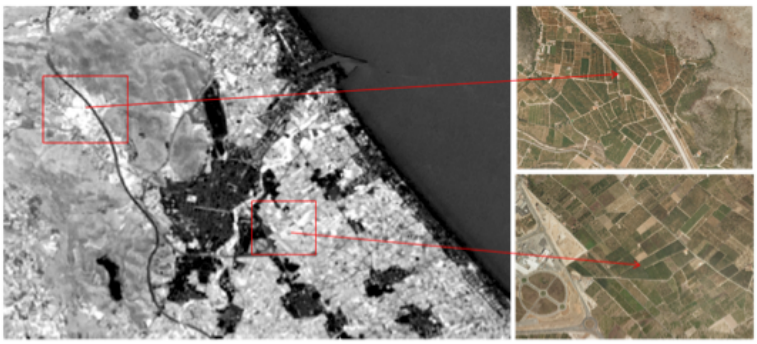

Figure 1: Component 2 image (left) and ortophotographs where agricultural areas of Gandia can be visualized with more detail (right).

In the second study area (Vallat), it was observed that the component 2 presents negative coefficients in the eigenvectors of the variance-covariance matrixes corresponding to the bands of the 1994 image (-0.0889 to -0.4388) and positive coefficients for the bands corresponding to the 2000 image ($0.0872 - 0.7479$) (Table 3). This contrast in signs is due to the fact that this component provides information of changes in the land's surface in the analysed period.

Year		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1994	1	0.2427	-0.1081	0.2133	0.3519	-0.5386	-0.1471	0.3647	0.1970	0.0075	-0.1640	-0.0082	0.4994
	2	0.1903	-0.0889	0.1462	0.1912	-0.2517	0.0549	-0.0067	0.2028	-0.0002	0.3193	0.6432	-0.5240
	3	0.2666	-0.1970	0.1690	0.1120	-0.2959	0.0165	-0.5334	-0.3506	-0.4328	0.2336	-0.3323	-0.0320
	4	0.2545	-0.1481	0.1559	0.4428	0.2726	0.5902	-0.1553	-0.0897	0.4199	-0.2397	-0.0638	-0.0037
	5	0.4426	-0.4388	0.1987	-0.4684	0.2790	-0.1331	0.1305	-0.0831	0.2384	0.3333	0.0819	0.2413
	6	0.2777	-0.2462	0.1527	-0.2560	0.0455	-0.0858	0.0788	0.2451	-0.2692	-0.6565	-0.1404	-0.4174
2000	1	0.2457	0.7479	0.5714	-0.2072	0.0488	0.0696	-0.0344	0.0016	-0.0163	0.0047	0.0083	0.0442
	2	0.1847	0.1269	-0.0708	0.2051	-0.0258	-0.3348	0.1529	0.1698	0.3906	0.2804	-0.5848	-0.4053
	3	0.2624	0.1512	-0.2093	0.1344	0.0262	-0.5655	-0.3116	-0.3754	0.2656	-0.3417	0.3156	0.0262
	4	0.2518	0.0872	-0.1097	0.4359	0.5915	-0.1576	0.1843	0.1139	-0.5209	0.1402	0.0579	0.1030
	5	0.4426	0.2003	-0.5786	-0.2104	-0.1458	0.2520	-0.2694	0.4412	-0.0026	0.0391	-0.0056	0.1762
	6	0.2703	0.1423	-0.3084	-0.0918	-0.1722	0.2764	0.5537	-0.5869	-0.1093	-0.0126	-0.0065	-0.1837

Table 3: Eigenvectors of the variance-covariance matrixes in the $1994 - 2000$ images.

In this case the land use change was due to the elimination of the vegetation cover due to a forest fire that occurred in 1999. In figure 2 we overlapped the forest fire cartography and the component 2 of the principal component analysis considering the 12 bands. A good correspondence is observed between the perimeter of the area affected by the forest fire (red colour) and the white colours of the component 2 image.

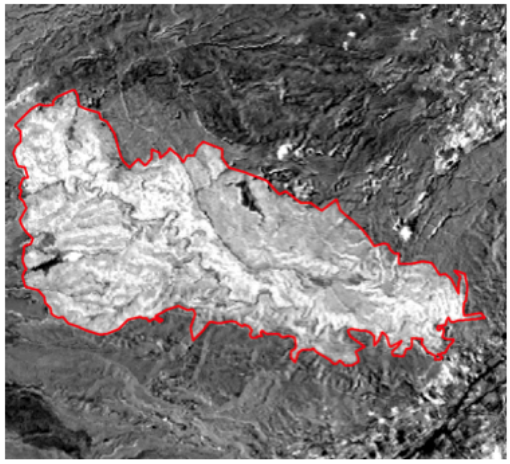

Figure 2: Overlapping between the forest fire cartography (red line) and the component 2 calculated from the 12 initial bands.

4 Conclusions

The findings already proven in other studies on the possibility of obtaining information from the land's surface using the method of PCA applied to satellite imagery. This is a good example of the importance of mathematics analysis to handle information and communication technology.

References

[1] E. Zhu, M. Kaplan. M. Landsat(12th ed.) In W. J. McKeachie (Ed.). Boston. MA: Houghton Mifflin. (2006).

[2] Fundamentals of Remote Sensing. Canada Centre for Remote Sensing. Accessed 23th October 2012. http://www.nrcan.gc.ca/earth-sciences/geography-boundary/remote-sensing/fundamentals/1814

[3] X. Jia, J. A. Richards. Segmented Principal Components Transformation for Efficient Hyperspectral Remote-Sensing Image Display and Classification. IEEE Transactions on Geoscience and Remote Sensing. **37**(1). pp. 538–542. (1999).

[4] J. R. Eastman, M. Filk. Long sequence time series evaluation using standardized principal components. Photogrammetric Engineering and Remote Sensing. **59**(6) 991–996. (1993).

[5] E. Chuvieco. Teledetección Ambiental. La observación de la Tierra desde el espacio. Barcelona. Editorial Ariel S.A. (2010).