:: Análisis de Componentes Principales en Teledetección. Consideraciones estadísticas para optimizar su interpretación ::

Fecha de Publicación: 19/08/2005

Autores del Artículo:

Prof. Susana Beatriz Ferrero[1], Prof. María Gabriela Palacio[1], Lic. Osvaldo R. Campanella[2]

Universidad Nacional de Río Cuarto
Facultad de Ciencias Exactas, Físico-Químicas y Naturales
Ruta 8 km. 603 - (5800) Río Cuarto - Córdoba
[1] Departamento de Matemática.
[2] Departamento de Geología.
Comentarios sobre el artículo: sferrero@exa.unrc.edu.ar

Los autores pertenecen al equipo de Río Cuarto, que está formado por **Osvaldo R. Campanella** (geólogo, especialista en SIG aplicado a la geología ambiental), **Marcelo Uva** (licenciado en computación y programador) y **Susana B. Ferrero** (matemática, y Magíster en estadística aplicada). Todos ellos poseen un elevado grado de cualificación profesional y pertenecen a la Universidad Nacional de Río Cuarto, Córdoba, Argentina.

Son contratables por cualquier usuario o institución que necesite la ejecución rápida y eficiente de algún trabajo en el ámbito de los Sistemas de Información Geográfica y Teledetección, poniendo a su disposición una elevada cualificación técnica y el dominio de un amplio abanico de herramientas, entre las que se encuentran ArcView (incluyendo trabajos de programación), ENVI, SAS, programación en IDL, Delphi y otros lenguajes. Me consta que trabajan rápido y bien.

Para pedir presupuestos y solicitar contrataciones contactar con Osvaldo R. Campanella.

Resumen:

En este trabajo se ha aplicado el Análisis de Componentes Principales (A.C.P.) a una subimagen LANDSAT 5 TM que comprende a la ciudad de Río Cuarto, provincia de Córdoba, Argentina. Se analizan en detalle consideraciones tales como: ponderación o "peso" de cada banda y número de componentes a ser usados.
Se presenta el desarrollo del A.C.P. basado en la matriz de varianza-covarianza y en la de correlación -- y sus correspondientes autovalores y autovectores -- y consideraciones para decidir cual de ellas usar. También se muestran criterios para seleccionar el número de componentes

Summary:

A Principal Components Analysis (PCA) to a LANDSAT 5 TM that includes the city of Río Cuarto, province of Cordoba, Argentina, has been performed. Considerations are analyzed in detail such as: "weight" of each band and number of components to be used.
It is presented the development of the PCA based on the covariance and correlation matrices -- and their eingenvalues and eigenvectors – and considerations in order to decide which one is the most adequated. It is also suggested criteria to select the number of main components to be used.If the goal is

principales a ser usados.
Si el objetivo es ponderar de la misma manera a todas las bandas intervinientes, debe usarse la matriz de correlación, en cambio si se desea dar mayor relevancia a aquellas bandas que tienen mayor varianza, debe usarse la matriz de varianza-covarianza.
Se presentan tres criterios estadísticos para decidir el número de componentes principales a ser usados: **(a)** En la curva que muestra los porcentajes de variación total explicada por cada componente, considerar aquellos anteriores al punto de inflexión (con lo cual se deberían usar los tres primeros componentes en el problema de aplicación); **(b)** Considerar los componentes cuyos autovalores son mayores que el autovalor promedio (con lo cual se deberían usar los dos primeros componentes en el problema); **(c)** Usar los componentes cuyos coeficientes de correlación con las bandas son grandes en valor absoluto (con lo cual se deberían usar los dos primeros componentes en el problema).

PALABRAS CLAVE: Teledetección – Análisis de Componentes Principales.

to assign the same weight to all the intervening bands, the correlation matrix must be used, whereas if it is desired to give greater relevance to those bands that have greater variance, the covariance matrix must be used.
Three statistical criteria have been shown in order to decide the number of main components to be used: **(a)** In the curve that shows the percentage of total variation explained by each component, it is necessary to consider the components previous to the saddle point (in this case the three first components should be use); **(b)** To consider the components whose eigenvalues are greater than the average (and then the two first components should be used); **(c)** To use those components whose correlation coefficients are high in absolute value (and then the two first components should be used, again).

KEY WORDS: Remote sensing – Principal Component Analysis

1. INTRODUCCIÓN

Teledetección Espacial

Las diferentes coberturas de la superficie terrestre (campos cultivados, roca desnuda, agua) reflejan la radiación electromagnética (REM) que les llega desde el sol, con distintas intensidades o niveles digitales (N.D.) de acuerdo a la región del espectro (firma espectral). Este fenómeno es el eje fundamental de la TELEDETECCIÓN.
Los datos son adquiridos en soporte digital y en formato numérico (N.D. para cada elemento de la superficie y para cada banda). Esto abre un enorme campo para la aplicación de la estadística a las ciencias de observación terrestre. Diversos métodos del análisis multivariado son utilizados en Teledetección.
Cualquier imagen puede pensarse como una matriz tridimensional, en la que cada una de las intersecciones de una fila y una columna corresponde a una posición geográfica discreta, y por lo tanto a un píxel, y la tercera dimensión está dada por la banda a la cual corresponde ese píxel. En otros términos, cada nivel digital asociado a un píxel puede denotarse como $ND_{i,j,k}$ donde i es el número de fila, j es el número de columna y k es la banda.
Teniendo presente este carácter matricial de cualquier imagen numérica, se pueden realizar sobre ella transformaciones y operaciones estadísticas. Por ejemplo, con los

datos de una imagen digital se pueden calcular medidas de tendencia central y dispersión en cada banda, aumentar el contraste, cambiar su orientación numérica (rotación de la matriz), realizar combinaciones aritméticas entre bandas, sintetizar varias bandas reduciendo la información redundante (componentes principales) o discriminar grupos con N.D. homogéneos dentro de la matriz (clasificación).

Transformaciones Lineales de los Datos de Imágenes

El carácter digital de las imágenes y su forma vectorial permite generar nuevas imágenes aplicando transformaciones lineales a una o varias imágenes. Las nuevas imágenes representan una descripción alternativa de los datos, en la cual los nuevos N.D. de un píxel están relacionados con sus N.D. originales a través de una operación lineal. La imagen transformada puede destacar ciertas características que no era posible discernir en los datos originales o, alternativamente, preservar la información esencial contenida en la imagen en un número menor de dimensiones. Las transformaciones pueden llevarse a cabo para realizar un mejoramiento de la imagen o como un análisis previo a la aplicación de técnicas de clasificación.

2. ANÁLISIS DE COMPONENTES PRINCIPALES

El objetivo del Análisis de Componentes Principales (A.C.P.) es resumir un grupo amplio de variables en un nuevo conjunto (más pequeño) sin perder una parte significativa de la información original (Chuvieco, 1996). Para el usuario final de productos de teledetección, el objetivo del A.C.P. es construir una o varias imágenes que incrementen su capacidad de diferenciar distintas coberturas. Es por ello que al realizar una composición color resulta interesante usar, en lugar de algunas bandas de la imagen, los componentes principales 1, 2 y 3 en la secuencia RGB respectivamente.

El A.C.P. puede aplicarse como realce previo a la interpretación visual o como procesamiento anterior a la clasificación. En general, esta técnica incrementa la eficiencia computacional de la clasificación porque reduce la dimensionalidad de los datos.

Por otra parte, desde el punto de vista estadístico, el A.C.P. facilita una primera interpretación sobre los ejes de variabilidad de la imagen, lo que permite identificar aquellos rasgos que aparecen en la mayoría de las bandas y aquellos otros que son específicos de algún grupo de ellas (Chuvieco, *opcit*). Este trabajo se refiere a casos en los que interesa identificar la información común a la mayoría de las bandas, que está presente en los primeros componentes.

El A.C.P. también es usado en aplicaciones multitemporales con el objeto de detectar cambios en distintas fechas. En este caso los primeros componentes resultantes del análisis no son los más interesantes ya que recogen información común a las distintas fechas (la estable). Los últimos componentes ofrecen la información no común (el cambio) que es lo que interesa en este contexto (Chuvieco, *opcit*).

Algebraicamente, el A.C.P. genera nuevas variables (componentes), mediante una combinación lineal de las p variables originales (bandas). Aunque se requieren los p componentes principales para reproducir la variabilidad total, muchas veces la mayor parte de ella está contenida en un número menor de componentes m. En ese caso,

reemplazando las p bandas por los m componentes, se reduce la dimensionalidad del problema conservando casi la totalidad de la información.

En teledetección, la adquisición de imágenes en bandas adyacentes del espectro implica, con frecuencia, detectar información redundante (en apariencia las bandas de la imagen se visualizan de manera similar) . Por ello, los N.D. de los píxeles de una banda pueden presentar una importante relación con los de otra, resultando una o más de una de ellas irrelevantes.

Aunque la imagen puede arreglarse en una matriz tridimensional, para realizar el A.C.P. se utiliza una matriz bidimensional. Formalmente, los N.D. de los *n* píxeles de una imagen en *p* bandas pueden arreglarse en una matriz X_{pxn},

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2n} \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ x_{p1} & x_{p2} & \cdot & \cdot & \cdot & x_{pn} \end{bmatrix}$$

La i-ésima fila de la matriz contiene los *n* niveles digitales de la i-ésima banda. Denominando $X_i = [x_{i1}\ x_{i2}\ ...x_{in}]^t$ para i=1,2,..., p, resulta $X^t = [X_1, X_2, ... , X_p]$.

Como el A.C.P. es un análisis descriptivo no requiere que *X* tenga distribución normal multivariada. Si *X* tuviera esta distribución se podría realizar inferencia (Mardia *et al*, 1982).

El estudio de la relación entre bandas, que es la base del A.C.P., puede realizarse de dos maneras:

· Con la matriz de varianza-covarianza Σ_x:

$$\Sigma_X = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \cdot & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdot & \cdot & \cdot & \sigma_{2p} \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \sigma_{p1} & \sigma_{p2} & \cdot & \cdot & \cdot & \sigma_{pp} \end{bmatrix}$$

en la que los elementos de la diagonal son las varianzas de los N.D. en cada banda:

$$\sigma_{ii} = \frac{1}{n} \sum_{k=1}^{n} \left[x_{ik} - E(x_i)\right]^2 \quad \text{con} \quad E(x_i) = \frac{1}{n} \sum_{k=1}^{n} x_{ik}$$

y los elementos fuera de la diagonal son las covarianzas entre los N.D. de dos bandas:

$$\sigma_{ij} = \frac{1}{n} \sum_{k=1}^{n} [x_{ik} - E(x_i)][x_{jk} - E(x_j)]$$

Como la covarianza entre la banda *i* y la *j* es la misma que entre la banda *j* y la *i* ($\sigma_{ij} = \sigma_{ji}$) la matriz Σ_x es simétrica. Cuando hay relación lineal entre los N.D. de dos bandas las covarianzas son grandes en comparación con las varianzas, por eso es que esta matriz sirve para estudiar la relación entre pares de bandas.

· Con la matriz de correlación ρ_x:

$$\rho_X = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdot & \cdot & \cdot & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdot & \cdot & \cdot & \rho_{2p} \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \cdot & & & & \cdot & \\ \rho_{p1} & \rho_{p2} & \cdot & \cdot & \cdot & \rho_{pp} \end{bmatrix}$$

en la que los elementos son los coeficientes de correlación lineal de Pearson:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}\ \sigma_{jj}}$$

Los elementos de la diagonal son unos porque son las correlaciones de cada banda consigo misma. Como la correlación entre la banda *i* y la *j* es la misma que entre la banda *j* y la *i* ($\rho_{ij} = \rho_{ji}$) la matriz ρ_x es simétrica. Cuando hay relación lineal entre pares de bandas las correlaciones son cercanas a 1 ó a –1.

Cuando no hay relación entre bandas ambas matrices son diagonales (los elementos fuera de la diagonal son ceros). En este caso cada banda aporta información diferente y por lo tanto el A.C.P. sería innecesario (en teledetección esta situación es poco común).

El objetivo del A.C.P. es generar un nuevo sistema de coordenadas en el espacio multiespectral en el cual los datos pueden ser representados sin correlación, de tal manera que la matriz de varianza-covarianza sea diagonal en el nuevo sistema de coordenadas.

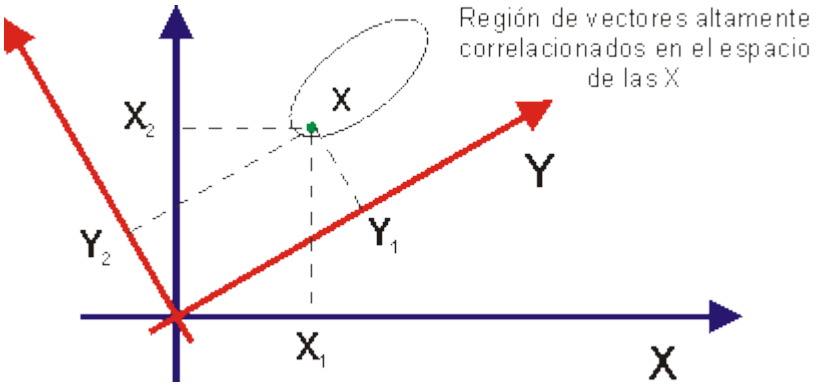

Figura 1: Ilustración de un sistema de coordenadas modificado en el cual los vectores tienen componentes no correlacionados

Componentes principales obtenidos usando la matriz de varianza-covarianza

Como se dijo, los componentes principales son nuevas variables $Y_1, Y_2, ..., Y_p$ que se obtienen como combinaciones lineales de las variables originales (bandas):

$$Y_1 = a_1^t X = a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p$$
$$Y_2 = a_2^t X = a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p$$
$$\vdots$$
$$Y_p = a_p^t X = a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{pp}X_p$$

Como los Y_i son combinaciones lineales de los X_i, tienen:

$$Var(Y_i) = a_i^t \Sigma_X a_i \qquad\qquad Cov(Y_i, Y_k) = a_i^t \Sigma_X a_k$$

De todas las combinaciones lineales posibles, los componentes principales son aquellas que no están correlacionadas y tienen máxima varianza. Como la varianza se incrementa multiplicando el vector de coeficientes a_i por una constante, para que esta combinación lineal sea única es conveniente usar los vectores de coeficientes normalizados, es decir con longitud 1 ($a_i^t a_i = 1$).

De esta manera:

⇨Primer componente principal = combinación lineal ($a_1^t X$) que maximiza

$Var(a_1^t X)$ sujeto a que $a_1^t a_1 = 1$.

⇨Segundo componente principal = combinación lineal ($a_2^t X$) que maximiza

$$\text{Var}(a_2^t X) \text{ sujeto a que } a_2^t a_2 = 1$$
$$\text{y que } \text{Cov}(a_1^t X , a_2^t X) = 0$$
$$\vdots$$

➪i-ésimo componente principal = combinación lineal ($a_i^t X$) que maximiza

$$\text{Var}(a_i^t X) \text{ sujeto a que } a_i^t a_i = 1$$
$$\text{y que } \text{Cov}(a_i^t X , a_k^t X) = 0$$
$$\text{para } k < i$$

Como en el caso de las bandas, la matriz de varianza-covarianza de los componentes (Σ_Y) es simétrica (por ser $\text{Cov}(Y_i,Y_k)= \text{Cov}(Y_k,Y_i)$, $\forall$ $i \neq k$). Además como $\text{Cov}(Y_i,Y_k)=0$ ($\forall$ $i \neq k$), Σ_Y ésta es diagonal.

El Resultado 8.1 de Johnson y Wichern (1992) muestra que el i-ésimo componente principal está dado por

$$Y_i = e_i^t X = e_{1i} X_1 + e_{2i} X_2 + ... + e_{pi} X_p \qquad \text{con } i = 1,...,p \qquad (1)$$

donde los li son los autovalores (con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$) y los e_i son los autovectores de Σ_Y.
A partir de esto:

$$\text{Var}(Y_i)= \text{Var}(e_i^t X) = e_i^t \Sigma_X e_i = \lambda_i \qquad i = 1,...,p$$
$$\text{Cov}(Y_i , Y_k)= e_i^t \Sigma_X e_k = 0 \qquad i \neq k$$

Es decir:

$$\Sigma_Y = \begin{bmatrix} \lambda_1 & \dots 0 \\ 0 & \lambda_2 \dots 0 \\ . & . \\ . & . \\ . & . \\ 0 & \dots \lambda_p \end{bmatrix}$$

Si algunos autovalores son iguales, los autovectores y por lo tanto los componentes no son únicos. Sin embargo, los autovectores correspondientes a autovalores iguales pueden elegirse de manera tal que sean ortogonales, y así los componentes son únicos.

La matriz Σ_Y contiene en la diagonal las varianzas (autovalores) de los N.D. de los píxeles en las coordenadas transformadas. Los autovalores son no crecientes, entonces la máxima varianza (en general) está en la primera componente y_1, la subsiguiente en y_2, y así sucesivamente.

Geométricamente, las combinaciones lineales representan la selección de un nuevo sistema de coordenadas $Y_1,Y_2,...,Y_p$ obtenido por rotación del sistema original con coordenadas $X_1,X_2,...,X_p$. Los nuevos ejes representan las direcciones de máxima variabilidad y proveen una descripción más simple y parsimoniosa de la estructura de covarianza. Los autovalores expresan la longitud de cada uno de los ejes (componentes), mientras que los autovectores dan la dirección de los mismos.

Por otra parte, cada *autovalor* representa la *proporción de información que retiene* el componente principal asociado (lo cual es útil para decidir qué componentes son más interesantes), en tanto que el autovector indica la *ponderación que debe aplicarse a cada una de las bandas para obtener el componente principal* (equivalentes a los coeficientes de regresión en una transformación lineal estándar, siendo las bandas de la imagen las variables independientes y los componentes principales, las dependientes). El valor absoluto del elemento e_{ji} del autovector e_i indica el *grado de contribución de la banda j al componente principal i* definido por la transformación lineal.(Chuvieco, *opcit*).

La transformación por componentes principales así definida recibe el nombre de **Transformación de Hotelling o de Karhunen-Loève** (Richards y Jia, 1999).

Como se mencionó al comienzo de esta sección, aunque se requieren los *p* componentes principales para reproducir la variabilidad total muchas veces la utilización de un número menor de componentes ($m<p$) conserva casi la totalidad de la información reduciendo la dimensionalidad del problema.

Para decidir el número de componentes principales a utilizar se requieren ciertos conceptos que serán presentados a continuación.
El siguiente resultado muestra que la variabilidad total de las *p* bandas es la variabilidad total de los *p* componentes, es decir la transformación preserva la varianza total (Resultado 8.2 de Johnson y Wichern, *opcit*).

$$\sigma_{11} + \sigma_{22} + \mathrm{L} + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 + \mathrm{L} + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$$

Esto se justifica de la siguiente manera:
La matriz Σ_X es definida positiva, entonces por el Teorema de Descomposición Espectral puede escribirse:

$$\Sigma_X = \sum_{i=1}^{p} \lambda_i \, e_i e_i^{t}$$

Si los autovectores normalizados son las columnas de una matriz ortogonal G (matriz de la transformación) y Λ es la matriz diagonal de los autovalores resulta $\Sigma_X = G \, \Lambda G^t$. Entonces

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \mathrm{tr}(\Sigma_X) = \mathrm{tr}(G \Lambda G^t) = \mathrm{tr}(\Lambda G G^t)$$

esto último por propiedades de la traza.
Además $GG^t = I$ (por ser G una matriz ortogonal) y $\Lambda = \Sigma_Y$ entonces

$$\sum_{i=1}^{p} Var(X_i) = tr(\Sigma_Y) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i)$$

Como la varianza poblacional total es $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \lambda_1 + \lambda_2 + ... + \lambda_p$, *la proporción de varianza total poblacional explicada por el k-ésimo componente es:*

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + L + \lambda_p}$$

Como los autovalores se ordenan en forma no creciente, la *eficiencia del ajuste* de los datos originales por los primeros *m* componentes ($m \leq p$) es:

$$\frac{\sum_{k=1}^{m} \lambda_k}{\lambda_1 + \lambda_2 + L + \lambda_p}$$

En particular cuando se consideran todos los componentes la proporción de variación explicada es 1.

El coeficiente de correlación entre el componente Y_k y la banda X_i es:

$$\rho_{Y_k, X_i} = \frac{\mathbf{e}_{ik} \sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$$

Los criterios presentados en este trabajo para decidir cuántos componentes principales se deberían seleccionar son:

> **(a)** En la curva que muestra los porcentajes de variación total explicada por cada componente versus los componentes, considerar aquellos anteriores al punto de inflexión.
> **(b)** Considerar los componentes cuyos autovalores son mayores que el autovalor promedio.
> **(c)** Usar los componentes cuyos coeficientes de correlación con las bandas son grandes en valor absoluto.

Estos criterios son utilizados en el Problema de Aplicación.
Antes de realizar el A.C.P. a una imagen real se presenta un ejemplo sencillo para mostrar el cálculo de los componentes principales.
Ejemplo (Richards y Jia, *opcit*): Suponga que los N.D. de 6 píxeles en 2 bandas son:

Banda 1	1	2	4	5	5	3	2
Banda 2	2	2	3	4	5	4	3

El sentido y la fuerza de la correlación lineal entre dos bandas puede representarse gráficamente mediante un diagrama de dispersión. Cuanto más se aproximan los puntos a una recta mayor será el grado de correlación entre bandas.

En la Figura 2 los N.D. muestran asociación lineal, que podría investigarse analíticamente con el coeficiente de correlación de Pearson, cuyo valor es 0.7609.

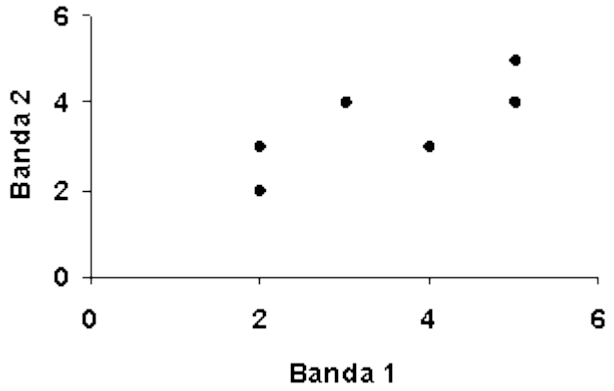

Figura 2: Diagrama de dispersión entre las bandas 1 y 2.

Para este ejemplo $X = \begin{bmatrix} 2 & 4 & 5 & 5 & 3 & 2 \\ 2 & 3 & 4 & 5 & 4 & 3 \end{bmatrix}$

La matriz de varianza-covarianza de los datos es $\Sigma_X = \begin{bmatrix} 1.9 & 1.1 \\ 1.1 & 1.1 \end{bmatrix}$

Para determinar los componentes principales es necesario encontrar los autovalores y autovectores de Σ_X. Los autovalores están dados por la solución de la ecuación $|\Sigma_X - \lambda| = 0$, es decir $\begin{vmatrix} 1.9-\lambda & 1.1 \\ 1.1 & 1.1-\lambda \end{vmatrix} = 0$ ó equivalentemente $(1.9-\lambda)(1.1-\lambda) - 1.1^2 = 0$, es decir $\lambda^2 - 3\lambda + 0.88 = 0$, que da por resultados λ_1=2.67 y λ_2=0.33.

Entonces la matriz de varianza-covarianza en el nuevo sistema es

$$\Sigma_Y = \begin{bmatrix} 2.67 & 0 \\ 0 & 0.33 \end{bmatrix}$$

Para encontrar la matriz G de la transformación, se deben calcular los autovectores normalizados asociados a los autovalores l1 y l2 . Considerando el primer autovalor λ_1=2.67, resulta que el vector solución de la ecuación $[\Sigma_X - \lambda_1 I]e_1 = 0$ es $e_1 = (e_{11}\ e_{21})^t$, y sustituyendo adecuadamente resulta el siguiente sistema

$$\begin{cases} -0.77e_{11} + 1.1e_{21} = 0 \\ 1.1e_{11} - 1.57\ e_{21} = 0 \end{cases}$$

del cual se obtiene $e_{11}=1.43\ e_{21}$, lo que indica que existen infinitas soluciones para el sistema. Como, además, los autovectores deben estar normalizados

$e_1^t\, e_1 = e_{11}^2 + e_{21}^2 = 1$. Esta ecuación conjuntamente con el sistema anterior da por resultado $e_1=(0.82\ 0.57)^t$. De manera similar $e_2=(-0.57\ 0.82)^t$. Por lo que:

$$G = \begin{bmatrix} 0.82 & -0.57 \\ 0.57 & 0.82 \end{bmatrix}$$ y los componentes principales son: $Y_1 = 0.82X_1 + 0.57X_2$, $Y_2 = -0.57X_1 + 0.82X_2$

Para los datos del ejemplo resulta:

$$Y = \begin{bmatrix} 2.78 & 4.99 & 6.38 & 6.95 & 4.74 & 3.35 \\ 0.50 & 0.18 & 0.43 & 1.25 & 1.57 & 1.32 \end{bmatrix}$$

Como los valores del primer componente son grandes comparados con los del segundo, la mayor variabilidad se da en la dirección del primer componente principal, lo cual indica que contiene la mayor parte de la información. Más específicamente, como $\lambda_1 =$ 2.67 el primer componente contiene el $\frac{2.67}{2.67+0.33}.100 = 89\%$ de la variación total. De esta manera, el primer componente muestra un alto contraste visual. Por otra parte el segundo componente es perpendicular al primero (porque no están correlacionados) lo que indica que contiene información no incluida en el primer componente. Todo esto se confirma en la Figura 3, donde se han graficado los datos en los dos sistemas (bandas y componentes principales).

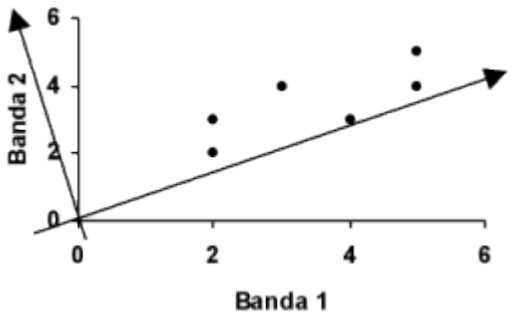

Figura 3: Diagrama de dispersión en los dos sistemas de coordenadas.

Componentes principales obtenidos usando la matriz de correlación

En la sección anterior se realizó el A.C.P. a partir de los datos originales. Para otorgar a todas las variables la misma importancia se deben estandarizar las mismas. Si X* es la matriz de variables estandarizadas, Σ_{X^*} coincide con ρ_X. Por lo tanto generar los componentes principales usando ρ_X es equivalente a usar las variables estandarizadas para el análisis.

Partiendo de la matriz ρ_X se otorga la misma importancia a todas las variables, lo cual es deseable o no dependiendo de la situación particular. En este caso se debe usar la matriz X* para expresar las combinaciones lineales que determinan los componentes principales.

Se recomienda examinar el efecto que tiene sobre el resultado del análisis la estandarización de los datos originales o la transformación de unidades de alguna o algunas de las variables a fin de homogeneizar las magnitudes. No hacerlo implica suponer que la variable con mayor variabilidad es la que más influye, o sea la determinante en el análisis (Plá, 1986).

En Teledetección s*ería conveniente generar los componentes principales usando la matriz de correlación* para atenuar el efecto de la diferencia de variabilidad entre bandas, sobre todo porque ésta puede deberse a diferencias entre las medias de los N.D. de los píxeles de distintas bandas. De esta manera se da la misma ponderación a todas las bandas.

3. PROBLEMA DE APLICACIÓN

Características de la imagen utilizada

Los datos utilizados en este trabajo corresponden a una subimagen extraída de una escena completa cuyas características se consignan en la Tabla 1. La subimagen abarca en su sector central a la ciudad de Río Cuarto, provincia de Córdoba (República Argentina).

CARACTERÍSTICAS DE LA IMAGEN: LANDSAT 5 TM 229/83			
Bandas	B1, B2, B3,B4, B5, B7	Nivel de preprocesamiento	8
Hora adquisición	13:36:08 GMT	Fecha de adquisición	02/MAY/1984
Tamaño del píxel	28.5 x 28.5 m		
Acimut del Sol	43,76º	Elevación del Sol	29,19º
Número de filas	400	Número de columnas	350

Tabla 1: Características de la imagen.

Los datos fueron procesados con la aplicación del paquete IDRISI for Windows.

En las Figuras 4 y 5 se muestran la banda 5 de la subimagen y la misma banda con ensanche de contraste por ecualización de histograma con la paleta GREY256 de IDRISI for Windows, respectivamente. En la Figura 4 el elemento lineal que recorre la subimagen desde el extremo superior izquierdo hasta el centro a la derecha es el río Cuarto y la mancha oscura en el centro de la imagen corresponde al microcentro de la ciudad de Río Cuarto.

Figura 4: Banda 5 original con paleta GREY256.

Figura 5: Banda 5 con ensanche de contraste por ecualización de histograma con paleta GREY256.

Estadística descriptiva para cada banda

En la Tabla 2 se muestran los estadísticos descriptivos de las bandas que intervienen en el trabajo:

BANDA	Mínimo	Máximo	Media	Desviación Estándar	Varianza
B1	46	210	60.9389	6.2891	39.5528
B2	15	106	25.5209	3.8425	14.7648
B3	12	125	29.8761	6.2358	38.8852
B4	10	103	44.3829	8.9281	79.7110
B5	6	184	61.7784	11.3242	128.2375
B7	1	142	27.3785	7.4158	54.9941

Tabla 2: Estadísticos descriptivos de la imagen.

Se puede observar en la tabla que las mayores variabilidades entre los N.D. (desviaciones estándar o varianzas) se presentan en las bandas 4, 5 y 7, mientras que las del espectro visible (1, 2 y 3) son las que presentan menos dispersión.

En las Figuras 6 y 7 se muestran los histogramas de las bandas 2 y 5, que son la de menor y la de mayor desviación estándar respectivamente. La localización de los histogramas permite deducir la tonalidad dominante, la amplitud está relacionada con el contraste y la presencia de picos relativos es un indicador de distintas coberturas. Así la banda 2 no ofrece prácticamente contraste mostrando predominio de colores muy oscuros, mientras la 5 muestra un mayor nivel de contraste y colores algo más claros. Por otra parte, ambos ocupan sólo una pequeña porción del rango de variación permitido por el sensor, lo cual hace pensar que (para una mejor visualización) podrían

usarse técnicas de mejoramiento y realce, como el ensanche de contraste por ecualización de histogramas que fuera presentado para la banda 5 en la Figura 5.

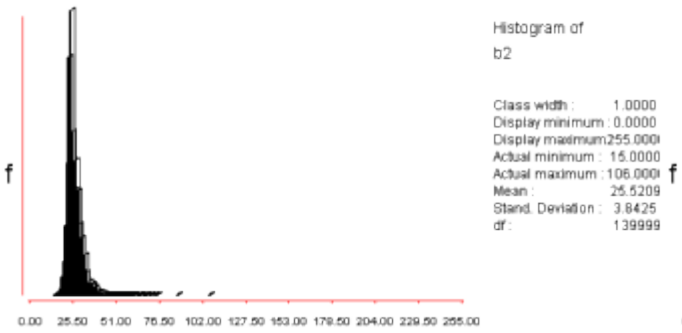

Figura 6: Histograma y estadística descriptiva banda 2.

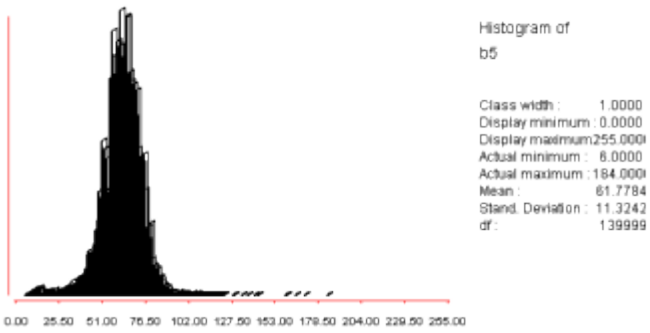

Figura 7: Histograma y estadística descriptiva banda 5.

Además de la consideración de cada banda por separado también resulta de interés tener en cuenta las relaciones entre pares de bandas, para analizar el grado de información original que aporta cada una. Este análisis puede realizarse gráficamente mediante diagramas de dispersión donde cada punto tiene como coordenadas los N.D. de cada píxel en cada una de las bandas consideradas, y completarse analíticamente calculando el Coeficiente de Correlación de Pearson. Las Figuras 8 y 9 presentan los diagramas de dispersión entre las bandas 1 y 2, y 1 y 4, respectivamente. Los valores de los coeficientes de correlación para todos los pares de bandas aparecen en la Tabla 3.

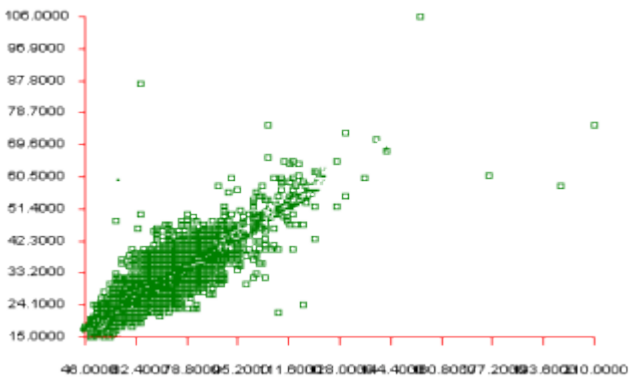

Figura 8: Diagrama de dispersión entre bandas 1 y 2.

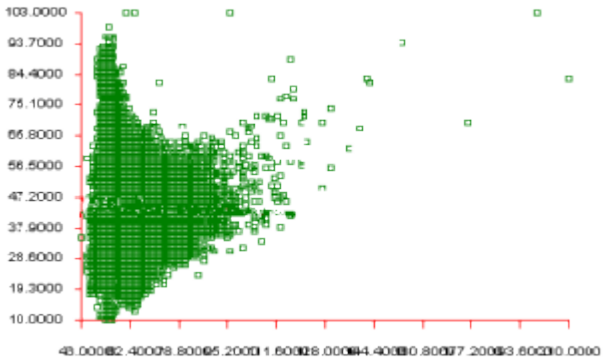

Figura 9: Diagrama de dispersión entre bandas 1 y 4.

COR MATRX	b1	b2	b3	b4	b5	b7
b1	1.000000	0.893030	0.845396	-0.024537	0.312310	0.588491
b2	0.893030	1.000000	0.911848	0.104067	0.358849	0.577399
b3	0.845396	0.911848	1.000000	-0.026410	0.462235	0.671396
b4	-0.024537	0.104067	-0.026410	1.000000	0.277525	-0.090058
b5	0.312310	0.358849	0.462235	0.277525	1.000000	0.721266
b7	0.588491	0.577399	0.671396	-0.090058	0.721266	1.000000

Tabla 3: Matriz de correlación entre bandas.

A partir de la tabla, se puede ver que la mayor correlación se presenta entre las bandas del espectro visible ($\rho > 0.84$), mientras que la banda 4 (infrarrojo cercano) es la que tiene menor correlación con las restantes ($-0.10 < \rho < 0.28$). La alta correlación entre las bandas 1 y 2 ($\rho=0.89$) coincide con lo que se observa en la Figura 8 (puntos cercanos a una recta). Por otra parte la nube (sin ningún patrón) de la Figura 9 es indicio de la falta de correlación entre las bandas 1 y 4 ($\rho = -0.02$).

Si los diagramas de dispersión entre todos los pares de bandas y los correspondientes coeficientes de correlación no dan indicios de correlación, el A.C.P. no cumpliría con el objetivo de disminuir la dimensionalidad de los datos.

Cálculo de los componentes principales a partir de la matriz de varianza-covarianza

En la Tabla 4 se presenta la matriz de varianza-covarianza para las seis bandas. En ella se puede observar que los valores de las varianzas para las bandas 4, 5 y 7 son más altos que el resto, lo que lleva a una mayor ponderación de éstas en el análisis.

VAR/COVAR	b1	b2	b3	b4	b5	b7
b1	39.55	21.58	33.15	-1.38	22.24	27.45
b2	21.58	14.76	21.85	3.57	15.61	16.45
b3	33.15	21.85	38.88	-1.47	32.64	31.05
b4	-1.38	3.57	-1.47	**79.71**	28.06	-5.96
b5	22.24	15.61	32.64	28.06	**128.24**	60.57
b7	27.45	16.45	31.05	-5.96	60.57	**54.99**

Tabla 4: Matriz de varianza-covarianza.

Los autovalores calculados a partir de la matriz de varianza-covarianza se presentan en la Tabla 5 . La última fila de dicha tabla muestra la proporción de varianza total explicada por cada uno de los componentes. Los tres primeros componentes sintetizan el 95.16% de la variabilidad total, mientras que con el cuarto retienen en conjunto el 98.16% de la variabilidad total. Teniendo presente que el objetivo es reducir la dimensionalidad de los datos, se podría pensar que los tres o cuatro primeros componentes conservan casi la totalidad de la información.

Componente	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Autovalor	197.22	89.92	51.75	10.68	5.45	1.13
% var	55.38	25.25	14.53	3.00	1.53	0.32

Tabla 5: Autovalores y porcentaje de la variación total explicada por cada componente.

Para generar los componentes principales se calculan los autovectores que se presentan en la Tabla 6.

COMPONENTE	C 1	C 2	C 3	C 4	C 5	C 6
b1	0.275912	-0.289462	0.520971	-0.104850	-0.716021	-0.212118
b2	0.179204	-0.127954	0.336464	-0.101244	0.112924	0.902941
b3	0.324220	-0.269487	0.397912	-0.278309	0.672617	-0.366135
b4	0.152983	0.839408	0.464789	0.223033	0.044426	-0.065154
b5	0.745432	0.239160	-0.480430	-0.376497	-0.114860	0.037120
b7	0.455595	-0.255760	-0.101024	0.842500	0.083561	-0.005003

Tabla 6: Matriz de autovectores.

Una vez generados los componentes principales, para determinar cuántos incluir se aplicarán los tres criterios presentados:

(a) Si se grafican los porcentajes de la variación total explicada por cada componente, o equivalentemente los autovalores, (Figura 10) y se consideran los componentes anteriores al punto de inflexión de la curva, se retendrían los tres primeros componentes.

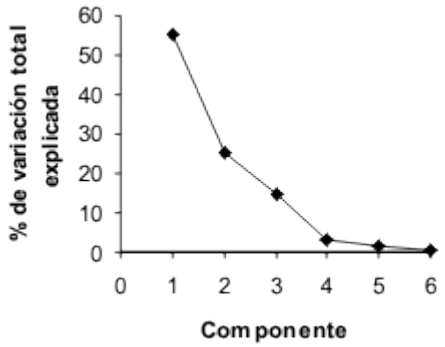

Figura 10: Porcentajes de la variación total explicada por cada componente.

(b) Si se consideran los componentes cuyos autovalores son mayores al promedio (en este caso 59.36) se deberían retener los dos primeros componentes (λ_1=197.22 y λ_2=89.92).

(c) Para examinar las correlaciones entre las bandas y los componentes se calculan los coeficientes de correlación. Por ejemplo, entre el componente 1 y la banda 1 es:

$$\rho_{C1,b1} = \frac{0.276\sqrt{197.22}}{\sqrt{39.55}} = 0.616$$

Éste y los restantes se presentan en la Tabla 7. De esta tabla se puede concluir que las bandas 5 y 7 son las que más contribuyen al componente 1 (ρ = 0.92 y ρ = 0.86), mientras que la banda 4 tiene el menor aporte (ρ = 0.24). Toda la contribución de la banda 4 está en el componente 2 (ρ = 0.89) que, por otro lado, tiene poca correlación con las restantes bandas. Los valores pequeños de los coeficientes de correlación del componente 3 y más aún del 4 en adelante, indicarían que sólo deberían incluirse los componentes 1 y 2.

LOADING	C 1	C 2	C 3	C 4	C 5	C 6
b1	0.616107	-0.436441	0.595933	-0.054480	-0.265901	-0.035781
b2	0.654946	-0.315762	0.629935	-0.086101	0.068636	0.249289
b3	0.730163	-0.409795	0.459055	-0.145843	0.251917	-0.062288
b4	0.240632	0.891523	0.374511	0.081632	0.011621	-0.007742
b5	0.924425	0.200263	-0.305205	-0.108644	-0.023689	0.003477
b7	0.862771	-0.327037	-0.098003	0.371251	0.026317	-0.000716

Tabla 7: Correlación entre los componentes principales y las bandas.

De acuerdo a los criterios (b) y (c) se eligen los componentes 1 y 2, cuyas expresiones son:

$$C1 = 0.276\ b1 + 0.179\ b2 + 0.324\ b3 + 0.153\ b4 + 0.745\ b5 + 0.456\ b7$$
$$C2 = -0.289\ b1 - 0.128\ b2 - 0.269\ b3 + 0.839\ b4 + 0.239\ b5 - 0.256\ b7$$

Cálculo de los componentes principales a partir de la matriz de correlación.

El A.C.P. partiendo de la matriz de correlación se realiza de manera análoga a la efectuada usando la matriz de varianza-covarianza. En este caso, se estaría dando la misma importancia a todas las bandas, lo cual es deseable si no se quiere que las que tienen mayores varianzas aparezcan como las que más aportan. En la Tabla 8 se muestran los autovalores y los autovectores; en la Tabla 9 se presentan las correlaciones entre los componentes y las bandas.

COMPONENT	C 1	C 2	C 3	C 4	C 5	C 6
% var.	59.77	19.56	14.69	2.86	2.18	0.94
eigenval.	3.59	1.17	0.88	0.17	0.13	0.06
b1	0.467720	-0.225803	-0.281254	0.355553	-0.683107	-0.241028
b2	0.481441	-0.116958	-0.358296	-0.121521	0.184466	0.759845
b3	0.495572	-0.145641	-0.130575	-0.490908	0.386979	-0.570442
b4	0.030839	0.792667	-0.529272	0.220257	0.143414	-0.146692
b5	0.336631	0.529742	0.488234	-0.424524	-0.412743	0.130778
b7	0.435419	0.071939	0.506921	0.623771	0.398321	-0.022718

Tabla 8: Autovalores y autovectores de la matriz de correlación.

LOADING	C 1	C 2	C 3	C 4	C 5	C 6
b1	0.885705	-0.244638	-0.264011	0.147412	-0.246895	-0.057330
b2	0.911689	-0.126714	-0.336330	-0.050382	0.066671	0.180733
b3	0.938449	-0.157789	-0.122570	-0.203529	0.139866	-0.135683
b4	0.058399	0.858786	-0.496824	0.091318	0.051834	-0.034892
b5	0.637467	0.573929	0.458302	-0.176007	-0.149177	0.031106
b7	0.824538	0.077940	0.475843	0.258614	0.143965	-0.005404

Tabla 9: Correlaciones entre componentes y bandas.

Se observa que el primer componente sintetiza el 59.77% de la variación total (porcentaje superior al 55.38% del otro análisis) y aparece fuertemente asociado con todas las bandas salvo la 4 (las bandas del visible aportan mucho a pesar de que sus varianzas son menores al resto). El segundo componente sintetiza el 19.56% de la variación total (inferior al 25.25% del otro análisis) y, de la misma manera que en el primer análisis, está altamente relacionado con la banda 4 ($\rho = 0.86$). Los restantes componentes aparecen escasamente correlacionados con todas las bandas, lo cual llevaría a pensar nuevamente en usar los dos primeros componentes que retienen en conjunto el 79.33% de la variabilidad total.

En las Figuras 11 y 13 se muestran los histogramas y la estadística descriptiva de los componentes mientras que las Figuras 12 y 14 presentan las imágenes de estos dos componentes sin y con ensanche de contraste, respectivamente.

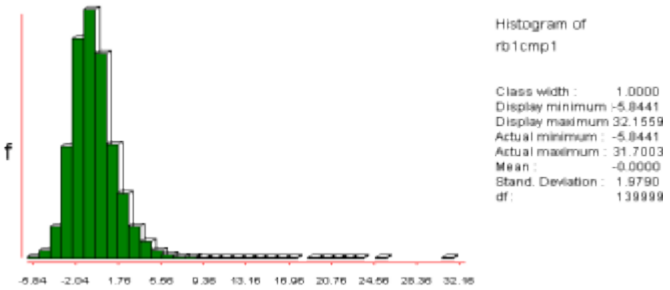

Figura 11: Histograma y estadística descriptiva del componente 1.

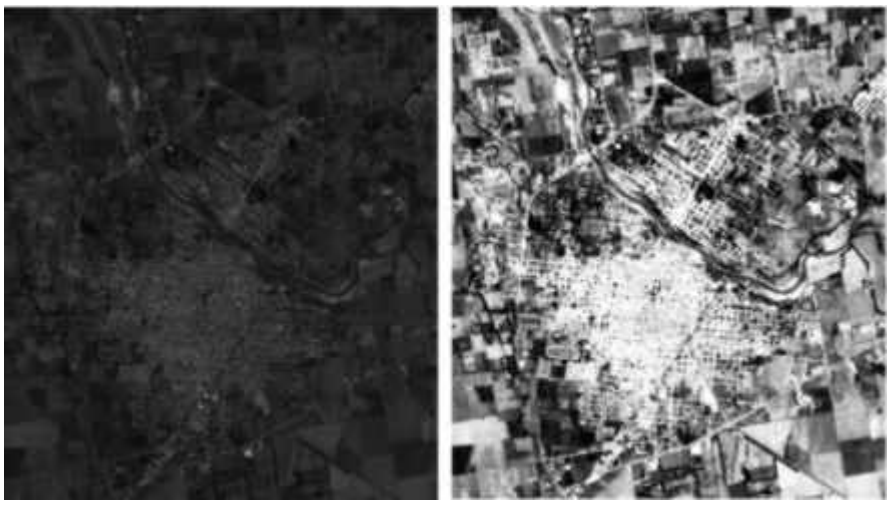

Figura 12A: Componente 1 original con paleta GREY256.

Figura 12B: Componente 1 con ensanche de contraste por ecualización de histograma con paleta GREY256.

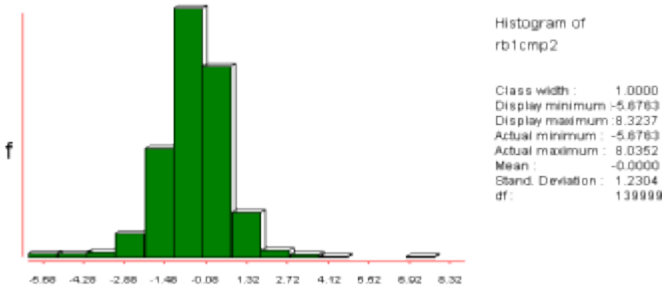

Figura 13: Histograma y estadística descriptiva del componente 2.

Figura 14A: Componente 2 original con paleta GREY256.

Figura 14B: Componente 2 con ensanche de contraste por ecualización de histograma con paleta GREY256.

El problema se ha reducido de seis a dos dimensiones tanto cuando se usa la matriz de varianza-covarianza como cuando se utiliza la matriz de correlación. En este último caso el componente 1 sintetiza la información de todas las bandas salvo la 4 que está asociada al componente 2.

Comparando la banda 5 con los componentes 1 y 2 (Figuras 5, 13 y 14, respectivamente) se pueden observar las siguientes características:

· lo edificado se visualiza en tonos oscuros en la banda 5, en tonos claros en el componente 1 y en tonos oscuros en el componente 2;
· el agua se ve en tonos oscuros en las tres imágenes;
· los lotes cultivados presentan menos variación de grises en la banda 5, la máxima variación de grises en el componente 1 y una variación intermedia en el componente 2;
· los elementos contrastantes lineales como la avenida de circunvalación

o el hipódromo se pueden identificar más claramente en el componente 1.

4. CONSIDERACIONES FINALES.

Debe tenerse en cuenta que el A.C.P. produce tantos componentes principales como bandas tiene la imagen. Cuando las variables originales no están correlacionadas, la información contenida en los componentes principales es esencialmente la misma que la de las variables originales. En teledetección, generalmente, las bandas están altamente correlacionadas, entonces con un número menor de componentes se conserva casi la totalidad de la variabilidad.

El A.C.P. aquí presentado es un análisis descriptivo que genera nuevas variables no correlacionadas que expresan la información contenida en el conjunto de datos. Para lograr reducir la dimensionalidad de los datos se presentaron tres criterios descriptivos que permiten decidir el número de componentes a utilizar. La decisión final de la cantidad de componentes no debería basarse en la aplicación de uno sólo de ellos. Más aún existen criterios basados en pruebas de hipótesis que son válidas bajo la distribución normal multivariada de las variables involucradas. Esto podría profundizarse en futuros trabajos.

Si bien el A.C.P. puede realizarse usando la matriz de varianza-covarianza o la de correlación, en este último caso se otorga a cada banda la misma importancia, independientemente de los valores relativos de sus varianzas.

En la interpretación de los resultados obtenidos, debe tenerse en cuenta que, por realizar transformaciones lineales, los valores no corresponden a N.D. y por tanto no deben asociarse a la respuesta espectral del terreno.

En teledetección el A.C.P en una técnica útil como paso previo para otros análisis. Cuando se realiza una composición color con los primeros componentes principales usualmente se distinguen más coberturas que si se utilizan las bandas. Por otra parte, en aplicaciones multitemporales en las que interesa detectar cambios, se seleccionan los últimos componentes porque ofrecen la información no común.

5. BIBLIOGRAFÍA.

· CHUVIECO SALINERO, E. 1996. *Fundamentos de Teledetección espacial*. Madrid. Ediciones RIALP. 568 pp.
· JOHNSON, R.; WICHERN,D. 1992. *Applied multivariate statistical analysis*. Nueva Jersey. Ed. Prentice Hall. 640 pp.
· MARDIA, K.V.; KENT,J.T.; BIBBY,J.M. 1982. *Multivariate analysis*. Londres. Ed. Academic Press. 521 pp.
· PLA, L.E. 1986. *Análisis multivariado: Método de componentes principales*. Washington. Ed. Organización de Estados Americanos (OEA). 95 pp.
· RICHARDS, J.A., JIA, X. 1999. *Remote sensing digital image analysis*. Alemania. Ed. Springer-Verlag. 363 pp.

Autores del Artículo:

Prof. Susana Beatriz Ferrero[1], Prof. María Gabriela Palacio[1], Lic. Osvaldo R. Campanella[2]

Universidad Nacional de Río Cuarto
Facultad de Ciencias Exactas, Físico-Químicas y Naturales
Ruta 8 km. 603 - (5800) Río Cuarto - Córdoba
[1] Departamento de Matemática.
[2] Departamento de Geología.
Comentarios sobre el artículo: sferrero@exa.unrc.edu.ar

Contrataciones del equipo de Río Cuarto:

Los autores pertenecen al equipo de Río Cuarto, que está formado por **Osvaldo R. Campanella** (geólogo, especialista en SIG aplicado a la geología ambiental), **Marcelo Uva** (licenciado en computación y programador) y **Susana B. Ferrero** (matemática, y Magíster en estadística aplicada). Todos ellos poseen un elevado grado de cualificación profesional y pertenecen a la Universidad Nacional de Río Cuarto, Córdoba, Argentina.

Son contratables por cualquier usuario o institución que necesite la ejecución rápida y eficiente de algún trabajo en el ámbito de los Sistemas de Información Geográfica y Teledetección, poniendo a su disposición una elevada cualificación técnica y el dominio de un amplio abanico de herramientas, entre las que se encuentran ArcView (incluyendo trabajos de programación), ENVI, SAS, programación en IDL, Delphi y otros lenguajes. Me consta que trabajan rápido y bien.

Para pedir presupuestos y solicitar contrataciones contactar con Osvaldo R. Campanella.