ANÁLISIS FACTORIAL CON COMPONENTES PRINCIPALES PARA INTERPRETACION DE IMÁGENES SATELITALES "LANDSAT TM 7" APLICADO EN UNA VENTANA DEL DEPARTAMENTO DE RISARALDA

Analysis factorial with principal components for interpretation of satelital images " landsat tm 7 " applied in a window of risaralda's department

RESUMEN

En este artículo se presenta una de las múltiples ventajas que trae consigo el análisis multivariante de datos, mostrando como a partir de la técnica de análisis factorial con componentes principales como método de extracción se pueden mejorar las imagen satelitales utilizadas en la exploración de áreas geográficas con el objetivo de identificar claramente la ubicación específica de ciertos materiales naturales como carbón y otros minerales.

PALABRAS CLAVES: Análisis Multivariante de Datos, Bandas Térmicas, Componentes Principales, Estadística Descriptiva, Métodos Factoriales, Píxeles.

ABSTRACT

In this article there appears one of the multiple advantages that brings the analysis multivariant of information, showing like from the skill of analysis factorial with principal components as method of extraction, the satelital image can be improved used in the exploration of geographical areas with the aim to identify clearly the specific location of certain natural materials as coal and other minerals

KEYWORDS: *Analysis Multivariant of Information, Thermal Bands, Principal Components, Descriptive Statistics, Methods Factoriales, Pixels.*

JUAN FERNANDO LOPEZ
Magíster en Investigación Operativa y Estadística.
Ingeniero Industrial.
Universidad Tecnológica de Pereira
jflopez@utp.edu.co

SERGIO FERNÁNDEZ HENAO
Ingeniero Industrial.
Estudiante de Maestría en Investigación Operativa y Estadística.
Universidad Tecnológica de Pereira
Docente Transitorio.
sfernandes@utp.edu.co

CARLOS LOZADA RIASCOS
Ingeniero Catastral y Geodesta
Esp.SIG - Director Sistema de información Regional
Estudiante de Maestría en Investigación Operativa y Estadística.
Universidad Tecnológica de Pereira

1. INTRODUCCIÓN.

Una imagen Satelital es una representación visual de la superficie terrestre capturada por un sensor (dispositivo) montado en un satélite, la cual puede ser utilizada para múltiples propósitos, entre ellos el de interpretar las características del territorio tales como la cobertura vegetal.

El satélite LandSat 7 ETM (Enhanced Thematic Mapper Plus) es un instrumento puesto en órbita por la NASA (National Space and Space Administration) y la producción y comercialización de imágenes depende de la USGS (United Sates Geological Survey). Este satélite esta en capacidad de tomar imágenes de un punto de la superficie terrestre cada 16 días (resolución temporal) compuestas por 8 bandas con resolución espacial a 30 metros y una banda Pancromática con resolución de 15 metros, las cuales pueden ser combinadas de distintas formas para obtener variadas composiciones de color u opciones de procesamiento.

El Análisis Factorial con método de extracción de Componentes principales sobre la imagen puede aplicarse como un procedimiento de realce previo a cualquier interpretación, identificando los rasgos comunes que aparecen en la mayoría de las bandas y los cuales se integran en los primeros componentes. Por otra parte los Niveles Digitales **ND** de los píxeles de una banda pueden presentar relación con los de

Fecha de Recepción: 25 de Enero de 2008
Fecha de Aceptación: 20 de Marzo de 2008

componentes principales en teledetección busca resumir un grupo de variables en un nuevo conjunto mas pequeño sin perder parte significativa de la información original.

Para el caso presentado en este artículo se trabajo sobre una ventana de la imagen LANDSAT 7 escena: L71009057_05720050224.742 del año 2005, con cubrimiento del Departamento de Risaralda y Zona Norte del Valle del Cauca.

Figura 1. Ventana de la imagen satelital en composición 345(Falso Color compuesto estándar I).

2. CONCEPTOS GENERALES.

Las bandas espectrales obtenidas por el sensor Landsat 7 ETM se detallan en la tabla 1. Cada una de estas bandas llegan con muy bajo nivel de contraste, esto quiere decir que en cuanto a resolución radiométrica los píxeles toman valores en la gama de gris medio.

En cuanto a la resolución radiométrica que se define como la cantidad de niveles de gris a partir de los cuales se registra la información recibida para luego ser procesada, las imágenes LANDSAT 7 cuentan con 256 niveles digitales **ND** de cada píxel de la imagen que van desde el negro pleno (0) al blanco pleno (255).

Banda	Longitud de onda (µm)	Espectro	Resolución espacial
1	0,450 a 0,515	Visible-Azul	30*30
2	0,525 a 0,605	Visible-Verde	30*30
3	0,630 a 0,690	Visible-Rojo	30*30
4	0,750 a 0,900	Infrarrojo Cercano	30*30
5	1,550 a 1,750	Infrarrojo Medio	30*30
6	10,400 a 12,500	Infrarrojo Térmico	60*60
7	2,090 a 2,350	Infrarrojo Medio	30*30
8	0,520 a 0,900	Pancromático	15*15

Tabla 1. Bandas Sensor Landsat 7 ETM.

Para el usuario final de productos de teledetección, el objetivo relevante para utilizar métodos factoriales 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, permitiendo mejorar significativamente la diferenciación de las zonas geograficas capturadas en la toma, lo cual se expondrá en el numeral 5.

3. ANÁLISIS MULTIVARIADO DE DATOS

3.1. Introducción.

Cada vez que se va a realizar un estudio sobre alguna temática en particular, el analista se puede encontrar con solo variables cuantitativas o solo cualitativas o una combinación de ambas, adicional a esta realidad puede obtener un número considerable de variables volviéndose el estudio aún más complejo respecto a los posibles resultados e interpretaciones correctas debido a la gran cantidad de variables que debe considerar. Es por eso que a partir de los años 70 científicos y estadísticos de la época empezaron a desarrollar técnicas de análisis multivariado como lo fueron los métodos factoriales y de clasificación[1].

Estas técnicas fueron acompañadas por los ordenadores facilitando y agilizando el desarrollo de las mismas, gracias a ello en la actualidad son muchos los estudios que se realizan basados en las técnicas de análisis multivariado en diferentes áreas del conocimiento como lo son: La medicina, las ciencias sociales y la economía entre otros.

3.2. Desarrollo.

El análisis multivariante de datos comprende el estudio estadístico de varias variables medidas en elementos de una población con los siguientes objetivos[2]:

- ✓ Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
- ✓ Encontrar grupos en los datos, si existen.
- ✓ Clasificar nuevas observaciones en grupos definidos.
- ✓ Relacionar dos conjuntos de variables.

Con base a los anteriores objetivos, al igual que la naturaleza de cada variable incluida en el estudio (cualitativa o cuantitativa) y la relación existente entre ellas (Explicativas y explicadas) se han desarrollado diferentes técnicas de análisis multivariante de datos como lo son:

- ✓ Métodos Factoriales: Análisis Factorial, Componentes Principales y Correspondencias.
- ✓ Escalamiento Multidimensional.
- ✓ Regresión: Simple, Múltiple y Canónica.
- ✓ Análisis Discriminante.
- ✓ Análisis Conjunto
- ✓ Análisis de Varianza: Anova, Ancova, Manova y Mancova.
- ✓ Segmentación: Clusters No Jerárquicos y Clusters Jerárquicos.

Para el estudio presentado en este artículo se utiliza la técnica de análisis factorial con el método de extracción de componentes principales debido a la naturaleza de las variables (mencionadas en el primer y segundo numeral) y a la relación de las mismas tal como se explica en los siguientes numerales.

4. MÉTODOS FACTORIALES[3].

4.1. Introducción.

El análisis factorial, al igual que el análisis en componentes principales, es una técnica multivariante que tiene como objetivo principal *reducir* la dimensión de una tabla de datos para pasar de "p" variables reales a "k" variables ficticias que aunque no observables, sean combinación de las reales y sinteticen la mayor parte de la información contenida en los datos.

Cuando se estudia el comportamiento de dos o más variables de naturaleza cualitativa se utiliza el análisis de correspondencia el cual se denomina simple para solo dos variables y múltiple para tres o más.

Como se mencionó anteriormente, el caso que se presenta en este artículo contiene seis variables cuantitativas (Banda 1, Banda 2,...., Banda 6), por lo tanto se aplica la técnica de análisis factorial con método de extracción de componentes principales para obtener dos o tres nuevas variables ficticias que permitan relacionar y resumir el nivel digital ND de cada banda en un píxel dado relacionado con una posición geográfica especifica, para obtener un realce de la imagen que permita identificar ciertos materiales naturales en determinada área geográfica.

4.2. El Modelo Factorial[4].

El análisis factorial como se ha mencionado, se dirige a establecer si las covarianzas o correlaciones observadas sobre un conjunto de variables pueden ser explicados en términos de un número pequeño no observable de variables no latentes (ficticias). De esta manera, considérese a X como un vector aleatorio de tamaño (p x 1) con media μ y matriz de covarianzas Σ; se trata entonces de indagar acerca del siguiente modelo:

$$X = \mu + \Lambda f + U \quad (1)$$

Donde:

- f es un vector (m x 1) de variables latentes o factores no observables.
- Λ es una matriz (p x m) de constantes desconocidas (m<p). Contiene los coeficientes que describen como factores "f", afectan a las variables observadas "**x**", y se denomina matriz de carga.
- μ es un vector (p x 1) de perturbaciones no observadas. Recoge el efecto de todas las variables distintas de los factores que influyen sobre **x.**

La ecuación (1) implica que dada una muestra aleatoria simple de n elementos generada por el modelo factorial, cada $\mathbf{x_{ij}}$ puede escribirse como:

$$x_{ij} = \mu_j + \lambda_{j1} f_{1i} + ... + \lambda_{jm} f_{mi} + \mu_{ij} \quad (2)$$

La ecuación (2) señala que la información contenida por cada variable "engloba" varios aspectos (los f's), compartidos en grado o intensidad distinta por las demás variables, y alguna información exclusiva de la variable. Los elementos de f son llamados los factores comunes y los elementos de U factores únicos o específicos.

Al aplicar la técnica del análisis factorial usando el método de componentes principales se busca estudiar y explicar las covarianzas y correlaciones existentes entre las diferentes variables.

5. APLICACIÓN DEL ANÁLISIS FACTORIAL EN IMÁGENES SATELITALES[6].

La base de datos empleada en el desarrollo de este estudio está conformada por seis variables cuantitativas y 3616256 registros como se observa en la siguiente figura.

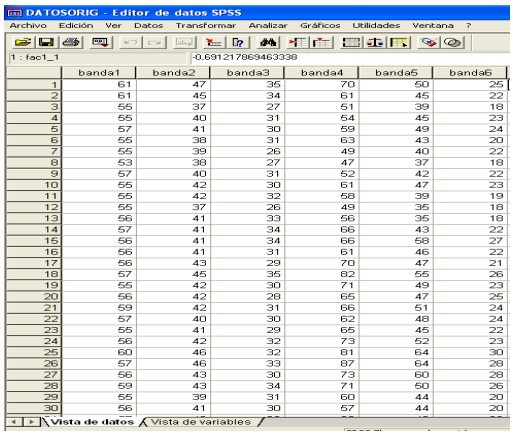

Figura 2. Tabla de datos en SPSS.

Se utilizó el paquete estadístico SPSS para realizar el análisis factorial con el método de extracción de componentes principales, de lo cual se habló en el numeral 4.

Antes de emplear la técnica multivariada se realizó un contrate de esfericidad de Barlett y de Medida de KMO para determinar si hay correlación entre las variables objeto de estudio y para determinar si la técnica de análisis factorial es aplicable en este caso.

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,719
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	32632254,344
	gl	15
	Sig.	,000

Tabla 2. Contraste de Barllet y KMO.

Al observar los resultados en la tabla anterior, el estadístico KMO tiene un valor de 0,719 que lo acerca a la unidad, lo que indica que los datos se adecuan para efectuar un análisis factorial y el contraste de Bartlett con p-valor 0.000 indica que se rechaza la hipótesis nula de que las variables iníciales **no** están correlacionadas, por lo tanto se puede efectuar un análisis factorial.

En este orden de ideas se continúa con un análisis descriptivo univariante para determinar el coeficiente de variación y así tener un primer plano de que variables presentan mayor variación que otras.

Con base a los resultados de la tabla 3 se puede concluir que entre las bandas del espectro visible (Bandas 1, 2 y3) la más constante es la 1 y la de mayor dispersión es la 3, y en el otro tipo (Bandas infrarrojo 4, 5 y 6) la banda 6 presenta la mayor variación.

VARIABLES	Media	Desv. típ.	Coeficiente de variación %
BANDA_1	63,55	9,637	15.16
BANDA_2	52,97	10,531	19.88
BANDA_3	43,44	13,901	**32.00**
BANDA_4	87,67	16,729	19.08
BANDA_5	83,37	21,345	25.60
BANDA_6	42,56	15,559	**36.56**

Tabla 3. Análisis Descriptivo.

Ahora se analiza la matriz de correlaciones para observar como se comporta cada variable frente a las otras y para observar su determinante el cual debe ser muy pequeño para poder decir que el grado de intercorrelación entre las variables es muy alto.

		B_1	B_2	B_3	B_4	B_5	B_6
Correlación	B_1	1,000	,948	,920	,092	,615	,754
	B_2	,948	1,000	,934	,227	,720	,793
	B_3	,920	,934	1,000	-,007	,705	,858
	B_4	,092	,227	-,007	1,000	,410	,093
	B_5	,615	,720	,705	,410	1,000	,893
	B_6	,754	,793	,858	,093	,893	1,000
Sig. (Unilateral)	B_1		,000	,000	,000	,000	,000
	B_2	,000		,000	,000	,000	,000
	B_3	,000	,000		,000	,000	,000
	B_4	,000	,000	,000		,000	,000
	B_5	,000	,000	,000	,000		,000
	B_6	,000	,000	,000	,000	,000	

Determinante = ,000

Tabla 4. Matriz de Correlaciones.

En la tabla anterior se observa que el determinante indica un alto grado de inercorrelación entre las variables, además las bandas pertenecientes al tipo de espectro visible están fuertemente correlacionadas entre sí, y en las tipo infrarrojo las bandas 5 y 6 son las más relacionadas quedando la banda 4 como la variable menos relacionadas con las otra cinco. También se puede apreciar que entre los dos tipos de bandas (visibles e infrarrojas) la banda 3 y la banda 6 presentan la mayor relación “0.858” indicando un factor común en relación al color rojo y al infrarrojo.

Al aplicar el método de extracción de componentes principales en el análisis factorial con rotación varimax se obtuvo los siguientes resultados:

C	Autovalores iniciales			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,301	71,688	71,688	4,178	69,631	69,631
2	1,107	18,447	90,135	1,230	20,504	90,135
3	,473	7,877	98,012			
4	,054	,907	98,920			
5	,041	,676	99,596			
6	,024	,404	100,000			

Tabla 5. Varianza total explicada.

Es así como los valores propios conocidos también como eigenvalores para cada componente se encuentran en la columna "Total" y en la siguiente columna se observa el porcentaje de varianza explicada con el método de extracción, sin embargo al aplicar la rotación de los ejes se ve como el porcentaje de explicación particular varía, pero el acumulado sigue siendo el mismo, esto se debe a que en el momento de realizar la rotación algunas variables cambian de componente, pero el objetivo sigue siendo el mismo, el cual es minimizar las distancias entre cada grupo perdiendo la mínima información posible a la vez que se aumenta la relación de las variables que quedan en cada factor. Con base a la teoría citada en los numerales 3 y 4 y al soporte bibliográfico presentado en este artículo, se puede concluir como con la técnica de análisis factorial se pasa de seis variables observables a dos "ficticias" con las cuales se explica el **90.135 %** de la variación total.

	Componente(a)	
	1	2
BANDA3	,980	-,058
BANDA2	,944	,157
BANDA1	,934	,003
BANDA6	,917	,129
BANDA5	,781	,477
BANDA4	,030	,979

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
(a) La rotación ha convergido en 3 iteraciones.
Tabla 6. Matriz Factorial.

La tabla 6 contiene las proyecciones de cada una de las variables sobre cada uno de los factores encontrados mediante el método de componentes principales, estas proyecciones reciben el nombre de saturaciones. Al sumar el cuadrado de cada saturación para cada componente "Factor" se obtiene su eigenvalor citado en la tabla 4, por lo tanto para el primer factor será: $0.980^2 + 0.944^2 + ... + 0.30^2 = 4.17$. De igual manera para el factor 2 será: $-0.058^2 + ... + 0.979^2 = 1.23$.

Esto significa que las seis variables son explicadas suficientemente con dos factores, en donde el primer factor agrupa las variables: Banda_3, Banda_2, Banda_1, Banda_6 y Banda _5, y el segundo factor contiene la variable Banda_4

Adicionalmente, con la siguiente tabla se puede obtener las transformaciones lineales que relacionan los componentes con las variables, y por lo tanto encontrar el resultado de las dos nuevas variables "ficticias" para cada registro, con lo cual se podrá utilizar estos valores en análisis posteriores (regresión, cluster, etc.) ya que estas variables sustituyen las variables iniciales que las resumen en virtud del análisis factorial que se acaba de realizar.

	Componente	
	1	2
BANDA1	,241	-,118
BANDA2	,224	,016
BANDA3	,261	-,177
BANDA4	-,119	,855
BANDA5	,140	,318
BANDA6	,220	-,005

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.
Tabla 7. Puntuaciones de componentes.

De tal manera que las formulas para las nuevas transformaciones lineales son:

$$C1 = 0.241*Banda1 + 0.224*Banda2 + 0.261*Banda3 – 0.119Banda4 + 0.140*Banda5 + 0.220*Banda6 \quad (3)$$

$$C2 = -0.118*Banda1 + 0.016*Banda2 – 0.177*Banda3 + 0.855*Banda4 + 0.318*Banda5 – 0.005*Banda6 \quad (4)$$

Con estas dos ecuaciones se puede encontrar las dos nuevas variables sustitutas que se encuentran relacionadas con las variables observables, por ejemplo, al reemplazar en cada formula el valor de las variables originales del primer registro presentado en la figura 1 se encuentra el valor de dichas variables sustitutas las cuales son:

$C1 = 38.534$ y $C2 = 62.984$

Esto indica que la Banda_4 incide fuertemente en el valor de la segunda transformación, tal como se observa en la tabla 6.

Por lo tanto se puede concluir que en el momento que se tenga un nivel digital ND en las tres primeras bandas en relación a las bandas

del espectro infrarrojo, se obtendrá un valor mayor en la primera transformación lineal es decir: **C1>C2**, y cada vez que se obtenga en la banda 4 un nivel digital mayor se presentará el efecto contrario es decir: **C2>C1**

Por otra parte el analista puede concluir sobre el comportamiento de las bandas térmicas en un momento dado a partir de los valores que se obtengan en las transformaciones lineales "C1 y C2", es decir, que si se encuentra frente al primer caso **C1>C2**, podrá indicar que en ese punto hubo mayor nivel digital ND en las bandas del espectro visible y en la banda 6, por el contrario si se encuentra con el caso **C2>C1,** se estará indicando que la banda infrarrojo 4 tuvo una mayor intensidad que las demás, por último se pude presentar el caso en que las dos transformaciones sean muy parecidas en sus valores, lo cual estará indicando que el nivel digital ND de las bandas 1, 3 y 4 son muy similares en dicho punto.

6. CONCLUSIONES.

Cuando el analista se enfrenta a una cantidad de variables las cuales no presentan una relación aparente es importante complementar la estadística univariante con la multivariante, ya que ésta le proporciona herramientas de contraste claras para observar una realidad que no se nota a simple vista.

En muchas ocasiones el análisis multivariente permitirá reducir considerablemente el tiempo de gestión o desarrollo del estudio final gracias a las alternativas de reducción respecto a las variables o a los registros.

Los métodos factoriales son una técnica multivariante muy fuerte para aplicarla en la reducción de la dimensión del estudio dado, sin embargo debe tenerse cuidado al tratar con análisis factorial y componentes principales, ya que se pueden aplicar como herramientas estadísticas por separado lo cual conllevaría a resultados diferentes o vincularlas utilizando el análisis factorial con método de extracción de los factores por medio de componentes principales, es por eso que se recomienda al lector profundizar en las diferencias y similitudes de estas dos técnicas multivariantes.

El análisis de componentes principales efectuado se convierte en un procedimiento útil antes de interpretar las imágenes satelitales, cuando se realiza una composición a color utilizando el resultado de los componentes principales (imagen 3) se puede diferenciar mas elementos que cuando se utilizan las bandas independientes en las mismas (imagen 2), además se puede concluir también que las bandas están altamente correlacionadas por lo tanto con los componentes seleccionados se conserva la variabilidad.

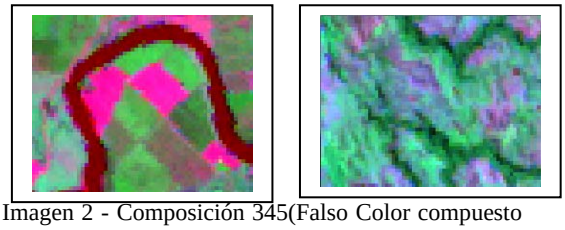

Imagen 2 - Composición 345(Falso Color compuesto estándar I).

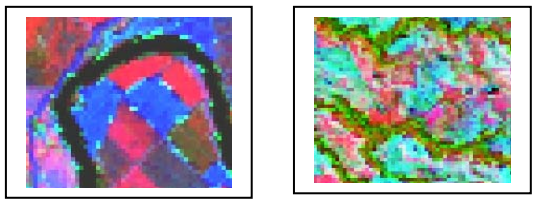

Imagen 3 - Composición 345(Falso Color compuesto estándar I)-Método Factorial.

Con la técnica de análisis factorial se pudo reducir la dimensión de seis variables originales a dos variables sustitutas con las que se podrán realizar estudios en dos dimensiones siendo más práctico para el analista.

7. BIBLIOGRAFÍA.

[1] PÉREZ. L. César. Técnicas de Análisis Multivariante de Datos. Edit. Pearson/Prentice Hall. 2006.

[2] PEÑA. Daniel. Análisis de Datos Multivariantes. Edit. Mc Graw Hill. 2002

[3] FERRÁN. A. Magdalena. SPSS para Windows – Análisis Estadístico. Edit. Mc Graw Hill. 2003.

[4] DÍAZ. Luís. Estadística Multivariada: Inferencia y Métodos. Universidad Nacional de Colombia. Departamento de Estadística. 2005.

[5] CHUVIECO SALINER0, E. Fundamentos de Teledetección Espacial. Madrid. Ediciones RIALP 1996.

[6] BUZAI, G.D. Baxendale, La construcción de regiones mediante técnicas geográficas cuantitativas. GERENCIA AMBIENTAL 2002.