



III CADI
IX CAEDI
2016



UNIVERSIDAD
TECNOLÓGICA NACIONAL
FACULTAD REGIONAL
RESISTENCIA



UNIVERSIDAD NACIONAL
DEL NORDESTE
FACULTAD
DE INGENIERÍA

ANÁLISIS DE COMPONENTES PRINCIPALES DESARROLLADO EN ENERGÍAS RENOVABLES. APLICACIÓN A SISTEMAS DINÁMICOS Y BIOMÉDICOS

Carlos Alvarez Picaza, Universidad Nacional del Nordeste, cpicaza@gmail.com

María Inés Pisarello, Universidad Nacional del Nordeste, mainespisarello@gmail.com

Jorge Emilio Monzón, Universidad Nacional del Nordeste, jemonzon2@gmail.com

Resumen— En la constante búsqueda del método automático que mejor clasifique las señales que arrojan las distintas clases de energías renovables, ya sea un algoritmo que contemple los aspectos temporales o morfológicos de la señal, es inevitable la acumulación de gran cantidad de datos. Como solución a este problema surgen técnicas estadísticas capaces de reducir el número de variables necesarias. Una consecuencia directa es la obtención de datos normalizados y procesos más eficientes que agilizan los tiempos computacionales y además economizan en espacio de almacenamiento. El análisis y la clasificación de dichas señales pueden realizarse a partir de estudios temporales y morfológicos. El Análisis de Componentes Principales (PCA) es un método que reduce la dimensión de los datos realizando un análisis de covarianza entre factores. A partir del estudio de una central eléctrica compuesta por paneles solares y aerogeneradores, conformamos una matriz dinámica; a la cual aplicamos las técnicas de PCA. Como resultado obtuvimos, en la gran mayoría, reducciones de hasta el 50 % en el número de variables. De un total de diez (10) variables originales, se logró concentrar más del 96 % de la información en sólo cinco (5) componentes principales. Además su aplicación es compatible a cualquier clase de sistemas, inclusive lo biomédicos.

Palabras clave— PCA, Compresión de Datos, Factores.

1. Introducción

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, si tomamos demasiadas variables sobre un conjunto de objetos, tendremos que considerar muchos posibles coeficientes de correlación, y va aumentando, si consideramos un número aún mayor de variables.

Otro problema que se presenta es la fuerte correlación que muchas veces se presenta entre las variables, si tomamos demasiadas variables (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio), lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, la presión sanguínea a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas.

Se hace necesario, pues, reducir el número de variables. Es importante resaltar el hecho de que el concepto de mayor información se relaciona con el de mayor variabilidad o varianza. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor información.

El Análisis de Componentes Principales o PCA (Principal Component Analysis), es una técnica estadística de síntesis de la información o reducción de la dimensión (número de variables). En bancos de datos de muchas variables, la técnica de PCA permite reducir el número de tales variables, sin perder información substancial. Los nuevos *factores* o *componentes* serán una combinación lineal de las variables originales, e independientes entre sí. Un aspecto clave en el Análisis de Componentes Principales es la interpretación de los factores, que no viene dada a priori, sino que se deduce tras observar la relación de los resultados con las variables iniciales. El propósito fundamental de la técnica consiste en la reducción de la dimensión de los datos con el fin de simplificar el problema en estudio.

El impacto de la contaminación del ambiente en el calentamiento global y los cambios climáticos resultantes pueden tener consecuencias desastrosas a largo plazo. El desarrollo tecnológico para la generación de energía eléctrica se está orientando hacia las fuentes seguras y renovables como la eólica y la fotovoltaica.

Un tercio de la población en el mundo vive lejos de la red de suministro de potencia eléctrica. Para ellos la energía eólica y la energía fotovoltaica son una alternativa muy importante.

Uno de los problemas de la energía eólica es que su disponibilidad es variable, y, por lo tanto necesita ser respaldada por otras fuentes de potencia. Los sistemas fotovoltaicos tienen la ventaja adicional de ser estáticos y de casi no requerir mantenimiento ni reparaciones. Sin embargo, la potencia fotovoltaica es típicamente cinco veces más cara que la potencia eólica. Actualmente, existen investigaciones y esfuerzos por desarrollar paneles fotovoltaicos de bajo costo para aplicaciones generales. La eficiencia de la conversión de la potencia solar es típicamente de 37.8% para paneles solares sin concentrador [Spectrolab].

Existen programas computacionales específicos que ayudan a simplificar el desarrollo del trabajo (XLSTAT, MATLAB).

2. Materiales y Métodos

El PCA se empleó inicialmente en la psicología, las ciencias sociales y naturales. Sin embargo, desde hace ya algunos años se ha extendido su aplicación a las ciencias físicas, la ingeniería, la economía, el reconocimiento de patrones, la compresión de datos, etc. [1].

El Análisis de Componentes Principales es un método que reduce la dimensión de los datos realizando un análisis de covarianza entre factores.

En muchas aplicaciones científicas, un conjunto de n objetos se representan a través de una colección de m descriptores, índices o parámetros. En algunos casos m es un número lo suficientemente grande como para que sea difícil analizar el conjunto de datos en toda su dimensionalidad. En términos más abstractos, se puede decir que se consideran los n objetos como n puntos ubicados en el espacio de m dimensiones [2]. Entonces, el objetivo perseguido es el de clasificar esos objetos, representados en un espacio de dimensión menor $p < m$, de tal manera que esa proyección en ese espacio sea óptima desde algún punto de vista, es decir, que la reducción de dimensionalidad no sea gratuita sino justificada. Podríamos decir que los datos sufren una reducción en su dimensión (Haykin, 1.996).

Esta técnica tiene tres efectos a destacar:

- * Ortogonaliza los componentes de los vectores de entrada de tal modo que estén no-correlacionados entre ellos.
- * Reordena los componentes ortogonales (*componentes principales*) de modo que aquellos con mayor variación estén en los primeros lugares.

* Elimina aquellos componentes que contribuyen menos a la variación del conjunto de datos.

Antes de iniciar una descripción más detallada, vamos a introducir algunos conceptos estadísticos que se usan en PCA. Ellos son, la desviación estándar, la covarianza, los autovectores y los autovalores. Para la definición de estos conceptos, vamos a asumir el hecho de que trabajamos con muestras de una población.

Desviación Estándar

Sobre un conjunto de n datos es posible realizar infinitud de mediciones [3], entre ellas la media aritmética cuya fórmula es:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

La media nos da una idea de cuál es el punto medio, pero no nos dice nada acerca de la distribución de los puntos.

La Desviación Estándar (SD – Standard Deviation) de un conjunto de datos es la medida de cómo están dispersos los puntos. La definición formal es: *La distancia promedio desde la media del conjunto de datos a un punto*. Su fórmula es:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}} \quad (2)$$

Varianza

Es otra medida de la dispersión de datos en el conjunto de datos. Es el cuadrado de la desviación estándar. Su fórmula es:

$$\text{var}(X) = \sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \quad (3)$$

Covarianza

Las últimas dos mediciones son puramente unidimensionales. Sin embargo, muchos conjuntos de datos poseen más de una dimensión y el objetivo de un análisis estadístico de estos datos es ver qué relaciones existen entre las dimensiones, si las hubiera. La covarianza [4] siempre se mide entre dos dimensiones. Si se calcula la covarianza entre una dimensión y ella misma, se obtiene la varianza. La fórmula de covarianza es muy similar a la fórmula de varianza de la ecuación (3.3) si expandimos el término cuadrático y está dada por:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad (4)$$

Autovectores y Autovalores de una matriz

Sea $\mathbf{A}$ una matriz $n \times n$ con coeficientes en el campo de los números complejos $\mathbb{C}$. Un escalar λ en $\mathbb{C}$ es llamado “autovalor” de $\mathbf{A}$ si existe un vector $\mathbf{x} \neq 0$ en $\mathbb{C}$ tal que

$$\mathbf{Ax} = \lambda \mathbf{x} \quad (5)$$

Todo vector $\mathbf{x} \neq 0$ que satisface la ecuación arriba citada se denomina “autovector” de $\mathbf{A}$ asociado al autovalor λ . Para hallar los autovalores de $\mathbf{A}$, escribimos $\mathbf{Ax} = \lambda\mathbf{x}$ como sigue

$$(\lambda\mathbf{I}-\mathbf{A})\mathbf{x} = 0 \quad (6)$$

Esta última ecuación tendrá una solución NO TRIVIAL, o sea $\mathbf{x} \neq 0$, si y solamente si el

$$\det(\lambda\mathbf{I}-\mathbf{A}) = 0 \quad (7)$$

Esto significa que un escalar λ es un autovalor de $\mathbf{A}$ si y solamente si es una solución de

$$\Delta(\lambda) \triangleq \lambda\mathbf{I}-\mathbf{A} = 0 \quad (8)$$

donde, $\Delta(\lambda)$ es un polinomio de grado “ n ” en λ llamado “polinomio característico” de $\mathbf{A}$.

Dado que $\Delta(\lambda)$ posee grado “ n ”, la matriz $\mathbf{A}$ $n \times n$ tendrá “ n ” autovalores (no necesariamente todos distintos).

Metodología

Este trabajo esta basado en la configuración de una central eléctrica compuesta por un panel fotovoltaico y un aerogenerador.

En la metodología de PCA se ordenan los descriptores en una matriz $\mathbf{A}$ de dimensión $n \times m$. El criterio matemático utilizado para conseguir la reducción de la dimensionalidad es tal que, para un valor prefijado de p , se retenga en ese subespacio la máxima varianza estadística total de los datos originales. Esto conduce a especificar una nueva serie de ejes ortogonales entre si, los *componentes principales* (CP). Cada CP es una combinación lineal de las variables o descriptores originales.

El primer tratamiento numérico que debe hacerse es el de escalar las columnas de descriptores de la matriz $\mathbf{A}$. Esto es así porque cada columna (cada variable) puede estar especificada en un sistema de unidades distinto. De hecho, cada variable no tiene porqué ser de la misma naturaleza que las otras. Hay varias posibilidades de escalado. La más común consiste en obtener vectores columnas centrados y normalizados adimensionales, así pues cada columna $\mathbf{a}_j$ de la matriz $\mathbf{A}$,

$$\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_m) \quad (9)$$

Se le calcula su media

$$\bar{a}_j = \frac{1}{n} \sum_i^n a_{ij} \quad (10)$$

y las desviaciones estándar multiplicadas por n

$$s_j = \sqrt{\sum_{i=1}^n (a_{ij} - \bar{a}_j)^2}, \quad (11)$$

obteniendo la matriz de variables adimensionales siguiente:

$$\mathbf{A} \rightarrow \mathbf{Z} = (\mathbf{z}_1 \ \mathbf{z}_2 \ \dots \ \mathbf{z}_m), \quad (12)$$

donde cada vector columna $\mathbf{z}_j$ se define a partir de la transformación

$$a_j \rightarrow z_j = \frac{a_j - \bar{a}_j}{s_j}, \quad (13)$$

La matriz de variables homogeneizadas adimensionales permite calcular la matriz de los coeficientes de correlación entre cada par de columnas de datos:

$$\mathbf{R} = \mathbf{Z}^T \mathbf{Z}, \quad (14)$$

esta matriz es de dimensión $m \times m$.

Los CP vienen dados por los vectores propios de la matriz $\mathbf{R}$:

$$\mathbf{R}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}, \quad (15)$$

donde

$$\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_m) ; \ \mathbf{\Lambda} = \text{Diag}(\lambda_1 \ \lambda_2 \ \dots \ \lambda_m) \quad (16)$$

Todos los valores propios son no negativos (recordemos que la matriz $\mathbf{Z}$ se obtiene de tal manera que es definida no negativa). Precisamente los valores propios de esta matriz son los parámetros que indican qué fracción de la varianza total original retiene cada nuevo CP.

$$f_i = 100 \frac{\lambda_i}{\sum_{j=1}^m \lambda_j} \% \quad (17)$$

Por ello, el ordenamiento, de mayor a menor, de los valores propios induce un orden de preferencia de los CP. A partir de ahora supondremos que

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \quad (18)$$

Ahora

$$\mathbf{R} \rightarrow \mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_m) \quad (19)$$

donde $\mathbf{x}_1$ es el vector propio asociado a λ_1 , $\mathbf{x}_2$ a λ_2 y así sucesivamente hasta m . El primer Componente Principal, $\mathbf{x}_1$ representa la mayor cantidad de varianza de los datos originales, $\mathbf{x}_2$ retiene la segunda mayor varianza, y así hasta m . El conjunto de los m Componentes Principales genera una nueva matriz de coordenadas. A los coeficientes de cada vector propio $\mathbf{x}_j$ se les llama pesos (*loadings*) e indican qué combinaciones lineales de las variables originales se deben construir para definir las nuevas coordenadas adimensionales [5].

Lo más usual, a efecto de reducir la dimensionalidad del problema, consiste en elegir $p = 2$ ó $p = 3$ primeros Componentes Principales. En nuestro caso $p = 5$.

XLSTAT

Cabe destacar que para nuestro análisis se tomaron las 24 horas de actividad de un día determinado. El objetivo es analizar las correlaciones entre las variables e identificar características que se distinguen fuertemente de los demás [6]. El análisis de Componentes Principales (PCA) es un método muy eficaz para el análisis de datos cuantitativos (continuos o discretos) que se presentan bajo la forma de cuadros de M observaciones / N variables.

Permite:

- Visualizar y analizar rápidamente las correlaciones entre las N variables
- Visualizar y analizar las M observaciones inicialmente descritas por N variables en un gráfico de dos o tres dimensiones, construido de tal forma que la dispersión entre los datos sea tanto preservada como posible.

- Construir un conjunto de P factores no correlacionados.

Señal electrocardiográfica

Se ha utilizado un extracto de 5 minutos de la señal N° 100 de la base de datos de arritmias del MIT-BIH. Esta señal pertenece al grupo de señales electrocardiográficas consideradas normales, es decir sin arritmias.

Algoritmo para la determinación de los puntos del QRS

La base de datos de Arritmias del MIT-BIH contiene 48 registros electrocardiográficos que corresponden a sujetos sanos y también sujetos con afecciones cardiacas. Cada señal en la base de datos es un complejo de tres señales, un archivo con extensión *.dat* que contiene los datos propiamente dichos, otro con extensión *.atr* que es un archivo de atributos, donde sólo aparecen como datos los complejos QRS detectados con el algoritmo de Jiapu Pan (1985) y un tercer archivo que permite visualizar la onda con la extensión *.hea* (header).

El algoritmo utiliza tanto la señal del electrocardiograma como el archivo de anotaciones extraído también de la base de datos de Arritmias del MIT-BIH correspondiente a esa señal. Este archivo de atributos contiene, en la primera columna, el tiempo en que sucede la onda Q en cada latido ventricular (lo que marca el comienzo del complejo QRS); en otra columna se indica el orden de la muestra que corresponde a ese dato; también se asienta el código de clasificación del latido ("N" si se trata de un latido normal). A continuación se ilustran las primeras dos filas del archivo.

0:00.194 70 N

0:01.005 362 N

.....

El algoritmo, diseñado con MATLAB® (The MathWorks, Natick, Massachussets, USA), combina los dos archivos, a partir de la concordancia entre el archivo de datos y el de anotaciones de la señal. Con los datos extraídos del archivo de datos propiamente dichos de la señal (*.dat*) generamos una matriz con los valores de muestras que corresponden al complejo QRS, una fila por cada latido. A continuación se muestra en la figura 1 el código mediante el cual carga la matriz mencionada. A continuación, el Código Fuente que carga la matriz de datos de los complejos QRS

```
*****
%% sqrs debe ser la señal de las anotaciones de la deteccion %%
%% de los qrs de la señal con el "sqrs.exe" del MIT-BIH. fs: %%
%% frecuencia de muestreo; 360 para las señales del MIT y 300 %%
%% para las IN-VIVO. %%
*****

[tiempo,x,tipo,y,z,w]=textread('100qrs.txt','%8c %u %c %u %u %u');

f=length(x); %longitud del vector que tiene los comienzos de los QRS
comfin(1:f,1:30)=0;
for i=1:f
    j=x(i);
    for k=1:30 %es como maximo la cantidad de muestras del QRS. Promedio 15
        h=j+k-1;
        comfin(i,k)=s(h); %comienzo del QRS
        if k>5
            if ((dl(h)<dl(h+1)) & (dl(h)<dl(h-1)))
                break
            end;
        end;
    end;
end;
end;
```

La matriz que contiene los datos que describen cada complejo QRS tiene un total de 30 columnas. En la mayoría de los casos los latidos no superan un promedio de 20 muestras, en el caso de latidos normales. Es decir que existe redundancia de datos. El método PCA pretende eliminar esa redundancia.

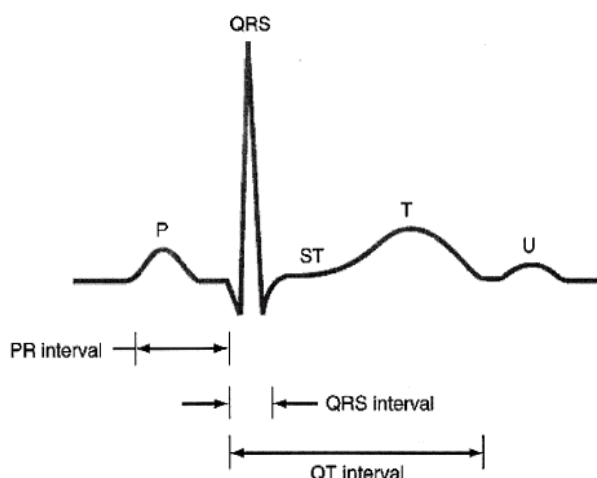


Figura a. Determinación del complejo QRS.

$$y(nT) = \frac{1}{2} [x(nT) + x(nT - T)] \quad (20)$$

Los límites del PCA vienen del hecho que es un método de proyección, y que la pérdida de información inducida por la proyección puede provocar interpretaciones erróneas. El buen discernimiento permite, sin embargo, evitar estos inconvenientes [7].

3. Resultados y Discusión

Incorporamos el Análisis de Componentes Principales realizados con el XLSTAT.

Tabla 1. Variables a lo largo de un día (elaboración propia)

Hour	Global Solar kW/m2	Incident Solar kW/m2	Wind Speed m/s	DC Primary Load kW	PV Power kW	SW AIR X kW	Excess Electricity kW	Bat Power kW	Bat State of Charge %	Bat Energy Cost \$/kWh
1	0	0	3,776	0,002	0	0,017	0,015	0	100	1
2	0	0	3,476	0,002	0	0,013	0,012	0	100	0
3	0	0	2,349	0,002	0	0,003	0,001	0	100	0
4	0	0	2,571	0,002	0	0,005	0,002	0	100	0
5	0	0	3,26	0,002	0	0,011	0,009	0	100	0
6	0	0	1,879	0,002	0	0	0	-0,002	99,92	0
7	0,014	0,005	3,012	0,011	0	0,008	0	-0,002	99,831	0
8	0,061	0,061	3,552	0,002	0,002	0,014	0,012	0,003	99,917	0
9	0,111	0,11	5,424	0,002	0,004	0,045	0,047	0,001	99,947	0
10	0,199	0,197	5,761	0,002	0,008	0,053	0,058	0,001	99,966	0
11	0,037	0,036	6,557	0,002	0,001	0,072	0,071	0	99,978	0
12	0,018	0,018	7,889	0,019	0,001	0,117	0,098	0	99,986	0
13	0,173	0,17	6,748	0,009	0,007	0,077	0,074	0	99,991	0
14	0,147	0,144	5,993	0,002	0,006	0,058	0,062	0	99,994	0
15	0,055	0,054	7,263	0,002	0,002	0,093	0,093	0	99,996	0
16	0,029	0,028	6,406	0,001	0,001	0,068	0,068	0	99,998	0
17	0,047	0,046	7,052	0,002	0,002	0,085	0,085	0	99,999	0
18	0,018	0,011	6,191	0,008	0	0,063	0,055	0	99,999	0
19	0	0	6,4	0,03	0	0,068	0,038	0	99,999	0
20	0	0	6,944	0,036	0	0,082	0,046	0	100	0
21	0	0	4,404	0,037	0	0,026	0	-0,012	99,553	0
22	0	0	3,525	0,038	0	0,014	0	-0,025	98,605	0
23	0	0	3,833	0,016	0	0,017	0,001	0	98,605	0
24	0	0	2,588	0,002	0	0,005	0	0,003	98,703	0

Análisis de Componentes Principales desarrollado en Energías Renovables. Aplicación a Sistemas Dinámicos y Biomédicos

XLSTAT 7.5.3 - Análisis de Componentes Principales (ACP) - el 24/10/2011 a 08:33:09 p.m.
 Tabla: libro = Ejemplo_1_XLSTAT.xls / hoja = Datos 24 hs / rango = \$B\$2:\$K\$25 / 24 filas y 10 columnas
 Ponderación uniforme (predeterminado)
 Ningún dato omitido detectado
 Coeficiente de correlación de Pearson (ACP normalizado, varianzas en 1/n)
 Sin rotación de los ejes
 Número de factores asociados a valores propios no triviales: 10

Prueba de esfericidad de Bartlett:

Chi-cuadrado (valor obse	491,322
Chi-cuadrado (valor crític	61,656
GDL	45
p-value unilateral	< 0,0001
Alpha	0,05

Conclusión:

Al umbral de significación Alfa=0,050 se puede rechazar la hipótesis nula de ausencia de correlación significativa entre las variables. Dicho de otro modo, la correlación entre las variables es significativa.

Media y desviación típica de las columnas:

	Media	Desviación típica
Global Solar kW/m2	0,038	0,058
Incident Solar kW/m2	0,037	0,058
Wind Speed m/s	4,869	1,813
DC Primary Load kW	0,010	0,012
PV Power kW	0,001	0,002
SW AIR X kW	0,042	0,034
Excess Electricity kW	0,035	0,034
Bat Power kW	-0,001	0,006
Bat State of Charge %	99,791	0,446
Bat Energy Cost \$/kWh	0,042	0,200

Tabla 2. Valor medio y desviación estándar (elaboración propia)

Tabla 3. Matriz de correlación (elaboración propia)

Matriz de correlación:

	Global Solar kW/m2	Incident Solar kW/m2	Wind Speed m/s	DC Primary Load kW	PV Power kW	SW AIR X kW	Excess Electricity kW	Bat Power kW	Bat State of Charge %	Bat Energy Cost \$/kWh
Global Solar kW/m2	1	0,999	0,412	-0,306	0,995	0,355	0,506	0,220	0,267	-0,136
Incident Solar kW/m2	0,999	1	0,411	-0,304	0,996	0,355	0,505	0,219	0,262	-0,133
Wind Speed m/s	0,412	0,411	1	0,164	0,412	0,982	0,933	0,162	0,359	-0,126
DC Primary Load kW	-0,306	-0,304	0,164	1	-0,283	0,141	-0,144	-0,651	-0,366	-0,130
PV Power kW	0,995	0,996	0,412	-0,283	1	0,362	0,505	0,206	0,251	-0,126
SW AIR X kW	0,355	0,355	0,982	0,141	0,362	1	0,950	0,190	0,379	-0,154
Excess Electricity kW	0,506	0,505	0,933	-0,144	0,505	0,950	1	0,288	0,458	-0,125
Bat Power kW	0,220	0,219	0,162	-0,651	0,206	0,190	0,288	1	0,494	0,051
Bat State of Charge %	0,267	0,262	0,359	-0,366	0,251	0,379	0,458	0,494	1	0,098
Bat Energy Cost \$/kWh	-0,136	-0,133	-0,126	-0,130	-0,126	-0,154	-0,125	0,051	0,098	1

En negrita, valores significativos (fuera diagonal) al umbral $\alpha=0,050$ (prueba bilateral)

Tabla 4. Valores propios (elaboración propia)

Valores propios:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Valor propio	4,622	2,007	1,632	0,913	0,503	0,299	0,019	0,003	0,001	0,000
% varianza	46,224	20,072	16,319	9,134	5,027	2,992	0,190	0,034	0,005	0,001
% acumulado	46,224	66,297	82,616	91,750	96,777	99,769	99,960	99,994	99,999	100,000

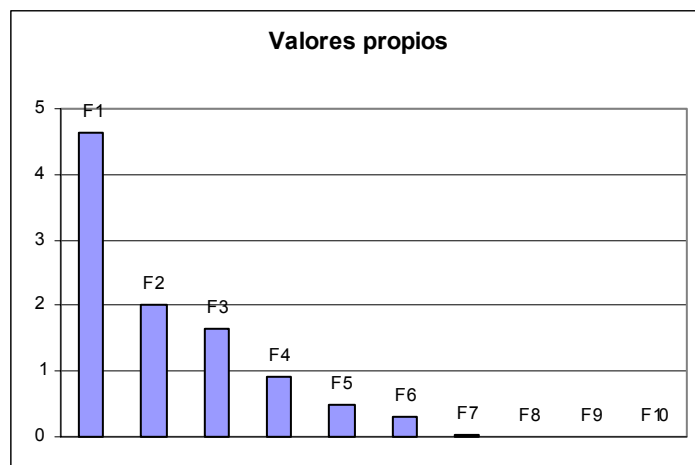


Figura 1. Valores propios (elaboración propia)

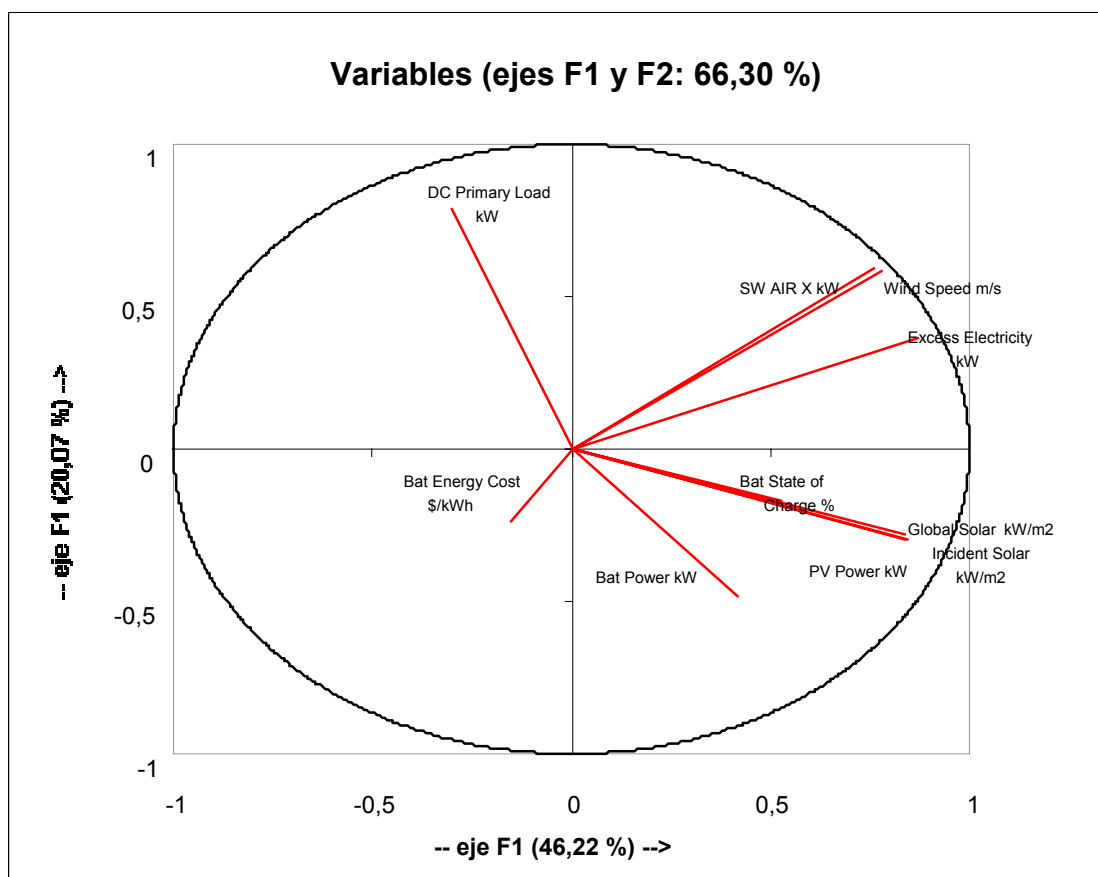


Figura 2. Coordenadas variables F1–F2 (elaboración propia)

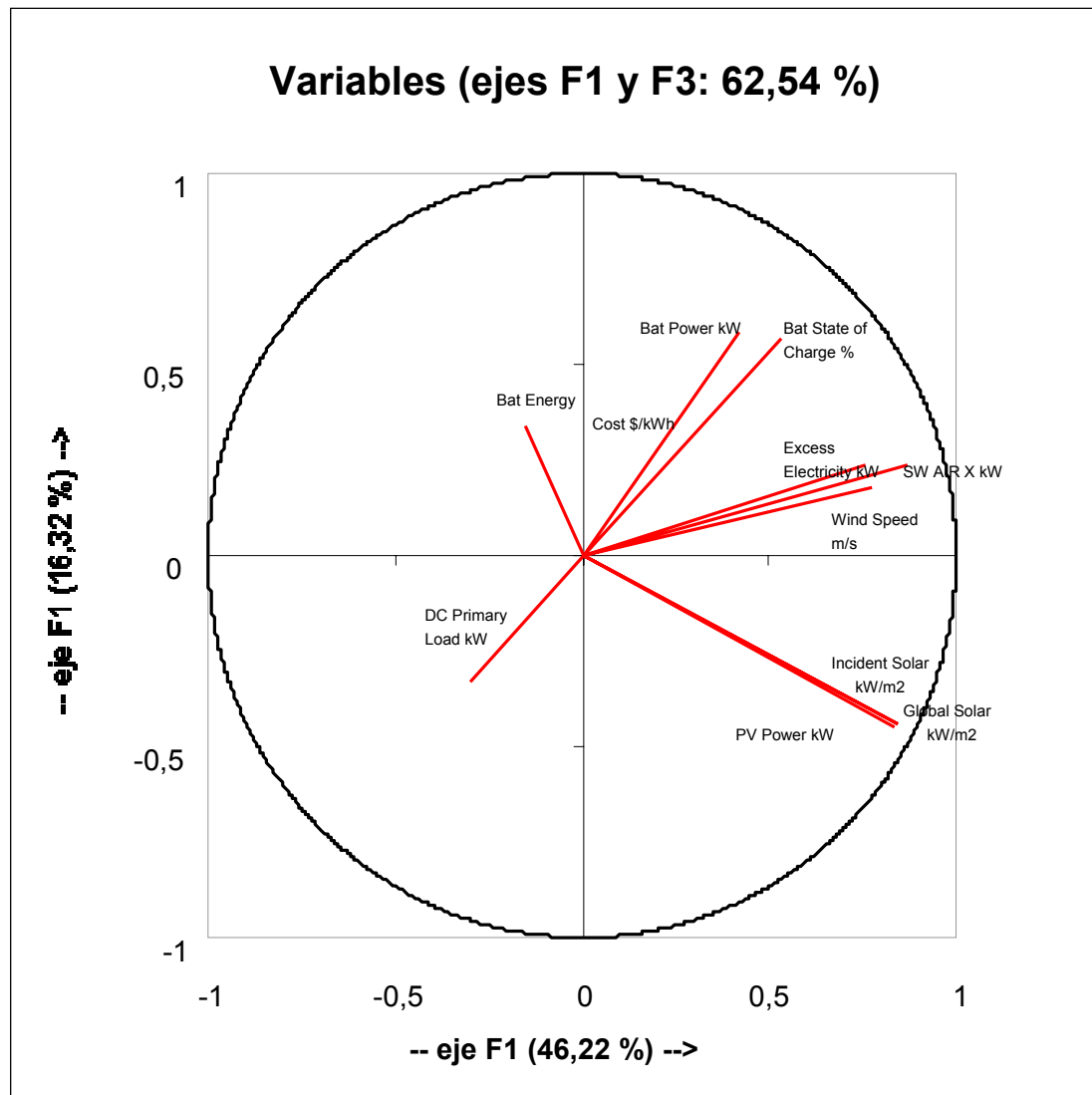


Figura 3. Coordenadas variables F1–F3 (elaboración propia)

DIAGRAMA DE PARETO

El diagrama de Pareto es una técnica gráfica que clasifica los elementos desde el más frecuente hasta el menos frecuente. Exhibe visualmente el orden de importancia, es decir, la contribución de cada elemento al efecto total.

Procedimiento:

- Seleccionar los elementos que serán analizados.
- Seleccionar el período de tiempo de los datos.
- Listar los elementos de izquierda a derecha en el eje horizontal en orden de magnitud decreciente.
- Dibujar dos (2) ejes verticales, uno de cada extremo del eje horizontal.
- La escala de la izquierda se calibra en la unidad de medición y su altura debe ser igual a la suma de todos los elementos.
- Arriba de cada elemento se dibuja una barra cuya altura representa la magnitud de la unidad de medición para ese elemento.

- Se elabora una línea de frecuencia acumulada (eje vertical derecho) sumando las magnitudes de cada elemento de izquierda a derecha.

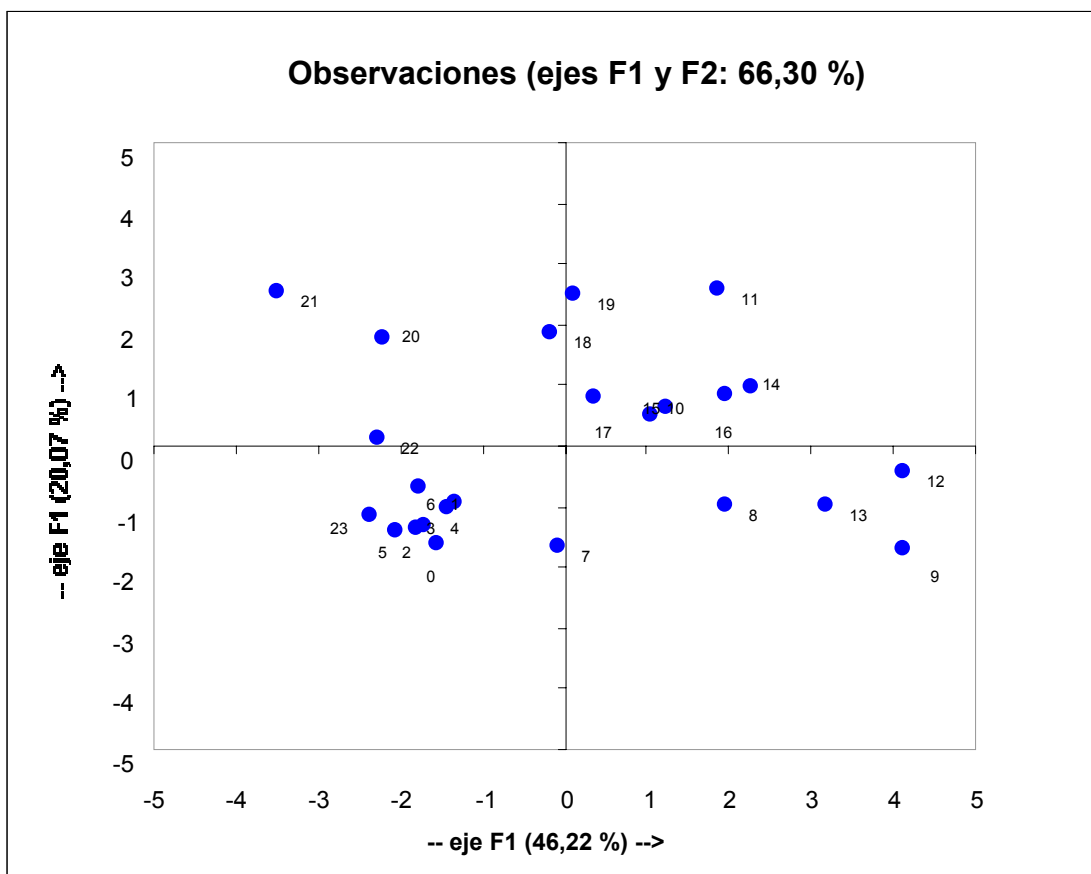


Figura 4. Coordenadas observaciones F1–F2 (elaboración propia)

Tabla 10. XLSTAT – Tabla de Pareto (elaboración propia)

X	Autovalor	% Acumulado
F1	4,622	46,224
F2	2,007	66,297
F3	1,632	82,616
F4	0,913	91,75
F5	0,503	96,777

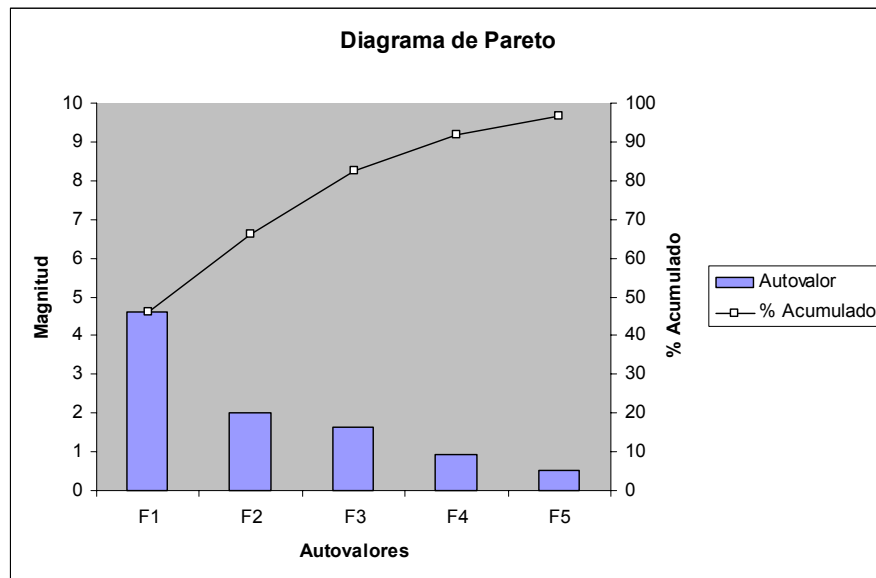


Figura 8. Diagrama de Pareto (elaboración propia)

DISCUSIÓN DE RESULTADOS

Matriz de Correlación

El primer resultado interesante a analizar es la matriz de la correlación (Tabla 3). Observamos el resultado evidente que los índices de Global Solar kW/m² e Incident Solar kW/m² son perfectamente correlacionados ($r \approx 1$); idem con PV Power y Global Solar kW/m² e Incident Solar kW/m²; ocurre lo mismo con SW Air x kW y Wind Speed m/s. Las cinco variables son entonces redundantes; y por última, en menor medida, Excess Electricity kW con Wind Speed m/s y SW Air x kW. Observamos que Bat Energy Cost \$/kWh, es muy poco correlacionada con las otras variables, lo mismo que DC Primary Load KW y Bat State of Charge %.

Valores Propios

El siguiente cuadro y el gráfico asociado son vinculados a un objeto matemático, los valores propios (Tabla 4), que son vinculados a un concepto muy simple: la calidad de la proyección cuando pasamos de N dimensiones (N siendo el número de variables; en este trabajo 10) a un número más débil de dimensiones. En nuestro caso, observamos que el primer valor propio vale 4,622 y representa aproximadamente el 46% de la variabilidad. Eso significa que si representamos los datos en un sólo eje, tendremos entonces siempre 46% de la variabilidad total que será preservada. En la Figura 1 a cada valor propio corresponde un factor. Cada factor es en realidad una combinación lineal de las variables de inicio. Los factores tienen la particularidad de no ser correlacionados entre ellos. Los valores propios y los factores son ordenados en orden descendente de variabilidad representada. Idealmente, los dos primeros valores propios corresponden a un porcentaje elevado de la variabilidad, de manera que la representación sobre los dos primeros ejes factoriales es de buena calidad. En nuestro ejemplo, ese no es exactamente el caso, de donde se ve la necesidad de utilizar además de los gráficos de los factores F1 y F2, los gráficos en F1 y F3. Vemos aquí que el número de factores (descriptores) es 5, cuando teníamos al principio 10 variables. Eso es debido a las cinco variables redundantes. Se decidió utilizar un quinto descriptor (5 autovalores) para acumular en 96,777 % de información, igualmente su pudo haber eliminado una variable más

(Excess Electricity kW) y trabajar solamente con cuatro (4), pero la información acumulada sería solamente del 91,750 %.

Variables F1-F2

La Figura 2 es un gráfico que representa el círculo de las correlaciones (el círculo en los ejes F1 y F2). Corresponde a una proyección de las variables iniciales sobre un plano de dos dimensiones constituido por los dos primeros factores. Cuando dos variables se encuentran lejos del centro del gráfico, entonces si están: cercas unas de las otras, se dice que están significativamente correlacionadas en forma positiva (r cerca de 1); si se encuentran ortogonales unas de las otras, entonces están significativamente no-correlacionadas (r cerca de 0); si se encuentran simétricamente opuestas con respecto al centro, entonces están significativamente correlacionadas en forma negativa (r cerca de -1). Cuando las variables están relativamente cerca del centro del gráfico, entonces cualquiera interpretación es arriesgada, y es necesario referirse a la matriz de correlaciones o a otros planos factoriales para interpretar los resultados.

Observaciones F1-F2

El gráfico de la Figura 4 a continuación corresponde a uno de los objetivos del PCA. Permite representar los individuos en una carta de dos dimensiones, y así identificar tendencias. Así observamos para los horarios correspondientes, aquellos que poseen las mismas características. A la derecha del eje de ordenadas vemos para el horario matutino y vespertino, y a la izquierda para el horario nocturno (observación de F1 y F2). El Análisis de Componentes Principales es a menudo utilizado ante una regresión ya que permite evitar el uso de variables redundantes, o ante una clasificación ya que permite identificar la estructura de la población y eventualmente determinar el número de grupos a construir; consiste en expresar un conjunto de variables en un conjunto de combinaciones lineales de factores no correlacionados entre sí. Este método permite representar los datos originales (individuos y variables) en un espacio de dimensión inferior del espacio original, mientras limite al máximo la pérdida de información.

Diagrama de Pareto

Luego de aplicar las herramientas que proporciona XLSTAT para la determinación de los Componentes Principales se obtuvo la matriz de correlaciones, la cual fue transformada arbitrariamente en el vector **X** en orden ascendente en función de los pesos de los valores propios de la matriz. La Figura 8 muestra el diagrama de Pareto obtenido en función de los Componentes Principales (valores propios) seleccionados. El análisis de Pareto es una comparación ordenada de factores relativos a un problema. Esta comparación ayuda a identificar y enfocar los pocos factores vitales diferenciándolos de los muchos factores útiles. La aplicación del mismo permite exhibir visualmente en orden de importancia, la contribución de cada elemento en el efecto total. En el gráfico sólo se visualizan los 5 primeros Componentes Principales debido a que los pesos de los 5 restantes son insignificantes respecto de los primeros, en los que se concentra más del 96 % de la información de la matriz original.

Señal Electrocardiográfica

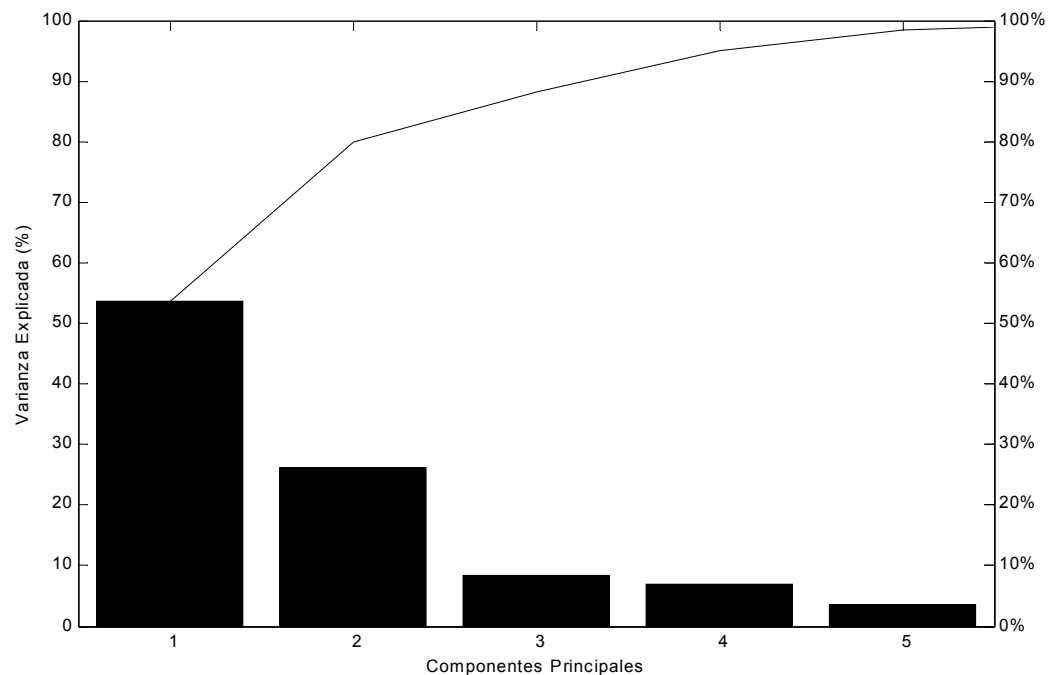


Figura 11. Diagrama de Pareto de los Componentes Principales.

Luego de aplicar las herramientas que proporciona Matlab para la determinación de los Componentes Principales se obtuvo la matriz de correlaciones que tiene una dimensión de 30 x 30, la cual fue transformada arbitrariamente en la matriz X en orden ascendente en función de los pesos de los valores propios de la matriz. La figura 11 muestra el diagrama de Pareto obtenido en función de los Componentes Principales (vectores propios) seleccionados. El análisis de Pareto es una comparación ordenada de factores relativos a un problema. Esta comparación ayuda a identificar y enfocar los pocos factores vitales diferenciándolos de los muchos factores útiles. La aplicación del mismo permite exhibir visualmente en orden de importancia, la contribución de cada elemento en el efecto total. En el gráfico sólo se visualizan los 5 primeros Componentes Principales debido a que los pesos de los 25 restantes son insignificantes respecto de los primeros, en los que se concentra más del 98% de la información de la matriz original.

4. Conclusiones

El método de Análisis de Componentes Principales (PCA) es efectivo, y permitió cumplir con los objetivos mencionados anteriormente. Se logró reducir una matriz de datos de 10 variables a una matriz de 5 variables para el caso de las fuentes de energías renovables y de 30 en 5 para el caso del sistema QRS de la matriz cardíaca, en las que se concentra casi el 100% de la información. Con esta herramienta se eliminó la redundancia de datos para agilizar los tiempos computacionales, lo que constituye un objetivo primordial en el procesamiento de información.

5. Referencias

- [1] **Comon P.**, 1994. *Independent component analysis, A new concept?*. Signal Processing, Vol. 36, pp. 287–314.
- [2] **Pisarello MI, Álvarez Picaza C, Veglia JI**, *Compresión de Bio-Datos – Utilización del MATLAB aplicada al Análisis de Componentes Principales*. 1er Congreso Latinoamericano de Ingeniería Prospectiva – Exposición Internacional de Ingeniería. Corrientes, 2005.
- [3] **Pisarello, MI**, *Wavelets, Bancos de Filtros y Análisis de Señales Biomédicas* (Tesis Doctoral) – Universidad Nacional del Nordeste – 2012.
- [4] **Anderson TW**, 1963. *Asymptotic Theory for Principal Components Analysis*. The Annals of Mathematical Statistics. Vol 34, Num 1, Pp 122-148.
- [5] **Pisarello MI, Álvarez Picaza C, Monzón JE**, *Reducción de Datos Electrocardiográficos a partir del Análisis de Componentes Principales*. Reunión de Comunicaciones Científicas y Tecnológicas. SGCyT. Universidad Nacional del Nordeste. Corrientes, 2005.
- [6] **Pisarello MI, Álvarez Picaza C, Monzón JE**, *Análisis de Componentes Principales para la Compresión del ECG*. Proceedings de la 5ta Conferencia Iberoamericana, Cibernética e Informática (CISCI 2006). Orlando - Florida - EE.UU.
- [7] **Hernandez M**, 1998. *Temas de Análisis Estadístico Multivariado*. Editorial Universidad de Costa Rica.