ANÁLISIS DE DATOS MULTIVARIANTE

"Verdaderamente no hay cultura más que cuando el espíritu se ensancha a la dimensión de lo universal"
J. Leclercq.

Prof. Esperanza Ayuga Téllez

DEFINICIÓN:

"El Análisis Multivariante (Cuadras, 1981) es la rama de la Estadística y del análisis de datos, que estudia, interpreta y elabora el material estadístico sobre un conjunto de n>1 de variables, que pueden ser cuantitativas, cualitativas o una mezcla."

OBJETIVOS:

1. Resumir los datos mediante un pequeño conjunto de nuevas variables con la mínima pérdida de información.
2. Encontrar grupos en los datos, si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables

APLICACIONES:

1. Administración de empresas: para construir tipología de clientes.
2. Agricultura: para clasificar terrenos de cultivo por fotografía aérea.
3. Arqueología: clasificar restos arqueológicos.
4. Biometría: identificar los factores que determinan la forma de un organismo vivo.
5. Computación: diseñar algoritmos de clasificación automática.
6. Educación: para investigar la efectividad del aprendizaje a distancia.

APLICACIONES:

7. Medio Ambiente: dimensiones de la contaminación ambiental.
8. Documentación: para clasificar revistas por su contenido.
9. Economía: dimensiones del desarrollo económico.
10. Geología: clasificar sedimentos.
11. Linguistica: encontrar patrones de asociación de palabras.
12. Medicina: para identificar tumores.
13. Psicología: para identificar factores que componen la inteligencia humana.

El análisis de datos multivariante puede plantearse a dos niveles:

❶ Queremos extraer la información que contienen los datos disponibles ⇒ **EXPLORACIÓN DE DATOS** (o *minería de datos)*

❷ Buscamos obtener conclusiones sobre la población que ha generado los datos lo que requiere construir un modelo que explique su obtención y permita prever valores futuros ⇒ **INFERENCIA**

CLASIFICACIÓN:

OBJETIVOS	DESCRIPTIVA	INFERENCIA
Resumir datos	Descripción de datos	*Constrcn. de modelos*
Obtener indicadores	Componentes principales *Escalado multidimens.* *Anl. de correspondencias*	Anl. Factorial
Clasificar	Anl. de conglomerados	Anl. Discriminante
Agrupar	Anl. de conglomerados	*Clasción. con mezcla*
Relacionar variables	Regresión múltiple Regresión multivariable	Correlac. Canónicas

Son necesarios algunos conocimientos de Álgebra matricial:

- La traspuesta de **A**, **A'**, se forma cambiando las filas de **A** por sus columnas.
- La traza de **A**, matriz cuadrada (nxn), es la suma de los elementos de la diagonal principal. Es un operador lineal.
- La inversa de **A**, $\mathbf{A}^{-1}$, es la matriz que cumple que $\mathbf{AA}^{-1}=\mathbf{I}$.
- Los vectores propios de **A**, matriz cuadrada (nxn), son aquellos cuya dirección no se modifica al transformarlos mediante la matriz. Cumplen que $\mathbf{Au}=\lambda\mathbf{u}$ con $||\mathbf{u}||=1$.
  El vector **u** es el vector propio.
  λ es un escalar denominado valor propio.
  La ecuación característica de la matriz es $|\mathbf{A}-\lambda\mathbf{I}|=0$.

DESCRIPCIÓN DE DATOS

El análisis descriptivo debe ser siempre un primer paso para comprender la estructura de los datos.

Pueden ser:

1. Cualitativos
   - Binarios
   - Numéricos si se pueden ordenar
2. Cuantitativos

DESCRIPCIÓN DE DATOS

- Los datos se expresarán mediante una matriz **X** de dimensiones nxp, llamada matriz de datos, de elementos x_{ij}, con i=1,...,n que representan el individuo y j=1,...,p que representan las variables.

p.e. si medimos en cinco individuos, tres características distintas, tenemos:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix}$$

DESCRIPCIÓN DE DATOS

Llamamos $\mathbf{X}= (x_1, x_2, ..., x_p)'$ a la variable multivariante, formada por las p variables escalares, que toman valores en los n elementos observados.

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

MEDIDAS DE ANÁLISIS UNIVARIANTE

- Conviene calcular las p medias y las medianas en el análisis inicial ya que si existen datos anormales o heterogeneidad en la distribución la media puede no ser una medida adecuada.
- Representación del histograma o gráfico de cajas para comprobar la distribución.
- Otras medidas:

Desviación $d_{ij} = \left(x_{ij} - \bar{x}_j\right)^2$

Coeficiente de homogeneidad $H_j = \dfrac{\frac{1}{n}\sum (d_{ij} - s_j^2)^2}{s_j^4}$

MEDIDAS DE ANÁLISIS MULTIVARIANTE

- Medidas de centralización: matriz de medias.

- Medidas de dispersión:

  – matriz de varianzas covarianzas: **S**

  la traza, el determinante y los valores propios son no negativos. Si el rango de **S**, r(**S**)=h<p y existen p-h variables redundantes que pueden eliminarse.

MEDIDAS DE ANÁLISIS MULTIVARIANTE

- Medidas de dispersión:

  – medidas globales que se usan con v. adimen.:

  V. Total: $\mathbf{T}=tr(\mathbf{S})$ no tiene en cuenta dependencia

  V. Media: $\bar{s}^2 = 1/p \sum_{i=1}^{p} s_i^2$ sin considerar dependencia

  V. Global: $VG=|\mathbf{S}|$, si p=2 mide el área, con p=3 volumen. No sirve para comparar conjuntos con distinto nº de v. Es distinta si usamos distintas unidades.

  V. Efectiva: $VE =|\mathbf{S}|^{1/p}$ para evitar lo anterior.

MEDIDAS DE ANÁLISIS MULTIVARIANTE

- Medidas de dependencia lineal:
  - Matriz de correlaciones: **R** de $r_{ij}=s_{ij}/(s_{ii}s_{jj})$
  - Coeficiente de correlación múltiple: $R_j^2 = 1 - \frac{s_r^2}{s_j^2}$
  - Coeficiente de correlación parcial entre x_j y x_i:

$$r_{ij,12\ldots p} = -\frac{s^{ij}}{\sqrt{s^{ii}s^{jj}}}$$

Donde s^{ij} es el elemento i, j de la matriz de precisión $\mathbf{S}^{-1}$

P = matriz de corr. parciales contiene los coef. de corr. parciales entre pares de var.

DISTANCIAS

- Distancia euclídea: $d_{ij} = \left[\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right]^{1/2}$

- Distancia de Mahalanobis: distancia entre un punto y su vector de medias. Tiene en cuenta la correlación y por tanto la estructura de los datos.

$$d_i = \left[(x_i - \bar{x})' \mathbf{S}^{-1} (x_i - \bar{x}) \right]^{1/2}$$

- Distancia promedio: $V_m = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x}) = \mathrm{Tr}(\mathbf{S})$

$$V_{mp} = \frac{1}{np} \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x}) = \bar{s}^2$$

TRANSFORMACIONES

- Lineales:
  - *estandarización univariante: $y_i = (x_i - \bar{x}) / s_i$
  - *estandarización multivariante: $\mathbf{y} = \mathbf{S}_x^{-1/2}(\mathbf{x} - \bar{\mathbf{x}})$
- No lineales:
  - *La más usada para datos positivos: y=lnx

Cuando mide tamaños, las diferencias relativas entre valores son importantes y queremos variabilidad independiente de las unidades de medida

*Familia Box-Cox: $\longrightarrow \begin{cases} y^{(\lambda)} = \dfrac{x^{\lambda} - 1}{\lambda}, \text{ para } \lambda \neq 0 \\ y^{(\lambda)} = \log x, \text{ para } \lambda = 0 \end{cases}$

DATOS ATÍPICOS

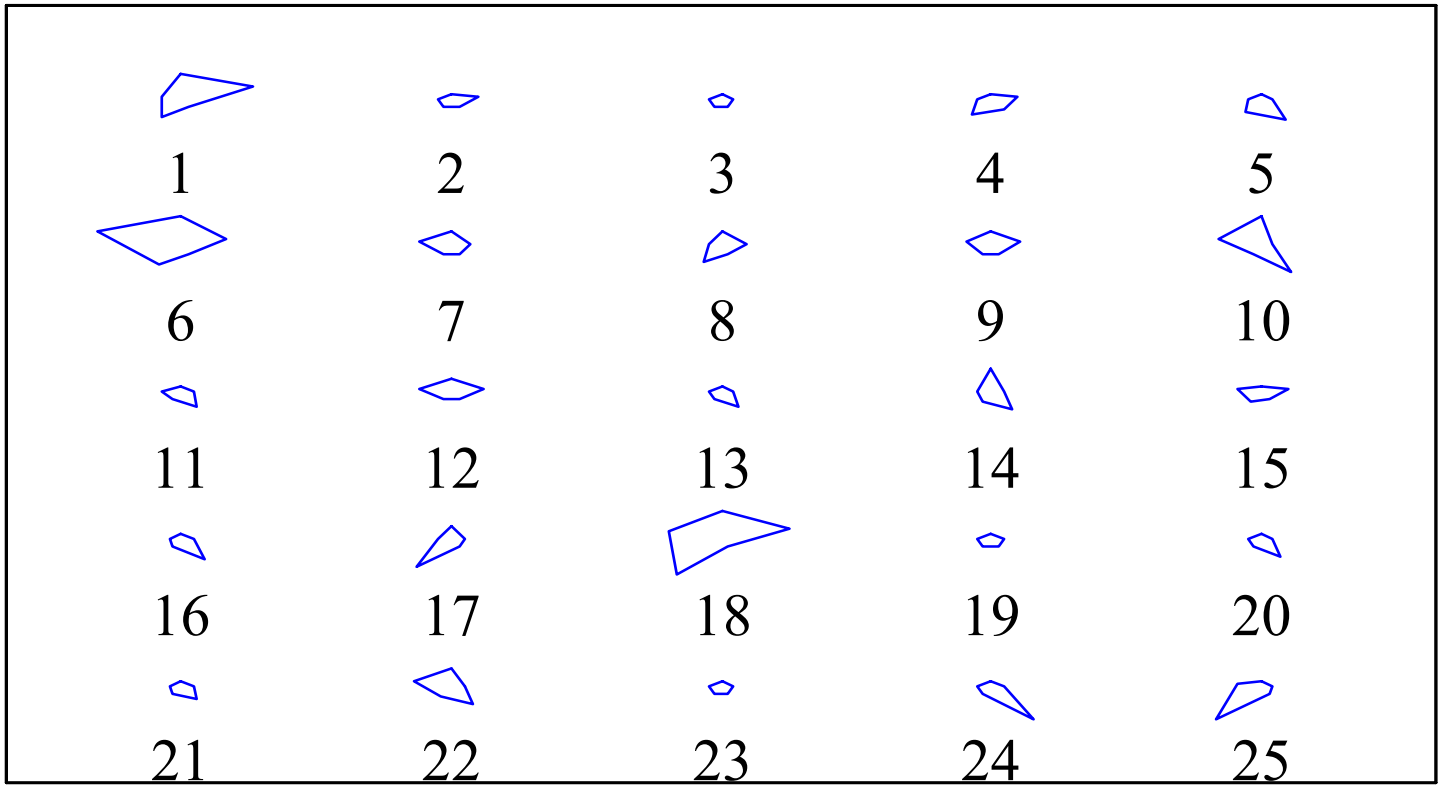
Son aquellos que parecen observados de forma distinta al resto (por errores de medida, cambio de instrumentos o heterogeneidad de los elementos

- Es frecuente entre 1-3% en experimentos controlados y alrededor de un 5% en experimentos incontrolados

-Las consecuencias de un dato atípico pueden ser graves:
distorsionan medias y desviaciones típicas

enmascaran las relaciones existentes entre ellas

DATOS ATÍPICOS



DATOS ATÍPICOS

Su identificación es necesaria para eliminarlos de los procedimientos habituales.

1. Para una variable (mediante gráficos de cajas)
2. Para multivariable (mediante proyección en la dirección de máxima curtosis o con gráficos de las observaciones.

Los datos detectados como atípicos deben ser estudiados con detalle $\Rightarrow$ pueden dar lugar a descubrimientos importantes

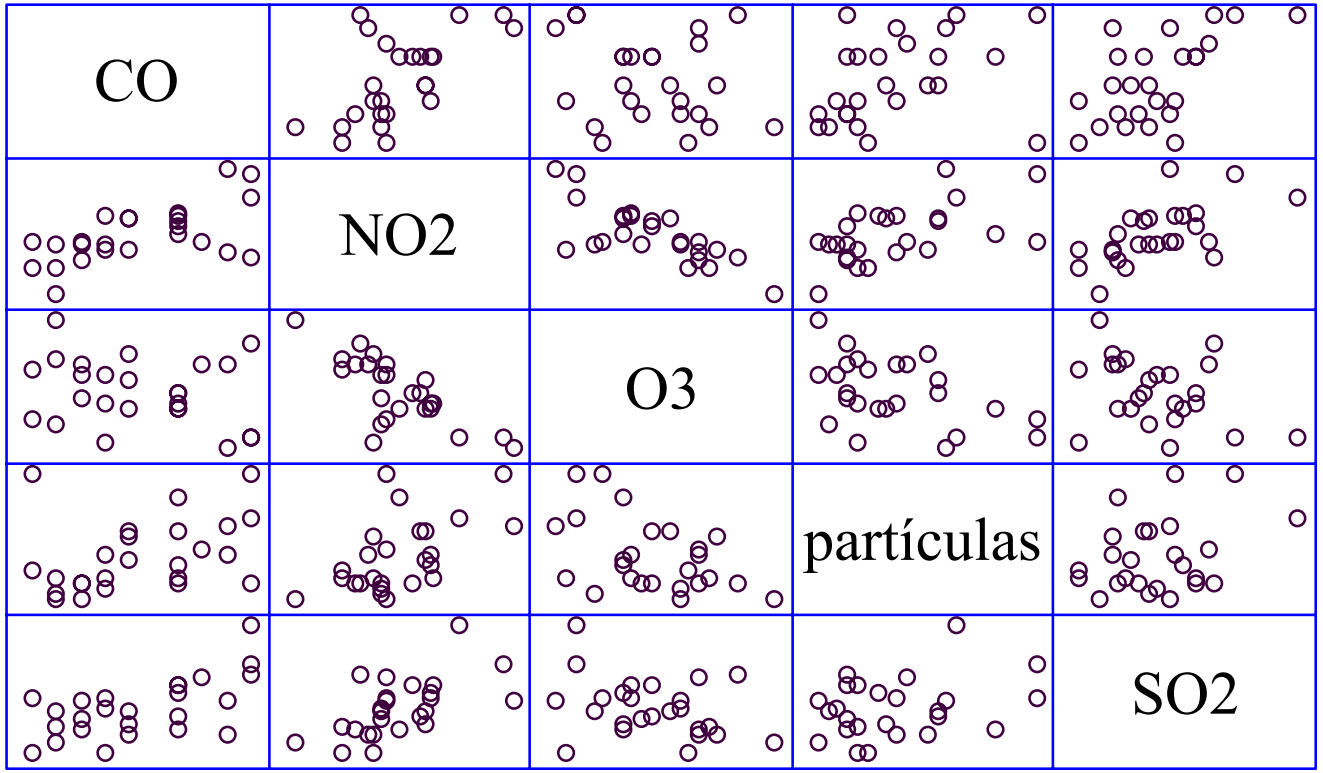
REPRESENTACIONES GRÁFICAS

- De dispersión 2 a 2: para detectar atípicos bivariantes y relaciones lineales.

- Tridimensional: de dispersión 3 a 3.

- Mediante figuras: para más de 3 variables. Se muestran los datos mediante figuras planas. En éstas los valores atípicos aparecen muy discordantes del resto.

REPRESENTACIONES GRÁFICAS



COMPONENTES PRINCIPALES

Técnica debida a Hotelling (1933) permite reducir las dimensiones del problema con mínima pérdida de información.

Tiene por objeto analizar si es posible representar adecuadamente un conjunto de n observaciones de p-variables con un número menor de variables construidas como combinaciones lineales de las originales.

p.e. Utilizamos observaciones de diferentes contaminantes atmosféricos para representar la calidad del aire. Podemos reducir las dimensiones del problema con un número menor de variables que combinen compuestos parecidos.

UTILIDAD

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio p-dimensional. Es un 1er paso para identificar variables generadoras de los datos.

2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas facilitando la interpretación de los datos.

PLANTEAMIENTO

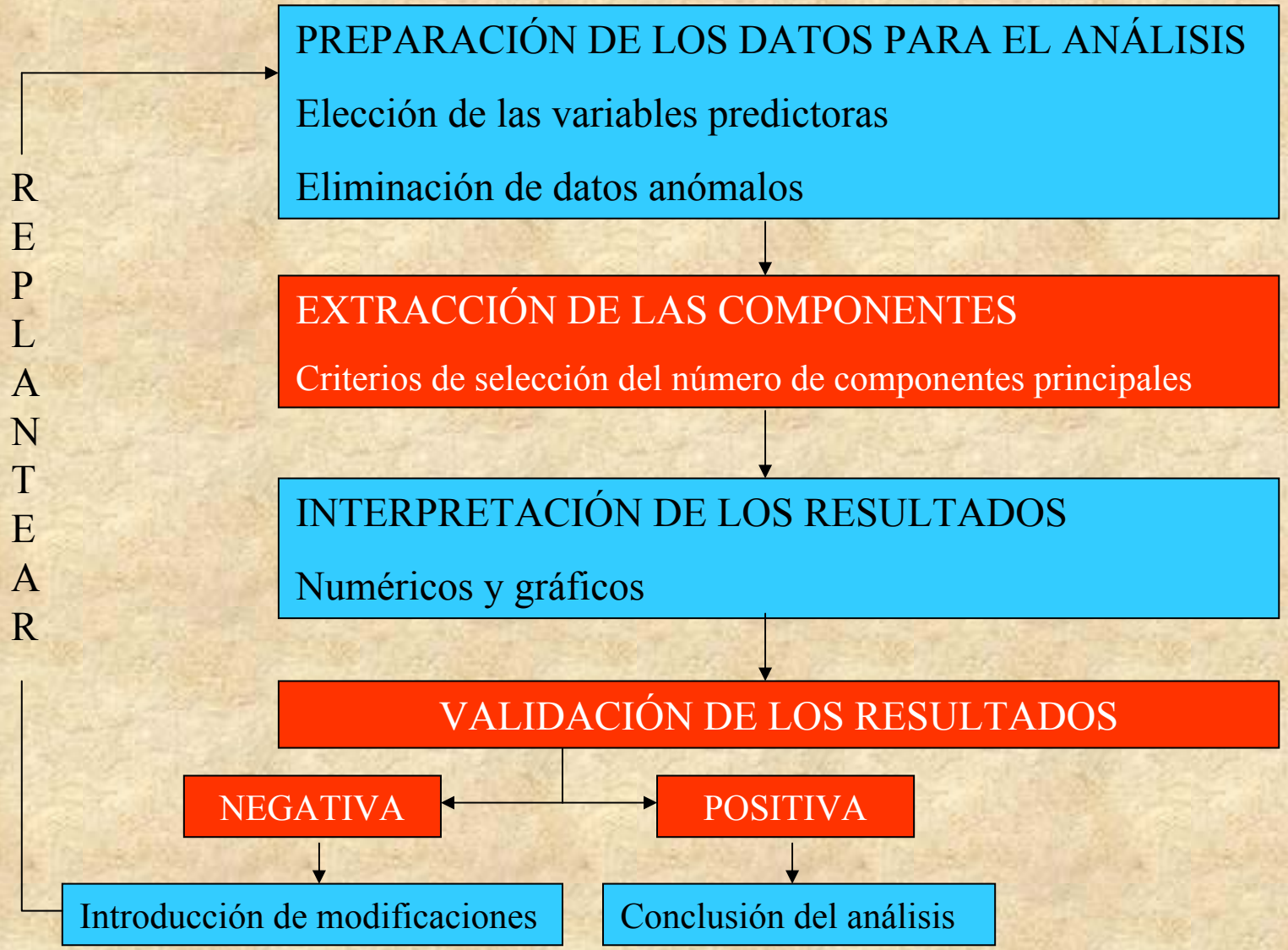
Se puede abordar desde tres puntos de vista:

- Descriptivo: encontrar un subespacio (dim<p) tal que al proyectar sobre él los puntos conserven la estructura (la menor distorsión posible)

- Estadístico: representar la v. con la min pérdida de información (máx correlación global con las originales)

- Geométrico: si los ptos se distribuyen en elipsoides su mejor descripción es la proyección sobre los ejes mayores de éste (valores incorrelados entre sí).

Fases de su aplicación



PREPARACIÓN DE LOS DATOS

- Elección de las variables

  Se seleccionan aquellas que resultan interesantes para el objetivo del estudio. Se suelen emplear todas.

- Estandarización de las variables

  Para evitar la influencia de las unidades de medida en la ponderación de los componentes

- Eliminación de datos anómalos

  Para evitar que se enmascaren relaciones existentes o se encuentren algunas inexistentes.

CÁLCULO

1ER Componente:

Se define como la combinación lineal de las variables originales que tienen V máxima: $\mathbf{z_1 = x a_1}$ con $\mathbf{a_1' a_1} = 1$.

Al maximizarla su solución es $\mathbf{a_1}$ igual al vector propio de **S** y si λ_1 es su valor propio, $V(\mathbf{z_1}) = \lambda_1$.

Por tanto, $\mathbf{a_1}$ (vector de coeficientes) es el vector propio de **S** asociado al mayor valor propio.

CÁLCULO

2ª Componente:

Se calcula max $V(\mathbf{z}_1)+ V(\mathbf{z}_2)$, si $\mathbf{z}_2 = \mathbf{x}\mathbf{a}_2$ y $\|\mathbf{a}_1\|=\|\mathbf{a}_2\|=1$.

También $\mathbf{a}_2$ es un vector propio de **S** tal que λ_2, su valor propio asociado es el 2º mayor de **S**. Se comprueba fácilmente que $\mathbf{a}_1$ y $\mathbf{a}_2$ están incorrelados.

Generalización:

Análogamente se puede calcular el espacio de dimensión r definido por los vectores propios asociados a los r mayores valores propios de **S** (**Z=XA** con **A'A=I**).

Calcular los C.P. Equivale a aplicar una transformación ortogonal a X para obtener las nuevas, Z, incorreladas entre sí.

SELECCIÓN DEL NÚMERO

1. Realizar un gráfico de valores propios frente a vectores propios. Seleccionar componentes hasta que los restantes tengan aproximadamente el mismo $\lambda_{i.}$

2. Seleccionar componentes hasta que se cubra una proporción determinada de varianza (80 o 90%). Se debe emplear con cuidado.

3. Desechar aquellos λ_i menores que la unidad (regla arbitraria).

PROPIEDADES

- Conservan la variabilidad inicial.
- La proporción de variabilidad explicada por un componente es $\lambda_h/\Sigma\lambda_i$.
- $Cov(z_i; x_1,..., x_p)= \lambda_i a_i$.
- $$\rho(z_i, x_j) = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$
- Las r C.P. proporcionan la predicción lineal óptima con r variables, del conjunto de variables X.
- Si estandarizamos los C.P. Se obtiene la estandarización multivariante de los datos originales.

ANÁLISIS NORMADO

Las C.P. se obtienen max la varianza de la proyección, cuando una v. tiene una varianza mucho mayor que las demás el 1er componente coincidirá aprox. con ésta v.

Para evitar esto, conviene estandarizar las v. antes de calcular los componentes $\Rightarrow$ los C.P. normados se obtienen calculando los vectores propios de **R**.

Si las diferencias entre v. son informativas no debemos estandarizar. En caso de duda conviene realizar ambos análisis y quedarse con el más informativo.

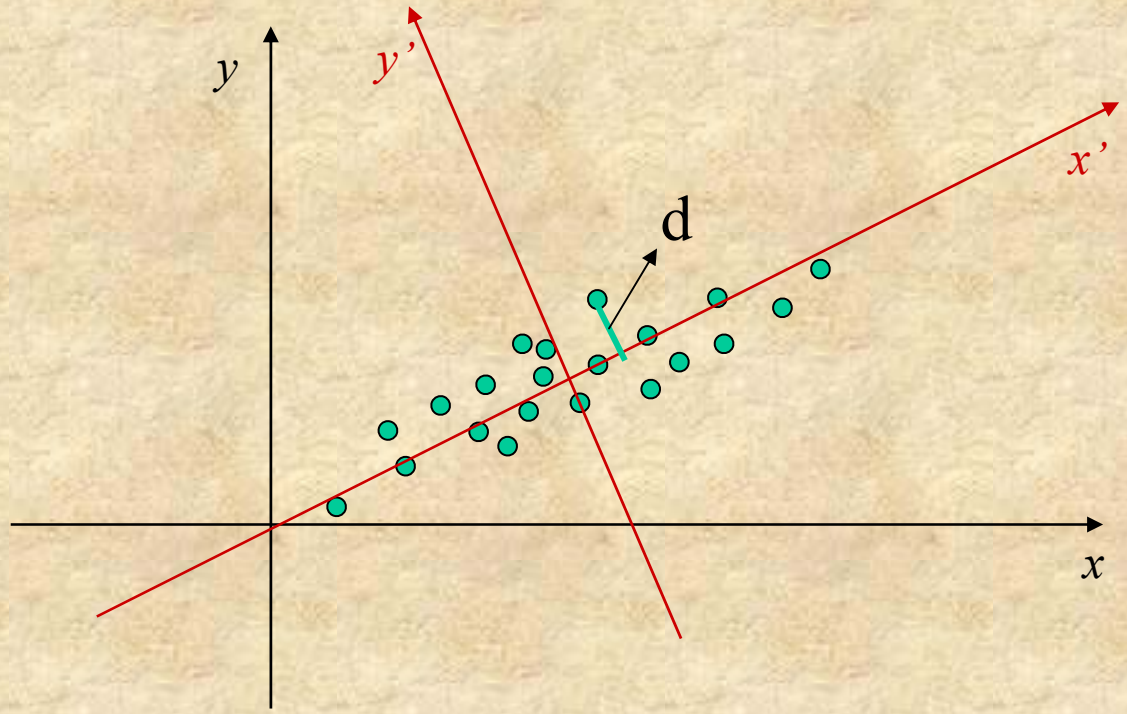
INTERPRETACIÓN

Cuando existe una alta correlación positiva entre todas las v., el 1er C.P. puede interpretarse como un factor global de tamaño, y los siguientes de forma (medias ponderadas de grupos contrapuestos por el signo).

La interpretación mejora con las proyecciones de las observaciones sobre los planos definidos por las parejas de componentes más importantes.

Si existen relaciones fuertes pero no lineales este análisis puede dar una información muy parcial.

INTERPRETACIÓN



Se minimizan los cuadrados de las distancias (d) al eje *x'*

El eje *y'* se calcula ortogonal al *x'*

OTRAS ACTUACIONES

1. Antes de obtenerlos conviene asegurarse de que no hay atípicos que distorsionen la matriz **S**.

2. Pueden verse como un conjunto nuevo de variables y estudiarse sus distribuciones (e investigar relaciones no lineales)

3. Las C.P. generalizados constituyen componentes con v. adicionales (x^2 y $x_i x_j$) que pueden detectar relaciones no lineales mediante λ_i próximos a 0. El inconveniente es que aumenta la dimensión.

ULTIMAS APLICACIONES

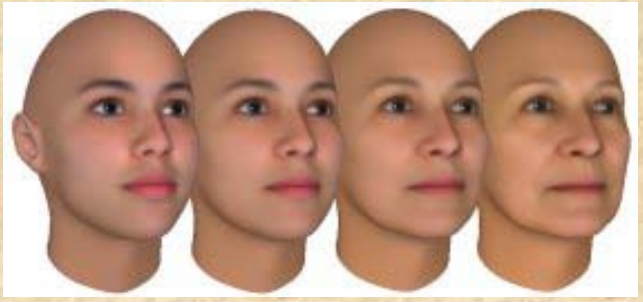
Para identificar personas mediante una base de datos de imágenes 3D de los rostros.

- 10,000 puntos en cada imagen
- x, y, z, R, G, B – 6 numeros para cada punto
- Por ello, cada imagen tiene 10,000*6 =**vector de 60.000 dimensiones**

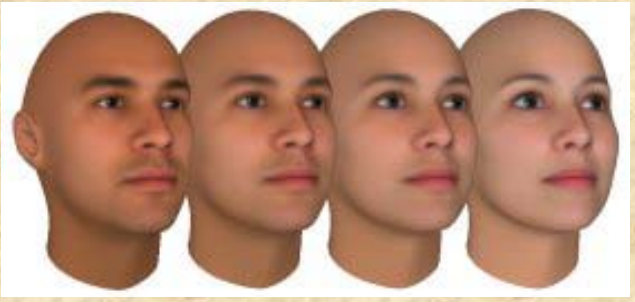
ULTIMAS APLICACIONES

¿Qué interés tiene un espacio de 60.000 dimensiones?.

Las CP reducen esta variabilidad en un número mucho menor de variables. Ejes de edad, genero...



Eje edad

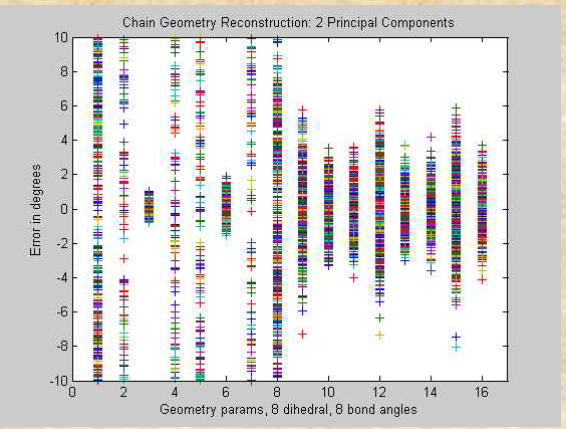


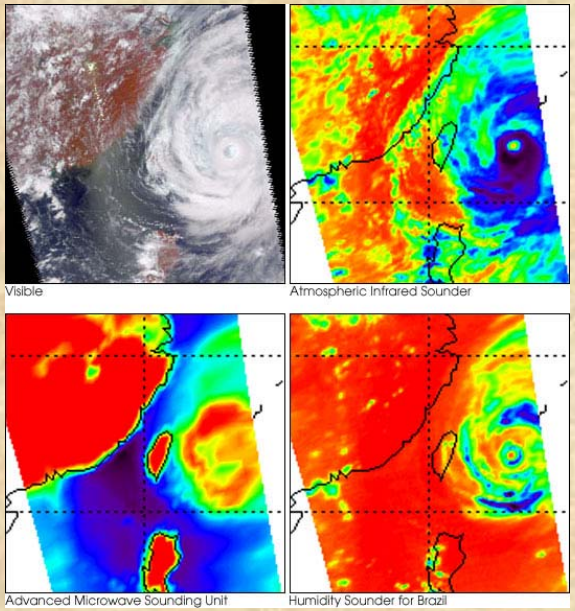
Eje genero

ULTIMAS APLICACIONES

Identificación de proteínas (2005):

Se identifican las proteínas responsables del movimiento de la espina dorsal basándose en los ángulos de éstas.





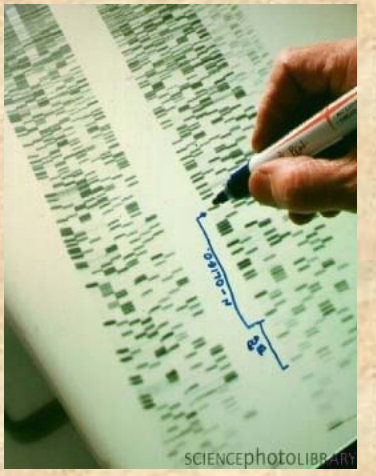
Estudios sobre catástrofes (2005):

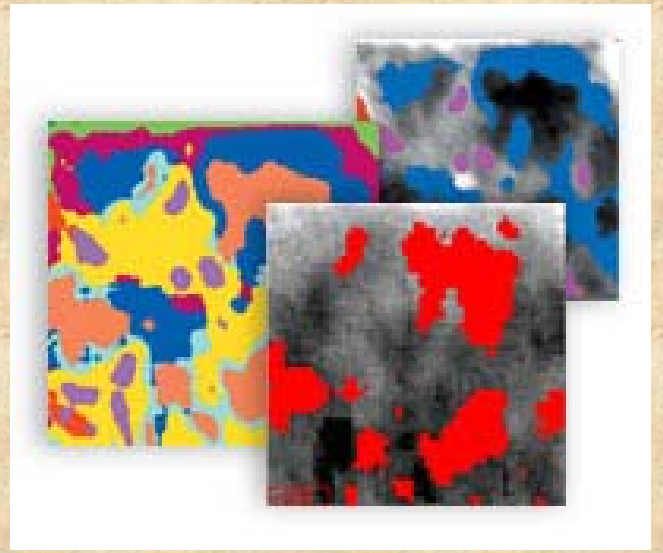
Para encontrar nuevas estructuras en los datos climáticos sobre huracanes que permitan prevenir sus efectos.

ULTIMAS APLICACIONES

Estudios sobre el cáncer (desde el año 2000) :

Para encontrar estructuras genéticas relacionadas con el cáncer.



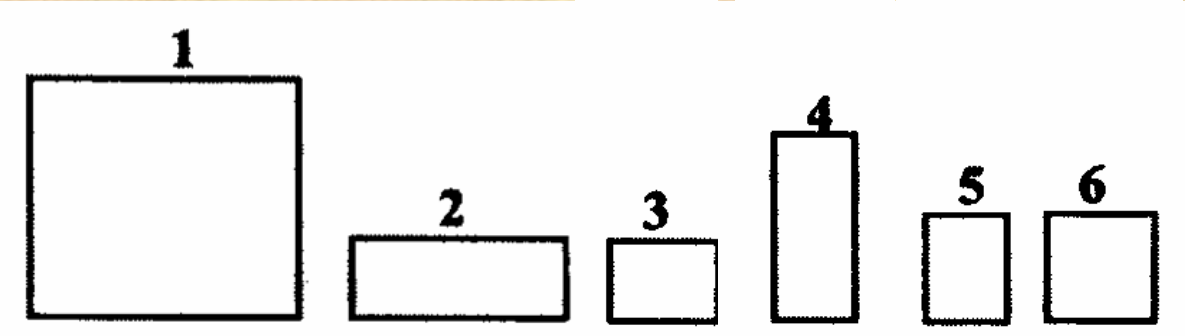


Estudios farmacológicos :

Para encontrar estructuras espaciales de la actividad de los medicamentos mediante imágenes.

Ejemplo 1:

Disponemos de 6 observaciones en 2 dimensiones, cada observación corresponde a un rectángulo y las variables son *longitud de la base* y *altura del rectángulo*. Gráficamente las observaciones son:



Que corresponde a la matriz de datos:

$$\mathbf{X} = \begin{bmatrix} 2 & 2 \\ 1{,}5 & 0{,}5 \\ 0{,}7 & 0{,}5 \\ 0{,}5 & 1{,}5 \\ 0{,}5 & 0{,}7 \\ 0{,}7 & 0{,}7 \end{bmatrix}$$

Ejemplo 1:

Aplicamos logaritmos a estos datos para facilitar la interpretación de los componentes. La matriz de covarianzas de log(**X**) es:

$$\mathbf{S} = \begin{bmatrix} 6{,}39 & 1{,}41 \\ 1{,}41 & 6{,}39 \end{bmatrix}.10^{-2}$$

Los autovalores y autovectores de la descomposición de esta matriz son: λ_1=0,78 λ_2=0,0498

$$\mathbf{a}_1 = \begin{bmatrix} 0{,}707 \\ 0{,}707 \end{bmatrix} \qquad \mathbf{a}_2 = \begin{bmatrix} 0{,}707 \\ -0{,}707 \end{bmatrix}$$

Ejemplo 1:

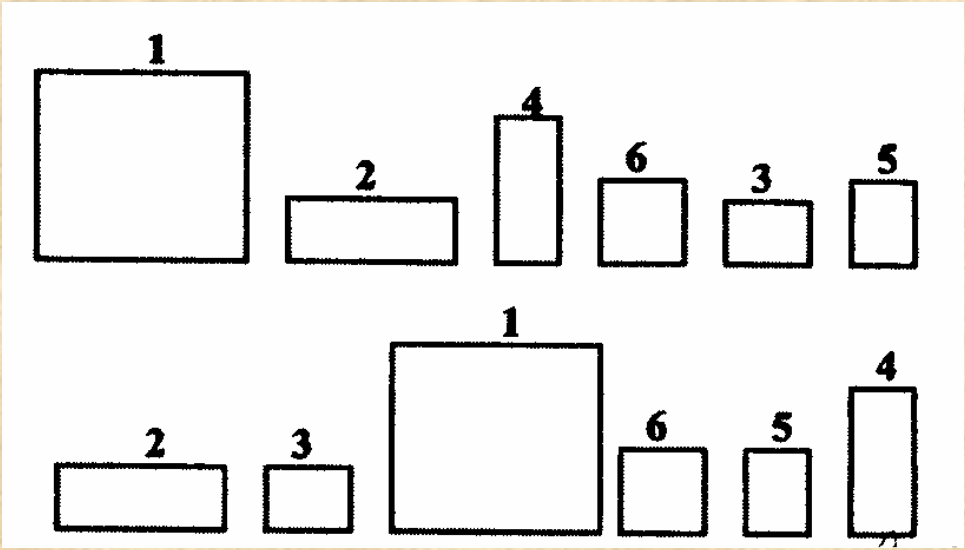
Las dos primeras componentes son :

$$Z_1 = Xa_1 = 0{,}707\log(X_1) + 0{,}707\log(X_2) = 0{,}707\log(X_1X_2) = \begin{bmatrix} 0{,}426 \\ -0{,}088 \\ -0{,}322 \\ -0{,}088 \\ -0{,}322 \\ -0{,}219 \end{bmatrix}$$

$$Z_2 = Xa_2 = 0{,}707\log(X_1) - 0{,}707\log(X_2) = 0{,}707\log(X_1/X_2) = \begin{bmatrix} 0 \\ 0{,}337 \\ 0{,}103 \\ -0{,}337 \\ -0{,}103 \\ 0 \end{bmatrix}$$

Ejemplo 1:

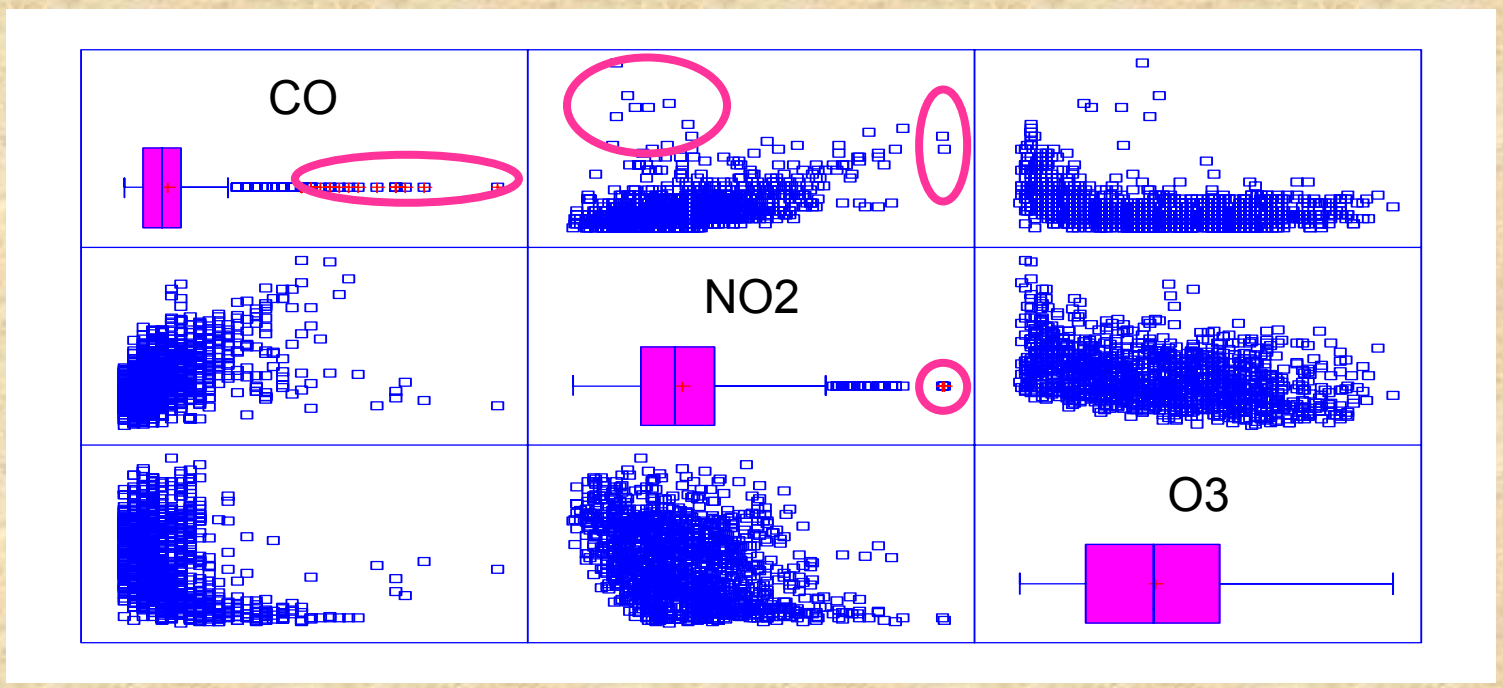
Si ordenamos los rectángulos según el valor de la primera y segunda componente obtenemos:



La 1ª ordenación coincide con la inducida por el área (describe el tamaño). El 2º componente relaciona la base con la altura y ordena en función de su forma.

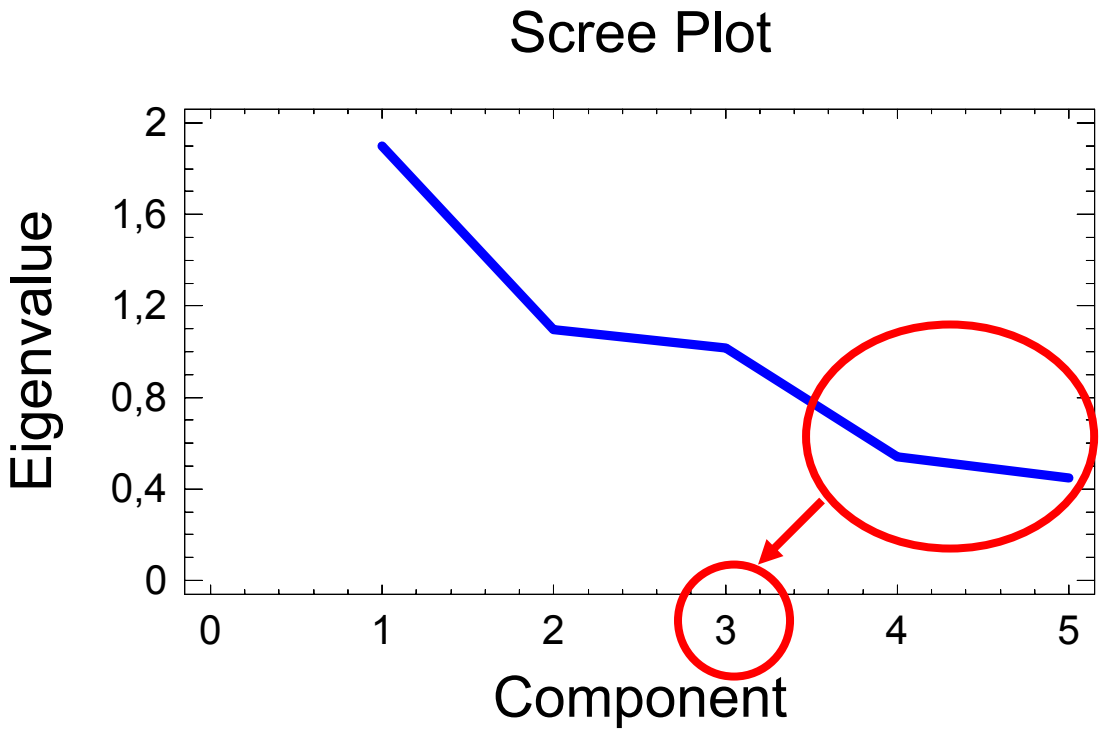
Ejemplo 2:

Disponemos de 1420 observaciones de datos de contaminación atmosférica en la provincia de Madrid. Se midió el ozono, el dióxido nitroso y el monóxido de carbono en diferentes meses y años. Su representación gráfica es la siguiente:



Ejemplo 2:

Después de la depuración y tipificación de los datos se escogen el nº de CP adecuado:



Se escoge el nº a partir del cual el autovalor disminuye muy poco

Ejemplo 2:

Number of components extracted: 5

Principal Components Analysis

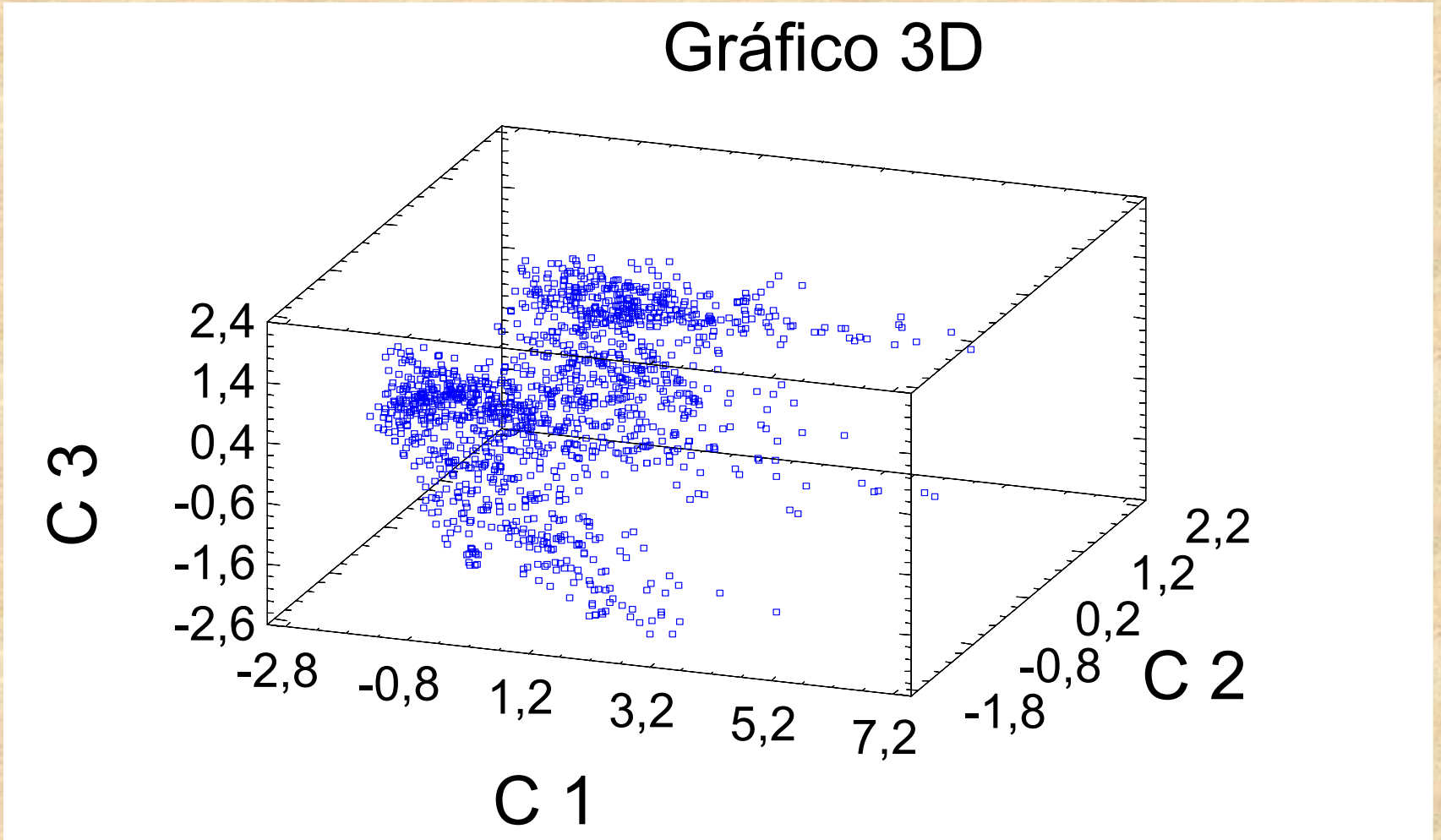
Comp. Nº	Eigenvalue	Percent of Variance	Cumulative Percentage
1	1,89739	37,948	37,948
2	1,09799	21,960	59,908
3	1,0163	20,326	80,234
4	0,539157	10,783	91,017
5	0,449163	8,983	100,000

Ejemplo 2:

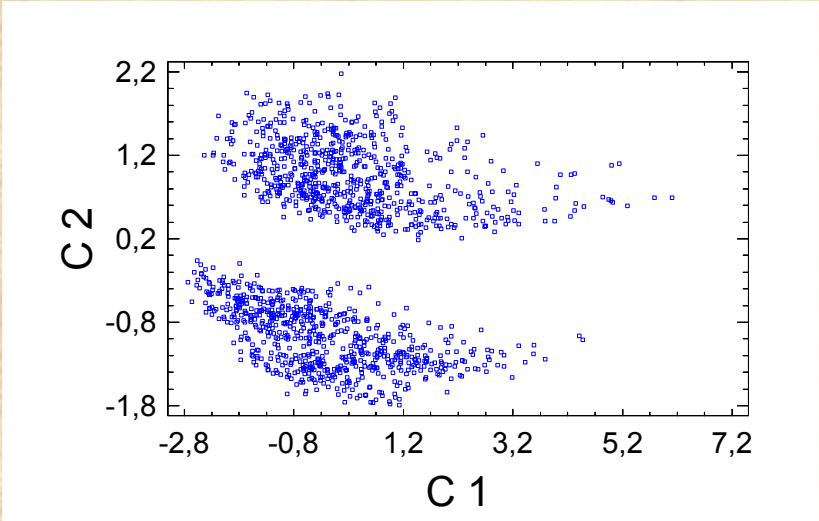
Tabla de pesos de los componentes

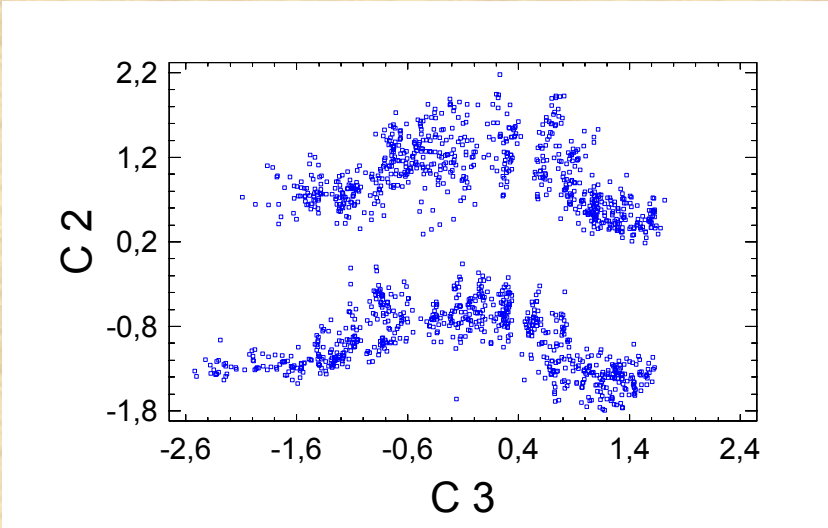
	Componente 1	Componente 2	Componente 3
CO	0,584737	0,0205001	-0,176301
NO2	0,604769	0,153802	0,100856
O3	-0,521156	0,422072	0,033207
mes	0,0490145	-0,104683	0,974821
año	0,13539	0,88703	0,0858305

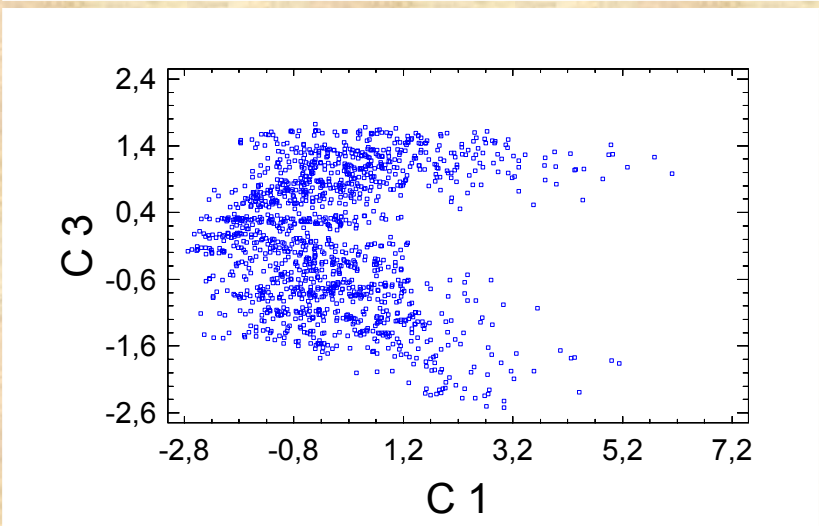
Ejemplo 2:



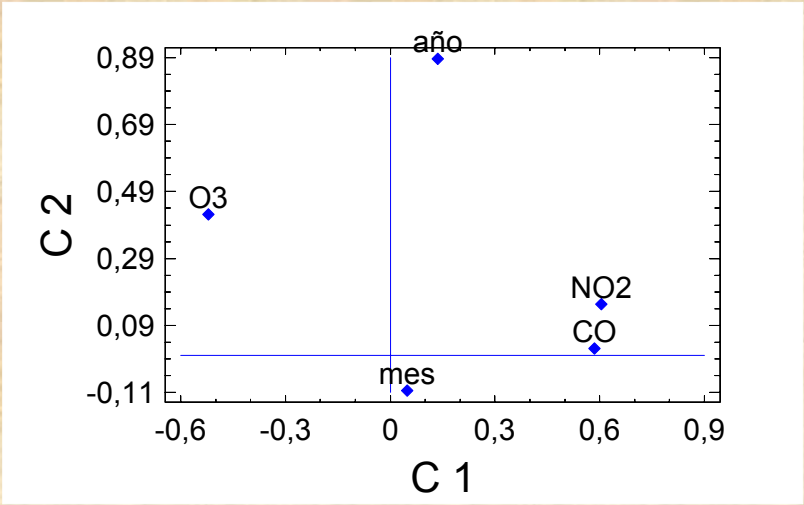
Ejemplo 2:

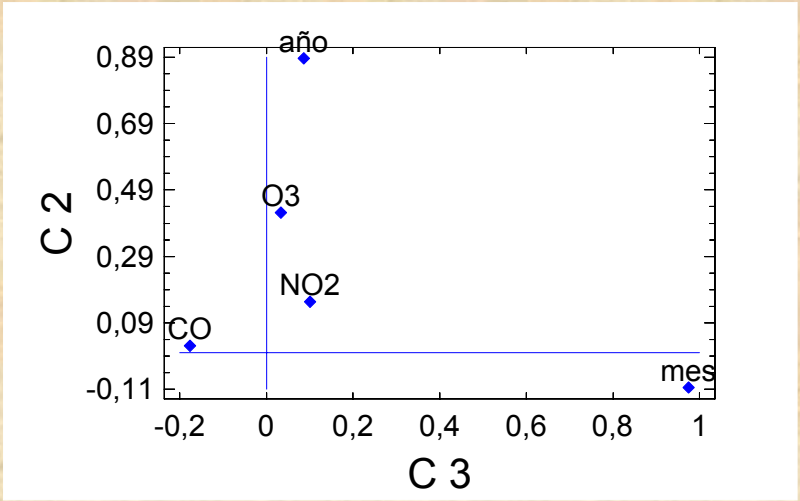

2,2
1,2
0,2
-0,8
-1,8
C 2
-2,8
-0,8
1,2
3,2
5,2
7,2
C 1

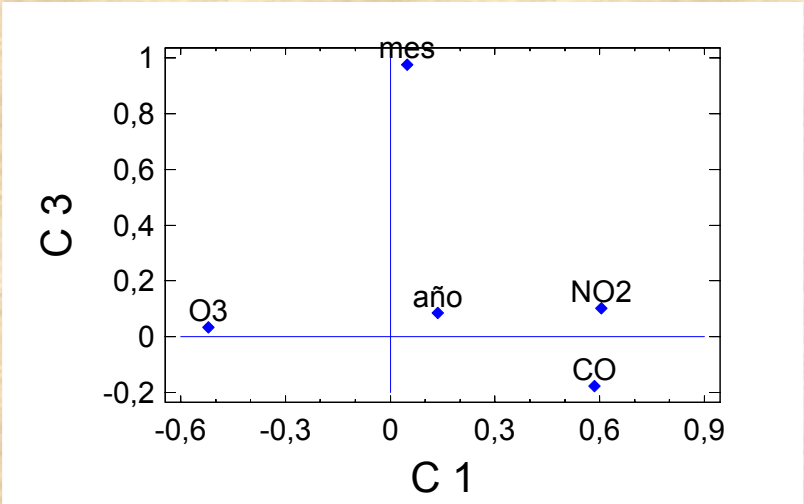

2,2
1,2
0,2
-0,8
-1,8
C 2
-2,6
-1,6
-0,6
0,4
1,4
2,4
C 3


2,4
1,4
0,4
-0,6
-1,6
-2,6
C 3
-2,8
-0,8
1,2
3,2
5,2
7,2
C 1

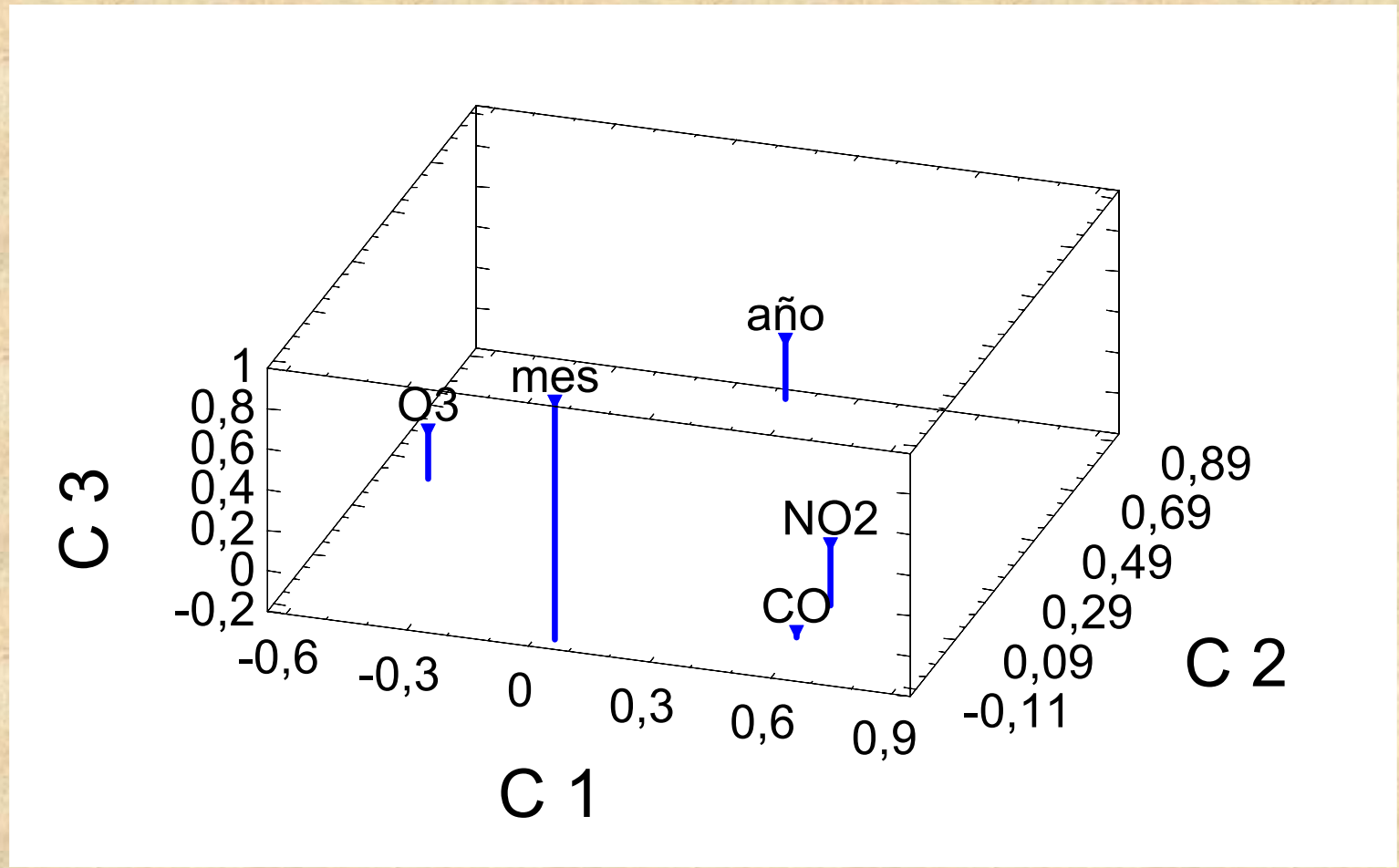
Ejemplo 2:


año
O3
NO2
CO
mes
0,89
0,69
0,49
0,29
0,09
-0,11
-0,6
-0,3
0
0,3
0,6
0,9
C 2
C 1

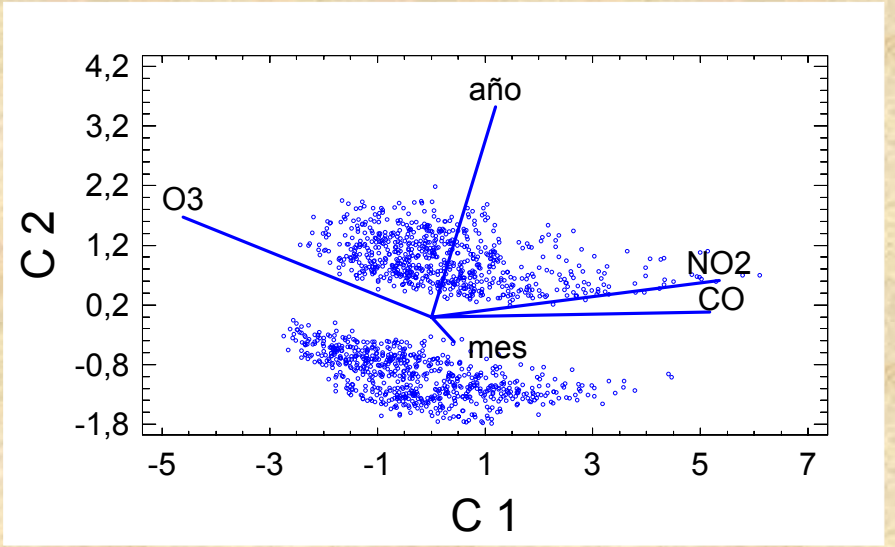

año
O3
NO2
CO
mes
0,89
0,69
0,49
0,29
0,09
-0,11
-0,2
0
0,2
0,4
0,6
0,8
1
C 2
C 3

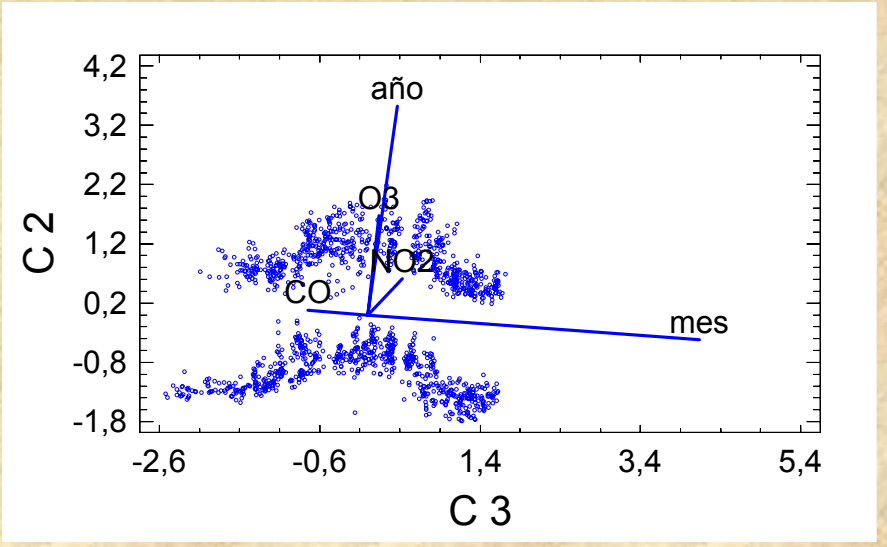

mes
O3
año
NO2
CO
1
0,8
0,6
0,4
0,2
0
-0,2
-0,6
-0,3
0
0,3
0,6
0,9
C 3
C 1

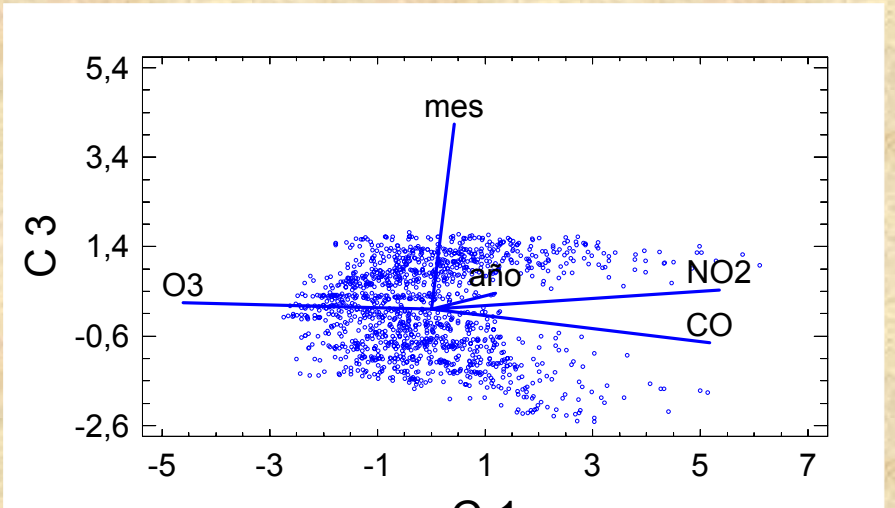
Ejemplo 2:


año
mes
O3
NO2
CO
C 3
1
0,8
0,6
0,4
0,2
0
-0,2
-0,6
-0,3
0
0,3
0,6
0,9
C 1
0,89
0,69
0,49
0,29
0,09
-0,11
C 2

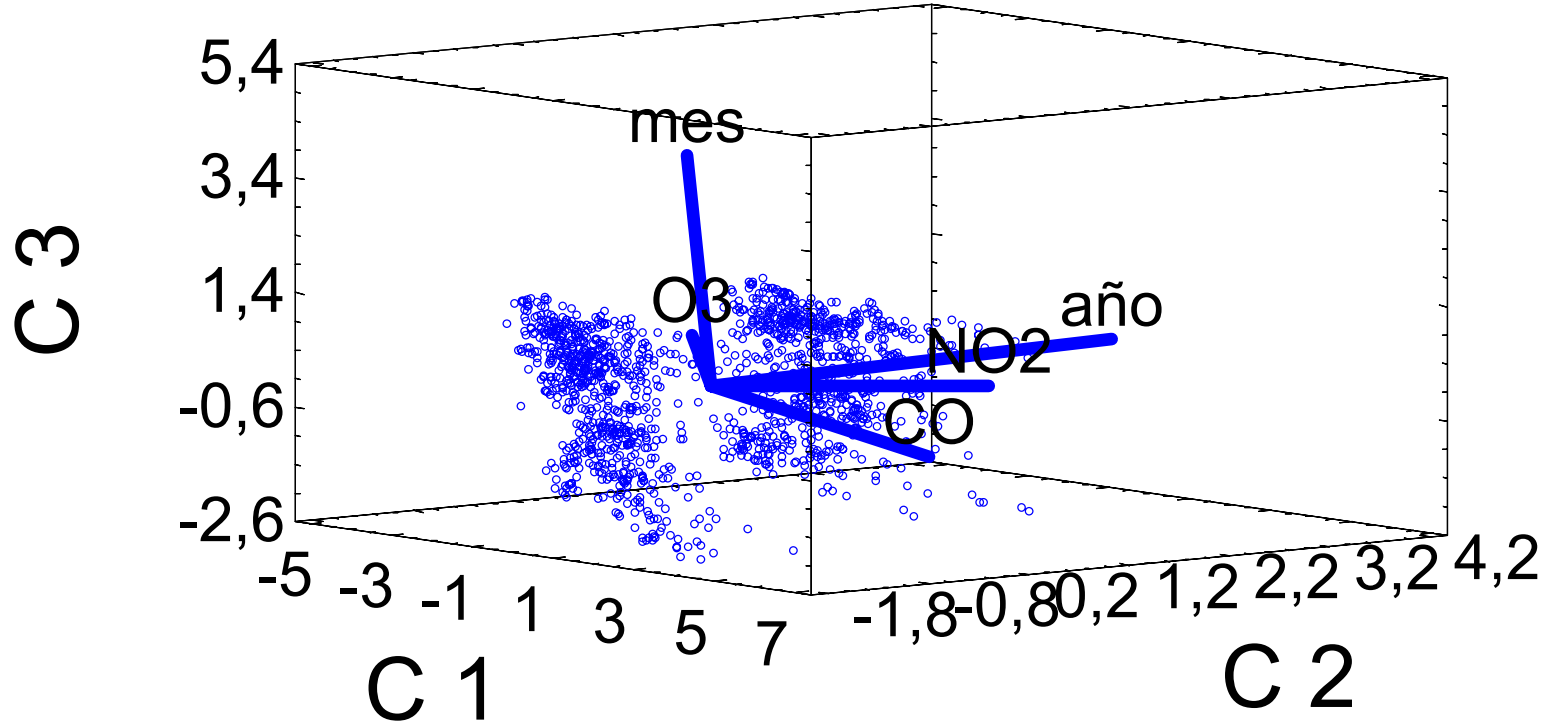
Ejemplo 2:


4,2
3,2
2,2
1,2
0,2
-0,8
-1,8
C 2
-5
-3
-1
1
3
5
7
C 1
año
O3
NO2
CO
mes


4,2
3,2
2,2
1,2
0,2
-0,8
-1,8
C 2
-2,6
-0,6
1,4
3,4
5,4
C 3
año
O3
NO2
CO
mes


5,4
3,4
1,4
-0,6
-2,6
C 3
-5
-3
-1
1
3
5
7
mes
O3
año
NO2
CO

Ejemplo 2:


C 3
5,4
3,4
1,4
-0,6
-2,6
mes
O3
año
NO2
CO
-5 -3 -1 1 3 5 7
C 1
-1,8 -0,8 0,2 1,2 2,2 3,2 4,2
C 2

Ejemplo 2:

La validación de los componentes principales se realiza con un análisis de componentes de la varianza.

Analysis of Variance for PCOMP_1

Source	SDC	Df	Mean Square	Var. Comp.	Percent
TOTAL (correg.)	2602,24	1419			
CO	1660,46	30	55,3488	1,27918	65,09
NO2	788,087	606	1,30047	0,487137	24,79
O3	153,362	761	0,201527	0,17423	8,87
mes	0,300948	18	0,0167194	0,0	0,00
año	0,0246149	1	0,0246149	0,0246149	1,25
ERROR	-3,63798E-12	3	-1,21266E-12	0,0	0,00

Ejemplo 2:

Analysis of Variance for PCOMP_2

Source	SDC	Df	Mean Square	Var. Comp.	Percent
TOTAL (correg.)	1563,43	1419			
CO	79,0473	30	2,63491	0,0239979	1,47
NO2	753,251	606	1,24299	0,0	0,00
O3	717,857	761	0,943308	0,0	0,00
mes	11,661	18	0,647832	0,0	0,00
año	1,61369	1	1,61369	1,61369	98,53
ERROR	9,09495E-13	3	3,03165E-13	3,03165E-13	0,00

El año es el factor que influye más en el 2º componente

Ejemplo 2:

Analysis of Variance for PCOMP_3

Source	SDC	Df	Mean Square	Var. Comp.	Percent
TOTAL (correg.)	1429,96	1419			
CO	81,2414	30	2,70805	0,0276561	1,77
NO2	615,977	606	1,01646	0,0	0,00
O3	704,668	761	0,925977	0,0	0,00
mes	28,0664	18	1,55924	1,52517	97,85
año	0,00583362	1	0,00583362	0,00583362	0,37
ERROR	1,59162E-12	3	5,30539E-13	5,30539E-13	0,00

El mes es el factor que influye en el 3er componente