

***Bases de datos,
descubrimiento de conocimiento,
análisis de datos:
primeros resultados del estudio de «El Niño»***

Sergio Camiz

Sapienza Università di Roma

www.camiz.net

sergio@camiz.net

Según la variabilidad de los datos almacenados, se habla de

- Bases de datos estáticas
de sólo lectura, para almacenar datos históricos y estudiarlos posteriormente.
Ejemplo: datos científicos.
- Bases de datos dinámicas
la información se modifica con el tiempo, t se emplea por consulta.
Ejemplo: el sistema informático de una compañía aérea.

Bases de datos

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

En este sentido, una biblioteca puede considerarse una base de datos, pero, debido al desarrollo tecnológico, la mayoría de las bases de datos están en formato digital.

Segundo su estructura y el programa de acceso se desarrollaron diferentes modelos:

- Bases de datos jerárquicas (años 1960)
- Bases de datos de red (años 1960)
- Bases de datos relacionales (1970)
- Bases de datos orientadas a objetos (años 1990)

Bases de datos jerárquicas

Cada nodo de información «hijo» se refiere a un nodo «padre», en donde esto puede tener varios hijos. La «navegación» puede hacerse revisando todas las «hojas» del árbol y subiendo hasta la «raíz», pero no al revés.

Ejemplo: de un pasajero se puede conocer el avión, pero no todos los que están en un avión.

Bases de datos relacionales

Se basan sobre el modelo *entidades-relaciones*: cada *entidad* es una tabla (*ejemplo:* «color de los ojos», «color del pelo») que contiene algunos atributos (los colores) y una *relación* entre tablas atara las filas de dos tablas (*ejemplo:* cada persona atara los colores de los ojos y del pelo). La interrogación se efectúa a través de consultas hechas en lenguaje *SQL*, resultando una tabla de datos rectangular.

Bases de datos de red

Los nodos están conectados entre sí por medio de enlaces en una red, así que un mismo nodo puede tener varios padres y se puede navegar en todas direcciones.

Ejemplo: de un pasajero se puede conocer el avión y de un avión los pasajeros. Asimismo todos los aviones que toma un pasajero.

Bases de datos orientadas a objetos

en estas se almacenen los objetos completos, o sea su estado y comportamiento, así que es posible de tratar objetos diferentes de manera adecuada.

Ejemplo: en una base de datos medicales el objeto «hombre» tiene atributos diferentes de un objeto «mujer»: por ejemplo, todos los datos relativos al sexo y a las gestaciones u los limites de valores de las análisis.

Minería de datos y descubrimiento de conocimiento

La *minería de datos* (DM, *Data Mining*) consiste en la extracción de información que reside de manera implícita en los datos, o sea *prepara, sondea y explora* los datos para sacar la información oculta en ellos.

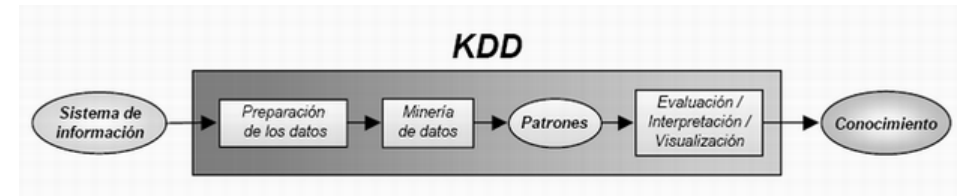
Este resultado se puede también definir como «*descubrimiento de conocimiento en bases de datos*» (KDD, *Knowledge Discovery in Databases*) y por esto objetivo se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos.

Análisis de datos

Bajo la denominación «*análisis de datos*» se engloba en estadística a un conjunto de métodos descriptivos multidimensionales finalizados a revelar estructuras y relaciones que se pueden encontrar extrayendolos de los datos mismos.

Desarrollada en Francia á partir de los años 1960, gracias a los trabajos de J.P. Benzécri y sus alumnos, la análisis de datos se distingue de la estadística porque no entente aplicar a los datos un modelo estadístico definido *a priori* ni tampoco sintetizarlos en un indicador de resumen (o sea una estadística), sino tiene como objetivo de llamar la atención sobre los datos mismos para revelar su contenido.

En realidad, *KDD* es un proceso que consta de un conjunto de fases, una de las cuales es la minería de datos, la fase que permite de encontrar patrones en los datos mismos.



Estas técnicas fueron desarrolladas principalmente en un contexto informático, porque esos fueron los que desarrollaron las bases de datos mismas, pero claro que se trata por la mayoría de técnicas de *análisis de datos y estadísticas*.

Física, Ecología y Ciencias Humanas

- *En física se emplearon modelos matemáticos desde Galilei:*
 - es posible construir experimentos;
 - los modelos son funcionales;
 - la matemática que se emplea puede ser muy compleja, para describir fenómenos muy difícil a comprender.
- *En ecología y ciencias humanas el empleo de modelos matemáticos es muy reciente:*
 - es difícil construir experimentos, pero se pueden sacar datos de campo o sino hacer encuestas;
 - los modelos describen tendencias, causadas a veces para muchos factores;
 - la matemática empleada debe tener en cuenta mucha incertitud.

La cantidad de datos que se encuentran en una investigación en ecología y en ciencias humanas es muy grande y los datos pueden ser heterogéneos.

Los objetivos pueden ser también heterogéneos.

Así es necesario un enfoque específico: —> *Análisis de Datos*.

Para su empleo, hay que estructurar en tres pasos las análisis y el estudio mismo, porque en cada paso hay diferentes objetivos, así que los métodos que se emplean en un paso no se pueden emplear en los otros.

Análisis exploratoria

En este paso es necesario de *estructurar* los datos segundo *ordenamiento* y *clasificación*, osea organizar-los de manera de identificar las fuentes de la diversidad y los grupos de datos homogéneos.

- Controle y primero estudio de los datos:
 - . Estadísticas descriptivas
- Búsqueda de relaciones y factores:
 - . Análisis factoriales (exploratorias)
- Búsqueda de estructuras:
 - . Clasificación (jerárquica)

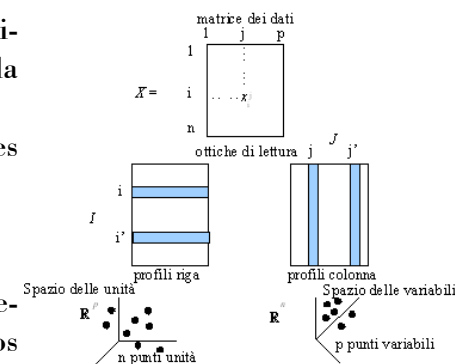
Al final se interpretan los resultados integrando factores y clases con lo que se conoce de los caracteres y las unidades.

- Las tres etapas de una investigación
 - *Exploratoria*:
 - . cuadro de referencia y objetivos del estudio, recogida de datos;
 - . búsqueda sobre existencia de estructuras y relaciones;
 - . formulación de hipótesis.
 - *Confirmatoria*:
 - . proyecto experimental;
 - . construcción de relaciones;
 - . test de hipótesis, inferencia estadística.
 - *Modelos*:
 - . formulación de un modelo matemático;
 - . implementación, calibración;
 - . simulación, previsiones.

Para aplicar estos métodos, se necesita que la información esté organizada en forma de tabla con:

- n filas que representan las unidades estadísticas, descritas por
- p caracteres o variables.

Así, en principio unidades y caracteres se pueden representar en dos espacios diferentes.



Análisis factoriales exploratorias

Permiten de reducir la dimensión de la tabla de datos de manera importante, pero reteniendo la mayor parte de la información. Además, representando caracteres y unidades en espacios geométricos, permite de visualizar las relaciones entre objetos de manera óptima.

La interpretación de estos ejes factoriales permite poner en evidencia la forma de las interrelaciones entre las variables estudiadas, y las semejanzas y diferencias entre los individuos con respecto a esas variables.

Como la inercia en respecto a la origen vale

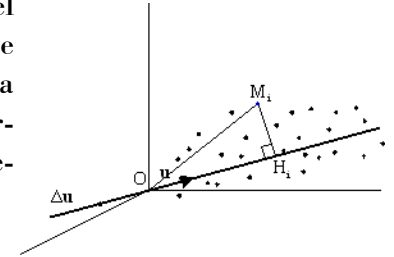
$$\sum_{i=1}^n p_i (OM_i)^2 = \sum_{i=1}^n p_i (M_i H_i)^2 + \sum_{i=1}^n p_i (OH_i)^2$$

como dirección mas importante ay que buscar la dirección que minimiza, en el segundo miembro, la primera cantidad o que maximiza la segunda.

Análisis de componentes principales

(Langrand y Pinzón, 2009; Jolliffe, 1986; Legendre and Legendre, 1998; Benzécri et al., 1973-82)

Se imaginan las unidades representadas en el espacio R^p generado para las caracteres, que forman una base ortogonal y se busca una nueva base ortogonal que maximiza la inercia de los datos proyectada sobre los “primeros” vectores.



En general, si X es la tabla de datos centrados (promedio = 0), N es la matriz diagonal de los pesos de las unidades ($\sum_{i=1}^n p_i = 1$) y M la matriz diagonal de la métrica del espacio (en ACP, $M = \text{diag}(1/\sigma_i^2)$), la inercia en respecto de una recta de versor \mathbf{u} es:

$$I_{r^{\perp}} = \|\mathbf{c}\|_N^2 = \mathbf{c}' N \mathbf{c} = (\mathbf{X} M \mathbf{u})' N (\mathbf{X} M \mathbf{u}) = \mathbf{u}' M \mathbf{X} N \mathbf{X} M \mathbf{u}$$

Para encontrar la dirección de inercia máxima, ay que resolver el problema de maximización

$$\begin{cases} \mathbf{u}'\mathbf{M}\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M}\mathbf{u} = \text{Max}_{\mathbf{u}} \\ \mathbf{u}'\mathbf{M}\mathbf{u} = 1 \end{cases}$$

y como la Lagrangiana

$$\mathcal{L} = \mathbf{u}'\mathbf{M}\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M}\mathbf{u} - \lambda (\mathbf{u}'\mathbf{M}\mathbf{u} - 1)$$

su solución es

$$\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M}\mathbf{u} = \lambda \mathbf{u}$$

con λ máximo valor propios de $\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M}$, y \mathbf{u} su vector propio correspondiente. λ es la variancia de los datos sobre la dirección \mathbf{u} .

Se puede demostrar que la matriz de los datos se puede reconstruir para

$$\mathbf{X}_{(n,p)} = \sum_{\alpha=1}^p \sqrt{\lambda_{\alpha}} \mathbf{v}_{\alpha} \mathbf{u}'_{\alpha}$$

Así si se ordenan valores y vectores propios en orden decreciente, ay que:

Teorema de Eckart y Young: la reconstrucción de la matriz hecha para los primeros $r < p$ valores y vectores propios

$$\mathbf{X}_{(n,r)} = \sum_{\alpha=1}^r \sqrt{\lambda_{\alpha}} \mathbf{v}_{\alpha} \mathbf{u}'_{\alpha}$$

es la mejor de rango r .

En general, para la descomposición a los valores propios

$$\mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

y se ordenan valores y vectores propios en orden decreciente de λ .

Resulta también que

$$\mathbf{N}\mathbf{X}\mathbf{M}\mathbf{X}' = \mathbf{X}\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\mathbf{X}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$$

así que los mismos valores propios sirven para la análisis en el espacio \mathbb{R}^n

Clasificaciones exploratorias

Permiten elaborar tipologías y agrupar individuos por clases en función de sus semejanzas con respecto al conjunto de las variables. Un criterio empleado a menudo desde el punto de vista técnico es el de buscar la clasificación que minimiza la varianza intraclase (variabilidad entre los individuos de una misma clase), y maximiza la varianza interclase (variabilidad entre las clases).

Clasificación jerárquica

(Langrand y Pinzón, 2009; Anderberg 1973; Benzécri et al., 1973-82; Legendre and Legendre, 1998; Gordon, 1999)

La construcción de un *dendrograma* permite de conocer el junto de relaciones entre los objetos que ay a clasificar a través de una taxonomía.

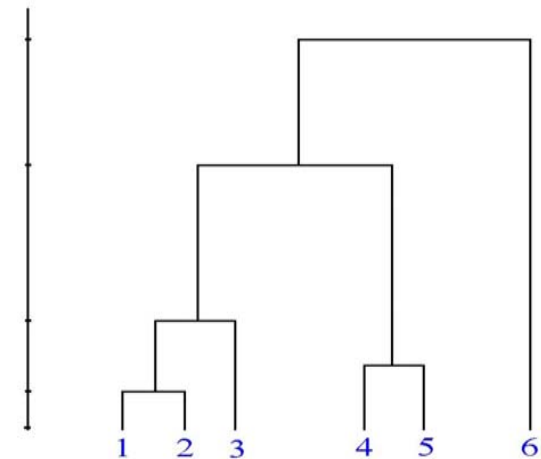
Cortando el dendrograma se consiguen las particiones que necesitan.

- Clasificación jerárquica ascendente

Para construir un dendrograma, ay que

- (1) decidir un criterio para asociar los objetos;
- (2) elegir un indice de asociación;
- (3) elegir un criterio para asociar objetos y grupos (función objetivo) a cada paso;
- (4) buscar los objetos y los grupos que agregando-se optimizan la función objetivo);
- (5) elegir un criterio para volver a calcular el indice entre objetos y grupos ya creados.

- Dendrograma



- Algoritmo iterativo

- (6) Al principio cada objeto es un grupo (singleton) y se construyen las relaciones de asociación entre objetos.
- (7) A cada paso:
 - (8) se elije la pareja de grupos que uniéndose optimizan la función objetivo;
 - (9) los dos grupos se combinan en un nuevo grupo;
 - (10) se calcula la asociación del nuevo grupo con los otros;
 - (11) se repite esta operación $n-1$ veces.
- (12) El proceso termina cuando no ay que un grupo solo.

Funciones objetivo para unidades:

- *Conexión completa*: la asociación entre dos grupos es la peor entre unidades de ambo grupos;
- *Conexión singla*: la asociación entre dos grupos es la mejor entre unidades de ambo grupos;
- *Conexión promedia*: la asociación entre dos grupos es el promedio entre unidades de ambo grupos;
- *Ward*: la asociación entre dos grupos es su variancia dentro de los grupos.

Como resultado a cada paso se encuentra una representación de todas las variables del grupo sobre un plan factorial, así que todas las unidades como representadas solamente para las variables del grupo.

Es interesante observar que cada grupo se presenta en forma de *dipolo*, porque el signo de la correlación entre variables no es importante sino su intensidad. Esto facilita la interpretación de las variables características, así que la distribución de las unidades.

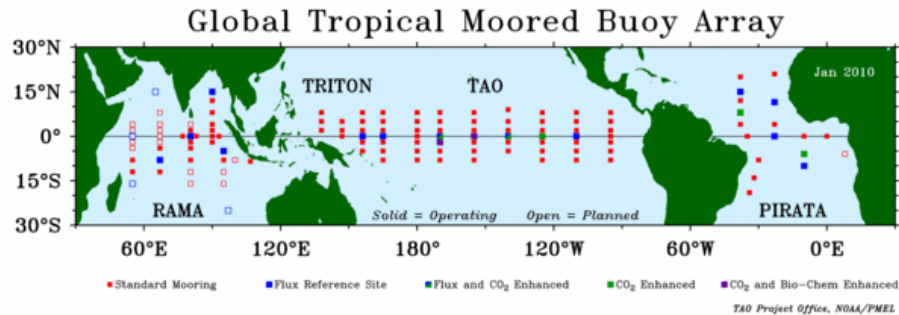
Clasificación factorial jerárquica para variables
(Denimal, 2007; Camiz et al., 2006; Camiz and Pillar, 2007)

- 1) cada variable estandarizada (promedio cero y variancia 1) forma un grupo y es representativa de su misma;
- 2) a cada paso
 - 3) se hace la *ACP* non-estandarizada de las variables representativas da cada pareja de grupos
 - 4) se elije la pareja por la cual el segundo valor-propio es el más bajo
 - 5) se juntan los dos grupos de variables
 - 6) se toma el primero vector-propio como variable representativa del nuevo grupo formado
 - 7) se toma el segundo valor-propio come índice de la jerarquía.

El Niño

«*El Niño*» y «*La Niña*» son parte del ciclo climático conocido internacionalmente com «El Niño Southern Oscillation» (*ENSO*). Durante El Niño en la zona Ecuatorial central y oriental del Océano Pacífico la temperatura superficial resulta más alta del promedio, mientras durante La Niña la temperatura resulta más baja. Consecuencias se encuentran en larga parte de América do Sur, particularmente en Peru. Por esto, la observación de las condiciones del Pacífico tropical se considera esencial para una previsión de breve término.

Como primera exploración, hemos encontrado la red de medidas implementada para el National Oceanic and Atmospheric Norte-Americano (NOAA). Se trata de un conjunto de bojas onde se median temperaturas, corrientes y vientos y se transmiten en tiempo real cada día a los investigadores.

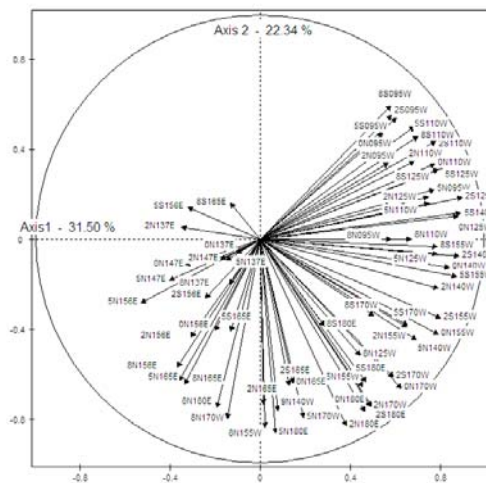


De su sitio web hemos bajado 88 series-temporales grabadas cada día desde el 1 de Marzo 1980 hasta el 31 Diciembre 2008. En realidad no resultan datos de 20 sitios y solo 27 estaban funcionando antes 1991, así que la nuestra tabla de datos esta limitada a 68 series diarias entre 1991 y 2008.

Sobre esta tabla de datos se hicieron una Análisis de Componentes Principales y una Clasificación Factorial Jerárquica. Sobre los planes factoriales se pudieron proyectar los años y los meses como modalidades ilustrativas para facilitar la interpretación de los resultados.

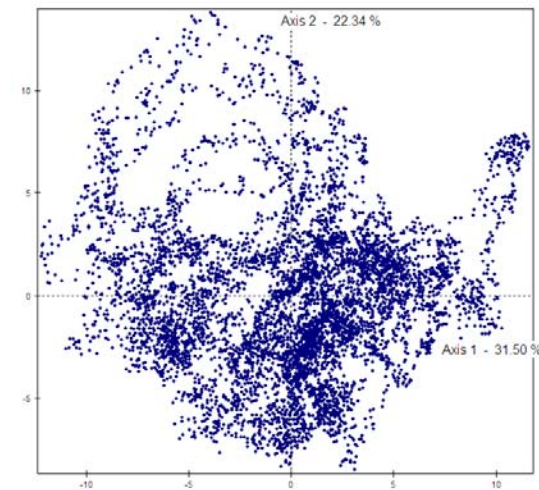
Análisis de las componentes principales - 68 series diarias 1991-2008

Representación de las series sobre el primero plan factorial

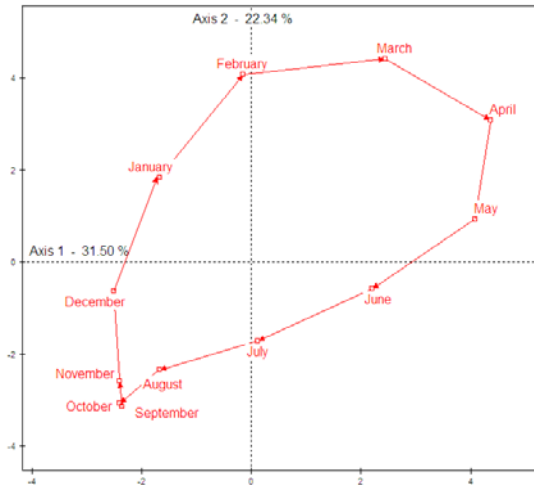


Análisis de las componentes principales - 68 series diarias 1991-2008

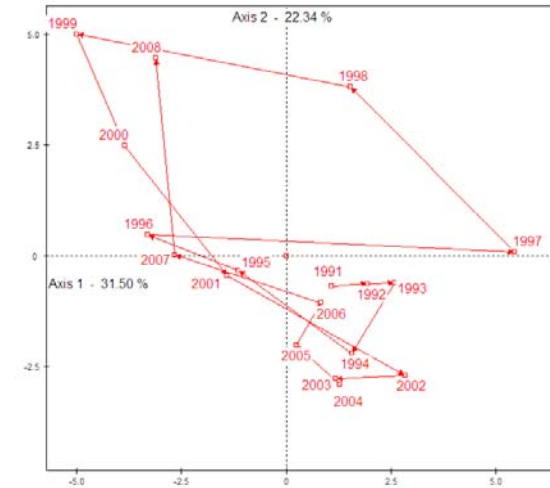
Representación de los días sobre el primero plan factorial



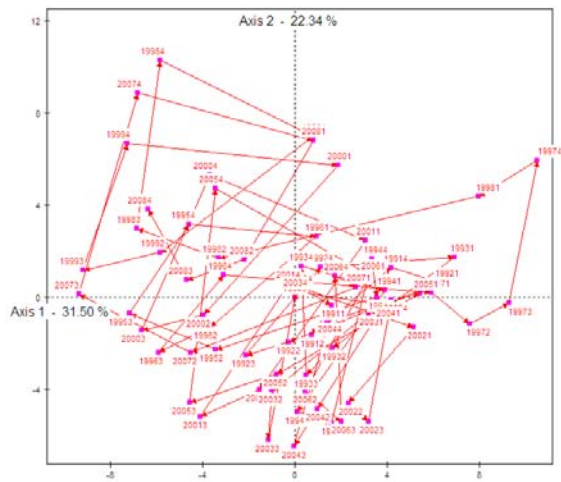
Análisis de las componentes principales - 68 series diarias 1991-2008
Representación de los meses sobre el primero plan factorial



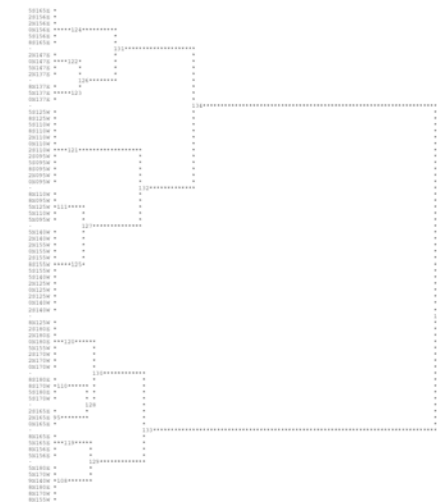
Análisis de las componentes principales - 68 series diarias 1991-2008
Representación de los años sobre el primero plan factorial



Análisis de las componentes principales - 68 series diarias 1991-2008
Representación de las temporadas sobre el primero plan factorial

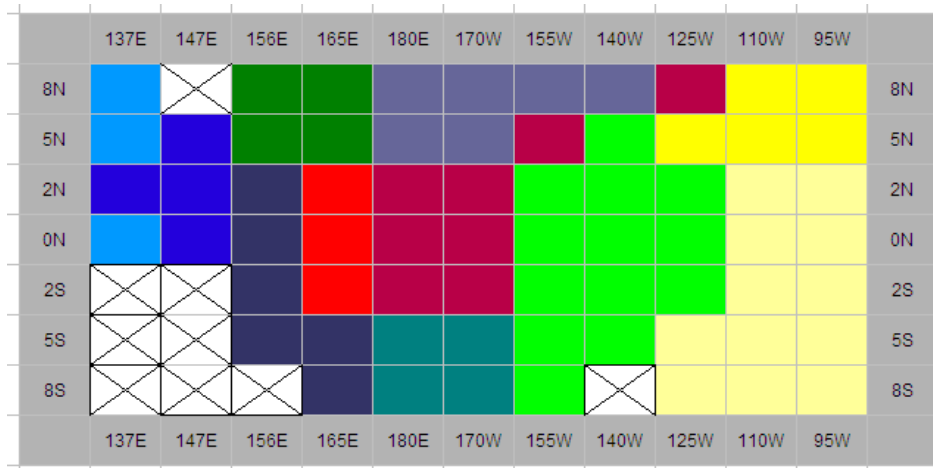


Clasificación factorial jerárquica - 68 series diarias 1991-2008
Dendrograma de la jerarquía limitado a los 10 nodos arriba



Clasificación factorial jerárquica - 68 series diarias 1991-2008

Partición de las series (sitios) en 11 classes



Bibliografía

- Anderberg M.R., 1973. *Cluster Analysis for Applications*. New York, Academic Press.
- Benzécri J.P., 1973-82. *L'analyse des données*. 2 vols., Paris, Dunod.
- Camiz S., 2001. «Exploratory 2- and 3-way Data Analysis and Applications». *Lecture Notes of TICMI*, Tbilisi University Press, vol. 2. <http://www.emis.de/journals/TICMI/lnt/vol2/lecture.htm>.
- Camiz S., A. Altieri, and F. Manes, 2008. «Pollution Bioindicators: Statistical Analysis Of A Case Study». *Water Air and Soil Pollution*, 194(1-4): pp. 111-139.
- Denimal J.J., 2007. «Classification factorielle optimisée d'un tableau de mesures». *Journal de la Société Française de Statistique - Revue de Statistique Appliquée*, 148.

Conclusión

- Los métodos de análisis exploratoria permiten un examen rápido de un conjunto muy grande de datos, ofreciendo herramienta gráfica de fácil comprensión.
- Algunos aspectos de El Niño ya se encontraron.
- Para una comprensión más profundizada y para previsiones precisan métodos de análisis confirmatoria.

- Jolliffe I.T., 1986. *Principal Components Analysis*. Berlin, Springer.
- Gordon A.D., 1999. *Classification*. London, Chapman and Hall.
- Langrand C. and L.M. Pinzón, 2009. *Análisis De Datos. Métodos y ejemplos*, Bogota, Escuela Colombiana de Ingeniería Julio Garavito.
- Legendre P. and L. Legendre, 1998. *Numerical Ecology*. 2nd Ed., Amsterdam, Elsevier.

además:

http://es.wikipedia.org/wiki/Base_de_datos

http://www.hypergeo.eu/article.php3?id_article=153

<http://www.monografias.com/trabajos72/base-datos/base-datos2.shtml>

<http://www.monografias.com/trabajos55/mineria-de-datos/mineria-de-datos.shtml>

<http://www.pmel.noaa.gov/tao/>