UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

DESARROLLO DE UN SISTEMA PARA MINERÍA DE DATOS BASADO EN LOS MÉTODOS BIPLOT

Valter Martins Vairinhos

2003

Dpto. de Estadística
Universidad de Salamanca

Mª PURIFICACIÓN GALINDO VILLARDON

Profesor Titular del Area de Estadística e I.O.
de la Universidad de Salamanca

CERTIFICA: **Que** D. Valter Martins Vairinhos**, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: "Desarrollo de un sistema para minería de datos basado en los métodos Biplot"; y para que conste, firma el presente certificado en Salamanca. 20 de Mayo 2003.**

DESARROLLO DE UN SISTEMA PARA MINERÍA DE DATOS BASADO EN LOS MÉTODOS BIPLOT.

Memoria que para optar al Grado de Doctor por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Valter Martins Vairinhos

Salamanca
2003

Agradecimientos

Al Profesor Rui Agonia Pereira, por haberme apuntado el camino de Salamanca: en buena hora lo hizo.

A la Dra. Purificación Galindo por haber aceptado dirigir el proyecto, por su constante apoyo, por sus enseñanzas, por su firme orientación y por el ambiente humano que supo crear en el Departamento de Estadística.

A los miembros del Departamento de Estadística y compañeros de doctorado por su apoyo constante.

A mi mujer y mi hijo, Irene y Carlos, sin cuyo apoyo, estímulo y sacrificio esta tarea muy difícilmente llegaría al final.

ÍNDICE

ÍNDICE.

INTRODUCCIÓN

INTRODUCCIÓN.

Al terminar la investigación que ahora se presenta y al mirar hacia atrás buscando identificar su génesis y su motivación inicial, el inicio del camino, creo poder atribuirlo a un *"fait divers"* ocurrido hace años.

Intentábamos, otros compañeros y yo, introducir en nuestra organización métodos objetivos de decisión en la actividad de definición de los contenidos funcionales de puestos de trabajo, usando métodos de análisis multivariante y el análisis cluster en particular.

Obtenido el mejor *software* estadístico entonces disponible, llegó el día de presentarle al jefe los resultados del primer estudio.

Se verificó que el volumen - en hojas de papel - de los resultados era muy superior al volumen de los datos... Peor que eso, el hecho de que el lenguaje de presentación de los resultados era esencialmente estadístico: no hubo el cuidado de transformar esos resultados en el lenguaje natural de quien debería decidir sobre el sistema real con base en ellos...

El resultado final fue que el estudio quedó inutilizado y ese incidente creó una actitud negativa en contra del proyecto que fue abandonado.

Al razonar sobre este incidente quedó clara en mi mente la necesidad de desarrollar *software* estadístico que no solo fuera capaz de tratar datos voluminosos sino también de presentar los resultados de forma adecuada a los procesos de decisión del usuario final, sin exigir que éste conociera los métodos estadísticos para poder entender los resultados y decidir sobre ellos; era necesario realizar una investigación, centrada en las cuestiones de

interpretación de los resultados obtenidos por análisis de datos multivariantes y transformar los resultados de esa investigación en *software* adecuado.

Iniciado el camino, fue fácil verificar la casi ausencia de trabajo en este dominio en las fronteras de la estadística con la informática y con la psicología cognitiva.

No solo era poca la literatura al respecto sino que además resultaba difícil interesar a los departamentos de estadística en un tema un poco excéntrico, para algunos mal definido y para muchos con pequeña probabilidad de éxito...

El paradigma en estadística es el de que la interpretación de los resultados es una tarea del investigador del dominio al que pertenecen los datos: a la estadística cabe generar los resultados, eventualmente de una forma que permita a los investigadores leerlos fácilmente, pero siempre en el presupuesto de que es necesario que el investigador conozca profundamente no solo los métodos sino también los datos que desea analizar.

Mientras tanto ocurría la revolución conocida como *data mining* (minería de datos) asociada a la necesidad de transformar los datos contenidos en grandes bases de datos en información utilizable. En minería de datos, una cuestión importante es, precisamente, la búsqueda de algoritmos que faciliten la interpretación de la gran cantidad de resultados generados por los análisis.

El problema de interpretar los resultados puede ser, en sí mismo, un nuevo y desafiante problema de análisis de datos. HASTIE *et al* (2001) comentan el hecho de que un algoritmo para detectar reglas de asociación encontró en un paquete de 6876 observaciones por 50 variables, 6288 reglas de asociación…

La oportunidad de transformar estas inquietudes en un proyecto de investigación la he encontrado en el Departamento de Estadística de la Universidad de Salamanca en la persona de su Directora, que ha aceptado dirigir el proyecto.

El hilo conductor de toda la investigación cuyos resultados ahora se presentan es, por lo tanto, la preocupación por los problemas de interpretación y de síntesis de resultados obtenidos en los procesos de análisis de datos multivariantes/minería de datos.

Esta preocupación se ha reflejado en la elección de la técnica de los biplots como núcleo de un sistema de minería de datos, considerando no sólo sus propiedades matemáticas, su generalidad como instrumento de análisis, su papel central en análisis multivariante y la facilidad con la que permite al ser humano **ver** en un plano las interrelaciones entre los actores principales de todo análisis de datos multivariantes: las variables y los individuos.

En esa línea de preocupación con el tema de la interpretación, se define, informalmente, en este trabajo, ***el proceso de análisis de datos*** *como un proceso que tiene como finalidad el descubrir el significado de los datos.*

También informalmente, se define ***interpretación*** *de los resultados de los análisis como la expresión de esos resultados en un lenguaje cercano al usado por los seres humanos sin grandes conocimientos de estadística.*

El contenido de esta tesis puede, por lo tanto, sintetizarse como un intento de obtener respuestas para las cuatro cuestiones siguientes:

Cuestión nº 1

¿Es posible y, siendo posible, es deseable centrar en los métodos de biplot un sistema de minería de datos?

Cuestión nº 2

¿Cómo formular los problemas de interpretación de resultados asociados a un sistema de minería de datos basado en biplots?

Cuestión nº 3

¿Qué problemas ocurren cuando se intenta centrar en el concepto de biplot un sistema de minería de datos? ¿Cuáles son las soluciones para esos problemas?

Cuestión nº 4

¿Cómo realizar minería de datos centrado en biplots?

El contenido de los distintos capítulos está, por lo tanto, relacionado con las respuestas encontradas para estas cuatro cuestiones.

Los **capítulos 1, 2 y 3** son metodológicos: si la tesis tiene que ver con el problema de construir un sistema de minería de datos basado en biplots,

entendemos que se debe empezar por definir esos conceptos y presentar sus propiedades.

Así, en el **capítulo 1** se busca definir lo que es la actividad de minería de datos y se revisan las relaciones de la minería de datos con las disciplinas de estadística matemática, análisis de datos multivariantes y la tecnología de las bases de datos.

En los **capítulos 2 y 3** se revisa el concepto de biplot buscando explicitar y dar relevancia a aquellas características y resultados – muchos de ellos obtenidos en el Departamento de Estadística de la Universidad de Salamanca - que justifican su elección para actividades de minería de datos, buscando, en síntesis, una respuesta para la **cuestión nº 1**: ¿es posible centrar en los métodos biplot un sistema de minería de datos?

En el **capítulo 4** se buscan respuestas para la **cuestión número 2** y se presenta un intento de teorización de los problemas de interpretación de resultados de análisis de datos multivariantes - incluidos los métodos de biplot. Este capítulo contiene una parte esencial de las contribuciones específicas de esta tesis.

En el **capítulo 5** se intenta contestar a las **cuestiones 3 y 4**: ¿cómo realizar la minería de datos basada en biplots?

Teniendo en cuenta que lo que se busca en minería de datos es descubrir patrones interpretables en los datos almacenados en grandes bases de datos, en este capítulo son analizados los problemas prácticos que hay que resolver y se presentan los principales algoritmos usados en la creación del sistema de minería de datos propuesto.

En este capítulo ocurren también algunas innovaciones de las cuales se destaca la explotación de los biplots basada en los conceptos de operación básica de interpretación (definición, caracterización y comparación interactiva de grupos) y en el concepto de proyección.

El **capítulo 6** está destinado a presentar lo que se considera el culminar de toda la investigación y su contribución principal: el prototipo de un sistema de minería de datos basado en biplots.

Finalmente, en el **capítulo 7** se presentan los resultados de la aplicación del sistema creado al estudio de datos reales.

Se consideran como principales contribuciones de esta tesis las siguientes:

1. La idea de basar en las técnicas de biplot un sistema de minería de datos.
2. La idea de que la interpretación es una fase autónoma del proceso de análisis de datos, con problemas específicos susceptibles de formalización.
3. Una formulación general del problema de interpretación de resultados de análisis de datos multivariantes.
4. Una formulación del problema de interpretación de resultados usando la teoría de los conjuntos *imprecisos* (*rought sets*) de PAWLAK (1991, 1998).
5. Una metodología, independiente de los métodos de análisis, para interpretar los resultados de los análisis de datos multivariantes.
6. Relevancia de la teoría de los grafos de intersección como base teórica para la formulación de los problemas de interpretación.
7. Una medida de afinidad entre átomos de grafos de intersección.

8. El prototipo de un sistema de minería de datos basado en biplots.

De estas contribuciones se destaca, en el **Capítulo IV**, la identificación y formulación general del problema de interpretación de resultados de métodos de análisis de datos multivariantes y su representación matemática sobre un grafo de intersección, cuestiones nuevas, no tratadas en la literatura actual.

C A P Í T U L O I

MINERÍA DE DATOS

1.1. INTRODUCCIÓN.

Siendo el objetivo de esta tesis la investigación de los problemas relacionados con el desarrollo de un sistema de minería de datos basado en biplots, empezamos por definir - en el apartado 1.2. - lo que entendemos por Minería de Datos (MD).

La actividad de MD está muy conectada a los problemas de explotación de bases de datos: por eso, presentamos en el apartado 1.3. una referencia al lenguaje básico usado hoy día por los sistemas de base de datos más frecuentes.

Los objetivos de la MD son muy similares a los objetivos de la actividad estadística designada por Análisis de Datos Multivariantes o, simplemente, Análisis de Datos. En el apartado 1.4. se busca encontrar los atributos distintivos de la Minería de Datos con relación a la Estadística.

En el apartado 1.5. se busca identificar y caracterizar las tareas básicas de minería de datos, situando la técnica de los biplots en este contexto.

1.2. DEFINICIÓN DE MINERÍA DE DATOS.

La actividad «Minería de Datos» como disciplina autónoma es reciente: FAYYAD (1997), en el editorial del primer número de la revista *Data Mining and Knowledge Discovery*, registra que la primera reunión internacional sobre este tema se realizó en Detroit en 1989.

En general, se admite que las actividades de minería de datos integran un proceso más general conocido universalmente por la designación anglosajona de KDD - *Knowledge Discovery in Data Bases* que, de ahora en adelante usaremos. FAYYAD *et al* (1996); BORGELT *et al* (2001).

Una definición en general aceptada y citada de KDD es la que ha formulado FAYYAD *et al* (1996), aquí transcrita:

> «*El descubrimento de conocimientos en bases de datos es el proceso* ***no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y comprensibles en los datos***».

En la referencia citada se precisa que:

Los **datos** a que se refiere esa definición son hechos registrados en una base de datos.

Los **patrones** son expresiones de un lenguaje específico, usados para designar subconjuntos de los hechos registrados en la base de datos. Esto significa que estas expresiones dependen de la técnica o método usado para analizar los datos. Por ejemplo, si la técnica es la clasificación aglomerativa, los patrones pueden ser grupos (*clusters*) o particiones.

La minería de datos es un **proceso** que incluye varias fases o pasos. Esto significa que es algo dinámico que puede envolver varias interacciones eventualmente convergentes para un resultado final.

Los patrones deben ser **válidos** en el sentido de que deben trascender los datos usados en su descubrimiento y tener significado para otros datos o poder ser usados en procesos de decisión.

Los patrones deben ser **novedosos** en el sentido de que deben ser inesperados, distintos de los valores esperados si se tiene en consideración la información más antigua. Al revés de lo que ocurre en estadística, interesa identificar valores lejos de los valores medios, considerados potencialmente interesantes.

Los patrones deben ser **potencialmente útiles** en el sentido de que deben poder ser usados para tomar decisiones.

Los patrones deben ser «**comprensibles**» por los seres humanos - lo que significa que, en primer lugar, deben ser **simples** e **interpretables**, con expresión fácil en el lenguaje informal e intuitivo de los seres humanos.

El término **interesante** en la definición significa una combinación de validez, novedad, utilidad y simplicidad y está conectada con el valor del patrón descubierto. Para una formulación matemática general de este concepto, puede verse SZYMON *et al* (2001).

El proceso *KDD*, tal como es definido por FAYYAD *et al* (1996), está formado por 8 fases o pasos:

1. Comprender el dominio de aplicación y los objetivos del usuario final.
2. Crear el paquete o conjunto de datos por analizar.
3. Limpiar y preprocesar los datos.
4. Aplicar técnicas de reducción de la dimensionalidad de los datos.
5. Elegir la función principal a realizar: buscar clasificaciones, predicción, descripción u otras.
6. Elegir los algoritmos y modelos para realizar la minería de datos.
7. **Realizar minería de datos - búsqueda de los patrones**.
8. Interpretar los patrones descubiertos, repitiendo los pasos 1 - 7 si necesario.
9. Incorporar los resultados del proceso en el sistema de decisión o, simplemente, presentar los resultados.

Como se verifica por esta descripción, la Minería de Datos es una de las ocho fases del KDD. Sin embargo, dada su importancia relativa, muchos autores - HASTIE *et al* (2001); BORGELT *et al* (2001) - identifican KDD y Data Mining.

Otra definición – HOSKING *et al* (1997) - es la de que «*Minería de datos es un conjunto de métodos para obtener inferencias a partir de los datos*». Esta definición significaría que la Minería de Datos y la Estadística tendrían los mismos objetivos siendo distinta la metodología.

En nuestro trabajo nos adherimos a la definición de Fayyad y mantendremos la distinción entre Minería de Datos y KDD, reconociendo

que todo el proceso de KDD depende crucialmente de la fase de Minería de Datos.

1.3. DATOS, BASES DE DATOS Y MINERÍA DE DATOS.

Datos son hechos que han sido registrados en resultado de experiencias planeadas con un objetivo científico (***datos experimentales***) o por conveniencias operacionales, en resultado del funcionamiento de un sistema (***datos observacionales***).

En sentido lógico, los datos representan proposiciones con el valor lógico de **verdad**. La ocurrencia de un dato equivale a la afirmación de que es verdad que algo ha ocurrido; que un aparato o una persona ha registrado y observado, en un lugar y tiempo específicos, un determinado valor; y que eso es cierto.

Hoy día, las organizaciones registran rutinariamente casi todo lo que ocurre: los flujos de dinero, los hechos relativos a la vida de las personas, los flujos de materiales y de decisiones, imágenes, sonidos. El mundo está representado, con realismo creciente, en las bases de datos.

El resultado es que no sólo el volumen de datos crece exponencialmente sino que aumenta en la misma proporción el volumen de datos por analizar.

Desde el punto de vista estadístico se ha pasado de una situación en la que a veces los métodos de análisis no podían aplicarse por no haber datos, a una situación en que los datos superabundantes no son analizados por no haber métodos adecuados. FAYYAD *et al* (1996).

Un carácter distintivo de la minería de datos es que se realiza sobre grandes bases de datos. Éstas no pueden ser confundidas con ficheros muy grandes; son, eso sí, verdaderos sistemas de datos organizados y gestionados según principios y métodos específicos.

Actualmente, el sistema más usado para la gestión de las bases de datos se basa en el modelo relacional. CODD (1970). Este modelo permite no sólo definir la estructura de las bases de datos relacionales sino también la programación de consultas (*queries*) tanto sobre bases de datos locales como sobre bases de datos distribuidas geográficamente.

El modelo relacional se basa en una representación del mundo por **entidades** (objetos) y por **relaciones entre esas entidades y en el cálculo relacional, creado por CODD (1970)**.

Para estructurar una base de datos es necesario empezar por un análisis de los flujos de datos de la organización, identificando y precisando con gran rigor el significado de todos los datos que son generados y circulan en esa organización. Qué datos, cuál es su significado; quién los genera; porqué y cuando son generados; quién los utiliza, cuando, para qué, porqué.

Este trabajo es crucial y tiene carácter estratégico no sólo por los recursos que moviliza sino también por sus consecuencias. Una mala identificación y definición del significado de los datos, de las entidades que interactúan y de sus relaciones puede generar un sistema que no corresponda a la realidad de la organización y a disfunciones que pueden comprometer su existencia.

Una vez identificadas y caracterizadas las **entidades y sus relaciones**, se define el ***esquema de la base de datos***: las entidades y las relaciones entre esas entidades son traducidas o materializadas en ficheros designados por **tablas**. Cada tabla corresponde a un tipo de **entidad** y tiene una estructura formada por **filas y columnas**.

Las filas corresponden a **instancias o realizaciones** del tipo de entidad representado en la tabla. En el sentido lógico, del punto de vista del cálculo relacional, representan instancias de predicados con tantos argumentos como número de atributos.

Las columnas representan los **atributos** del tipo de entidad representada por la tabla.

Las bases de datos son más que imágenes de la organización: son representaciones de la organización para los procesos de decisión a todos los niveles. Las personas toman decisiones sobre esas representaciones y las modifican a cada instante, actuando sobre las representaciones como si fueran las entidades reales.

1.4. MINERÍA DE DATOS Y ESTADÍSTICA.

Puede afirmarse que los conceptos fundamentales de la minería de datos actual tienen sus raíces conceptuales en las ideas de TUKEY (1962, 1966, 1984), acerca del ***Data Analysis*** y en el concepto francés, BENZÉCRI (1973, 1992) de ***Analyse des Donnés***.

En TUKEY (1962), puede leerse

... «*All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make the analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.*»

En estas palabras Tukey anticipaba muchas de las cuestiones fundamentales de la actividad entonces designada ***data análisis***.

Según TUKEY (1966):

... «*The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationship and information and to exhibit the results in such a way as to make them recognizable to the data analyst and recordable for posterity*».

y, mas adelante:

... «*Two considerations are important in implementing data analysis: **first** that the process of analysis usually involves a volume or output much greater then the original body of data. **Second**, that there is no clear barrier between output and input in the overall process of data analysis...*»

Si comparamos estas observaciones de Tukey con la definición más aceptada actualmente de minería de datos - FAYYAD (1996) - se puede verificar que prácticamente coinciden.

En síntesis, aquello que John Tukey designaba, en 1966, como proceso de análisis de datos - o de aprender de los datos - es lo que hoy día se designa por proceso de descubrimiento de conocimiento a partir de bases de datos o minería de datos. En la **tabla 1.4.1** se puede ver el paralelismo de conceptos**.**

Proceso de Análisis de Datos	**Proceso de Minería de Datos / KDD**
Datos	Datos (almacenados en Bases de Datos)
Análisis de Datos	Minería de Datos
Resultados	Patrones y Modelos

Tabla 1.4.1. Análisis de datos versus KDD.

¿Que es lo que distingue la actividad estadística en general del análisis de datos?

Una vez más, TUKEY (1972) da una ayuda al afirmar que

> ... «*Data Analysis instead of "statistics" is a name that allows us to use probability where it is needed and avoid it when we should ...*».

Tukey identificaba la estadística con la estadística matemática y consideraba la actividad de Análisis de Datos más general que la actividad Estadística. *Data Analisis, including Statistic*, es el título de un paper de J. Tukey en 1966.

BENZÉCRI (1973), en la obra que presenta los principios de las concepciones francesas de análisis de datos multivariantes, por oposición a

la estadística matemática tradicional, adopta como segundo principio, el siguiente:

> ... «*...Le modèle doit suivre les données, non l'inverse...*»
> Se puede decir que la práctica de análisis de dados de la Escuela Francesa de Benzécri ha anticipado mucho de lo que hoy día se designa de minería de datos.

HAND (1998), considera la disciplina de Análisis de Datos (*data analysis*) como una ciencia interdisciplinaria, producto de la unión de distintas disciplinas: **Estadística, Informática, Inteligencia artificial**.

Según HAND (1998), Análisis de datos (*Data Analysis*)

> ... «*... is what we do when we turn data into information*».

La minería de datos - que se ocupa del análisis de datos almacenados en grandes bases de datos - sería una rama de la actividad del análisis de datos y ésta algo más general que la Estadística.

Consideremos el problema de investigación en minería de datos o el de desarrollar un sistema para realizar minería de datos (análisis de datos almacenados en grandes bases de datos) y consideremos los conocimientos y pericias necesarias a estas clases de proyectos.

No son suficientes los conocimientos de Estadística Matemática ni los conocimientos que integran hoy día los currícula de un curso de estadística.

Son necesarios conocimientos profundos de informática en diferentes ramas: material (*Hardware*), sistemas operativos, programación, bases de datos.

Es difícil reunir, en una misma profesión, la diversidad y profundidad de los conocimientos necesarios al desarrollo de proyectos de este tipo.

Si los conocimientos de estadística y análisis de datos son cruciales, los de informática no son menos importantes. Ver CLEVELAND (2001).

Por lo tanto, al nivel de la investigación, creación y desarrollo de nuevos sistemas, tiene todo sentido identificar una actividad de minería de datos.

De estas consideraciones, resulta que adoptemos la **figura 1.4.1.** para representar las relaciones entre esas disciplinas; en esa figura la zona azul representa la minería de datos.

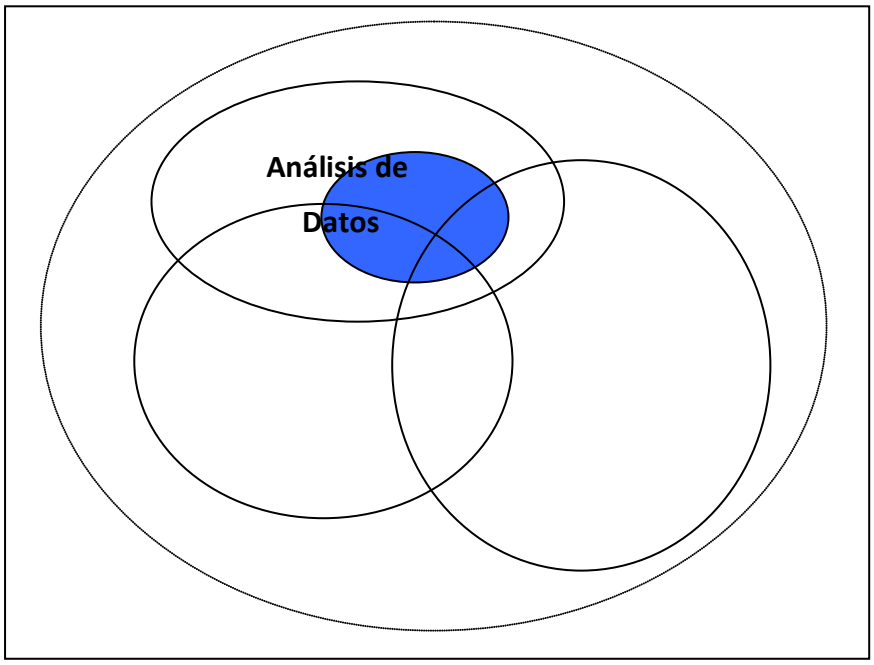


Figura 1.4.1. La actividad de Minería de Datos (azul) es una actividad de análisis de datos con la participación de la Estadística Matemática y la Ciencia de La Información.

En esta tesis, se considera la «Minería de Datos» como una actividad especializada del Análisis de Datos que se ocupa de la investigación y resolución de los problemas que ocurren al analizar los datos almacenados en grandes bases de datos.
El primero de estos problemas es el hecho de que esos datos, no pueden ser cargados en la memoria central de los ordenadores - lo que crea la necesidad de crear algoritmos que tengan ese hecho en consideración.

En síntesis, los aspectos específicos de la minería de datos con relación al análisis de datos multivariantes en general, son:

- Dimensión de los datos: los problemas de minería de datos se refieren a grandes bases de datos, que no pueden ser cargadas en la memoria central de los ordenadores.
- El desarrollo es guiado por los datos (*data driven*). Casi siempre, los sistemas de minería de datos son desarrollados para resolver problemas de datos con estructuras específicas que exigen soluciones específicas. Ver HUBER *et al* (1996); BRADLEY *et al* (1998).
- Los problemas de preparación de los datos - limpiar los datos, recodificar las variables existentes, crear nuevas variables relevantes para el problema, combinar tablas - son aspectos mucho más relevantes que en análisis de datos en general. Ver HUBER *et al* (1996), y KNOBBE *et al* (2001).
- Búsqueda de lo que es interesante. Ver, por ejemplo, JAROSZEWICZ *et al* (2001). En minería de datos - en contraste con lo que ocurre en estadística clásica - DODGE (1996) - se privilegia la búsqueda de observaciones discordantes, de situaciones lejos de los valores esperados o centrales, lo inesperado, sorprendente, *interesante*. Esto es

importante dada la cantidad enorme de resultados generados – lo que implica una selección basada en medidas del grado de interés.

En BURNHAM *et al* (2002), puede verse, en contraste, la perspectiva de muchos estadísticos sobre las actividades de análisis de datos y la minería de datos, considerada como «*data dredging*». La crítica principal se basa en el hecho de que, en minería de datos, no existe la preocupación por la formulación a priori de un modelo, antes de intentar un análisis de datos, de donde resultaría que los modelos obtenidos son sobrestimados, demasiado dependientes de los datos y con poca validez.

1.5. TAREAS Y MÉTODOS DE MINERÍA DE DATOS.

Parece existir consenso acerca de las tareas fundamentales de los sistemas de minería de datos / KDD, en la literatura de minería de datos revisada, incluyendo la más reciente. Ver KLÖSGEN *et al* (2002).

Así, FAYYAD *et al* (1996), HAN *et al* (2001), ALUJA *et al* (1999); ALUJA (2001), FAYYAD *et al* (2002), RHODES (2002), BORGELT *et al* (2001) son casi unánimes en reconocer que esas tareas básicas son las siguientes:

Clasificación.
Es el problema de atribuir a una nueva observación la etiqueta de una de las clases en la que han sido clasificadas las observaciones existentes.

Mas precisamente: si algunos de los objetos observados han sido divididos en *k* clases de equivalencia (han sido considerados indistinguibles o

equivalentes del punto de vista de la información existente) ¿en cuál de esas clases debemos clasificar un nuevo objeto observado, una nueva observación, para la cual no sea conocida la clase?

Los datos de partida tienen la estructura general siguiente:

	$\boldsymbol{X_{(1)}}$	$\boldsymbol{X_{(2)}}$	...	$\boldsymbol{X_{(j)}}$	...	$\boldsymbol{X_{(p)}}$	$\boldsymbol{Y}$
$\boldsymbol{x_1}$	$\boldsymbol{x_{11}}$	$\boldsymbol{x_{12}}$	...	$\boldsymbol{x_{1j}}$	...	$\boldsymbol{x_{1p}}$	$\mathbf{y_1}$
$\boldsymbol{x_2}$	$\boldsymbol{x_{21}}$	$\boldsymbol{x_{22}}$	...	$\boldsymbol{x_{2j}}$	...	$\boldsymbol{x_{2p}}$	$\mathbf{y_2}$
...	...	...	...	...	...	...	…
$\boldsymbol{x_i}$	$\boldsymbol{x_{i1}}$	$\boldsymbol{x_{i2}}$	...	$\boldsymbol{x_{ij}}$	...	$\boldsymbol{x_{ip}}$	$\mathbf{y_i}$
...	...	...	...	...	...	...	…
$\boldsymbol{x_m}$	$\boldsymbol{x_{m1}}$	$\boldsymbol{x_{m2}}$	...	$\boldsymbol{x_{mj}}$	...	$\boldsymbol{x_{mp}}$	$\mathbf{y_m}$
x_{m+1}	$x_{m+1,1}$	$x_{m+1,2}$	…	$x_{m+1,i}$	…	$x_{m+1,p}$	?
…	…	…	…	…	…	…	…
x_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,j}$	…	$x_{n,p}$	?

El paquete de datos está dividido en dos partes: **los datos de entrenamiento** formados por *m* observaciones para cada una de las cuales se conoce el valor de las variables $\boldsymbol{X_{(1)}}$ ... $\boldsymbol{X_{(p)}}$ y la clase, recogida por la variable $\boldsymbol{Y} \in \{1...k\}$ y las *n-m* observaciones para las cuales el valor de $\boldsymbol{Y}$ es desconocido.

El problema de clasificación es el de obtener - usando las primeras *m* observaciones - una regla o modelo que permita atribuir a $\boldsymbol{Y}$ valores $\in\{1...k\}$ para las observaciones $\boldsymbol{x_{m+1}}... \boldsymbol{x_n}$.

Este problema, conocido en Estadística Multivariante como un problema de Análisis Discriminante - ver, por ejemplo, JOHNSON *et al* (1998) - ha sido

designado en Inteligencia Artificial como un problema de ***clasificación supervisada***, ***aprendizaje asistido por un profesor***, ***problema de aprendizaje estadístico*** o ***problema de reconocimiento de patrones.*** Ver VAPNIK (1998), HASTIE *et al* (2001).

Segmentación o Análisis de Clusters.

Dado un conjunto de objetos observados bajo *p* variables, el objetivo es descubrir grupos homogéneos de objetos tales que los objetos que están en cada uno de esos grupos sean más semejantes entre sí que los objetos que están en grupos distintos.

La semejanza entre objetos puede ser calculada en función de los valores de las variables observadas sobre esos objetos o puede ser un dato de observación. El número de grupos puede o no ser especificado antes del proceso de análisis.

Es uno de los procedimientos más conocidos y estudiados en estadística multivariante. En inteligencia artificial es conocido por ***clasificación o aprendizaje no supervisada***. Ver HASTIE *et al* (2001).

La estructura de los datos de partida es:

	$X_{(1)}$	...	$X_{(j)}$	...	$X_{(p)}$	
x_1	x_{11}	…	x_{1j}	…	x_{1p}	y_1
…	…	…	…	…	…	…
x_i	x_{i1}	…	x_{ij}	…	x_{ip}	y_i
…	…	…	…	…	…	…
x_n	$x_{n,1}$	…	$x_{n,j}$	…	$x_{n,p}$	y_n

En resultado del proceso de análisis, a cada una de las observaciones es atribuida una clase cuya etiqueta puede ser recogida en una variable ***Y*** con valores en {1...*k*} en donde *k* es el número de clases.

En contra de lo que ocurre con un problema de clasificación - en donde ***Y*** es un dato conocido - ***Y*** es aquí el resultado del proceso.
En un problema de clasificación se parte del conocimiento de una relación de equivalencia entre los objetos – definida por una variable cualitativa ***Y*** - para construir una función que permite clasificar los otros objetos en esas clases. En un problema de segmentación, al revés, el resultado es una relación de equivalencia - cuyas clases son identificadas por los valores de una variable cualitativa ***Y*** definida por el proceso.

Descripción de Grupos.
Dado un grupo de objetos que ha sido definido, descubierto o identificado según un criterio predefinido en un paquete de datos, ¿cómo caracterizar el grupo usando los atributos considerados relevantes? BENZÉCRI (1973, 1992).

Dados dos grupos que han sido identificados en el proceso de análisis, ¿cómo comparar esos grupos y resumir lo que los distingue? Ver HAN *et al* (2001).

El problema de expresar, mediante un proceso automático, el resumen de un grupo o la comparación de dos grupos en un lenguaje próximo del lenguaje humano ha sido poco tratado en la literatura estadística revisada- a no ser en el contexto del análisis conceptual de datos. Ver MICHALSKY *et al* (1983) y GANTER *et al* (1996).

Importa descubrir cuales son los atributos relevantes para describir un grupo o para comparar dos grupos de objetos.

Interesa definir reglas que permitan construir descripciones o sumarios más abstractos - menos restrictivos pero aún válidos y consistentes con los datos - olvidando algunas restricciones. Ver MICHALSKI *et al* (1983).

Predicción.

El problema de predicción ha sido siempre un problema fundamental en Estadística. Basta considerar el relieve que en esa disciplina tienen las técnicas relacionadas con la previsión: análisis de series cronológicas uni o multivariantes, regresión (simple, múltiple, árboles de regresión) técnicas de previsión.

El problema de clasificación también puede ser mirado como el problema de predecir, con la información disponible, cual va a ser la clase de una observación futura.

Análisis de Asociaciones.

En Estadística el problema de independencia y, por lo tanto, el problema de dependencia entre variables es crucial y muy estudiado.

Los instrumentos fundamentales para ese estudio son el concepto de probabilidad condicionada, el concepto de independencia condicional y los conceptos de correlación lineal.

En Minería de Datos el término «análisis de asociaciones» se refiere al descubrimiento de las relaciones entre variables que resultan de la ocurrencia simultánea de grupos de objetos.

Un problema típico de análisis de asociaciones es el que ocurre en el contexto de las compras realizadas por clientes en supermercados. AGRAWALL *et al* (1993), HASTIE *et al* (2001).

El objetivo es buscar sub-conjuntos de las variables X= (X_1,..., X_p) que ocurren con más frecuencia en una base de datos.
En particular, si $X_j \in \{0,1\}$, ese problema es conocido como el problema de «análisis de la cesta de compras».

La estructura de datos típica es la siguiente:

Transaciones	$X_{(1)}$	$X_{(2)}$	...	$X_{(j)}$	...	$X_{(p)}$
T_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
T_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…
T_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…	…
T_n	$x_{n,1}$	$x_{n,2}$	…	$x_{n,j}$	…	$x_{n,p}$

$$x_{ij} = \begin{cases} 1 & \text{significa que el item } j \text{ ha sido incluido en la transacción } T_i. \\ 0 & \text{significa que el item } j \text{ no ha sido incluido en la transacción } T_i. \end{cases}$$

El objetivo es obtener conjuntos de valores v_l... v_L cuya ocurrencia conjunta tenga una probabilidad significativa. Ver HASTIE *et al* (2001).

Las tareas características antes identificadas pueden ser realizadas por distintas técnicas de análisis. Ver ,por ejemplo, BORGELT *et al* (2001).

De la literatura revisada - ver KLÖSGEN (2002) y BRADLEY, *et al*.(1998), por ejemplo - se concluye que estas técnicas deben tener ciertas características, las principales de las cuales son:

- **Interpretabilidad** - los resultados generados por la técnica deben poder ser fácilmente interpretados por los seres humanos.
- **Incrementabilidad** - los resultados producidos por la técnica deben poder ser modificados, cuando ocurran nuevos datos, sin necesidad de procesar de nuevo los datos usados para obtener esos resultados.
- **Escalabilidad** - las técnicas deben poder aplicarse bien tanto a pequeños como a grandes paquetes de datos.

Algunas técnicas - como los árboles de clasificación – se han revelado muy bien adaptadas, según todos estos criterios, a las actividades de minería de datos.

Las referencias a la posibilidad de utilizar técnicas de tipo factorial son menos frecuentes en la literatura revisada. Ver, por ejemplo, ALUJA *et al* (1999), o ALUJA (2001).

Se verificará en los capítulos siguientes - en particular en los **capítulo III** y **IV** - que el biplot, por estar teóricamente muy relacionado con todas las técnicas más importantes de análisis multivariante y permitir la representación gráfica plana de casi todos los resultados producidos por esas técnicas es, potencialmente, según todos los criterios identificados, un instrumento natural de minería de datos.

C A P Í T U L O II

LOS MÉTODOS BIPLOT

2.1. INTRODUCCIÓN Y RESUMEN HISTÓRICO.

GABRIEL (1971), define un biplot como una representación gráfica de una matriz de datos $X_{(n\times p)}$ resultante de observar *n* individuos en *p* características numéricas.

Un *biplot* es una representación gráfica plana o tridimensional.

El bi en la palabra *biplot* se refiere al hecho de que en ese gráfico existen dos tipos de marcadores correspondientes a los dos tipos de información: los marcadores para los individuos o filas y los marcadores para las variables o columnas.

Usando la misma lógica, para un conjunto de datos de tres vías $X_{(n\times p\times q)}$ resultante de observar *n* individuos con *p* características numéricas en *q* ocasiones, la representación gráfica plana o de tres dimensiones en que ocurran 3 tipos de marcadores (individuos, variables y ocasiones) debería designarse por *triplot.* Sin embargo esto fue llamado por GABRIEL un Bimodel y actualmente este término no se usa.

En la **figura 2.1.1.** se presenta un *biplot* en donde los marcadores están representados por puntos en un *espacio de representación* de dimensión *3*.

Si consideráramos las proyecciones de estos marcadores en un plano, el *biplot* quedaría ahora representado en un espacio de dimensión 2.

Los *biplots* permiten, por inspección visual, identificar relaciones entre variables, relaciones entre individuos y relaciones entre variables e individuos.

		b_1 ↓	b_2 ↓	b_3 ↓	b_4 ↓
		$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$
$a_1\rightarrow$	1	x_{11}	x_{12}	x_{13}	x_{14}
$a_2\rightarrow$	2	x_{21}	x_{22}	x_{23}	x_{24}
$a_3\rightarrow$	3	x_{31}	x_{32}	x_{33}	x_{43}
$a_4\rightarrow$	4	x_{41}	x_{42}	x_{34}	x_{44}
$a_5\rightarrow$	5	x_{51}	x_{52}	x_{35}	x_{45}

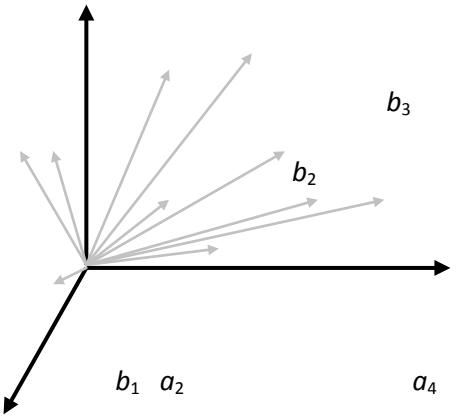


Figura 2.1.1. Un biplot puede estar representado en un espacio de representación de dos o más dimensiones.

BENZÉCRI (1973), al presentar el análisis factorial de correspondencias, pensado para tablas de contingencia de dos vías, relacionando las categorías de dos variables cualitativas, sintetiza las conclusiones de esos análisis en gráficos planos designados «*representations simultanées*». Esos gráficos tienen marcadores para las categorías en filas y marcadores para las categorías en columnas; conociendo las coordenadas de los marcadores de las filas es posible calcular las coordenadas de los marcadores de las columnas usando las fórmulas de transición.

GREENACRE (1984), en la obra que presenta a los lectores anglosajones el Análisis Factorial de Correspondencias de Benzécri, introduce entre muchas otras innovaciones, el concepto de descomposición en valores singulares generalizados (GSVD), base para el concepto de biplot generalizado.

GALINDO (1985), generaliza el concepto de representación simultánea creando un nuevo tipo de biplot - el HJ Biplot - que se aplica a todo conjunto de datos de dos vías y permite representar los individuos y las variables con igual calidad de representación - lo que no ocurre con los biplots de GABRIEL (1971).

En secuencia de GALINDO (1985), se ha desencadenado en el Departamento de Estadística de la Universidad de Salamanca un conjunto de investigaciones relativas al concepto de *biplot* - tanto en el aspecto teórico como de aplicaciones - que han culminado en tesis de doctorado y otras publicaciones de entre las cuales se citan:

VICENTE-VILLARDÓN (1992) – demuestra el papel central del concepto de biplot en análisis de datos multivariantes y presenta una generalización: el biplot generalizado.

VICENTE-TAVERA (1992) - estudia las relaciones entre los biplots y los métodos de análisis cluster, desarrollando un método de análisis cluster basado en el concepto de inercia.

MARTIN-RODRIGUEZ (1996, 2000)- generaliza a los biplots de Gabriel, HJ-biplot y biplot generalizado, los métodos de integración de análisis en componentes principales de KRZANOWSKI, (1979).

FERNANDEZ-GOMEZ (1995)- presentan una alternativa al análisis canónico de correspondencias basada en los métodos de biplot, con importantes aplicaciones en ecología.

DIAZ-LENO (1995) – desarrolla, basándose en trabajos anteriores de BRADU y GABRIEL (1978), métodos gráficos para diagnóstico visual de modelos logaritmo – lineares y de independencia condicional.

CÁRDENAS (2000) - Investiga los aspectos inferenciales de los biplots utilizando la metodología de los Modelos Bilineales Generalizados, generalizando el ajuste de los Biplots con Información Externa para variables de la familia exponencial.

AMARO-MARTÍN (2001) - Realiza una investigación teórica de las propiedades de los MANOVA BIPLOT en el contexto del MODELO LINEAL GENERAL MULTIVARIANTE, desarrollando métodos de interpretación de los MANOVA-BIPLOTS.

VARELA-NUALLES (2002) – aplica los métodos de biplot al análisis de las interacciones de orden elevada que ocurren en modelos bilineares al conjunto de datos de tres modos o más.

2.2. BIPLOTS DE GABRIEL.

2.2.1. DEFINICIÓN.

GABRIEL (1971), introduce el término biplot usando la definición siguiente:

*«Toda matriz de rango dos puede ser representada gráficamente como un **biplot** que consiste en un vector para cada fila y en un*

vector para cada columna, elegidos de modo que cada elemento de la matriz, sea exactamente el producto interno de los vectores que corresponden a esa fila e a esa columna. Si una matriz tiene rango superior a 2, se puede representar esa matriz, de modo aproximado, por un biplot de una matriz de rango 2 - que es una aproximación de la matriz original».

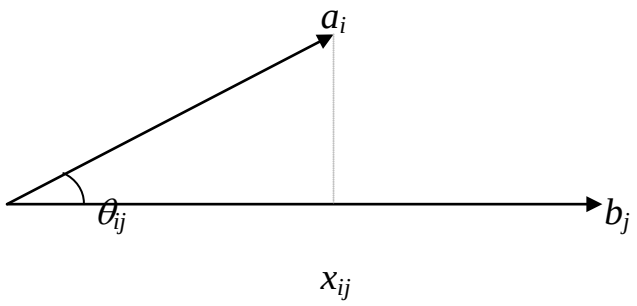
Esto significa que el término biplot ha sido asociado, explícitamente, por su creador, al concepto de la descomposición de una matriz de datos numéricos $X_{(n\times p)}$ con las mediciones de p variables numéricas sobre n individuos y a una representación gráfica plana.

Explícitamente

$$X_{(n\times p)} = \left[x_{ij}\right] = \left[a_i^T b_j\right] = \left[< a_i\, b_j >\right] = \left[\| a_i \| b_j \| \cos\theta_{ij}\right]$$

donde a_i y b_i son vectores de un espacio de dimensión r - el espacio de representación - usados, respectivamente, para representar los individuos y las variables, siendo r = rango(X) $\leq$ min (n, p).

Los vectores a_i (i= 1…n) son designados marcadores de los individuos o filas de X y los vectores b_j son designados marcadores de las variables o columnas de X.



$x_{ij} = | a_i | | b_j |$ cos θ_{ij}, en donde θ_{ij} es el ángulo formado por el marcador de la fila (individuo) *i* con el marcador de la columna *j*.

Si *r*= rango(*X*) ≤ 3, entonces es posible una representación gráfica perfecta (sin pérdida de información) para el biplot - sobre una recta, un plano o un espacio 3D.

Cuando *r* = rango(*X*) > 3, no es posible representar gráficamente, sin pérdida de información, a la matriz *X*. Si fijamos arbitrariamente la dimensión del espacio de representación en *d*= 2, entonces los marcadores - tanto para las variables cuanto para los individuos - son puntos o vectores de $\Re^2$.
Cuando *r* = rango(*X*) > 2, esto significa que la representación gráfica correspondiente no es perfecta en el sentido de que no representa toda la información de *X*.

En la definición de GABRIEL (1971), no está especificado el método de obtención de los marcadores*;* en ese trabajo fundador, Gabriel usa el método de los mínimos cuadrados y la descomposición en vectores y valores singulares de *X*. En artículos posteriores - GABRIEL (1998) - usa también el método de los mínimos cuadrados alternados (CRISS-CROSS) para obtener esos marcadores.

Sea $X= U\Sigma V^T$ la descomposición de $X_{(n\times p)}$ en vectores y valores singulares - ECKART *et al* (1936) - en donde $U_{(n\times r)}$ es una matriz cuyas columnas son los vectores singulares a la izquierda de *X*, $V_{(p\times r)}$ una matriz cuyas columnas son los vectores singulares a derecha de *X* y $\Sigma_{(r\times r)}$ es una matriz diagonal de valores singulares $\sigma_1 \ldots \sigma_r$ de *X*.

Las columnas $u_{(1)} \ldots u_{(r)}$ de U son ortogonales:

$$u_{(i)}^T u_{(j)} = 0 \quad i \neq j \quad y \qquad u_{(i)}^T u_{(i)} = \left\| u_{(i)} \right\|^2 = 1$$

Las columnas $v_{(1)} \ldots v_{(r)}$ de V son vectores ortogonales:

$$v_{(j)}^T v_{(j)} = \left\| v_{(j)} \right\|^2 = 1$$

GABRIEL (1971), elige como marcadores para las filas de X los vectores que forman las filas de A y para marcadores de las variables los vectores que forman las filas de B, en la expresión de la descomposición:

$$X = \left(U\Sigma^{\alpha}\right)\left(V\Sigma^{1-\alpha}\right)^T = AB^T \textbf{ con :}$$
$$A = U\Sigma^{\alpha} \quad \textbf{y} \quad B = V\Sigma^{1-\alpha}, \ \alpha \in [0\ ,1]$$

en donde

$A_{(n\times r)}$ y $B_{(p\times r)}$ con $r = \text{rango}(X) \leq \min(n, p)$.

Esto significa que los biplots de Gabriel forman una familia infinita - indexada por $\alpha \in [0,1]$.

Cuando pretendemos visualizar estos biplots hay que elegir la dimensión $d \leq r$ del espacio de representación.

En el caso $d = 2$ (biplots planos), se utilizan para marcadores de los individuos las filas de las submatrices formadas por las dos primeras columnas de A y para marcadores de los individuos las filas de las de B (para las variables).

Designemos por $A_{(2)}$ y por $B_{(2)}$ a esas submatrices de A y de B y por $\Sigma_{(2)}$ a la matriz diagonal correspondiente.

Cuando r= rango(X) $\geq$ min(n, p) es superior a 2, la representación

$$X_{(2)} = A_{(2)}\, \Sigma_{(2)}\, B_{(2)} \neq X$$

es una aproximación inferior, verificándose pérdidas de información.

Sea Y la matriz que resulta centrando las columnas de X . Entonces Y^TY= n * *COV(X),* donde *COV(X)* es la matriz de covarianzas de las variables-columnas de X.

YY^T es entonces una matriz cuyos elementos son los productos internos de las filas:

$$YY^T = [d_{il}] = \left[(x_i - \bar{x})^T (y_l - \bar{x})\right]$$

en donde x_i y x_l son las filas de X correspondientes a los individuos i y l.

El ángulo θ entre los vectores (x_i - $\bar{x}$) y (x_l - $\bar{x}$) es dado por

$$\cos\theta = \frac{(x_i - \bar{x})^T \ (x_l - \bar{x})}{\|x_i - \bar{x}\| \times \|x_l - \bar{x}\|}$$

con

$$\|x_i - \bar{x}\| = \sqrt{(x_i - \bar{x})^T \ (x_i - \bar{x})}$$

2.2.2. PROPIEDADES DE LOS BIPLOTS DE GABRIEL.

En GABRIEL (1971, 1995a, 1995b) pueden verse las propiedades geométricas y estadísticas de estos biplots.

En este apartado se da relieve únicamente a aquellas propiedades y aspectos que se consideran más interesantes para el objetivo de construir un sistema de minería de datos basado en biplots.

Sea Y la matriz resultante de centrar las columnas de la matriz de datos $X_{(n,p)}$.

Descomponiendo esta matriz:

$$Y = U\,\Sigma V^T = U\,\Sigma^{\alpha}\left(V\,\Sigma^{1-\alpha}\right) = A\,B^T$$

Con $A = U\,\Sigma^{\alpha}$, $B = V\,\Sigma^{1-\alpha}$ y $\alpha \in [0,1]$

Sean a_i el marcador de la fila número i y b_j el marcador de la columna número j de Y $(i = 1 \ldots n\ ;\ j = 1 \ldots p)$

$$\left\|a_i\right\|^2 = \sum_{k=1}^{r} u_{ik}^2\ \sigma_k^{2\alpha} = \sum_{k=1}^{r} u_{ik}^2\ \lambda^{\alpha}$$

$$\left\|b_j\right\|^2 = \sum_{k=1}^{r} v_{jk}^2\ \sigma_k^{2(1-\alpha)} = \sum_{k=1}^{r} v_{jk}^2\ \lambda_k^{1-\alpha}$$

en donde $\lambda_i = \sigma_i^{\,2}$ son los valores propios de Y^TY, v_{ij} el elemento genérico de V y u_{ij} el elemento genérico de U.

Si se consideran solo las dos primeras componentes de los marcadores, entonces

$$Y_{(2)} = U_{(2)}\,\Sigma_{(2)}\,V_{(2)}^{T} = U_{(2)}\,\Sigma_{(2)}^{T}\left(V_{(2)}\,\Sigma_{(2)}^{1-\alpha}\right) = A_{(2)}\,B_{(2)}^{T}$$

Cuando r= rango(Y)= 2, $Y \equiv Y_{(2)}$.

Cuando r> 2, $Y_{(2)}$ es una aproximación inferior de Y.

Para la fila $y_{i,}$

$$y_i = \left[a_i^T b_1, ..., a_i^T b_j, ..., a_i^T b_p\right] = Ba_i \quad (i = 1...n)$$

Para la columna $y_{(j)}$,

$$y_{(j)} = \begin{bmatrix} a_1^T\, b_j \\ ... \\ a_i^T\, b_j \\ ... \\ a_n^T\, b_j \end{bmatrix} = Ab_j \quad (j = 1, ..., p)$$

$$y_{ij} = a_i^T\, b_j = \left[u_{i1}\,\sigma_i^{\alpha}, ..., u_{ik}\,\sigma_i^{\alpha}, ..., u_{ir}\,\sigma_i^{\alpha}\right] \begin{bmatrix} v_{j1}\,\sigma_{1i}^{1-\alpha} \\ ... \\ v_{jk}\,\sigma_{ki}^{1-\alpha} \\ ... \\ v_{jr}\,\sigma_r^{1-\alpha} \end{bmatrix}$$

$$= u_{i1}\,v_{j1}\,\sigma_i^{\alpha} + ... + u_{ir}\,v_{jr}\,\sigma_r^{\alpha}$$

Si consideramos solamente las dos primeras componentes obtenemos la aproximación $y_{ij}^{(2)}$.

$$y_{ij}^{(2)} = u_{i1}\,v_{j1}\,\sigma_1^{\alpha} + u_{i2}\,v_{j2}\,\sigma_2^{\alpha}, \quad \alpha \in [0, 1]$$

GABRIEL (1995a), muestra que toda combinación lineal de filas o columnas de Y tiene como imagen en el biplot, la transformación lineal correspondiente de los marcadores de esas filas o columnas. Como se puede ver abajo - ver 2.3.2. - esta propiedad también se verifica para el biplot de GALINDO.

2.2.3. CASO PARTICULAR α= 0. EL CMP BIPLOT.

Cuando en

$$Y = U\,\Sigma^{\alpha}\left(V\,\Sigma^{1-\alpha}\right)^{T} = A\,B^{T}$$

sustituimos α = 0, obtenemos el caso particular originalmente designado *GH - biplot* por Gabriel (1971):

$$Y = U(V\Sigma)^{T}, A\,B^{T} \text{ con } A = U \;\; y \;\; B = V\Sigma$$

La designación *CMP-Column Metric Preserving* para este *biplot* ha sido introducida por GREENACRE (1984), y aceptado por GABRIEL (1995a), para ilustrar el hecho de que, en ese biplot, los productos internos de los marcadores de las columnas reproducen las covarianzas y las correlaciones de las variables.

En efecto - ver GABRIEL (1971, 1995a) - la matriz de covarianzas de los datos es:

$$S = \frac{1}{n}\left[y_{(j)}^T\ y_{(k)}\right] = \frac{1}{n}Y^T Y = \frac{1}{n}\left(U\,\Sigma V^T\right)^T U\,\Sigma V^T = \frac{1}{n}V\,\Sigma U^T\,U\,\Sigma V^T$$
$$= \frac{1}{n}V\,\Sigma\left(V\,\Sigma\right)^T = \frac{1}{n}B\,B^T = \frac{1}{n}\left[b_j^T\ b_k\right]$$

en donde $B = V\,\Sigma^{1-\alpha}$ es $V\,\Sigma$ cuando $\alpha = 0$.

Esto quiere decir que, en ese caso:

$$s_j^2 = s_{jj} = Var\,x_{(j)} = \frac{1}{n}b_j^T\ b_j = \frac{1}{n}\left\|b_j\right\|^2$$
$$s_j = \frac{1}{\sqrt{n}}\left\|b_j\right\|$$

La desviación-típica de una variable queda representada, en este biplot, por el tamaño del marcador respectivo salvo la constante $\frac{1}{\sqrt{n}}$.

Sí $\quad s_j^2 \approx \left\|b_j\right\|^2$, $s_k^2 \approx \left\|b_k\right\|^2$, $s_{jk}^2 \approx b_2^T\ b_k$

la correlación r_{jk} entre las variables j y k es

$$r_{jk} = \frac{s_{jk}}{s_j \times s_k} = \frac{b_j^Y\ b_k}{\left\|b_j\right\|\left\|b_k\right\|} = \cos\left(b_j, b_k\right)$$

El coeficiente de correlación entre las variables j y k queda representado en el biplot por el coseno del ángulo entre los marcadores respectivos.

En este biplot, las coordenadas de los marcadores de las variables son las proyecciones sobre las direcciones principales de inercia de las columnas de *Y*. Esto **no** ocurre con los marcadores de los individuos, una vez que

$Y\ V$ = Proyecciones de las filas (individuos) sobre las direcciones principales

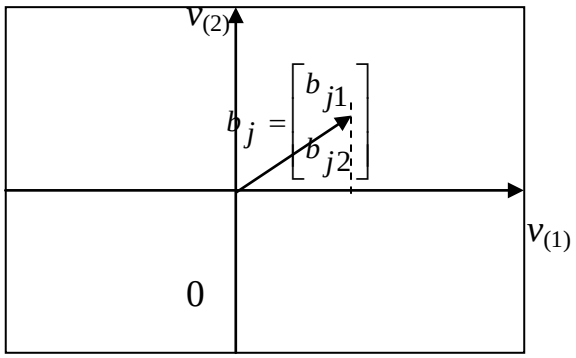
$$= \left(U\,\Sigma^{\alpha}\right)\left(V\,\Sigma^{1-\alpha}\right)^{T} V = U\,\Sigma \neq A$$

En efecto, $U\,\Sigma$ es distinto de *A*, una vez que, en este biplot, $A = U = U\,\Sigma^{0}$.

Por lo tanto, en el CMP-biplot, las proyecciones de los individuos sobre los ejes principales no coinciden con las coordenadas de los marcadores de los individuos sobre esos ejes.

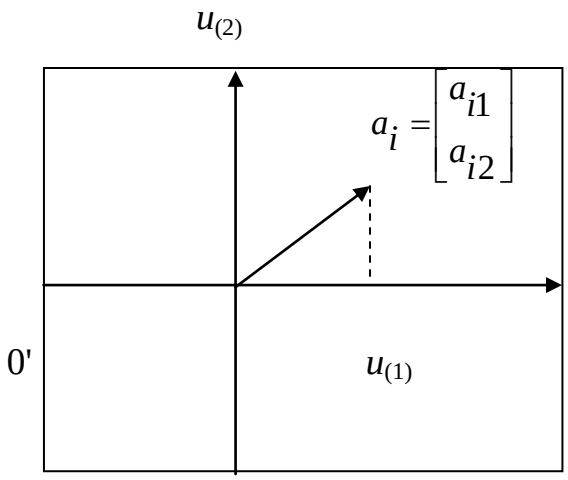
En este caso, el biplot es simplemente una superposición de gráficos (representación simultanea) cuyos ejes usan escalas distintas cuando se trata de individuos o de variables. Ver **figura 2.2.3.1. a, b, c**.

a)



Unidades: σ_1 para el eje 1; σ_2 para el eje 2.

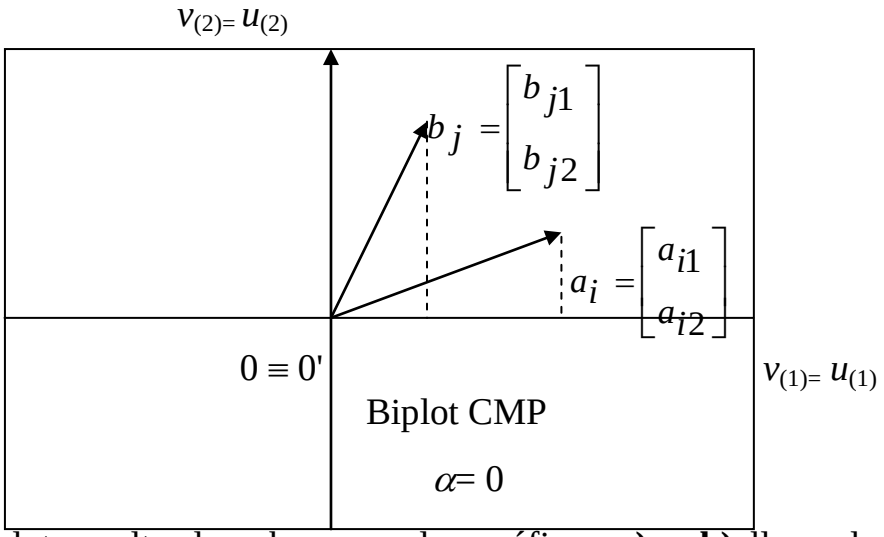
b)



$a_{i1} = u_{i1} \times 1 \neq y_i^T v_{(1)} \qquad a_{i2} = u_{i2} \times 1 \neq y_i^T v_{(2)}$

Unidades: 1 para el eje 1; 1 para el eje 2.

c)



El biplot resulta de sobreponer los gráficos **a)** y **b)** llevando a coincidir 0 con 0' y los ejes con el mismo significado. Pero las escalas son distintas.

Figura 2.2.3.1. Construcción de CMP-biplot.

Una propiedad interesante del punto de vista estadístico es la de que - ver GABRIEL (1971, 1995a) - la distancia de Mahalanobis entre dos filas de Y - distancia de métrica S^{-1} - es dada por la distancia euclidiana entre los respectivos marcadores. Esto significa que las distancias entre marcadores de individuos no reproducen las distancias (disimilitudes) entre esos individuos.

En efecto, si y_i e y_e son dos filas de Y,

$$d^2_{S^{-1}}(y_i, y_e) = (y_i - y_e) S^{-1} (y_i - y_e)$$
$$= \langle y_i - y_e, y_i - y_e \rangle_{S^{-1}} = (B a_i - B a_e)^T S^{-1} (B a_i - B a_e)$$

Recordando que

$$S = \frac{1}{n} Y^T Y = (U \Sigma V^T)^T U \Sigma V^T = \frac{1}{n} V \Sigma^2 V^T$$
$$S^{-1} = n V \Sigma^{-2} V^T$$

Por lo tanto:

$$d^2_{S^{-1}}(y_i, y_e) = n (a_i - a_e)^T B^T V \Sigma^{-2} V^T V \Sigma (a_i - a_e)$$
$$= n (a_i - a_e)^T (a_i - a_e) = n \| a_i - a_e \|^2_I$$

En conclusión: este biplot reproduce bien las propiedades estadísticas y geométricas de las variables, pero los individuos quedan mal representados.

2.2.4. CASO PARTICULAR α= 1. EL RMP BIPLOT.

Cuando en la descomposición

$$Y = U\,\Sigma^{\alpha}\left(V\,\Sigma^{1-\alpha}\right)^{T} = A\,B^{T}$$

substituimos α= 1, obtenemos el caso particular designado originalmente por JH-biplot en GABRIEL (1971).

$$Y = U(V\Sigma)^{T} \quad , \quad A = U\Sigma \quad , \quad B = V$$

La designación RMP - Row Metric Preserving - introducida por GREENACRE (1984) - es justificada por el hecho de que, ahora, el biplot preserva las distancias euclidianas entre individuos.

Si y_i e y_e son dos filas de Y, entonces

$$\begin{aligned}(y_i - y_e)^{T}(y_i - y_e) = \|y_i - y_e\|_I^2 &= (Ba_i - Ba_e)^{T}(Ba_i - Ba_e) \\ &= (a_i - a_e)^{T}\,B^{T}\,B(a_i - a_e)\end{aligned}$$

$$\|y_i - y_e\|_I^2 = (a_i - a_e)^{T}\,V^{T}\,V(a_i - a_e) = (a_i - a_e)^{T}(a_i - a_e) = \|a_i - a_e\|_I^2$$

Pero, ahora, B= V. Por eso:

Así, en este biplot, las distancias euclidianas entre filas de Y (individuos) quedan bien representadas por las distancias euclidianas entre los marcadores respectivos. Si la dimensión del gráfico es $d = 2 < r$ = rango(Y) esas distancias son aproximaciones inferiores a las verdaderas distancias:

$$\|y_i - y_e\|_I^2 \geq \|a_i^{(2)} - a_i^{(2)}\|_I^2$$

Esto significa que las proximidades entre individuos (medidas por las distancias euclidianas) quedan bien representadas en el JK biplot por las distancias euclidianas entre los marcadores correspondientes - lo que explica que la designación RMP - Row Metric Preserving para este biplot.

Las variables, en cambio, no quedan bien representadas.

Así, las covarianzas entre dos variables $y_{(j)}$ e $y_{(k)}$ serían representadas por

$$Cov\left(y_{(j)}, y_{(k)}\right) = \frac{1}{n} y_{(j)}^T y_{(k)}^T = \frac{1}{n}\left(Ab_j\right)^T \left(Ab_k\right) = \frac{1}{n} b_j^T A^T A b_k$$
$$= \frac{1}{n} b_j^T \left(U\Sigma\right)^T \left(U\Sigma\right) b_k = \frac{1}{n} b_j^T \Sigma U^T U \Sigma b_k$$
$$= \frac{1}{n} b_j^T \Sigma^2 b_k$$

valor que es distinto del verdadero valor:

$$\frac{1}{n} b_j^T b_k$$

Si consideramos un producto interno de variables basado en $(Y^T Y)^{-1}$ - semejante a la métrica de Mahalanobis para los individuos -

$$\left\langle y_{(j)}, y_{(k)} \right\rangle_{\left(Y Y^T\right)^{-1}} = \frac{1}{p} y_{(j)}^T \left(Y Y^T\right)^{-1} y_{(k)} = \frac{1}{p}\left(Ab_j\right)^T \left(y y^T\right)^{-1} Ab_k$$

Una vez que $Y Y^T = U \Sigma^2 U^T$,

se tiene $(Y Y^T)^{-1} = U \Sigma^2 U^T$, de donde resulta:

$$d^2\left(y_{(j)}, y_{(k)}\right) = \frac{1}{p}(Ab_j)^T U \Sigma^{-2} U^T Ab_k = \frac{1}{p} b_j^T A^T U \Sigma^{-2} U^T Ab_k$$
$$= \frac{1}{p} b_j^T \Sigma \underbrace{U^T U}_{I} \Sigma^{-2} \underbrace{U^T U}_{I} b_k = \frac{1}{p} b_j^T b_k$$

Como lo reconoce GABRIEL (1995a), una métrica basada en este producto interno tiene poco interés estadístico.

En síntesis: del punto de vista estadístico, este biplot no es tan interesante como el CMP una vez que este produce una buena representación de las correlaciones entre variables.

Tal como en el CMP, se verifica en el RMP una asimetría en la calidad de representación de las filas y de las columnas:

Los marcadores a_i de las filas (i= 1, …, n), son las proyecciones de las filas sobre las direcciones principales.

En efecto: $Y\,V = U\,\Sigma\,V^T\,\mathrm{V} = U\,\Sigma = A$, lo que significa que el marcador a_i de la fila i se proyecta sobre la dirección principal $v_{(j)}$ por a_{ij} (j= 1, …, r), como se puede representar en la **figura 2.2.4.1.a)**).

En cambio, el marcador b_j de la columna $y_{(j)}$ tiene, en el sistema de ejes principales, coordenadas que **no son** las proyecciones de la variable sobre esos ejes. Ver **figura 2.2.4.1.b)**.

En efecto: $X^T U$= Proyecciones de las variables sobre las direcciones principales

$= V \Sigma$.

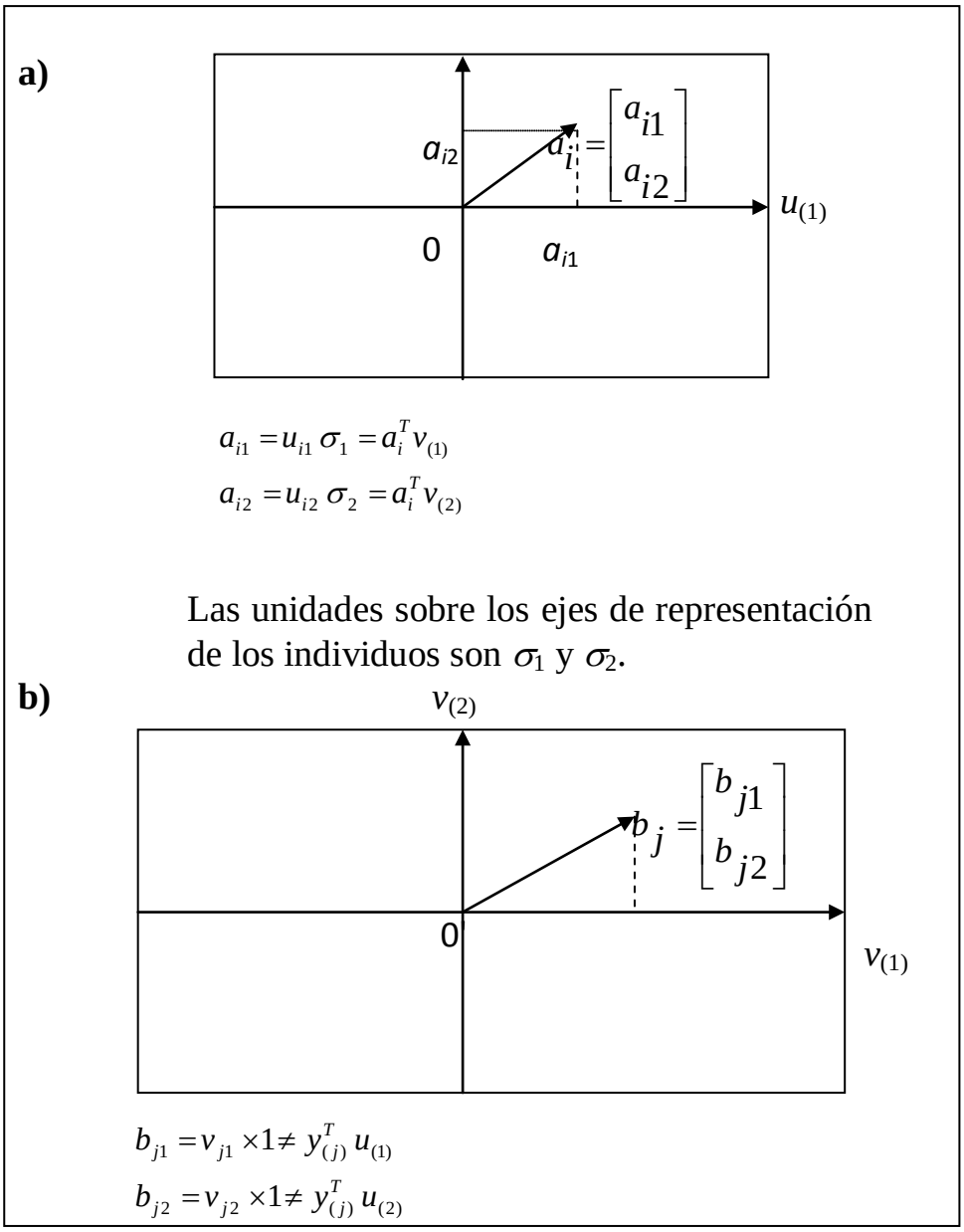


Figura 2.2.4.1. Construcción del RMP-biplot.

La escala de los ejes en la representación de las variables es la misma (1) para todos los ejes - al contrario de lo que ocurre en la representación de los individuos, en donde las escalas son $\sigma_1, \ldots, \sigma_r$.

Por lo tanto, el gráfico que resulta de superponer los dos gráficos anteriores **no es** una verdadera representación simultánea, sino, simplemente una representación conjunta de las dos nubes de individuos y variables. Ver GALINDO (1985).

2.3. BIPLOTS DE GALINDO.

2.3.1. DEFINICIÓN.

GALINDO (1985) propone un nuevo tipo de biplot – denominado HJ-biplot - que, aunque no verifica $y_{ij} = a_i^T b_j$ (i= 1, …, n; j= 1, …, p) garantiza que tanto variables como individuos están representados en un biplot plano con la misma, y máxima, calidad de representación:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{r} \lambda_i}$$

La definición del HJ-Biplot dada por GALINDO (1985, 1986), es la siguiente:

> … *«El HJ-Biplot es una representación gráfica multivariante de las filas de una matriz* $X_{n \times p}$ *mediante marcadores* $j_1, \ldots, j_n$ *para sus filas y* $h_1, \ldots, h_p$ *para sus columnas, elegidos de forma que ambos marcadores pueden ser superpuestos en un mismo sistema de referencia con máxima calidad de representación».*

Si Y es la matriz de datos después de centrada, su descomposición en valores y vectores singulares es

$$\underset{(n \times p)}{Y} = \underset{(n \times r)}{U}\ \underset{(r \times r)}{\Sigma}\ \underset{(r \times p)}{V^T}$$

Los marcadores para las filas de Y (individuos) usados en esta representación biplot son las filas $a_1, \ldots, a_n$ de

$$\underset{(n\times p)}{A} = \underset{(n\times r)}{U}\,\underset{(r\times r)}{\Sigma}$$

y los marcadores usados para representar a las variables son las filas $b_1, \ldots, b_p$ de la matriz

$$\underset{(p\times r)}{B} = \underset{(p\times r)}{V}\,\underset{(r\times r)}{\Sigma}$$

Este criterio para elegir los marcadores equivale a usar para marcador de una fila de Y al vector cuyas coordenadas son las proyecciones de esa fila sobre los vectores singulares a derecha de Y.

Para marcadores de una columna se usa el vector cuyas coordenadas son las proyecciones de esa columna sobre los vectores singulares a izquierda de Y.

En efecto, de:

$$Y = U\,\Sigma\,V^T$$

resulta

$$YV = U\,\Sigma = A \quad \text{(Marcadores de las filas)}$$

De:

$$U^T\,Y = \Sigma\,V^T$$

resulta

$$Y^T\,U = V\,\Sigma = B \quad \text{(Marcadores de las columnas)}$$

Con esta elección de marcadores los valores y_{ij} **no** son productos internos de los nuevos marcadores.

En efecto, $y_{ij} \neq a_i^T b_j$. En cambio, la calidad de la representación plana de individuos y variables es ahora máxima:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{r} \lambda_i}.$$

2.3.2. PROPRIEDADES DEL BIPLOT DE GALINDO.

P1 - El biplot de GALINDO preserva la geometría de las filas y de las columnas:

$A = U\,\Sigma$ - contiene en sus filas los marcadores de los individuos.

$B = V\,\Sigma$ - contiene en sus filas los marcadores de las variables.

$Y^T Y$= [Productos cruzados de columnas j, k]

$$= \left[n\, s_{jk}\right]$$
$$= \left(U\,\Sigma V^T\right)^T \left(U\,\Sigma V^T\right) = V\,\Sigma U^T\, U\,\Sigma V^T$$
$$= V\,\Sigma^2\, V^T = \left(V\,\Sigma\right)\left(V\,\Sigma\right)^T = B\,B^T = \left[b_j^T\, b_k\right]$$

$Y^T Y$= [Productos cruzados de filas]

= [Disimilitudes entre filas]

$$= \left(U\,\Sigma V^T\right)\left(U\,\Sigma V^T\right)^T = U\,\Sigma^2\, U^T$$
$$= \left(U\,\Sigma\right)\left(U\,\Sigma\right)^T = A\,A^T = \left[a_i^T\, a_e\right]$$

Esto significa que:

$$n\, s_{jj} = n\, s_j^2 = b_j^T\, b_j = \left\| b_j \right\|^2$$

(La norma del marcador b_j de una columna representa - salvo un factor $\frac{1}{\sqrt{n}}$ *- la desviación típica de la variable correspondiente a esa columna).*

También:

$$n\,s_{jk} = b_j^T\, b_k$$

(El producto interno de dos marcadores - de variables representan - salvo un factor $\frac{1}{n}$ *- la covarianza de las variables correspondientes).*

Por lo tanto, si θ_{jk} es el ángulo entre los marcadores b_j y b_k,

$$\cos(\theta_{jk}) = \frac{b_j^T\, b_k}{\|b_j\| \times \|b_k\|} = \frac{n\,s_{jk}}{\sqrt{n}\,s_j \times \sqrt{n}\,s_k} = r_{jk}$$

= Coeficiente de correlación entre las variables *j* y *k*, representadas por las columnas $y_{(j)}$ e $y_{(k)}$.

Si $\|y_i - y_e\|^2$ es la distancia euclidiana entre dos filas, representando la disimilitud entre los objetos correspondientes, se verifica:

$$\begin{aligned}\|y_i - y_e\|^2 &= (y_i - y_e)^T (y_i - y_e) = y_i^T - y_i - y_i^T - y_e + y_e^T - y_i + y_e^T - y_e \\ &= \|y_i\|^2 - 2\, y_i^T - y_e + \|y_e\|^2 \\ &= a_i^T\, a_i - 2\, a_i^T\, a_e + a_e^T\, a_e \\ &= (a_i - a_e)^T (a_i - a_e) = \|a_i - a_e\|^2\end{aligned}$$

Esto significa que el biplot de Galindo preserva la geometría euclidiana de las columnas (interpretada estadísticamente en términos de covarianzas) y la geometría euclidiana de las filas (interpretada estadísticamente en términos de disimilitudes).

Esta propiedad justifica la designación *RCMP (Row Colum Metric Preserving*) que puede ser adoptada para el biplot de Galindo.

P2 - Sean $I = \{1, \ldots, n\}$ y $J = \{1, \ldots, p\}$ los conjuntos de índices para los individuos y para las variables, respectivamente

Sean $F \subseteq I$ con $|F| = n_f$ y $C \subseteq J$ con $|C| = n_c$

Si $\sum_{f \in F} w_f \, y_f$ es una combinación lineal de las filas y_f, con pesos w_f, $f \in F$

y

$\sum_{c \in C} w_c \, y_{(c)}$ es una combinación lineal de las columnas $y_{(c)}$, con pesos $w_{(c)}$, $c \in C$

entonces estas combinaciones lineales, son representadas en el biplot de Galindo por los marcadores

$\sum_{f \in F} w_f \, a_f$ para las filas

y $\sum_{c \in C} w_c \, b_c$ para las columnas.

En particular, para $w_f = \frac{1}{n_f}$ y para $w_{(c)} = \frac{1}{n_c}$ se puede afirmar que:

«*Si seleccionamos conjuntos de filas o columnas de Y, los marcadores que en el RCMP-Biplot representan las medias de esas filas y columnas son, respectivamente:*

$$\bar{a}_f = \frac{1}{n_f} \sum_{f \in F} a_f \qquad \text{y} \qquad \bar{b}_c = \frac{1}{n_c} \sum_{c \in C} b_c$$

(medias de los marcadores respectivos)».

P3 - Fórmulas de Transición

Conociendo los marcadores de las filas podemos obtener los marcadores de las columnas y recíprocamente, usando las fórmulas de transición siguientes:

$$B = Y^T A \Sigma^{-1}$$
$$A = Y B \Sigma^{-1}$$

Demostración (GALINDO (1985))

De

$$Y = U \Sigma V^T$$
$$Y^T Y = \left(U \Sigma V^T\right)^T U \Sigma V^T = V \Sigma^2 V^T$$

Entonces

$$Y^T Y V = V \Sigma^2$$
$$Y^T Y V \Sigma^{-1} = V \Sigma = B$$

Pero

$$Y V = U \Sigma = A$$

Por lo tanto,

$$B = Y^T A \Sigma^{-1}$$

Del mismo modo, de:

$$Y = U \Sigma V^T$$
$$Y Y^T = \left(U \Sigma V^T\right)\left(U \Sigma V^T\right)^T = U \Sigma^2 U^T$$

Resulta sucesivamente que:

$$YY^T U = U \Sigma^2$$

$$YY^T U \Sigma^{-1} = U \Sigma = A$$

Pero

$$Y^T U = V \Sigma = B$$

Por lo tanto :

$$A = Y B \Sigma^{-1}$$

Por ejemplo - GALINDO (1985) - en la figura siguiente, se conocen las coordenadas ($a_{1,\alpha}$, …, $a_{5,\alpha}$) de los cinco individuos en el eje α, a que corresponde el valor singular $\sigma_\alpha = \sqrt{\lambda_\alpha}$.

Los valores y_{ij} de la variable j sobre los individuos i= 1, 2, 3, 4, 5 están representados en esa figura por trazos verticales proporcionales a esos valores.

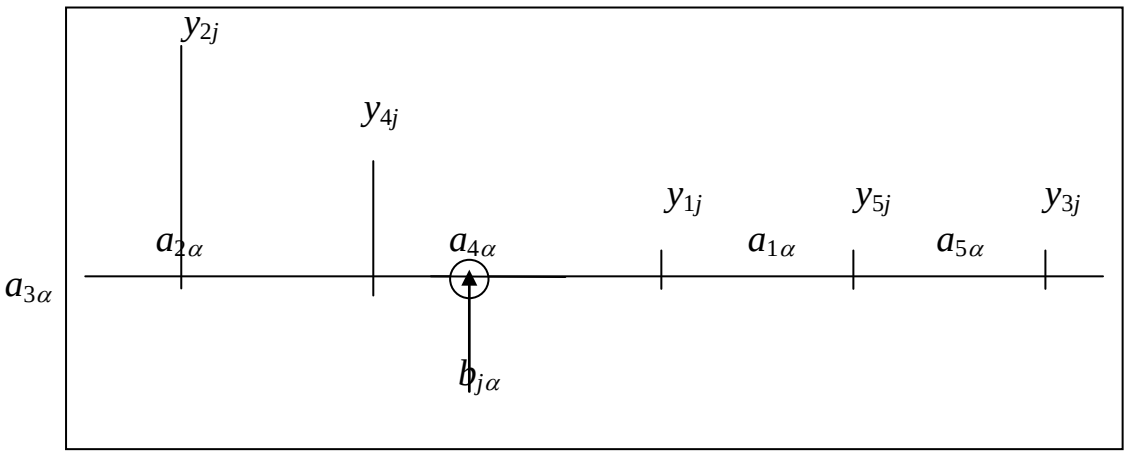


Figura 2.3.2.1. Ordenadas: Coordenadas de 5 individuos sobre el eje α.
Abscisas: Valores de la variable j para los 5 individuos.

La coordenada $b_{j\alpha}$ de la variable *j* en el eje α se obtiene multiplicando las coordenadas $a_{i\alpha}$ (*i*= 1, ..., 5) por los valores que la variable *j* tiene para cada uno de los objetos.

$$b_{j\alpha} = \frac{1}{\sqrt{\lambda_n}} \left(y_{1j} \times a_{1\alpha} + ... + y_{nj} \times a_{n\alpha} \right)$$

$a_{1\alpha}$ → Coordenada del individuo *i* en el eje α

y_{1j} → Valor de la variable *j* en el individuo *i*

Esto significa que la coordenada $b_{j\alpha}$ de la variable j en el eje α tiene tendencia a acercarse a las coordenadas de los objetos para los que esa variable tiene valores elevados.

En el gráfico del HJ-biplot, el marcador de una variable tiene tendencia a acercarse a los marcadores de los objetos para los que toma valores más altos.

Del mismo modo, consideremos ahora un ejemplo con *p*= 3 variables cuyas coordenadas en el eje están marcadas en la **figura 2.3.2.2.** por ◯ .

En esa figura están representados por segmentos verticales los valores de las variables $y_{(1)}$, $y_{(2)}$, $y_{(3)}$ sobre el objeto *i*.

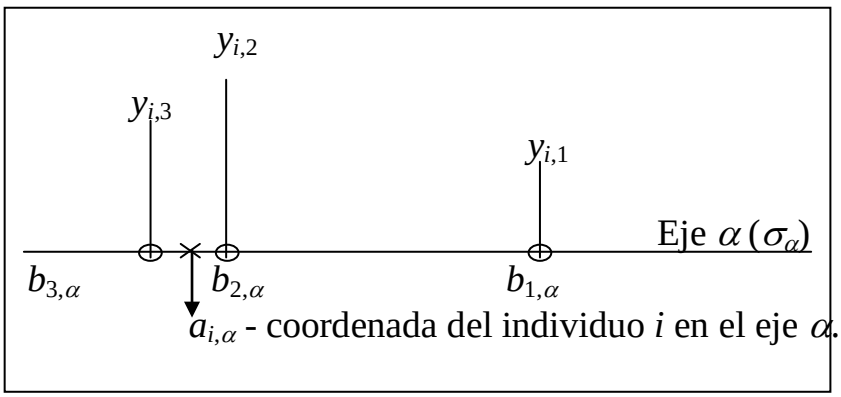


Figura 2.3.2.2. Ordenadas: coordenadas de las variables *j*= 1, 2, 3 en el eje α.
Abscisas: valores de las variables *j*= 1, 2, 3 para el mismo individuo *i*.

Ahora, la coordenada del individuo i se acerca a las coordenadas de las variables que tienen preponderancia (valores más grandes) sobre ese individuo.

En el gráfico, el marcador del individuo i estará cercano a los marcadores de las variables que tienen grandes valores sobre ese individuo.

2.4. INTERPRETACIÓN DE LOS BIPLOTS.

2.4.1. INTRODUCCIÓN.

Interpretar un biplot es expresar el significado de ese biplot en lenguaje geométrico, estadístico o en un lenguaje más cercano al lenguaje humano.

Algunos autores - GOWER *et al* (1996), por ejemplo - consideran que un biplot debe presentar puntos para representar los individuos y *ejes* para representar las variables. El biplot es visto como una generalización del concepto de gráfico de dispersión, correspondiente a más de dos variables ($p > 2$). En este caso, los biplots no tienen explícitamente marcados los ejes principales asociados a la descomposición en valores y vectores singulares.

Esto tiene tanto más sentido cuando se considera que es posible construir biplots usando otras técnicas - como lo admite GABRIEL (1971).

Sin embargo, una vez que la técnica más usada para construir biplots es la descomposición de la matriz de datos X en vectores y valores singulares, es natural que los ejes factoriales sean marcados en los gráficos de biplots

obtenidos según esa técnica particular y que sean presentados métodos para interpretación no sólo de esos ejes sino también para la interpretación de los planos por ellos definidos.

2.4.2. CALIDAD DE LA REPRESENTACIÓN.

Consideremos los biplots de GABRIEL indexados por $\alpha \in [0, 1]$ en la descomposición:

$$Y = \left(U\,\Sigma^{\alpha}\right)\left(V\,\Sigma^{1-\alpha}\right)^{T} = A\,B^{T} \quad \text{con} \quad \alpha \in [0,1]$$

Cuando se construye un biplot usando un espacio de representación de dimensión $d = 2 < r$= rango(Y), la matriz $Y_{(2)}$ es una aproximación inferior de Y.

$$Y_{(2)} = \left(U_{2}\,\Sigma_{2}^{\alpha}\right)\left(V_{2}\,\Sigma_{2}^{1-\alpha}\right)^{T} = A_{(2)}\,B_{(2)}^{T}$$

La variabilidad, inercia, o información contenida en Y es:

$$\|Y\|^{2} = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}^{2} = tr\left(Y^{T}\,Y\right) = tr\left(Y\,Y^{T}\right) = \sum_{j=1}^{r} \sigma_{i}^{2} = \sum_{i=1}^{r} \lambda_{i}$$

Hay que comparar esta información o variabilidad de los datos con la variabilidad explicada por la variabilidad de los marcadores de las variables y de los individuos en la representación biplot.

En los biplots de Gabriel, los individuos están representados por marcadores

$$a_{i} = \begin{bmatrix} a_{i1} \\ \dots \\ a_{id} \\ \dots \\ a_{ir} \end{bmatrix} = \begin{bmatrix} u_{i1}\,\sigma_{1}^{\alpha} \\ \dots \\ u_{id}\,\sigma_{d}^{\alpha} \\ \dots \\ u_{ir}\,\sigma_{r}^{\alpha} \end{bmatrix} \in R^{r}$$

y las variables están representadas por marcadores

$$b_j = \begin{bmatrix} b_{j1} \\ \dots \\ b_{jd} \\ \dots \\ b_{jr} \end{bmatrix} = \begin{bmatrix} v_{j1}\,\sigma_1^{1-\alpha} \\ \dots \\ v_{jd}\,\sigma_d^{1-\alpha} \\ \dots \\ v_{jr}\,\sigma_r^{1-\alpha} \end{bmatrix}$$

En el caso de biplots bidimensionales (d= 2) (ver **figura 2.4.2.1.**)

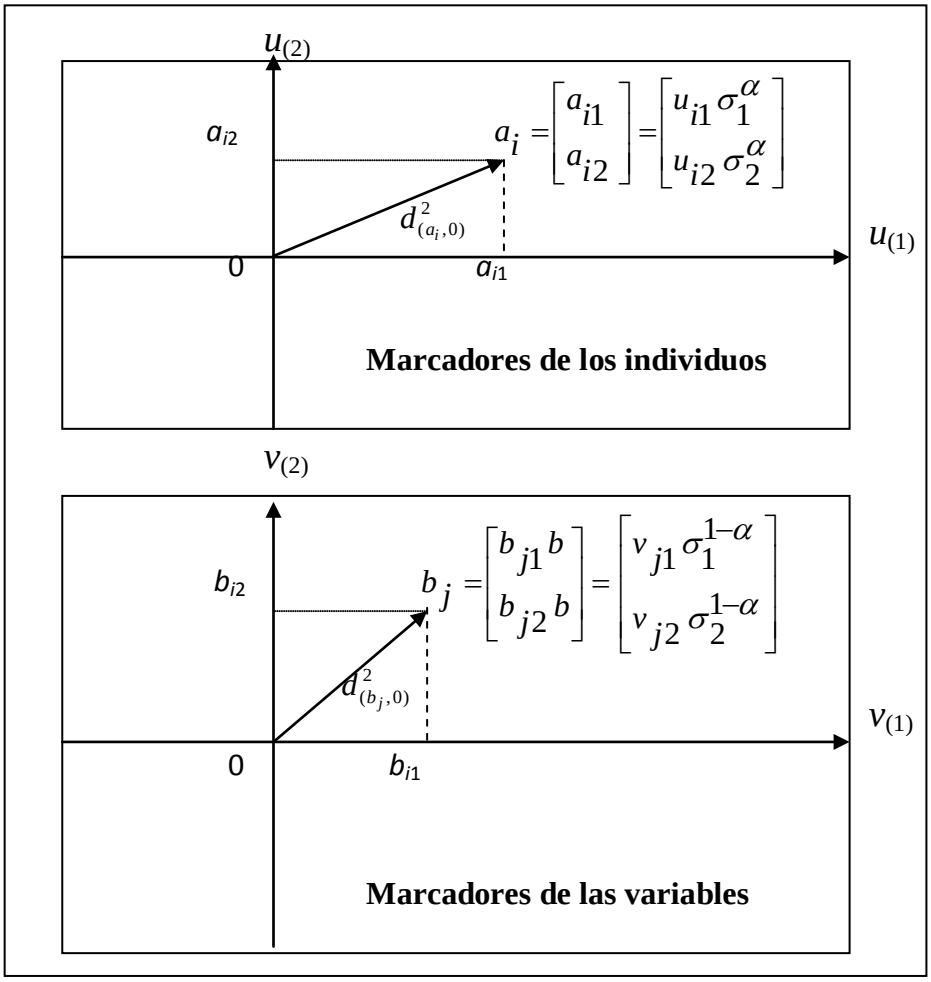


Figura 2.4.2.1. La variabilidad de los marcadores en la representación biplot se obtiene sumando los cuadrados de las distancias de esos marcadores al centro.

La variabilidad total de los marcadores de los individuos es la suma de los cuadrados de las distancias de a_i al marcador medio o centro; lo mismo

ocurre con la variabilidad total de los marcadores de las variables, con relación al respectivo marcador medio u origen.

Para un individuo,

$$d^2\left(a_i,0\right)=d^2\left(x_i,CG_I\right)=u_{i1}^2\,\sigma_1^{\,\alpha}+u_{i2}^2\,\sigma_2^{\,\alpha}=u_{i1}^2\,\lambda_1^{\alpha}+u_{i2}^2\,\lambda_2^{\alpha}$$

Sumando para todos los individuos (**inercia o varianza** total):

$$\sum_{i=1}^{n}d^2\left(a_i,0\right)=\sum_{i=1}^{n}u_{i1}^2\,\lambda_1^{\alpha}+\sum_{i=1}^{n}u_{i2}^2\,\lambda_2^{\alpha}=\lambda_1^{\alpha}\sum_{i=1}^{n}u_{i1}^2+\lambda_2^{\alpha}\sum_{i=1}^{n}u_{i2}^2$$

Pero de $\left\|u_{(j)}\right\|^2=1$ resulta $\sum_{i=1}^{n}d^2\left(a_i,0\right)=\lambda_1^{\alpha}+\lambda_2^{\alpha}$.

Para las variables, por un razonamiento similar:

$$\sum_{j=1}^{p}d^2\left(b_j,0\right)=\lambda_1^{1-\alpha}\sum_{j=1}^{p}v_{j1}^2+\lambda_2^{1-\alpha}\sum_{j=1}^{p}v_{j2}^2=\lambda_1^{1-\alpha}+\lambda_2^{1-\alpha}$$

En conclusión: para los biplots de Gabriel, la calidad de la representación de los individuos es

$$\frac{\lambda_1+\lambda_2}{\sum_{i=1}^{r}\lambda_i}\qquad\alpha\in[0,1]$$

La calidad de representación para las variables es

$$\frac{\lambda_1^{1-\alpha}+\lambda_2^{1-\alpha}}{\sum_{i=1}^{r}\lambda_i}\qquad\alpha\in[0,1]$$

Casos particulares:

α= 0 (CMP), la calidad es:

$$\frac{2}{\sum_{i=1}^{r}\lambda_i} \text{ para los individuos.}$$

(Si los datos han sido reducidos, la calidad es $\frac{2}{p}$, una vez que $\sum \lambda_i = p$).

En este caso, las variables están representadas con calidad $\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{r}\lambda_i}$.

y, en el caso reducido, la calidad es $\frac{\lambda_1 + \lambda_2}{p}$.

Cuando $\alpha = \frac{1}{2}$, la calidad de representación de individuos y variables es la misma y vale

$$\frac{\lambda_1^{1/2} + \lambda_2^{1/2}}{\sum_{i=1}^{r}\lambda_i} \qquad \left(\frac{\lambda_1^{1/2} + \lambda_2^{1/2}}{p} \text{ en caso de reducción} \right).$$

La calidad es la misma pero no tan buena como la que es posible obtener para individuos y variables por separado:

$$\frac{\lambda_i + \lambda_2}{\sum \lambda_i}$$

Para los biplots de GALINDO una vez que los marcadores de los individuos son las filas de

$A = U\,\Sigma$

y los marcadores de las variables son las filas de

$$B = U\Sigma,$$

todas las consideraciones hechas para los individuos en el biplot de Gabriel son válidas para los de GALINDO desde que se substituya $\alpha = 1$ en las expresiones respectivas.

Del mismo modo, las consideraciones hechas para las variables en el biplot de Gabriel son válidas para el de GALINDO desde que en las expresiones se tenga $\alpha = 0$.

Específicamente, para los individuos, la variabilidad de los marcadores en la representación de GALINDO vale:

$$\lambda_1^{\alpha} + \lambda_2^{\alpha} = \lambda_1 + \lambda_2 \quad (\alpha = 1)$$

Para los individuos, esa variabilidad es exactamente la misma:

$$\lambda_1^{1-\alpha} + \lambda_2^{1-\alpha} = \lambda_1 + \lambda_2 \quad (\alpha = 0)$$

De aquí resulta que, en este biplot, individuos y variables están representados en el plano con la calidad de representación máxima e igual para los dos:

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{r} \lambda_i}$$

Sobre este asunto, puede verse el reciente estudio comparativo de GABRIEL (2002), acerca de los distintos tipos de biplots.

2.4.3. INTERPRETACIÓN DE LOS EJES.

Para la interpretación de los ejes se usa, en general, el lenguaje de la escuela francesa desarrollada en análisis factorial de correspondencias, basada en los conceptos de contribución (absoluta y relativa). Ver BENZÉCRI (1973, 1992), JAMBU (1991), GALINDO (1985), GALINDO y CUADRAS (1986). En lo que sigue, α representa el índice de la dimensión y no el grado de RMP-CMP de los biplots de Gabriel.

En lo que respeta a la interpretación de los ejes, las consideraciones son comunes a los dos tipos de biplots.

Recordando que INERCIA ≡ VARIANZA, sean:

$A_\alpha(i)$= Coordenada del individuo i en el eje factorial α (α= 1, 2,…,r).

= Lo mismo que $a_{i\alpha}$ (i= 1 ,…, n; α= 1 ,…, r).

$B_\alpha(j)$= Coordenada de la variable j en el eje factorial α (α= 1, 2, …, r).

= Lo mismo que $b_{j\alpha}$ (j= 1 ,…, p; α= 1 ,…, r).

Inercia (varianza) total de la nube de individuos

$$= tr\left(Y\,Y^T\right) = tr\left(Y^T\,Y\right) = \sum_{\alpha=1}^{r} \lambda_\alpha = \sum_{\alpha=1}^{r} \sigma_\alpha^2 = \sum_{j=1}^{p} d^2\left(b_j, 0\right)$$

$$= \sum_{i=1}^{p} d^2\left(y_{(j)}, 0\right) = \sum_{\alpha=1}^{r}\sum_{j=1}^{p} b_{j\alpha}^2 = \sum_{\alpha=1}^{r}\sum_{j=1}^{p} B_\alpha^2(j)$$

$$= \sum_{i=1}^{n} d^2\left(y_i, 0\right) = \sum_{\alpha=1}^{r}\sum_{i=1}^{n} a_{i\alpha}^2 = \sum_{\alpha=1}^{r}\sum_{j=1}^{n} A_\alpha^2(i)$$

= Inercia (Varianza) total de la nube de variables

$CAE_i\ F_\alpha = A_\alpha^2(i) \quad (i = 1, ..., n)$ Son las contribuciones absolutas de los individuos para la inercia (varianza) del eje α (α= 1, …, r).

$CAE_j\ F_\alpha = B_\alpha^2(j) \quad (j = 1, ..., p)$ Son las contribuciones absolutas de las variables para la inercia (varianza) del eje α (α= 1, …, r).

$\sum_{i=1}^{n} A_\alpha^2(i) = \sum_{i=1}^{n} CAE_j\ F_\alpha = \lambda_\alpha = \sigma_\alpha^2$ Inercia total del factor α obtenida considerando las contribuciones de todos los individuos.

$\sum_{j=1}^{p} B_\alpha^2(j) = \sum_{j=1}^{p} CAE_j\ F_\alpha = \lambda_\alpha = \sigma_\alpha^2$ Inercia total del factor α obtenida considerando las contribuciones de todos las variables.

$CRE_i\ F_\alpha = \dfrac{CAE_i\ F_\alpha}{\lambda_\alpha} \quad (\alpha = 1, ..., r)$ Contribución relativa del elemento *i* para el factor α.

$CRE_j\ F_\alpha = \dfrac{CAE_j\ F_\alpha}{\lambda_\alpha} \quad (\alpha = 1, ..., r)$ Contribución relativa de la variable *j* para el factor α.

$$CRF_\alpha\ E_i = \frac{A_\alpha(i)}{\sum_{\alpha=1}^{r} A_\alpha(i)} = \frac{A_\alpha^2(i)}{d^2(a_i, 0)} = \cos^2(\theta_i)$$

= Contribución relativa del factor α (α= 1 ,…, r) para el elemento *i*.

$$CRF_\alpha\ E_j = \frac{B_\alpha^2(j)}{\sum_{\alpha=1}^{r} B_\alpha^2(j)} = \frac{B_\alpha^2(j)}{d^2(b_j, 0)} = \cos^2(\beta_j)$$

= Contribución relativa del factor α (α= 1 ,…, r) para la variable *j*.

En la **figura 2.4.3.1.** puede verse el significado de estas cantidades referente al marcador de un individuo.

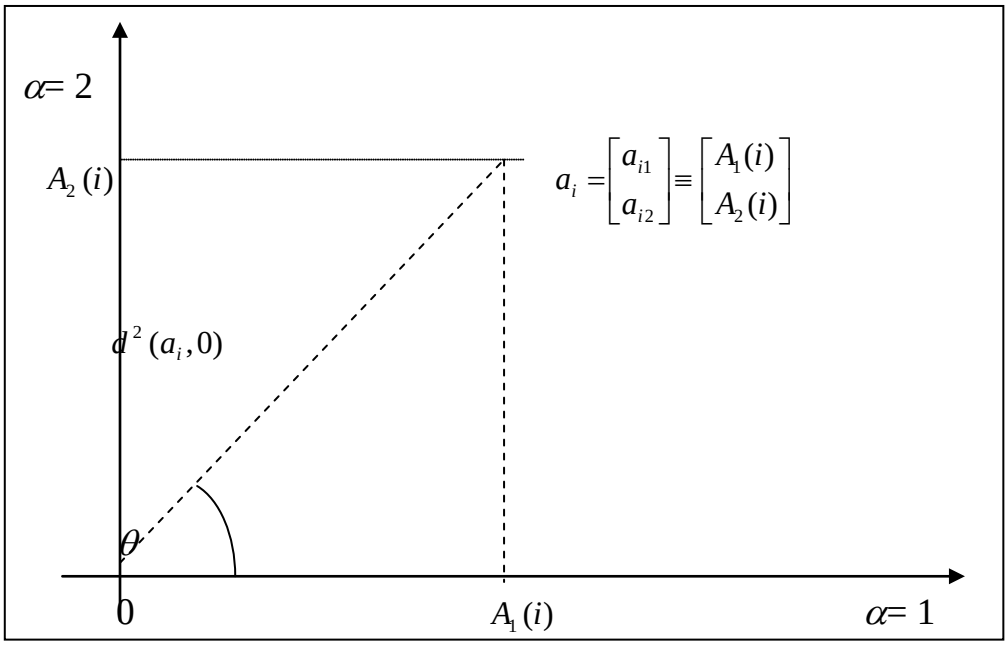


Figura 2.4.3.1. Interpretación del marcador de un individuo.

El significado de estas contribuciones es el siguiente:

$CRE_i F_\alpha$ / $CRE_j F_\alpha$ — En que medida la formación del factor α puede ser explicada por el individuo *i* o por la variable *j*, respectivamente.

$CRF_\alpha E_i$ / $CRF_\alpha E_j$ — En que medida el significado del factor α está asociado al significado del individuo *i* o de la variable *j*, respectivamente.

Puede ocurrir que un individuo - o variable - contribuya significativamente para la formación del factor pero que no sea interesante para interpretar el factor.

Por ejemplo, en la figura siguiente, el elemento *j* tiene una gran contribución relativa para el factor correspondiente a α= 1, pero el

elemento *j*' es preferible a la hora de buscar el significado de F_1 que debe estar más relacionado con el significado de *j*' que con el de *j*.

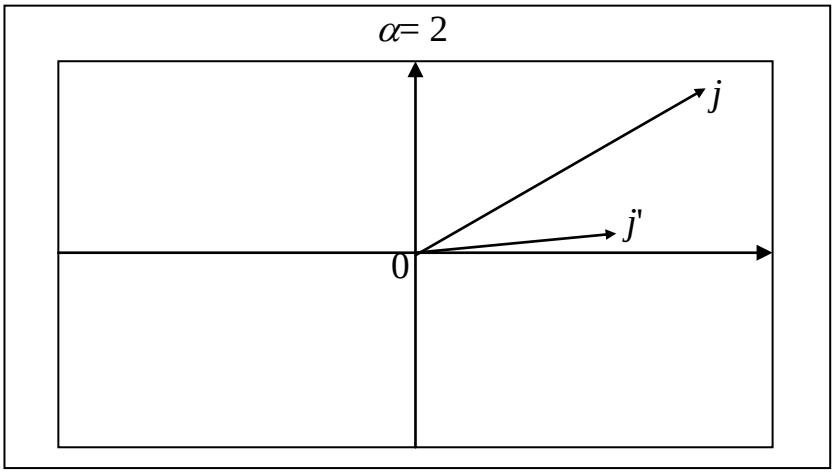


Figura 2.4.3.2. El factor α= 1 tiene un significado más próximo de *j*' que de *j*, aunque *j* tenga una contribución relativa más grande.

2.4.4. INTERPRETACIÓN DE ÃNGULOS Y DISTANCIAS ENTRE MARCADORES.

Como se ha visto en los apartados anteriores, en un biplot CMP (α= 0) son las variables las que están mejor representadas tanto del punto de vista geométrico como estadístico: los cosenos de los ángulos entre vectores representativos de las variables son exactamente las correlaciones entre variables; las distancias entre puntos representativos de los individuos son, aproximadamente, las disimilitudes entre individuos. Variables con vectores formando ángulos pequeños son variables con comportamientos muy similares; marcadores de individuos próximos corresponden a individuos semejantes, marcadores de individuos muy alejados en el gráfico corresponden a individuos disimilares.

La proyección de un individuo en la dirección de una variable nos permite conocer el valor de la variable para el individuo: individuos con altos **valores para esa** variable y recíprocamente.

La proyección de una variable sobre un individuo ($b_j^T a_i = y_{ij} = a_i^T b_j$) significa la importancia de la variable en la representación del individuo.

Si entre una variable y un individuo existe un ángulo pequeño eso significa que el individuo es significativo para explicar la variable y que la variable tiene un gran valor sobre el individuo.

En los biplots de Gabriel, la distancia entre individuos y variables no tiene significado especial - una vez que las nubes de individuos y variables están en escalas distintas (escala 1 para los individuos e σ_f (f= 1, …, r) para las variables).

En el RMP se verifica que las disimilitudes - medidas por la métrica euclidiana - están muy bien representadas por las distancias entre marcadores de individuos pero que las correlaciones entre variables solo están representadas de modo aproximado por los ángulos entre los marcadores respectivos. Desde el punto de vista de una inspección visual del biplot, los razonamientos son los mismos.

Cuando $\alpha = \frac{1}{2}$ la calidad de representación para individuos y variables es la misma pero no es el máximo que puede ser conseguido separadamente para variables e individuos.

EN EL BIPLOT DE GALINDO:

En este biplot, tanto variables como individuos están representados por sus respectivas coordenadas sobre los ejes principales de inercia (varianza máxima) correspondientes a las componentes principales de $Y^T Y$ (o $Y Y^T$).

En efecto,

$$Y^T Y = \mathrm{V}\, \Sigma^2 V^T = V \Sigma (V \Sigma)^T = B B^T$$

$$Y^T Y = \mathrm{U}\, \Sigma^2 U^T = U \Sigma (U \Sigma)^T = A A^T$$

y

$Y V = U \Sigma = A$ (Proyección de las filas sobre las direcciones principales)

$Y^T U = V \Sigma = B$ (Proyecciones de las columnas sobre las direcciones principales).

Por lo tanto, para cada eje, las escalas utilizadas, tanto para individuos como para variables, son las mismas: σ_f (f= 1 ... r).

Por lo que hemos visto, las distancias entre marcadores de individuos se interpretan como disimilitudes entre individuos: marcadores de individuos semejantes están próximos en el biplot; marcadores de individuos disimilares están alejados.

Los cosenos de los ángulos entre marcadores de variables representan correlaciones entre las variables: ángulos pequeños representan correlaciones elevadas; ángulos grandes representan correlaciones pequeñas.

Las fórmulas de transición permiten afirmar que el marcador de un individuo tiende a situarse cerca de los marcadores de las variables que son preponderantes para ese individuo; el marcador de una variable tiende a

acercarse a los marcadores de los individuos sobre los cuales tiene valores más elevados.

C A P Í T U L O III

PAPEL CENTRAL DE LOS MÉTODOS BIPLOT EN ANÁLISIS PRELIMINARES DE DATOS

3.1. INTRODUCCIÓN.

El objetivo del capítulo es, en síntesis, presentar los argumentos que justifican la elección de los métodos de biplot para núcleo de un sistema de minería de datos.

Como se ha visto en el capítulo – ver KLÖSGEN (2002) y BRADLEY, *et al*.(1998) - las técnicas a usar deben tener las características siguientes:

Escalabilidad: aplicabilidad a datos de pequeña y gran dimensión.
Incrementalidad: los resultados deben poder ser actualizados sin repetir los cálculos cuando se obtienen nuevos datos.
Interpretabilidad: los resultados deben poder ser fácilmente interpretados por el analista humano.

En nuestro proyecto se atribuye importancia crucial a la **interpretabilidad**, una vez que el sistema que se pretende desarrollar debe ser interactivo y colaborar con el usuario en la búsqueda de interpretaciones.

La interpretabilidad tiene que ver no solo con la expresión fácil de dos resultados en lenguaje próximo del lenguaje humano sino también con la posibilidad de comparar resultados obtenidos por métodos distintos.

En este capítulo se recuerda que los métodos biplot tienen una base teórica común con muchos de los métodos de análisis de datos multivariantes más importantes; incluso, como demostró VICENTE VILLARDÓN (1992), familias importantes de métodos como las componentes principales, análisis factorial de correspondencias y análisis canónico pueden verse como casos particulares de biplots generalizados.

Este hecho permite presentar los resultados obtenidos con esos métodos como configuraciones de puntos en biplots y, por lo tanto, un relacionamiento fácil de esos resultados, a la hora de realizar su interpretación.

También se va a verificar, usando ejemplos, que para otros métodos - como el MDS y la Clasificación - los resultados siguen expresándose fácilmente como configuraciones de marcadores en biplots.

Todo esto prueba la interpretabilidad de las técnicas biplot.

La estructura del capítulo es la siguiente: para cada familia de métodos de análisis de datos multivariantes serán referidos resultados de la investigación teórica revisada que se consideran relevantes para el doble objetivo de relacionar la teoría de los biplots con la teoría de esos métodos y para mostrar que los resultados respectivos pueden ser interpretados visualmente en biplots adecuados.

Para los métodos más conocidos se presentan ejemplos, todos basados en los mismos datos de la **tabla 3.1.1,** en donde las ideas anteriores son ilustradas, buscando preparar el terreno para las ideas a desarrollar en los capítulo IV y V.

Esos datos son los resultados reales obtenidos en una disciplina de Estadística por 58 alumnos de un curso de Psicología.

COLUMNA	**SÍMBOLO**	**SIGNIFICADO**
Número	Número	Identificación del alumno
Turno	Turno	D - Día N - Noche
X_1	G1A	Valoración (0, 1, 2) obtenida en la pregunta 1a), relativa a histogramas.
X_2	G1B	Valoración (0, 1, 2) obtenida en la pregunta 1b), relativa a frecuencias relativas.
X_3	G2A	Valoración (0, 1, 2) obtenida en la pregunta 2a), relativa a la distribución normal (probabilidad).
X_4	G2B	Valoración (0, 1, 2) obtenida en la pregunta 2b), relativa a la distribución normal (cuantiles).
X_5	G3A	Valoración (0, 1, 2) obtenida en la pregunta 3a), relativa a gráficos de dispersión.
X_6	G3B	Valoración (0, 1, 2) obtenida en la pregunta 3b), relativa al concepto de correlación.
X_7	G3C	Valoración (0, 1, 2) obtenida en la pregunta 3c), relativa al concepto de regresión.
X_8	G4A	Valoración obtenida (0, 1, 2) en la pregunta 4a), relativa a tablas de contingencia (frecuencias conjuntas).
X_9	G4B	Valoración (0, 1, 2) obtenida en la pregunta 4b), relativa a tablas de contingencia (esperanza marginal).
X_{10}	G4C	Valoración (0, 1, 2) obtenida en la pregunta 4c), relativa a tablas de contingencia (probabilidades condicionales).
X_{11}	TOTAL	Valoración (0, 1, 2, ..., 20) obtenida en el test. SUMA: $X_1 + ... + X_{10}$

Tabla 3.1.1. Significado de las columnas.

NÚMERO	***TURNO***	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	D	2	2	2	0	2	2	2	2	1	2	16
2	D	1	0	2	0	1	0	0	0	0	0	4
3	D	2	1	2	0	1	0	0	0	0	0	6
4	D	2	2	2	0	1	2	0	1	0	1	10
5	D	1	0	2	0	2	2	2	0	2	2	13
6	D	2	1	2	1	2	2	2	2	2	2	18
7	D	2	0	2	1	2	2	2	1	0	1	14
8	D	1	0	0	0	2	2	0	0	0	0	6
9	D	2	2	2	2	1	2	2	2	2	2	20
10	D	1	1	0	0	2	2	0	0	0	0	6
11	D	2	1	2	2	2	2	2	2	2	2	19
12	D	1	0	1	0	2	2	2	1	0	1	10
13	D	1	0	2	0	0	2	0	0	0	0	5
14	D	2	1	0	1	1	2	2	1	2	1	12
15	D	2	1	2	2	2	2	2	0	1	0	13
16	D	2	2	0	0	0	0	0	0	0	0	4
17	D	1	0	0	0	1	0	0	0	0	0	2
18	D	2	1	2	1	2	1	1	0	0	0	9
19	D	2	1	1	0	1	1	0	0	0	0	6
20	D	2	0	2	2	2	2	2	1	2	0	13
21	D	2	2	2	2	2	2	2	2	2	2	20
22	D	1	0	0	0	1	1	0	0	0	0	3
23	D	2	0	0	0	2	1	0	1	0	0	6
24	N	1	0	2	1	1	2	1	0	0	1	8
25	N	0	1	2	0	0	0	1	1	2	1	8
26	N	2	1	0	0	0	1	2	1	1	1	9
27	N	2	2	2	0	1	2	1	0	0	0	10
28	N	2	1	1	0	1	1	2	2	2	0	10
29	N	2	2	0	0	0	1	1	1	2	1	9
30	N	2	0	0	0	0	0	0	0	0	0	2
31	N	2	1	0	0	1	2	2	0	0	0	8
32	N	2	0	2	2	0	1	0	0	0	0	7
33	N	2	0	0	0	1	1	2	0	0	0	6
34	N	1	1	1	0	1	0	0	0	0	0	4
35	N	1	0	0	0	0	0	0	0	0	0	1
36	N	2	2	2	0	1	0	0	0	0	0	6
37	N	2	1	1	0	0	0	0	1	2	2	8
38	N	1	0	1	0	1	0	0	0	0	0	3
39	N	1	0	0	0	0	0	1	0	0	0	1
40	N	1	0	0	0	0	0	0	0	0	0	1
41	N	1	0	2	0	0	1	2	0	0	1	6
42	N	1	1	0	0	0	0	0	0	0	0	2
43	N	2	2	2	1	1	1	0	0	0	0	9
44	N	1	0	0	0	1	0	0	1	0	1	5
45	N	2	2	2	2	1	2	0	1	0	1	12
46	N	2	0	0	0	0	0	0	0	0	0	2
47	N	2	2	2	1	1	2	2	2	2	2	18
48	N	2	1	1	1	2	2	2	0	0	1	11
49	N	2	0	0	0	0	0	0	0	0	0	2
50	N	2	2	2	2	1	2	2	1	1	1	16
51	N	2	2	2	2	1	2	0	1	0	0	12
52	N	2	0	0	0	2	2	0	0	0	0	6
53	N	2	0	0	0	0	0	0	0	0	0	2
54	N	2	2	2	0	2	1	1	1	0	2	12
55	N	2	2	0	0	1	0	0	2	0	2	8
56	N	2	2	0	0	1	1	0	1	0	1	8
57	N	1	2	0	0	1	2	2	0	0	0	7
58	N	2	2	1	2	2	1	1	1	2	1	14

Tabla 3.1.1.(cont) Valoración de las 10 respuestas a un test de estadística y valoración global.

3.2. BIPLOTS Y MÉTODOS FACTORIALES.

3.2.1. SÍNTESIS TEÓRICA.

GABRIEL (1971), en el artículo fundador de la teoría de los biplots observa, en el apartado 1, que los espacios vectoriales generados por las filas y las columnas de la matriz de datos *X* pueden ser dotados de métricas que garanticen la unicidad de los biplots.

Si el espacio vectorial generado por las filas x_1, …, x_i, …, x_n de la matriz *X* fuera dotado de una geometría definida por una métrica de matriz definida positiva simétrica Φ, y el espacio vectorial generado por las columnas $x_{(1)}$, …, $x_{(j)}$, …, $x_{(p)}$ de *X* fuera dotado de una geometría definida por una métrica dada por la matriz definida positiva Ω, podría escribirse:

Para las filas x_i e x_e de *X*:

$$x_i^T \, \Phi \, x_e = \left\langle x_i \, , x_e \right\rangle_\Phi$$

y, por lo tanto,

$$X \, \Phi \, X^T = \left[\left\langle x_i \, , x_e \right\rangle_\Phi \right]$$

Para las columnas $x_{(j)}$ y $x_{(k)}$ de *X*:

$$x_{(j)}^T \, \Omega \, x_{(k)} = \left\langle x_{(j)} \, , x_{(k)} \right\rangle_\Omega$$

y, por lo tanto,

$$X \, \Omega \, X = \left[\left\langle x_{(j)} \, , x_{(k)} \right\rangle_\Omega \right]$$

Descomponiendo *X* en valores y vectores singulares (SVD):

$$X = U \, \Sigma \, V^T$$

resulta:

$$X\,\Phi\,X^T = \left[\left\langle x_i\,,\,x_e\right\rangle_\Phi\right] = \left(U\,\Sigma V^T\right)\Phi\left(U\,\Sigma V^T\right) = U\,\Sigma V^T\,\Phi V\,\Sigma U^T$$

Para garantizar que esta descomposición es única, debe imponerse la condición:

$$V^T\,\Phi V = \left[< v_{(j)}^T\,\Phi\,v_{(k)} >\right] = \left[< v_{(j)}\,,v_{(k)} >_\Phi\right] = \left[< \delta_{jk} >\right] = I$$

siendo δ_{jk} el símbolo de Knonecker.

Si $V^T\,\Phi\,V = I$ entonces

$$X\;\Phi\;X^T = U\,\Sigma^2\,U^T$$

es única.

Del mismo modo, para las columnas:

$$\begin{aligned} X^T\,\Omega\,X &= \left[\left\langle x_{(j)}\,,\,x_{(k)}\right\rangle_\Omega\right] \\ &= \left(U\,\Sigma V^T\right)^T\,\Omega\left(U\,\Sigma V^T\right) \\ &= V\,\Sigma U^T\,\Omega U\,\Sigma V^T \end{aligned}$$

Si imponemos

$$U^T\,\Omega\,U = \left[\delta_{jk}\right] = I$$

entonces,

$$X^T\,\Omega\,X = V\,\Sigma^2\,V^T$$

es única.

De aquí resulta la idea - GREENACRE (1984) - de definir una descomposición en valores singulares generalizada (GSVD) de la matriz *X*:

$$X = U\,\Sigma V^T$$

s.a.

$$U^T\,\Omega U = I$$

$$V^T\,\Phi V = I$$

en donde Ω y Φ son matrices simétricas definida positivas. *U* y *V* contienen ahora, en columnas, los v*ectores singulares generalizados*.

Como lo verifica el propio GREENACRE (1984), si transformamos la matriz *X* en $\Omega^{1/2} X \Phi^{1/2}$, entonces la SVD de la matriz transformada coincide con la GSVD de la matriz original *X*.

O sea:

$$\Omega^{1/2} X \Phi^{1/2} = U \Sigma V^T$$

s.a.

$$U^T U = I$$
$$V^T V = I$$

en que $\Omega^{1/2}$ *y* $\Phi^{1/2}$ pueden obtenerse de las respectivas descomposiciones simples en valores y vectores singulares (SVD's).

Este hecho permite a GREENACRE (1984), mostrar que métodos factoriales importantes como el análisis en componentes principales y el análisis canónico pueden ser formulados usando el mismo referencial teórico que el análisis factorial de correspondencias. Ver GREENACRE (1984), Table A1, pg. 348.

Basándose en las propiedades de la GSVD, VICENTE-VILLARDÓN (1992), define los biplots de Gabriel generalizados (GCMP, GRMP) usando la definición siguiente:

> «*Un biplot generalizado de la matriz mediante marcadores* $g_1 \ldots g_n$ *para sus filas y marcadores* $h_1 \ldots h_p$ *para sus columnas de forma que el producto escalar* $g_i^T h_j$ *represente el elemento* x_{ij} *de la matriz de*

partida; la unicidad de las factorizaciones se consigue introduciendo métricas, distintas de las asociadas a la matriz identidad, tanto en el espacio de las filas como en el espacio de las columnas».

En VICENTE-VILLARDÓN (1992) puede verse que esta definición generaliza las propiedades correspondientes de los biplots de Gabriel. Del mismo modo, los biplots de Galindo - RCMP - pueden generalizarse introduciendo métricas para las filas y para las columnas.

VICENTE-VILLARDÓN (1992), introduce el GRCMP-biplot (designado originalmente por GHJ-biplot) con la definición siguiente - op. cit. pg. 94.

«*Llamaremos HJ-Biplot generalizado (GHJ-biplot) a una representación gráfica multivariante de las líneas de una matriz* $X_{n\times p}$ *mediante marcadores* $j_1 \ldots j_n$ *para sus filas y* $h_1 \ldots h_p$ *para sus columnas, elegidos de forma que se consiga una representación simultánea en sentido estricto, preservando una métrica cualquiera* Φ *otra métrica* Ω *introducida en el espacio de las columnas*».

Usando una notación más actualizada, donde X es la matriz original, con GSVD

$$X = U\,\Sigma V^T$$

s.a.

$$U^T\,\Omega U = I$$
$$V^T\,\Phi V = I$$

Los marcadores elegidos para las filas son, ahora, las filas de

$$A = U\,\Sigma$$

y los marcadores elegidos para las columnas son, ahora, las filas de

$$B = V\,\Sigma$$

En las expresiones anteriores, U y V son los vectores singulares generalizados. De aquí resulta que:

$$X\,\Phi X^{T} = U\,\Sigma^{2}\,U^{T} = U\,\Sigma\left(U\,\Sigma\right)^{T} = A\,A^{T}$$
$$X\,\Omega X^{T} = V\,\Sigma^{2}\,V^{T} = V\,\Sigma\left(V\,\Sigma\right)^{T} = B\,B^{T}$$

Usando esta definición VICENTE-VILLARDÓN (1992), demuestra que se preservan todas las propiedades del RCMP-biplot, incluyendo las fórmulas de transición que permiten pasar de las coordenadas de los individuos a las coordenadas de las variables y recíprocamente; lo que permite afirmar que, con esta nueva definición, se garantiza además, una representación conjunta.

Los biplots de Galindo se obtienen del biplot GRCMP/GHJ-biplot como el caso particular correspondiente a $\Omega = \Phi = I$.

Sea $F_{n\times p} = \left[f_{ij}\right]$ una tabla de contingencia, $r = F1_p$ un vector cuyas n componentes son las sumas de las filas de F, $c = F^T 1_p$ el vector de p componentes formadas por las sumas de las columnas de F, D_r = Diag(r) y D_c = Diag(c) las matrices diagonales formadas por las componentes de r y c, respectivamente.

VICENTE-VILLARDÓN (1992), pueba que:

Si en el espacio de filas se define una métrica de matriz $\Phi = D_e$ y en el espacio de columnas una métrica de matriz $\Omega = D_n$, la representación GHJ-biplot de la matriz $E = N^{1/2} D_r^{-1} F D_c^{-1}$ coincide con el Análisis Factorial de correspondencias de la matriz F, en donde $N = \sum_{i=1}^{I} \sum_{j=1}^{J} f_{ij}$.

O sea, si realizamos la GSVD de *E*, se tiene

$$E = U^T \Sigma V^T = A B^T$$

con

$$A = U \Sigma$$
$$B = V \Sigma$$

y

$$U^T \Omega U = 1$$
$$V^T \Omega V = 1$$

El biplot generalizado de Galindo tendría ahora por marcadores de las filas - las filas de $A = U \Sigma$ y por marcadores de las columnas las filas de $B = V \Sigma$, en donde *U* y *V* son los vectores singulares generalizados de *X*.

Pero realizar la GSVD de *E* es equivalente a realizar la SVD de $\Omega^{1/2} y \Phi^{1/2}$, con $\Omega = D_r$ y $\Phi = D_c$. Esto significa que realizar el análisis Factorial de Correspondencias de *F* equivale a realizar el RCMP de

$$D_r^{1/2} \left(N^{1/2} D_r^{-1} F D_c^{-1} \right) D_c^{1/2} = N^{1/2} D_r^{-1/2} F D_c^{-1/2}$$

Lo que equivale a usar como marcadores de las filas de *F* a las filas de $A = U \Sigma$ y como marcadores de las columnas de *F* las filas de $B = V\Sigma$, con

$$N^{1/2} D_r^{-1/2} F D_c^{-1/2} = U \Sigma V^T$$

y

$$U^T U = V^T V = I$$

Aún en VICENTE-VILLARDÓN (1992), puede verse que el Análisis de Correlación Canónica de *Hotelling* puede ser presentado como caso particular del GHJ-biplot.

En VAZQUEZ (1995) puede verse una formalización algebraica de los Biplots Generalizados vistos como Modelos Bilineales.

En síntesis: los trabajos de GABRIEL (1971), GALINDO (1985), GREENACRE (1988), VICENTE-VILLARDÓN (1992), permiten afirmar que las técnicas de análisis de datos de tipo factorial - todas ellas basadas en la DVS o la GDVS - son casos particulares de biplots generalizados o producen resultados que pueden ser expresados por biplots adecuados.

Para otra visión del concepto de biplot generalizado, ver GOWER (1995) y GOWER *et al* (1996).

En otras palabras, desde el punto de vista del objetivo central de esta tesis, las investigaciones referidas significan, al final, que todos los métodos importantes de análisis de tipo factorial pueden ser expresados gráficamente usando biplots.

3.2.2. ANÁLISIS EN COMPONENTES PRINCIPALES. PROBLEMAS DE INTERPRETACIÓN.

En el epígrafe anterior se ha visto que el análisis en componentes principales corresponde a un caso particular del biplot de Gabriel (GH / CMP).

Sobre este método puede consultarse, por ejemplo, LEBART *et al* (2002); JACKSON (1991); JOHNSON *et al* (1998); LEBART *et al,* (1998); RAO (1999); KRZANOWSKI (1998, 2000); JOLLIFFE (2002).

Las componentes principales son combinaciones lineales de las variables observadas, no correlacionadas, que explican la mayor parte de la información/varianza/inercia contenida en las matrices de covarianza o correlación de las variables observadas.

Si las variables observadas están poco correlacionadas, las componentes principales coinciden prácticamente con esas variables. En este caso, el problema de interpretación es trivial: las componentes principales tienen el mismo significado que las variables observadas.

Interpretar las componentes principales es buscar el significado de las nuevas variables, designadas componentes principales, en el dominio de donde han sido obtenidos los datos.

Esto puede realizarse de dos modos. Uno, es examinar los coeficientes de las variables observadas en la combinación lineal que forma la componente a interpretar; lo que permite asociar el significado de las componentes principales al significado de las variables observadas que más han contribuido para la formación de la componente principal - lo que equivale a formar grupos de variables asociadas a las componentes principales. Este

método tiene inconvenientes, como lo reconoce KRZANOWSKI (1979, 1998, 2000).

El otro modo consiste en representar los resultados del análisis en un biplot (GH/CMP u otro) en donde las componentes principales corresponden a los ejes principales.

El significado del eje a interpretar queda asociado al significado de los objetos y de las variables con mayor contribución relativa para ese eje. Esas variables e individuos pueden obtenerse, en casos simples, por inspección visual del gráfico.

Al final, hay que identificar y caracterizar dos grupos de individuos y variables que se oponen a lo largo del eje correspondiente a la componente principal a interpretar: el significado de la componente principal debe buscarse en esa oposición /discriminación.

Ejemplo 3.2.2.1.

Analizando los datos de la **tabla 3.1.1.** se obtiene el biplot de la **figura 3.2.2.1.**, correspondiente a un biplot de GABRIEL con α= 0.5, y matriz de datos centrada.

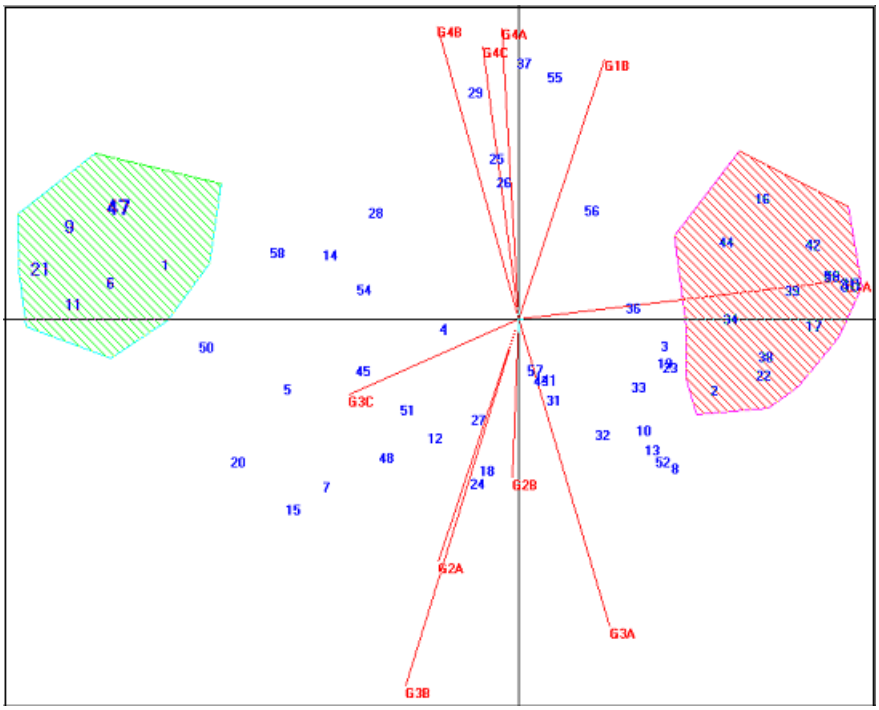


Figura 3.2.2.1. Biplot de Gabriel (α= 0.5) de los datos de la **tabla 3.1.1.**, excluida la variable TOTAL.

El significado del eje 1, correspondiente a la primera componente principal, puede ser explicado, interpretado, expresado por la oposición entre los significados del grupo de objetos pintados de verde – a la izquierda - y el significado del grupo de objetos pintados de rojo a la derecha. Al grupo verde esta asociada la variable G4B, relativa a tablas de contingencia. Al grupo rojo (derecha) esta asociada la variable G1A relativa a histogramas, en el ejemplo que nos ocupa y que fue descrito en las páginas 75 y 76.

Este eje separa los “buenos” alumnos (izquierda) de los “malos” alumnos (derecha).

VERDE= {1, 6, 9, 11, 21, 27, 47}

ROJO= {2, 16, 17, 22, 30, 34, 35, 38, 39, 40, 42, 44, 46, 49, 53}

En síntesis, la interpretación de una componente implica:

1. Proyectar los individuos sobre el eje e identificar los grupos que se oponen a lo largo de ese eje, elegidos de entre los mejor representados (mayor contribución relativa del factor al elemento).
2. Proyectar sobre el eje los marcadores de las variables identificando los grupos de variables relevantes para la caracterización de los grupos de individuos identificados.
3. Caracterizar los dos grupos separados por el eje.

Cada grupo de individuos identificado puede ser definido por su función característica δ_G o variable indicadora.

$$\delta_G(x_i) = 1 \quad \text{si } x_i \in G \qquad \delta_G(x_i) = 0, \text{ caso contrario.}$$

En alternativa, la componente principal a interpretar puede ser representada por una variable simbólica de valores {-, +, *} en donde "-" significa que un individuo está de un lado (izquierda, Verde), "+" significa que está del lado opuesto - Rojo, derecha; y "*" significa que no tiene una posición definida.

Síntesis: el problema de interpretación de una componente principal es equivalente al problema de interpretar una nueva variable cualitativa que divide el conjunto de puntos de un biplot en clases de equivalencias con etiquetas {-, +, *}.

3.2.3. ANÁLISIS FACTORIAL DE CORRESPONDENCIAS. PROBLEMAS DE INTERPRETACIÓN.

La técnica de Análisis Factorial de Correspondencias está descrita en BENZÉCRI (1973, 1992), JAMBU (1991), CUADRAS (1981), GREENACRE (1984), GREENACRE *et al* (1994), GREENACRE *et al* (1996), GOWER *et al* (1986), entre otros.

Se ha verificado en el epígrafe 3.2.1. que realizar un AFC equivale a realizar el RCMP-biplot de una matriz que se obtiene por transformación de la matriz de frecuencia inicial.

Tanto en BENZÉCRI (1973, 1992), como en JAMBU (1991), pueden verse reglas que sistematizan la interpretación de los resultados respectivos. En GALINDO *et al* (1996) puede verse una detallada exposición de reglas prácticas de interpretación de las representaciones simultáneas asociadas al AFC.

Una vez que el resultado fundamental de un AFC es un biplot - un HJ / RCMP-biplot - cuando se intenta interpretar el resultado de un AFC ocurren exactamente los mismos problemas que han sido identificados en el párrafo anterior a propósito del ACP. Si el problema es el de interpretar un eje, hay que identificar los grupos de individuos y variables con elevadas contribuciones absolutas y relativas para ese eje y que se oponen a lo largo del eje. El significado del eje hay que buscarlo en el significado de la oposición entre grupos.

Ejemplo 3.2.3.1.

Analizando un AFC de los datos de la **tabla 3.1.1.**, se obtiene la **figura 3.2.3.1.** en la que han sido identificados los grupos Verde y Rojo cuya oposición da el significado del eje factorial 1.

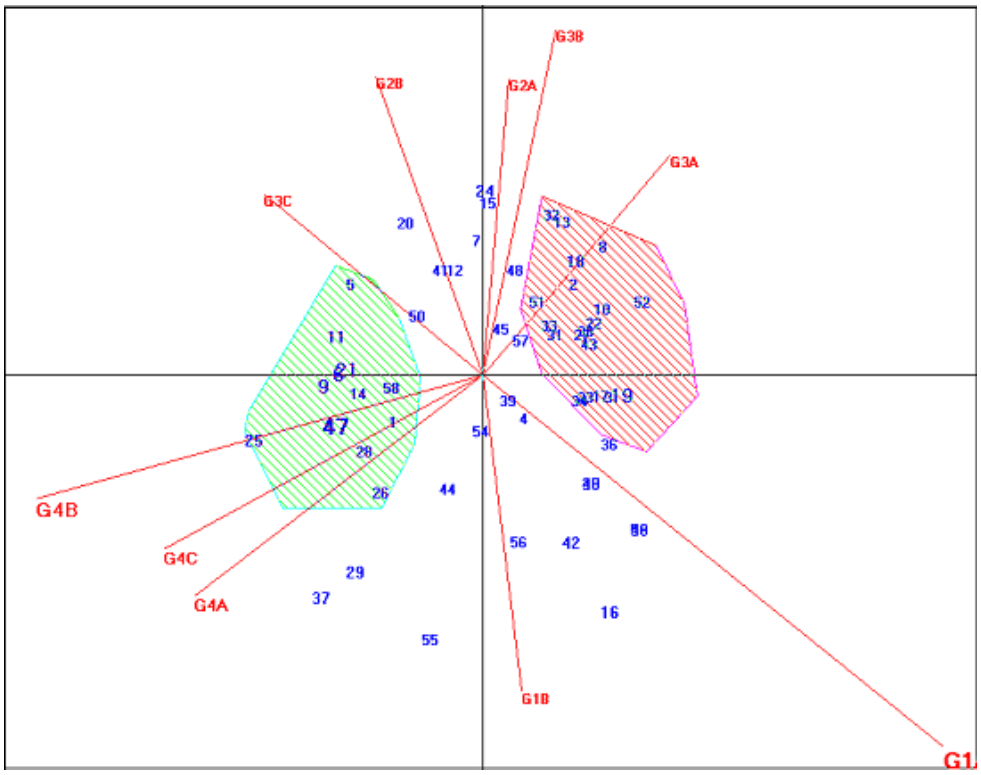


Figura 3.2.3.1. Biplot con el resultado de un AFC de los datos de la **tabla 3.1.1.**, excluida la variable TOTAL.

Verde= {1, 5, 6, 9, 11, 14, 21, 25, 26, 28, 47, 58}

Rojo= {2, 3, 8, 10, 13, 17, 18, 19, 22, 23, 27, 31, 32, 33, 34, 38, 43, 51, 52}

En síntesis: el problema de interpretación de un eje factorial en AFC es representable como el problema de interpretar una nueva variable cualitativa que divide el conjunto de individuos de un biplot en clases de equivalencia con etiquetas $\in$ {-, +, *}.

3.3. BIPLOTS vs MDS.

3.3.1. ASPECTOS TEÓRICOS.

Dada una matriz cuadrada $\Delta = [\delta_{ij}]$ formada por las disimilitudes observadas entre los $\frac{n \times (n-1)}{2}$ pares de n individuos, el objetivo del MDS es obtener una configuración $X_{n \times d}$ de puntos de un espacio euclídeo de dimensión d (en general d= 1, 2, 3) tal que las distancias euclideas entre los puntos de esa configuración (representando filas de X) reproduzcan «lo mejor posible» las disimilitudes originales. Ver CUADRAS (1981), COX *et al* (1994), KRZANOWSKI (2000), GOWER *et al* (1996), BORG *et al* (1997), MARTIN-CASADO (1992),entre otros. Ver **figura 3.3.1.1.**

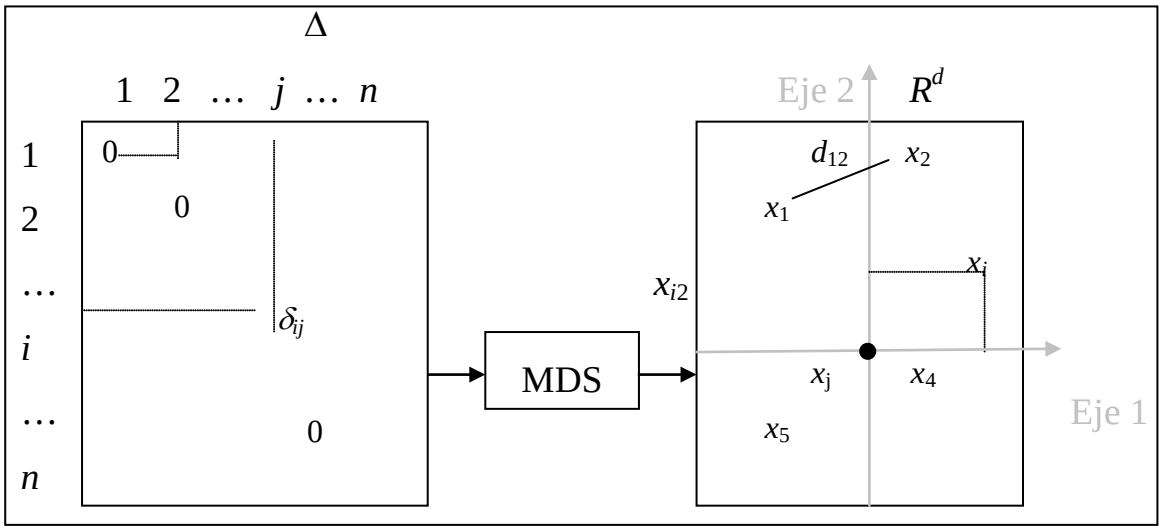


Figura 3.3.1.1. El MDS transforma las disimilitudes δ_{ij} entre individuos en distancias d_{ij} entre puntos de una configuración representada en un sistema de coordenadas.

Los datos de partida pueden ser, o no, distancias euclídeas pero no están referidos a un sistema de ejes, como ocurre con casi todos los otros métodos de análisis multivariante.

El resultado de un MDS es una configuración de marcadores $x_1 \ldots x_n$ (uno para cada individuo) en un espacio métrico. Esta configuración, por supuesto, está referida a un sistema de ejes formando un referencial. El referencial es uno de los resultados y no un dato de partida. Ver GOWER *et al* (1996).

La teoría del MDS está, desde sus orígenes, muy relacionada con la descomposición en valores y vectores singulares de ECKART *et al* (1936), en la que se basa el biplot. YOUNG y HOUSEHOLDER (1938) prueban, usando la SVD que, dadas las distancias entre *n* individuos, es siempre posible obtener las coordenadas de esos individuos en un espacio métrico de *n* - 1 dimensiones de tal modo que las distancias entre los puntos de la configuración reproduzcan exactamente las distancias de partida.

Si *A* es una matriz cuadrada simétrica formada por los elementos $\left[-\frac{1}{2}\delta_{ij}^2\right]$ y $B = H\ A\ H$ con $H = \left(I - \frac{1}{n}11^T\right)$ la matriz de centrado, entonces, realizando la SVD de *B*, se obtiene

$$B = V\,\Sigma^2\,V^T = V\,\Sigma\left(V\,\Sigma\right)^T = X\ X^T$$

con Σ = Matriz diagonal ($r \times r$) con los *r* valores singulares no nulos $\underset{n\times r}{X} = V\,\Sigma$.

Las filas $x_1 \ldots x_n$ de *X* son vectores de R^r en donde *r*= rango (*X*) ≤ *n* - 1.

Estas filas forman una configuración de n puntos en un espacio de dimensión $r \leq n$ - 1 cuyas distancias $d_{ij} = \left[\left(x_i - x_j \right)^T \left(x_i - x_j \right) \right]^{\frac{1}{2}}$ coinciden con las disimilitudes δ_{ij} observadas. Por otro lado, estas filas están representadas en un referencial ortogonal formado por los ejes principales (coordenadas principales) una vez que $X^T X = \left(V \Sigma \right)^T \left(V \Sigma \right) = \Sigma^2$. Si $x_{(j)}$ y $x_{(k)}$ son dos columnas de X - correspondientes a los ejes j y k - entonces

$x_{(j)}^T x_k = 0$ para $j \neq k$ y

$x_{(j)}^T x_j = \left\| x_{(j)} \right\|^2 = \sigma_j^2$ (j= 1 … r).

Cuando $\Delta = [\delta_{ij}]$ no es una métrica, sea

$A = \left[-\frac{1}{2} \delta_{ij}^2 \right]$ y $B = H A H$ (obtenida por doble centrado de A).

Puede verse en COX *et al* (1994), que, si $|B| > 0$, el procedimiento definido sigue válido. Cuando $|B| < 0$ - hay valores propios negativos; en este caso, si transformamos Δ sumando una constante - obtenida por el método de Caillez - COX *et al* (1994), pg. 36 - es aún posible aplicar la metodología anterior (MDS Clásico, Métrico o Análisis en Coordenadas Principales) para obtener una configuración de puntos en un espacio métrico.

Como pone de manifiesto VICENTE-VILLARDÓN (1992), el GRCMP-biplot es un procedimento tal que, dadas dos métricas Φ, en el espacio de filas y Ω en el espacio de columnas, permite obtener configuraciones de marcadores en un espacio de baja dimensión $r \leq$ rango(X) que preservan tanto la métrica Φ de las filas como la métrica Ω de las columnas.
Por lo tanto, el MDS métrico puede verse como un caso particular de GRCMP-biplot.

Pero también es posible la perspectiva de que un biplot es un caso particular de MDS. Es ésta la perspectiva de GOWER *et al* (1996), que desarrolla un concepto de biplot considerado como una representación gráfica que se obtiene a partir de las distancias entre filas de una matriz de datos por un proceso designado **interpolación** y sobre el cual pueden posicionarse ejes - correspondientes a variables - usando un procedimiento designado **predicción**.

En este trabajo seguiremos la perspectiva de Gabriel y de la Escuela de Salamanca.

3.3.2. EL MDS. PROBLEMAS DE INTERPRETACIÓN.

De acuerdo con BORG *et al* (1997), dado el gráfico formado por la configuración final de puntos producidos por uno de los métodos MDS - métricos o no - la interpretación consiste en buscar en ese gráfico grupos de objetos, regiones significativas o ejes.

De la obra citada - pg. 5 - se recogen las observaciones siguientes a propósito de la interpretación de un conjunto de datos:

> … «*What does the MDS picture in figure 1.1. tells us? It shows that crimes are primarily distributed along a horizontal dimension that could be interpreted as "violence Vs property" crimes. Moreover, the "property crimes" are less homogeneous, exhibiting some spread along the vertical axis, a dimension that could be interpreted as "hidden Vs street" crimes.*»

Y, más adelante:

> … «*Although here we looked at dimensions, it is important to keep in mind that any property of the MDS representation that appears unlikely to result from chance can be interesting. The points may, for example, form certain groupings or clusters. Or, they may fall into different region such as the centre region surrounded with bands.*»

Dada una configuración de puntos resultantes del MDS, las **operaciones básicas de interpretación** son las mismas que hemos identificado para análisis de tipo factorial: hay que buscar el significado de las dimensiones espaciales de la configuración y, para eso, intentar descubrir cuales son los grupos de objetos/individuos que se oponen a lo largo de esas dimensiones. Pero eso presupone que identifiquemos, antes, los **grupos de objetos** cuya oposición da significado a esa dimensión.

Ejemplo 3.3.2.1.

Aplicando el MDS clásico a una matriz de distancias euclideas construida con las distancias entre filas de la matriz de datos de la **tabla 3.1.1.** se obtiene la **figura 3.3.2.1.**

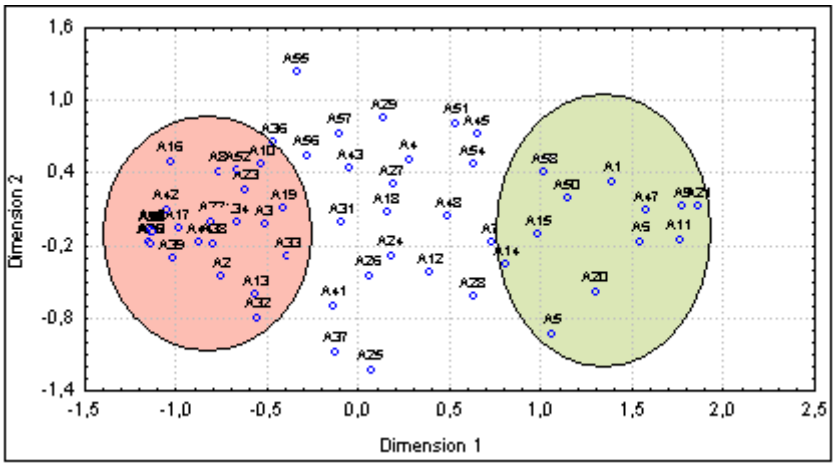


Figura 3.3.2.1. MDS de las distancias entre individuos de la **tabla 3.1.1.**

Si intentamos interpretar el eje horizontal de la **figura 3.3.2.1.**, hay que buscar que es lo que opone el grupo Rojo (a la izquierda) al grupo Verde (a la derecha).

VERDE = {1, 5, 6, 9, 11, 14, 15, 21, 47, 50, 58}
ROJO= {2, 3, 10, 13, 16, 17, 19, 22, 23, 32, 33, 39}

Una vez caracterizados los grupos Verde y Rojo, el significado de la dimensión horizontal puede ser expresado en función del significado de cada uno de los grupos que se oponen a lo largo de esa dimensión.

Ocurre aquí, exactamente, el mismo problema que ha sido identificado al interpretar los resultados de tipo factorial: *los resultados pueden ser representados por variables indicadoras o por variables cualitativas representativas de particiones*.

Si queremos representar la dimensión horizontal de la **figura 3.3.2.1.**, por ejemplo, eso equivale a definir una variable cualitativa con valores {-, +, *} en donde "-" es la etiqueta del grupo Rojo; "+" es la etiqueta del grupo Verde; "*" es la etiqueta que representa «ni una cosa ni otra».

Cuando se identifica un grupo de objetos con el propósito de interpretación, estamos, implícitamente considerando que, desde ese punto de vista, los objetos que integran al grupo son indistinguibles o equivalentes. Esto se representa matemáticamente atribuyendo al grupo una etiqueta común a todos sus elementos.

3.4. BIPLOTS Y MÉTODOS DE CLASIFICATIÓN.

3.4.1. SÍNTESIS TEÓRICA.

Los métodos de clasificación buscan identificar grupos homogéneos de individuos, es decir, particiones del conjunto de individuos de forma que los individuos que integran un grupo tengan más semejanzas entre sí que con los individuos que forman un grupo distinto.

Los métodos de clasificación integran una literatura demasiado voluminosa para poder ser sintetizada en unas pocas hojas.

En SNEATH *et al* (1973), BENZÉCRI (1973), CUADRAS (1981), DIDAY *et al* (1982); GAUL *et al* (*eds.*) (1998), GORDON (1999), HAIR *et al* (1995), BOCK *et al* (1999), GAUL *et al* (2000), LEBART *et al* (2002), pueden verse los fundamentos lógicos, la teoría básica asociada a los principales algoritmos de clasificación y los resultados de la investigación reciente.

Los métodos de clasificación son tradicionalmente divididos en métodos jerárquicos y no jerárquicos.

Entre los jerárquicos se consideran los aglomerativos y los divisivos.

Los métodos aglomerativos parten de los objetos a clasificar. Estos se van agregando según criterios específicos hasta que todos los objetos están reunidos en una sola clase. Cuando dos clases se funden para formar una nueva clase, el valor del índice de disimilitud a que esa fusión se produce es registrado.

En los métodos divisivos, se parte del todo y este es dividido en grupos sucesivamente más específicos. Ver BREIMAN *et al* (1993), HASTIE *et al* (2001).

Consideremos los métodos jerárquicos aglomerativos - los más usados - y analicemos la posibilidad de representar sus resultados usando gráficos planos como el biplot.

El resultado final de un algoritmo de clasificación aglomerativo es un dendograma como el ejemplificado en la **figura 3.4.1.1.a)**

La **figura 3.4.1.1.a)** representa un dendograma obtenido por la clasificación aglomerativa de 6 individuos. Para cada clase (vértices 7, 8, 9, 10, 11) la escala a la izquierda indica el nivel de disimilitud en la que esa clase se ha formado.

Un dendograma puede representarse numéricamente usando una ultramétrica. Ver, por ejemplo, BENZÉCRI (1973), o DIDAY *et al* (1982). Dado un conjunto de objetos O, una ultramétrica es una aplicación

$u: O \times O \longrightarrow \Re^{+}$ tal que

(1) $u(i, j) = 0 \Leftrightarrow i = j$ para todo $(i, j) \in O \times O$

(2) $u(i, j) = u(j, i)$ para todo $(i, j) \in O \times O$

(3) $u(i, j) \leq \max (u(i, k)\, u(k, j))$ para $(i, j, k) \in O \times O \times O$

En DIDAY *et al* (1982) puede verse que

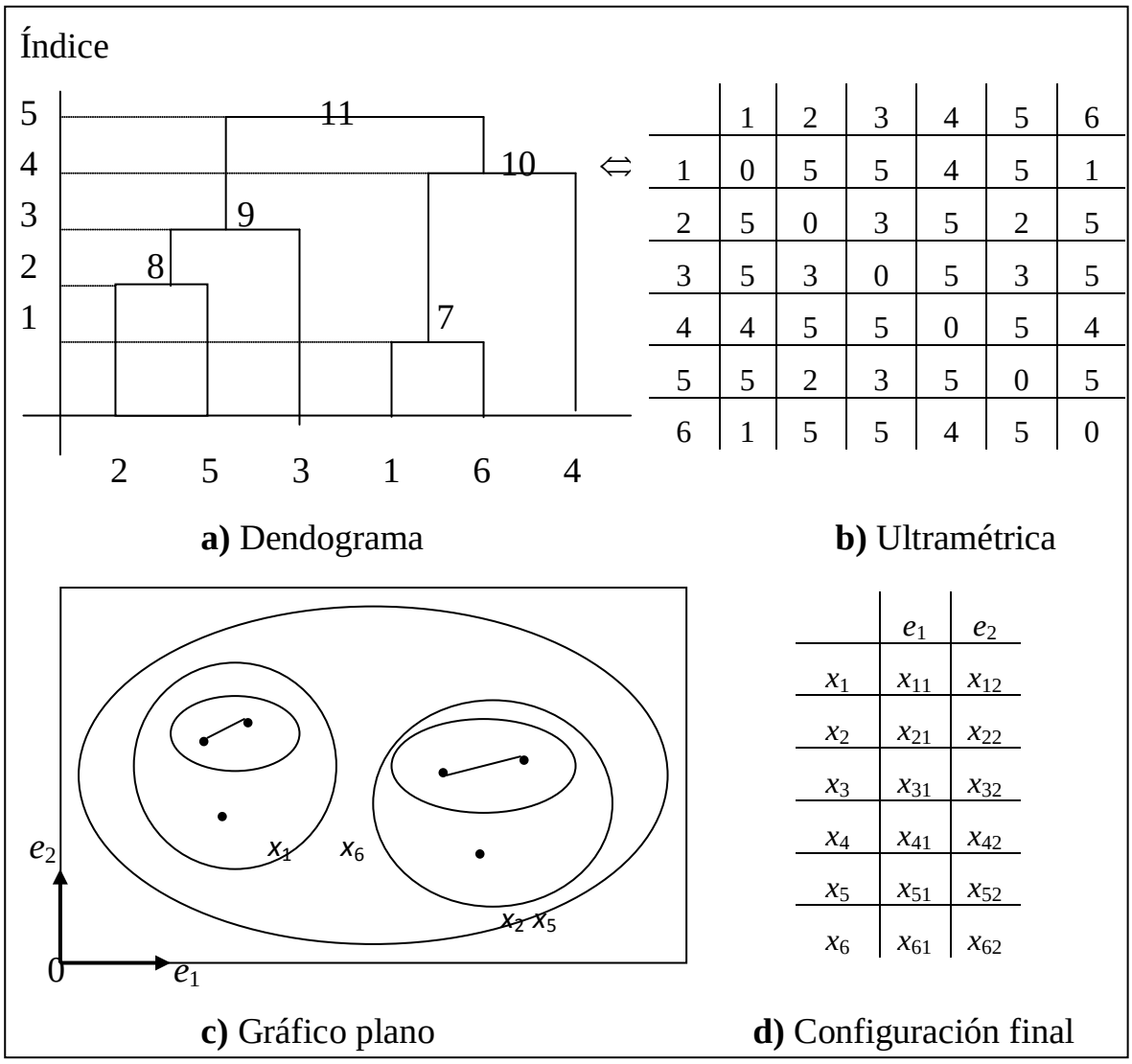
si

$$u(i, j) \leq \max (u(i, k),\ u(k, j))$$

entonces

$$u(i, j) \leq d(i, k) + u(k, j)$$

Esto quiere decir que una ultramétrica es una métrica específica en donde todos los triángulos son isósceles.



	1	2	3	4	5	6
1	0	5	5	4	5	1
2	5	0	3	5	2	5
3	5	3	0	5	3	5
4	4	5	5	0	5	4
5	5	2	3	5	0	5
6	1	5	5	4	5	0

	e_1	e_2
x_1	x_{11}	x_{12}
x_2	x_{21}	x_{22}
x_3	x_{31}	x_{32}
x_4	x_{41}	x_{42}
x_5	x_{51}	x_{52}
x_6	x_{61}	x_{62}

Figura 3.4.1.1. Clasificación aglomerativa de 6 individuos y su relación con la ultramétrica.

Dado un dendograma como el de la **figura 3.4.1.1.a)** podemos representar numéricamente ese dendograma por una ultramétrica (ver **figura 3.4.1.1.b)**) asociando a cada par de objetos (*i, j*) el nivel *u* (*i, j*) al que es necesario subir en el dendograma para que los individuos del par (*i, j*) pertenezcan a la misma clase.

En DIDAY *et al* (1982) puede verse la demostración del hecho de que existe una correspondencia biunívoca entre ultramétricas y dendogramas: dado un dendograma se puede construir una ultramétrica y para toda ultramétrica existe un dendograma que lo representa.

Una vez que una ultramétrica es una métrica, dada la ultramétrica que representa un dendograma obtenido por un método cualquiera (a partir de disimilitudes observadas o calculadas usando un criterio de agregación específico) es posible - usando el biplot generalizado (RCMP-biplot) o el MDS métrico o clásico - representar ese dendograma por una configuración de puntos en un espacio métrico. Ver en la **figura 3.4.1.1.** la tabla **d)**.

Se observa que, aún en este entorno de clasificación, un modo natural de representar gráficamente los resultados es usar configuraciones de puntos en biplots.

Esto significa que los árboles obtenidos por distintos métodos de clasificación aglomerativa, usando índices de disimilitud y criterios de agregación distintos, pueden ser comparados comparando las configuraciones que les corresponden sobre biplots.

En GREENACRE (1984, 1998) pueden verse las relaciones entre métodos de clasificación y los biplots particulares asociados al análisis factorial de correspondencias y el uso de un criterio de agregación de clases basado en la noción de inercia de un grupo de marcadores con relación a su centro de gravedad.

En VICENTE-TAVERA (1992), esta idea es desarrollada y aplicada en la interpretación del HJ / RCMP-biplot, definiendo un método de agregación

basado en la inercia. Este método consiste en fundir, en cada fase de la construcción del árbol, las dos clases que produzcan la mínima reducción de inercia entre la partición existente y la que resulta de la fusión de las dos clases de la partición presente que tengan la mínima inercia entre clases, calculadas esas inercias considerando los centros de gravedad de cada una de las clases candidatas y el centro de gravedad común.

El examen, en el biplot, de las configuraciones de marcadores representando los clusters resultantes, permite identificar las variables que caracterizan cada cluster.

3.4.2. CLASIFICATIÓN. PROBLEMAS DE INTERPRETACIÓN.

Los resultados de los métodos de clasificación son de dos tipos: grupos específicos de individuos y particiones del conjunto de individuos.

Los problemas de interpretación de esos resultados son, por eso, los siguientes:

1. ¿Cuál es el número de clases «naturales» subyacentes a esos grupos o particiones? Se trata del problema de la validación de grupos.

En el caso particular de las clasificaciones aglomerativas, este problema equivale a identificar el nivel óptimo que debe ser usado para «cortar» el árbol de clasificación (dendograma).

Como puede verse en GORDON (1999) se trata de un problema de inferencia aún no totalmente resuelto.

2. Dado un cluster o grupo de objetos, ¿cómo resumir o sintetizar ese grupo?
¿Cómo expresar sintéticamente el significado de ese grupo?

3. Dados dos «clusters» o grupos de objetos, ¿qué es lo que separa, opone o distingue los dos grupos uno del otro? ¿Cómo descubrir y expresar lo que distingue un grupo de otro?

Desde el punto de vista de la interpretación de resultados en un contexto de análisis preliminar de datos, se verifica que los problemas son exactamente los que hemos identificado para el MDS o para los análisis de tipo factorial.

Ejemplo 3.4.2.1. (Interpretación de resultados de un análisis cluster aglomerativo).
Volviendo a los datos de la **tabla 3.1.1.** realicemos ahora un análisis aglomerativo de los individuos (alumnos de un curso de estadística) usando como medida de disimilitud la distancia euclidea entre marcadores de esos individuos en un biplot de GABRIEL (α= 1/2) y como criterio de agregación el criterio de Ward. Los resultados pueden verse en la **figura 3.4.2.1.**

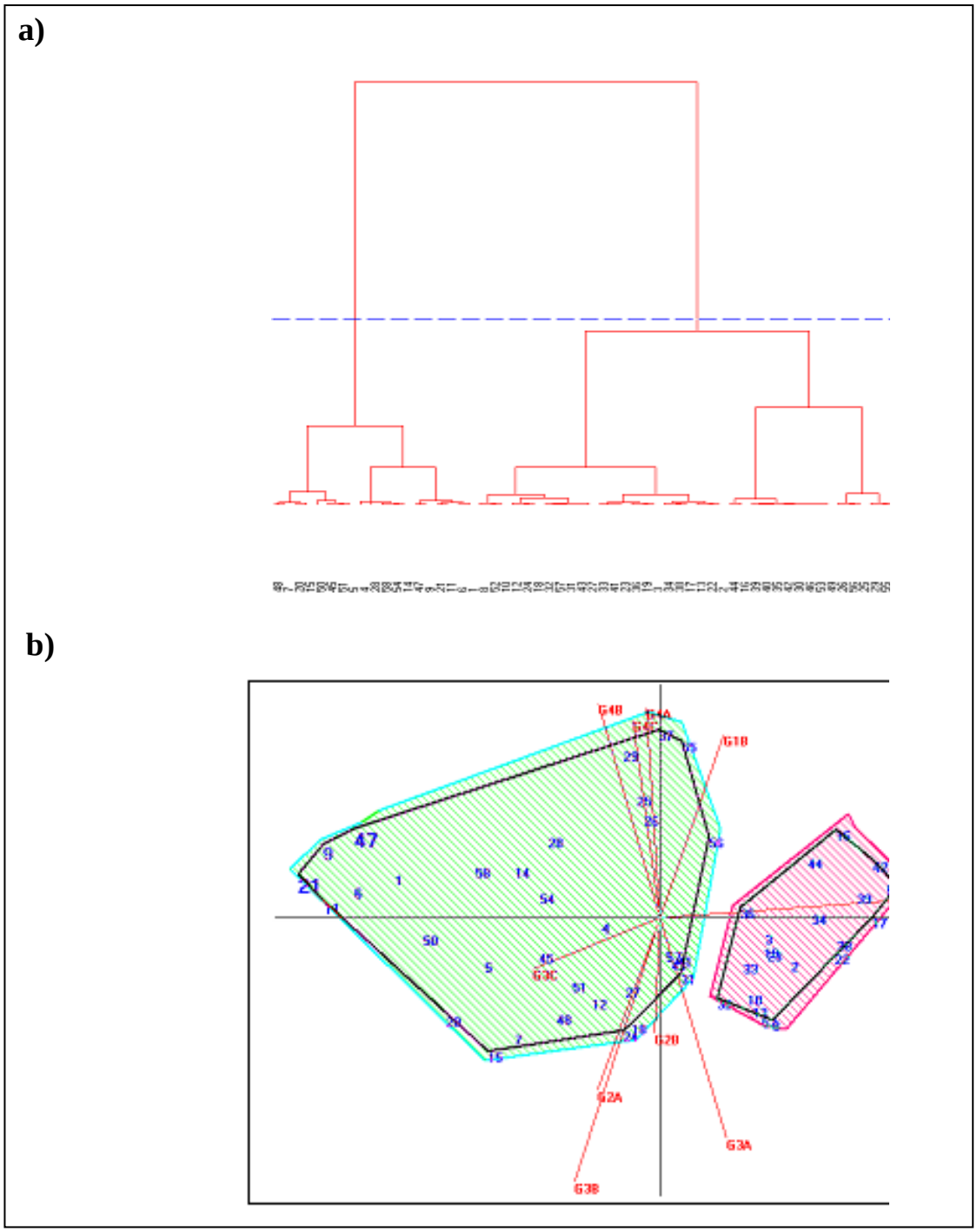


Figura 3.4.2.1. **a)** Resultado de un cluster análisis de los datos de la **tabla 3.1.1.** y su representación sobre el biplot, en **b)**.

Cortando el dendograma a un nivel correspondiente a una partición con solo dos grupos se obtiene la **figura 3.4.2.1.a)**. Del lado izquierdo figura el grupo Verde y del lado derecho el grupo Rojo.

Verde= {1, 4, 5, 6, 7, 9, 11, 14, 15, 20, 21, 28, 45, 47, 48, 50, 51, 54, 58}.

Rojo= {2, 3, 8, 10, 12, 13, 16, 17, 18, 19, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 46, 49, 52, 53, 55, 56, 57}.

Construyendo los cierres convexos de estos grupos sobre el biplot se obtiene la representación de esta partición sobre el biplot. Ver **figura 3.4.2.1.b)**.

Los grupos ROJO o VERDE podrían representarse matemáticamente por la función característica respectiva; la partición podría representarse por una variable cualitativa con 3 valores, como en los casos del MDS o análisis factorial.

3.5. BIPLOTS PARA DATOS CUALITATIVOS Y MIXTOS.

3.5.1. SÍNTESIS TEÓRICA.

Los biplots de GABRIEL (1971) han sido definidos como representaciones geométricas de datos correspondientes a variables tipo cuantitativo.

Para los datos de tipo cualitativo, una solución posible para la aplicación de la técnica de los biplots es construir tablas de contingencia de dos vías para pares de variables y aplicar a la tabla resultante la técnica de los biplots o un análisis factorial de correspondencias.

Cuando se construye un biplot a partir de una tabla de contingencia, la representación gráfica resultante contiene marcadores para las categorías de

la variable en filas, marcadores para las categorías de la variable en columnas pero no contiene marcadores para los individuos. Ver GABRIEL (1995a).

En otras palabras: el biplot resultante permite relacionar las variables cualitativas dos a dos pero no presenta información acerca de los individuos.

Este procedimiento puede ser útil en un contexto de minería de datos mixtos o cualitativos una vez que permite resumir un gran conjunto de datos por una tabla de contingencia, pero es muy limitado, dadas las pérdidas de información.

Cuando es necesario relacionar las categorías de más de dos variables, hay dos soluciones:

1. Crear biplots para tablas de contingencia de dos vías asociadas a variables cuyas categorías se obtienen por concatenación de las categorías de otras variables.
2. Transformar cada variable cualitativa en un vector booleano (o vector de variables indicadoras) con dimensión igual al número de categorías de la variable, calcular la tabla de BURT (con todas las tablas de contingencia correspondientes a esas variables) - ver BENZÉCRI (1973), GREENACRE (1984) - y, a continuación, crear biplots de GALINDO (1985), o biplots generalizados a partir del cuadro de BURT[·]. Ver **figura 3.5.1.1.**

[·] Las variables continuas o cuantitativas discretas también pueden ser transformadas en variables discretas cuantitativas con un número limitado de valores. Ver, por ejemplo, GABRIEL (1995a).

Figura 3.5.1.1. Los datos mixtos son discretizados en variables indicadoras y con estas se genera cuadro BURT que puede ser analizado por AFC, usando el biplot de GALINDO.

En el caso del cuadro de BURT, el biplot contiene, marcadores para todas las categorías de todas las variables del conjunto de datos pero no contiene marcadores para los individuos.

Para matrices de datos con variables indicadoras de las categorías de variables cualitativas, GABRIEL (1995a), sugiere las variantes CMP_f y RMP_f de los biplots GH / CMP y JK / RMP. El índice ***f*** se refiere a las frecuencias relativas, usadas para centrar las variables indicadoras. Estos biplots permiten la representación simultánea de los individuos y de las categorías de las variables cualitativas.

Dada una variable cualitativa $x_{(k)}$ con categorías indexadas por j= 1, 2, ., m_k, entonces (k | j) representa la categoría j de la variable k (k= 1, 2, ., p), usando la notación de GABRIEL (1995a). Por ejemplo, en la **figura 3.5.1.1**, la variable X_2 genera dos variables indicadoras - X_{21} y X_{22} - correspondientes a las categorías *A* y *B*. En la notación de GABRIEL (1995a) esas columnas se designan por $X_{1|2}$ y $X_{2|2}$

$x_{i,\,k|j}$= 1 si $x_{(k)}$ tiene el valor número j ; *0* en el caso contrario.

$$\sum_j x_{i,k|j} = 1, \qquad i = 1, 2, ..., n; \qquad k = 1, 2, ..., q$$

$$\sum_{k,\,j} x_{i,k|j} = q, \qquad i = 1, 2, ..., n$$

Sea $m = \sum_k m_k$ el número total de variables indicadoras de las categorías.

En el ejemplo de la **figura 3.5.1.1,** *m = 2 + 2 +3 = 7 y q = 3.*

$$X_{n\times m}=\left[x_{i,k|j}\right]$$

$$p_{k|j}=\frac{1}{n}\sum_i x_{i,k|j}\ ,$$ media de la variable $x_{(k|j)}$

$$\mathbf{p}^T=\left[p_{1|1}..p_{1|m_1},...,p_{q|1},...,p_{q|m_q}\right]$$

$$\mathbf{D_p}=diag\,(p)$$

$$Y_{n\times m}=\left[x_{i,k|j}-p_{k|j}\right],$$ datos centrados.

En GABRIEL (1995b) puede verse que, realizando la SVD de la matriz $YD_p^{-1/2}=U_f\,\Sigma_\phi\,V_f^T$ se pueden obtener el GH / CMP-biplot y el JK / RMP-biplot. En esa referencia también se consideran los biplots resultantes de la descomposición de $Z=\left(X-1^T\ p\right)D_s^{-1}=U_Z\,\Sigma_\xi\,V_Z^T$ en donde D_s es, ahora, la matriz diagonal cuyos elementos diagonales son las desviaciones típicas $\sqrt{p_{k|j}(1-p_{k|j})}$ de las variables indicadoras. La letra z se refiere a la estandarización realizada sobre cada una de las variables indicadoras en consecuencia de esas transformaciones.

En el caso de datos mixtos puede ser necesario discretizar las variables continuas y tratar las variables resultantes del mismo modo que las variables cualitativas.

Sobre este asunto, puede también verse GREENACRE *et al* (2001).

Esto significa que el carácter cualitativo de las variables y los datos mixtos no limitan la aplicación de los biplots a esos datos.

3.5.2. INTERPRETACIÓN DE BIPLOTS RESULTANTES DE VARIABLES INDICADORAS MÚLTIPLES.

Como puede verse en GABRIEL (1995a), existe un paralelismo casi perfecto entre la interpretación estadística de los biplots resultantes de matrices de variables indicadoras múltiples y la interpretación estadística de los biplots resultantes de variables cuantitativas.

Ejemplo 3.5.2.1.

Si cada una de las 10 variables del conjunto de datos de la **tabla 3.1.1.** es reemplazada por un vector de tres variables indicadoras correspondiente a las categorías 0, 1, 2 (valoración posible de cada una de las respuestas) obtenemos, ahora, una matriz de 30 variables indicadoras en lugar de las 10 variables iniciales. La matriz *X* tiene, ahora, 58 filas y 30 columnas.

Descomponiendo la matriz centrada por columnas *Y*, se obtiene el biplot de la **figura 3.5.2.1.** cuya macroestrutura se explica por la oposición entre los grupos VERDE y ROJO ("buenos" y "malos" alumnos) detectada en los ejemplos anteriores de este capítulo.

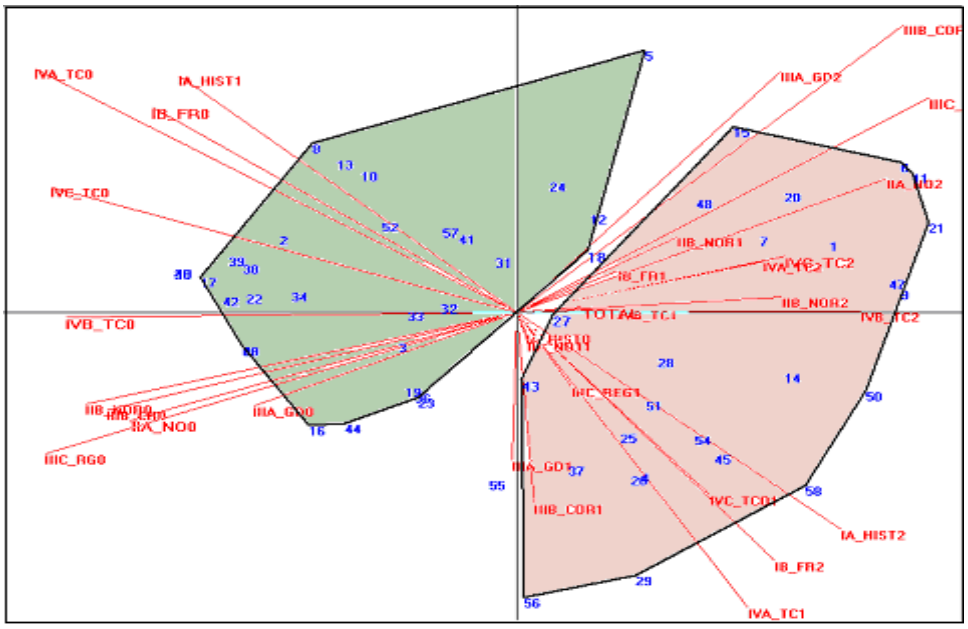


Figura 3.5.2.1 Grupo VERDE- "Buenos alumnos"; Grupo ROJO- "malos alumnos".

VERDE= {1, 4, 7, 9, 11, 14, 15, 20, 21, 25, 26, 27, 28, 29, 37, 43, 45, 47, 48, 50, 51, 54, 55, 56, 58}

ROJO= {2, 3, 5, 8, 10, 12, 13, 16, 17, 18, 19, 22, 23, 24, 30a, 31, 33, 34, 35, 36, 38, 39, 40, 41, 42, 44, 45, 49, 52, 53, 57}

Los problemas de interpretación son, exactamente, los que han sido descritos para los métodos anteriores.

3.6. BIPLOTS Y DIAGNOSIS DE MODELOS EN TABLAS DE CONTINGENCIA.

3.6.1. SÍNTESIS TEÓRICA.

BRADU y GABRIEL (1978) muestran como pueden usarse los biplots de Gabriel en diagnósticos de modelos log-lineales.

DIAZ-LENO (1995) presenta una investigación sistemática de las relaciones entre patrones de marcadores de filas y columnas de tablas de contingencia en el biplot de Gabriel y los modelos de independencia condicional log-lineales para las frecuencias.

En ese trabajo se demuestra que si en un biplot de Gabriel se observa que un grupo dado de marcadores para las categorías en columnas - b_j - está sobre una dirección, entonces $b_j = b_. + n_j\, v$, en donde n_j es un número real, v es el vector unitario con la dirección $(b_j - b.)$

y $$b. = \frac{1}{n} \sum_{j=1}^{n} b_j .$$

Para una fila *i*, representada por su marcador a_i, y para los marcadores de las columnas del grupo con la dirección definida por *v*,

$$\begin{aligned} y_{ij} = \log\ f_{ij} &= a_i^T\ b. \\ &= \left(a. + \left(a_i - a.\right)\right)^T \left(b. + n_j\ v\right) \\ &= a_.^T\ b. + a_.^T\ n_j\, v + \left(a_i - a.\right)^T b. + \left(a_i - a.\right)^T n_j\ v \end{aligned}$$

Esta expresión tiene la forma

$$y_{ij} = \log f_{ij} = \mu + \alpha_i + \beta_i + \delta_i\, \beta_i$$

con

$$\begin{aligned} &\mu = a_.^T\, b_. \\ &\left(a_i - a.\right)^T b. = \alpha_i \\ &n_j\, a_.^T\, v = \beta_j \\ &\frac{\left(a_i - a.\right)^T v}{a_.^T\, v} = \delta_j \end{aligned}$$

Esto significa que, si se observa sobre el biplot de Gabriel un conjunto de marcadores de columnas colineales, eso es compatible con el diagnóstico de que los datos $\left[y_{ij}\right]_{I\times J} = \left[\log f_{ij}\right]_{I\times J}$ siguen el modelo

$$\left[\mu + \alpha_i + \beta_i + \delta_i\ \beta_i\right]_{I\times J}$$

Aún en DIAZ-LENO (1995) se pueden ver reglas de diagnóstico para las hipótesis de independencia en modelos bifactoriales, lo que hace intervenir patrones lineales paralelos o perpendiculares en los biplots.

Así, como ejemplo de una de esas reglas -DIAZ-LENO (1995) - se tiene:

> *«La hipótesis de aditividad en una tabla bidimensional – $i \amalg j$ – requiere que los marcadores* ***a*** *y los marcadores* ***b*** *estén aproximadamente sobre* ***lineas rectas perpendiculares****»*.

La diagnosis de modelos de independencia condicionada en tablas trifactoriales – por ejemplo, la hipótesis $i \amalg j \mid k$ – hace intervenir patrones más complejos.

Como ejemplo, se citan las reglas sugeridas en DIAZ-LENO (1995) y GABRIEL *et al* (1998) para diagnosticar el modelo $i \amalg j \mid k$ – independencia condicionada de *i* y *j* dada la categoría *k*. De esta última referencia se cita la regla:

> «*Diagnosticar* $i \amalg j \mid k$ *si, para un k dado, los* a_i*'s y* b_j*'s forman aproximadamente* ***rectas perpendiculares***»

En la misma referencia, la ***Regla del Paralelogramo*** para diagnosticar el modelo de independencia condicionada $j \amalg k \mid i$ es formulada del modo siguiente:

> «*Diagnosticar* $\boldsymbol{j \amalg k \mid i}$ *si para cada j, j'; k, k' los marcadores (*b_{jk}*,* $b_{j'k}$*,* $b_{jk'}$*,* $b_{j'k'}$*) están próximos a un* ***paralelogramo****, o sea, si el marcador* $b_{jk}+b_{j'k'}$ *está próximo de* $b_{jk'}+b_{j'k}$».

En BLÁZQUEZ-ZABALLOS (1998) se formulan reglas de diagnosis para modelos logarítmico lineales y modelos logit con variables de respuesta.
Estas reglas refieren patrones geométricos en tres dimensiones. Como ejemplo se cita, en el referido trabajo, la regla:

> «*Los datos siguen un modelo de efectos columna si, y solo si, admiten una representación en rango tres en la que los marcadores para las filas y para las columnas son* ***coplanarios****, ambos* ***planos son perpendiculares*** *y además las* ***proyecciones*** *de los marcadores fila sobre la intersección están ordenadas*».

Por lo tanto, esta regla de diagnosis implica detectar sobre los biplots grupos de marcadores (de individuos y de variables) que se aproximen de ***planos perpendiculares*** en un espacio de dimensión 3.

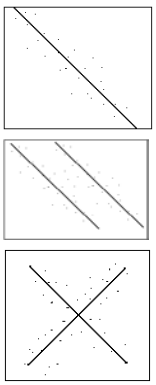
En VARELLA-NUALLES (2002) son sugeridas reglas de diagnosis para patrones de interacción de orden 3 basadas en biplots interactivos – ver KROONENBERG (1983).

En SEPULVEDA (2003) pueden verse reglas de diagnosis para modelos de clases latentes basadas en biplots para modelos lineales generalizados.
Aún aquí se buscan grupos de marcadores próximos de líneas rectas, para realizar la diagnosis de esos modelos.

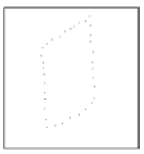
3.6.2. INTERPRETACIÓN DE BIPLOTS PARA DIAGNOSIS DE MODELOS.

Se ha visto por el examen de algunas de las reglas de diagnosis citadas en el apartado 3.6.1, que estas reglas implican la necesidad de «reconocer» en las representaciones biplots de dimensión dos o tres, la existencia de **grupos de marcadores** – de variables e individuos – con configuraciones «próximas» de:

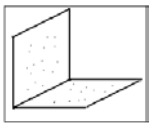
- Rectas
- Rectas paralelas
- Rectas perpendiculares



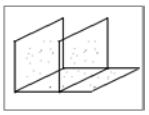
- Paralelogramos



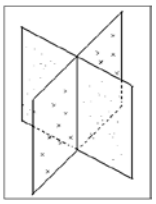
- Planos en espacios de dimensión 3



- Planos paralelos



- Planos perpendiculares



Esto significa que el proceso de diagnosis o reconocimiento de esos modelos estadísticos a partir de los biplots adecuados, implica crear instrumentos que faciliten la detección, sobre los biplots, de grupos de marcadores que se aproximen de los patrones geométricos característicos.

Una vez que las investigaciones teóricas citadas han puesto de manifiesto la correspondencia entre modelos de comportamiento estadístico y patrones geométricos característicos en los biplots, el problema fundamental que precede la eventual diagnosis de un **modelo** o **patrón** de comportamiento estadístico es el de identificar, mirando a los biplots, ***grupos*** de variables e individuos que sean el soporte, en los datos observados, de esos modelos teóricos.

En el ejemplo siguiente, basado aún en los datos de la **tabla 3.1.1**, se sugiere la utilización del mecanismo de proyección incorporado en el sistema prototipo descrito en el capítulo VI, para detectar grupos de marcadores próximos de patrones lineales paralelos y perpendiculares, que puedan sugerir la presencia de los modelos referidos.

Ejemplo 3.6.2.1.

Empleando el sistema prototipo descrito en el Capitulo VI con los datos de la **tabla 3.1.1** se ha formado una tabla de contingencia cuyas filas resultaron de la concatenación de los valores de las variables *X5* y *X6* y cuyas columnas corresponden a los valores de la variable *X9*. Esa tabla es la siguiente:

	X9=0	X9=1	X9=2	ni.
X5=0\X6=0	9		2	11
X5=0\X6=1	2	1	1	4
X5=0\X6=2	1			1
X5=1\X6=0	8			8
X5=1\X6=1	5		1	6
X5=1\X6=2	7	1	3	11
X5=2\X6=1	3		1	4
X5=2\X6=2	6	2	5	13
n.j	41	4	13	58

Seguidamente, después de calcular los logaritmos de las frecuencias absolutas se ha construido un biplot de GABRIEL con $\alpha = \frac{1}{2}$ (columnas estandarizadas).Ver la **figura 3.6.2.1.**

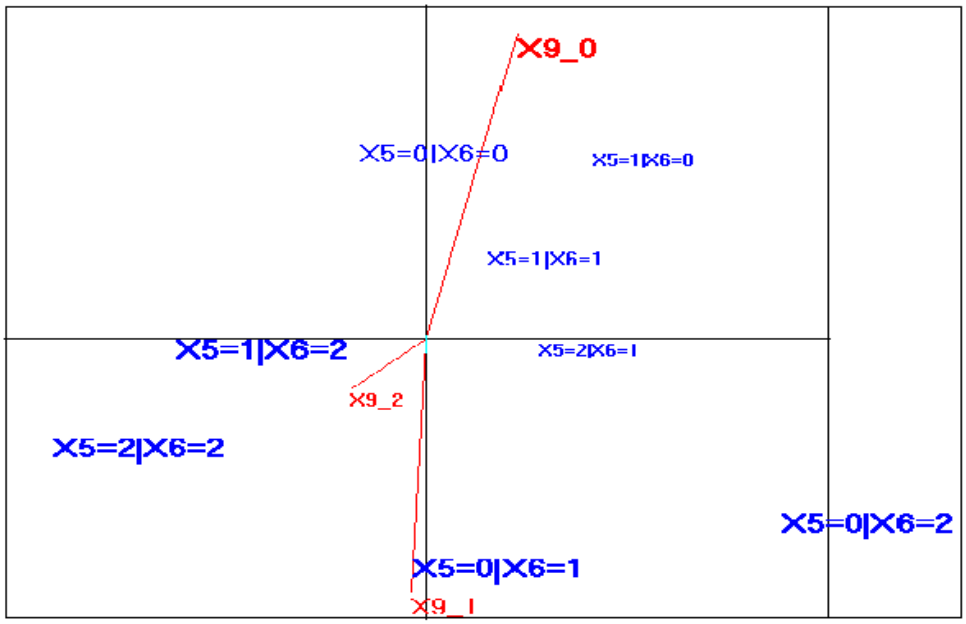


Figura 3.6.2.1. Biplot de Gabriel de los logaritmos de las frecuencias de la tabla de contingencia resultante de cruzar los valores concatenados de (X_5,X_6) con los de X_9.

¿Existirán patrones lineales entre los marcadores de las filas de esa tabla de contingencia?

Si proyectamos, sucesivamente, los individuos sobre una dirección arbitraria – ver **capítulo V**, apartado 5.5.3. – obtenemos, después de algunos experimentos, el gráfico de la **figura 3.6.2.2.** que sugiere alguna proximidad a una línea recta de los marcadores de filas del conjunto ROJO en esa figura.

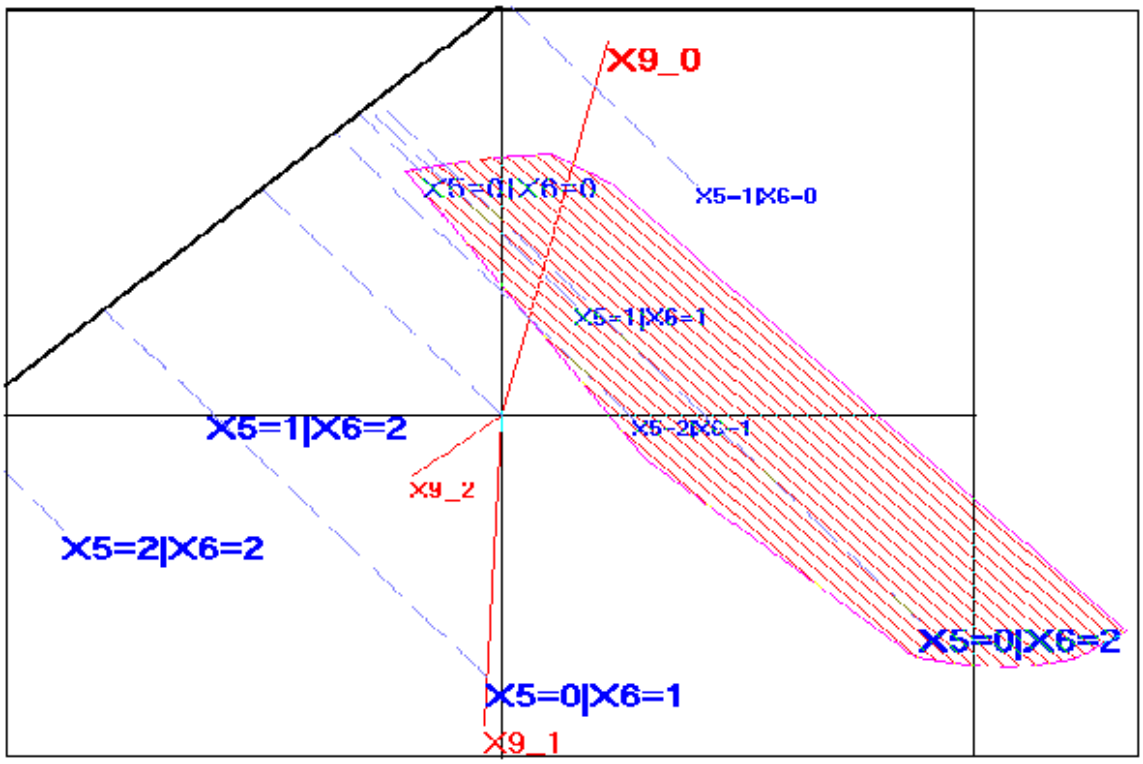


Figura 3.6.2.2. ¿Existe una línea recta entre un subconjunto de los marcadores?

En otras palabras: el grupo de marcadores del conjunto

Rojo = *{(X5 = 0 | X6 = 0), (X5 = 1 | X6 = 1), (X5 = 2 | X6 = 1), (X5 = 0 | X6 = 2)}*

sugiere que podría existir un patrón lineal entre esos marcadores, soporte de un eventual modelo, relacionando subconjuntos de las categorías de las variables *(X5,X6) y X9.*

Desde el punto de vista de la interpretación y diagnosis, el problema incidiría, ahora, sobre el grupo de filas de la tabla de contingencia identificado sobre el biplot. Una vez que estos identificadores corresponden a la concatenación de valores de variables observadas, podríamos ahora intentar descubrir si el patrón, aproximadamente lineal puesto de manifiesto en el biplot de la tabla anterior, tiene algún significado que pueda expresarse en función de las observaciones.

Calculando un biplot de GALINDO – datos estandarizados – con los datos originales y usando las posibilidades del sistema prototipo descrito en el **Capitulo VI** se verifica que las filas cercanas del patrón lineal en la tabla de contingencia- conjunto ROJO- tienden a ubicarse en una región específica de ese biplot. Ver **figura 3.6.2.3.**

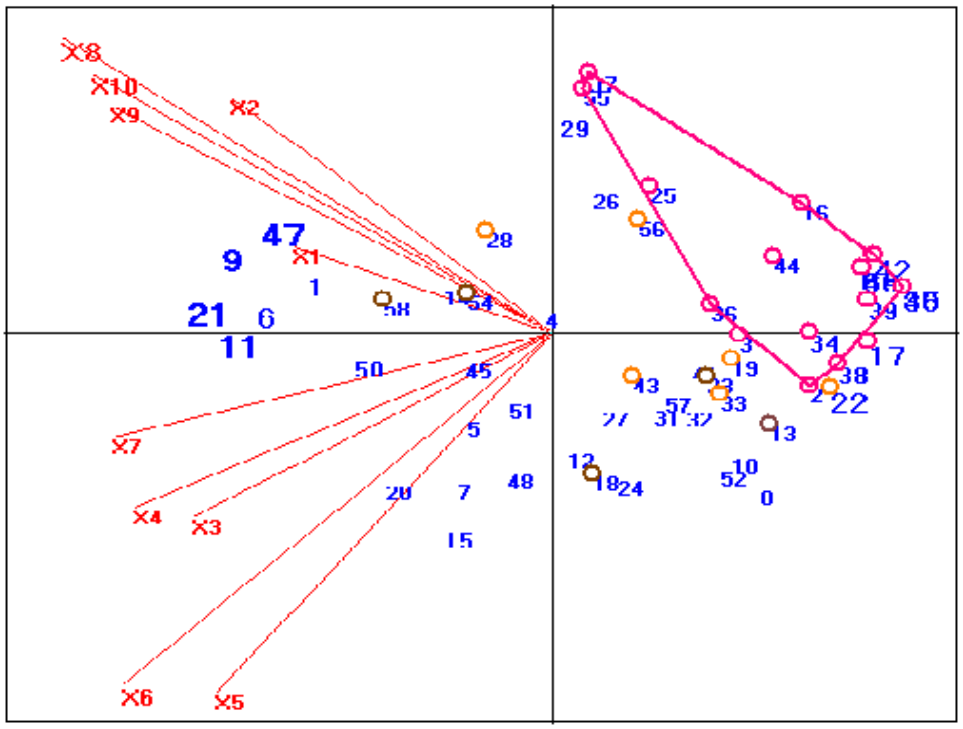


Figura 3.6.2.3. Las filas de la tabla que corresponden a un patrón lineal tienden a ubicarse en una zona específica del biplot de las observaciones originales.

En esa figura se puede ver por su cierre convexo de color rojo una caracterización aproximada de ese grupo; esa caracterización corresponde, aproximadamente a la expresión $(X_4 = 0) \wedge (X_6 = 0)$.

Por lo tanto, también en este dominio de diagnóstico de modelos, llegamos a la necesidad de interpretar y caracterizar los grupos de individuos observados que son el soporte en los datos de los modelos por diagnosticar.

C A P Í T U L O IV

MODELOS PARA INTERPRETACIÓN DE RESULTADOS DE ANÁLISIS DE DATOS MULTIVARIADOS

4.1. INTRODUCCIÓN.

El objetivo de este capítulo es el de formular matemáticamente los problemas de interpretación de resultados de análisis de datos multivariantes, considerando la interpretación como una fase independiente de la fase de análisis, con problemas específicos que implican la necesidad de metodología, teoría y técnicas propias.

La idea fundamental que se intenta transmitir es la de que, si es verdad que al analizar un conjunto de datos se puede y se debe utilizar diferentes técnicas - mirar a los datos desde diferentes perspectivas, permitir que los datos sean examinados por diferentes expertos - los resultados generados por estas diferentes técnicas de análisis pueden, sin imponer demasiadas restricciones, ser expresados usando un lenguaje común.

Hemos puesto de manifiesto en el capítulo anterior que, a la hora de realizar la interpretación de los resultados obtenidos por los principales métodos de análisis de datos multivariantes, las categorías que siempre están presentes en casi todos los razonamientos son *grupos de individuos*, *grupos de variables* o *grupos de individuos y variables*.

Conociendo los objetos (individuos y variables) que integran a un grupo, ese grupo puede, a su vez, expresarse por configuraciones de marcadores en biplots - lo que, en resumen, significa que, al interpretar los resultados obtenidos por métodos de análisis distintos, podemos razonar sobre configuraciones de marcadores en biplots apropiados. Ver en la **figura 4.1.1.** un diagrama que intenta expresar esta idea.

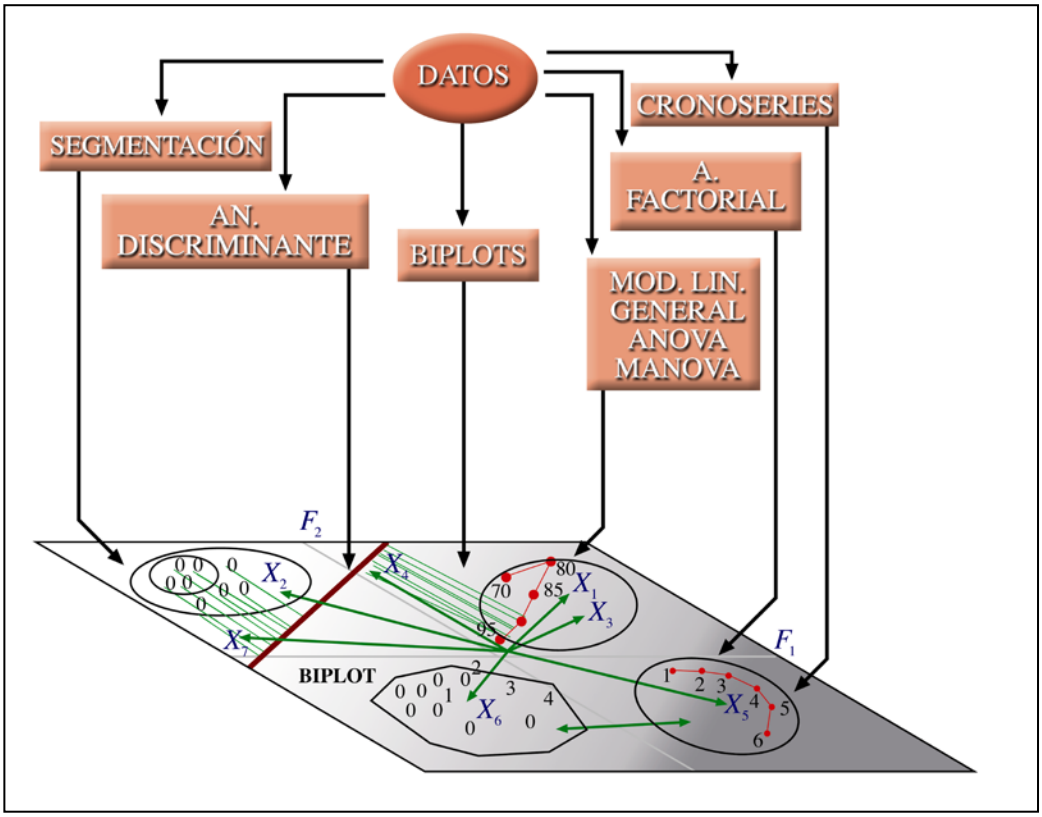


Figura 4.1.1. Los resultados más importantes de casi todos dos métodos de análisis de datos son grupos de objetos (individuos y variables) que pueden ser expresados por configuraciones de marcadores en biplots.

Estas observaciones son el fundamento empírico para los desarrollos teóricos de este capítulo.

La interpretación es una etapa crucial de todo análisis una vez que corresponde a la búsqueda del significado de los resultados; o sea, la correspondencia entre esos resultados y algún aspecto de la realidad considerada importante para el problema en investigación o de decisión.

Los problemas de interpretación siempre han sido considerados importantes en la literatura de análisis de datos pero son pocas y muy recientes las referencias relacionadas con la interpretación de resultados.

Son de TUKEY (1962) estas palabras:

> ... *«All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make the analysis easier, more precise or more accurate, and all the machinery and results of mathematical statistics which apply to analysing data».*

La realidad es que al desarrollo de procedimientos de análisis de datos se ha atribuido un esfuerzo de investigación sin correspondencia con el esfuerzo dedicado a la investigación de técnicas para interpretar los resultados.

En general, la interpretación es considerada, en la literatura de análisis de datos multivariantes - implícitamente - una responsabilidad exclusiva de los analistas.

Todo se desarrolla como si se considerara que, para la interpretación de los resultados fuera suficiente la percepción que el analista tiene de la realidad donde provienen los datos, el conocimiento de la técnica o técnicas usadas y este fuera capaz de realizar sin error las múltiples tareas de síntesis, combinación de resultados, elaboración de resúmenes - sin tener necesidad de ser orientado por principios, reglas, metodología o alguna teoría de interpretación.

En la literatura consultada hay dos grandes excepciones a esta regla general: la literatura relacionada con el MDS y la literatura de la escuela francesa de análisis de datos ligada a BENZÉCRI (1973).
Véanse los artículos publicados en la revista francesa «*Les Cahiers de l'Analyse des Données*», editada bajo la dirección de J.P. Benzécri.

En BENZÉCRI (1973, 1992) las preocupaciones acerca de la interpretación de los resultados de los análisis factoriales de correspondencias y de los resultados de taxonomía numérica (análisis de clusters) van al punto de desarrollar un conjunto de *ayudas a la interpretación* y de reglas básicas para interpretar a los ejes factoriales y las particiones de los árboles resultantes de clasificación aglomerativa.

Es instructivo leer, por ejemplo, en el volumen I de la obra citada, el artículo [**Peurs**], TIC nº 13 §§ 4-4 respetando la interpretación de un árbol de clasificación de los miedos infantiles….

> *«…4. Validité d'une interpretación verbale.*
> *Nous voici parvenues, non sans peine, à munir d'une légende la figure produite par l'ordinateur. Que vaut cette légende? En quel sens peut-on dire q'un nom est bien choisi pour une classe? De quelle utilité peut être un tel nom ?*
>
> *Le nom doit cerner, définir la classe; il doit à la fois s'accorder avec le contenu de cette clase (les individus, ici des thèmes de peur, don't elle se compose) et être incompatible avec les éléments (thèmes de peurs) qui n'appartiennent pas à la clase…»*

En la misma obra, Volumen II - *L'Analyse des Correspondences*, Parte C-Application de l'Analise des Correspondences, se puede estudiar un conjunto de casos de aplicación en donde las preocupaciones de interpretación son muy evidentes.

En esta misma línea - que aquí designamos por Escuela Francesa - está JAMBU (1991). En esta referencia se busca, con base en el concepto de ayudas a la interpretación, formular - sin preocupaciones teóricas - una metodología para la interpretación de los ejes factoriales y de las particiones resultantes de la taxonomía numérica.

De esta referencia se extrae el comentario siguiente (cf.pg 210):

> *«…After twenty years of experience in applying correspondence analysis and other factor analysis models, it appears clear that rules must be specified to help in selecting significant axes and elements of I and J because of the size of the sets involved. Twenty years ago, a data set whose size was(50,50), was considered large, so it was easy to select "by hand" each explained point and each explicative point.The data sets involved now are larger; a standard questionnaire can have 2000 to 3000 individuals and can involve 100 variables…»*

Esto tenía como consecuencia que los programas deberían facilitar las tareas de interpretación sugiriendo los ejes y puntos (individuos) más significativos - para lo que era necesario crear reglas a implementar en los programas.

A pesar de estas preocupaciones con las cuestiones de interpretación, no se vislumbra en estos trabajos ningún intento de crear una metodología general de interpretación. Tampoco se percibe en estos trabajos ninguna preocupación por definir reglas que tengan en cuenta los aspectos de psicología cognitiva relevantes para las interacciones de los analistas con los sistemas informáticos y los datos.

La literatura acerca del MDS es otra notable excepción, quizás porque, como lo reconocen BORG y GRONEN (1997):

> *«…MDS has been used not only as a tool for data analysis but also as a framework for modeling psychological phenomena. This is made clear by equating an MDS space with the notion of psychological space. A metric geometry is interpreted as a model that explains percepcions of similarity…»*

El MDS sirve, frecuentemente, de modelo a fenómenos físicos y psicológicos, estableciendo una correspondencia intuitiva y muy sugestiva entre fuerzas o tensiones mecánicas o psicológicas (estrés) y el concepto estadístico de estrés, básico en la técnica del MDS.

Es en el dominio del MDS donde se ha asistido a un intento de creación de una teoría de interpretación - relacionando los resultados de los análisis con la realidad a la que el MDS sirve de modelo: la teoría de las facetas (Facet Theory) presentada en el capítulo 5 de la referencia citada. Ver también BORG *et al* (1998).

En las palabras de los referidos autores BORG *et al* (1997):

> *«... Interpreting an MDS solution means linking geometric propreties of the configuration to substantive features of the represented objects. A very general approach is to interpret regions of an MDS space…»*

En BORG *et al* (1997), puede verse también una gran preocupación por relacionar el MDS con otras técnicas como los Análisis de Procrustes y el Análisis Factorial de Correspondencias razonando con base en el hecho de que los resultados de todos estos métodos pueden ser expresados, en un ***lenguaje común,*** por configuraciones de puntos en espacios métricos.

En la línea de una preocupación creciente con los problemas de interpretación podemos citar a KRZANOWSKI (1998, 2000), artículo que tiene el sugestivo título de: *Principles of Data Analysis: A user's Perspective****.***

La preocupación por los problemas de interpretación lleva a este autor a escribir todo el volumen privilegiando las cuestiones de interpretación y de búsqueda de relaciones entre los diferentes métodos.

De esta referencia se destaca - dadas las consecuencias que tiene para la interpretación de resultados - el problema de comparación e integración de subespacios.

KRZANOWSKI empieza por utilizar el método de Procrustes de GOWER (1975) para comparar configuraciones de puntos resultantes de análisis en componentes principales con configuraciones de puntos resultantes de análisis MDS. En los dos casos, los individuos observados son

representados por configuraciones de puntos en espacios métricos. En esa obra KRZANOWSKI refiere también métodos anteriores - KRZANOWSKI (1979) - para comparar configuraciones obtenidas por componentes principales a partir de muestras distintas de las mismas variables. Esas configuraciones de puntos pueden ser comparadas, después de transformadas por el grupo de transformaciones formado por Translación, Rotación/Reflexión y Dilatación. Esto significa que se compara lo que es invariable: la ***forma*** que queda después de transformadas las configuraciones usando ese grupo de transformaciones.

Significativamente, ***estas comparaciones son realizadas sin ninguna referencia a los métodos que han generado las configuraciones: se trata, por lo tanto, de operaciones de interpretación realizadas de modo independiente del método que ha generado los resultados.***

Estos trabajos de Krzanowski han sido generalizados por MARTIN--RODRIGUEZ (1996, 2000), al mostrar que se aplican, casi sin modificación, a todas las configuraciones de marcadores de individuos y de variables en los distintos tipos de biplots (CMP, RMP, RCMP y biplots generalizados), obtenidos a partir de datos resultantes de observar los mismos individuos en instantes distintos. Ver también MARTÍN-RODRIGUEZ *et al* (2002).

Este capítulo se estructura según las líneas siguientes:

En el apartado 4.2 se identifican y caracterizan ***operaciones básicas de interpretación***, presentes no solamente cuando se intenta interpretar resultados de análisis de datos multivariantes, sino también al razonar sobre el mundo, de un modo general. Estas operaciones integran el lenguaje de

interpretación y en torno a ellas se estructura el ambiente interactivo del sistema de minería presentado en el **capítulo VI**.

En 4.3. se precisa el lenguaje y los límites de la investigación a realizar: ahí se presenta el concepto de interpretación a desarrollar y se dan los primeros pasos de la formalización. Se intenta mostrar que analizar datos multivariantes es establecer relaciones de equivalencia entre los individuos observados, y que interpretar esos resultados significa caracterizar esas clases de equivalencia o las particiones correspondientes, usando las variables observadas consideradas relevantes. Si es cierto que esta abstracción no cubre la totalidad de los problemas de interpretación, también es cierto que una parte muy relevante del problema de interpretación está incluida por ese paradigma.

En el apartado 4.4 el problema de interpretación es representado usando un grafo de intersección específico, construido con los átomos de significado asociados a un conjunto de datos: las expresiones atómicas del tipo ***(V= valor)***, en donde *V* es una variable observada.

En el apartado 4.5 se presenta la descripción de un método interactivo desarrollado para obtener aproximaciones del significado de resultados representados por marcadores en biplots, usando expresiones de tipo conjuntivo.

En el apartado 4.6 se analiza la posibilidad de usar el MDS clásico o métrico - por lo tanto un biplot - para desarrollar un algoritmo gráfico interactivo de aproximación de resultados por conceptos simples asociados a los valores de las variables observadas.

En el apartado 4.7 se explota la posibilidad de usar los árboles de regresión para expresar los resultados – representados por variables cualitativas - en función de las variables observadas relevantes.

Cuando se intenta interpretar un grupo usando exclusivamente el significado de las variables observadas, lo que se busca es expresar, de modo lo más aproximado que sea posible, el significado del grupo, usando el significado conocido de esas variables, de sus valores y de los individuos observados. Teniendo en cuenta que casi nunca se logra una aproximación perfecta, un modo de formular el problema es usar el concepto de conjunto impreciso (***rough set***) de PAWLAK (1991). Es este, esencialmente, el contenido del apartado 4.8

En síntesis, este capítulo pone de manifiesto que, lejos de ser un problema intratable y ajeno a la formulación matemática, una parte relevante de los problemas de interpretación de los resultados de análisis de datos multivariantes puede formularse objetivamente.

4.2. OPERACIONES DE INTERPRETACIÓN DE RESULTADOS.

Al razonar sobre lo que es la actividad de interpretación de resultados, surge la cuestión siguiente:

> *¿Qué es lo que un analista hace cuando intenta interpretar los resultados de los análisis de datos multivariantes - representados o no- sobre biplots?*

En otras palabras: ¿Existen operaciones universales o básicas tales que las operaciones más complejas de interpretación sean combinaciones de estas operaciones básicas?

En esta tesis se postulan las operaciones básicas siguientes:

Op1 – Identificación de un grupo.

Op2 – Caracterización de un grupo.

Op3 – Comparación de dos grupos.

Op4 – Creación de un concepto.

Op5 – Reconocimiento de un patrón en los datos.

Op1- **Identificación de un grupo.**

La identificación sobre el plano del biplot de un grupo de marcadores es algo básico y natural: si varios marcadores están próximos sobre un plano es natural que el sistema visual humano los considere como relacionados.

A su vez - ver **capítulo II** - la teoría de los biplots muestra que si los marcadores de las filas están próximos, esto significa que las filas correspondientes están próximas y que los individuos que esas filas representan son **semejantes**.

Es decir, la impresión visual tiene correspondencia con la realidad.

Con las variables esto también se verifica; ahora, a grupos de marcadores de variables «próximos», formando ángulos pequeños, en el biplot, corresponden variables muy correlacionadas en la realidad.

De un modo general, para las otras técnicas de análisis de datos multivariantes, en los razonamientos de interpretación, las categorías siempre presentes son grupos de objetos (individuos o variables), como se ha verificado en los ejemplos presentados en el **capítulo III**.

Op2- **Caracterización de un grupo.**

Después de reconocida la existencia de un grupo se sigue la caracterización de ese grupo.

La caracterización de un grupo puede atender a distintos criterios y realizarse de modos distintos, como:

- Búsqueda de las características comunes a todos los objetos del grupo.
- Búsqueda de lo que hay de común a la *mayor parte* de los objetos del grupo.
- Búsqueda del carácter distintivo del grupo.
- Búsqueda de lo que distingue un grupo del todo al que pertenece.
- Caracterización de un grupo por la media de las variables sobre el grupo.
- Caracterización de un grupo por una moda.

……

Op3- **Comparación de dos grupos.**

Dados dos grupos de objetos G_1 y G_2 ,

¿Qué es lo que distingue a G_1 de G_2?

¿Cuáles son las variables que más varían cuando se pasa de G_1 a G_2?

¿Cuáles son las variables que no varían cuando se pasa de G_1 a G_2?

Cuando razonamos sobre varias entidades, una actividad natural de la mente es intentar precisar «aquello» que distingue un caso de otro, un concepto de otro, un grupo de otro.

Esta actividad de interpretación corresponde a una actividad psicológica básica: la búsqueda de distinciones entre los objetos del razonamiento.

También en inferencia estadística, una de las actividades fundamentales es la operación de comparación de grupos usando los contrastes estadísticos apropiados.

Eventualmente, un proceso de comparación puede llevar a la conclusión de que no hay distinciones entre los dos grupos – lo que significaría que esos grupos podrían ser considerados como un único grupo.

Op4- **Creación de un concepto: atribución de un nombre a un grupo de objetos.**

La creación y caracterización de un grupo son estados intermedios en el proceso de creación de un concepto.

Cuando a un grupo, identificado y caracterizado, el investigador decide atribuirle un nombre (*label* o etiqueta) lo que sucede es que a ese grupo le está asociando el significado denotado por la etiqueta, nombre o *label*; el grupo pasa a tener un significado denotado por la etiqueta y es integrado en el conocimiento del analista.

Este acto equivale a establecer una conexión del grupo con el mundo real, asociando al grupo el significado que, en el mundo real, es denotado por el nombre, etiqueta o *label*.

La etiqueta o nombre no solo identifica y simboliza el nuevo concepto – grupo con significado - sino que substituye ese **significado** en todas las referencias al concepto.

Este acto de etiquetar o **nombrar** al grupo marca la intervención del conocimiento del mundo – más allá de los datos – por parte del investigador. Significa introducción de información exterior a los datos en el proceso de interpretación.

Op5- **Reconocimiento de patrones en los datos.**

Un patrón es un concepto predefinido o modelo de comportamiento estadístico, de propiedades matemáticas conocidas.

Una de las actividades básicas de la estadística es intentar descubrir si existen patrones de dependencia lineal o dependencia condicional en grupos de variables.

El primer paso es buscar grupos de observaciones/individuos que sirvan de soporte a esos patrones en los datos.

Un patrón es una regularidad en los datos o, en un contexto más preciso, una regularidad con propiedades conocidas y características.

El sistema visual humano es un poderoso sistema de reconocimiento de patrones que permite identificar patrones incluso donde no existen.

En el contexto de la interpretación esta actividad de «reconocer» patrones (eventualmente inexistentes) permite concentrar la atención y los recursos («llamar la atención») sobre posibilidades de patrones que un análisis más profundo puede mantener o descartar.

4.3. FORMULACIÓN DEL PROBLEMA DE INTERPRETACIÓN.

4.3.1. DEFINICIONES Y PRINCIPIOS BÁSICOS.

Las cuestiones de interpretación y de búsqueda de significado siempre han sido cuestiones sensibles en Lógica, Filosofía y Ciencia.

El objetivo de este apartado es reducir, lo más posible, las ambigüedades sobre el concepto de interpretación de resultados usado en este trabajo.

Se adopta la definición siguiente de análisis de datos:

Definición 4.2.1. (Análisis de Datos)
Proceso que tiene como objetivo obtener el significado de los datos.

Justificación:
Los análisis de datos no son un fin en sí mismos; todo análisis de datos tiene un objetivo específico y busca respuestas en los datos para cuestiones practicas en diversos dominios de la realidad. Eso quiere decir que las

respuestas no pueden olvidar las cuestiones formuladas y éstas dependen de los problemas de quien las formula.

Definición 4.3.2. (Interpretación)

Interpretar los resultados de los análisis de datos multivariantes significa integrar en una síntesis significativa, expresada en un lenguaje próximo del lenguaje humano, los resultados y conocimientos obtenidos en los análisis a los que se ha sometido un conjunto de datos.

Justificación:

El destinatario final de estos resultados es casi siempre alguien que no sabe estadística; alguien que no conoce el lenguaje y las particularidades de los métodos estadísticos. Por eso, interesa que los resultados sean expresados en un lenguaje próximo del lenguaje que los seres humanos, no especializados en estadística, puedan usar en sus razonamientos.

Principio Número 1 - Principio Básico de Interpretación.

La interpretación de un resultado debe ser expresada en función del significado de las variables observadas, del significado de los individuos observados y del significado de los valores observados, integrantes del conjunto de datos.

Justificación:

En este trabajo se asume que los únicos soportes de significado son los individuos, las variables y los valores que las variables toman sobre los individuos.

No se considera la información que el analista pueda poseer acerca de los datos más allá de los propios datos.

Todas las sugerencias de interpretación generadas por los programas deben ser expresadas en función de los datos.

Si los datos existen es porque han sido observadas unas variables sobre ciertos individuos específicos. Esto presupone que el significado de las variables es conocido de modo explícito o implícito: es conocido el instrumento de observación o medición, sus capacidades, sus limitaciones; es conocido el proceso de observación; son conocidos los valores resultantes del proceso de observación y su significado.

Del mismo modo, si ciertos individuos han sido observados es porque su significado es conocido de modo implícito o explícito.

O sea, se asume que, asociada a los identificadores de las variables y de los individuos existe información (meta-información) que puede estar explícita o no, y que traduce o expresa el significado de las variables y de los individuos. Referir el identificador de una variable o de un individuo equivale a referir el significado del individuo o de la variable.

Principio Número 2 - Asimetría Individuos vs Variables.

Desde el punto de vista de la interpretación de los resultados, las variables y los individuos no tienen un papel simétrico: si los individuos representan puntos de un espacio, las variables representan funciones definidas sobre esos puntos.

Justificación:

El punto de partida para todo análisis de datos multivariantes es un conjunto o matriz de datos de dos o tres vías que resulta de observar los

valores que las variables de un conjunto asumen sobre un conjunto de individuos, en una o más ocasiones.

Desde el punto de vista de la interpretación, por *comportamiento no simétrico* queremos significar la misma distinción que existe entre substantivos y adjetivos. Los individuos corresponden a los sustantivos: es aquello de que se habla; las variables corresponden a los adjetivos: son las características de los individuos de los que se habla.

Los individuos, representados por los respectivos identificadores, son cosas únicas, invariables al tiempo y al lugar. Aunque dos de esos individuos tengan los mismos valores para las mismas variables eso no significa que sean la misma cosa. Significa que el proceso de observación no tiene la precisión suficiente para distinguirlos pero que podrían ser distinguidos si se consideraran más atributos o una precisión superior de los aparatos o instrumentos de observación.

Al revés, las variables son funciones que están definidas para todos los individuos del mismo tipo, observados o no observados.

Matemáticamente, una *variable* es efectivamente una función que atribuye a cada individuo un valor de un cierto conjunto de valores.

4.3.2. FORMULACIÓN DEL PROBLEMA DE INTERPRETACIÓN.

Hemos visto en el **capítulo III** que los resultados más importantes de casi todos los métodos de análisis de datos multivariantes son grupos de individuos o particiones del conjunto de individuos. Sintetizando: *analizar*

datos multivariantes es definir clases de equivalencia y particiones sobre el conjunto de individuos. Los individuos que integran la misma clase son indistinguibles por el método de análisis usado.

Cuando el resultado es un grupo de individuos podemos representar ese resultado por la función característica o una variable indicadora del grupo: si un individuo pertenece al grupo identificado la variable asume el valor *1*, sino, el valor *0*.

Cuando el resultado es una partición o relación de equivalencia, podemos representar el resultado por una *variable cualitativa cuyos valores son los identificadores de las distintas clases*: para todos los individuos de la misma clase, el valor de la variable es el identificador de la clase.

No todos los resultados obtenidos por los métodos de análisis de datos multivariantes pueden ser representados de este modo; pero esta clase es lo suficientemente importante para ser tratada de manera especial.

Todo lo que se sigue se aplica a los resultados que puedan ser expresados por grupos de individuos o por particiones del conjunto de individuos.

Se asume, también, informalmente, que interpretar los resultados es aplicar las operaciones básicas de interpretación definidas en el número 4.1., buscando expresar el significado de los resultados (grupos y particiones) en un lenguaje próximo del lenguaje humano.

Desde el punto de vista lógico, el significado de un grupo es la lista de individuos que integran el grupo; por lo tanto, cuando un método de

análisis de datos identifica un grupo de individuos, su significado en esta acepción está definido sin ambigüedades.

Una corta lista de individuos puede ser aceptable para una persona como expresión del significado de un grupo. Una larga lista con docenas, centenares, o miles de individuos es inaceptable e inútil.

Para conceptos cuyo significado es una larga lista de individuos, es necesario que ese significado formal sea expresado en un lenguaje próximo al lenguaje humano, de forma que el ser humano pueda razonar sobre el significado del resultado.

Matemáticamente, interpretar un resultado es expresar la variable cualitativa que representa ese resultado en función del significado de las variables observadas y del significado de los individuos.

Son estas ideas las que se pretende sintetizar en la definición siguiente:

Definición 4.3.2.1. (Problema de Interpretación)
Dado un resultado representado por una variable cualitativa R, definida sobre el conjunto de individuos, interpretar ese resultado consiste en aproximar el significado de las clases equivalencia definidas por R usando una función de las variables observadas relevantes, de modo que se maximice la calidad de la aproximación, medida por una función de pérdida adecuada.

Más específicamente, si *R* es una variable cualitativa representativa de un resultado, el problema de interpretación es el de obtener una función *f( )* tal

que $R \cong f(X_1,X_2,\ldots X_p)$, en que $X_1,\ldots X_r$ son las variables observadas consideradas relevantes para la interpretación del resultado.

Ejemplo 4.3.3.1.

Usando los datos de la **tabla 3.1.1**, supóngase que un análisis cluster ha identificado el grupo de alumnos que designamos provisionalmente por la variable cualitativa *R* tal que *(R=1)* para los alumnos del conjunto {1, 6, 9, 11, 21, 47} y *(R=0)* para los restantes. Abreviadamente, podemos escribir *(R=1)* = {1, 6, 9, 11, 21, 47}. Para un ser humano, este modo de expresar el significado de ese resultado es inaceptable; especialmente cuando el cardinal de *(R=1)* sobrepasa una media docena de ítems. Ver ANDERSON (1990).

Sea ahora $f(X_1, X_2, ...) = (G1A = 2) \wedge (G4A = 2)$.

Se verifica que esta expresión, obtenida automáticamente (véase 4.6 y 4.7) o por tentativa y error (véase 4.5) no coincide con (*R*=1) pero se aproxima lo suficiente para que, en el contexto de un análisis preliminar de datos, pueda ser aceptada como una descripción del resultado.

El error cometido en la aproximación $(R=1) \cong (G2A= 2) \wedge (G4A= 2)$ podría medirse, por ejemplo, por el número

$|(R=1) - (R= 1) \cap (G2A= 2) \cap (G4A= 2)| = |\{58\}| = 1$.

Mirando la composición de esta expresión conjuntiva, podría el analista concluir que se trata de los «Alumnos que conocen las tablas de contingencia» y considerar conveniente atribuir al resultado (*R*=1) la

designación de «Buenos alumnos», convirtiendo el grupo (R=1) en el concepto «Buenos alumnos. Conocen bien las tablas de contingencia».

Obsérvese que esta designación traduce la percepción que el analista pueda tener de f(G2A, G4A)= (G2A = 2) $\wedge$ (G4A = 2), considerando su experiencia personal, pero, desde el punto de vista del automatismo integrado en un programa, no tiene ninguna influencia en el proceso de la caracterización

(R= *1*) $\cong$ (G2A= 2) $\wedge$ (G4A= 2).

4.3.3. LENGUAJE PARA EXPRESAR INTERPRETACIONES.

Las variables, como se ha visto, son entidades con significado para todos los individuos de una población, observados o no; son conceptos generales. Los individuos, al revés, son entidades con una especificidad potencialmente infinita.

De aquí resulta que es más natural, cuando se intenta construir un lenguaje para hablar de los resultados, que ese lenguaje sea basado en las variables y sus valores.

Consideremos un resultado que corresponda a un grupo dado de individuos; por ejemplo, el vértice de un árbol obtenido por clasificación aglomerativa. El significado del grupo/vértice es la lista de los centenares de individuos que forman ese grupo. Pero este modo de expresar el significado no es ni práctico ni apropiado para el ser humano.

Al ser humano no le interesa tanto la *extensión* de los grupos o conceptos que los análisis puedan revelar sino una síntesis de ese significado. La *intención*, en el sentido aristotélico del término. Ver SOWA (2000).

Dado un conjunto de datos, sean X_1, X_2, …, X_p las variables observadas.

El lenguaje de interpretación de resultados que se asume es formado por disyunciones de expresiones del tipo conjuntivo, cada una de las cuales tiene la forma

$(X_{i1} = v_{i1}) \wedge (X_{i2} = v_{i2}) \wedge \ldots \wedge (X_{ik} = v_{ik})$, en que :

$$(i_1, i_2, ..., i_k) \subset \{1,2,...p\}$$

es un subconjunto de identificadores de variables y v_{i1}, v_{i2},…v_{ik} valores observados de esas variables.

4.4. REPRESENTACIÓN DEL PROBLEMA DE INTERPRETACIÓN USANDO GRAFOS DE INTERSECCIÓN.

4.4.1. GRAFOS DE INTERSECCIÓN.

En GROSS, *et al* (1999), puede verse la definición siguiente:

Definición (Grafo de Intersección)

Sea S una familia de subconjuntos de un conjunto A. El grafo de intersección para la familia S tiene un vértice por cada miembro de la familia F y dos vértices son adyacentes si los conjuntos correspondientes tienen una intersección no - vacía.

Se trata de un tópico reciente de la investigación en teoría de los grafos. La primera monografía sobre este asunto es McKEE y McMORRIS (1999).

En esa referencia - pag 4 - pueden verse las demostraciones de los resultados siguientes:

Teorema 1: Todo grafo es un grafo de intersección.

Teorema 2: Todo grafo G es el grafo de intersección de una familia de subgrafos de un grafo.

Ejemplo 4.4.1.1. (Tablas de contingencia como grafos de intersección)
Consideremos un conjunto de datos donde existen las variables $X \in \{1, 2, 3\}$, $Y \in \{a, b, c\}$.

La tabla de contingencia habitual entre estas dos variables es

	$Y = a$	$Y = b$	$Y = c$	
$X = 1$	f_{11}	f_{12}	f_{13}	$f_{1.}$
$X = 2$	f_{21}	f_{22}	f_{23}	$f_{2.}$
$X = 3$	f_{31}	f_{32}	f_{33}	$f_{3.}$
	$f_{.1}$	$f_{.2}$	$f_{.3}$	

Esta tabla también puede ser vista, del punto de vista del significado de las frecuencias absolutas que integran su cuerpo, como:

	(Y = a)	***(Y = b)***	***(Y = c)***	
(X = 1)	$\lvert(X= 1) \cap (Y= a)\rvert$	$\lvert(X= 1) \cap (Y= b)\rvert$	$\lvert(X= 1) \cap (Y= c)\rvert$	$\lvert(X = 1)\rvert=f_1.$
(X = 2)	$\lvert(X= 2) \cap (Y= a)\rvert$	$\lvert(X= 2) \cap (Y= b)\rvert$	$\lvert(X= 2) \cap (Y= c)\rvert$	$\lvert(X = 2)\rvert = f_2.$
(X = 3)	$\lvert(X= 3) \cap (Y= a)\rvert$	$\lvert(X= 3) \cap (Y= b)\rvert$	$\lvert(X= 3) \cap (Y= c)\rvert$	$\lvert(X = 3)\rvert= f_3.$
	$\lvert(Y= a)\rvert= f_{.1}$	$\lvert(Y= b)\rvert= f_{.2}$	$\lvert(Y= c)\rvert= f_{.3}$	

En la inspección anterior hemos hecho explícito el significado de las frecuencias absolutas conjuntas y marginales.

Olvidando por el momento las frecuencias absolutas, vemos que, subyacente a la tabla de frecuencias existe la tabla siguiente, en donde figuran ahora los conjuntos relacionados y, en lugar de operaciones de contar/sumar están las operaciones básicas de unión e intersección de conjuntos que las anteceden:

	(Y = a)	***(Y = b)***	***(Y = c)***	
(X = 1)	$(X= 1) \cap (Y= a)$	$(X= 1) \cap (Y= b)$	$(X= 1) \cap (Y= c)$	$U\,(X= 1) \cap (Y= y)$ $y \in \{a, b, c\}$
(X = 2)	$(X= 2) \cap (Y= a)$	$(X= 2) \cap (Y= b)$	$(X= 2) \cap (Y= c)$	$U\,(X= 2) \cap (Y= y)$ $y \in \{a, b, c\}$
(X = 3)	$(X= 3) \cap (Y= a)$	$(X= 3) \cap (Y= b)$	$(X= 3) \cap (Y= c)$	$U\,(X= 3) \cap (Y= y)$ $y \in \{a, b, c\}$
	$(Y= a) = (\bigcup_{i=1}^{3}(X = i) \cap (Y = a))$	$(Y= b) = (\bigcup_{i=1}^{3}(X = i) \cap (Y = b))$	$(Y= c) = (\bigcup_{i=1}^{3}(X = i) \cap (Y = c))$	

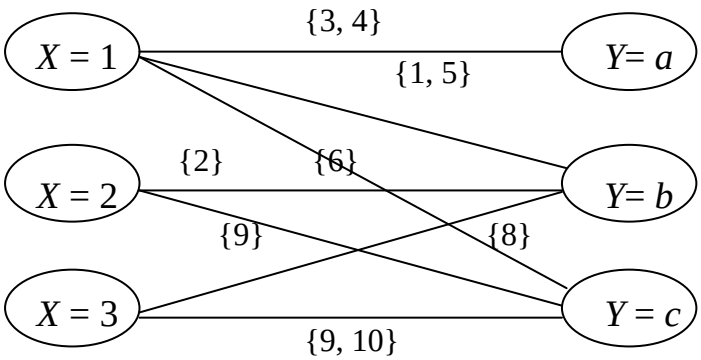
Esta tabla podría ser vista como una ***generalización de la tabla de contingencia*** una vez que la tabla habitual resulta de ella contando el número de elementos de los conjuntos que integran sus células, sin que

ocurra lo inverso: no podemos, dadas las frecuencias absolutas, recuperar los conjuntos.

Si la tabla es:

	(Y = a)	***(Y = b)***	***(Y = c)***	
***(X* = 1)**	{3, 4}	{1, 5}	{6}	{1, 3, 4, 5, 6}
***(X* = 2)**	∅	{2}	{8}	{2, 8}
***(X* = 3)**	∅	{7}	{9, 10}	{7, 9, 10}
	{3, 4}	{1, 2, 5, 7}	{6, 8, 9, 10}	

El grafo de intersección correspondiente es:



Este ejemplo permite verificar que una tabla de contingencia de dos vías puede representarse por un grafo de intersección cuyos vértices son los átomos correspondientes a las categorías de las dos variables en filas y columnas. Existe un arco entre dos de esas categorías cuando la frecuencia conjunta es superior a cero – lo que significa que esas dos categorías tienen intersección no vacía. No existen arcos entre vértices de la misma variable.

Las tablas de 3 y más vías pueden, de modo similar, ser representadas sobre grafos de intersección, considerando ahora caminos de 3 y más vértices con

intersección no vacía. Los cardinales de esas intersecciones son las frecuencias absolutas conjuntas de 3 o más categorías de las variables distintas a que pertenecen esos vértices; los cardinales de los vértices son las frecuencias marginales.

4.4.2. REPRESENTACIÓN DE UM CONJUNTO DE DATOS POR UN GRAFO DE INTERSECCIÓN.

Se ha visto en 4.3 la conveniencia de utilizar como lenguaje de interpretación un lenguaje que emplea expresiones conjuntivas.

Para facilitar los cálculos simbólicos inherentes a la interpretación, se asume, en lo que sigue, que el conjunto de datos está descompuesto en átomos del tipo $(X = v)$, en donde v es un valor observado de la variable X.

Se trata de una operación que puede realizarse en un solo pasaje secuencial sobre el conjunto de datos, mientras se van calculando las estadísticas básicas de las variables. En el **capítulo VI** puede verse el algoritmo respectivo.

Admitamos que un conjunto de datos resulta de observar el conjunto O de individuos, identificados por los números enteros $\{1,2, \ldots n\}$, en p variables. Sean $X_1, X_2,..X_p$ las variables observadas.

El significado, soporte o extensión del concepto $(X = v)$ es el conjunto
$\{o \in O: X(o) = v\}$

Una vez que el concepto (X = v) tiene el significado {o ∈ O: X(o) = v}, en este trabajo se resume ese hecho usando la abreviatura:
(X = v) = {o ∈ O: X(o) = v}.

Los conceptos de tipo *(X = v)* se designan atómicos porque su significado es simple, en el sentido de que no puede descomponerse. Los significados (extensiones) de los otros conceptos pueden expresarse o aproximarse usando el significado de estos conceptos simples mediante las operaciones de unión e intersección de conjuntos.

El conjunto de átomos correspondiente a una variable forma una partición de *O*, conjunto de objetos observados: si X_j es una variable y V_X es el conjunto de valores observados de X_j,

$$O = \bigcup_{v \in V_{X_j}} (X_j = v)$$

La expresión (X_j = *v*) representa también un predicado que puede verificarse o no para un objeto dado $o \in O$.

Como se ha visto, en este lenguaje, una expresión del tipo
$(X_1 = x_1) \wedge (X_2 = x_2)$
tiene por extensión o significado el conjunto de objetos que están en la intersección
$\{o \in O: X_1(o) = x_1\} \cap \{o \in O: X_2(o) = x_2\}$

y que, abreviadamente podríamos expresar por
$(X_1 = x_1) \cap (X_2 = x_2)$.

De estas consideraciones resulta la definición siguiente:

Definición 4.4.2.1. (Grafo de intersección asociado a la interpretación) *Dado un conjunto de datos $X_{(nxp)}$ resultante de observar los objetos identificados por O = {1,2…n}, en p variables X_1, X_2,..X_p. Se designa por Grafo de Intersección asociado a la interpretación el grafo G = (V, A) cuyos vértices - V - son los átomos en que ha sido descompuesto el conjunto de datos y cuyos arcos - A - son definidos del siguiente modo: existe un arco entre dos átomos (X_1= x_1) y (X_2= x_2), cuando (X_1= x_1) ∩ (X_2= x_2) no es vacío.*

Ejemplo 4.4.2.1.

Consideremos, por ejemplo, los «*datos*» artificiales siguientes

Objetos	***X***	***Y***	***Z***
1	*a*	2	t
2	*b*	1	q
3	*b*	1	q
4	*b*	1	q
5	*a*	2	t
6	*c*	1	q
7	*a*	2	t

Los átomos son:

(X= a) = {1, 5, 7}

(X= b) = {2, 3, 4}

(X= c) = {6}

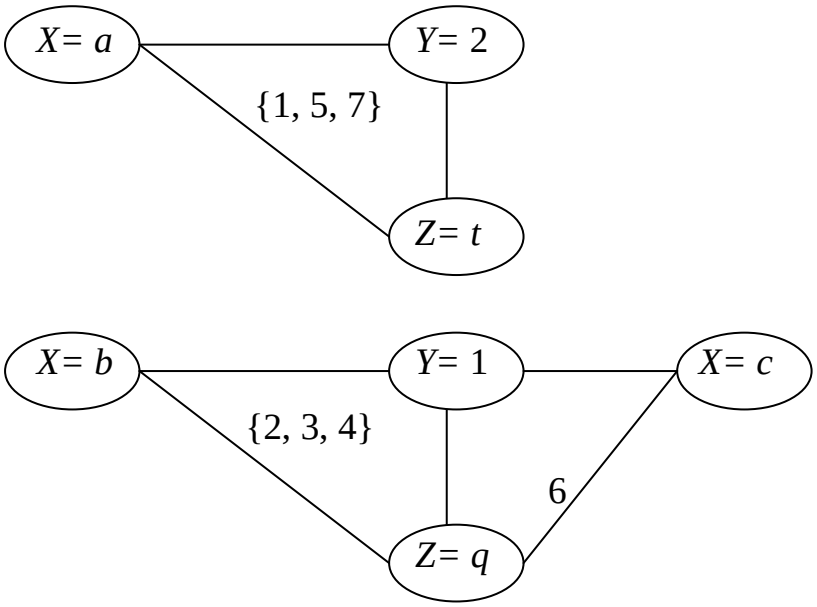
(Y=1) = {2, 3, 4, 6}

(Y= 2) = {1, 5, 7}

(Z= t) = {1, 5, 7}

(Z= q) = {2, 3, 4, 6}

El *grafo de intersección* correspondiente es:



Dada una expresión de tipo conjuntivo

$$(X_{i1} = v_{i1}) \wedge (X_{i2} = v_{i2}) \wedge ... \wedge (X_{ik} = v_{ik}),$$

en el grafo de intersección, a esa expresión le corresponde un camino de *k* vértices (átomos), con *k* <= *p*, de intersección no vacía. Eso quiere decir que todos los vértices correspondientes a esa expresión están conectados por arcos y que, por eso, los vértices correspondientes forman un *clique*[1] de orden *k*. Ver, por ejemplo, BERGE (1970), GROS *et al* (1999).

Recíprocamente, dado un *clique* de orden *k* <= *p*, de intersección no vacía, a ese *clique* corresponde una expresión disyuntiva, formada por átomos correspondientes a variables distintas. Ese *clique* describe un concepto con soporte en los datos.

Así, en el ejemplo anterior, existen 3 *cliques* de orden *k*= 3, correspondientes a conceptos cuyas descripciones y significado son:

[1] Un clique de orden *k* de un grafo es un subgrafo completo: todos sus *k* vertices están conectados por arcos (relacionados).

$(X = a) \wedge (Y = 2) \wedge (Z = t) = \{1,5,7\}$

$(X = b) \wedge (Y = 1) \wedge (Z = q) = \{2,3,4\}$

$(X = c) \wedge (Y = 1) \wedge (Z = q) = \{6\}$

Cuando las variables observadas son de tipo continuo las extensiones de los átomos tienen cardinales pequeños: las frecuencias absolutas de los valores observados son pequeñas y los cardinales de las expresiones atómicas correspondientes son pequeños. Esto puede ocurrir también para variables cualitativas con un número elevado de categorías.

En estos casos, el número de átomos es muy grande, dificultando la expresión del significado de los resultados usando las expresiones disyuntivas.

Puede ser necesario combinar valores de las variables observadas para formar categorías de "mejor poder expresivo", más adecuadas a las necesidades de expresión de la fase de interpretación.

Esto puede realizarse mediante dos operaciones: la unión de átomos de una misma variable y la concatenación de los valores de 2 o más variables.

Estas dos operaciones pueden realizarse fácilmente sobre el grafo de intersección anteriormente definido.

Así, para la ***unión de categorías***, supongamos que deseamos crear una nueva categoría de una variable *V* por unión de dos categorías existentes, como se puede ver en la **figura 4.4.2.1.**

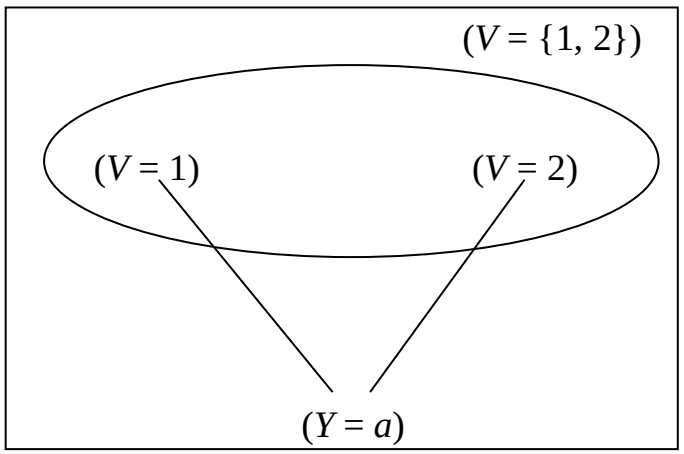


Figura 4.4.2.1. El nuevo átomo (V= {1, 2}) tiene por significado la unión de los significados de (V= 1) y (V= 2).

(V= {1,2}) es el nuevo átomo correspondiente al conjunto de valores anteriores.

Esta operación corresponde a formar un nuevo vértice

(V= {1,2}) = (V= 1) ∪ (V= 2).

Si las categorías *(V= 1) y (V= 2)* estuvieran conectadas al vértice (Y= a), entonces el nuevo vértice también quedaría conectado a (Y= a) y a este nuevo arco quedaría asociado el conjunto

$$(V= \{1,2\}) \cap (Y= a) = ((V= 1) \cup (V= 2)) \cap (Y= a)$$

$$= ((V= 1) \cap (Y= a)) \cup ((V= 2) \cap (Y= a))$$

Consideremos, ahora, la operación de ***concatenación de los valores*** de X con los valores de Y.

Admitamos que en el conjunto de datos existen las variables

$X \in \{1, 2, 3\}$

$Y \in \{a, b, c\}$

$Z \in \{u, v, w\}$

Si concatenamos las variables *X* y *Y*, obtenemos una variable *C* con valores:

$$C \in \{1*a, 1*b, 1*c, 2*a, 2*b, 2*c, 3*a, 3*b, 3*c\}.$$

El átomo ($C= x_i * y_j$) tiene como extensión al conjunto

$$\{o \in O: C(o)= x_i * y_j\}= \{ o \in O: (X(o) = x_i) \cap (Y(o)= y_j)\}.$$

Esto significa que si en el grafo de intersección, antes de la concatenación, no existe el arco (($X= x_i$) , ($Y = y_i$)), tampoco va a existir el vértice ($C = x_i * y_i$).

En otras palabras: los vértices correspondientes a la variable resultante de la concatenación son los arcos del grafo original que conectan los átomos de las variables a concatenar.

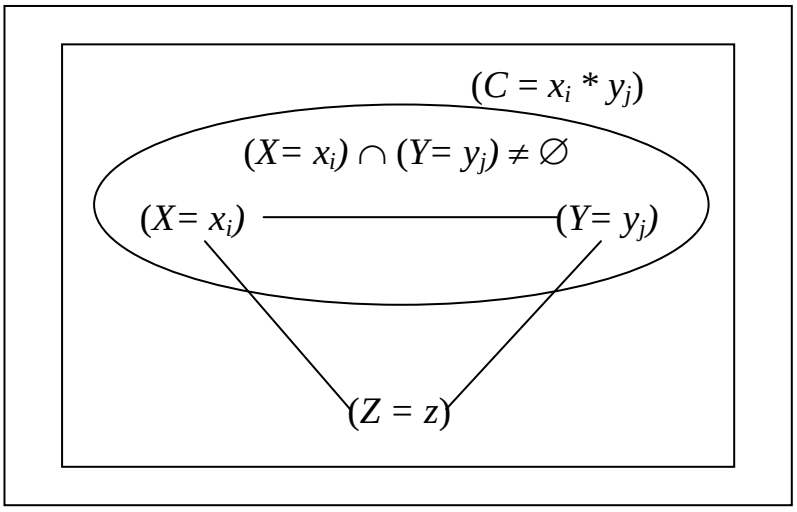
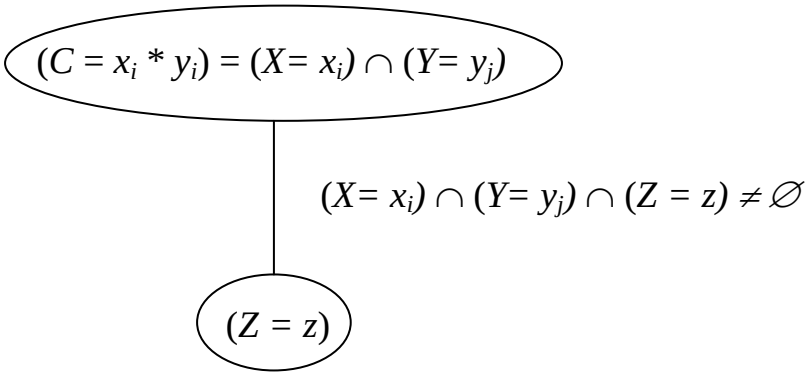


Figura 4.4.2.2. Si los átomos ($X= x_i$) y ($Y= y_i$) estan conectados por un arco, es posible que exista ($C= x_i * y_i$).

Después de concatenar:



Si existe el arco $(X= x_i) \cap (Y= y_j)$ entonces existe el nuevo vértice $C= x_i * x_j$.

El arco $((C= x_i * y_j) , (Z= z))$ existirá si $|(X= x_i) \cap (Y= y_j) \cap (Z= z)| > 0$.

4.4.3. EL PROBLEMA DE INTERPRETACIÓN EN EL GRAFO DE INTERSECCIÓN.

En el apartado 4.3.2 se ha definido el problema de interpretación como el problema de aproximar el significado *R* de un resultado – grupo de individuos o partición – usando el significado de las variables observadas.

Sean R_1, R_2, ..., R_r los resultados, representados por variables cualitativas, generados por un conjunto de análisis realizados sobre el mismo conjunto de datos, hasta un instante dado.

Teniendo en cuenta que es necesario expresar los significados/extensiones de estos resultados en función del significado de los átomos presentes en el grafo de intersección y relacionar los resultados unos con los otros, hay que

empezar por añadir al grafo de intersección actual los vértices correspondientes a las clases de equivalencia definidas por esos resultados.

Estos nuevos vértices no son considerados átomos dado que no corresponden a variables observadas y, por eso, sus significados son descomponibles. Sus significados deben ser expresados usando los significados de los átomos definidos por las variables observadas o los que han resultado de esos por las operaciones de concatenación o fusión.

Por ejemplo, $(Ri = r_i)$ tiene por significado formal la lista de individuos que han sido agregados en ese resultado por el particular método de análisis empleado. Pero $(Ri = r_i)$ no es un átomo porque que su significado va ser expresado en función del significado irreductible de los átomos correspondientes a variables observadas.

Dado que interpretar $(Ri = r_i)$ es aproximar su extensión/significado por los conceptos simples correspondientes a los átomos de las variables observadas, las variables relevantes para esa interpretación son aquellas que tienen átomos conectados al nuevo vértice $(Ri = r_i)$.

La aproximación que designamos por interpretación depende, ahora, del algoritmo particular elegido. En el apartado siguiente se presentan algunas posibilidades.

4.5. INTERPRETACIÓN INTERACTIVA DE GRUPOS.

En el apartado 4.3. hemos definido el problema de interpretación de un resultado R como el problema de aproximación del conjunto $\{o \in O: R(o)= 1\}$, usando una función del significado de los átomos de las variables observadas.

En este apartado se va a mostrar como se puede obtener una interpretación aproximada de un resultado interactuando con el sistema prototipo que describiremos en el **capítulo VI**.

Ese sistema interactivo permite realizar, dinámicamente, las operaciones siguientes (ver 4.4.2.):

- Descomposición del conjunto de datos en átomos de las variables observadas
- Creación del grafo de intersección
- Fusión de átomos y concatenación de variables
- Actualización del grafo de intersección con los resultados de los análisis
- Creación interactiva de expresiones disyuntivas, usando los átomos
- Visualizar las expresiones disyuntivas sobre el biplot corriente.

Dado el conjunto de datos $X_{n\times p}$, con individuos identificados por $O \in \{1, 2, ..., n\}$ y variables X_1 ... X_p, el esquema general del procedimiento interactivo referido es el siguiente:

- Crear el biplot.
- Si R es un resultado, representar ese resultado por una configuración de marcadores en el biplot.
- Sea $\{o \in O: R(o)= 1\}$ el significado formal del resultado. Una vez que esta representación del significado no es adecuada para el ser humano, es necesario traducir este significado por una expresión conjuntiva construida con átomos de las variables observadas consideradas relevantes para la interpretación.
- Observando en el biplot la posición de $(R=1)$, identificar las variables observadas que son relevantes para la interpretación. Dado que solo los átomos de estas variables relevantes interesan, esto reduce substancialmente la combinatoria del problema.
- De entre estas variables relevantes para la interpretación, elegir solamente aquellas cuyos átomos contengan totalmente o la mayor parte del significado del resultado. Esto permite reducir aún más la combinatoria del problema.
- Sean $(X_{i1}, ..., X_{ik})$ con $\{i_1 ... i_k\} \subseteq \{1 ... n\}$ las variables relevantes para la interpretación, detectadas en los pasos anteriores. Las expresiones disyuntivas que se buscan son construidas con los átomos de algunas de esas variables. Para descubrir cuales son las que interesan, pueden visualizarse sobre el biplot, con el auxilio del programa, los cierres convexos de esos átomos y sus intersecciones.
- Cuando, por examen visual, la expresión disyuntiva actual sea considerada suficientemente próxima del resultado $(R=1)$, el proceso es interrumpido.
- Sea $(X_{i1} = v_{i1}) \wedge ... \wedge (X_{ie} = v_{ie})$ con $\{i_1... i_e\} \subseteq \{1 ... p\}$ y $l \leq k \leq p$ la expresión disyuntiva que se ha obtenido. La expresión disyuntiva actual:

$$F\ (X_{i1} = v_{ie}) = (X_{i1} = v_{i1}) \wedge \ldots \wedge (X_{ie} = v_{ie})$$

es la interpretación del resultado.

Ejemplo 4.5.1.

Ilustremos los pasos del procedimiento anterior con los datos de la **tabla 3.1.1.**, usando el sistema prototipo. Ver Capitulo VI.

En la **figura 4.5.1.** aparece representado el HJ-biplot (RCMP), construido con las variables X_1 ... X_{10} de la **tabla 3.2.1.** En ese biplot los alumnos están representados por sus identificadores $\in \{1 \ldots 58\}$.

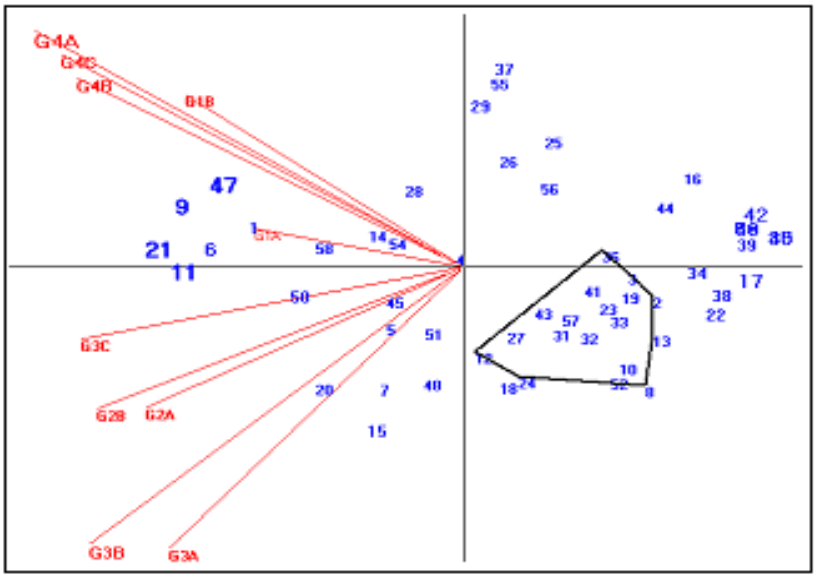


Figura 4.5.1. El cierre convexo identifica un resultado G obtenido por análisis cluster. Este resultado debe interpretarse usando las variables observadas.

Usando un algoritmo de clasificación aglomerativa, se ha identificado el grupo de 18 alumnos

$G = \{3, 8, 10, 12, 13, 18, 19, 23, 24, 27, 31, 32, 33, 36, 41, 43, 53, 57\}$

Interesa reemplazar esta lista - el significado del grupo *G* - por una expresión conjuntiva de átomos de algunas de las variables observadas que aproxima *G* «lo mejor posible».

Inspeccionando el biplot, se verifica que el grupo (*R*= 1) está situado en una zona de valores bajos para las variables que representan las preguntas del grupo IV - Tablas de Contingencia - y valores bajos para la pregunta sobre frecuencias relativas.

La posición de (*R*= 1) en el biplot sugiere, aún, que para algunos alumnos del grupo (*R*= 1) pueden verificarse resultados elevados en las variables G3B (grupo 3 - correlación) y gráficos de dispersión.

En suma, las variables candidatas a variables relevantes para la intersección son:

$\{\text{G3A , G3B , G4B , G4A}\} = \{X_5, X_6, X_8, X_9\} \subseteq \{X_1, ..., X_{10}\}$

(Ver **tabla 3.1.1.**)

Pintando, sucesivamente, en el biplot, los átomos de esas variables podemos verificar que el grupo (*G*= 1) está contenido en zonas homogéneas correspondientes a las variables X_6= G3B y a X_8= G4A.

Las variables que finalmente se consideran más relevantes para interpretar (*G*= 1) son:

$\{X_6, X_8\} \subset \{X_5, X_6, X_8, X_9\} \subset \{X_1, ..., X_{10}\}$.

Visualizando los cierres convexos de las expresiones,

(*G*= 1), (G3B $\in \{1, 2\}$) y (G4A= 0),

se obtiene la **figura 4.5.2.**

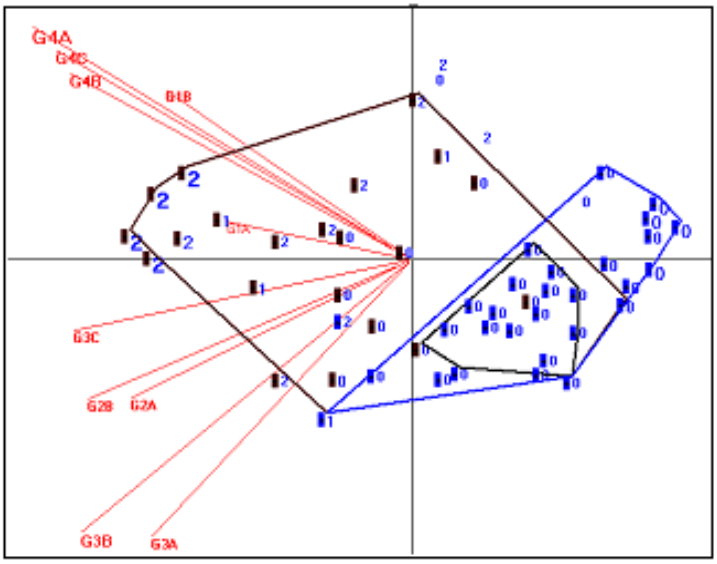


Figura 4.5.2. Visualización de átomos.

Calculando la intersección

$$(G = 1), (\text{G3B} \in \{1, 2\}) \cap (\text{G4A} = 0)$$

y vizualizando si cierre convexo, se obtiene la **figura 4.5.3.**

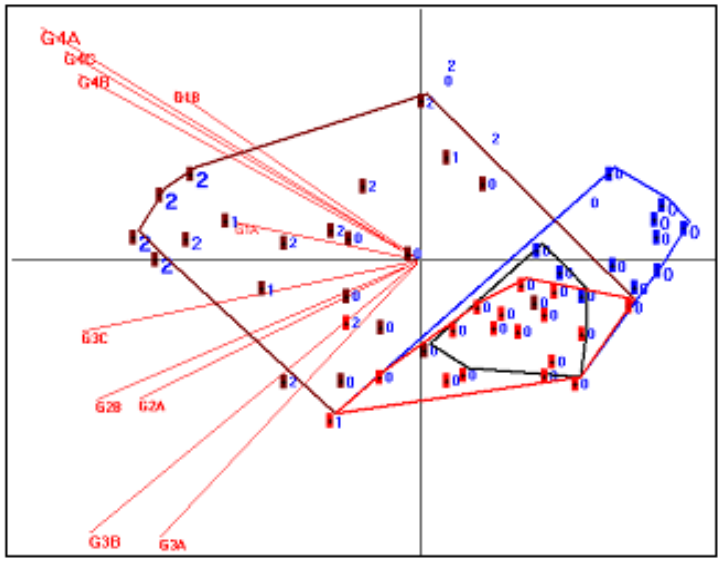


Figura 4.5.3. Visualización de la intersección de átomos relevantes para interpretar el grupo en negro.

La aproximación obtenida - en la figura anterior, corresponde al polígono rojo - es:

$$f(X_6, X_8) = (X_6 \in \{1, 2\}) \cap (X_8 = 0)$$
$$= \{5, 8, 10, 13, 15, 18, 19, 22, 24, 27, 31, 32, 33, 41, 43, 48, 52, 57\}$$

Se verifica que la aproximación *no es perfecta*:

$(G = 1) \neq f(X_6, X_8)$.

Se tiene:

$(G = 1) \cap f(X_6, X_8) = \{8, 10, 13, 18, 19, 24, 27, 31, 32, 33, 41, 43, 57\}$

Solamente 13 de los 18 alumnos están cubiertos por esta descripción.

Por otro lado, $f(X_6, X_8)$ incluye los alumnos 5, 15, 22, 48, 52 **que no pertenecen** a $(G = 1)$.

Este método no permite afirmar que la solución encontrada sea la mejor posible, pero es una descripción adecuada a la fase de análisis preliminar de datos, una vez que corresponde al 70% del verdadero significado de *G*.

La traducción al lenguaje natural podría ser:
«*El grupo G corresponde a alumnos que no saben de tablas de contingencia pero saben algo de correlación*».

La formulación de esta proposición en lenguaje natural representa la intervención en el proceso del conocimiento externo a los datos que el

analista posea. Pero, como se ha puesto de manifiesto, ese conocimiento no influye en la obtención de las expresiones conjuntivas.

4.6. INTERPRETACIÓN: APROXIMACIÓN DE RESULTADOS BASADA EN UNA MEDIDA DE AFINIDAD.

4.6.1. AFINIDAD ENTRE ÁTOMOS.

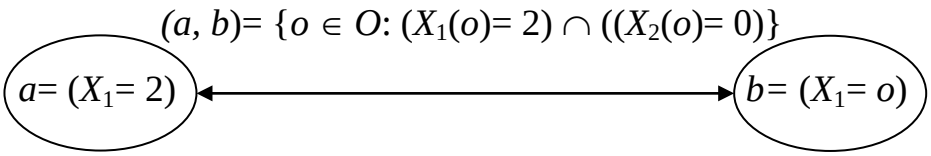
En el apartado 4.3.2. se ha definido el problema de interpretación como un problema de aproximación de conjuntos. Interpretar un resultado es aproximar las categorías del resultado usando categorías simples, cuyo significado se conoce: los átomos correspondientes a las variables observadas.

También se ha visto que este problema puede ser representado sobre un grafo de intersección. Buscar una interpretación en el lenguaje de las expresiones conjuntivas definidas por esos átomos es buscar caminos alternativos de intersección no vacía (cliques) en ese grafo de intersección.

En este apartado se propone un algoritmo que pretende facilitar la búsqueda de esos caminos.

En el grafo de intersección (ver 4.4) a cada arco está asociado el conjunto de individuos que tienen las propiedades definidas por los dos vértices conectados por el arco.

Por ejemplo, si a= (X_1= 2) y b= (X_2= o) están conectados por un arco (a, b)



a este arco esta asociada la lista

$\{o \in \text{Obj}: (X_1(o)= x_1) \wedge (X_2(o)= x_2)\}$.

Con base en estas intersecciones podemos definir, de muchas maneras, pesos asociados a los arcos.

En nuestro problema de interpretación de un resultado ($R= 1$) hay que obtener soluciones para la cuestión siguiente:

De entre todos los átomos que están conectados al resultado (R= 1) en el gráfico de intersección, ¿cuál debemos elegir como la mejor aproximación de (R= 1)?

Ver la **figura 4.6.1.1.** En esta figura hay que elegir entre (X_1=x_1) y (X_2= x_2).

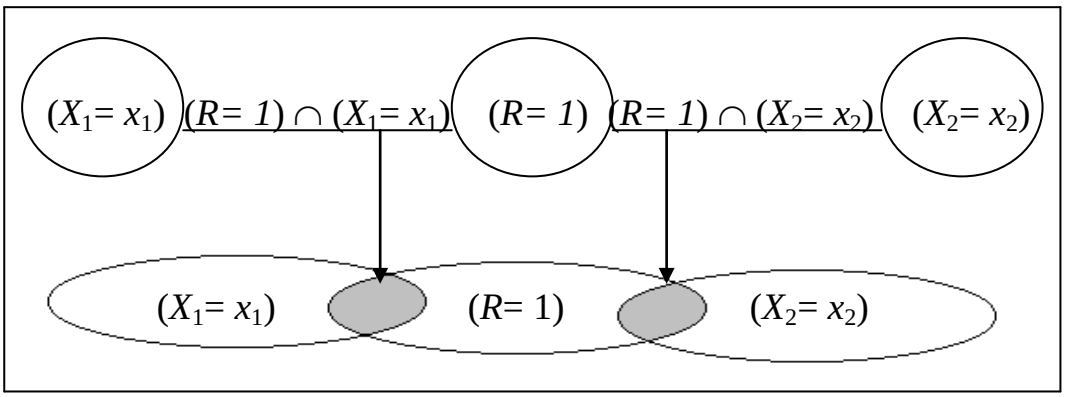


Figura 4.6.1.1. ¿De entre los dos átomos (X_1= x_1) y (Y_2= y_2) conectados con (R=1), cual es el que más aproxima (R= 1)?

Cuando se busca aproximar el significado de ($R= 1$) usando el significado de un átomo, la situación ideal corresponde a la existencia de uno o más átomos que coincidan con nuestro resultado. Si eso ocurre con el átomo

(X_0=x_0) entonces (R= *1*)= (X_0= x_0) y el significado de (R= *1*) es simplemente (X_0= x_0), para una variable X_0 y valor x_0 respectivo.

Por ejemplo: si (X_0= x_0) es el conjunto de personas con ojos azules, entonces el significado de (R= *1*) seria «Personas con ojos azules».

Lo que ocurre casi siempre es que un resultado puede ser parcialmente cubierto por átomos distintos.

Los átomos de una misma variable definen particiones del resultado, como se puede ver en la **figura 4.6.1.2.**

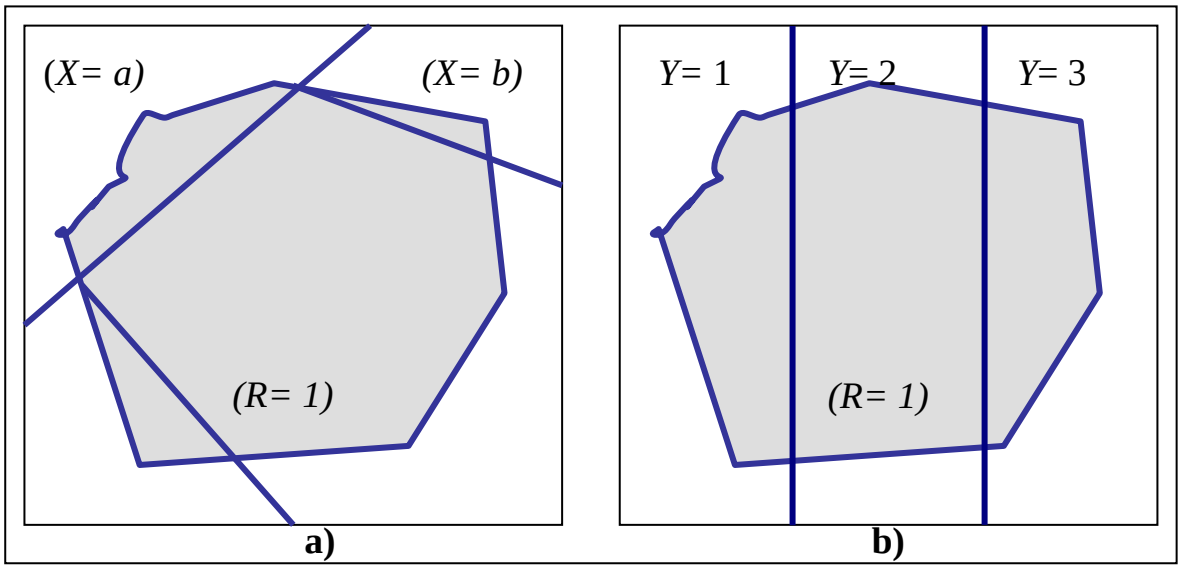


Figura 4.6.1.2. El resultado (R=1) está cubierto por los átomos de $X \in \{a, b, c\}$ y también por los átomos de la variable $Y \in \{1, 2, 3\}$.

Si (R = 1) intercepta por lo menos uno de los átomos de cada una de las p variables $\{X_1 \ldots X_p\}$ entonces existirían tantas particiones de (R= 1) como las variables.

Al elegir un átomo para aproximar el resultado (R= 1), interesa que éste sea cubierto por el átomo candidato «lo mejor que sea posible».

En otras palabras: el grado de cobertura de (*R*= 1) por (*X*= *x*) debe ser tan cercano a 1 como sea posible. Si definimos esa cantidad por:

$$Cobert\,(R=1)=\frac{|(R=1)\cap(X=x)|}{|(R=1)|}$$

ese valor debe ser tan cercano a 1 como sea posible.

La expresión anterior tiene el valor máximo cuando todos los elementos de (*R*= 1) - el significado del resultado - son elementos de (*X*= *x*); o sea, cuando los individuos de (*R*= 1) tienen la propiedad (*X*= *x*). En ese caso:

$$Cobert\,(R=1)=\frac{|(R=1)|}{|(R=1)|}=1$$

En síntesis: el significado de (*R*= 1) está contenido en el significado de (*X*= *x*).

Pero esto no implica, necesariamente, que la aproximación de (*R*= 1) por (*X*= *x*) sea «buena»: puede ocurrir que «muchos» de los objetos que forman el significado de (*X*= *x*) no pertenezcan al resultado (*R*= 1). Ver **figura 4.6.1.3.**

Dicho de otro modo: puede suceder que (*X*= *x*) sea poco específico como aproximación del significado de (*R*= 1).

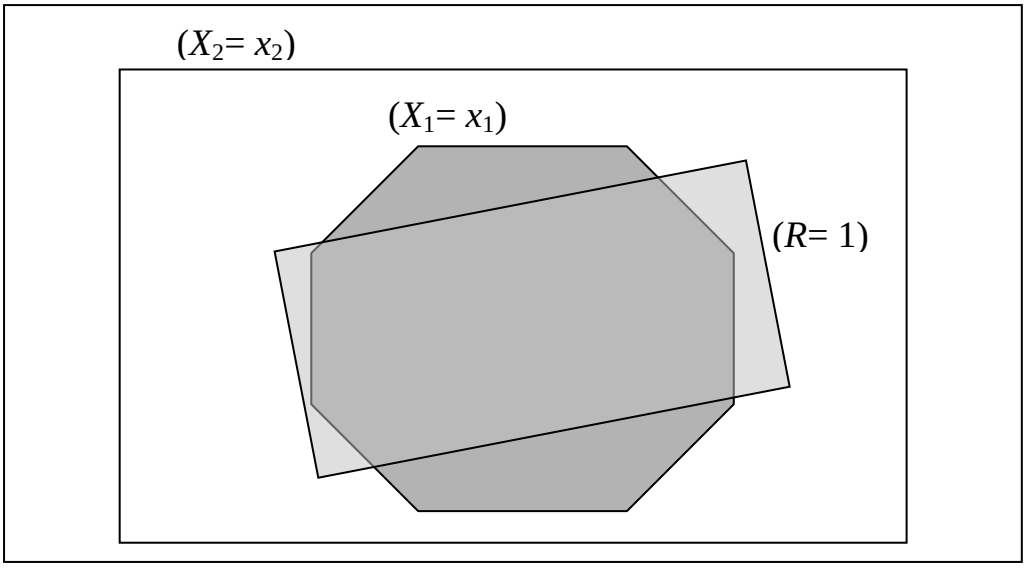


Figura 4.6.1.3. $(X_1 = x_1)$ aproxima mejor $(R = 1)$ que $(X_2 = x_2)$ aunque $(X_1 = x_1)$ no contenga $(R = 1)$

En consecuencia,

$$Esp(X = x) = \frac{|(R = 1) \cap (X = x)|}{|(X = x)|}$$

debe ser tan grande como sea posible.

Interesa que la cobertura de (R= 1) por (X= x) sea «buena» - próxima a 1 - pero interesa también que la especificidad de (X= x) en relación a (R= 1) sea tan grande como sea posible.

En síntesis: dado el resultado ($R = 1$) - cuyo significado buscamos aproximar por los átomos ($X = x$) - debemos elegir aquella variable y aquel valor respectivo tales que:

$$\frac{|(R = 1) \cap (X = x)|}{|(R = 1)|} \times \frac{|(R = 1) \cap (X = x)|}{(X = x)} = Esp(X = x) \times Cobert(R = 1)$$

sea tan grande como sea posible.

Formalizando estas ideas,

Buscar entre las variables X_1 ... X_p aquella variable X_0 y valor

$$(X_0 = x_0) = \underset{\substack{X \in \{X_1 \ldots X_p\} \\ x \in V_X}}{\arg\max} \frac{|(R=1) \cap (X=x)|^2}{|(R=1)| \times |(X=x)|}$$

respectivo x_0 tales que:

Este criterio significa que buscamos los átomos que tengan una **afinidad** máxima para el resultado (*R= 1*) que intentamos caracterizar. Esto no significa que, necesariamente, esos átomos estén contenidos en (*R= 1*). De estas consideraciones resulta:

Definición 4.6.1. (Afinidad)

Dada una familia de conjuntos *A*, sean *a, b* elementos de esa familia.

Se define afinidad entre esos dos conjuntos por la función

$$\tau : A \times A \longrightarrow R^+$$

$$(a, b) \longrightarrow \tau(a, b)$$

$$\tau(a,b) = \frac{|a \cap b|^2}{|a| \times |b|}$$

con

Si $a = \varnothing$ o $b = \varnothing$ se define $\tau(a, b) = 1$.

1. $\tau \geq 0$:

Una vez que $|a \cap b|$, $|a|$, $|b|$ son cantidades positivas o nulas,

$$\tau(a,b) = \frac{|a \cap b|^2}{|a| \times |b|} \geq 0$$

2. $\tau \leq 1$:

$$\frac{|a\cap b|^2}{|a|\times|b|}=\frac{|(a\cap b)|}{|a|}\times\frac{|(a\cap b)|}{|b|}.$$

Una vez que

$$|a\cap b|\leq|a| \qquad \text{y} \qquad |a\cap b|\leq|b|$$

entonces

$$\frac{|a\cap b|^2}{|a|\times|b|}\leq 1.$$

3. $(\tau = 1) \Leftrightarrow (a = b)$:

$$(\tau=1)\Leftrightarrow\left(\frac{|(a\cap b)|}{|a|}\times\frac{|(a\cap b)|}{|b|}\right)=1$$

La única posibilidad de que esto ocurra es:

$$\frac{|a\cap b|}{|a|}=1 \qquad \text{y} \qquad \frac{|a\cap b|}{|b|}=1.$$

Pero $\frac{|a\cap b|}{|a|}=1\Leftrightarrow(a\cap b=a)\Leftrightarrow(a\subseteq b)$

y $\left(\frac{|a\cap b|}{|b|}=1\right)\Leftrightarrow(a\cap b=b)\Leftrightarrow(b\subseteq a).$

Por lo tanto, $\left(\frac{|(a\cap b)|^2}{|a|\times|b|}=1\right)\Leftrightarrow(a=b)$

Esta función define una medida de similitud entre conjuntos. Ver, por ejemplo, COX & COX (1994).

Usando τ podemos, ahora, definir una medida de distancia entre conjuntos mediante la expresión $\delta(a, b) = 1 - \tau(a, b)$.

Por las propiedades de τ se verifica que

$(\delta \geq 0)$, $(\delta = 0) \Leftrightarrow (\tau = 1) \Leftrightarrow (a = b)$

$(\delta = 1) \Leftrightarrow (\tau = 0) \Leftrightarrow (a \cap b = \varnothing)$

Sin embargo, δ no es necesariamente una métrica: puede ocurrir que existan *a*, *b*, *c* tales que

$$1-\frac{|a\cap b|^2}{|a|\times|b|} > 1-\frac{|a\cap c|^2}{|a|\times|c|} + 1-\frac{|b\cap c|^2}{|b|\times|c|} = 2-\left(\frac{|a\cap c|^2}{|a|\times|c|}+\frac{|b\cap c|^2}{|b|\times|c|}\right).$$

Basta, para eso, que

$\frac{|a\cap c|^2}{|a|\times|c|}$ y $\frac{|b\cap c|^2}{|b|\times|c|}$ tengan valores cercanos de 1

y $\frac{|a\cap b|^2}{|a|\times|b|}$ sea cercano de cero.

Sin embargo, si $\delta(a, b)$ es reemplazado - para $a \neq b$ - por

$\theta(a, b) = 1 + \delta(a, b) = 2 - \tau(a, b)$

se obtiene, entonces:

$$2-\frac{|a\cap b|^2}{|a|\times|b|} \leq 4-\left(\frac{|a\cap c|^2}{|a|\times|c|}+\frac{|b\cap c|^2}{|b|\times|c|}\right)$$

y teniendo en cuenta que el valor mínimo del lado derecho es

4 - (1 + 1)= 2, valor que es siempre superior a

$$2-\frac{|a\cap b|^2}{|a|\times|b|} = \theta(a, b).$$

Esto significa que

$\theta(a, b) = 1 + \delta(a, b) = 2 - \alpha(a, b)$ $\quad (a \neq b)$

$= 0$ para $(a = b)$

Por tanto, cumple con la desigualdad del triángulo y es una métrica.

Ejemplo 4.6.1.1.

Volviendo a los datos de la **tabla 3.1.1.**, supongamos que deseamos caracterizar el resultado

$(R= 1)=\{3, 8, 10, 12, 13, 18, 19, 23, 24, 27, 31, 32, 33, 36, 41, 43, 52, 57\}$

$|(R= 1)|= 18$

En el grafo de intersección, existen los arcos indicados en la **figura 4.6.1.4.**

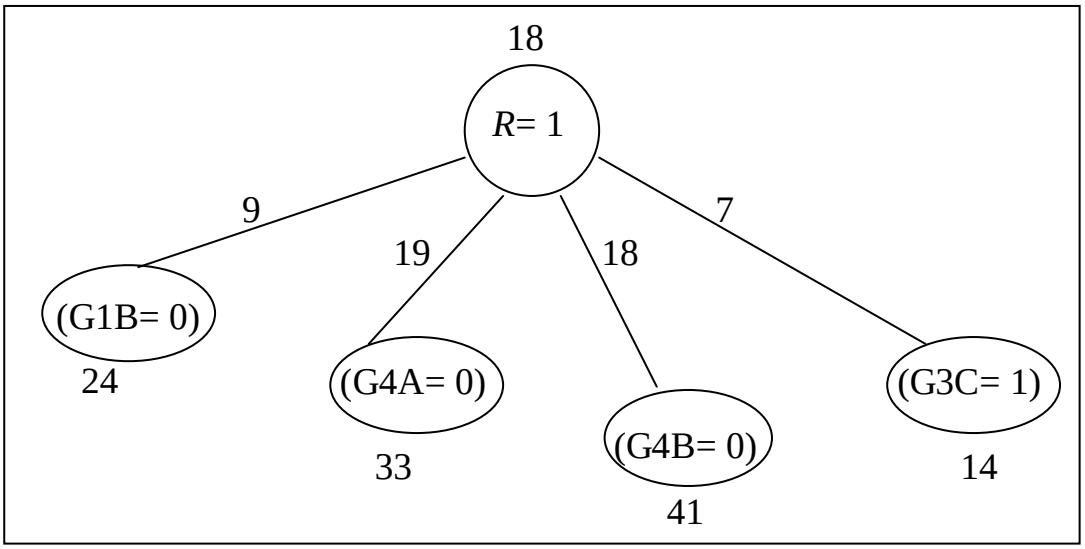


Figura 4.6.1.4. Los valores asociados a los arcos y vértices son los cardinales de los conjuntos respectivos.

Examinando la tabla de datos se puede escribir:

	Vecinos de (R= 1)				
	(G1B= 0)	**(G4A= 0)**	**(G4B= 0)**	**(G3C= 0)**	
(R= 1)	$\|(R=1)\cap(G1B=0)\|$ = 9	$\|(R=1)\cap(G4B=0)\|$ = 19	$\|(R=1)\cap(G4B=0)\|$ = 18	$\|(R=1)\cap(G3B=0)\|$ = 7	$\|(R=1)\|$ = 18
	$\|(G1B=0)\|= 24$	$\| (G4A=0)\|= 33$	$\| (G4B=0)\|= 41$	$(G3C=0)\|= 14$	

Los valores del coeficiente de afinidad $\alpha\,(a, b)$, de $\delta\,(a, b)$ y de $\theta\,(a, b)$ son:

	Vecinos de (*R*= 1)				
	(G1B= 0)	**(G4A= 0)**	**(G4B= 0)**	**(G3C= 0)**	
(*R*= 1)	α= 0.19 δ= 0.81 θ= 1.81	α= 0.43 δ= 0.57 θ= 1.57	α= 0.4 δ= 0.56 θ= 1.56	α= 0.19 δ= 0.81 θ= 1.81	|(*R*=1)| = 18
	|(G1B=0)|= 24	|(G4A=0)|= 33	|(G4B=0)|= 41	|(G3C=0)|= 14	

Se verifica que existen dos átomos (G1B= 0) y (G3C= 0) que maximizan el criterio δ, una vez que

$$\delta\,((G1B= 0)\,,\,(R= 1))= 0.81 \Leftrightarrow \theta\,((G1B = 0),\,(R= 1))= 1.81$$

$$\delta\,((G3C= 0)\,,\,(R= 1))= 0.81 \Leftrightarrow \theta\,((G3C = 0),\,(R= 1))= 1.81$$

En este caso, el mejor sería el que garantiza mejor cobertura.

O sea, (*G1B*= 0).

Interpretación probabilística del concepto de afinidad.

Sean a = (*R*= *1*) un resultado por interpretar y b = (*X*= *x*) un átomo candidato a aproximar el significado de (*R* = 1).

Entonces

$$\tau(a,b)=\frac{|(a\cap b)|^2}{|a|\times|b|}=\frac{|(a\cap b)|}{|a|}\times\frac{|(a\cap b)|}{|b|}.$$

Podemos interpretar

$\frac{|(a\cap b)|}{|a|}$ como

$$P(b|a) = P((X = x)|(R = 1)) = \frac{P((X = x) \cap (R = 1))}{P(R = 1)}$$

probabilidad, dado el significado de (*R= 1*), de que, elegido un elemento al azar entre los que pertenecen al significado de (*X=x*), este tenga el significado (*R= 1*).

De modo similar, podemos interpretar

$\frac{|(a \cap b)|}{|a|}$ como

$$P(a|b) = P((R = 1)|(X = x)) = \frac{P((X = x) \cap (R = 1))}{P(X = x)}$$

Entonces

$$\tau(a,b) = P(a|b) \times P(b|a) = \frac{P(a \cap b)}{P(b)} \times \frac{P(a \cap b)}{P(a)} = \frac{P^2(a \cap b)}{P(b)\,P(a)}.$$

4.6.2. ALGORITMO DE APROXIMACIÓN BASADO EN LA MEDIDA DE AFINIDAD.

Consideremos la situación representada en la **figura 4.6.2.1.**

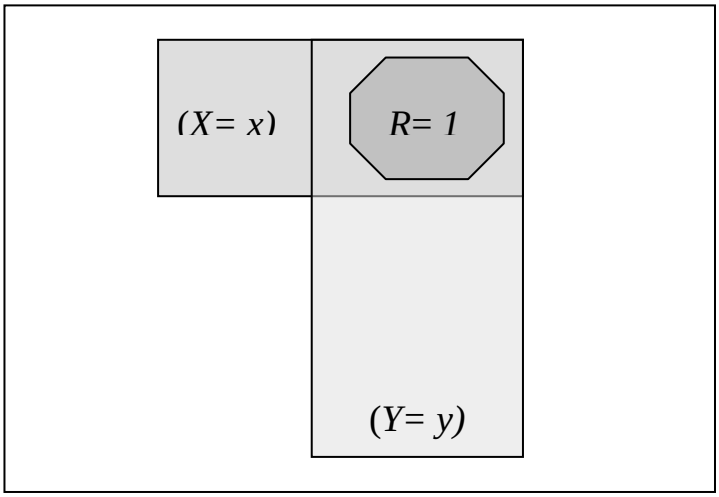


Figura 4.6.2.1. (*X= x*) aproxima (*R*= 1) mejor que (*Y= y*). Pero (*X= x*) ∩ (*Y= y*) es aún mejor.

La aproximación de (R= *1*) por (X= x) es mejor que la aproximación de (R= 1) por (Y= y).

En los dos casos, $(R= 1) \subseteq (X= x)$

y $(R= 1) \subseteq (Y= y)$

pero (X= x) es una descripción mucho más específica de (R= *1)* que la descripción (Y= y).

Si hay que elegir entre $a = (X= x)$ y $b = (Y= y)$ deberíamos preferir a, ya que $\alpha((R= 1), a) \geq \alpha((R= 1), b)$.

Pero eso no significa que hayamos encontrando la mejor solución.

Como puede verse en la **figura 4.6.2.1.**, (X= x) $\cap$ (Y= y) es una solución aún mejor que (X= x).

En efecto,

$$\tau((R= 1), (X= x) \cap (Y= y)) = \frac{|(R=1) \cap (X= x) \cap (Y= y)|^2}{|(R=1)| \times |(X= x) \cap (Y= y)|}.$$

Examinando la **figura 4.6.2.1.**, en donde el área representa el cardinal, el grado de cobertura de (R= 1) por (X= x) $\cap$ (Y= y) es el mismo que el obtenido por (X= x) - o sea: 1; la especificidad de (X= x) $\cap$ (Y= y) es más grande que la especificidad de (X= x) o la de (Y= y).

Entonces

$\tau((R= 1), (X= x) \cap (Y= y)) \geq \tau((R= 1), (X= x)) \geq \tau((R= 1), (Y= y))$

y, por eso, δ y θ tienen valores inferiores.

Esto implica que el significado de (*R= 1*) queda, en este caso, mejor aproximado, según el criterio de afinidad, por el significado de la expresión conjuntiva ($X= x$) $\wedge$ ($Y= y$) que por el significado de la expresión ($X= x$) o por la del significado ($Y= y$).

Lo que este ejemplo evidencia es que, si buscamos la expresión conjuntiva que aproxima lo mejor posible el resultado (*R= 1*), seria necesario examinar todas las $\sum_{i=1}^{v}\binom{v}{i}=2^v-1$ expresiones formadas por 1, 2, ..., *v* átomos, en que *v*= número total de átomos que es posible formar con las $\{X_1, ..., X_p\}$ variables.

Entre estos átomos no interesan las expresiones conjuntivas que tengan más de un átomo de la misma variable; para esos átomos, ya que, para ($a\neq b$), ($X= a$) $\cap$ ($X= b$) = $\varnothing$, se tiene α (a, b)= 0 y, no es necesario ningún examen.

Esto significa que las expresiones conjuntivas que interesan tienen, como máximo, *p* átomos.

La búsqueda sería, entonces, entre las

$$\binom{v}{1}+\binom{v}{2}+...+\binom{v}{p}=\sum_{i=1}^{p}\binom{v}{i}$$ combinaciones.

Por consideraciones psicológicas relacionadas con la capacidad de la memoria de trabajo - ver ANDERSON (1990) - no debemos incluir en las expresiones conjuntivas más que un número máximo *c* de átomos. Por ejemplo $c = 5$.

En consecuencia, las expresiones conjuntivas que interesa examinar tienen un número de átomos *min (p, c).*

Eso significa que el número de expresiones a examinar se reduce a

$$\sum_{i=1}^{\min(c,p)} \binom{v}{i}.$$

Por ejemplo, si p = 10 variables, cada una con tres valores y v= 30.

Fijando c= 5, entonces min (5, 10)= 5, y el número de expresiones a examinar sería:

$$\binom{30}{1}+\binom{30}{2}+\binom{30}{3}+\binom{30}{4}+\binom{30}{5} = 174\ 436 \text{ expresiones.}$$

!Lo que es aún demasiado!

La explosión combinatoria implica que es necesario un algoritmo de búsqueda aproximada o heurística que, aunque no garantice la solución optima, produzca una solución satisfactoria.

Una posibilidad para esa heurística es la siguiente:

Algoritmo 4.6.2.1.

1. Sean V_1, V_2, ..., V_k los vecinos de (R= 1) en el grafo de intersección, átomos con los cuales (R= 1) tiene intersección no vacía.

2. Buscar los c vértices más cercanos de (R= *1*) según el criterio de distancia

$$\theta(a, b) = 1 + \delta(a, b) \text{ con } \delta(a,b) = 1 - \frac{|(a \cap b)|^2}{|a| \times |b|} .$$

Sean esos vértices:

$V_{(1)}$, $V_{(2)}$, ..., $V_{(c)}$

ordenados según el valor decreciente de $\theta(a, b)$.

3. Considerar todas las expresiones conjuntivas que se pueden formar con 1, 2, 3, ..., c átomos elegidos entre los $V_{(1)}$, $V_{(2)}$, ..., $V_{(c)}$. Elegir aquellas que producen el valor máximo de θ.

Por ejemplo, si c= 5, este algoritmo examina solamente

$$\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5}=31 \text{ expresiones conjuntivas.}$$

Sea $C_0 = (X_{i1} = x_{i1}) \wedge ... \wedge (X_{ik} = x_{ik})$ con $\{i_1 ... i_k\} \subseteq \{1...c\}$.

Si el significado de C_0 es representado por $(X_{i1} = x_{i1}) \cap ... \cap (X_{ik} = x_{ik})$ entonces (R= *1*) $\cap$ C_0 queda bien aproximado por C_0.

Si aplicamos a (R= *1*) - ((R= *1*) $\cap$ C_0) = (R= *1*) $\cap$ $(\overline{C_0})$ el mismo **algoritmo 4.6.2.1.**, sea C_1 la nueva expresión conjuntiva.

Si $C_0 \cup C_1$ cubre ($R= 1$) entonces la interpretación de ($R= 1$) seria $C_0 \vee C_1$.

El procedimiento puede aplicarse sucesivamente hasta garantizar la cobertura de ($R= 1$) por $C_0 \vee C_1 \vee ... \vee C_k$.

Una vez que lo que se pretende es una expresión aproximada del significado de ($R= 1$), ocurre frecuentemente que la primera expresión - C_0 - obtenida por el **algoritmo 4.6.2.1.** puede ser suficiente.

Ejemplo 4.6.2.1.

Con los datos de la **tabla 3.1.1.** se ha obtenido el biplot GALINDO de la figura siguiente, trabajando con datos reducidos.

Considerando que ($R= 1$) es uno de los vértices de un análisis cluster, supongamos que deseamos interpretar

($R= 1$)= {3, 8, 10, 12, 13, 18, 19, 23, 24, 27, 31, 32, 33, 36, 41, 43, 52, 57}

Los $c= 5$ vértices más próximos de ($R= 1$) y los valores del criterio de afinidad son:

	Átomo	|Átomo|	θ
$V_{(1)}=$	($X_9= 0$)	41	1.561
$V_{(2)}=$	($X_8= 0$)	33	1.569
$V_{(3)}=$	($X_{10}= 0$)	10	1.621
$V_{(4)}=$	($X_4= 0$)	40	1.728
$V_{(5)}=$	($X_6= 0$)	14	1.810

Examinando estos vértices se verifica que el «mejor» es ($X_9= 0$) con θ=1.561.

El siguiente mejor es $V_{(2)}$= (X_8= 0) con θ= 1.569.

En la figura siguiente están representados los cierres convexos de (R= *1*) en negro, (X_9= 0) en marron y (X_8= 0) en verde.

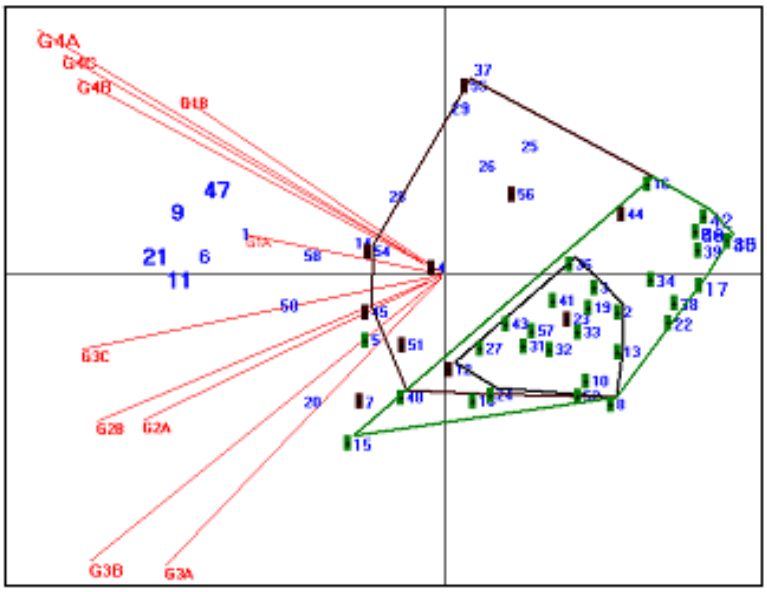


Figura 4.6.2.2. Un resultado (negro) cubierto por dos átomos: (X_9= 0) y (X_8= 0).

Se observa que la intersección (X_9= 0) ∩ (X_8= 0) aún cubre (R= *1*) y es más específico.

Por eso, (X_9= 0) ∧ (X_8= 0) es una expresión que describe (R= *1*) mejor que (X_8= 0) o que (X_9= 0).

Realizando el examen de las

$$\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5}= 31 \text{ expresiones conjuntivas,}$$

se verifica, como se observó anteriormente, que la mejor es C_0= (X_8= 0) ∧ (X_9= 0) y corresponde a θ= 1.541 < 1.561, valor que correspondía a (X_9= 0) y 1.541 < 1.569, valor que correspondía a (X_8= 0).

El significado de $(X_8 = 0) \wedge (X_9 = 0)$ es el conjunto:

{2, 3, 8, 10, 13, 16, 17, 18, 19, 22, 24, 27, 30, 31, 32, 33, 34, 35, 36, 38, 39, 40, 41, 42, 43, 46, 48, 49, 52, 53, 57}.

Entonces

$(R = 1) \cap C_0$= {3, 8, 10, 13, 18, 19, 24, 27, 31, 32, 33, 36, 41, 43, 52, 57}, con $|(R=1) \cap C_0| = 16$.

El grado de cobertura de (R= 1) por C_0 es

$$\frac{|(R=1) \cap C_0|}{|(R=1)|} = \frac{16}{18} = 0.89.$$

Pero la especificidad de C_0 es solamente de

$$\frac{|(R=1) \cap C_0|}{|C_0|} = \frac{15}{31} = 0.53.$$

4.6.3. USO DEL MDS BASADO EN LA AFINIDAD PARA BUSCAR ASOCIACIONES DE VALORES DE VARIABLES CUALITATIVAS.

Las consideraciones hechas en 4.6.1. permiten definir, para toda la familia C de conjuntos, una topología métrica. O sea, transformar la operación de intersección de conjuntos en una noción de proximidad entre puntos de un espacio métrico y definir vecindades entre conjuntos.

En particular, sea C la familia de todas las extensiones (significados) de los conceptos atómicos, asociados a un conjunto de datos $X_{n \times p}$.

Sea $|C| = v$ = Número de átomos que pueden ser definidos con las variables $X_1...X_p$ y sus valores sobre los individuos 1...n

= Número de vértices del grafo de intersección.

Los elementos $a \in C$ son conjuntos de objetos, tales que:

$$a = \{o: X_j(o) = x_j, X_j \in \{X_1...X_p\}, o = 1...n, x_j \in V_{X_j}\},$$

en donde V_{X_j} es el conjunto de valores de X_j.

Si $a, b \in C$, entonces

$$\theta(a,b) = 2 - \frac{|a \cap b|^2}{|a||b|} = 2 - \frac{|(X_i = x_i) \cap (X_j = x_j)|^2}{|(X_i = x_i)| \times |(X_j = x_j)|} = \delta(b,a)$$

en donde X_i y X_j son dos variables observadas.

Sea la matriz cuadrada $A_{v \times v} = [\theta(a, b)]$ en que $a, b \in C$.

En el apartado 3, relativo al MDS - ver también COX & COX (1994), pag. 24 - se observó que, si transformamos la matriz $A_{v \times v}$ en una matriz B por doble centrado, se tiene:

$$B_{v \times v} = H A H^T \text{ con}$$

$$H = I - \frac{1}{n} 1 1^T \text{ y}$$

$$1^T_{1 \times v} = (1,1,...,1)^T .$$

Realizando la D.V.S. de B, se obtiene

$$B_{v \times v} = U_{(n \times v)} \wedge_{(r \times r)} U^T_{v \times r} = U \Sigma^2 U^T = (U \Sigma)(U \Sigma)^T = X X^T$$

en que

$\wedge$= diag (λ_i), i= 1, ..., r, siendo λ_i los valores propios de B

Σ= diag (σ_i), i= 1, ..., r, siendo σ_i los valores singulares de B

con $\sigma_i = \lambda^{1/2} \quad i = 1...r$

r = rango (B)

$U_{(v\times r)}$ = Vectores singulares de B

$X_{v\times r} = U\Sigma =$ Matriz cuyas filas son los marcadores de las filas (y columnas) de A en el RCMP biplot (Galindo).

Por este procedimiento se obtienen v marcadores para los átomos en un espacio euclídeo - cuyas distancias coinciden con las distancias iniciales, lo que corresponde a realizar un MDS métrico. Ver VICENTE-VILLARDÓN (1992).

Cuando, en lugar de un espacio de dimensión r = rango (B) se representan los átomos en un espacio de dimensión 2, el resultado es solamente aproximado.

En este biplot - correspondiente a un MDS métrico cuya métrica **no** resulta de ningún sistema de coordenadas sino de la topología definida por la distancia θ - las proximidades entre pares de marcadores de átomos reflejan las relaciones semánticas entre átomos sobre el grafo de intersección.

Así, a dos marcadores muy próximos deben corresponder, en principio, dos conceptos atómicos con significados (extensiones) conectados en el grafo de intersección por un arco (conjunto) cuyo cardinal debe ser «significativo».

A tres marcadores muy próximos en el biplot deben corresponder, sobre el grafo de intersección, un camino de intersección no vacía o *clique* de intersección no vacía.

Lo mismo para 4, 5, ..., *p* átomos.

Como se ha visto en el apartado anterior, no interesa examinar vecindades de más que *p* átomos dado que átomos de la misma variable tienen intersección vacía.

El hecho de que dos marcadores estén muy próximos en el biplot no significa, necesariamente, que correspondan a átomos muy próximos en realidad: es necesario recordar que el biplot de 2 dimensiones es solo una aproximación; puede ocurrir que dos átomos muy alejados en el espacio de dimensión r = rango(B) tengan proyecciones cercanas en el plano. Lo inverso puede también ocurrir.

Pero este procedimiento tiene interés una vez que **llama la atención** del analista para las zonas de alta densidad en donde es más probable encontrar cadenas de átomos correspondientes a cliques de intersección no vacía.
Usando, ahora, las coordenadas de la configuración de marcadores representados por las filas de $X_{(v\times r)}$ es también posible construir un árbol por cluster aglomerativo que, mecánicamente, **llame la atención** del analista para ciertos grupos de átomos que deben ser inspeccionados.

Explicitando esta idea, supongamos que, visualmente hemos identificado un conjunto de 4 átomos $\{a, b, c, d\}$ muy próximos.

Entonces hay que examinar todas las

$$\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4}=15 \text{ combinaciones de átomos}$$

$\{a, b, c, d, (a\wedge b)\ , (a\wedge c)\ , (a\wedge d)\ , (b\wedge c)\ , (b\wedge d)\ , (c\wedge d)\ , (a\wedge b\wedge c)\ ,$
$, (a\wedge c\wedge d)\ , (a\wedge b\wedge d)\ , (b\wedge c\wedge d)\ , (a\wedge b\wedge c\wedge d)\}$

para verificar si corresponden a conceptos «interesantes» - con soporte «significativo» en el conjunto de datos.

Este procedimiento puede tener interés en si mismo y puede ser visto como una realización visual[2] de aquello que se designa por «algoritmo de análisis de la cesta de compras». AGRAWALL *et al* (1993).

En el contexto de nuestra investigación, lo que nos interesa es la búsqueda de cadenas de átomos que puedan servir de aproximaciones para los resultados obtenidos por otros análisis.

Ejemplo 4.6.3.1.

Realizando un RCMP-biplot de las distancias entre los átomos correspondientes a los resultados de la **tabla 3.1.1.**, se obtiene la **figura 4.6.3.1.** en donde se han superpuesto los cierres convexos de una partición obtenida por análisis cluster aglomerativo, usando el criterio de Ward.

[2] Este procedimiento puede, evidentemente, servir de base a un algoritmo puramente automático, no visual, de búsqueda de asociaciones.

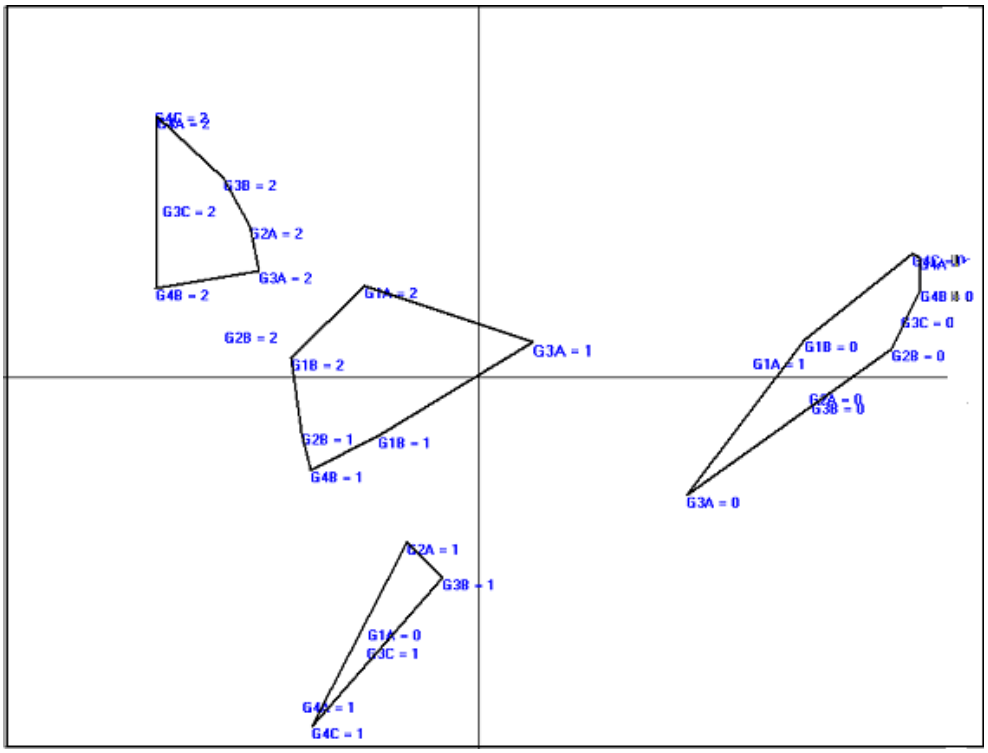


Figura 4.6.3.1. MDS obtenido con base en la afinidad entre átomos, usando el biplot de Galindo.

En esta figura se sugieren posibles asociaciones a investigar.

Por ejemplo:

$(G4C = 2) \wedge (G4A = 2)$

$(G1A = 1) \wedge (G1B = 0)$.

4.7. LA INTERPRETACIÓN COMO UN PROBLEMA DE REGRESIÓN.

Se ha visto en el apartado 4.6. una heurística para obtener aproximaciones del significado de un resultado usando una medida de afinidad entre los átomos de las variables observadas y el resultado.

Una alternativa a esa heurística se basa en la formulación del problema de interpretación - **definición 4.3.2.1.** - como un problema de regresión cualitativa.

El objetivo es aproximar, mediante un árbol de regresión, las clases definidas por una variable resultado *R* usando los átomos definidos por las variables observadas.

Específicamente, consideremos un resultado representado por una variable cualitativa *R* cuyos valores $\{r_1, r_2, ..., r_{1k}\}$ identifican o corresponden a clases disyuntivas de individuos. O sea:

$$(R= r_e)= C = \{o \in O: R(o)= r_e\}, \qquad e= 1, 2, ..., k.$$

En el caso particular de que el resultado sea un grupo,

$$R \in \{0, 1\}$$

$$(R= 1) = C_1= \{o \in O: R(o)= 1\}$$

$$(R= 0) = (R= 1)= C_0= \{o \in O: R(o)= 0\}$$

El problema de interpretación, tal como es definido por 4.3.2.1., consiste en aproximar, lo mejor que sea posible, las categorías de los resultados, usando conjuntos $(X_j = x_j) = \{o \in O: X_j(o) = x_j\}$ definidos por los valores de las variables observadas.

En este contexto, el objetivo de la interpretación es el de obtener el conjunto mínimo de descripciones del tipo

$$(X_{e1} = x_{e1}) \wedge ... \wedge (X_{ed} = x_{ed}) \subseteq C_j,\ 1 \leq d \leq p,\ j= 1 ... k$$

tales que

(1) $S= \{ o \in O: X_{e1}(o) = x_{e1}, X_{e2}(o) = x_{e2}, ..., X_{ed}(o) = x_{ed}\} \subseteq C_j,$
$j= 1, ..., k$

(2) $\frac{|S|}{|C_j|}$ sean la más grande que sea posible.

Estas ideas están ilustradas en las **figuras 4.7.1.a)** y **4.7.1.b)**.

En la **figura 4.7.1.a)** el conjunto *O* está representado por las tres clases c_1, c_2, c_3 de un resultado *R*, obtenido por un análisis a los datos, resultantes de observar dos variables X_1 (con valores 1, 2, 3, 4) y X_2 (con valores *a*, *b*, *c*, *d*).

En la **figura 4.7.1.b)** se busca expresar las clases c_1, c_2, c_3, c_4 usando expresiones disyuntivas construidas con los átomos de las variables observadas X_1 y X_2.

Por ejemplo, $(X_1 = 3) \wedge (X_2 = d) \subset c_2$ y también $(X_1 = 4) \wedge (X_2 = d) \subset c_2$.

De aquí resulta que una posible caracterización o interpretación de la clase $c_2 = (R = r_2)$ seria por la disyunción de dos expresiones conjuntivas:

$$((X_1 = 3) \wedge (X_2 = d) \vee (X_1 = 4) \wedge (X_2 = d)) \Rightarrow c_2.$$

Tal como es sugerido por el diagrama, estas interpretaciones pueden no ser perfectas; puede no ser posible cubrir de modo perfecto el significado de un resultado por expresiones de este tipo.

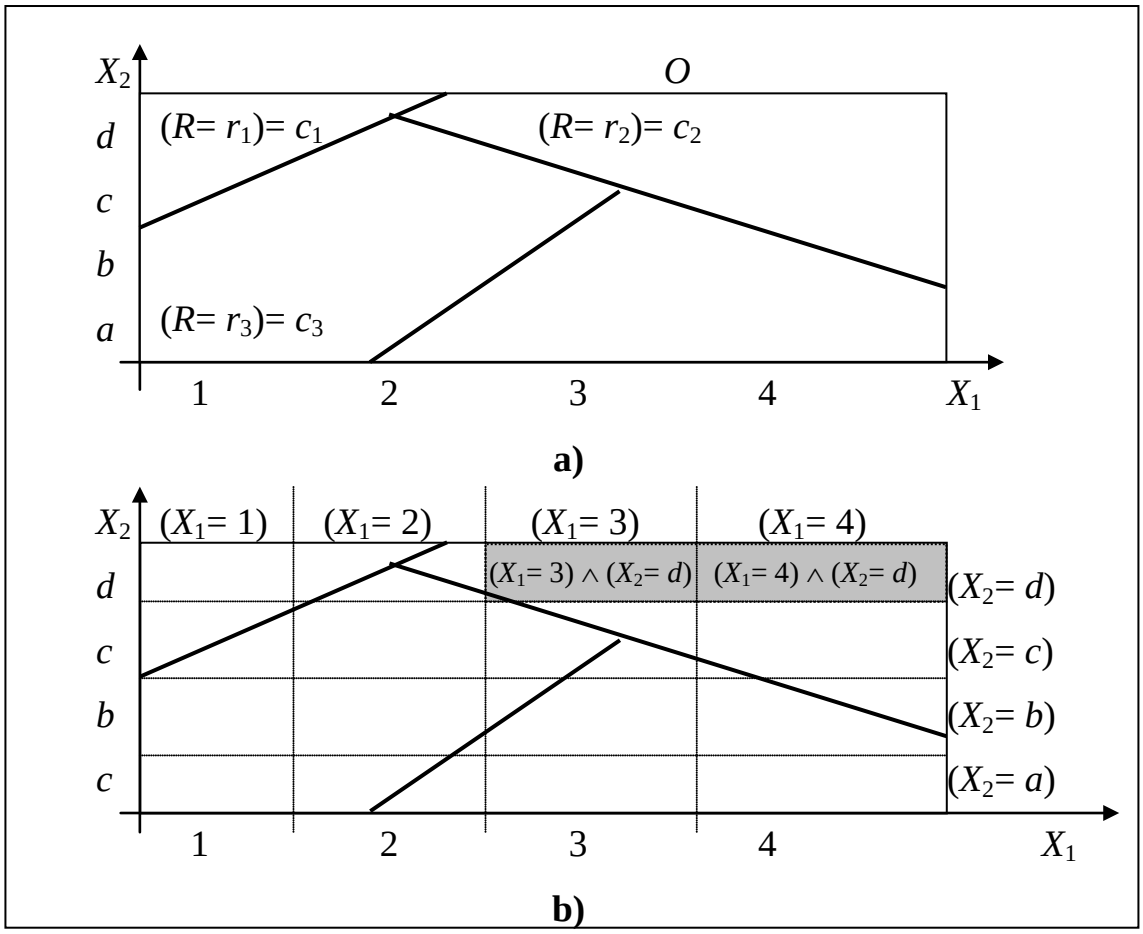


Figura 4.7.1. El resultado a interpretar es una partición formada por 3 clases que definan la variable cualitativa $R \in \{c_1, c_2, c_3\}$. En **b)**, la clase $(R= c_2)$ es aproximada por una expresión definida por X_1 y X_2.

Un modo de obtener sistemáticamente estas expresiones es construir un árbol de clasificación (regresión) de los individuos de O en la clases c_1, ..., c_k del resultado $R \in \{ c_1, ..., c_k\}$.

Ese árbol puede ser construido usando los algoritmos bien conocidos de QUINLAN (1993) o entonces el algoritmo CART de BREIMAN *et al* (1984).

BREIMAN *et al* (1984) - también HASTIE *et al* (2001) - dan preferencia a árboles binarios. Para eso, los átomos de las variables observadas o experimentales son agrupados en dos clases, a la hora de elegir las variables a usar para dividir los nudos no terminales.

En el contexto de interpretación de resultados obtenidos por distintas técnicas de análisis de datos, el lenguaje a usar en esas interpretaciones debe ser el mismo – lo que permite comparar resultados usando las mismas categorías para expresar esas aproximaciones.

Esto implica que, **antes del proceso de interpretación**, es necesario que el analista defina las categorías a usar, agrupando, eventualmente, valores de las variables observadas. Ver el apartado 4.4.2.

Por ejemplo, si la variable observada $X \in \{1, 2, 3, 4, 5, 6, 7\}$, puede ser necesario, antes de la interpretación, definir las categorías

(X= Pequeño)= ($X \in \{1, 2\}$) y

(X= Medio)= ($X \in \{3, 5\}$) y

(X= Grande)= ($X \in \{5, 6\}$).

En el acto de interpretar los resultados, los átomos a usar en las expresiones de aproximación serian (X= Pequeño) y (X= Grande).

De aquí resulta que, para nuestro contexto específico, nos parezca más adecuado el algoritmo de QUINLAN (1993) en donde el número de ramas de cada vértice no es necesariamente dos. El número de ramas en que se divide un vértice no terminal depende del número de átomos de la variable elegida para dividir ese vértice.

El algoritmo que se ha desarrollado para realizar este objetivo, basado en el algoritmo de QUINLAN (1993), es el que enseguida se desribe:

Inicialmente, se busca la variable X_1, con valores $X_1 \in \{x_{11}, x_{12}, ..., x_{n1}\}$, entre las variables consideradas relevantes para la interpretación del resultado $R \in \{c_1, c_2, ..., c_k\}$ e que maximiza la reducción de entropía mutua entre las clases $\{(X_1 = x_{11}), (X_1 = x_{12}), ..., (X_1 = x_{1n1})\}$ y las clases $\{(R= c_1), ..., (R= c_k)\}$ correspondientes al resultado R.

Esto significa que el conjunto O de individuos queda dividido en n_1 clases. Ver **figura 4.7.2.**

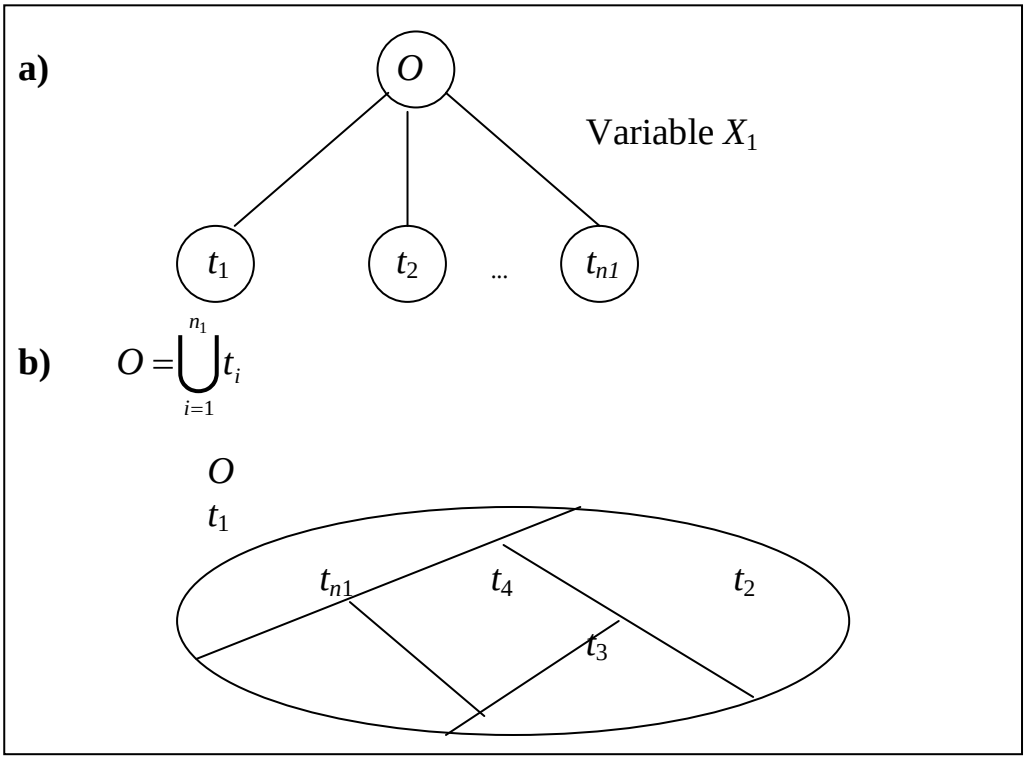


Figura 4.7.2. Situación inicial. Los átomos de la variable X_1 forman una partición del conjunto de individuos O.

En el caso de que alguna de estas clases iniciales esté integramente contenida en una de las clases del resultado R, eso significaría que ese nudo sería un terminal del árbol.

Por ejemplo, si $t_3 = (X_1 = x_{13}) \subseteq (R= c_2)$ entonces eso significaba que se verificaría la regla:

$(X_1 = x_{13}) \Rightarrow (R = c_2)$.

Cuando eso ocurre, el nudo correspondiente del árbol es puro: integra solamente objetos de una clase del resultado.

Cuando eso no ocurre, entonces el nudo es **impuro**, con un grado variable de impureza. Ver BREIMAN *et al* (1984).

Si el nudo es impuro y la división aun es posible, entonces ese nudo es dividido usando una de las variables relevantes aún no usadas en las divisiones anteriores: se aplica al nudo lo que se ha hecho inicialmente para la totalidad de los objetos.

Sea X_2 la variable usada para dividir el nudo *t*. Entonces el árbol de la **figura 4.7.1.a)** se transforma en el árbol de la **figura 4.7.2.**

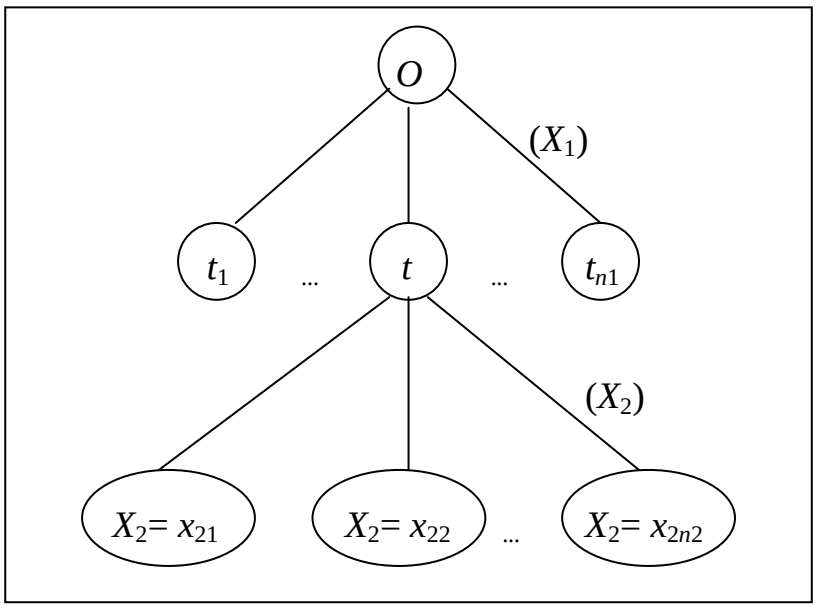


Figura 4.7.2. Aplicación al nudo *t* del procedimiento aplicado a *O*.

El proceso termina cuando todos los nudos son terminales.

A cada nudo del árbol en construcción corresponde, por este algoritmo, un descritor de tipo conjuntivo

$$(X_{j1} = x_{j1}) \wedge ... \wedge (X_{jd} = x_{jd})$$

que integra solamente átomos de las variables relevantes para la interpretación.

Estas ideas pueden ser formalizadas en el **algoritmo 4.7.1**

Algoritmo 4.7.1 - Estructuras de Datos

Este algoritmo se basa en las estructuras de datos siguientes:

ÁrbolFin - Es una lista de nudos con la estructura NUDO

ÁrbolEC - Es una lista de nudos con la estructura NUDO

Cada NUDO tiene la estructura siguiente:

Padre	***Descriptor***	***Extensión***	***Clase***

Padre: Es el identificador del nudo que, al dividirse, genera este nudo

Descriptor: Es una lista de identificadores de los átomos que describen este nudo.

Extensión: Es la lista de individuos que corresponden al ***Descriptor***

Clase: Identifica la clase del resultado(partición) a que pertenece el nudo.

Inicialmente:

ArbolEC tiene solamente un nudo cuya ***Extensión*** es el conjunto de todos los individuos, cuyo ***Padre*** es 0, ***Descriptor*** = 0, ***Clase*** Indefinida.

ArbolFin : Vacía.

Al final:

ArbolEC : Está vacía;

ArbolFin: Contiene los nudos terminales con ***Descriptor Extensión*** y ***Clase*** definidas por el proceso.

Algoritmo 4.7.1- Descripción

```
Iniciar
    Crear el primer nudo de ArbolEC
    Crear ArbolFin

Repetir Mientras ( Existen nudos en ArbolEC)
  Inicio
     NudoActual  := Primer nudo de ArbolEC;
     Si Terminal ( NudoActual, Clases, Tipo) entonces
        Caso
           Tipo =  1 :  TrataNudoTerminal ( NudoActual, Clases, 1 );
           Tipo = 2 :   TrataNudoTerminal ( NudoActual, Clases, 2 );
           Tipo = 3  :  TrataNudoTerminal ( NudoActual, Clases, 3 );
           Tipo = 4  :  TrataNudoTerminal ( NudoActual, Clases, 4 );
        FimCaso;
     Sino
        TrataNudoNoTerminal ( NudoActual,  ArbolEc, ArbolFin,  Clases);
     Retirar (ArbolEC, NudoActual);
  Fin.
```

Notas:

1- ***Terminal***: Función booleana que toma el valor 1 cuando el ***NudoActual*** es terminal y el valor 0 cuando no lo es. Si es terminal, establece el ***Tipo*** de terminal.

Tipo = 1- Contenido en una de las ***Clases***
Tipo = 2- El descritor tiene el numero máximo de átomos permitido
Tipo = 3- El numero de individuos de su extensión es el mínimo admisible
Tipo = 4- No puede ser dividido con los átomos de las variables aún no usadas

2- ***TrataNudoTerminal***- Procedimiento para tratar el ***NudoActual*** cuando este es terminal. Actualiza ***ArbolFin*** con el nudo terminal después de clasificado en una de las ***Clases***.

3- ***TrataNudoNoterminal*** – Procedimiento para tratar el ***NudoActual*** cuando este no es terminal. Busca la variable cuyos átomos producen la máxima entropía y divide el ***NudoActual*** en nudos-hijos, actualizando sus ***Descriptores*** y ***Extensiones***, añadiendo esos hijos a ***ArbolEC***.

Como se ha visto, el algoritmo de construcción del árbol detecta un nudo cuando se verifican los casos siguientes:

Caso 1 - El nudo está integralmente contenido en una de las clases del resultado que se pretende interpretar.

Caso 2 - El nudo no puede ser dividido porque en su descritor ya están representadas por átomos todas las variables relevantes para la interpretación.

Caso 3 - El nudo no puede ser dividido porque el número de objetos que lo integran es inferior al número admisible.

Caso 4 - El nudo no puede ser dividido, porque, aunque no tenga aún todas las variables relevantes para la interpretación representadas en el descritor corriente, no es posible encontrar, entre las variables aún no utilizadas, átomos que tengan intercesiones no vacías con ese nudo.

El tratamiento dado por el algoritmo a cada uno de estos casos es el siguiente.

Caso 1 - En este caso ha sido encontrada una regla «pura». Si un objeto satisface el descritor correspondiente a ese nudo, entonces pertenece a una de las clases del resultado por interpretar.

Casos 2, 3, 4 - En estos casos, una parte de los elementos del nudo pertenece a una de las clases pero el resto no. Ver la **figura 4.7.3.**

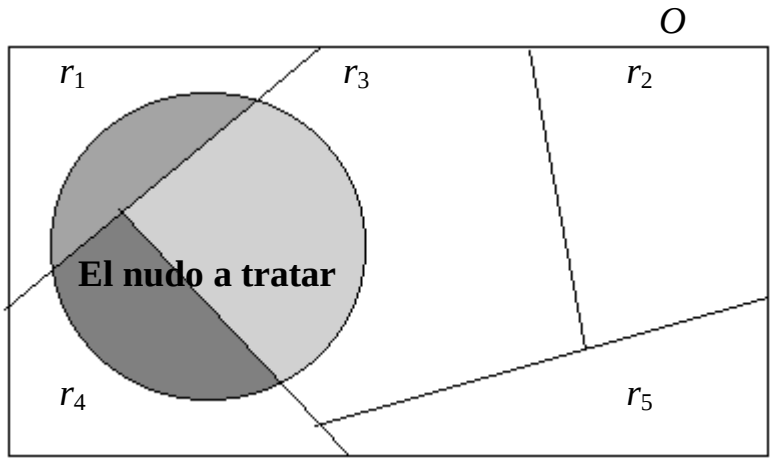


Figura 4.7.3. ¿Cuál es la clase del resultado con la cual el resultado tiene más afinidad?

Una regla habitual en estos casos es la regla de Bayes: decidir que un nudo es clasificado con la etiqueta de la clase más frecuente.

En este trabajo hemos implementado una regla un poco distinta, basada en el concepto de afinidad, medida por la función definida en 4.6.1.

Por este criterio, el nudo debe ser clasificado en la clase *con la cual tenga una afinidad más grande o, lo que es equivalente, en la clase más cercana, de acuerdo con la función de distancia definida en 4.6.1.*

Ejemplo 4.7.1.

Retomando el **ejemplo 4.6.2.1,** consideremos el problema de construir una aproximación para el significado del resultado

(R= 1)= {3, 8, 10, 12, 13, 18, 19, 23, 24, 27, 31, 32, 33, 36, 41, 43, 52, 57}

Si construimos un árbol para clasificar el conjunto de objetos

O= {1, 2, ..., 58} en las dos clases (R= 1) y (R= 0), las dos reglas con soporte mas elevado son:

$$(G4B= 0) \wedge (G3B= 1) \qquad y \qquad (G4B= 0) \wedge (G3B= 1) \wedge (G2B= 0).$$

Representando sobre el biplot los cierres convexos de los átomos que representan la primera regla, se obtiene el polígono marrón en la **figura 4.7.4.**

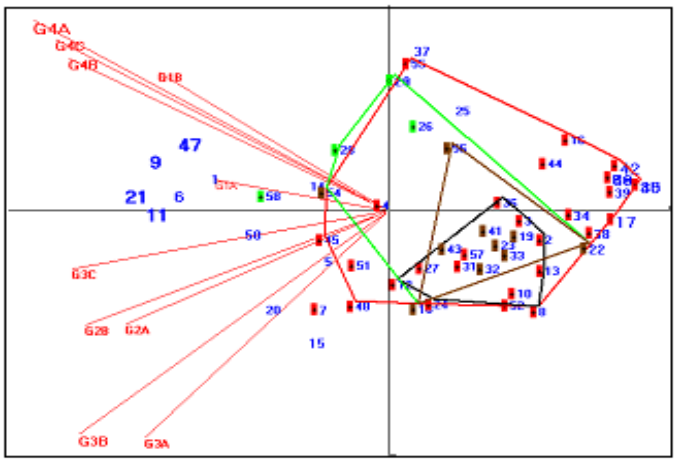


Figura 4.7.4. En esta figura puede verse el resultado (R= 1) - negro - y la aproximación (G4B= 0) ∩ (G3B= 1) a marrón.

Esta figura representa el cierre convexo de la extensión representada por la zona triangular a marrón.

$$\{18, 19, 22, 23, 32, 33, 41, 43, 54, 56\}= (G4B= 0) \wedge (G3B= 1).$$

Se verifica que esta expresión cubre la parte

$$(R= 1) \cap (G3B = 1) \cap (G4B = 0) = \{18, 19, 23, 32, 33, 41, 43\}$$

con

$$|(R = 1) \cap (G3B = 1) \cap (G4B = 0)| = 7.$$

El grado de cobertura es $\frac{7}{18} \cong 0.4$.

En el significado de (G3B = 1) $\wedge$ (G4B = 0) existen tres individuos:

{22, 54, 56}

que no pertenecen a (*R = 1*).

El grado de especificidad es: $\frac{7}{10} \cong 0.7$.

La afinidad es, finalmente, de $0.4 \times 0.7 \cong$ **0.28**.

4.8. LA INTERPRETACIÓN COMO APLICACIÓN DE LA TEORIA DE LOS CONJUNTOS IMPRECISOS.

4.8.1. INTRODUCCIÓN.

En este apartado se formula el problema de interpretación, definido en 4.3.2, usando la teoría de los conjuntos imprecisos (*Rought Sets*) desarrollada por PAWLAK (1991, 1998).

Aunque no iremos usar esta nueva formulación para desarrollar algoritmos de interpretación, pretendemos con su inclusión reforzar la idea de que el problema de interpretación de resultados puede ser formulado matemáticamente.

El concepto básico de esa teoría (resumido en el apartado 4.8.2.) es el de que no podemos distinguir entre dos objetos con las mismas descripciones, construidas usando un dado conjunto de variables.

Dado un grupo de objetos, resulta de esta verificación que puede no ser posible describir exactamente ese grupo, usando descripciones construidas con las variables observadas. Ese hecho genera un tipo de incertidumbre que puede ser modelada por el concepto de ***conjunto impreciso*** (*rought set*). PAWLAK (1991, 1998).

En un contexto de análisis preliminar de datos, lo que se pretende es traducir «en un lenguaje próximo del lenguaje humano, el significado de un resultado».

En los apartados 4.5, 4.6 y 4.7 se ha verificado que, usando como lenguaje de descripción las expresiones conjuntivas de conceptos atómicos, combinadas por disyunciones, obtenemos aproximaciones aceptables de los resultados por interpretar.

En este apartado se expresa el problema de interpretación usando la teoría de los ***conjuntos imprecisos*** de PAWLAK (1991, 1998). Las referencias POLKOWSKI *et al* (*eds.*) (1998a, 1998b), POLKOWSKI *et al* (*eds.*) (2000) presentan una actualización de la investigación en este tema.

El concepto de ***conjunto impreciso*** (*rough set*) permite formular matemáticamente la idea intuitiva de que «dos resultados son prácticamente lo mismo», en contextos donde no es posible aplicar pruebas estadísticas multivariantes.

La principal ventaja que en nuestro trabajo se atribuye a esta formulación es de carácter metodológico, formal, lógico. Es una confirmación más de que el problema de interpretación de los resultados puede ser formulado matemáticamente.

En lo que respecta a aplicar directamente esa teoría a la construcción de algoritmos y heurísticas para nuestro problema específico, el autor es menos optimista. Se cree que la explosión combinatória asociada al cálculo de algunos de los conceptos fundamentales de la teoría - como por ejemplo, el concepto de *reduct* - pone aun demasiados problemas.

4.8.2. SÍNTESIS DE LA TEORÍA DE LOS CONJUNTOS IMPRECISOS.

La teoría de los *rough sets* - aquí traducida por teoría de los conjuntos imprecisos - ha sido desarrollada por PAWLAK (1991, 1998).

Versiones actualizadas de esta teoría pueden verse en LIN *et al* (1997), PAWLAK (1998), POLKOWSKI *et al* (1998a, 1998b), POLKOWSKI *et al* (2000), STEPANUIK *et al* (2000), BAZAN *et al* (2000).

En PEÑA *et al* (1999) pueden verse los problemas que ocurren al aplicar la teoría a problemas reales.

Designemos por $O = \{1, 2, ..., i, ..., n\}$ el conjunto de etiquetas o identificadores de los individuos observados y por $A = \{X_1, X_2, ..., X_p\}$ el conjunto de etiquetas o identificadores de los atributos o variables.

Entre las variables se distinguen dos subconjuntos *C y D, tales que*:

$A = C \cup D$

C = Conjunto de variables de *condición* o *independientes*.

D = Conjunto de variables de *decisión* o *dependientes*.

En nuestro contexto de interpretación de resultados, las variables independientes son las variables observadas y las variables dependientes son variables indicadoras o cualitativas representativas de resultados obtenidos por distintos métodos de análisis.

Si X es una variable, $X \in A$, V_X designa los valores de X.

Una variable X es una función

$$X : O \longrightarrow V_X$$
$$i \longmapsto V(i)$$

El concepto básico de la teoría de los *conjuntos imprecisos* es el concepto de ***indistinguibilidad***.

Dado un conjunto $B \subseteq A$, formado por variables observadas, dos individuos son indistinguibles o equivalentes cuando esas variables asumen los mismos valores sobre esos individuos.

Específicamente, sea $X \in B$ una de las variables y sean $i, j \in O$, dos individuos.

Si $X(i) = X(j)$ para $X \in B$, entonces i y j son indistinguibles usando solamente las variables del conjunto B. No pueden ser distinguidos usando expresiones construidas con esas variables.

Esta verificación permite definir la relación de equivalencia I_B tal que

$$i \, I_B \, j \Leftrightarrow (X(i) = X(j), \; X \in B).$$

La clase de equivalencia a que pertenece el individuo de identificador *i* se designa por

$[i]_{I_B}$.

La partición O/I_B se obtiene agrupando los individuos que no pueden ser distinguidos por las variables $X \in B$.

Ejemplo 4.8.2.1.

En nuestra **tabla 3.1.1,** consideremos B = {G1A, G1B} en que

G1A= «Resultados obtenidos en la cuestión 1A»

G1B= «Resultados obtenidos en la cuestión 2A».

Si ordenamos las 58 alumnos por los resultados que han obtenido en el grupo de preguntas B = {G1A, G1B}, obtenemos las clases:

O/I_B = {(G1A= 0) $\wedge$ (G1B= 0), (G1A= 0) $\wedge$ (G1B= 1),
(G1A= 0) $\wedge$ (G2B= 2), (G1A= 1) $\wedge$ (G2A= 0),
(G1A= 1) $\wedge$ (G2A= 1), (G1A= 1) $\wedge$ (G2A= 0),
(G1A= 2) $\wedge$ (G2A= 0), (G1A= 2) $\wedge$ (G2A= 1),
(G1A= 2) $\wedge$ (G2A= 2)}

Por ejemplo,

(G1A= 0) $\wedge$ (G2A= 2)= {1, 2, 3, 4, 6, 7, 9, 11, 14, 15, 16, 18, 19, 20, 21, 23}.

Las clases de equivalencia para la relación I_B forman los *conceptos B-elementales*[3].

[3] Obsérvese que lo que aquí se designa por *conceptos B-elementales* son intersecciones de *conceptos atómicos*, usando nuestra designación, adoptada en el párrafo 4.3. La designación de *concepto elemental* atribuida a un conjunto resultante de la combinación de otros por operaciones de intersección, nos parece discutible.

Podemos realizar la aproximación de un conjunto $R \subseteq O$ empleando dos conjuntos designados por **aproximación inferior** y **aproximación superior** de *R* en *B*. Ver PAWLAK (1991, 1998).

La aproximación inferior de *R* en *B* – simbolizada por $B_*(R)$ - es formada por la unión de todos los individuos cuyas clases de equivalencia están contenidas en *R*.

O sea: $B_*(R) = \{ i \in O: [i]_{I_B} \subseteq R\}$.

La aproximación superior de *R* en *B* – simbolizada por $B^*(R)$ - es formada por la unión de todos los individuos cuyas clases interceptan *R*

$B^*(R) = \{i \in O: [i]_{I_B} \cap R \neq \varnothing\}$.

La frontera $Front_B(R) = B^*(R) - B_*(R)$.

Un conjunto es impreciso cuando no puede ser expresado de modo exacto usando conceptos *B-elementales*. Esto ocurre cuando $Front_B(R) \neq \varnothing$. En ese caso, el conjunto, designado por conjunto impreciso, es representado por el par ordenado ($B_*(R)$, $B^*(R)$).

Si $B^*(R) = B_*(R)$ entonces $Front_B(R) = \varnothing$ y el concepto *R* puede ser expresado de modo exacto, usando el conocimiento formado por los conceptos *B*-elementales: no existe ninguna *imprecisión* o *incertidumbre* en esa expresión.

El otro concepto básico de la teoría de los conjuntos imprecisos es el concepto de **reducto** (*reduct*), basado en la idea de dependencia entre variables.

Dadas dos variables X_1 y X_2, ¿en que medida podemos aproximar las categorías de X_2 usando las categorías de X_1?

En particular, dado un atributo/variable de decisión *X* y un atributo/variable de decisión *R* - un resultado *R* - ¿en que medida podemos aproximar las categorías de *R* usando las categorías de *X*?

La definición de dependencia usada en la teoría de los conjuntos imprecisos es la siguiente. Ver PAWLAK (1998), POLKOWSKI *et al* (eds.) (2000):

Definición 4.8.2.1. (Dependencia)

Dado el conjunto de atributos de condición *C* y el conjunto de atributos de decisión *D*, se dice que *D* depende de *C* en grado *k* - con $0 \leq k \leq 1$ - cuando:

$$k = \gamma(C, D) = \frac{|Pos_C(D)|}{|O|}$$

en donde

$$Pos_C(D) = \bigcup_{E \in O/I_D} C_*(E)$$

se designa por Región Positiva de O/I_D con respecto a *C*.

En la expresión anterior, POS_C (*D*) es formada por el conjunto de objetos que pueden ser clasificados en las clases de *D* usando la información suministrada por las clases de *C*.

Con base en esta definición de dependencia se define entonces el concepto de REDUCTO (*REDUCT*): dados dos conjuntos *C* y *D* de atributos de condición y decisión, se dice que $C' \subseteq C$ es un *D*-reduct (reducto con respecto a *D*) cuando *C*' es minimal en *C*, en el sentido de que

$$\gamma(C, D) = \gamma(C', D).$$

Esto significa que toda la información a respecto de *D* contenida en las clases de *C* está en las clases de *C*' - lo que significa que los atributos de la diferencia (*C* - *C*') son *dispensables* o *superfluos* del punto de vista de caracterizar las clases de *D*.

No existe un reducto único para un dado conjunto de atributos. PAWLAK (1991) muestra que la intersección de dos reductos es un reducto - lo que implica que existe un reducto mínimo designado NUCLEO: la intersección de **todos** los reductos.

La computación de los reductos implica construir una matriz cuadrada, de dimensión igual al número de objetos observados, designada por matriz de indistinguibilidad. POLKOWSKI *et al* (*eds*) (2000).

4.8.3. FORMULACIÓN DEL PROBLEMA DE INTERPRE-TACIÓN DE RESULTADOS USANDO LA TEORÍA DE LOS CONJUNTOS IMPRECISOS.

Admitamos que un conjunto de datos de *n* individuos por *p* columnas o variables ha sido sometido a una o más análisis de datos multivariantes.

Se ha verificado en el capitulo III que una clase muy importante de resultados pueden ser expresados por un conjunto D= {R_1, R_2, ..., R_r} de variables cualitativas R_1 ... R_r.

El problema de interpretación de los resultados R_1 ... R_r puede ser visto como el problema de aproximar - en el contexto de la teoría de los *conjuntos imprecisos* - las categorías de las variables R_1 ... R_r usando los conjuntos elementales definidos por las variables observadas C= {X_1... X_p}, con A= $C \cup D$. Ahora, D= {R_1 ... R_r} son las variables de decisión o dependientes y las variables observadas C= {X_1 ... X_p} las variables de condición o independientes.

Puede ocurrir que los conceptos C-elementales (formados por todas las expresiones conjuntivas, de intersección no vacía, de conjuntos de tipo (X= x) con 1, 2, ..., p factores) no sean suficientes para aproximar los resultados de modo exacto.

Si esto ocurre para un resultado R, entonces ese resultado seria representado por un par de conjuntos precisos o exactos: sus aproximaciones inferior y superior - $C_*(R)$, $C^*(R)$.

Esto permitiría definir en el conjunto de resultados D= {R_1 ... R_r} una relación de equivalencia: dos resultados serian equivalentes cuando fueran representados por el mismo par ($C_*(R)$, $C^*(R)$), designado por ***conjunto impreciso.***

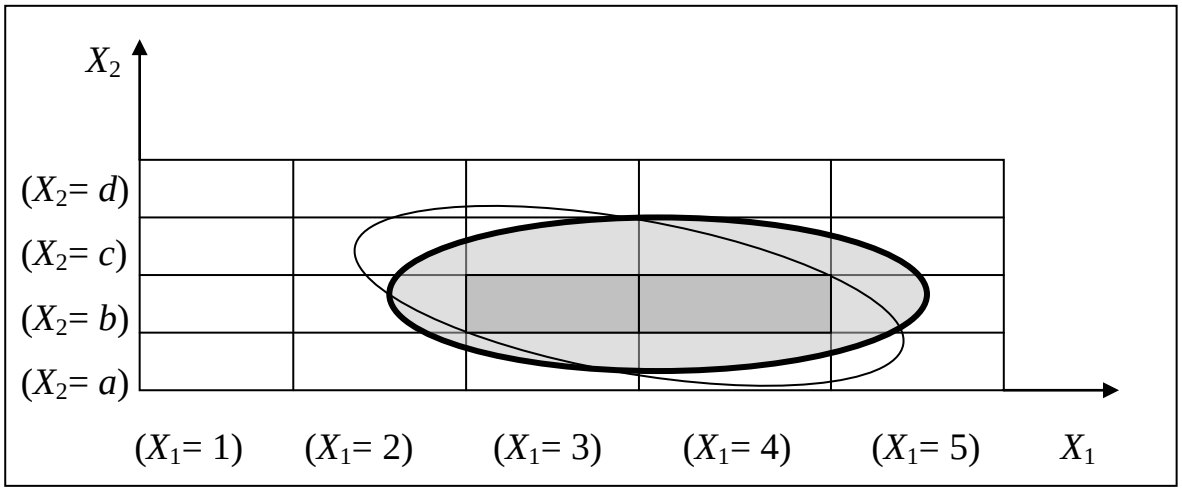


Figura 4.8.3.1. Aproximación de un resultado $D= \{R_1, R_2\}$ por las variables X_1 e X_2.

En la figura anterior, $C= \{X_1, X_2\}$, $D= \{R_1, R_2\}$

con $X_1 \in \{1, 2, 3, 4, 5\}$

$X_2 \in \{a, b, c, d\}$

Los conceptos elementales que se puede definir con estas dos variables son:

$(X_1= x)$, $x \in \{1, 2, 3, 4, 5\}$

$(X_2= y)$, $y \in \{a, b, c, d\}$

$\{(X_1= x) \cap (X_2= y), x \in \{1, 2, 3, 4, 5\}, y \in \{a, b, c, d\}\}$.

En el ejemplo,

$$C_* (R_1) = ((X_1 = 3) \cap (X_2 = b)) \cup ((X_1 = 4) \cap (X_2 = b))$$
$$= C_* (R_2)$$
$$C^* (R_1)= C_* (R_1) \cup ((X_1 = 1) \cap (X_2 = b)) \cup ((X_1 = 1) \cap (X_2 = c))$$
$$\cup ((X_1 = 3) \cap (X_2 = c)) \cup ((X_1 = 4) \cap (X_2 = a))$$
$$\cup ((X_1 = 4) \cap (X_2 = c)) \cup ((X_1 = 5) \cap (X_2 = a))$$
$$\cup ((X_1 = 5) \cap (X_2 = c)) = C^* (R_2).$$

Eso quiere decir que R_1 y R_2 no pueden ser distinguidos con la información suministrada por las categorías de $C = \{X_1, X_2\}$ y, por eso, deberían ser considerados «*la misma cosa*».

Aunque el problema de interpretación obtiene así una formulación matemática precisa, el problema de computación de los reductos en un contexto de minería de datos, con un número muy elevado de variables y individuos, es NP-complejo. POLKOWSKI *et al* (*eds.*) (2000), STEPANIUK (2000).

En las aplicaciones - ver, por ejemplo, PEÑA *et al* (1999) - la complejidad asociada al cálculo de los reductos puede ser limitada considerando solamente las variables que son *relevantes para la aproximación*, o entonces limitando el número de términos en las expresiones conjuntivas, por consideraciones psicológicas.

En este trabajo, no se presenta ningún algoritmo de interpretación basado en esta teoría, reservándose esa posibilidad para futuras investigaciones.
Específicamente: el concepto de conjunto impreciso, representativo de una clase de resultados equivalentes o indistinguibles, puede ser representado en el plano de un biplot por las configuraciones de marcadores de individuos que forman sus aproximaciones inferior y superior.

Esto significa que es fácil, conceptualmente, integrar en el contexto de una minería de datos basada en biplots, los razonamientos de interpretación de resultados tal como los hemos formulado en este apartado.

C A P Í T U L O V

LOS MÉTODOS BIPLOT COMO TÉCNICA DE MINERÍA DE DATOS. POSIBILIDADES Y LIMITACIONES

5.1. INTRODUCCIÓN.

En el capítulo III se ha PUESTO DE MANIFIESTO la gran generalidad de la técnica de los biplots. Una vez que los métodos más conocidos de análisis de datos multivariantes pueden ser vistos como casos particulares de biplots y generan resultados que pueden ser representados mediante configuraciones de marcadores con los biplots apropiados, los biplots son candidatos naturales a ser instrumentos de minería de datos. Ver **figura 4.1.1.**

En este capítulo se buscan respuestas para las cuestiones 3 y 4 que han sido formuladas en la Introducción:

Cuestión número 3
¿Qué problemas ocurren cuando se intenta centrar en el concepto de biplot un sistema de minería de datos?. ¿Cuáles son las soluciones para esos problemas?

Cuestión número 4
¿Cómo realizar minería de datos centrada en biplots?.

Los problemas que ocurren han sido agrupados en las categorías siguientes, cada una de ellas objeto de un apartado en este capítulo:

- Modelo a emplear para realizar minería de datos basada en biplots.
- Cuestiones computacionales.
- Cuestiones gráficas.
- Cuestiones de interactividad entre el usuario y el sistema.

En el apartado 5.2 se presenta un modelo genérico para realizar minería de datos centrada en biplots.

En el apartado 5.3 se analizan las cuestiones computacionales, buscándose mostrar que las técnicas biplot cumplen con los criterios básicos, definidos en el apartado 1.4, para que pueda ser considerada una técnica de minería de datos.

Según GABRIEL (1971, 1995a,1995b), un biplot es un gráfico en dos o tres dimensiones. Eso significa que un biplot es un instrumento de *visualización* de las relaciones que puedan existir entre variables, entre individuos y entre variables e individuos.

En consecuencia, un sistema de minería de datos basado en biplots es, esencialmente, un sistema de visualización de información.

En el apartado 5.4 se revisa la literatura acerca de los problemas gráficos creados por grandes conjuntos de datos, formulándose la estrategia seguida en nuestro proyecto.

Un sistema de minería de datos basado en biplots es, en principio, un sistema interactivo, en donde el analista toma decisiones que condicionan el proceso de análisis. En el apartado 5.5 se analizan y proponen soluciones para aspectos básicos de esta interacción, coherentes con los conceptos teóricos desarrollados en el **Capítulo IV**.

5.2. UN MODELO PARA MINERÍA DE DATOS BASADA EN BIPLOTS.

Un sistema de minería de datos basada en biplots (SMDBB) puede admitir una gran diversidad de técnicas de análisis, actuando los biplots como instrumento de visualización y síntesis de los resultados obtenidos por esas distintas técnicas. No significa que se trate de un sistema en la que la única técnica de minería empleada sea la técnica de los biplots.

En efecto, se ha verificado en el **capítulo III** que casi todas las técnicas generales de análisis de datos multivariantes son casos particulares de biplots o, no siéndolo, sus resultados pueden expresarse por configuraciones de marcadores biplots.

En este contexto, los biplots suministran un lenguaje común que permite relacionar los resultados más importantes producidos por las distintas técnicas. Ver **capítulo IV**, **figura 4.1.1.**

En un SMDBB, los datos pueden, por supuesto, ser analizados por técnicas de biplots pero también por técnicas de análisis cluster (aglomerativas o divisivas), análisis discriminante, análisis factorial (componentes principales, correspondencias simples y múltiples, análisis canónicos), análisis de regresión, modelos log-lineales, MDS y otras.

Eso si, en un SMDBB todos los resultados producidos por esas técnicas distintas son representados en una forma común: por grupos de individuos particiones del conjunto de individuos. Matemáticamente, todos esos resultados son representados por variables cualitativas, según vimos en el **capítulo IV**.

A los grupos y particiones generados automáticamente por las distintas técnicas se añaden los grupos y particiones que el analista identifique al examinar los resultados o que correspondan a hipótesis que desee comprobar.

Todos estos grupos y particiones, sea cual sea su origen, son representados, por configuraciones de marcadores, en los biplots que forman la interface gráfica con el analista.

Una vez representados los grupos/particiones que interesa estudiar por configuraciones de marcadores en biplots, el analista puede caracterizar y comparar visualmente esos resultados, usando técnicas estadísticas u otras.

La **figura 5.2.1.** presenta estas ideas en forma gráfica.

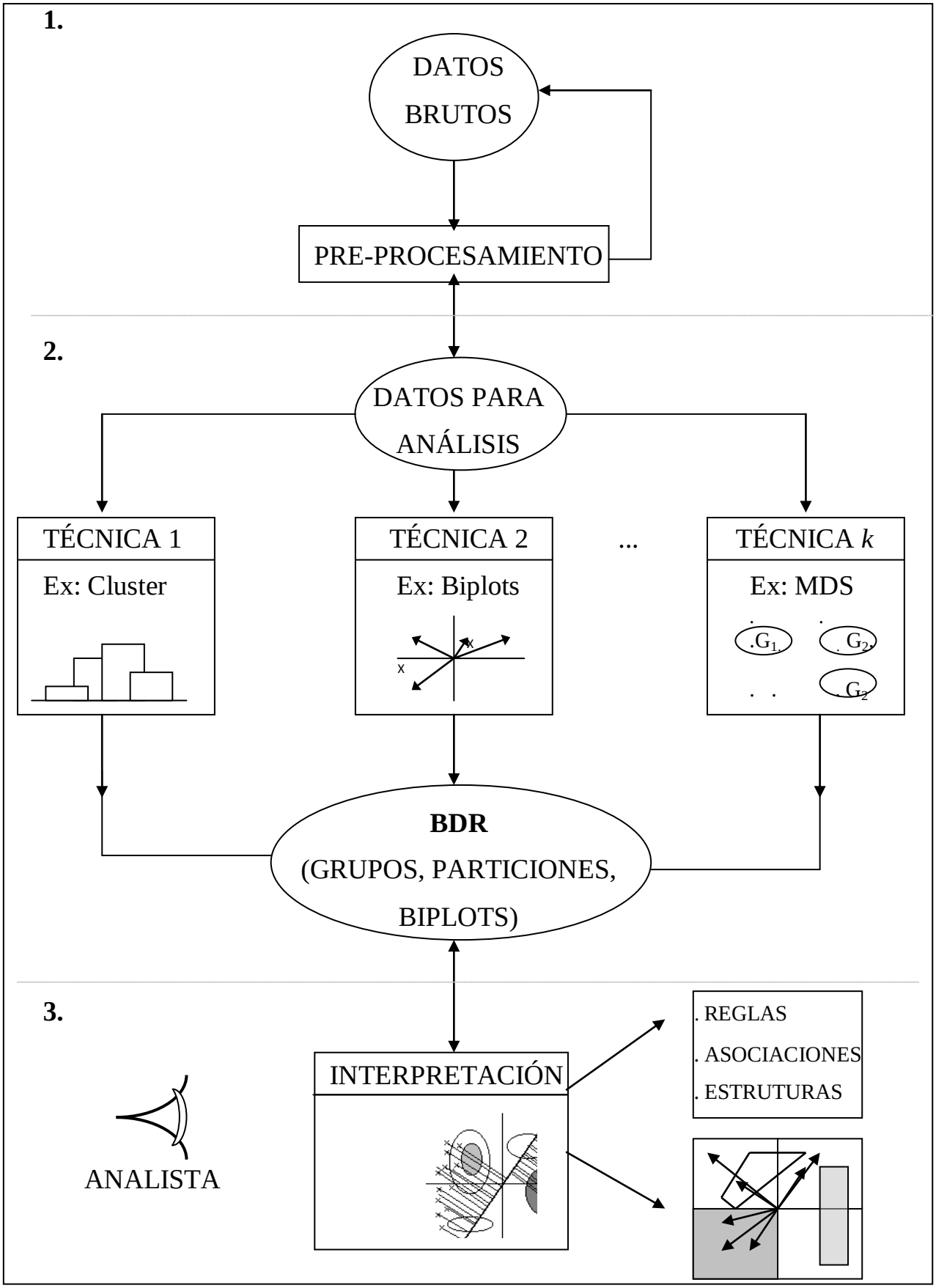


Figura 5.2.1. Un SMDBB permite integrar resultados obtenidos por distintas técnicas, incluidas las técnicas de biplot. Los resultados son integrados en una BDR - Base de Datos de Resultados.

Al crear un SMDBB se asumen los postulados siguientes, basados en las consideraciones teóricas desarrolladas en el **capítulo IV**:

1. La minería de datos centrada en biplots es interactiva. Esa interacción es basada en representaciones biplot.

2. Los resultados de los distintos métodos de análisis de datos multivariantes se expresan por variables cualitativas que representan grupos y particiones de individuos y por grupos de variables.

3. Al intentar interpretar los resultados de los distintos análisis de datos multivariantes, el analista realiza las operaciones básicas siguientes, descritas en el **capítulo IV**, apartado 4.2.
   - Identifica grupos
   - Caracteriza grupos
   - Compara grupos
   - Nombra grupos (crea conceptos)
   - Reconoce patrones

4. Interpretar un resultado es aproximar el significado de ese resultado por el significado de los valores de las variables observadas. Ver **capítulo IV**, apartado 4.3.

Usando la **figura 5.2.1.** como referencia, la estructura básica y el funcionamiento global de un SMDBB típico son los siguientes:

1. **PRE PROCESAMIENTO.**

Los **Datos Brutos** contenidos en una base de datos son pre - procesados con en el fin último de obtener datos validados listos para la fase de minería.

El contenido de los datos brutos es, en general, mal adaptado a la aplicación directa de las técnicas de análisis. Por eso es necesario, realizar operaciones de limpieza y adaptación, como: eliminar registros, recodificar y/o eliminar variables, calcular los estadísticos básicos, decidir qué hacer con los datos faltantes.

La obtención de las estadísticas básicas puede ser simultánea a la descomposición de los datos en átomos, y creación del grafo de intersección correspondiente. Ver **capítulo IV**.

Con base en estos átomos pueden realizarse las operaciones siguientes:

- **1.1.** Recodificación de variables: equivale a agregar los distintos valores observados en clases, definidas por el analista para efectos de análisis y de interpretación.
- **1.2.** Decidir qué hacer con valores nulos, faltantes, excepcionales, y tipo «NS / NR».
- **1.3.** Eliminar registros que correspondan a condiciones definidas por el analista.
- **1.4.** Eliminar variables no relevantes para el problema.
- **1.5.** Cuando los datos son mixtos, generar nuevos datos en la forma TDC (Tabla Disyuntiva Completa). LEBART *et al* (2002).
- **1.6.** Descomponer la tabla de datos en datos parciales obedeciendo a criterios convenientes.

1.7. Seleccionar los datos para análisis. En particular, esos datos pueden ser elegidos usando un método de muestreo, cuando los datos sobrepasan a los recursos computacionales. Ver, en este capítulo, párrafo 5.3.

1.8. Definir variables e individuos suplementarios.

El resultado de la fase 1 es un conjunto de **Datos Para Análisis**.

2. **ANÁLISIS / MINERÍA DE LOS DATOS.**

Estos dados pueden ser analizados por distintas técnicas. Por ejemplo: Análisis Cluster, Componentes Principales, Biplots, MDS,... Todos los resultados generados por estas técnicas son almacenados en una **Base de Datos de Resultados** (**BDR**) en la forma de descripciones de Grupos (de individuos y de variables), Particiones y Biplots (configuraciones de marcadores de individuos y variables en los sistema de coordenadas definidas por los distintos tipos de biplots).

Entre estos resultados se incluyen los grupos y particiones identificados por el analista, los obtenidos automáticamente por los distintos métodos y los que correspondan a hipótesis enunciadas por el analista.

3. **INTERPRETACIÓN/SÍNTESIS.**

El proceso de interpretación, controlado por el analista, actúa sobre la BDR. Ver **figura 5.2.1.**

3.1. Representación, sobre uno o más biplots, de los resultados (grupos y particiones) obtenidos por las distintas técnicas.

3.2. Identificación por el analista, sobre los distintos planos del biplot, de nuevos grupos y particiones.

3.3. Caracterización de los grupos y particiones existentes.

El sistema debe sugerir caracterizaciones estadísticas o entonces descripciones en el lenguaje de las expresiones disyuntivas descrita en el **capítulo IV.**

3.4. Comparación de resultados.

Estas comparaciones pueden ser realizadas con base en métodos estadísticos (por ejemplo: comparación de medias / ANOVA, Análisis Discriminante) o, informalmente, con base en las estadísticas descriptivas de las variables.

3.5. Interacción del analista con los datos y el gráfico del biplot, identificando nuevos grupos, eliminando grupos, ordenando los datos, formulando nuevas cuestiones.

4. En resultado de la interacción Analista/Biplots, en la fase 3, puede ser necesario repetir las fases 2 y 3 un número indeterminado de veces hasta que el analista se considere satisfecho.

5. El resultado final del proceso 1, 2, 3, es la estructura presumida de los datos. Esta estructura o modelo de los datos es usada para:

5.1. Clasificación de nuevas observaciones, como elementos suplementarios.

5.2. Predicción de los valores de variables en zonas especificas de los biplots.

6. Cuando ocurren nuevas observaciones, es necesario analizar su impacto sobre la estructura asumida o modelo en uso y evaluar la necesidad de su revisión.

5.3. PROBLEMAS COMPUTACIONALES.

5.3.1. INTRODUCCIÓN.

En el número 1.4. han sido identificados los criterios a que debe obedecer una técnica para que pueda ser considerada una técnica de minería de datos. Eses criterios son los siguientes:

- Interpretabilidad de los resultados
- Incrementabilidad
- Escalabilidad

Las cuestiones de interpretación de los resultados forman el núcleo central de nuestra investigación y han sido tratadas de modo autónomo en el **capítulo IV**. Los biplots no solo son fácilmente interpretados y por tanto permiten interpretar los resultados obtenidos por otros métodos.

En este capítulo se consideran los problemas de incrementabilidad y escalabilidad que ocurren al intentar construir sistemas de minería de datos basados en biplots (SMDBB).

En la literatura revisada, las cuestiones de incrementabilidad y escalabilidad ocurren a veces muy mezcladas, como por ejemplo en KÖLSGEN *et al* (2002).

En nuestro trabajo, se busca mantener la separación entre esos criterios.

La incrementabilidad - tratada en el apartado 5.3.2. - tiene que ver con la posibilidad de incorporar o integrar observaciones adicionales sin tener que repetir todos los cálculos realizados con los datos anteriores.

Esta propiedad es crucial para la aplicabilidad de la técnica a grandes masas de datos que no puedan ser cargadas en la memoria central del ordenador.

La escalabilidad – tratada en el apartado 5.3.3- también tiene que ver con el tamaño del conjunto de datos pero considerando aspectos distintos:

¿La técnica puede aplicarse a grandes y pequeños conjuntos de datos de forma que los resultados obtenidos con unos y otros sean igualmente interesantes?.

¿Cómo aumenta el tiempo de ejecución cuando aumenta el tamaño del conjunto de datos?.

¿Es posible relacionar los resultados obtenidos con la misma técnica a escalas distintas?.

Un problema crucial es el de definir lo que se considera un «gran» conjunto de datos.

En este trabajo se utiliza una clasificación creada por Huber y referida por WEGMAN (1995). Esta clasificación - ver WEGMAN (1995) - se presenta en la **tabla 5.3.1.1**

WEGMAN (1995) analiza la posibilidad de procesamiento interactivo de los distintos volúmenes de datos con algoritmos de complejidad conocida, en ordenadores con potencias de cálculo expresadas en megaflops (1megaflop = 10^6 operaciones de cálculo en coma flotante por segundo).

Designación	**Volumen (Bytes)**	**Ejemplo (Almacenamiento)**
Muy pequeño (*Tiny*)	10^2	Hoja de papel
Pequeño (*Small*)	10^4	Algunas hojas de papel
Medio (*Medium*)	10^6	Una disquete
Grande (*Large*)	10^8	Disco rígido
Enorme (*Huge*)	10^{12}	Múltiples discos rígidos

Tabla 5.3.1.1. Adaptado de WEGMAN (1995). Este autor considera aún una clase que designa por *ridiculous* (10^{12} bytes).

Según WEGMAN (1995) para un ordenador de 1000 gigaflops, una operación de complejidad O (n^2) podría realizarse en un segundo (considerado el límite de la posibilidad de interactividad) si el volumen de datos fuera del orden de 10^6 (Medio) y necesitaría de **3 horas** si el volumen de datos fuera de 10^8= 100 Mb (Grande).

Los algoritmos de descomposición en valores y vectores propios (SVD), base del cálculo de los biplots clásicos, tienen complejidad computacional $O(n^2p)$ en donde n es el número de individuos del conjunto de datos y p el número de variables o columnas. Ver WEGMAN (1995).

Esto significa que, para grandes conjuntos de datos ($\geq 10^6$ bytes) y ordenadores de uso general como los ordenadores de tipo personal, se trata de operaciones demasiado lentas, para que sea considerada la hipótesis de interactividad.

En una cierta visión de la minería de datos se considera que la interactividad no es posible ni deseable. En esta perspectiva, todos los datos disponibles deben ser procesados de un modo completamente automático. La intervención del analista se limitaría a definir, en el inicio del proceso,

los parámetros de la sesión: los datos y las opciones. En el final de la sesión, el sistema entregaría los patrones descubiertos.

Se cree que esta visión tiene limitaciones conceptuales importantes.

Al ritmo a que son generados los datos, la distancia entre el volumen de datos por analizar y los recursos disponibles seguirá aumentando. El mundo, en cierto sentido - texto, imágenes, sonidos - está pasando para el interior de los ordenadores.

Esto significa que los analistas están, hoy día, en relación a los datos almacenados en bases de datos, en la misma situación en que siempre se han considerado los estadísticos: la imposibilidad de procesar toda la población. El reconocimiento de esta realidad llevó a los estadísticos a la creación de los métodos de muestreo y la inferencia estadística.
Con los ordenadores, es simple y casi gratis extraer muchas muestras de dimensión adecuada a partir de una gran base de datos, sea cual sea su dimensión, al paso que, tradicionalmente, los estadísticos siempre han realizado todas sus inferencias usando una única muestra.

En esta perspectiva, una buena parte de la solución para la minería de grandes masas de datos puede estar en la aplicación de métodos estadísticos, abandonando, por lo menos en un significativo número de casos, la pretensión de procesar todos los datos.

En contra de esta perspectiva está el hecho de que las técnicas de muestreo no detectan pequeños grupos de observaciones, soporte de relaciones importantes, como ocurre, por ejemplo, en problemas de detección de

fraudes económicos y en problemas de seguridad informática. Ver KLÖSGEN *et al* (*eds.*) (2002).

Aunque fuera posible transformar una base de datos gigante en un biplot, ese biplot seria inútil para el analista: contendría, simplemente, demasiada información y estructura artificial generada por la interacción de factores como la precisión limitada de los cálculos, *n* y *p* muy grandes y el número limitado de pixel en los dispositivos de presentación.

Para ser útil como instrumento de interacción, un biplot, no puede contener demasiada información. Una minería de datos basada en biplots puede implicar que esos biplots se refieran a partes, resúmenes o muestras del conjunto de datos.

En ese contexto, el problema es el de combinar o integrar los biplots parciales, correspondientes a esos resúmenes o muestras. Ese problema ha sido considerado, también, para otras técnicas como puede verse en PROVOST *et al* (2002) y BRADLEY *et al* (1998). Estos autores hacen referencia sistemas que procesan múltiples muestras sucesivamente, combinando los resultados.

En un proceso de descubrimiento basado en una gran base de datos en que no sea posible procesar todos los datos, lo que importa es que la búsqueda de nueva información relevante para el problema sea orientada por la información o el estado de conocimientos obtenidos hasta el momento actual. Eso significa que pueden revelarse útiles los procedimientos de muestreo adaptativo, descritos, por ejemplo, en THOMPSON *et al* (1996) y LOHR (1999).

Estos métodos permiten usar la información de una muestra para orientar la elección de la muestra siguiente, explotando las relaciones conocidas entre esas observaciones. Este tipo de muestreo se utiliza, precisamente, en situaciones en las que se busca caracterizar pequeñas poblaciones dispersas en enormes espacios como el océano THOMPSON (1992).

5.3.2. INCREMENTABILIDAD Y BIPLOTS.

La complejidad computacional del cálculo de la D.V.S. de un conjunto de datos de n filas por p columnas es $O(n^2p)$. Ver GOLUB *et al* (1983), WEGMAN (1995), KOLDA *et al* (1999). Eso significa que la complejidad computacional del cálculo de un biplot clásico es, por lo menos, $O(n^2p)$, lo que comprometería, en principio, la utilización de esta técnica como instrumento de minería de datos - tal como las técnicas de clasificación cuya complejidad es $O(n^3)$.

Este hecho y la naturaleza gráfica de los biplots implica que la aplicabilidad de los biplots a tareas de minería de datos es dependiente de sus características incrementales.

En un proceso incremental, si R_1 son los resultados obtenidos con esa técnica al procesar el conjunto de datos X, de n_1 filas por p columnas, si R_2 son los resultados obtenidos al procesar el conjunto de datos Y, de n_2 filas por p columnas y si R es el resultado obtenido por la técnica al procesar el conjunto $Z=\begin{bmatrix} X \\ Z \end{bmatrix}$ con $n = n_2 + n_1$ filas por p columnas, entonces puede obtenerse R empleando solamente la información contenida en R_1 y R_2, aún en la ausencia de los datos X y Y.

La técnicas con esa propiedad permiten obtener los resultados *R* usando un ordenador que no pueda contener en su memoria central al conjunto *Z*, procesando sucesivamente *X* e *Y*, combinando R_1 y R_2.

El carácter incremental de una técnica – por ejemplo el cálculo de la media- permite, aún, combinar la información de muestras sucesivas - obtenidas eventualmente por un método de muestreo adaptable - para aumentar la información acerca de la estructura de los datos.

KRZANOWSKY (1979, 2000) demuestra que, dados dos grupos de individuos- de cardinales n_1 y n_2, observados sobre el mismo grupo de variables $x_{(1)}$ ... $x_{(p)}$, si estos dos grupos son sometidos a dos análisis en componentes principales y quedan referidos a sistemas de ejes ortogonales $y_{(1)}$ ... $y_{(p)}$ para el primer grupo y $z_{(1)}$ ... $z_{(p)}$ para el segundo, entonces es posible definir un tercer espacio que resume los resultados de estos dos. Ver **figura 5.3.2.1.**

En la **figura 5.3.2.1.** está representado el caso particular en que los subespacios *X* y *Y*, definidos por los componentes principales (x_1, x_2) y (y_1, y_2) de los dos grupos, son planos del espacio generado por las variables iniciales (z_1, z_2, z_3).

En esa figura

$$y_1 = e_{11}\, z_1 + e_{12}\, z_2 + e_{13}\, z_3$$

$$y_2 = e_{21}\, z_1 + e_{22}\, z_2 + e_{23}\, z_3$$

(o entonces: $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \end{bmatrix}}_{L} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$).

$$x_1 = m_{11}\, z_1 + m_{12}\, z_2 + m_{13}\, z_3$$

$x_2 = m_{21}\, z_1 + m_{22}\, z_2 + m_{23}\, z_3$

(o entonces: $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underbrace{\begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix}}_{M} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$).

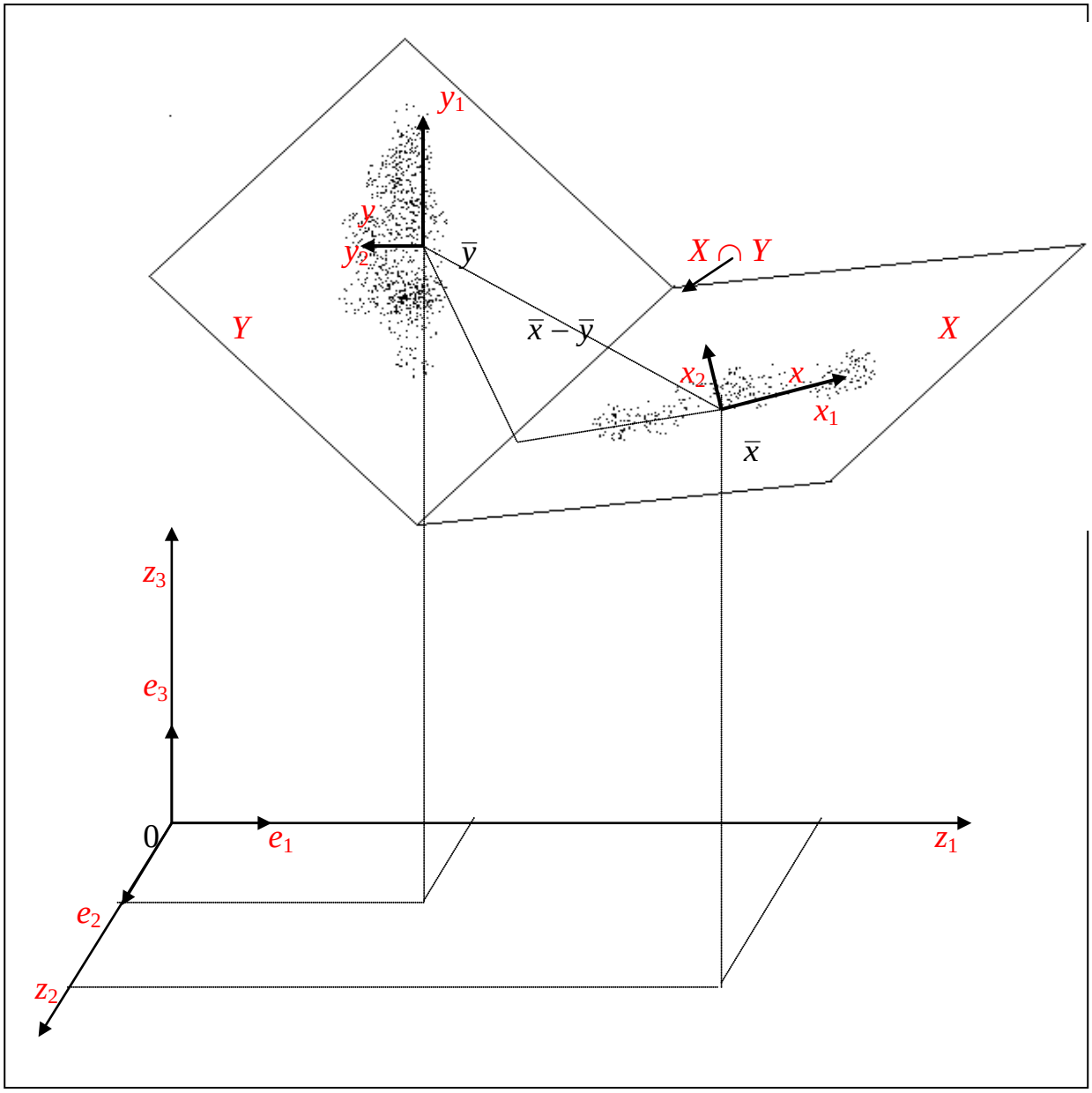


Figura 5.3.2.1. Las dos muestras - que corresponden a la observación de las mismas variables en los mismos individuos o individuos distintos referidos a sistemas ortogonales definidos por sus respectivas componentes principales y al referencial original (variables observadas z_1, z_2, z_3).

La integración en biplots de consenso de biplots resultantes de la observación de los mismos individuos en instantes distintos ha sido

investigada en trabajos de la Escuela de Salamanca. Ver MARTIN-RODRÍGUEZ(1996) y MARTÍN-RODRÍGUEZ *et al* (2002).

Estos trabajos pueden considerarse, en la perspectiva que aquí nos ocupa, como contribuciones para el problema de la incrementabilidad de los biplots una vez que permiten la integración, en un biplot de consenso, de la información adicional obtenida en observaciones parciales.

Pero la construcción de biplots de consenso – MARTÍN-RODRÍGUEZ (1996, 2000) y MARTÍN-RODRÍGUEZ *et al* (2002) – no garantiza, ni es ese su objetivo, que el biplot obtenido por integración de dos o más biplots sea el mismo que se obtendría procesando simultáneamente toda la información.

HALL *et al* (2000, 2002) muestran que la descomposición de una matriz simétrica - la matriz de covarianzas, por ejemplo - en valores y vectores propios (EVD) y también la descomposición de una matriz rectangular en valores y vectores singulares (SVD) tienen esas propiedades.

En HALL *et al* (2002) se presentan esos resultados y sus aplicaciones; en HALL *et al* (2000) se presentan las demostraciones relativas a la (EVD).

Como los biplots son configuraciones de puntos en espacios vectoriales definidos por los vectores singulares a izquierda y a la derecha de la matriz de datos, estos resultados permiten afirmar que los biplots son indudablemente técnicas incrementables en el sentido anteriormente definido.

Específicamente, sea *B* el biplot que se obtendría usando la SVD de $\underset{(n\times p)}{D} = \begin{bmatrix} X \\ \overline{Y} \end{bmatrix}$ en donde $X_{(n_1 \times p)}$ y $X_{(n_2 \times p)}$, con $n = n_1 + n_2$.

Entonces, la SVD de $\begin{bmatrix} X \\ \overline{Y} \end{bmatrix}$ puede obtenerse de la SVD de *X* y de la SVD de *Y*, sin conocer ni *X* ni *Y*. HALL *et al* (2002). O sea, podemos construir el biplot *B* de $\begin{bmatrix} X \\ \overline{Y} \end{bmatrix}$ cargando en la memoria, sucesivamente, *X* y *Y*.

5.3.3. ESCALABILIDAD Y BIPLOTS.

La técnica de descomposición de una matriz *X* en vectores y valores singulares (SVD) se aplica tanto a pequeños como a grandes conjuntos de datos y, como se ha visto en el apartado anterior, es incrementable. Por lo tanto, la técnica de los biplots, basada en esa descomposición, es, en principio, escalable.

Cuando n y p son muy grandes, pueden emplearse formas alternativas de biplots, cada una adaptada a la dimensión del problema parcial, siendo posible relacionar los resultados obtenidos en escalas distintas.

Por ejemplo puede ser útil empezar por usar la técnica MANOVA - biplot - GABRIEL (1998), VICENTE-VILLARDÒN (1992), AMARO-MARTÍN (2002). - para construir un biplot de medias de grupos definidos por un parámetro- tiempo, espacio u otro.

Así se obtiene una estructura macroscópica a gran escala, en la que los individuos son indistinguibles y solo es posible relacionar grandes grupos.

Es esa la idea que se pretende ilustra en la **figura 5.3.3.1**, en donde los 13 individuos pueden considerarse agrupados por los valores de una variable - parámetro que permiten definir las clases C_1, C_2, C_3, C_4, C_5.
Calculando las medias de las restantes variables en cada una de estas clases, podemos usar estas medias para realizar un biplot de medias que permite detectar proximidades entre las clases representadas por esas medias y relacionar esas clases con las variables. Los individuos son indistinguibles.

Al examinar el biplot de medias puede ser posible identificar grupos de clases que convenga, ahora, estudiar con detalle.

Por ejemplo: si las clases – las medias de las clases – C_1 y C_5 forman un grupo y las clases C_3 y C_4 forman otro grupo, puede interesar ahora una visión microscópica, realizando un biplot sobre los individuos que integran eses dos grupos. Los dos análisis han sido realizados a escalas distintas, usando la misma técnica y de modo que los resultados se relacionan.

Otra alternativa consiste en crear tablas de contingencia para ciertas variables estructurales y estudiar los biplots correspondientes, identificando enseguida grupos de clases cuya proximidad indique la necesidad de nuevos biplots, condicionados a esos valores.

En síntesis, se puede decir que la técnica de los biplots es escalable, una vez que la misma técnica (biplots) se aplica a datos agregados - lo que conviene a una visión macro - y a observaciones individuales - lo que conviene a una visión microscópica - siendo posible relacionar los resultados obtenidos a escalas distintas.

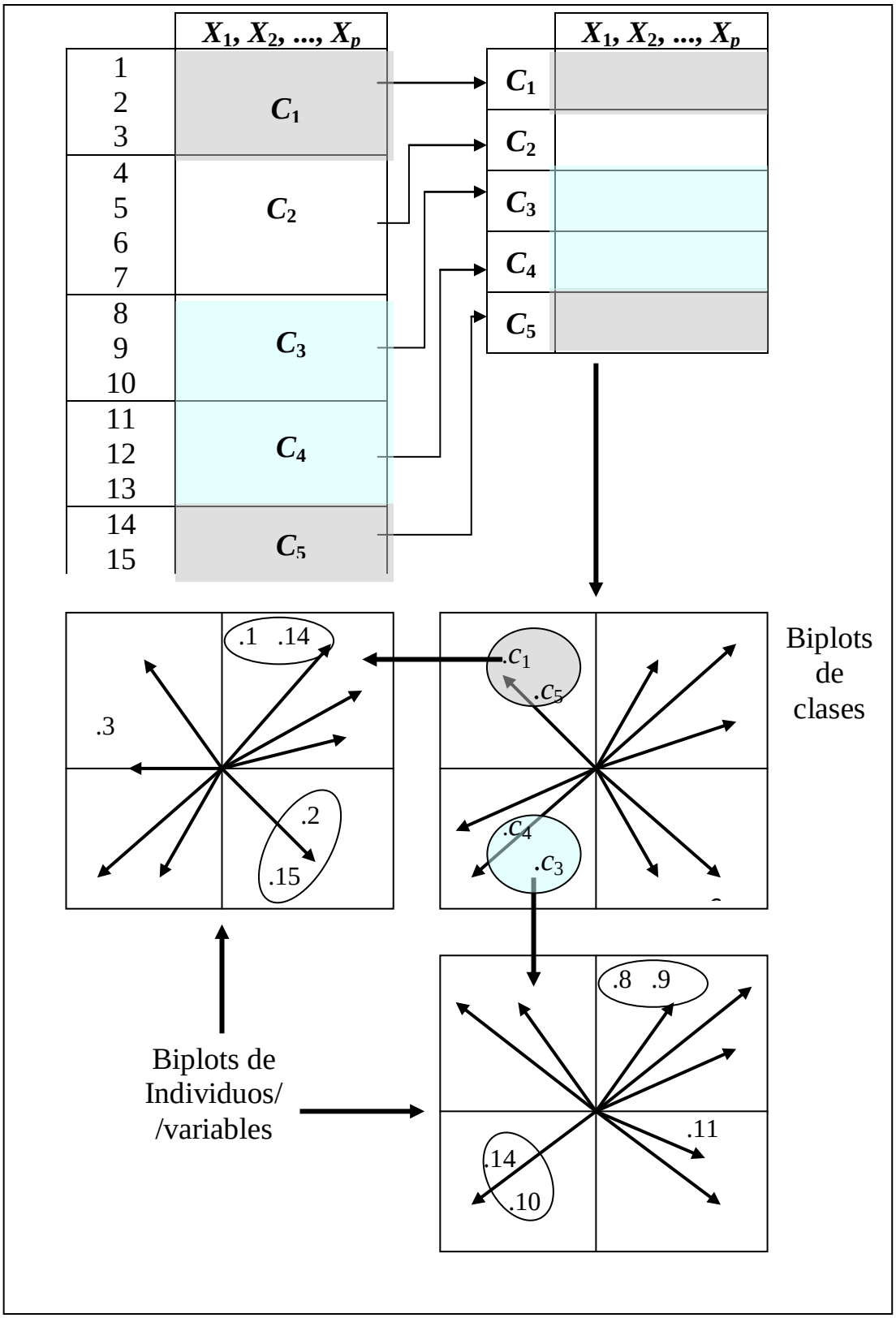


Figura 5.3.3.1. Ejemplo de escalabilidad de los biplots. Un gran conjunto agregado en 5 clases. Un biplot de medias de clases descubre 2 grupos de clases que pueden ser analizados por biplots, específicos, relacionando individuos.

5.4. PROBLEMAS GRÁFICOS.

5.4.1. IDENTIFICACIÓN DE PROBLEMAS.

La creación de buenos gráficos usando el ordenador ha desencadenado la investigación teórica acerca de este asunto. Ver SCOTT (1992), CLEVELAND (1993a, 1993b), COOK *et al* (1993), FURNAS *et al* (1994), CARD *et al* (1999), POSSE (1995), TREMMEL (1995), UNWIN *et al* (1996), BECKER *et al* (1996), BERTIN (1977), HUANG *et al* (1997), WILLS (1999), WILKINSON (1999), VALOIS (2000), WARE (2000), SPENCE (2001).

La creación de buenos gráficos, usando el ordenador, para volúmenes de datos que se miden en terabybes es un problema reciente que ha desencadenado una intensa investigación teórica.

Según CIPRA (1999):

> « *... To flesh a terabyte of data on a 1000 × 1000 screen, you need to cram a megabyte of data into each pixel.*
> *Something's got to give and the researchers haven't yet figured out what's esential and what's superfluores.*
> *Today's massive graphs don't fit the memory ...*».

El lenguaje expresivo del párrafo anterior sintetiza muy bien la dimensión y naturaleza, aún mal definida, de los problemas que se crean cuando se intenta visualizar enormes conjuntos de datos.

Con los conocimientos actuales, las soluciones específicas son compromisos entre factores discrepantes: factores humanos (psicológicos y físicos, relacionados con el sistema visual), volumen de los datos, características del procesador, dimensión de la pantalla, número de pixel.

Al crear un SMDBB, hay que distinguir entre problemas generales de visualización y los problemas específicos asociados a la visualización basada en biplots.

La revisión de la literatura sobre este asunto ha permitido identificar los siguientes problemas generales, objeto de investigaciones específicas aún no fundidas en una teoría general:

Separabilidad - En esta designación hemos incluido los problemas de **superposición** de marcadores, los problemas de **separabilidad** de los símbolos y la creación de gráficos de densidad cuando los problemas de superposición y separabilidad no pueden resolverse.

HUANG *et al* (1997) analizan los problemas de superposición que ocurren en gráficos de dispersión y llegan a la conclusión de que la elección entre símbolos individuales y un símbolo aglomerativo (representando más que un individuo) debe ser función del volumen local de superposiciones.

Esto significa que la decisión de usar para marcadores de las observaciones símbolos individuales o símbolos aglomerativos es un problema local que depende del número de observaciones, del tamaño y forma de los símbolos individuales y del área disponible para el gráfico. La solución encontrada por esos autores - VAREBIPLOTS – consiste en variar la representación de las observaciones, en el mismo gráfico, entre símbolos individuales,

símbolos aglomerativos (representando más de una observación). La decisión de usar uno u otro se basa en el cálculo de una función de transferencia (*transfer function*) que calcula la cantidad de «tinta» a atribuir a cada división del gráfico.

TREMMEL (1995) investiga el problema de la separabilidad de marcadores en gráficos de dispersión, en función de la forma de los símbolos, de su contenido (llenos o vacíos) y del contraste entre símbolos considerando la forma y el contenido.

Cuando el número de objetos por representar sobre un gráfico supera un límite dado, la mejor solución es usar un gráfico de densidad: los objetos dejan de poder tener individualidad y sólo influyen por su número. Ver HUANG *et al* (1997).

Valores faltantes.

En UNWIN *et al* (1996) puede verse un análisis de los problemas de los valores faltantes y sus efectos en la construcción de gráficos. En particular se presentan las soluciones encontradas para desarrollar el sistema MANET (*Missing Are Now Equally Treated*).

Conexión (*linking*).

Este concepto ha sido objeto de investigaciones recientes y asume relevancia en la literatura revisada. Ver CLEVELAND (1993), BECKER *et al* (1996).

El concepto básico es el de «*treillis*»: familia de gráficos, indexada por una variable, de modo que cada gráfico representa un aspecto particular de los

datos. Los mismos objetos son representados en gráficos distintos de la familia por símbolos o valores idénticos.

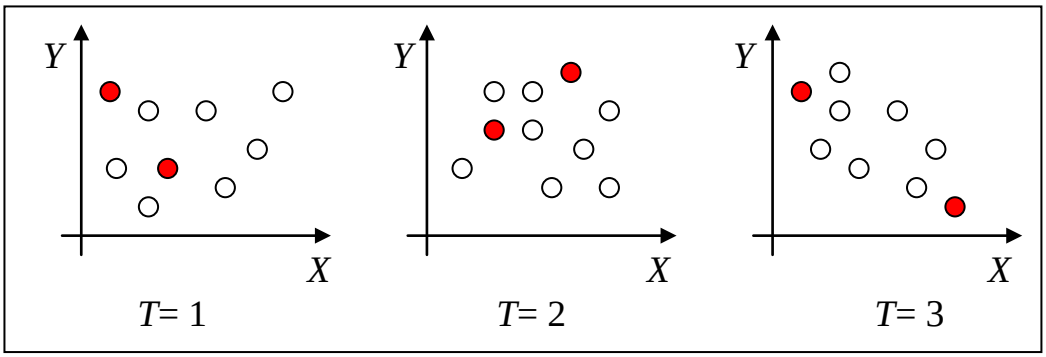


Figura 5.4.1.1. Una familia de gráficos de dispersión indexada por la variable $T \in \{1, 2, 3\}$. Los mismos objetos están representados en todos los gráficos por los mismos símbolos y colores.

Proyecciones óptimas.

Gran parte de la literatura revisada respeta, de un modo u otro, a gráficos de dispersión planos.

La idea básica de la técnica conocida por "***proyección pursuit***" consiste en examinar datos de dimensión muy elevada buscando encontrar proyecciones de dimensión *1*, *2*, *3* que sean «interesantes» según un criterio.

El concepto de «interesante» en general se mide por el grado de **alejamiento** con relación a lo normal: en general se buscan las proyecciones más alejadas de la normal.

Otra técnica íntimamente relacionada con ésta se designa por «*gran tour*» y permite una variación continua de las proyecciones representativas de todas las proyecciones de los datos. COOK *et al* (1995).

La investigación acerca de este paradigma ha permitido desarrollar importantes conceptos de interactividad y el *software* Xgobi. Ver COOK *et al* (1995), POSSE (1995), BUJA (1996).

Estas investigaciones prueban la gran importancia atribuida al problema de la representación plana de datos multivariantes - aunque en una perspectiva distinta de la de los biplots.

Una distinción importante en el plan teórico y práctico es la distinción entre proyección y sección - y el concepto asociado de *prosección*: proyección de una sección de los datos. Ver FURNAS *et al* (1994). Ver **figura 5.4.1.2.**

La idea de prosección puede verse en la **figura 5.4.1.2.**

Datos

i	*X*	*Y*	*Z*
1	0	7	5
2	5	12	10
3	10	10	12
4	15	14	7
5	12	10	5
6	15	7	15
7	18	10	7
8	0	1	10
9	5	7	10
10	11	5	7

Sección

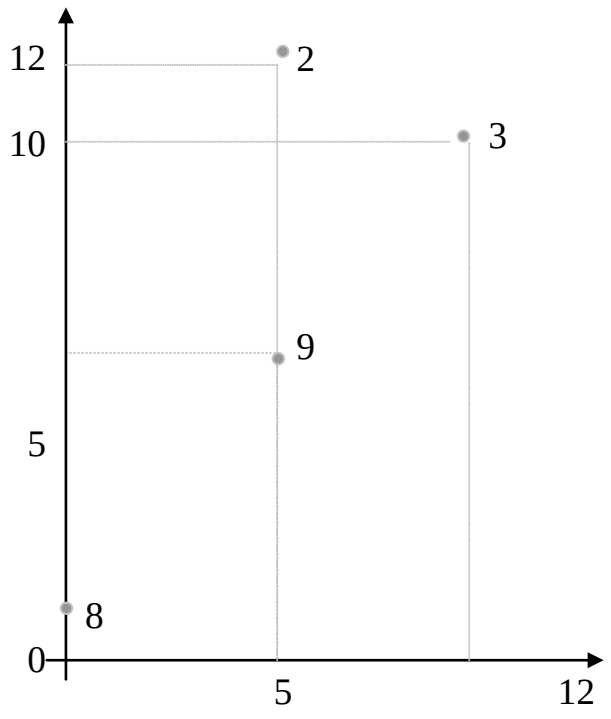


Figura 5.4.1.2. Proyección sobre el plan *XY* de los datos que pertenecen a la sección $9 \leq Z \leq 12$.

La explotación de datos multivariantes basada en el concepto de "*prosección*" se ha revelado importante a la hora de construir la interactividad gráfico/datos en el *software* gráfico.

Cuando se intenta representar en un mismo plano dos tipos de información (individuos y variables), construyendo un biplot, la complejidad del problema de crear buenos gráficos aumenta considerablemente.

Desde luego, es necesario distinguir, en todos los biplots, los individuos de las variables. Esto no es una tarea trivial: si representamos a los individuos por puntos y a las variables por vectores, por ejemplo, cuando el número de variables es grande, el gráfico es ilegible, aún para un número moderado de variables.

Por otro lado, una vez que el biplot resulta de la superposición de dos gráficos de dispersión cuyos ejes son los ejes factoriales, distintos tipos de biplot crean problemas distintos, que hay que abordar.

Puede ocurrir que en un biplot - el CMP, por ejemplo - las variables tengan una buena representación pero que casi todos los individuos formen un "borrón" como el la **figura 5.4.1.3.**

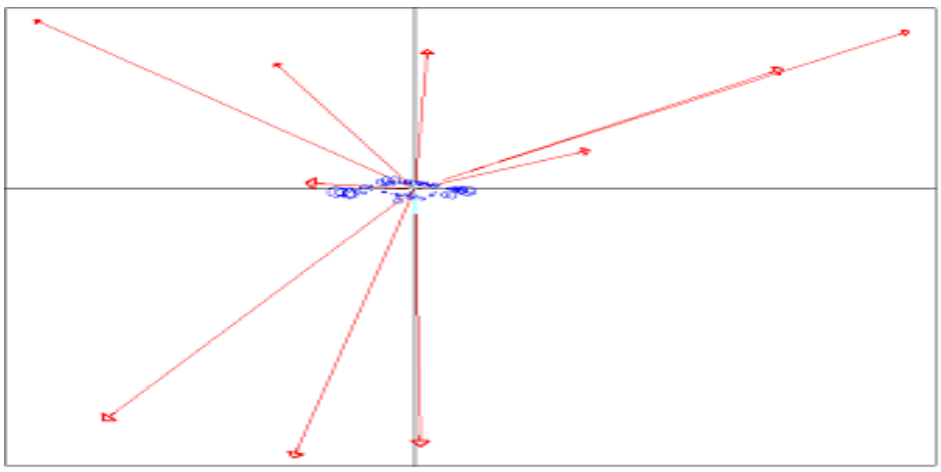


Figura 5.4.1.3. Un GH/ CMP-Biplot (α=0).Los individuos están todos en el centro.

Otra dificultad con los biplots de grandes conjuntos de datos ocurre en la interactividad datos/gráfico: lo que es bueno para los individuos no es necesariamente bueno para las variables.

En efecto, individuos y variables son entidades semánticamente distintas y, por otro lado, sucede muy frecuentemente que $n >> p$.

Por ejemplo: si $p = 20$ y $n = 2000$, tiene sentido presentar en el gráfico los identificadores de las variables pero no tiene sentido presentar los identificadores de los individuos.

En este tipo de gráfico es muy importante usar las características de los símbolos - la dimensión del símbolo, por ejemplo - para reflejar la calidad de representación, función de la contribución relativa de los ejes a los objetos (individuos y variables). Ver **capítulo II**.

Pero cuando el número de objetos en el gráfico sobrepasa un valor dado, esa regla puede ser impracticable.

Al considerar biplots para datos de tres vías - por ejemplo, biplots de consenso - la complejidad de los problemas aumenta aún más; ahora, es necesario separar tres tipos de información: individuos, variables, tiempos.

Ver KROONENBERG (1983), CARLIER *et al* (1998), MARTÍN-RODRÍGUEZ (1996), MARTÍN-RODRÍGUEZ *et al* (2002).

5.4.2. UNA PERSPECTIVA "ECOLÓGICA" DE LOS GRÁFICOS.

El término ecológico en el título significa que la ubicación en la pantalla de un ordenador de símbolos con los cuales intentamos transmitir información útil acerca de los datos, crea problemas semejantes a los encontrados en sistemas biológicos y sociales. También aquí hay que condicionar el comportamiento de los símbolos a los objetivos globales, a la escasez de recursos computacionales (tiempo y espacio) y al numero de símbolos.

El problema de la definición de una interfaz gráfica para un sistema de minería de datos en general, y para un SMDBB en particular, es un problema complejo que exige mucha investigación y experimentación.

Las soluciones adoptadas son siempre compromisos entre factores contradictorios.

De la literatura revisada concluimos que, para el caso de gráficos que correspondan a grandes conjuntos de datos - con dimensiones y contenidos muy variados - la idea de adaptar la solución a las condiciones locales puede ser interesante.

Ver, por ejemplo, el problema de los VAREBI PLOTS en HUANG *et al* (1997) en donde se verifica que la solución para la representación depende de la concentración local de observaciones: si la densidad local es muy grande, las observaciones pierden la individualidad gráfica y lo que se presenta al analista son símbolos colectivos. En zonas del gráfico en donde la densidad es baja, las observaciones no pierden su individualidad gráfica.

Podríamos generalizar la solución encontrada por HANG *et al* (1997) a otros aspectos de la construcción de gráficos, formulando el principio de que la solución debe depender de las condiciones locales.

Estas observaciones nos han llevado a formular, para este proyecto, una estrategia que podríamos designar por **ecológica**, explicada a continuación:

Podemos pensar en los marcadores - tanto de individuos como de variables - como objetos cuyo comportamiento en el proceso de construcción del gráfico es limitado por las condiciones siguientes:

1. Los marcadores *compiten* unos con los otros en la obtención de recursos de la pantalla del ordenador (pixel, brillo, etc.) «*buscando captar*» la atención del analista.
2. Los recursos de la pantalla son limitados: hay un número limitado de pixel (por ejemplo, 1000 × 1000) y el tiempo de procesamiento de cada uno de los símbolos gráficos también es limitado.
3. La capacidad que tiene el analista para distinguir entre dos marcadores - TREMMEL (1995) y WEGMAN (1995)- es también muy limitada: depende de la distancia de los ojos a la pantalla, de las estructuras del sistema visual humano y de las características gráficas de los símbolos.

De aquí resulta que, cuando el número de marcadores a representar es pequeño, es posible que los marcadores «*se presenten*» representados por sus nombres, sin que eso afecte la capacidad de los demás para hacer lo mismo.

Cuando el número de objetos (individuos y variables) a presentar aumenta, deja de ser posible que los marcadores *se identifiquen* por sus nombres. Pueden, entonces, representarse por símbolos gráficos de dimensión sucesivamente más pequeña, en función del número de marcadores presentes y, por lo tanto, de los recursos disponibles en términos de pixel, de tiempo de procesamiento y de capacidad de separación del analista. Este deja de poder distinguir dos formas cuando la dimensión de los símbolos es muy reducida.

Todo esto significa que, para un número de marcadores por encima de determinado límite, es necesario que los marcadores dejen de tener individualidad gráfica y solo contribuyan para el gráfico por su número en un reticulado: la influencia de un marcador en determinada zona de la pantalla es, entonces, función de la distancia del marcador a ese punto y del número de marcadores que son relevantes en ese punto. Lo que el analista ve es, entonces, la suma de estas influencias representadas, por ejemplo, por una densidad estimada por un estimador KERNEL. Ver **figura 3.5.2.1.**

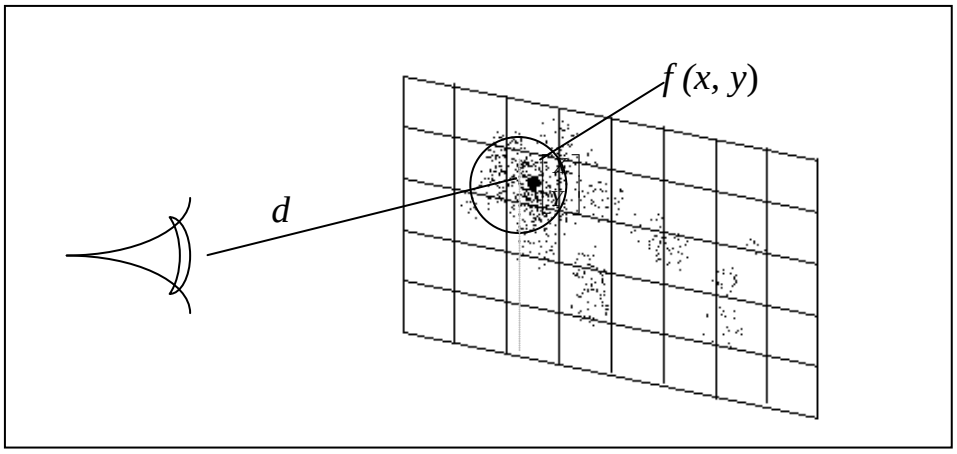


Figura 3.5.2.1. La ecología del gráfico en un punto depende del punto [*x, y*] depende de *d*, del número de puntos *n*, del número de pixel, de la densidad de puntos f(x,y) y del tiempo para pintar un símbolo.

El problema ecológico en un punto $\begin{bmatrix} x \\ y \end{bmatrix}$ de la pantalla se caracteriza, esencialmente, por:

d - Distancia de la pantalla a los ojos del analista (capacidad de separación de dos marcadores).

f(x, y) - Densidad de los marcadores en el punto (*x, y*).

m -Número total de marcadores por representar (*m = n + p*).

pix - Número total de pixel de la pantalla.

t- Tiempo necesario para pintar un símbolo.

La solución gráfica a emplear en la célula que contenga (*x, y*) depende de *d*, *f (x, y)*, *m*, *pix*, *t*.

En este trabajo se designa por visión ecológica el reconocimiento de que la solución gráfica debe depender de estos factores.

La solución gráfica engloba las decisiones siguientes:

1. Representación de las observaciones por símbolos individuales o densidades.
2. Símbolo a usar para representar cada tipo de marcador.
3. Color de los símbolos.
4. Dimensión de los símbolos.
5. Tipo de interacción gráfico/datos.

La posibilidad de modelar los marcadores como objetos que, pueden adecuar su *comportamiento* a las condiciones ecológicas locales, está relacionada con el actual paradigma de programación basada en el concepto de clases y objetos (programación orientada por objetos).

En este paradigma es posible definir una clase de objetos a la que podríamos llamar MARCADOR con propiedades como color, dimensión, forma, contenido y comportamientos - métodos o procedimientos - función de los parámetros indicados.
La obtención de soluciones óptimas para el problema tal como ha sido formulado exige, por supuesto, investigaciones que sobrepasan el objetivo de este proyecto.

Los compromisos/decisiones implementadas en este trabajo han sido obtenidos con base en esta visión ecológica pero en bases empíricas.

De entre estas decisiones se destacan el hecho de que la dimensión máxima de un símbolo es función del número de observaciones y del numero de variables; la exclusión de símbolos cuyo tiempo de presentación sea grande; la imposición de usar sistemáticamente un símbolo y color distintos para los individuos y para las variables; la posibilidad de substituir la nube de puntos por un gráfico de densidad.

5.5. INTERACTIVIDAD BIPLOT-DATOS.

5.5.1. INTRODUCCIÓN.

Dado que los biplots son gráficos, un SMDBB es casi siempre un sistema visual de minería de datos, en el que la intervención del analista en el desarrollo del proceso es crucial.

Cuando el conjunto de datos por analizar es muy grande, este hecho crea dos tipos de problemas, como se ha visto en los apartados 5.3 y 5.4: el

primero tiene que ver con la capacidad de cálculo del ordenador; el segundo tiene que ver con las capacidades del cerebro humano (sistema visual y capacidad de la memoria de trabajo).

Del punto de vista de la capacidad de cálculo del ordenador, debe atenderse a que la realización, en tiempo real, de todos los cálculos necesarios a la producción de los gráficos (biplots) y a su actualización en resultado de las decisiones del analista, exigen gran potencia de cálculo y, en muchos problemas, ordenadores con procesamiento paralelo. Ver KOLDA *et al* (1999).

En lo que respeta a la capacidad de cálculo, es evidente que las tareas de minería de datos para volúmenes de datos gigantes (terabytes) no pueden ser realizadas por ordenadores de mesa, a pesar de sus capacidades crecientes.

Por otro lado, la interpretación sin la ayuda de *software* adecuado, de los resultados que ocurren al analizar conjuntos de datos muy grandes implicaría que el analista fuera capaz de memorizar y ordenar miles de identificadores y realizar en segundos, sin error, las múltiples operaciones lógicas necesarias para decidir que hacer enseguida.

Estas realidades implican que el ambiente interactivo realice esas operaciones auxiliares de interpretación, en respuesta a decisiones del analista.

La interactividad del sistema SMDBB ha sido desarrollada en consistencia con las consideraciones teóricas del **capítulo IV**, en particular con los

postulados acerca de las operaciones básicas de interpretación, definidas en el apartado 4.2.

A continuación se presentan las soluciones adoptadas para algunos de estos problemas en el contexto del desarrollo de nuestro sistema prototipo.

5.5.2. OPERACIONES INTERACTIVAS DE COMPARACIÓN DE GRUPOS.

En el **capítulo IV**, apartado 5.4, se ha postulado que, al interpretar los resultados, los analistas realizan, básicamente, las operaciones siguientes:

- Identifican grupos.
- Caracterizan grupos.
- Comparan grupos.
- Nombran grupos.
- Reconocen que un grupo es el soporte de un patrón.

En este apartado se especifican los principios y procedimientos que permiten la realización de las operaciones de identificación, caracterización y comparación al implementar un SMDBB.

1. **Identificación de Grupos.**

   Los grupos de individuos tienen origen muy distintos pero, en el contexto de un SMDBB, todos esos grupos pueden ser visualizados por configuraciones de marcadores en el biplot adecuado.

   Los orígenes de estos grupos son:

   - Selección en la base de datos de individuos que obedezcan a un criterio que pueda ser expresado por variables observadas (*query*).

- Conjunto de individuos que sea el soporte de una hipótesis enunciada por el analista.
- Conjunto arbitrario de individuos marcados - incluso aleatoriamente.
- Conjunto de individuos definido automáticamente por un algoritmo de análisis cluster.
- Conjunto de individuos marcados sobre el biplot por el analista tras el examen visual del mismo.

Una vez definido el grupo de individuos, su presentación sobre el biplot es fundamental una vez que permite situar ese grupo en el contexto de otros grupos con significados, obtenidos por otros métodos.

Esta presentación puede realizarse de modos distintos, de acuerdo con el contexto:

1. Presentación del cierre convexo de los marcadores de los individuos del grupo (usando colores para distinguir los grupos).
2. Presentación del centro de gravedad de los marcadores de los individuos del grupo, usando un símbolo especial para representar estos centros.
3. Presentación de los marcadores de los individuos del grupo usando un color común para todos los marcadores del grupo.
4. Presentación de los marcadores de los individuos del grupo usando un símbolo común.

Todas esas técnicas refuerzan la idea de que los individuos del grupo tienen «algo» en común.

2. **Caracterización de grupos.**

La caracterización de un grupo tiene por objetivo, como se ha visto en el apartado 4.2, expresar en un lenguaje cercano del lenguaje humano, lo que hay de común entre los individuos del grupo. Estas expresiones, como se ha postulado en ese apartado, deben ser construidas usando las variables observadas.

La formulación matemática de este problema de aproximación de conjuntos ha sido realizada en el apartado 4.2.

En este apartado presentamos una alternativa que se puede designar por empírica o experimental.

Cuando los datos son numéricos y forman muestras de distribuciones normales, existen procedimientos estadísticos bien establecidos que podrían aplicarse. Ver SEBER (1983), ANDERSON (1984), RENCHER (1995), KRZANOWSKI(2000).

En minería de datos la situación más frecuente es la de datos mixtos. Este hecho, asociado a la gran dimensión de las muestras y al hecho de que el contexto es el análisis preliminar de datos, tornan problemático el uso de inferencia estadística multivariante como el MANOVA.

La construcción automática de síntesis o sugestiones de síntesis exige el empleo de métodos que permitan «razonar» sobre los hechos (*findings*) revelados por los análisis.

En ese contexto se reconoce la relevancia de modelos construidos usando el concepto de *conjunto fluido* (*fuzy set*), propuesto por ZADEH (1965).

Basado en ese concepto son muchos los modelos que han sido construidos para el control de máquinas y análisis de datos. Ver BEZDEK *et al* (1992).

En nuestro trabajo – que consideramos experimental- hemos preferido investigar la posibilidad de realizar esa caracterización sin abandonar el lenguaje de la estadística.

La caracterización de un grupo se realiza en dos fases:

1ª fase - Caracterización estadística del grupo, usando la estadística descriptiva univariante y, eventualmente, las matrices de covarianza y correlación para las variables cuantitativas y para relacionar las variables cualitativas.

El resultado final de esta fase es una síntesis formada por «hechos relevantes» que buscan evidenciar lo que separa el grupo del conjunto de datos al que pertenece.

Las variables que integran esa síntesis forman la lista de «variables relevantes» para caracterizar el grupo.

2ª fase - Búsqueda de una expresión de tipo conjuntivo o disyunción de expresiones conjuntivas (ver apartado 4.6) que aproxime el significado del grupo «lo mejor posible».

Estas expresiones utilizan solamente las «variables relevantes» y el número de conceptos atómicos que las integra es limitado por consideraciones psicológicas: la capacidad de la memoria de trabajo. Ver apartado 4.6.

La estrategia general adoptada para construir la síntesis estadística relativa a la **1ª fase** es la siguiente:

La síntesis final es formada por un conjunto de mensajes en castellano (u otro idioma) que expresan conclusiones fácilmente interpretables y con interés para la toma de decisiones.

Como ejemplo de tales mensajes se presentan las siguientes:

1. «*La media tiene tendencia a aumentar*».
2. «*La media aumenta significativamente*».
3. «*La media tiene tendencia a disminuir*».

…….

14. «*Aumenta la variabilidad*».
15. «*Gran aumento de la variabilidad*».

…….

31. «*Las modas son muy distintas*».

…….

Para cada mensaje - tras un análisis de la semántica del mensaje en castellano, y de como traducir esa semántica en lenguaje estadístico- se definen condiciones (predicados) cuyo valor lógico es una función de las estadísticas básicas de una variable o más, calculadas sobre el grupo y el conjunto de datos completo.

Por ejemplo, el mensaje número *m* integrará la síntesis, si se verifican ciertas condiciones $cond_1$, $cond_2$, ..., $cond_k$.

Esto se traduce por la regla:

generar (m) **si** $cond_1 \wedge cond_2 \wedge ... \wedge cond_k$.

(la decisión de integrar el mensaje en la síntesis depende de la verificación de las condiciones).

A su vez, las condiciones son predicados que pueden traducir resultados de pruebas estadísticas o reglas empíricas.

Ejemplo: generar el mensaje «*La media aumenta significativamente*» cuando se verifica la condición «*el resultado de la prueba de Student para comparar 2 medias (cuando aplicable) es: rechazo de H0 al nivel 0.01*».

El trabajo de análisis para un número significativo de mensajes, para definir las condiciones de su aplicabilidad y para crear las condiciones respectivas es muy grande e implica un grupo que abarque varias disciplinas. En nuestro sistema prototipo se incluyeron, para demostración, solamente algunos de esos mensajes.

3- **Comparación de grupos.**

La estrategia adoptada para realizar la comparación de grupos es semejante. Se definen y caracterizan desde el punto de vista estadístico y lingüístico mensajes en castellano (u otro idioma) relevantes para la construcción de síntesis de esas comparaciones. Esos mensajes deben resumir «*lo que distingue un grupo de otro*».

La inclusión o no de un mensaje específico en la síntesis depende del valor lógico de predicados, calculados en función de estadísticas relevantes para la comparación.

5.5.3. PROYECCIÓN COMO INSTRUMENTO DE INTERPRETACIÓN DE BIPLOTS E INTERACCIÓN CON LOS DATOS.

Se puso de manifiesto en el **capítulo II** que la interpretación de los biplots se basa en cinco reglas principales.

R1 Distancias entre marcadores de individuos: representan semejanzas entre eses individuos.

R2 Ángulos entre variables: representan coeficientes de correlación entre esas variables.

R3 En los biplots de Galindo, la distancia entre un marcador de un individuo y la dirección de una variable: está relacionada con la preponderancia de esa variable en ese individuo.

R4 Proyecciones de los individuos sobre la dirección de una variable: representan los valores de esa variable para todos los individuos.

R5 Proyecciones de las variables sobre la dirección definida por un individuo: representan los valores de las variables para ese individuo.

Nos proponemos, en este apartado, profundizar en el significado de estas reglas y extraer las consecuencias para la creación de ambientes de interacción con el gráfico y los datos.

Consideremos la regla **R4**. Si fijamos una de las variables - que define una dirección específica - representada en el biplot por su marcador b_j, entonces, si los marcadores de los individuos son a_1, ..., a_i, ..., a_n, estas proyecciones son:

$$\left[a_1^T b_j, ..., a_i^T b_j, ..., a_n^T b_j\right]$$

y representan la columna (variable) $x_{(j)}$ - exactamente en el caso de los biplots de Gabriel salvo una isometría en el caso de los biplots de Galindo.

Si ahora consideramos una dirección arbitraria definida por un marcador adicional *d* en ese biplot y proyectamos todos los marcadores de los individuos activos (usados para construir el biplot) sobre ese nuevo marcador, obtenemos los valores

$$y^T = \left[a_1^T d,\ a_2^T d, ..., a_n^T d\right]$$

de las proyecciones en esa dirección.

O sea: hemos definido **una nueva variable** suplementaria, cuyo marcador es, precisamente, *d*. Ver **figura 5.5.3.1.,** en donde *n*= 6

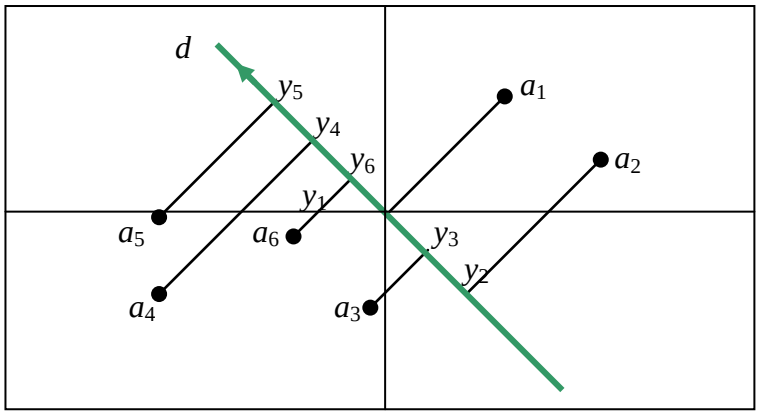


Figura 5.5.3.1. Proyección de los individuos activos a_1 ... a_6, sobre una dirección arbitraria definida por *d*.

Algebraicamente, la situación de la **figura 5.5.3.1.** corresponde a la **figura 5.5.3.2.**

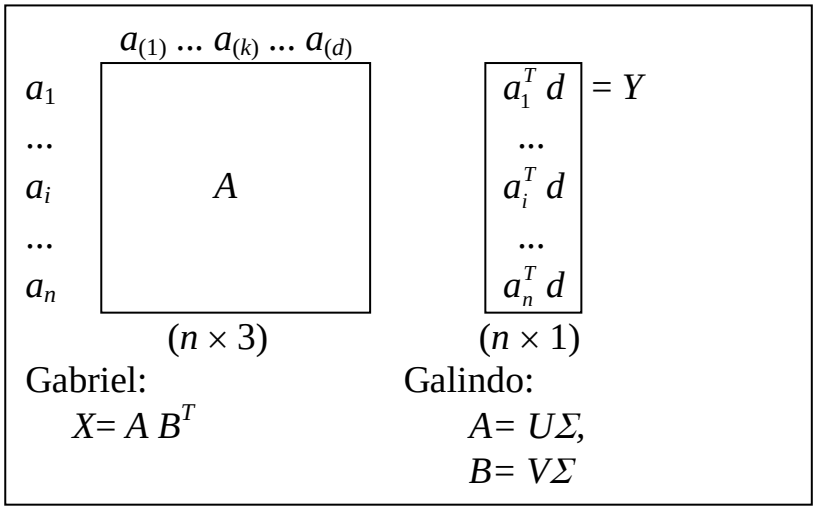


Figura 5.5.3.2. Las proyecciones de los marcadores de los individuos activos - - filas de *A* - sobre una dirección arbitraria de marcador *d* en el biplot - - define una **nueva variable** *Y* (variable suplementaria). En esta figura, $a_i \in \Re^d$, *d* - dimensión del espacio de representación.

La regla **R5** podría ser vista de modo semejante.

Sea, ahora, a_i (i= 1 ... n) uno de los individuos activos usados para la construcción del biplot y b_j (j= 1 ... p) los marcadores de las variables.

Los valores

$$[\, a_i^T b_1,\ a_i^T b_2, \ldots, a_i^T b_j, \ldots, a_i^T b \,]$$

representan el individuo *i* (fila *i* de la matriz *X* de los datos).

Si consideramos una dirección arbitraria definida por un marcador adicional *r* y proyectamos sobre esa dirección *r* todos los marcadores $b_1 \ldots b_p$ de las variables activas, obtenemos el vector

$$Z^T = \left[a_i^T r,\ a_i^T r, \ldots, a_i^T r, \ldots, a_i^T r\right] \in \Re^p$$

Este vector define un nuevo individuo - individuo suplementario *Z*, cuyo marcador es *r*. Ver **figura 5.5.3.3.** , caso p= 3 (3 variables activas).

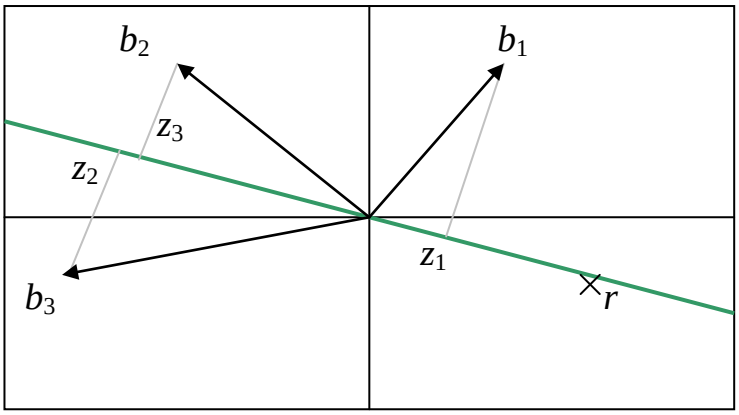


Figura 5.5.3.3. Proyección.

Algebraicamente, la situación corresponde, ahora, a la **figura 5.5.3.4.**

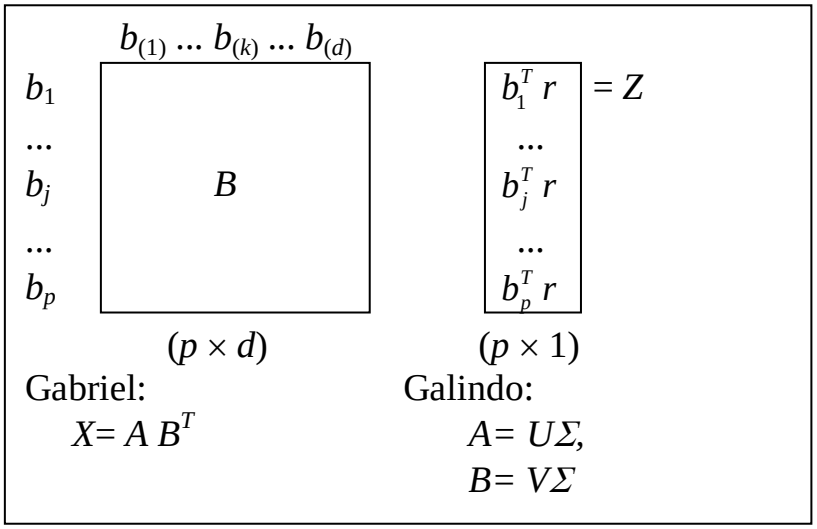


Figura 5.5.3.4. Las proyecciones de los marcadores de las variables activas sobre la dirección arbitraria de marcador *r* - filas de *B* - define el nuevo objeto (suplementario) *Z*. En esta figura, $b_j \in \Re^d$, *d* - dimensión del espacio de representación.

En síntesis:

- Proyectar los marcadores de los individuos activos sobre una dirección arbitraria *d*, equivale a definir una nueva variable - *Y* - observada sobre esos individuos.

- Proyectar los marcadores de las variables activas sobre una dirección arbitraria *r* equivale a definir un nuevo individuo suplementario.

La variable adicional - *Y* - definida por las proyecciones de los individuos sobre la dirección arbitraria tiene especial interés para la interpretación de biplots en los 2 casos siguientes:

Caso 1 (Interpretación de un grupo).

Si la dirección conecta el centro del gráfico con el centro de un grupo, entonces el significado de *Y* está asociado al significado de las variables relevantes para interpretar al grupo.

Caso 2 (Oposición entre 2 grupos G_1 y G_2).

Si la dirección conecta los centros de 2 grupos, entonces la variable *Y* tiene por significado «*lo que opone el grupo G_1 al grupo G_2*»; o sea: está asociada a las variables relevantes para explicar esa oposición.

Los valores de *Y* pueden ser definidos según varias alternativas.

Dado que proyectar según una dirección equivale (desde el punto de vista de la interpretación) a proyectar según una dirección paralela, podemos definir los valores de *Y* por las distancias hasta un punto de referencia; por ejemplo, el punto que representa la proyección del origen de coordenadas del biplot sobre la dirección. Ver **figura 5.5.3.5.**

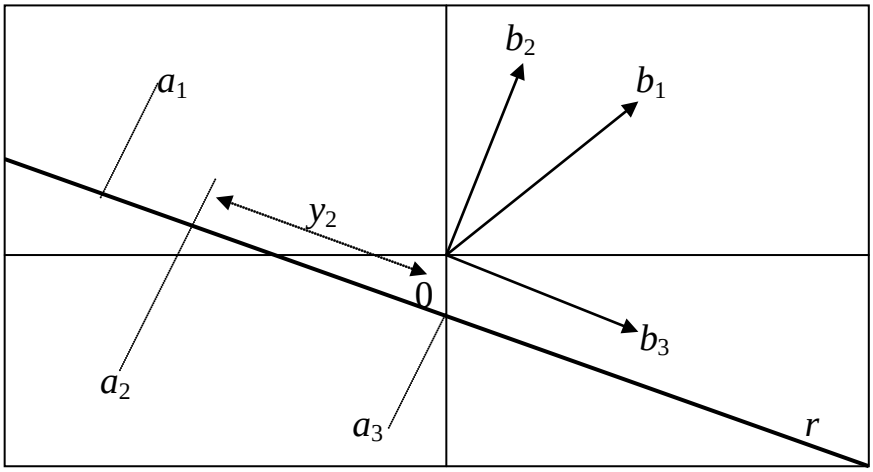


Figura 5.5.3.5. y_2 - valor de Y para a_2 es la distancia entre la proyección de a_2 y de 0 sobre le recta r. 0 es la proyección del origen del biplot.

Una vez definidos los valores $y^T = [y_1 \ldots y_n]$, para buscar el significado de Y en función de los significados de las variables activas podemos correlacionar Y con cada una de las variables activas $x_{(1)} \ldots x_{(p)}$, representadas por las columnas de X.

Se obtienen así los valores de esos coeficientes de correlación:

$$[\, r\,(Y, x_{(1)})\,,\, r\,(Y, x_{(2)}),\, \ldots,\, r\,(Y, x_{(p)})].$$

Las variables activas **más relevantes** para interpretar Y son aquellas que corresponden a los valores más elevados (en valor absoluto) de estos coeficientes de correlación.

Esta interpretación tiene sentido cuando la calidad de representación de las variables activas es elevada.

En efecto, en la **figura 5.5.3.6.** se presenta la situación en donde $x_{(j)}$, de marcador b_j, está mal representada en el plano (F_1, F_2). El marcador b_j de $x_{(j)}$ se proyecta sobre el plano (F_1, F_2) por b'_j.

Se verifica que, aunque la dirección r forma con la proyección b'_j.de b_j. sobre ese plano un ángulo pequeño, la correlación de Y con b_j. es pequeña una vez que $x_{(j)}$ está muy mal representada en (F_1, F_2).

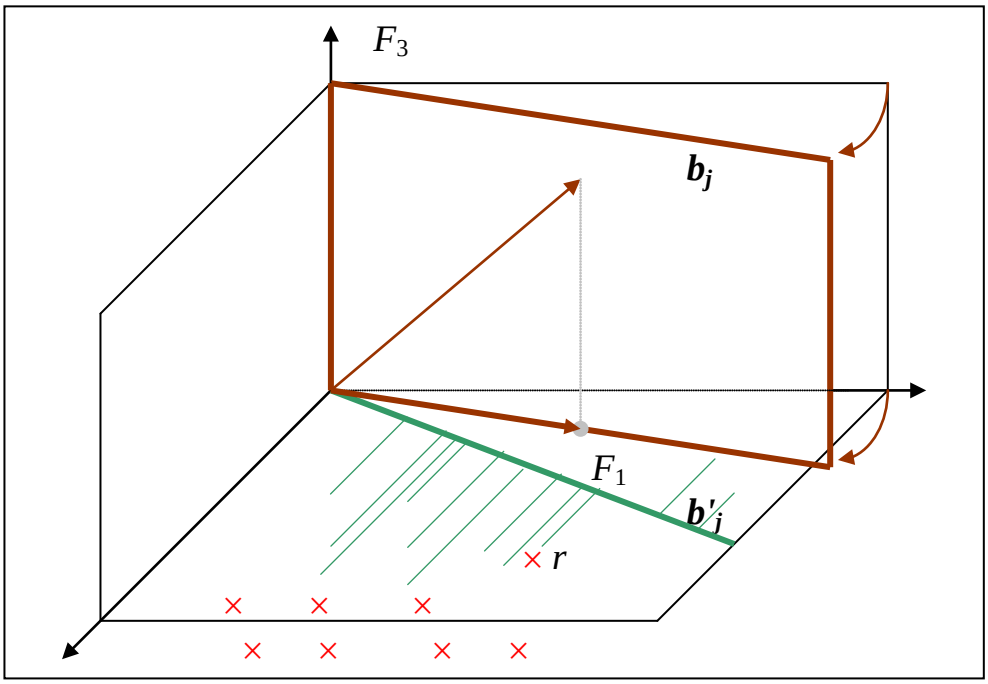


Figura 5.5.3.6. $x_{(j)}$ está muy mal representado en (F_1, F_2) pero su proyección en (F_1, F_2) forma un ángulo pequeño con la dirección r. La correlación (r, $x_{(j)}$) es pequeña pero la observación del biplot (r forma ángulo pequeño con b_j) sugiere lo contrario.

Dada una dirección r y la variable suplementaria Y correspondiente a las proyecciones de los individuos sobre esa dirección, la interacción con los datos se obtiene colocando los datos por orden (creciente o decreciente) de los valores de Y.

Según esta ordenación, quedarán cerca las filas correspondientes a individuos cuyas proyecciones sobre r sean próximas; esto equivale, literalmente, a ordenar los datos según la dirección r; a «mirar» los datos según la «perspectiva» r.

Por ejemplo, si r es la dirección que separa dos grupos, esta ordenación equivale a «mirar los datos según la perspectiva de lo que separa G_1 de G_2».

Ejemplo 5.6.1.

Consideremos los datos presentados en la **tabla 3.1.1.** - resultados obtenidos en un examen de estadística elemental.

En el biplot correspondiente - ver **figura 5.5.3.7.** - se busca estudiar los grupos G_1 (Verde) y G_2 (Rojo). Si usamos los centros aproximados de esos grupos para definir la recta *r*, las proyecciones forman la columna *Y* de la **tabla 5.5.3.1.** Los datos han sido ordenados según los valores de *Y*.

Obsérvese que este conjunto de operaciones, al definir una relación de orden sobre los datos, define una relación de orden sobre cada una de las variables - incluso las categóricas. Transforma esas variables en «series cronológicas».

El estudio estadístico (correlación serial, por ejemplo) de estas «series cronológicas» puede revelar aspectos interesantes de los datos.

Por ejemplo, en la **tabla 5.5.3.1**, se verifica, por una mera inspección visual, que en la perspectiva de lo que separa G_1 de G_2 (Rojo de Verde), los valores de la variable G3B disminuyen sistemáticamente - lo que no ocurre, por ejemplo, para la variable G1A.

En síntesis, el autor cree que este mecanismo simple de proyección puede permitir otros desarrollos interesantes en la perspectiva de interpretación de biplots, que se reservan para trabajos futuros.

Correlacionando estas proyecciones con las variables activas, se obtiene la **tabla 5.5.3.2.,** con los datos ordenados por orden creciente de las proyecciones.

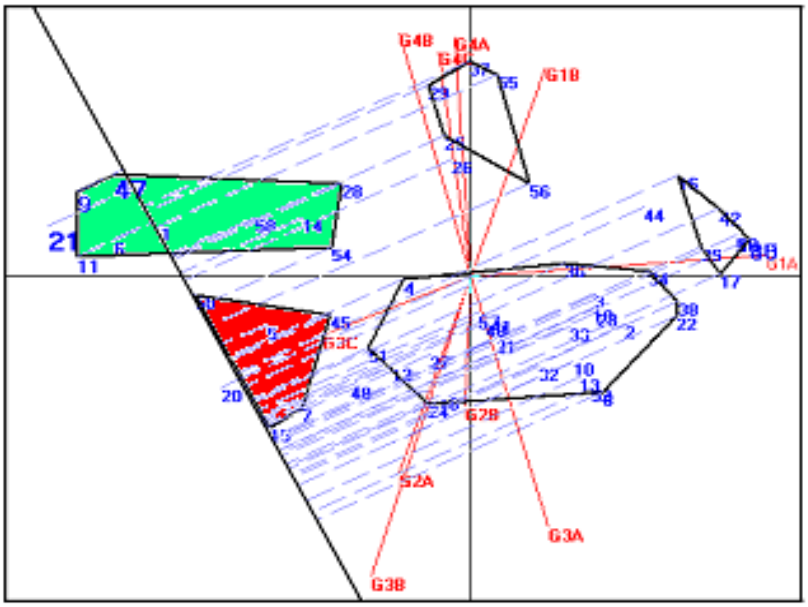


Figura 5.5.3.7. Comparando los grupos Verde y Rojo proyectando todos los Individuos sobre una dirección que separa sus centros.

NºVariable	*Nombre*	*Correlación*
7	G3C	-0.740
6	G3B	-0.739
3	G2A	-0.590
4	G2B	-0.563
9	G4B	-0.531
5	G3A	-0.500
10	G4C	-0.441
8	G4A	-0.378
2	G1B	-0.019
1	G1A	0.102

Tabla 5.5.3.1. Correlaciones de las proyecciones sobre la dirección definida por los centros de los grupos ROJO y VERDE con las variables.

NUMALUM	SIMGRAF	G1A	G1B	G2A	G2B	G3A	G3B	G3C	G4A	G4B	G4C	Y G_1 G_2
20	0	2	0	2	2	2	2	2	1	2	0	**-174**
5	0	1	0	2	0	2	2	2	0	2	2	**-164**
11	0	2	1	2	2	2	2	2	2	2	2	**-140**
24	1	1	0	2	1	1	2	1	0	0	1	**-138**
15	0	2	1	2	2	2	2	2	0	1	0	**-129**
21	0	2	2	2	2	2	2	2	2	2	2	**-117**
9	0	2	2	2	2	1	2	2	2	2	2	**-114**
25	1	0	1	2	0	0	0	1	1	2	1	**-114**
41	1	1	0	2	0	0	1	2	0	0	1	**-114**
12	0	1	0	1	0	2	2	2	1	0	1	**-111**
6	0	2	1	2	1	2	2	2	2	2	2	**-108**
7	0	2	0	2	1	2	2	2	1	0	1	**-107**
14	0	2	1	0	1	1	2	2	1	2	1	**-91**
47	1	2	2	2	1	1	2	2	2	2	2	**-83**
32	1	2	0	2	2	0	1	0	0	0	0	**-71**
48	1	2	1	1	1	2	2	2	0	0	1	**-69**
13	0	1	0	2	0	0	2	0	0	0	0	**-60**
28	1	2	1	1	0	1	1	2	2	2	0	**-39**
1	0	2	2	2	0	2	2	2	2	1	2	**-36**
45	1	2	2	2	2	1	2	0	1	0	1	**-30**
8	0	1	0	0	0	2	2	0	0	0	0	**-28**
18	0	2	1	2	1	2	1	1	0	0	0	**-28**
51	1	2	2	2	2	1	2	0	1	0	0	**-20**
57	1	1	2	0	0	1	2	2	0	0	0	**-20**
33	1	2	0	0	0	1	1	2	0	0	0	**-11**
50	1	2	2	2	2	1	2	2	1	1	1	**-9**
26	1	2	1	0	0	0	1	2	1	1	1	**-4**
31	1	2	1	0	0	1	2	2	0	0	0	**-3**
58	1	2	2	1	2	2	1	1	1	2	1	**0**
10	0	1	1	0	0	2	2	0	0	0	0	**16**
39	1	1	0	0	0	0	0	1	0	0	0	**19**
22	0	1	0	0	0	1	1	0	0	0	0	**23**
27	1	2	2	2	0	1	2	1	0	0	0	**29**
38	1	1	0	1	0	1	0	0	0	0	0	**33**
54	1	2	2	2	0	2	1	1	1	0	2	**33**
29	1	2	2	0	0	0	1	1	1	2	1	**36**
52	1	2	0	0	0	2	2	0	0	0	0	**37**
37	1	2	1	1	0	0	0	0	1	2	2	**40**
43	1	2	2	2	1	1	1	0	0	0	0	**48**
44	1	1	0	0	0	1	0	0	1	0	1	**48**
4	0	2	2	2	0	1	2	0	1	0	1	**55**
23	0	2	0	0	0	2	1	0	1	0	0	**68**
2	0	2	0	2	0	1	0	0	0	0	0	**69**
34	1	1	1	1	0	1	0	0	0	0	0	**79**
17	0	1	0	0	0	1	0	0	0	0	0	**82**
19	0	2	1	1	0	1	1	0	0	0	0	**91**
3	0	2	1	2	0	1	0	0	0	0	0	**104**
56	1	2	2	0	0	1	1	0	1	0	1	**130**
36	1	2	2	2	0	1	0	0	0	0	0	**134**
35	1	1	0	0	0	0	0	0	0	0	0	**139**
55	1	2	2	0	0	1	0	0	2	0	2	**143**
42	1	1	1	0	0	0	0	0	0	0	0	**168**
30	1	2	0	0	0	0	0	0	0	0	0	**204**
46	1	2	0	0	0	0	0	0	0	0	0	**204**
49	1	2	0	0	0	0	0	0	0	0	0	**204**
53	1	2	0	0	0	0	0	0	0	0	0	**204**
40	1	1	0	0	0	0	0	0	0	0	0	**239**
16	0	2	2	0	0	0	0	0	0	0	0	**244**

Tabla 5.5.3.2. La última columna contiene las proyecciones de los marcadores de los individuos sobre la recta que separa los grupos G_1 y G_2 (Verde y Rojo) en la **figura 5.5.3.7.**

5.5.4. ANÁLISIS DISCRIMINANTE INTERACTIVO.

Consideremos el problema de distinguir entre dos grupos G_1 (Verde) y G_2 (Rojo) usando el análisis discriminante.

Desde el punto de vista del análisis preliminar de datos, el objetivo es el de obtener una función discriminante entre dos grupos G_1 (Verde) y G_2 (Rojo). Esto puede conseguirse, por ejemplo, con la función discriminante de Fisher, clasificador que no depende de la distribución de la población. Ver RENCHER (1995).

Desde el punto de vista de la predictibilidad e inferencia estadística, el objetivo es decidir a que grupo G_1 o G_2 atribuir una nueva observación x controlando el error del clasificador. RENCHER (1995).

Estas operaciones - de gran interés en tareas de minería de datos - pueden ser realizadas visualmente, de modo interactivo, sobre un biplot. Ver **figura 5.5.41.**

Si sobre un biplot se identifican dos grupos G_1 (Verde) y G_2 (Rojo) (automáticamente o entonces visualmente) supongamos que los individuos observados corresponden a las filas ($x_{1,1}$ ... $x_{1,n1}$) y ($x_{2,1}$ ... $x_{1,n2}$) del conjunto de datos.

La función discriminante de Fisher viene dada por la expresión - RENCHER (1995):

$a^T x$, en donde

$$a = \left(\overline{x}_1 - \overline{x}_2\right) S_{comb}^{-1}$$

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \qquad (i = 1, 2)\sim$$

$$S_{comb} = \left(\frac{n_1 - 1}{n_1 + n_2 - 2} \right) S_1 + \left(\frac{n_1 - 1}{n_1 + n_2 - 2} \right) S_2$$

x - es una nueva observación - o individuo - a clasificar.

en donde S_1 y S_2 son las covarianzas observadas en los grupos G_1 y G_2, admitiendo que, a nivel poblacional, $\Sigma_1 = \Sigma_2 = \Sigma$.

En el biplot, $\bar{x}_i$ (i= 1, 2) son los marcadores de los centros de los grupos G_1 y G_2: las medias de los marcadores ($x_{i1} \ldots x_{in_i}$) i= 1, 2 correspondientes a los grupos G_1 y G_2, por las propiedades de los biplots descritas en el **Capitulo III**.

Teniendo cuenta que S_{comb}^{-1} introduce solamente cambios de escala, eso significa que, en el biplot, ese vector a está representado por el marcador de la dirección que conecta los marcadores de $\bar{x}_i$ (i= 1, 2).

La transformación lineal $a^T x$ proyecta x sobre la dirección a.

Dado un individuo observado x_i, a que corresponda el marcador a_i en el biplot, $a^T a_i$, queda representado visualmente por la proyección a.

La regla de clasificación asociada a la función discriminante de Fisher es - ver RENCHER (1995) - la siguiente:

«Clasificar *x* en el grupo G_1 cuando

$$a^T x = (\bar{x}_1 - \bar{x}_2)^T \, S_{comb}^{-1} \, x > \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \, S_{comb}^{-1} \, (\bar{x}_1 - \bar{x}_2) + \ln \frac{p_2}{p_2}$$

y en G_2 en el caso contrario».

En esa expresión, $p_1 = \frac{n_1}{n_1 + n_2}$, $p_2 = \frac{n_2}{n_1 + n_2}$ son los estimadores máximo-verosímiles de π_1 y π_2 , las probabilidades de G_1 y G_2, cuando se asumen distribuciones poblacionales normales. Cuando la normalidad no se verifica, la regla de clasificación anterior no es óptima, pero es asintoticamente optima: aproxima la decisión óptima cuando n_1 y n_2 aumentan.

Esta operación de clasificación puede fácilmente visualizarse sobre el biplot pintado en colores distintos las proyecciones de los individuos clasificados en G_1 y en G_2.

Estas ideas están materializadas sobre la **figura 5.5.4.1.** generada por el sistema protótipo descrito en el **capítulo VI**.

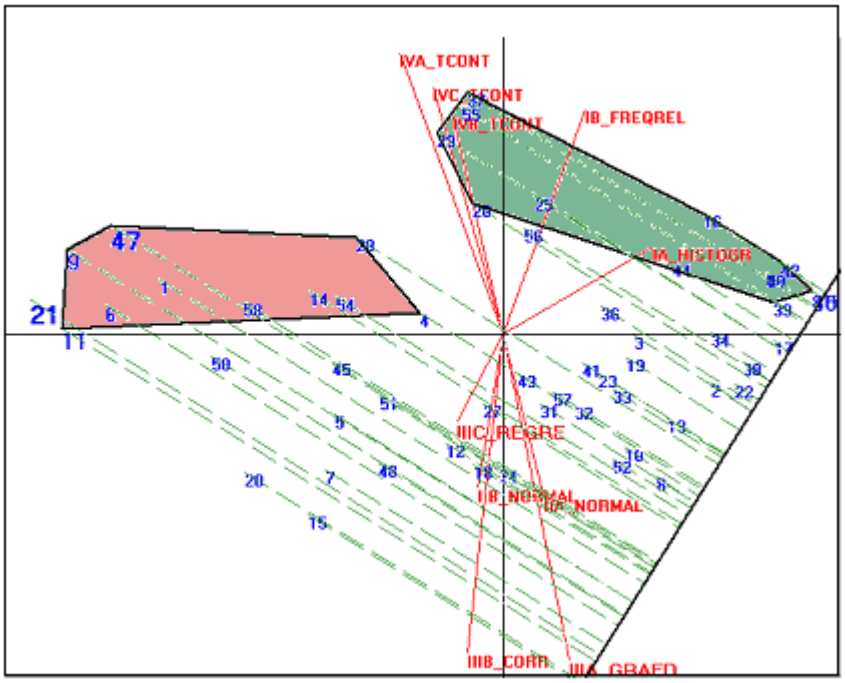


Figura 5.5.4.1. Proyección de dos grupos G1 y G2 sobre la dirección que separa sus respectivos centros.

5.5.5. REGRESIÓN MÚLTIPLE, ELEMENTOS SUPLEMENTARIOS Y PREDICCIÓN.

Consideremos un biplot creado con el conjunto $x_{(1)}$... $x_{(p)}$ de columnas activas y con el conjunto de filas (individuos) x_1 ... x_n, formando la matriz X $(n \times p)$.

Para los biplots de GABRIEL,

$$\underset{(n\times p)}{X} = \underset{(n\times d)}{A}\,\underset{(p\times d)}{B^T} = U\,\Sigma V^T$$

$$A = U\,\Sigma^{\alpha}$$

$$B = V\,\Sigma^{1-\alpha}$$

$$0 \le \alpha \le 1$$

y para el biplot de GALINDO

$$A = U\,\Sigma$$

$$B = V\,\Sigma.$$

Sea Y una variable dependiente de $x_{(1)}$... $x_{(p)}$ cuyos valores han sido observados sobre todos los individuos.

El problema clásico de regresión múltiple - ver RENCHER (1995), SEBER (1983), ANDERSON (1984) - es el de obtener los estimadores $\hat{\beta}$ de β tal que $\|\varepsilon\|^2 = \|Y - \beta X\|^2$ sea mínima, con

$$\underset{(n\times 1)}{Y} - \underset{(n\times p)}{X}\ \underset{(p\times 1)}{\beta} = \underset{(n\times 1)}{\varepsilon} \ .$$

La solución, obtenida por el método de mínimos cuadrados (y también por el método de máxima verosimilitud) es

$$\underset{(p\times 1)}{\hat{\beta}} = \underset{(p\times p)}{(X^T X)^{-1}}\ \underset{(p\times 1)}{X^T Y}$$

Esto significa que

$$\hat{Y} = \hat{\beta}_1 x_{(1)} + \hat{\beta}_2 x_{(2)} + ... + \hat{\beta}_p x_{(p)} + E$$

en donde E son los residuos

$$E = Y - \hat{Y} \ .$$

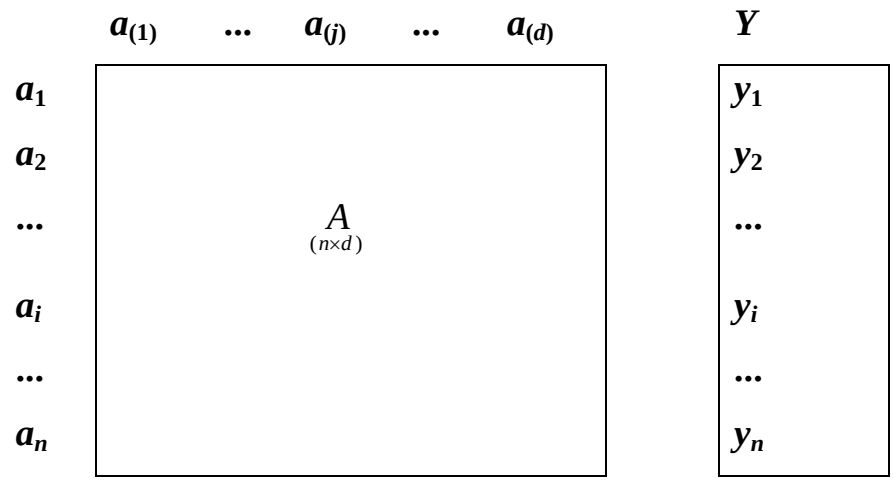
Desde el punto de vista de la representación biplot, el problema consiste en, conociendo los marcadores b_1, b_2,..., b_p para las variables, obtener el marcador d para $\hat{Y}$.

Por las propiedades del biplot (tanto de Gabriel como de Galindo) el marcador d es

$$d = \hat{\beta}_1 b_1 + \hat{\beta}_2 b_2 + ... + \hat{\beta}_p b_p$$

en que $b_i \in \Re^d$, d - dimensión del espacio de representación.

Esta perspectiva está representada en el siguiente esquema:



Aquí, a_i (i= 1 ... n) son los marcadores de los individuos en el espacio de representación, de dimensión d= 1, 2, 3.

Buscar el marcador a_y de Y equivale a resolver el problema de regresión lineal múltiple

$$\underset{(p\times 1)}{Y} = \underset{(n\times d)}{A}\,\underset{(d\times 1)}{\gamma} + \underset{(n\times 1)}{\varepsilon}$$

por el método de mínimos cuadrados, cuya solución es

$$\underset{(d\times 1)}{\hat{\gamma}} = \underset{(d\times d)}{(A^T A)^{-1}}\,\underset{(d\times 1)}{A^T Y}$$

Por lo tanto, $\underset{(d\times 1)}{\hat{\gamma}}$ contiene las coordenadas del marcador de $\hat{Y}$ - la mejor aproximación de Y en espacio lineal generado por las columnas de A.

O sea: $\hat{y} = a_y$ es el marcador de Y en ese espacio.

Esta perspectiva es también la que corresponde a la obtención de marcadores para las variables suplementarias, como se verificará seguidamente - GABRIEL (1995a). **Ver figura 5.5.5.1.**

Dada la matriz X ($n \times p$) formada por los individuos y variables activos, sea

$$X^{+} = \left[x_{(p+1)} \ldots x_{(p+c)} \ldots x_{p+C} \right]$$

la matriz centrada o reducida, matriz cuyas columnas - en número de C - son las variables adicionales cuyos marcadores en el biplot se buscan.

Sea

$$X_{+} = \left[x_{n+1} \ldots x_{n+f} \ldots x_{n+F} \right]$$

la matriz, después de centrada o reducida, cuyas filas - en número de F - son los individuos adicionales cuyos marcadores en el biplot se buscan.

Sea $x_{(p+c)}$ el vector ($n \times 1$) representativo de la columna o variable adicional c (c= 1 ... C).

Podemos representar $x_{(p+c)}$ (c= 1 ... C) por el marcador que aproxima lo mejor posible a $x_{(p+c)}$ por una combinación lineal de las columnas activas.

O sea, por el vector:

$$\underset{(d\times 1)}{\hat{\beta}} = \underset{(d\times d)}{(A^T A)^{-1}} \underset{(n\times d)}{A^T} \underset{(n\times 1)}{x_{(p+c)}}, \; c = 1 \ldots C$$

usando el razonamiento de la regresión múltiple. Ver **figura 5.5.5.1.b)**.

Por un razonamiento semejante, considerando, ahora, la matriz $B_{(p\times d)}$ cuyas filas son los marcadores de las variables activas, un elemento suplementario x_{n+f} (f= 1 ... F) puede ser representado por la combinación lineal de **columnas** de B que mejor aproximan al elemento suplementario x_{n+f} ; o sea, por el marcador cuyas componentes son:

$$\underset{(d\times 1)}{\hat{\gamma}} = \underset{(d\times d)}{(B^T B)^{-1}} \underset{(p\times d)}{B^T} \underset{(p\times 1)}{x_{n+f}} \qquad (f = 1 \ldots F)$$

Las expresiones específicas de $\hat{\beta}$ y $\hat{\gamma}$ para los distintos tipos de biplots clásicos pueden, ahora, obtenerse, recordando que, para los biplots de GABRIEL:

$$X = A B^T = U \Sigma V^T$$
$$A = \left(U \Sigma^{\alpha}\right)$$
$$B = \left(V \Sigma^{1-\alpha}\right)$$
$$0 \le \alpha \le 1$$

y para el biplot de GALINDO:

A= U Σ
B= V Σ.

Para marcadores de las variables suplementarias («olvidando» a n), se obtiene:

Biplots de GABRIEL ($0 \le \alpha \le 1$)

$$\hat{\beta}_{(d\times 1)} = (A^T A)^{-1} A^T x_{(p+c)} \qquad (c = 1 \ldots C)$$
$$= ((U \Sigma^{\alpha})^T (U \Sigma^{\alpha}))^{-1} (U \Sigma^{\alpha})^T x_{(p+c)}$$
$$= (\Sigma^{\alpha} U^T U \Sigma^{\alpha})^{-1} (U \Sigma^{\alpha})^T x_{(p+c)}$$
$$= \Sigma^{-2\alpha} \Sigma^{\alpha} U^T x_{(p+c)} = \Sigma^{-\alpha} U^T x_{(p+c)}$$
$$0 \le \alpha \le 1$$

Biplot de GALINDO:

$$\hat{\beta}_{(d\times 1)} = (A^T A)^{-1} A^T x_{(p+c)} = ((U\Sigma)^T U\Sigma^{\alpha}))^{-1} (U\Sigma)^T x_{(p+c)}$$
$$= \Sigma^{-2\alpha} \Sigma U^T x_{(p+c)} = \Sigma^{-1} U^T x_{(p+c)}$$
$$(c = 1 \ldots C)$$

Para los marcadores de los individuos suplementarios, se tiene:

Biplot de GABRIEL ($0 \leq \alpha \leq 1$)

$$\hat{\gamma}_{(d\times 1)} = (B^T B)^{-1} B^T x_{n+f} \qquad (f = 1 \ldots F)$$
$$= ((V\Sigma^{1-\alpha})^T V\Sigma^{1-\alpha}))^{-1} (V\Sigma^{1-\alpha})^T x_{n+f}$$
$$= (\Sigma^{1-\alpha} V^T V \Sigma^{1-\alpha})^{-1} \Sigma^{1-\alpha} V^T x_{n+f}$$
$$= \Sigma^{-2(1-\alpha)} \Sigma^{1-\alpha} V^T x_{n+f} = \Sigma^{-(1-\alpha)} V^T x_{n+f}$$
$$(f = 1 \ldots F) \qquad 0 \leq \alpha \leq 1$$

Biplot de GALINDO:

$$\hat{\gamma}_{(d\times 1)} = (B^T B)^{-1} B^T x_{n+f} = ((V\Sigma)^T V\Sigma))^{-1} (V\Sigma)^T x_{n+f}$$
$$= (\Sigma^T V^T V\Sigma)^{-1} (V\Sigma)^T x_{n+f} = \Sigma^{-2} \Sigma V^T x_{n+f}$$
$$= \Sigma^{-1} V^T x_{n+f}$$
$$f = 1 \ldots F$$

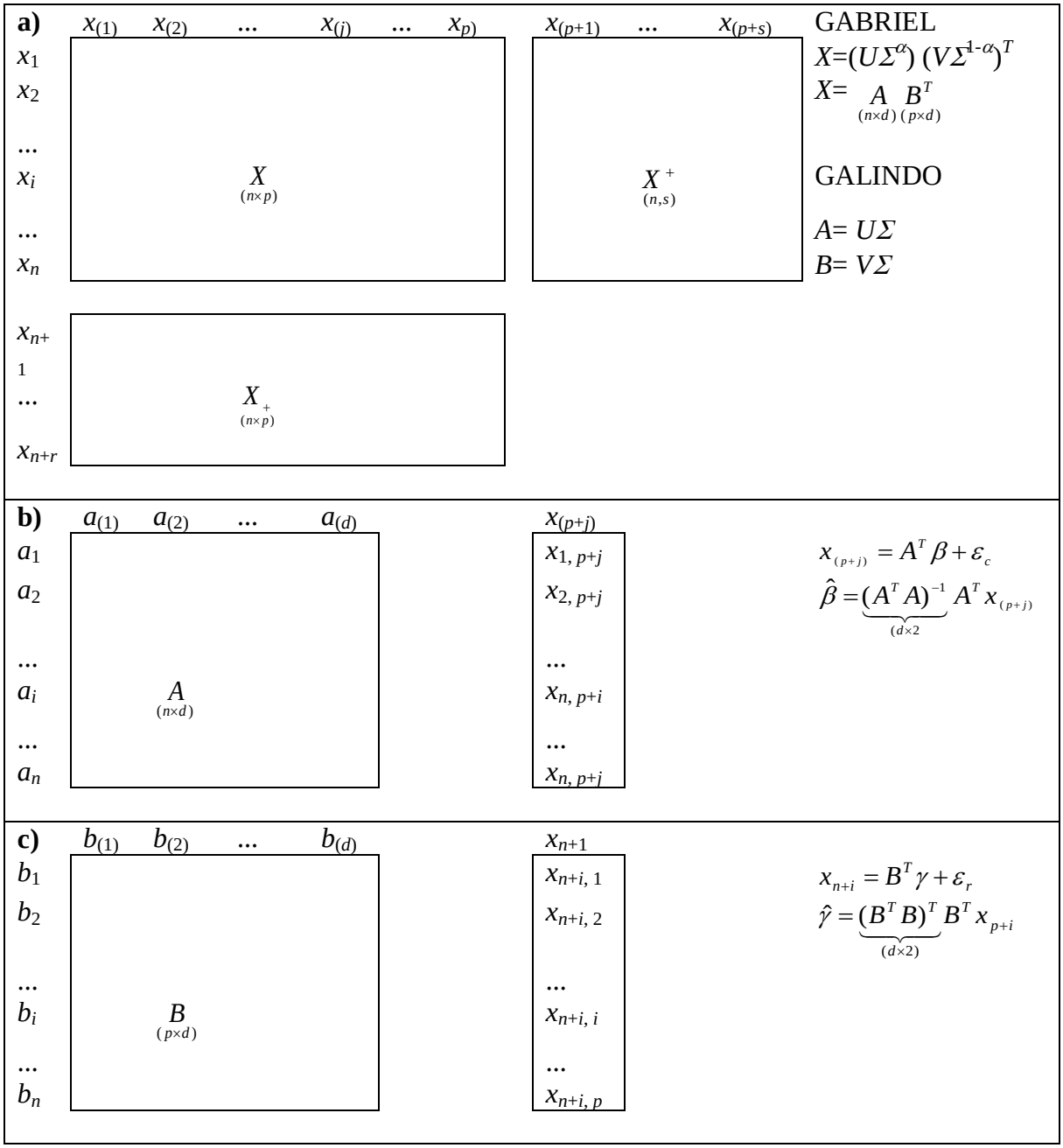


Figura 5.5.5.1. Elementos suplementarios (filas y columnas).**a)** Datos originales y descomposiciones. **b) c)** Elementos suplementarios con relación a los marcadores activos.

C A P Í T U L O VI

BIPLOTS PMD – UN SISTEMA DI MINERÍA DE DATOS BASADO EN BIPLOTS

6.1. INTRODUCCIÓN.

El objetivo de este capítulo es el de presentar un prototipo de sistema de minería de datos basado en biplots que en adelante designaremos por Biplots PMD (Biplots Para Minería de Datos).

Es un sistema de carácter experimental, con el que se busca concretar las ideas teóricas desarrolladas a lo largo de los capítulos anteriores - en especial las ideas acerca de interpretación desarrolladas en el **capítulo IV**.

En el **capítulo V** se han identificado los principales problemas prácticos y teóricos que se plantean en el desarrollo de sistemas de este tipo: problemas computacionales, gráficos y de interactividad. Estos problemas resultan del gran volumen de datos (dimensionalidad y número de observaciones) de las limitaciones de los recursos computacionales y también de las limitaciones humanas.

La gran dimensión de los datos y el gran número de observaciones cambia la naturaleza de los problemas de análisis de datos. La mera manipulación de la meta-información (símbolos, designaciones, significados y tipos de datos) asociada a un sistema de minería de datos de carácter industrial exige la funcionalidad de un sistema de gestión de bases de datos y productos especializados, como diccionarios de datos.

El desarrollo de un sistema de minería exige soluciones orientadas y optimizadas para los datos a los que se destina (orientado para los datos) y ordenadores con gran capacidad de memoria y potencia de cálculo. En KLÖSGEN *et al* (2002) pueden verse nuevos ejemplos y aplicaciones.

Se considera que el programa prototipo que se ha creado para este fin permite mostrar la adecuación de los biplots como instrumento de minería interactiva/visual de datos y también la relevancia práctica de las investigaciones del **capítulo IV**. Por eso, se considera una contribución original.

Las ideas básicas que han orientado la concepción y desarrollo del sistema son las siguientes:

- Facilitar las operaciones de interpretación de biplots de grandes conjuntos de datos.
- Servir de laboratorio a las ideas teóricas del **Capítulo IV**.
- Ser un programa interactivo, permitiendo que el analista determine la secuencia del proceso de análisis en función de los resultados intermedios, presentados en forma gráfica sobre biplots.
- Poseer buenas capacidades gráficas y de interacción datos/gráfico.
- Trabajar con los datos cargados en la memoria central del ordenador, aunque fuera necesario trabajar con muestras.
- Funcionar sobre ordenadores de bajo costo.

Las características del producto final que ahora se presenta son las siguientes:

- Sistema operativo: Windows (98, ME, 2000, NT, XP).
- Lenguaje de programación: Pascal en su versión Delphi de BORLAND.
- Dimensiones del paquete de datos: limitado por la capacidad de la memoria central del ordenador, velocidad del procesador y capacidad del sistema gráfico de presentación.
- Dimensión del programa ejecutable: 1600 Kb.

El sistema ha sido desarrollado con un ordenador de tipo secretaria (desk top) con 256 Mb de memoria central, procesador INTEL PENTIUM IV de 2 GHZ de velocidad.

El ejecutable y manual del usuario con ejemplos de utilización están integrados en el CD ROM que forma el anexo a esta tesis.

El texto del manual del usuario forma el sistema de ayuda del programa.

En los apartados 6.2. a 6.15 de este capítulo se presentan los aspectos considerados más relevantes de su desarrollo y funcionalidad.

6.2. ESTRUCTURA Y FUNCIONES PRINCIPALES.

La **figura 6.2.1.** presenta la ventana principal del sistema.

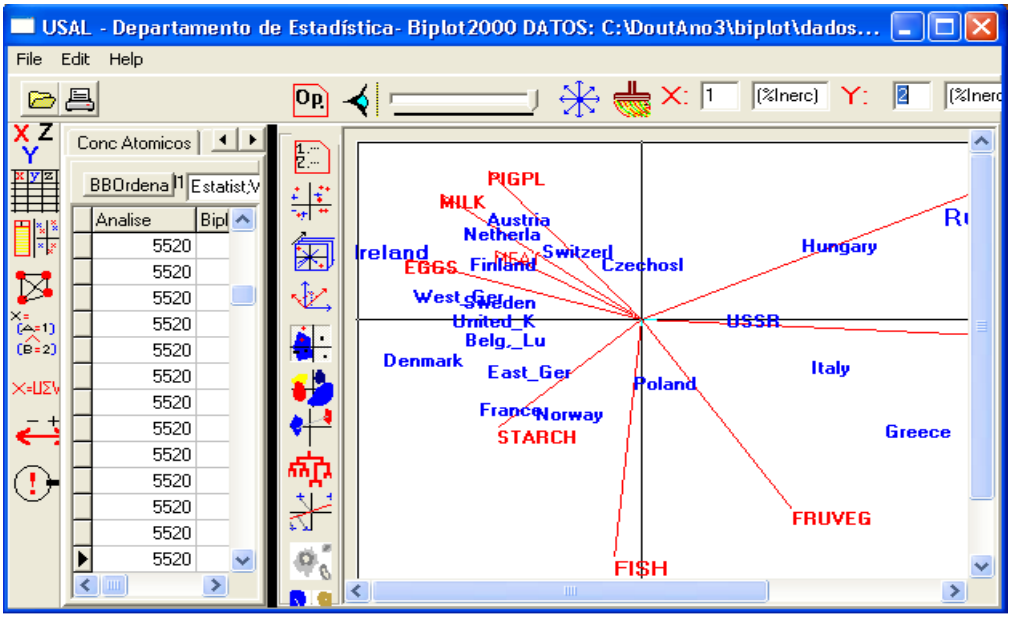


Figura 6.2.1. Ventana principal del Sistema de Minería de Datos Basado en Biplots.

En síntesis, se puede decir que el programa está pensado para cooperar con el analista en la obtención del significado de los grupos de individuos y variables que han sido descubiertos por los programas de análisis de datos

multivariantes (biplots y otros), usando las operaciones básicas de interpretación de resultados definidas en el apartado 4.2.

En coherencia con el modelo teórico desarrollado en el **capítulo IV**, se expresan los resultados obtenidos por los métodos de análisis de datos multivariantes por grupos y particiones del conjunto de individuos y de variables. Ver apartado 4.3.

Estos grupos y particiones son representados por variables cualitativas cuyos valores son los identificadores de los grupos obtenidos y cuyo significado hay que aproximar usando el significado de las variables observadas. Ver apartados 4.4. y 4.6.

El modelo de minería de datos usado para desarrollar el programa ha sido presentado en el apartado 5.2. y en la **figura 5.2.1.**

Básicamente, el ciclo de funcionamiento del programa es el siguiente:

1. **Edición de los datos**

   Creación de los Datos en Análisis (DEA) en distintos formatos: Tablas, Tablas de Contingencia, Tablas Disyuntivas Completas (TDC), Distancias entre átomos (TDEA).
2. **Análisis de los DEA.**

   Empleo de distintos métodos de análisis de datos multivariantes (biplots y análisis cluster), representando los resultados por grupos y particiones del conjunto de individuos.
3. **Interpretación de los resultados**

   Los grupos y particiones son representados como configuraciones de puntos en biplots. Interacción con los biplots mediante un modelo basado en las operaciones básicas de interpretación.

Tabla 6.2.1. Ciclo de funcionamiento.

La **figura 6.2.2.** presenta las principales estructuras de datos mencionadas en la **figura 5.2.1,** pero ahora con más detalle.

Inicialmente, los datos por analizar son extraídos de una base de datos (Datos Brutos) y son editados.

Creados los datos en análisis (DEA), sobre estos datos pueden actuar distintos métodos de análisis.

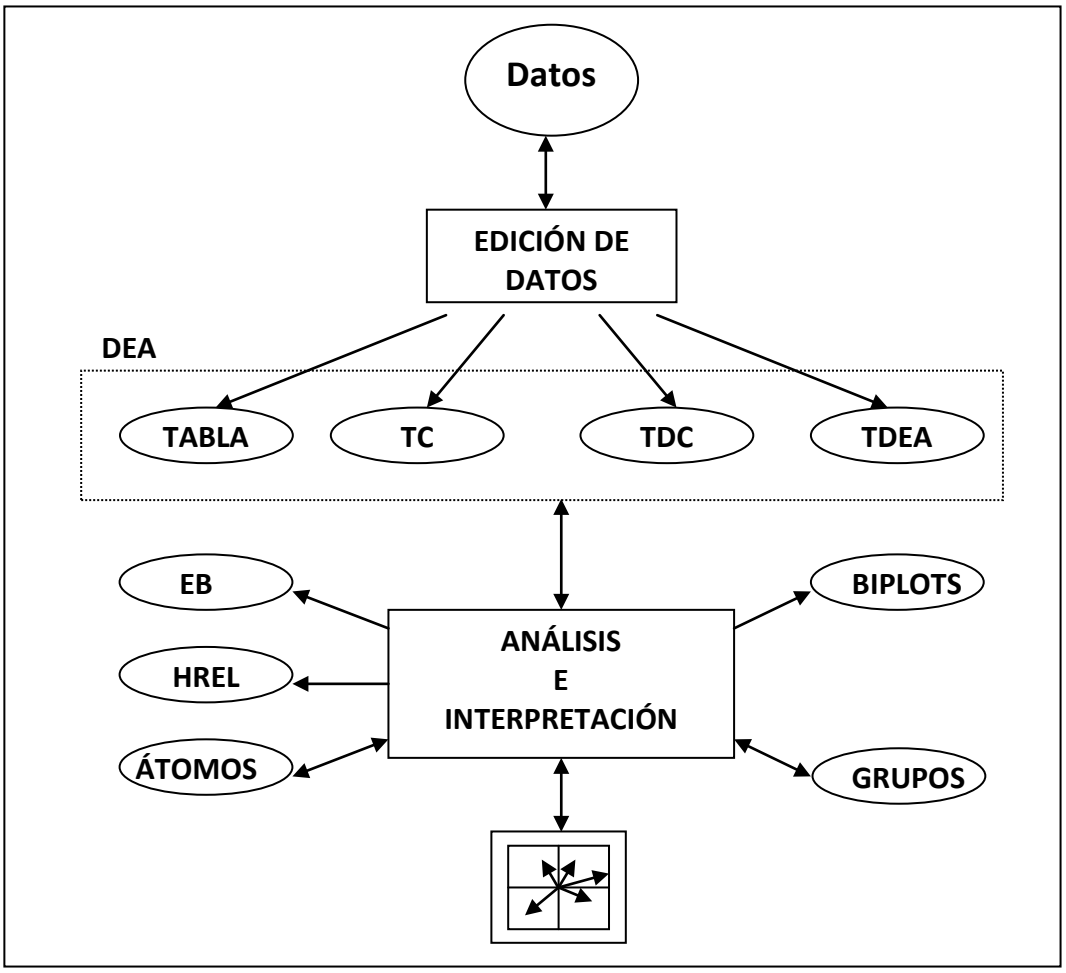


Figura 6.2.2. Esta figura completa la **figura 5.2.1.** con detalles acerca de las estructuras de datos.

En nuestro sistema experimental, más allá de los biplots, se usa - para ilustrar la intervención de otros métodos de análisis - un programa de análisis cluster que genera, automáticamente, grupos y particiones de individuos, representados sobre los biplots.

Los resultados del proceso de análisis multivariante/minería son **grupos** de **individuos**, **particiones** del conjunto de individuos y **biplots**.

Unos y otros son almacenados en estructuras de datos (bases de datos) designadas en la **figura 6.2.1.** por GRUPOS y BIPLOTS.

Los BIPLOTS contienen configuraciones de marcadores de individuos y variables en sistemas de referencia definidos por los ejes factoriales resultantes de la SVD.

La base de datos GRUPOS contiene la composición de todos los grupos resultantes de los distintos análisis, sin ninguna referencia a sistemas de coordenadas. La ubicación de estos grupos/resultados en biplots (transformación de un grupo en configuración) es determinada por la identificación de los objetos (individuos y variables) y por las coordenadas de los marcadores respectivos en biplots existentes.

Más allá de las estructuras BIPLOTS y GRUPOS, existen las estructuras EB, HRE, ATOMOS cuyos contenidos se resumen a continuación:

EB Estadísticas Básicas. Para cada variable, sobre cada grupo, están guardados los valores de las estadísticas básicas de las variables observadas. Los datos completos forman un grupo con número cero. Estas estadísticas son usadas en las síntesis de resultados.

HREL Hechos relevantes. Esta estructura guarda los hechos relevantes (ver apartado 5.5.) en la construcción de síntesis finales. Por ejemplo: el hecho de que la diferencia de medias de una variable en dos grupos sea «significativa» es relevante para comparar los grupos.

ATOMOS Contiene el resultado de la descomposición de los datos en átomos usados para la construcción del grafo de intersección. Ver apartado 4.4. Ver también, en el presente capítulo - en el apartado 6.5. - el algoritmo usado para obtener esa descomposición.

La realización de las tareas de interpretación y la interactividad exigidas al sistema implican que todas las estructuras de datos necesarias estén cargadas en la memoria central del ordenador.

Por otro lado, las operaciones de cálculo, búsqueda, ordenación y selección que es necesario realizar sobre esas estructuras son muy frecuentes y abarcan a miles de ítems.

Todo esto llevó a la decisión de implementar estas estructuras usando bases de datos residentes en la memoria central del ordenador.

Esto significa que está asociada a cada una de las estructuras mencionadas toda la funcionalidad propia de una base de datos: índices, filtros, búsqueda, ordenación, selección.

Esto crea condiciones de base - no totalmente explotadas en este prototipo - para la realización, en proyectos futuros, de sofisticadas operaciones de razonamiento e interpretación.

El sistema ha sido desarrollado en el lenguaje de programación PASCAL, usando el sistema Delphi de BORLAND 7 en un ordenador con el sistema operativo WINDOWS XP. Este sistema de programación por objetos integra un sistema de datos que permite, con relativa facilidad, crear y

modificar tablas (en tiempo de ejecución) sin necesidad de cambiar de lenguaje.

6.3. FUNCIONES PRINCIPALES DEL SISTEMA.

6.3.1. PANTALLA PRINCIPAL.

La figura siguiente presenta el esquema de la pantalla principal del programa durante una sesión de análisis de los datos de GABRIEL (1981).

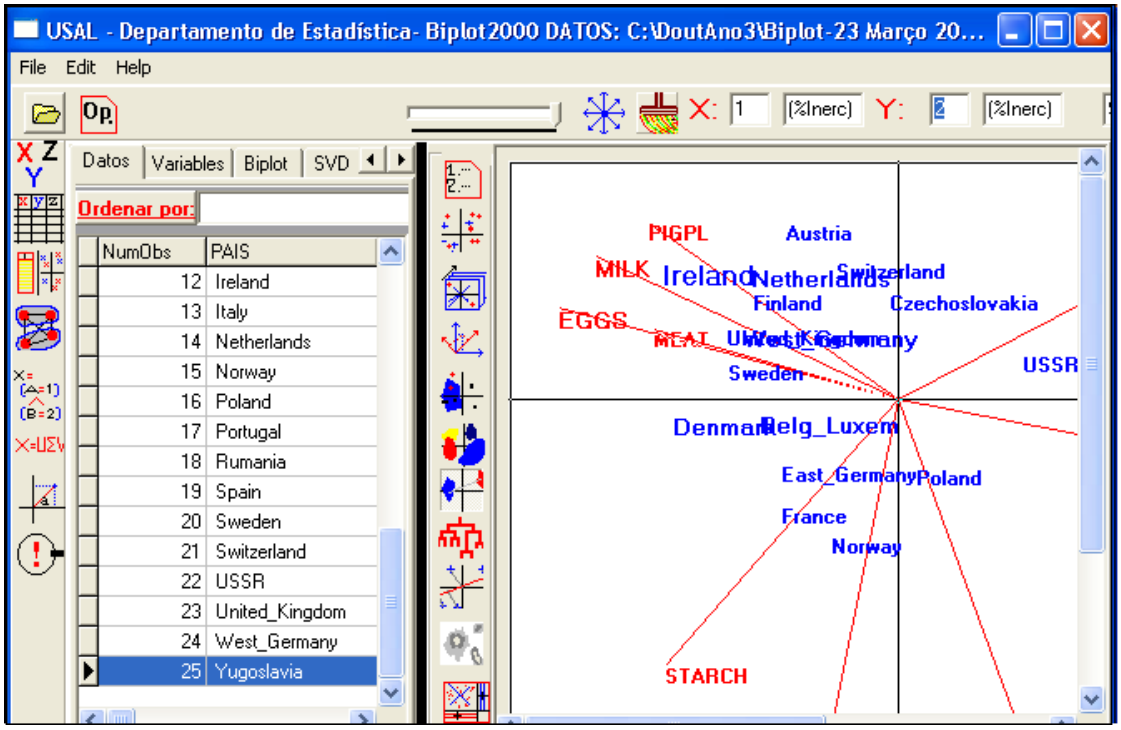


Figura 6.3.1.1. Pantalla principal del sistema.

En el esquema de la **figura 6.3.1.2.** se pueden ver las funciones que se han asociado a cada uno de los componentes de la pantalla principal.

La pantalla se divide en las regiones indicadas en la figura siguiente:

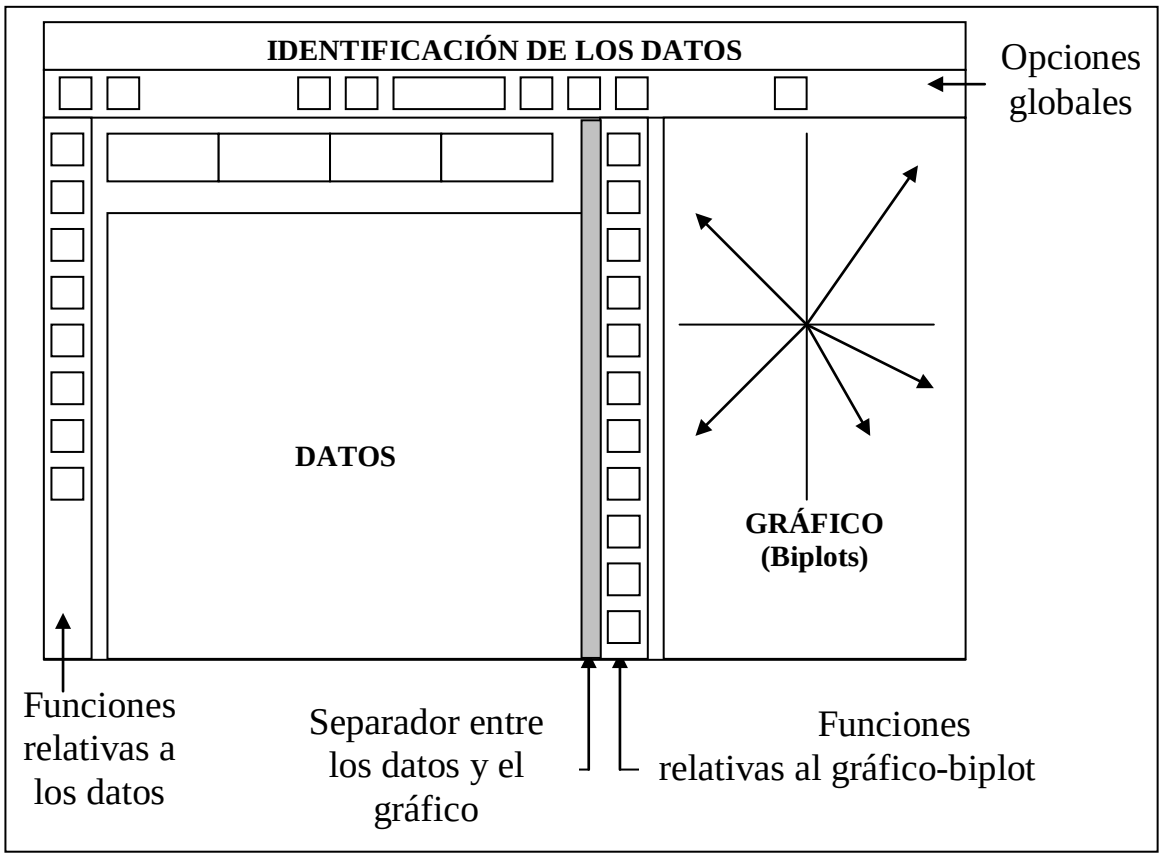


Figura 6.3.1.2. Esquema de la pantalla principal.

6.3.2. FUNCIONES DE LOS BOTONES PRINCIPALES.

A continuación se resumen las funciones de los botones que forman la pantalla principal.

 Botón para elegir y leer los datos por analizar.

 **Opciones globales**.

Permite elegir la dimensión del gráfico del biplot y otras opciones globales.

 Permite fijar el nivel de calidad de representación de los elementos (individuos de los elementos) que son pintados sobre el biplot.

Por ejemplo, si se elige 0.5, solamente se presentan individuos y variables con una calidad de representación por encima de 0.5.

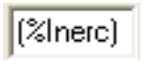
 **Dilatación**.

Pulsando sucesivamente este botón, los individuos son pintados sobre posiciones radiales sucesivamente más alejadas del centro.

 **Pintar el biplot de nuevo**, eliminando los elementos gráficos superfluos, usando las opciones definidas en el inicio de la sesión, sin recalcular el biplot.

 Estos botones permiten elegir el plano factorial actual.

 Por ejemplo, *X*= 1, *Y*= 2 representa el plan formado por los ejes factoriales (1, 2).

 **Porcentaje** de inercia: varianza o variabilidad asociada al plan elegido.

 Coordenadas de la posición actual del apuntador del ratón relativa al sistema de ejes.

 Elegir las variables y los individuos activos para el análisis.

 **Editor de datos.**

Este botón permite realizar las operaciones de limpieza, recodificación y producción de ficheros a someter al análisis

Interacción datos/biplot.

Permite ver sobre el biplot los marcadores de los individuos que han sido elegidos directamente sobre los datos y responder a cuestiones del tipo «*¿Dónde están sobre el biplot los individuos... "tal y tal"*?»

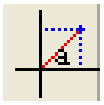
Agrupamiento de átomos con pocos individuos en átomos de soporte con cardinal más elevado.

Definición visual de sub-conjuntos de datos.

Este botón permite ver sobre el biplot los soportes de expresiones de tipo conjuntivo.

Responde a cuestiones del tipo: ¿dónde están sobre el biplot actual, los individuos tales que $(A= 1) \wedge (B= 2)$, en donde A y B son dos variables observadas?

Este botón permite ver el resultado de la descomposición de los datos en valores y vectores singulares, base del biplot actual.



Estudio de la calidad de representación de los individuos y variables en el biplot actual, ordenando variables por la contribución de los elementos para los ejes y de los ejes para los elementos.

Resumen y conclusiones.

Permite ver las principales conclusiones del estudio, antes de la impresión

Opciones gráficas de los biplots.

Permite elegir los símbolos y etiquetas a usar para pintar variables e individuos sobre los biplots generados en la sesión actual

 **Crear Biplots**.

Este botón permite elegir el tipo de biplot (GABRIEL o GALINDO) y el tipo de transformación por realizar sobre los datos antes de crear el biplot.

 **Crear familias de biplots**.

Este botón permite definir un conjunto de biplots (todos con las mismas características globales) que corresponde al conjunto de valores de una variable - parámetro.

En particular, esa variable puede corresponder a una sucesión de tiempos (años, meses, ocasiones).

 **Transformaciones geométricas: rotación y reflexión**.

El sistema permite realizar dos de las transformaciones clásicas en el plano: **rotación** y **reflexión**.

 **Definición de grupos**.

Permite que el usuario defina, visualmente, un grupo de individuos, de variables o de individuos y variables que le parezcan interesantes interesantes.

 **Comparación de dos grupos**.

Dados dos grupos definidos sobre el biplot actual este botón permite comparar esos grupos usando las variables observadas.

 **Estudio de grupos**.

Dados todos los grupos que han sido definidos a lo largo de la sesión (por el usuario o en su caso por los métodos de clasificación automática) este botón permite visualizar esos grupos **sobre el biplot actual** y recordar sus propiedades.

Permite, además, obtener, para un conjunto de grupos - una partición, por ejemplo - las reglas que caracterizan esos grupos.

 **Arbol de clasificación**.

Permite obtener clasificaciones jerárquicas de los individuos, generando particiones por análisis cluster aglomerativa.

Los grupos generados por análisis cluster pueden ser representados sobre el biplot actual tanto por sus cierres convexos como por colores, siendo interpretados como los demás grupos.

 **Gráficos de densidad**.

Genera un gráfico de densidad por el método Kernel.

 **Ver los elementos suplementarios**.

Este botón permite visualizar sobre el biplot **actual** tanto variables como los individuos que actualmente están calificados como suplementarios.

6.4. LOS DATOS.

6.4.1. ESTRUCTURA DE UN FICHEIRO DE DATOS.

El sistema asume que los datos están organizados según la estructura de los datos usados como ejemplo a lo largo de este capítulo y también en el sistema de ayuda del programa.

Estos datos se basan en los datos de GABRIEL (1981) relativos a los alimentos – origen de las proteínas consumidas en los distintos países de Europa.

Para atender a transformaciones históricas mientras tanto ocurridas y también a la necesidad de ilustrar las funciones del programa relativas a variables cualitativas, hemos añadido a esos datos originales las columnas siguientes:

SimbGr Símbolo Gráfico. Es un carácter (número, letra de otro símbolo) que representa la fila (país) en el gráfico cuando no es posible presentar el identificador de la fila.

CortHier Cortina de Hierro. Es una variable cualitativa con dos valores: *n* = NO (El país no pertenecía a la denominada Cortina de Hierro) - *s* = SI (El país pertenecía a la denominada Cortina de Hierro).

Area Area Geográfica de Europa.

n NORTE

c CENTRO – Países continentales

s SUL – Países del Sur, o mediterráneos.

Observaciones Esta última columna contiene anotaciones acerca del significado de las filas. Actualmente, el programa lee esos comentarios pero no los utiliza en el procesamiento.

Las columnas restantes de esa tabla de datos representan los nombres en inglés de los alimentos – origen de las proteínas consumidas en los distintos países.

En el cruce de la fila – país número *i* con la columna *j* del alimento está el porcentaje de proteínas que en el país *i* tendría su origen en el alimento *j*.

Los países – filas – son identificados en el programa por un índice *i*= 1, 2, …, 25, en que

i= 1 corresponde a Albania y

i= 25 corresponde a Yugoslavia.

Las variables están numeradas por *j*= 1, 2,..., 11, en que

j= 1 corresponde a CORTHIERR y

j= 11 corresponde a FRUVE (Frutas y Vegetales)

Las columnas SimbGraf y Observaciones no son tratadas como variables por el programa.

Pais	Simb Graf	Cort Hierr	AREA	MEAT	PIGPL	EGGS	MILK	FISH	CEREAL	STARCH	NUTS	FRUVEG	Observaciones
Albania	n	s	c	10,1	1,4	0,5	8,9	0,2	42,3	0,6	5,5	1,7	No-CEE
Austria	n	n	c	8,9	14	4,3	19,9	2,1	28	3,6	1,3.	4,3	CEE
Belg_Luxem	n	n	n	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4	CEE
Bulgaria	n	s	c	7,8	6	1,6	8,3	1,2	56,7	1,1	3,7	4,2	No-CEE
Czechoslovakia	n	s	c	9,7	11,4	2,8	12,5	2	34,3	5	1,1	4	No-CEE
Denmark	n	n	n	10,6	10,8	3,7	25	9,9	21,9	4,8	0,7	2,4	CEE
East_Germany	n	s	c	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6	No-CEE
Finland	n	n	n	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1	1,4	CEE
France	n	n	s	18	9,9	3,3	19,5	5,7	28,1	4,8	2,4	6,5	CEE
Greece	n	n	s	10,2	3	2,8	17,6	5,9	41,7	2,2	7,8	6,5	CEE
Hungary	n	s	c	5,3	12,4	2,9	9,7	0,3	40,1	4	5,4	4,2	No-CEE
Ireland	n	n	n	13,9	10	4,7	25,8	2,2	24	6,2	1,6	2,9	CEE
Italy	n	n	s	9	5,1	2,9	13,7	3,4	36,8	2,1	4,3	6,7	CEE
Netherlands	n	n	n	9,5	13,6	3,6	23,4	2,5	22,4	4,2	1,8	3,7	CEE
Norway	n	n	n	9,4	4,7	2,7	23,3	9,7	23	4,6	1,6	2,7	CEE
Poland	n	s	c	6,9	10,2	2,7	19,3	3	36,1	5,9	2	6,6	No-CEE
Portugal	n	n	s	6,2	3,7	1,1	4,9	14,2	27	5,9	4,7	7,9	CEE
Rumania	n	s	c	6,2	6,3	1,5	11,1	1	49,6	3,1	5,3	2,8	No-CEE
Spain	n	n	s	7,1	3,4	3,1	8,6	7	29,2	5,7	5,9	7,2	CEE
Sweden	n	n	n	9,9	7,8	3,5	24,7	7,5	19,5	3,7	1,4	2	CEE
Switzerland	n	n	s	13,1	10,1	3,1	23,8	2,3	25,6	2,8	2,4	4,9	No-CEE
USSR	n	s	c	9,3	4,6	2,1	16,6	3	43,6	6,4	3,4	2,9	No-CEE
United_Kingdom	n	n	n	17,4	5,7	4,7	20,6	4,3	24,3	4,7	3,4	3,3	CEE
West_Germany	n	n	n	11,4	12,5	4,1	18,8	3,4	18,6	5,2	1,5	3,8	CEE
Yugoslavia	n	s	s	4,4	5	1,2	9,5	0,6	55,9	3	5,7	3,2	No-CEE

Tabla 5.4.1.1. Estrutura de los conjuntos de datos.

6.4.2. TIPOS FICHEIRO DE DATOS.

En la versión actual, el programa espera que los datos estén organizados en tablas de tipo «DBF», con extensión .DBF (sistema WINDOWS), con la estructura ilustrada en el apartado 6.4.1.

Al grabar un fichero, el sistema graba también en ese formato.

En la **figura 6.4.2.1.** se ilustra el momento de grabación de un fichero, usando el tipo .DBF.

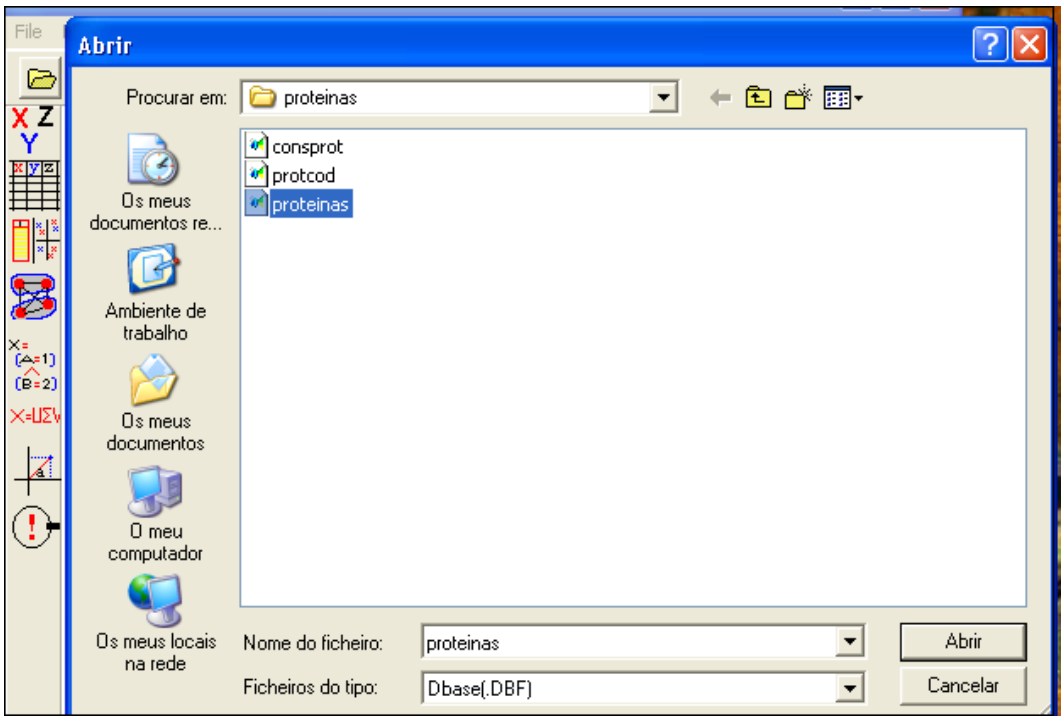


Figura 6.4.2.1. En la carpeta Proteínas existen los ficheros ConsProt.DBF, ProtCod.DBF, Proteinas.DBF.

Los datos son creados con una hoja de cálculo - usando la estructura ilustrada - y seguidamnete **exportada** en formato .DBF. Además de las columnas para las variables debe reservarse la ultima columna para observaciones.

La **figura 6.4.2.2.** ilustra el resultado de la apertura del fichero Proteínas .DBF, pudiendo verse parte de su contenido.

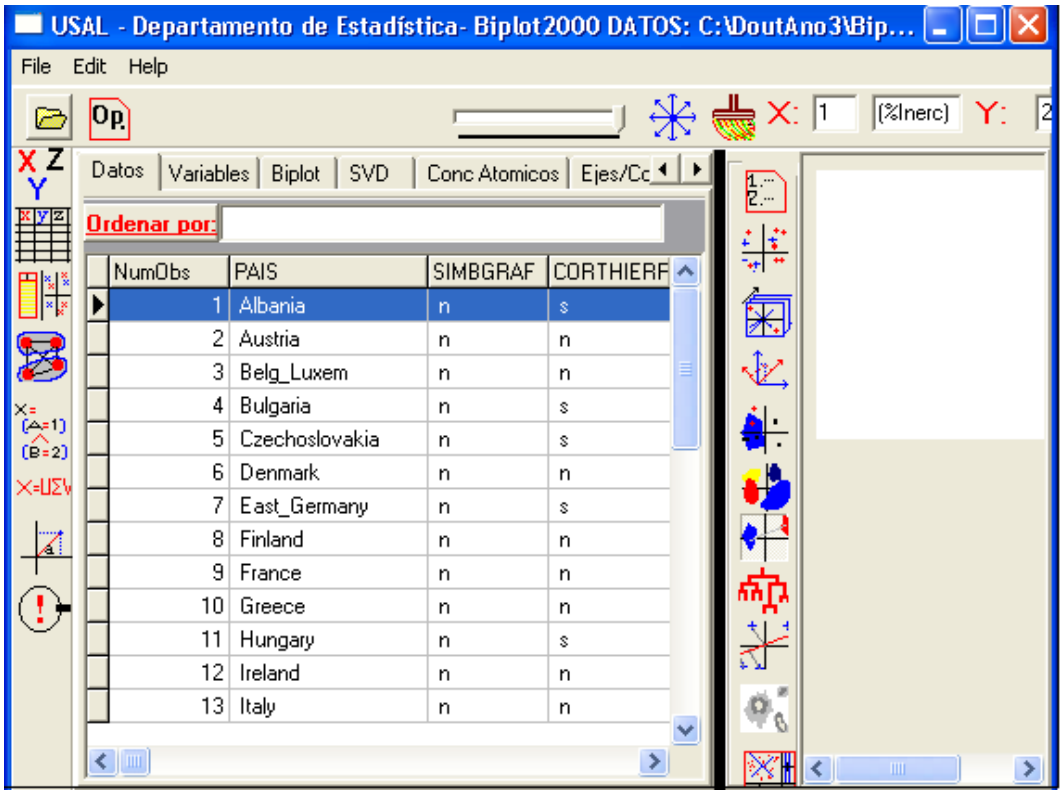


Figura 6.4.4.2. Contenido del fichero Proteínas.

6.5. EDICIÓN DE LOS DATOS.

6.5.1. EL EDITOR DE DATOS.

El editor de datos, cuya ventana principal puede verse en la **figura 6.5.1.1.**, realiza las funciones siguientes:

- Inspección de los datos de un fichero para conocer su contenido.
- Modificación de los valores almacenados para corregir esos valores o experimentar otros.
- Modificación del texto que describe el significado de las variables y sus valores.

- Inspección de la composición de los átomos formados con los valores de las variables.
- Recodificación de las variables.
- Eliminación de registros.
- Eliminación de variables.
- Creación de tablas de contingencia por concatenación de variables.
- Crear los datos por analizar.
- Creación de datos en la forma disyuntiva completa (TDC).
- Creación de tablas con afinidades entre átomos.
- Guardar los datos creados por el editor.

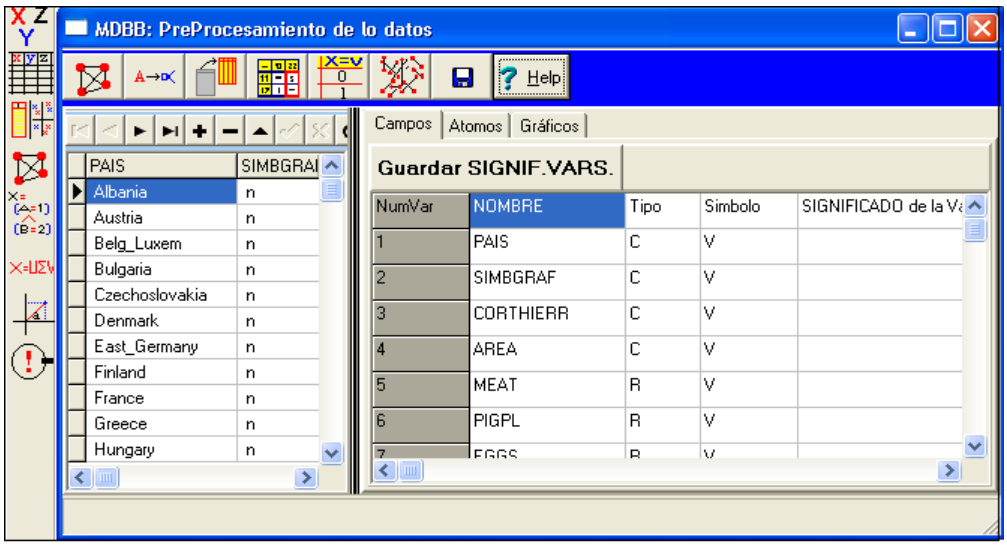


Figura 6.5.1.1. Ventana principal del editor de los datos

Todas estas funciones son implementadas por algoritmos que funcionan sobre el grafo de intersección formado por los átomos, en los que se descomponen los datos brutos. Ver **capitulo IV**, apartado 4.4.1.

Por eso, se presenta aquí el algoritmo que realiza esa descomposición.

El grafo de intersección tiene por vértices los conjuntos de individuos que son las extensiones de los átomos, expresiones de tipo ($X = v$). Ver apartado 4.4.

En ese grafo, existe un arco entre dos de estos vértices, V_1 y V_2, cuando $V_1 \cap V_2 \neq \varnothing$.

El grafo de intersección es implementado por dos estructuras.

VÉRTICES contiene los vértices
ARCOS contiene los arcos.

La estructura VÉRTICES es una matriz (o lista dinámica) cuyas filas tienen la estructura siguiente:

V Número del vértice
Var Número de la variable
Val Valor de la variable
Inds Lista de las observaciones para las que (Variable = Valor).

La estructura ARCOS es una matriz (o lista) cuyas filas tienen la estructura siguiente:

V_1 Número del vértice 1
V_2 Número del vértice 2
$V_1 \cap V_2$ Intersección de la lista de los individuos del vértice 1 con la lista de individuos con el vértice 2.

En principio no sería necesario guardar $V_1 \cap V_2$, una vez que se puede obtener fácilmente por cálculo a partir de VÉRTICES, conociendo V_1 y V_2.

La decisión de guardar o no $V_1 \cap V_2$ en la estructura ARCOS depende de consideraciones operacionales (tiempo y espacio) al implementar un sistema específico.

Ejemplo 6.5.5.1.

Para ilustrar estos conceptos consideremos los «datos»

	VARIABLES		
IND	*X*	*Y*	*Z*
1	*a*	1	5
2	*b*	2	4
3	*a*	1	5
4	*a*	1	5

Las estructuras finales VÉRTICES y ARCOS (al final del proceso secuencial que empieza en el individuo 1 y termina en el individuo 4) son las siguientes:

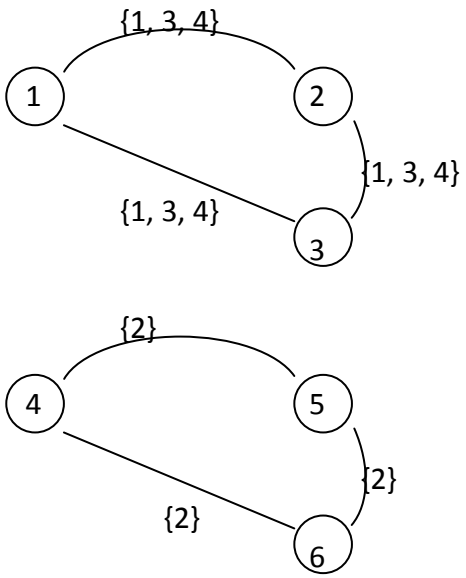
VÉRTICES

V	***VAR***	***VAL***	***INDS***
1	*X*	*a*	{1, 3, 4}
2	*Y*	1	{1, 3, 4}
3	*Z*	5	{1, 3, 4}
4	*X*	*b*	{2}
5	*Y*	2	{2}
6	*Z*	4	{2}

ARCOS

V_1	V_2	$V_1 \cap V_2$
1	2	{1, 3, 4}
1	3	{1, 3}
2	3	{1, 3, 4}
4	5	{2}
4	6	{2}
5	6	{2}

El grafo de intersección es:



En este caso particular, existen dos conceptos importantes expresados por tres átomos:

$(X= 1) \wedge (Y= 1) \wedge (Z= 5) = \{1, 3, 4\}$

$(X= b) \wedge (Y= 2) \wedge (Z= 4) = \{2\}$

Usando las estructuras indicadas, el algoritmo siguiente realiza la extracción de los átomos.

En la **figura 6.5.1.2.** se presenta el algoritmo que realiza, secuencialmente, la descomposición de un conjunto de datos en átomos de la forma $(X= x) = \{i: X(i) = x\}$.

Estos átomos son los vértices del grado de intersección en el que se basa todo el proceso de edición de datos e interpretación de resultados.

Algoritmo 6.6.1. Descomposicón de los Datos en àtomos o Vértices del Grafo de Intersección.

```
nv = 0              {Número de vértices}
Repetir para i variando de 1 a n  {individuos}
  Repetir para j variando de 1 a p {variables}
Inicio
                    Valor = Datos [i, j];
Encontrado =  Existe (Vértices, j, Valor, Pos)
Si (Encontrado) entonces
   Actualizar (Vértices, j, Pos, i, j, Valor)
                    Sino
                      Inicio
                nv= nv + 1;
                        Crear (Vértice);
      Añadir (Vértice);
                      Fin;
         Fin;
Fin.
```

Figura 6.5.1.2. Algoritmo de extracción de átomos.

El significado de las expresiones **Existe**, **Actualizar**, **Crear**, **Añadir** es el siguiente:

Existe (Vértices, *j*, Valor, Pos).

Representa una función booleana que busca en los Vértices un átomo (vértice) en la que la variable *j* tenga el valor «valor». Si existe, registra su posición en Pos y toma el valor Verdadero; si no existe, Pos está indefinido y la función toma el valor Falso.

Actualiza (*Vértice*, *Pos*, *i*, *j*, *Valor*).

Es un procedimiento que actualiza un vértice existente en la posición *Pos* y con el *Valor* de la variable *j* en el individuo *i*. Básicamente, añade *i* a la extensión de (X_j = Valor).

Crear (Vértice).

Crea una estructura para contener los valores de una fila de la matriz vértice.

Añadir (Vértices, Vértice).

Añade a la estructura Vértices el nuevo Vértice.

Aplicando este algoritmo a los «datos» del **ejemplo 6.5.1.1.**, la evolución de la estructura VÉRTICES es:

(i= 1) → (X = a), {1}
→ (Y = 1), {1}
→ (Z = 5), {1}

(i= 2) → (X = a), {1}
→ (Y = 1), {1}
→ (Z = 5), {1}
→ (X = b), {2}
→ (Y = 2), {2}
→ (Z = 4), {2}

(i= 3) → (X = a), {1, 3}
→ (Y = 1), {1, 3}
→ (Z = 5), {1, 3}
→ (X = b), {2}
→ (Y = 2), {2}
→ (Z = 4), {2}

(i= 4) → (X = a), {1, 3, 4}
→ (Y = 1), {1, 3, 4}
→ (Z = 5), {1, 3, 4}
→ (X = b), {2}
→ (Y = 2), {2}
→ (Z = 4), {2}

El algoritmo anterior es desencadenado por el botón  del editor.

Una vez que el funcionamiento del editor depende tan profundamente del grafo de intersección, conviene que sea esta la primera operación que se realice siempre que se utiliza el editor.

6.5.2. EXAMEN INICIAL DE LOS DATOS.

El editor permite un examen inicial de los datos brutos para familiarizar el usuario con su contenido y determinar las necesidades de recodificación, las variables a eliminar, las observaciones que deben ser eliminadas, que valores fueran observados y con que frecuencia ocurren.

El EDITOR permite que el usuario realice las operaciones siguientes:

1. Inspección visual de los datos.

   Como se puede ver en la **figura 6.5.2.1.**, en la ventana del lado izquierdo, se presentan los datos brutos, el número *n* de observaciones y el número *p* de variables. Ver **figura 6.5.2.1.**, lado izquierdo.
2. Introducción del significado de las variables. Ver **figura 6.5.2.1.**
3. Inspección de los átomos.

   El usuario puede ver cuales son los valores que han sido observados para cada variable, incluyendo valores faltantes y espacios vacíos. Esta inspección permite, después, decidir que hacer con los registros en donde ocurren. Ver **figura 6.5.2.2.**, lado derecho.
4. Para cada variable, ver un gráfico de barras de los valores observados. Ver **figura 6.5.2.3.**

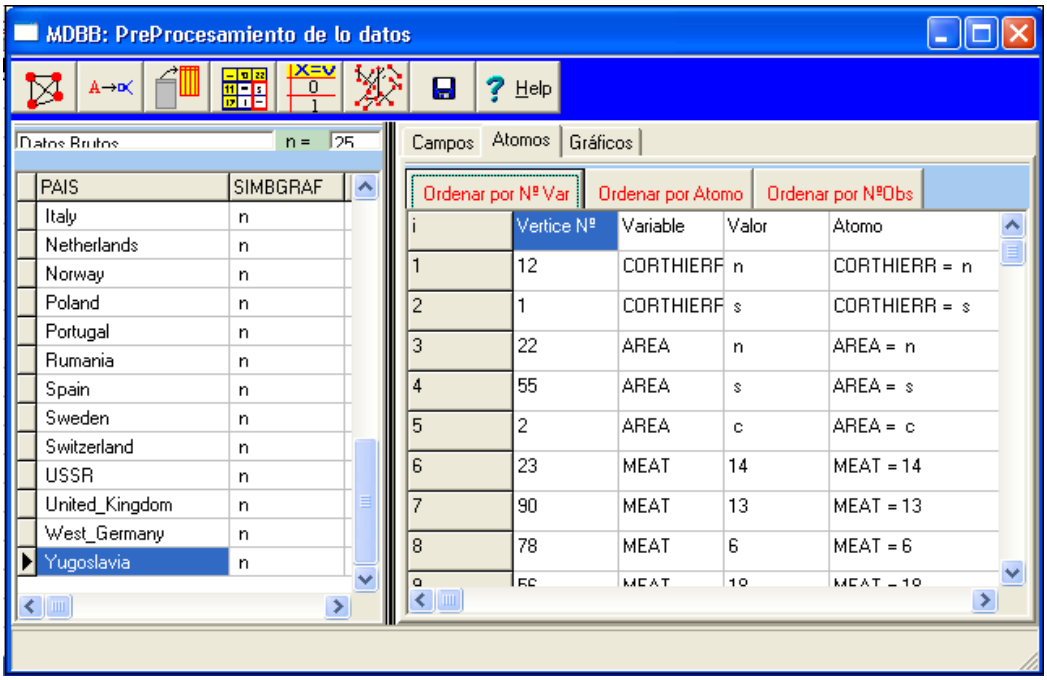


Figura 6.5.2.1. Introducción del significado de las variables.

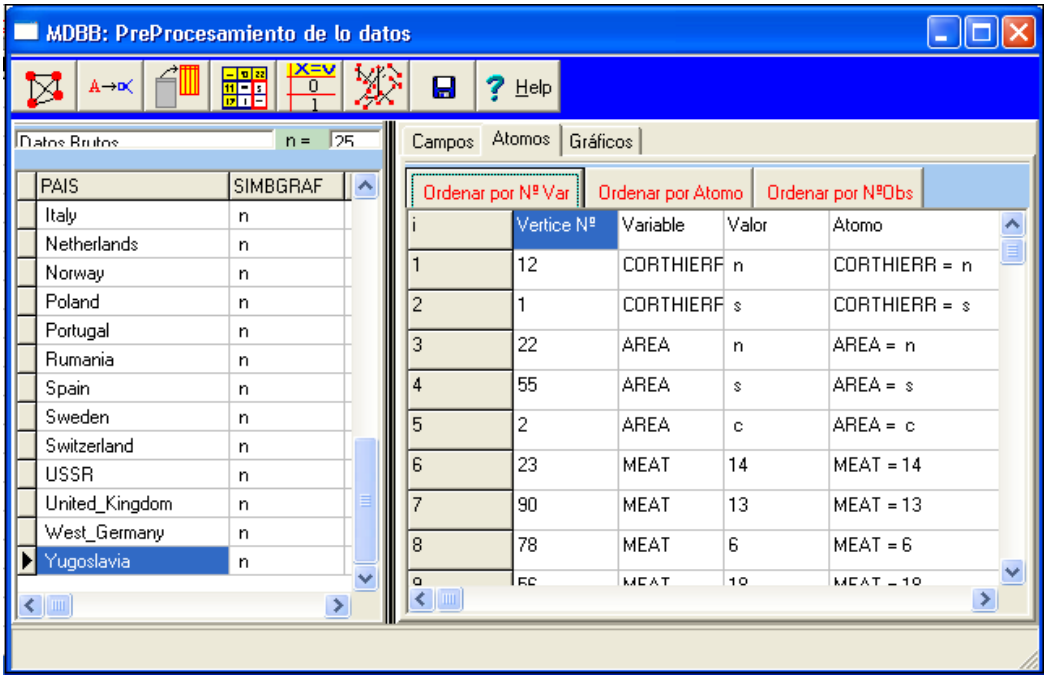


Figura 6.5.2.2. El átomo (MEAT = 14), forma el vértice nº 6 del gráfico de intersección y tiene un soporte formado por 6 países.

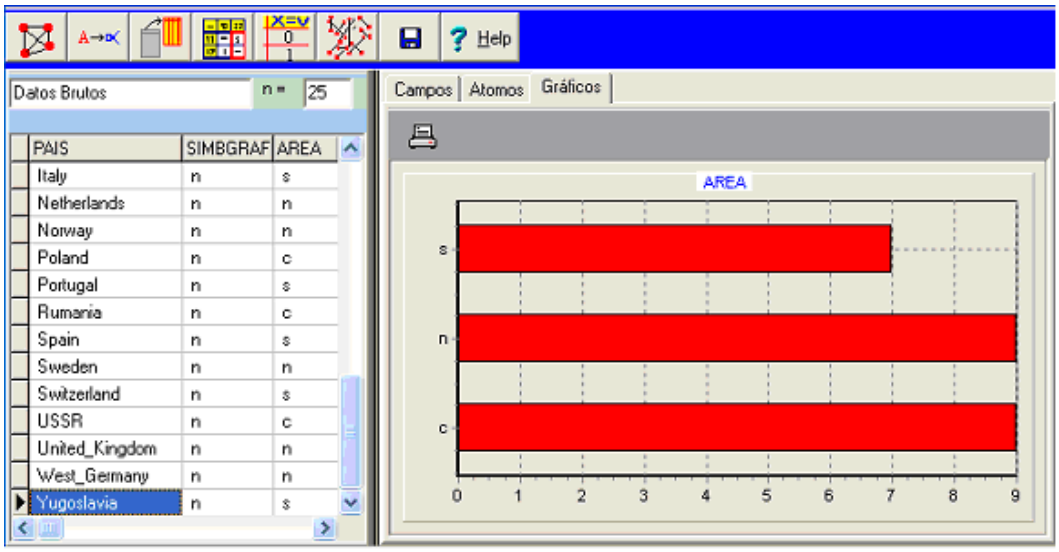


Figura 6.5.2.3. Gráfico de barras de los valores observado de la variable AREA.

6.5.3. RECODIFICIÓN Y LIMPIEZA DE DATOS.

La inspección de los datos - ver apartado 6.5.2. - permite identificar las necesidades de:

1. Recodificar variables.

   Puede ser conveniente agrupar los valores de una variable en nuevas categorías.

   Puede ser necesario substituir valores faltantes por otros que el usuario indique.
2. Eliminar registros que correspondan a conjuntos de valores identificados por el usuario.

   Por ejemplo, a veces es necesario eliminar todos los registros con espacios vacíos en algunas variables.
3. Eliminar variables. A veces ocurren "variables" que de hecho no lo son: presentan el mismo valor para todos los individuos y, por eso, no aportan información útil.

En otros casos las variables tienen tantos valores faltantes que, conservarlas, condicionaría todo el proceso de análisis.

Todas estas operaciones son realizadas por el sistema en una única ventana llamada, dentro del editor, por el botón  .

Estas operaciones se basan en algoritmos que funcionan sobre el grafo de intersección de los átomos.

Este grafo de intersección es actualizado siempre que se verifica una operación de **recodificación**, eliminación de registros o eliminación de variables.

El modo de realizar esas operaciones puede verse en el sistema de ayuda del programa.

La secuencia de figuras que enseguida se describe, ilustra esas funciones.

En las **figuras 6.5.3.1.** y **6.5.3.2.** se ilustra el procedimiento que consiste en agrupar los valores {4, 5, 6, 7, 8, 9} de la variable MEAT en el valor 1, y en agrupar los valores {10, 11, 13, 14, 17, 18} en el nuevo valor 2.

Por supuesto, los átomos correspondientes a los valores anteriores son eliminados y creados los átomos correspondientes a los valores nuevos.

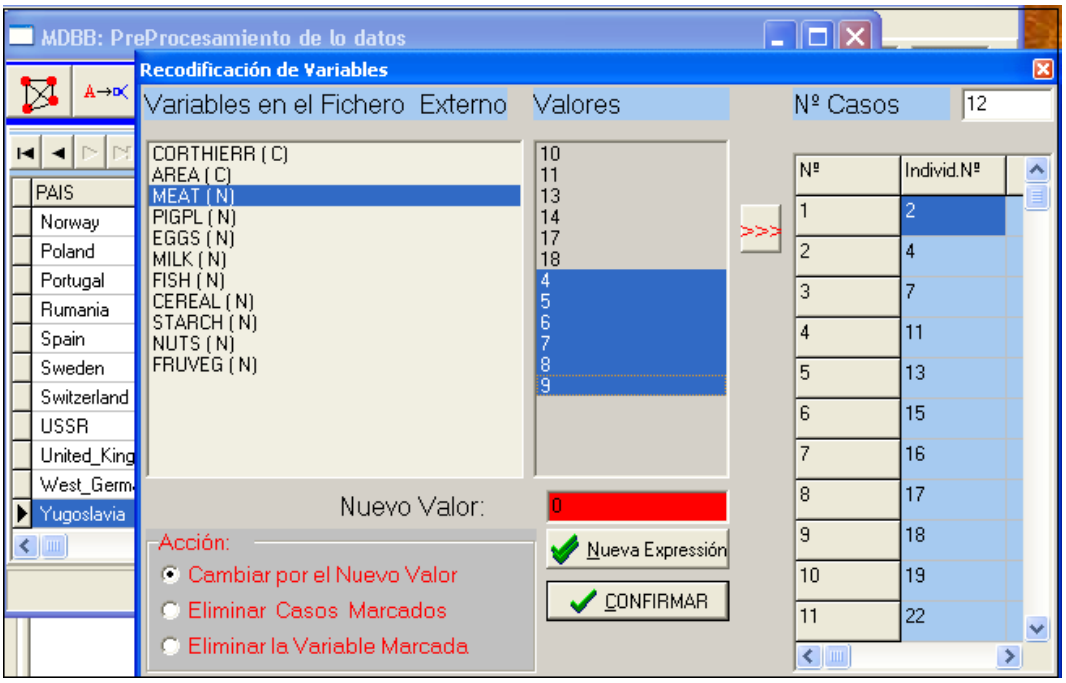


Figura 6.5.3.1. La variable MEAT tiene los valores indicados. Los valores 4, 5, 6, 7, 8, 9 han sido marcados. Van a ser reemplazados por el valor 0 + 1= 1. Cada conjunto de valores marcados es reemplazado por NUEVO VALOR +1.

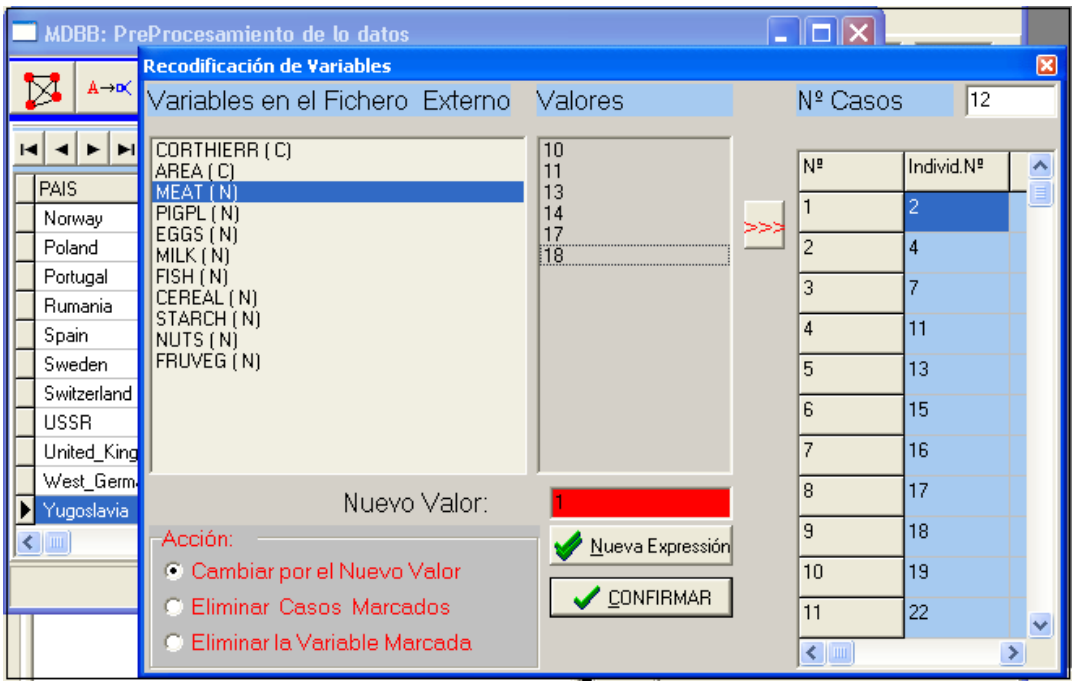


Figura 6.5.3.2. Los valores marcados han sido agrupados en el nuevo valor 1. Las observaciones transformadas han sido 12 y sus identidades están en la lista a la derecha.

La **figura 6.5.3.3.** presenta un ejemplo en que se procede a la eliminación de todos los registros en que la variable NUTS $\in \{1, 2, 3, 4\}$.

El grafo de intersección deja de incluir los átomos

(Nuts = 1), (Nuts = 2), (Nuts = 3), (Nuts = 4).

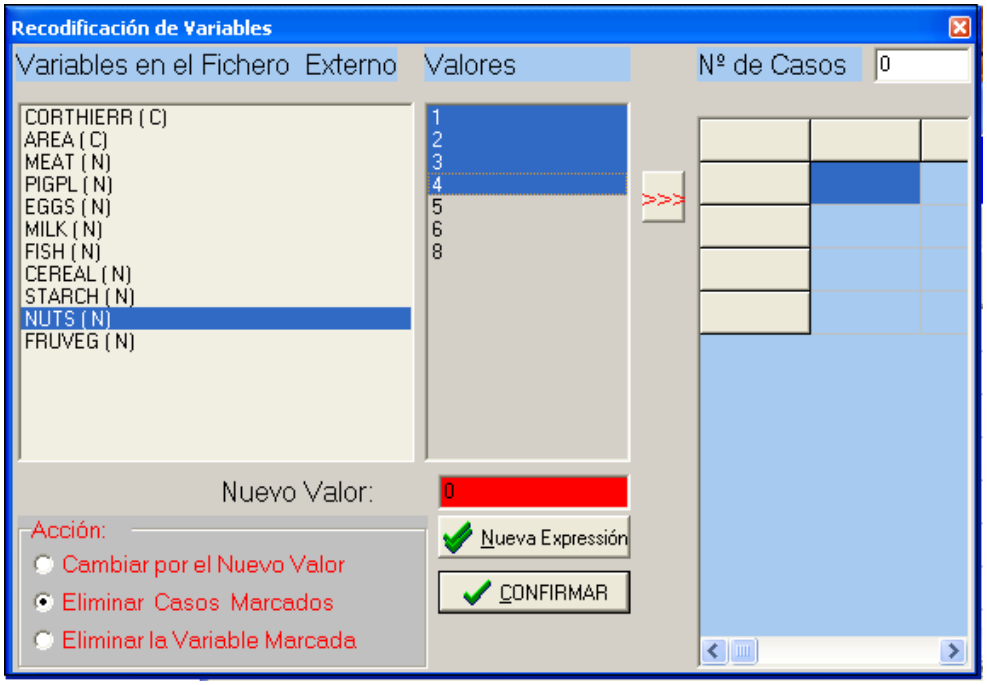


Figura 6.5.3.3. Se van a eliminar los registros para los cuales NUTS= 1, 2, 3, 4.

En la **figura 6.5.3.4.** se observa la realización de esa operación, verificándose que el número de observaciones eliminadas es 18 y pudiendo verse también sus identificadores.

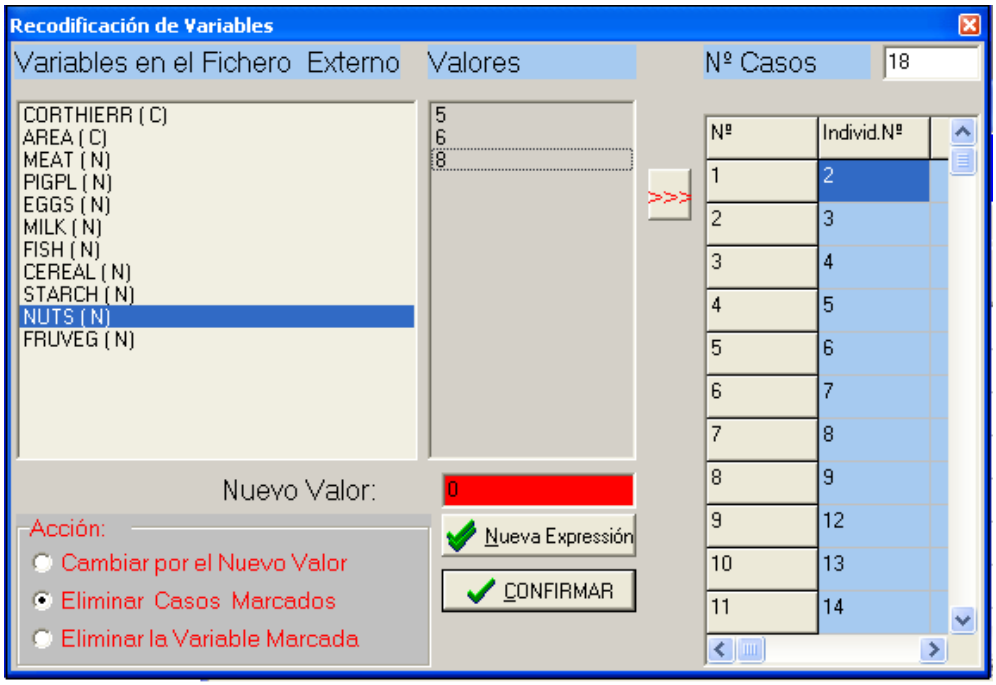


Figura 6.5.3.4. Apretando el botón CONFIRMAR las 18 observaciones para los cuales NUTS ∈ {1, 2, 3, 4} han sido marcados son eliminadas.

En la **figura 6.5.3.5.** puede examinarse el proceso de eliminación de las 3 variables MEAT, MILK, NUTS. Para eso hay que marcar las variables por eliminar y CONFIRMAR. Los átomos de las variables marcadas son eliminados del grafo de intersección.

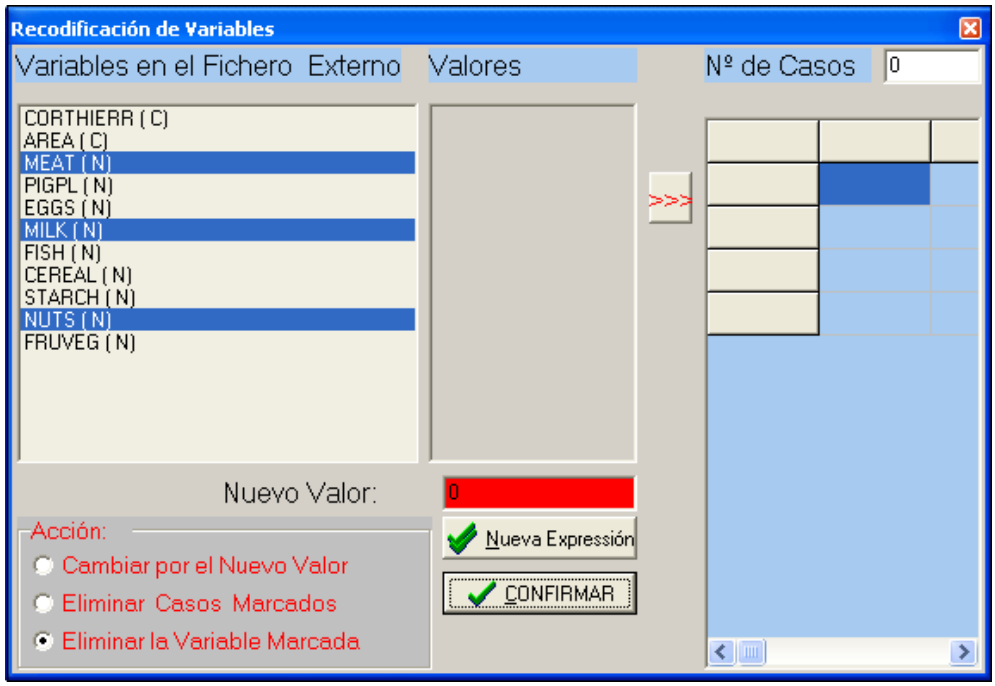


Figura 6.5.3.5. Al apretar el botón CONFIRMAR serán excluidas las columnas MEAT, MILK, NUTS.

6.5.4. CREAR E GRABAR LOS DATOS POR ANALIZAR.

 Cuando se abre un fichero, ese fichero es inmediatamente considerado el fichero actual por analizar.

Puede suceder que ese fichero no esté listo para análisis, por necesitar operaciones de limpieza como las realizadas por el editor de datos. En esas condiciones, realizar el análisis produce errores.

Cuando terminan las operaciones de limpieza y recodificación es necesario comunicar al sistema que los datos por analizar son los que han resultado del proceso de edición. Eso se realiza apretando el botón  .

Por ejemplo: después de eliminar las variables MEAT, MILK, NUTS (ver **6.5.3.**), al apretar el botón  aparece la **figura 6.5.4.1.**

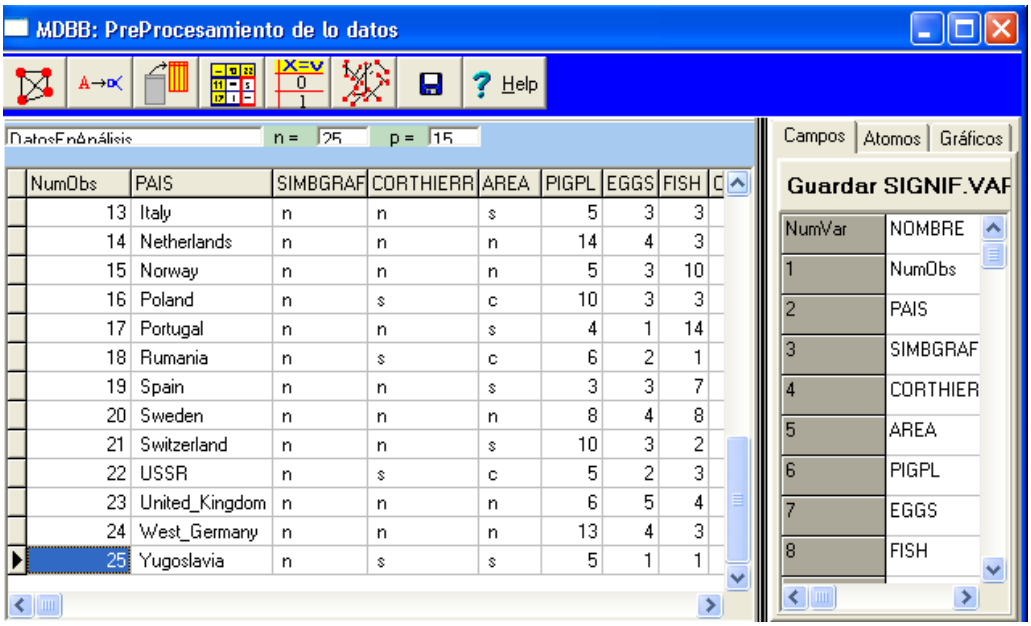


Figura 6.5.4.1. El nuevo fichero por analizar no contiene las columnas antes eliminadas por el editor.

Examinando la **figura 6.5.4.1.** se puede verificar que las columnas MEAT, MILK, NUTS han desaparecido. Si ahora creamos el biplot, estas variables no son consideradas.

Conviene guardar en un nuevo fichero de datos el resultado obtenido por el largo proceso de edición.

Eso se realiza mediante el botón  del editor: el sistema guarda en el fichero indicado por el usuario el contenido del fichero por analizar.

6.5.5. CREACIÓN DE TABLAS DE CONTINGENCIA.

 Dada la importancia de las tablas de contingencia en los problemas de diagnóstico de modelos loglineales, se permite crear y estudiar con el sistema tablas de contingencia.

Las tablas creadas pueden ser guardadas como ficheros de datos que posteriormente pueden ser tratados por el sistema como datos a estudiar por los métodos biplot.

Las tablas pueden ser creadas indicando la variable cuyos valores forman las categorías de las filas y la variable cuyas categorías forman las columnas.

Las categorías de las filas y de las columnas pueden también resultar de la concatenación de las variables que el usuario indique.

El algoritmo de creación de estas tablas de contingencia funciona sobre el grafo de intersección usando la interpretación de las tablas presentadas en el **capítulo IV**, apartado 4.4.1.

En las **figuras 6.5.5.1.** y **6.5.5.2.** se ver puede el proceso de elección y variables de frecuencias absolutas que resulta del proceso.

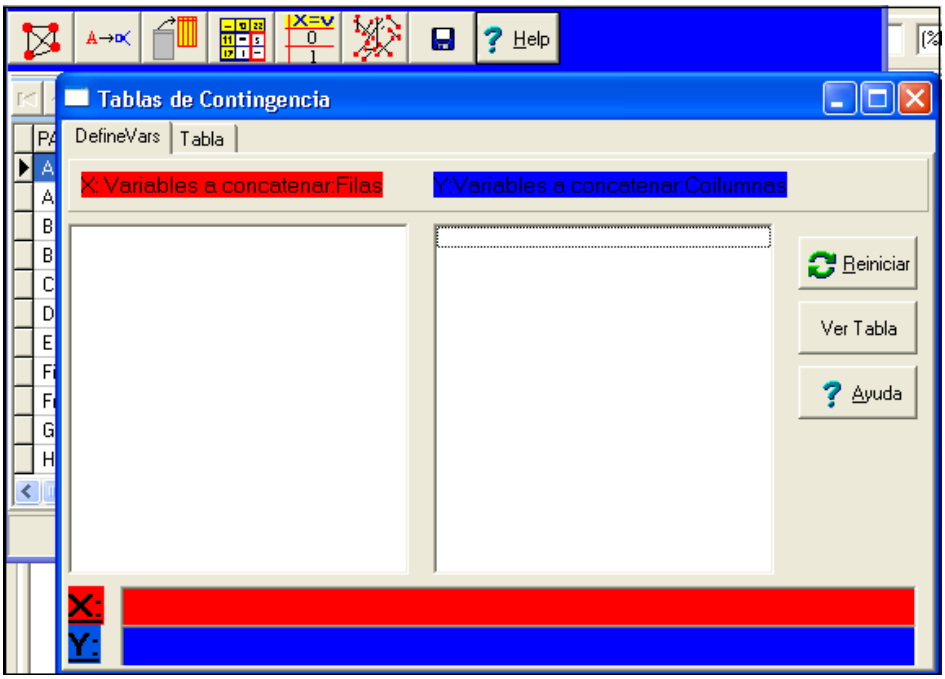


Figura 6.5.5.1. Ventana para definir las variables de una tabla de contingencia de dos vías.

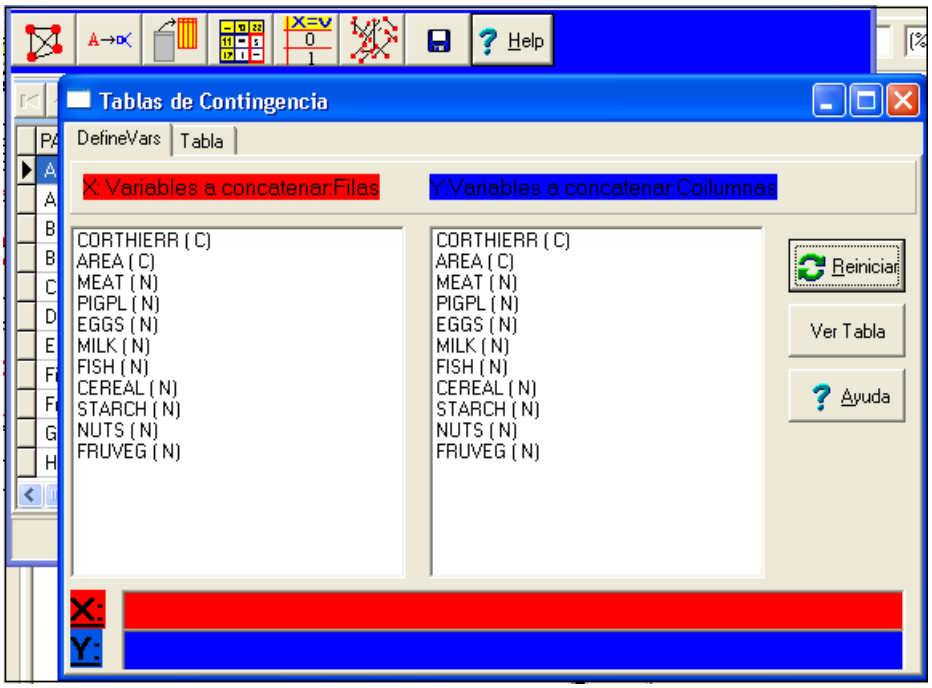
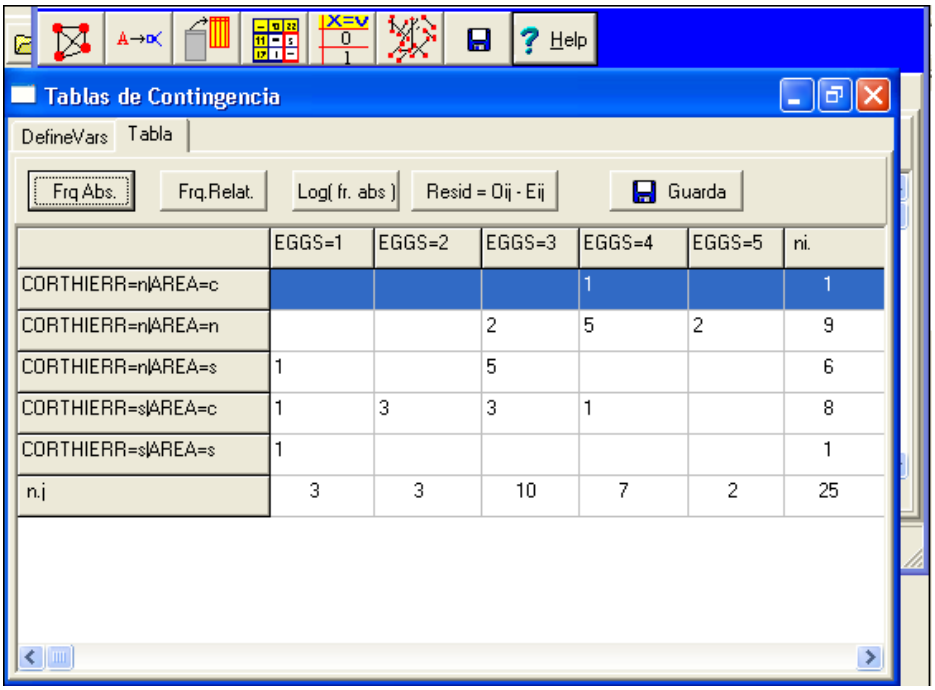


Figura 6.5.5.2. Las filas y las columnas de la tabla resultan de concatenar uno o más variables.



	EGGS=1	EGGS=2	EGGS=3	EGGS=4	EGGS=5	ni.
CORTHIERR=n|AREA=c				1		1
CORTHIERR=n|AREA=n			2	5	2	9
CORTHIERR=n|AREA=s	1		5			6
CORTHIERR=s|AREA=c	1	3	3	1		8
CORTHIERR=s|AREA=s	1					1
n.j	3	3	10	7	2	25

Figura 6.5.5.3. Las filas (*X*) resultan de concatenar CORTHIERR con AREA. Las columnas (*Y*) corresponden a los valores de EGGS.

En la **figura 6.5.5.3.**, obsérvese que las categorías de las filas (variable *X*) resultan de las categorías de las variables CORT HIERR $\in \{n, s\}$ de ÁREA $\in \{n, c, s\}$. Las categorías son, por eso: $\{nn, nc, ns, sn, sc, ss\}$.

Por ejemplo, la categoría *nc* está identificada por

(CORT HIERR= *n*) | (ÁREA= *c*).

Inicialmente, la tabla presenta las frecuencias absolutas (conjuntas y marginales). Existen botones que permiten realizar las operaciones más frecuentes sobre esas tablas (cálculo de frecuencias relativas, logaritmos de las frecuencias, residuos en relación a la hipótesis de independencia, contraste de chi-cuadrado).

Los logaritmos de las frecuencias absolutas tienen especial interés una vez que el análisis biplot de la tabla resultante permite identificar patrones lineales correspondientes a modelos log-lineales. Ver BLÁZQUEZ-ZABALLOS (1998).

En el cálculo de los logaritmos de frecuencias, la regla implementada para el caso de frecuencias conjuntas nulas es la de substituir una frecuencia nula por 0.6931= ln (1/2).

Después de examinar y realizar el estudio preliminar habitual con las tablas de contingencia, si esta va a ser objeto de un estudio por biplot - para identificar, eventualmente, algún modelo log-lineal o detectar asociaciones entre filas y columnas - entonces debe ser grabada en disco como un fichero a estudiar.

El algoritmo de grabado implementa las reglas siguientes:

1. Al nombre indicado por el usuario es añadido el prefijo «TC-» que significa que el fichero contiene una tabla de contingencia.

En esta versión experimental del sistema, esta información no es usada en el análisis de una tabla de contingencia. Es el usuario quien debe saber que, para las tablas de contingencia, algunas operaciones tienen sentido y otras no.

2. Cuando el nombre de la fila o de la columna resultante de la concatenación excede de 15 caracteres, el sistema designa automáticamente las filas por F_1, F_2, ..., F_I y las columnas por C_1, C_2, ..., C_J en donde I y J son los números de categorías en fila y columna.

Por ejemplo, si indicamos como nombre del archivo Tabla1, el sistema crea el fichero Tc-Tabla1.DBF.

Ver en la **figura 6.5.5.4.** el resultado de abrir el fichero TC-Tabla1.DBF.

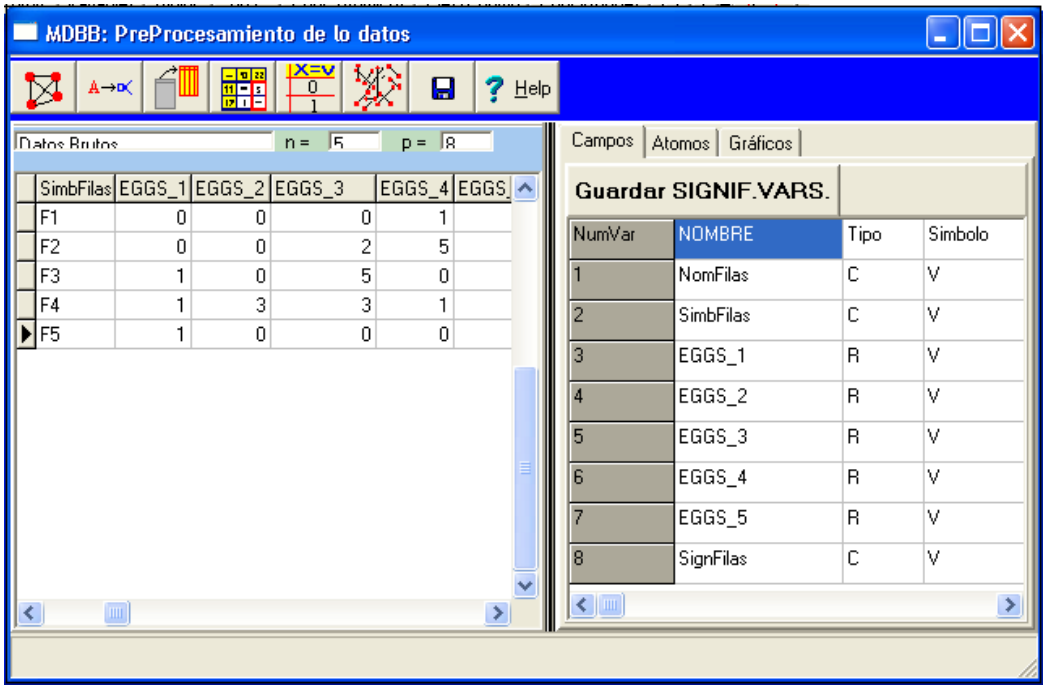


Figura 6.5.5.4. La tabla de datos ha sido transformada en un paquete de datos que puede ser analizado usando los biplots.

Obsérvese en esa figura que las filas han sido designadas por F_1, F_2, ..., F_5.

En el lado derecho de la figura aparecen, ahora, las identificaciones de las columnas (variables) del nuevo conjunto de datos.

En la **figura 6.5.5.5.** se ve un biplot de estos datos (GALINDO, Análisis Factorial de Correspondencias). Puede verse, por ejemplo, la asociación de gran consumo de EGGS a los países

F_1, F_2= (No pertenecientes a la cortina de hierro, Norte o Centro).

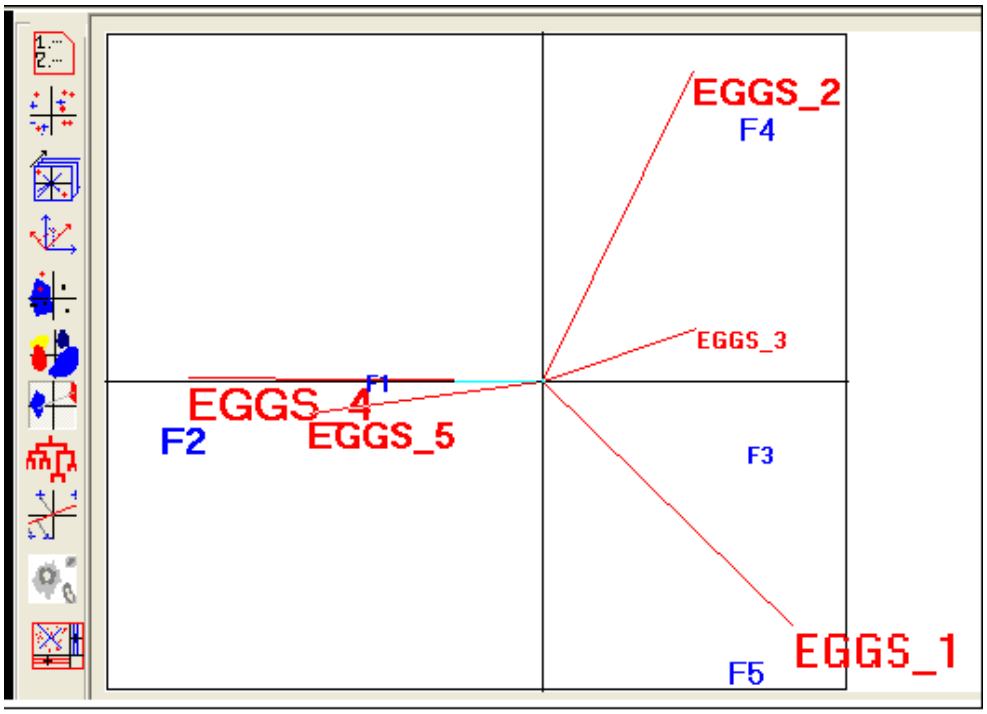
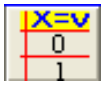


Figura 6.5.5.5. Biplot de la taba de contingencia mostrada en la **figura 6.5.5.4.**

6.5.5. CREACIÓN DE FICHEROS EN FORMA DE TDC, TABLA DUSYUNTIVA COMPLETA.



Para datos mixtos (variables cuantitativas y cualitativas) interesa muchas veces transformar los datos de forma que, si para una variable han sido observados *k* valores (cantidades o calidades), a esa variable corresponden *k* variables indicadoras.

Por ejemplo, la variable cuantitativa NUTS genera las 7 variables indicadores siguientes:

NUTS= 1, NUTS= 2, NUTS= 3, NUTS= 4, NUTS= 5

NUTS= 6, NUTS= 7, NUTS= 8.

El resultado de esta codificación es grabado en el fichero indicado por el usuario, extensión .DBF.

El sistema añade al nombre indicado por el usuario el prefijo «TDC-», lo que permite reconocer el contenido de estos ficheros de datos por su nombre.

Por ejemplo, si el usuario indica el nombre prot, el sistema crea el fichero TDC-prot.DBF con el resultado de esa codificación.

Una vez grabado, el fichero TDC-prot.DBF puede ser analizado por los métodos biplot.

En nuestro caso, leyendo ahora el nuevo fichero, obtenemos la **figura 6.5.6.1.**

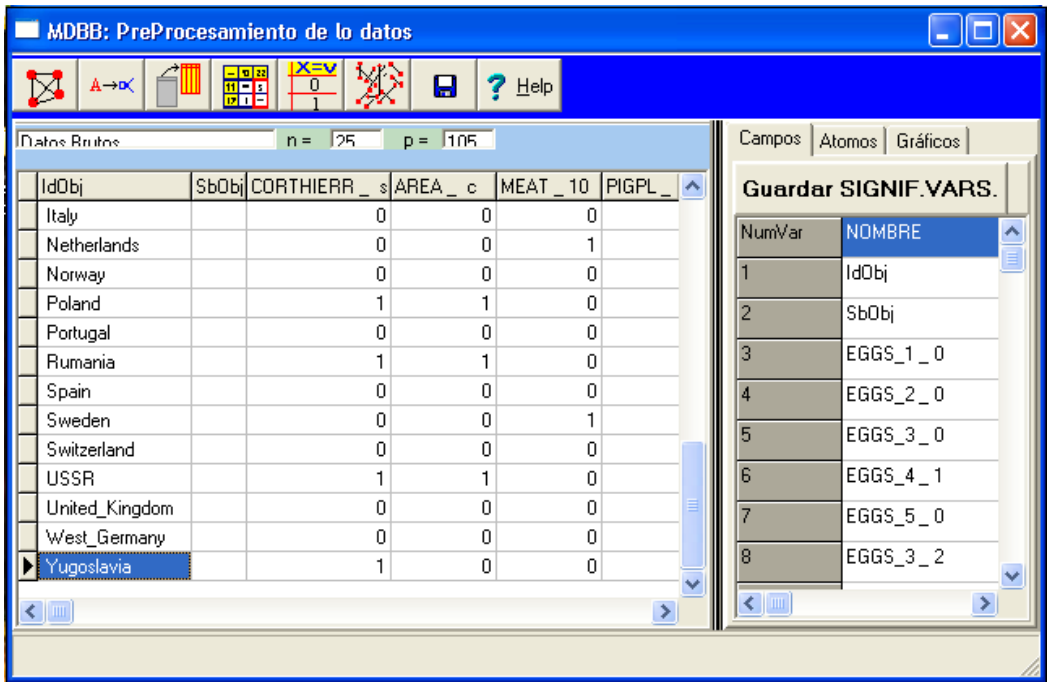


Figura 6.5.6.1. La tabla después de codificada en la forma TDC (Tabla Disyuntiva Completa).

En esta figura son visibles algunas de las columnas de la nueva tabla.

Los individuos (países) siguen identificados por los nombres en el fichero original.

En la **figura 6.5.6.2**. está el biplot (GALINDO, Estandardizar Columnas) en donde las nuevas variables binarias están representadas por «**x**».

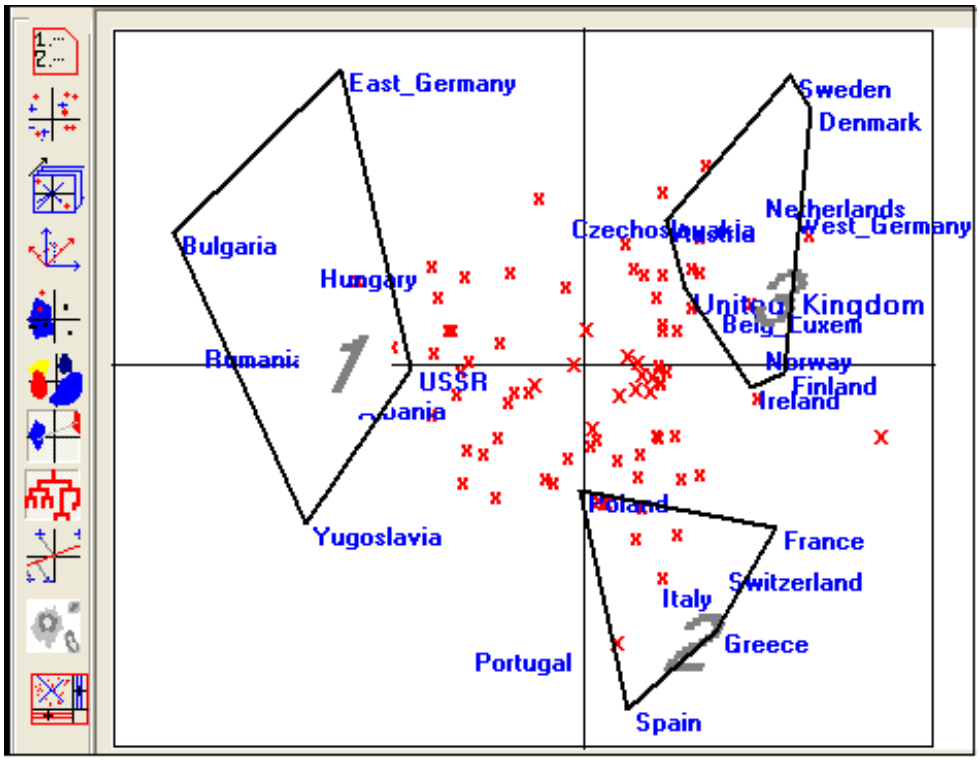


Figura 6.5.6.2. Análisis de los datos de Proteínas después de transformados en datos binarios, con el resultado de un análisis cluster superpuesto.

6.5.6. CREACIÓN DE FICHEROS CON AFINIDADES ENTRE ÁTOMOS.

 Creado el grafo de intersección de los átomos en el que se descompone el fichero de datos, es posible calcular las afinidades entre átomos usando

$$d(a,b) = \frac{|a \cap b|^2}{|a||b|}$$

la expresión:

en que *a* y *b* son dos átomos.

Ver **Capítulo IV**, apartado 4.6.1.

Con los valores de las afinidades calculadas entre todos los pares de átomos existentes, se obtiene una matriz cuadrada que el sistema graba en el fichero señalado por el usuario.

El nombre del fichero es prefijado con el símbolo AEA de Afinidades Entre Átomos.

Leyendo este fichero se puede realizar un análisis MDS - Clásico o Métrico - ver VICENTE VILLARDÓN (1992), mediante un biplot de GALINDO sobre datos doblemente centrados.

La **figura 6.5.7.1.** muestra el resultado de aplicar ese método a una matriz de afinidades producida a partir de los datos de proteínas, después de transformadas en discretas (3 valores) las variables continuas.

Examinando las asociaciones producidas, podrían descrubrirse asociaciones potencialmente interesantes, como hemos explicado en el **Capítulo IV**, apartado 4.6.3.

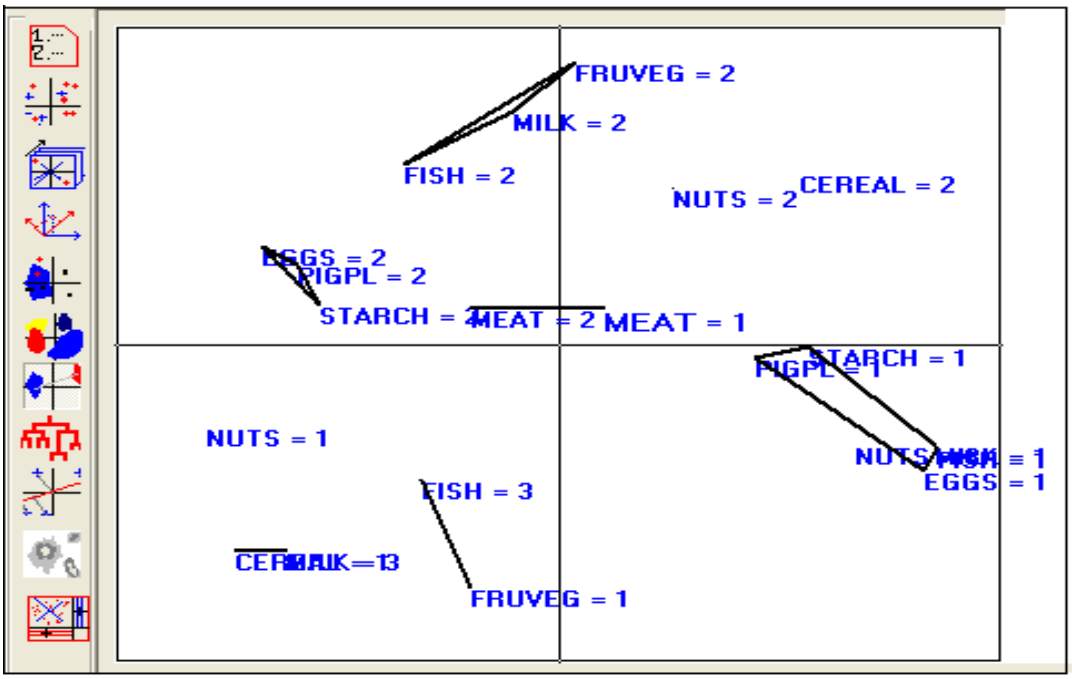


Figura 6.5.7.1. Biplot (GALINDO, Doble Centrado) obtenido a partir de una matriz de afinidades entre átomos, al que se ha superpuesto el resultado de un análisis cluster.

6.6. CRIACIÓN DE BIPLOTS.

El sistema permite crear biplots en estrecha conexión con la exposición teórica del **Capítulo II** tomando en consideración lo que se ha puesto de manifiesto en el **Capítulo V**.

El biplot calculado y presentado por el sistema resulta de tres decisiones a realizar por el usuario:

1ª Decisión Qué variables y individuos usar (elementos activos). Ver el apartado 6.6.1.

2ª Decisión Qué opciones gráficas emplear al pintar el biplot. Ver el apartado 6.6.2.

3ª Decisión Qué tipo de biplot y qué transformación de datos a emplear. Ver el apartado 6.6.3.

6.6.1. SELECCIÓN DE VARIABLES E INDIVIDUOS.

La definición de los individuos y de las variables activas se realiza en la ventana ilustrada en las **figuras 6.6.1.1.** y **6.6.1.2.**

En esas figuras se puede ver que el sistema permite elegir al azar un porcentaje de los individuos o bien señalar explícitamente cuales son las variables y los individuos activos.

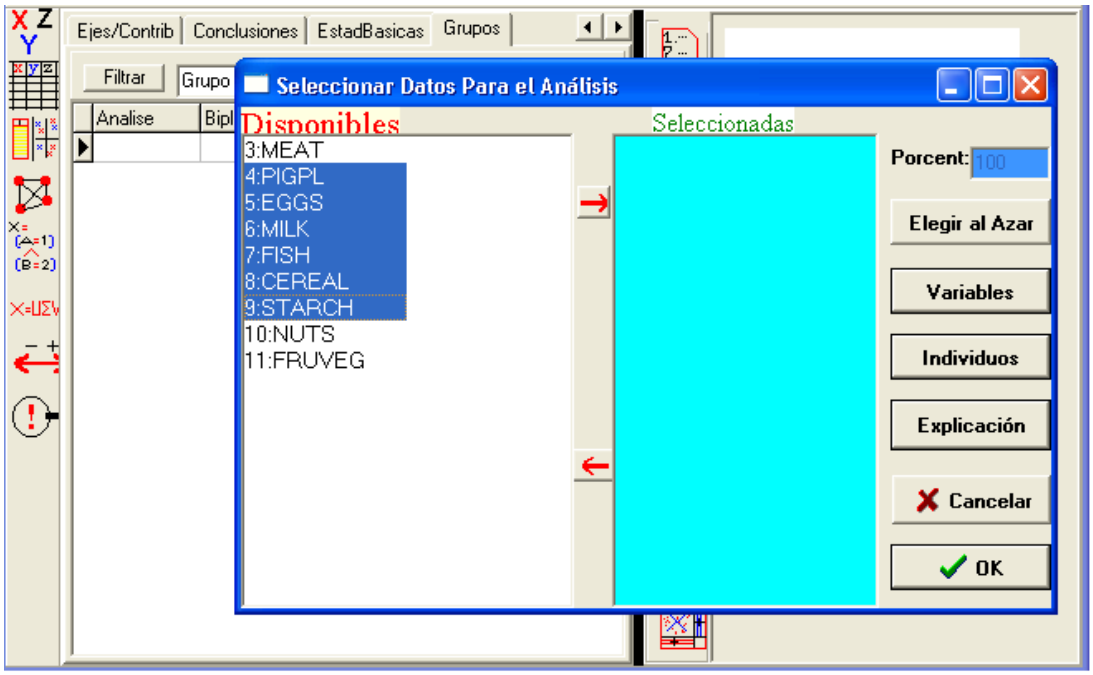


Figura 6.6.1.1. Selección de las variables e individuos activos.

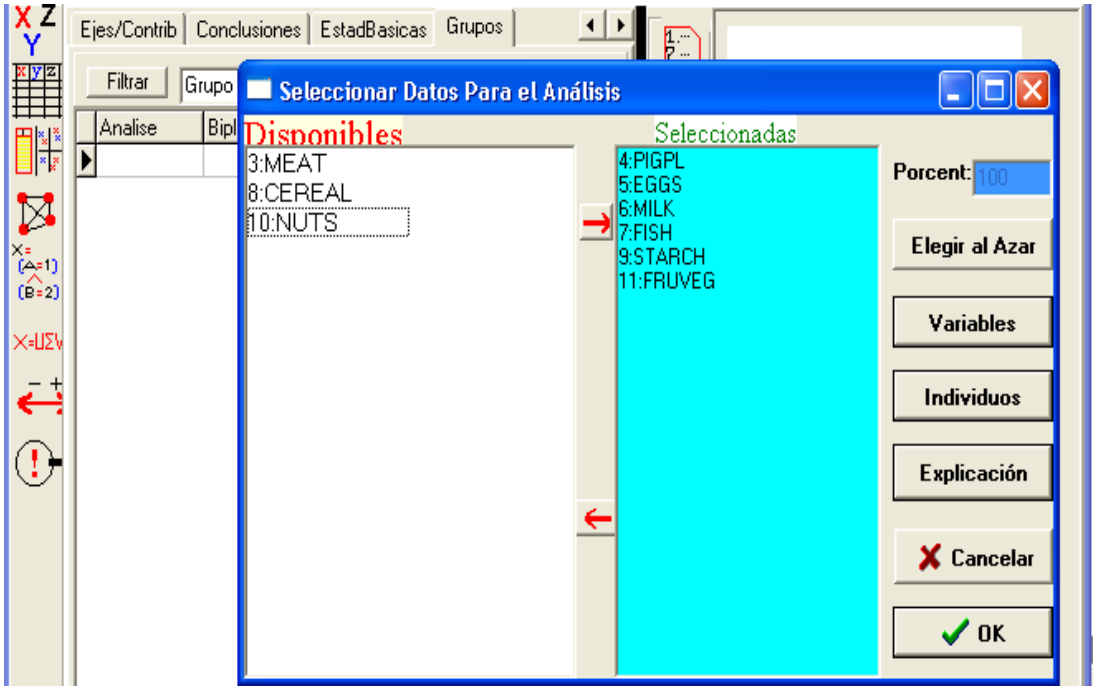


Figura 6.6.1.2. Seleccionar como variables activas las variables 4, 5, 6, 7, 9, 11.

Por ejemplo, en la **figura 6.6.1.3.** se han considerado como individuos activos los países 10, 13, 17, 19.

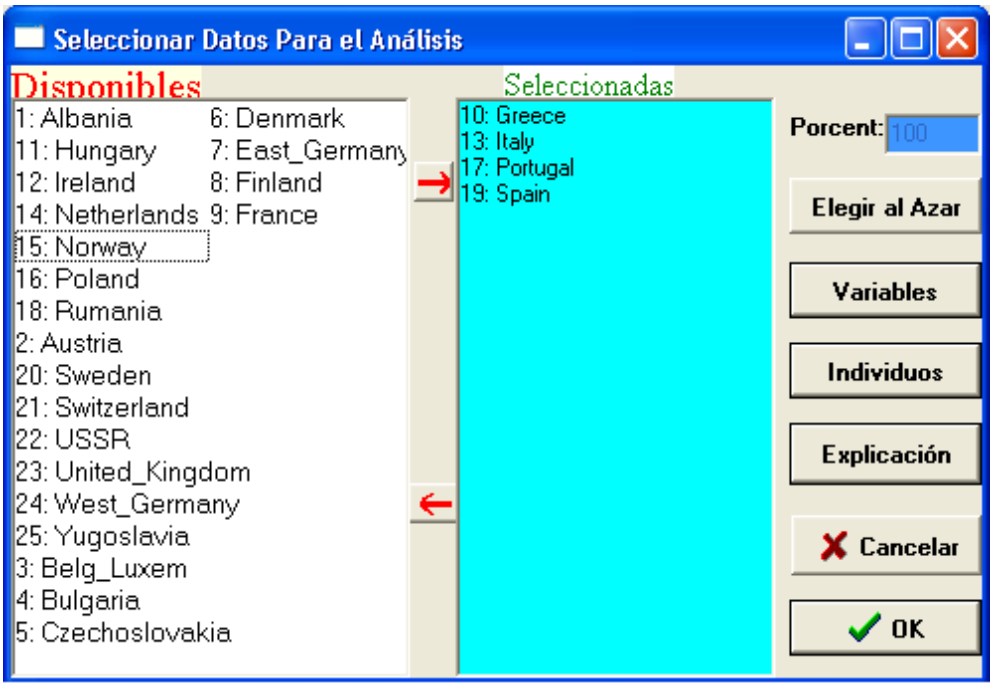


Figura 6.6.1.3. Seleccionar como individuos activos los países 10, 13, 17, 19.

Esto significa que se puede combinar una lista de variables de interés (variables activas) con la elección al azar de una parte de los individuos (individuos activos).

Los individuos y las variables que **no** han sido elegidos como activos son considerados, potencialmente, como suplementarios con relación al biplot actual. Es entre esos individuos y variables donde son seleccionados los elementos suplementarios.

Cuando el análisis engloba todos los individuos y variables no es necesario proceder a esta selección.

6.6.2. OPCIONES GRÁFICAS.

Antes de iniciar la construcción del biplot deben definirse los parámetros gráficos adecuados a la naturaleza de los datos.

Esta definición se realiza en la ventana de la **figura 6.6.2.1.**

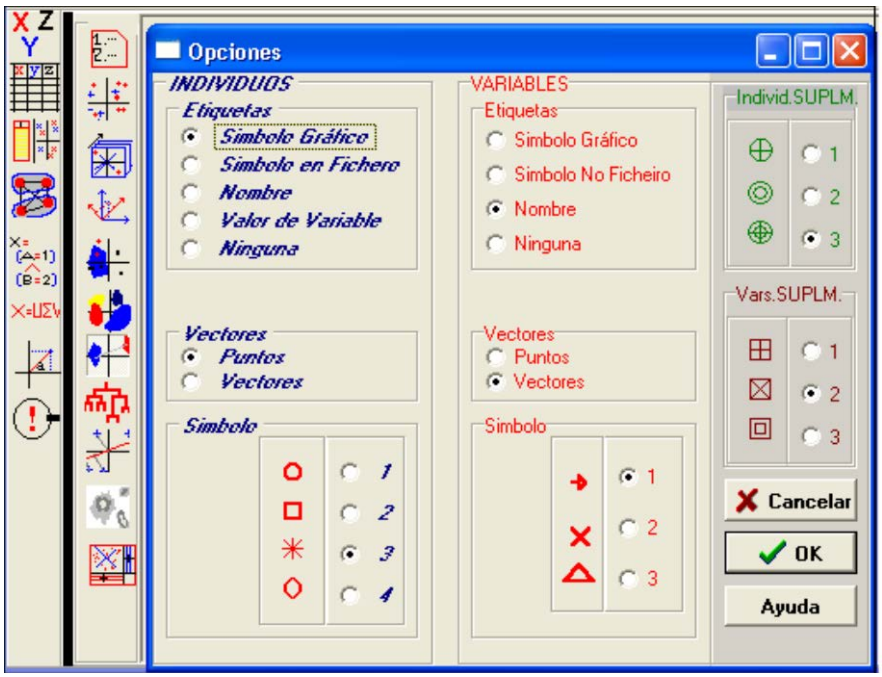


Figura 6.6.2.1. Opciones gráficas.

Tanto para individuos como para variables activos y suplementarios puede elegirse entre un **Símbolo Gráfico** (indicado en la parte inferior de la ventana), el símbolo en el fichero (para las variables este símbolo es *X*), el nombre del individuo (identificador en la columna 1 del fichero) o bienninguno de los tres.

Para los individuos es además posible reemplazar el identificador por el valor de una variable - numérica o cualitativa - lo que permite visualizar la distribución espacial – en el biplot - de toda variable observada.

En general, los individuos son representados por puntos y las variables por vectores a partir del centro de gravedad de las nubes.

6.6.3. ELECCIÓN DE LA TRABSFORMACIÓN DE LOS DATOS.

El sistema considera que se inicia una nueva sesión de análisis cuando se abre el fichero de datos para analizar.

A cada sesión corresponden distintos biplots; a cada uno de esos biplots corresponde una Transformación de los datos y un Tipo de Biplot.

En la **figura 6.6.3.1.** puede verse la ventana de selección del tipo de biplot y tipo de transformación.

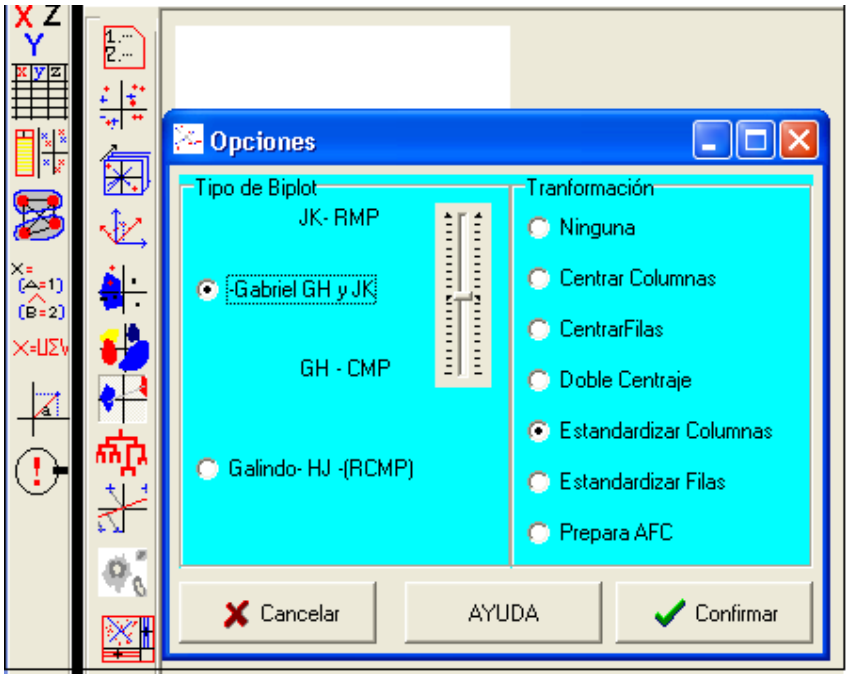


Figura 6.6.3.1. Selección del tipo de biplot y tipo de transformación.

Se consideran dos tipos de biplot:

GABRIEL (GH/CMP – JK/RMP)

y GALINDO.

Para los biplots de GABRIEL hay que elegir el grado - α - que define el carácter GH/CMP o JK/RMP de esos biplots. Ver **capítulo II**.

6.7. OPERACIONES GEOMÉTRICAS SOBRE EL BIPLOT.

El programa permite realizar sobre el gráfico de un biplot tres transformaciones geométricas clásicas:

- Dilatación
- Rotación
- Reflexión

Ninguna de estas transformaciones altera la interpretación del biplot. Su inclusión se destina a facilitar la comparación de biplots obtenidos por métodos distintos y a mejorar su lectura.

6.7.1. DILATACIÓN.

 A veces, la nube de individuos tiende a concentrarse en el centro del gráfico, lo que complica, o, incluso, impide su examen visual. Ver, por ejemplo, la **figura 6.7.1.1.**

Para los biplots de GABRIEL - ver **Capítulo II** - la distancia entre los marcadores de los individuos y de las variables no tiene significado y por eso, la dilatación no altera la interpretación.

Para los biplots de GALINDO, esta distancia es interpretada como preponderancia de la variable para el individuo o del individuo para la variable.

La dilatación mantiene los ángulos entre los marcadores de los individuos y de las variables y por eso no influye en la interpretación de esas relaciones.

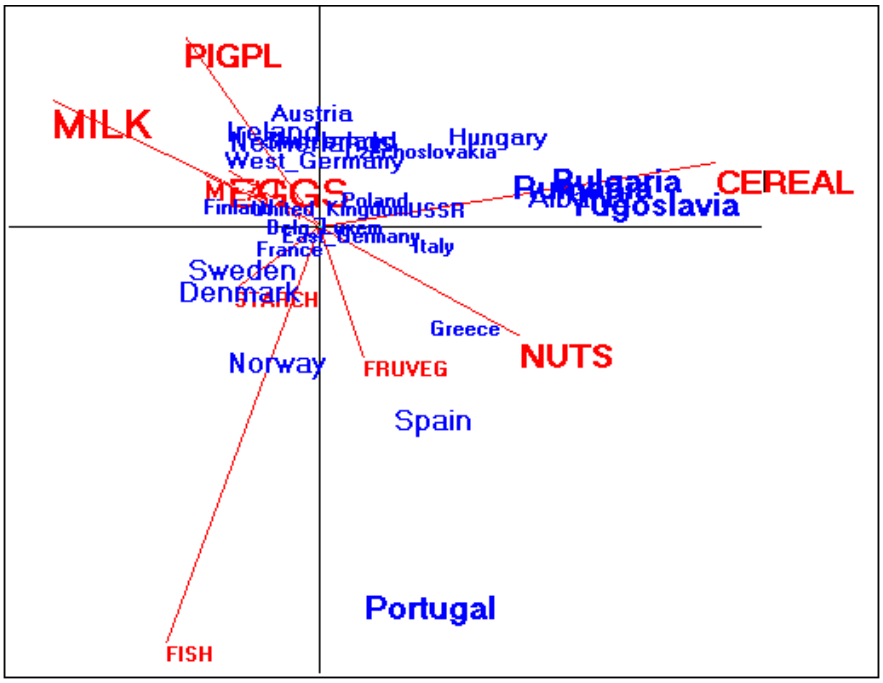


Figura 6.7.1.1. Antes de dilatar.

En la **figura 6.7.1.2.** está el resultado de dilatar el biplot de la **figura 6.7.1.1.**

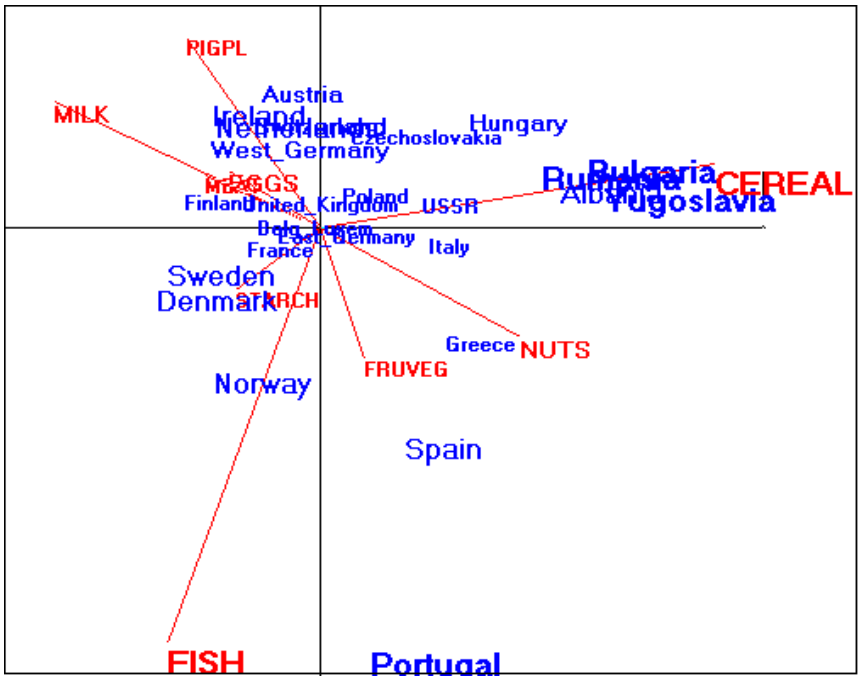
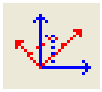


Figura 6.7.1.2. Resulta de **6.7.1.1.** por dilatación de la nube de individuos.

6.7.2. ROTACIÓN Y REFLEXIÓN.

 El sistema permite girar el biplot de un ángulo arbitrario - hacia la izquierda o hacia la derecha.

La definición del movimiento se realiza en la ventana que se ve en la **figura 6.7.2.1.**

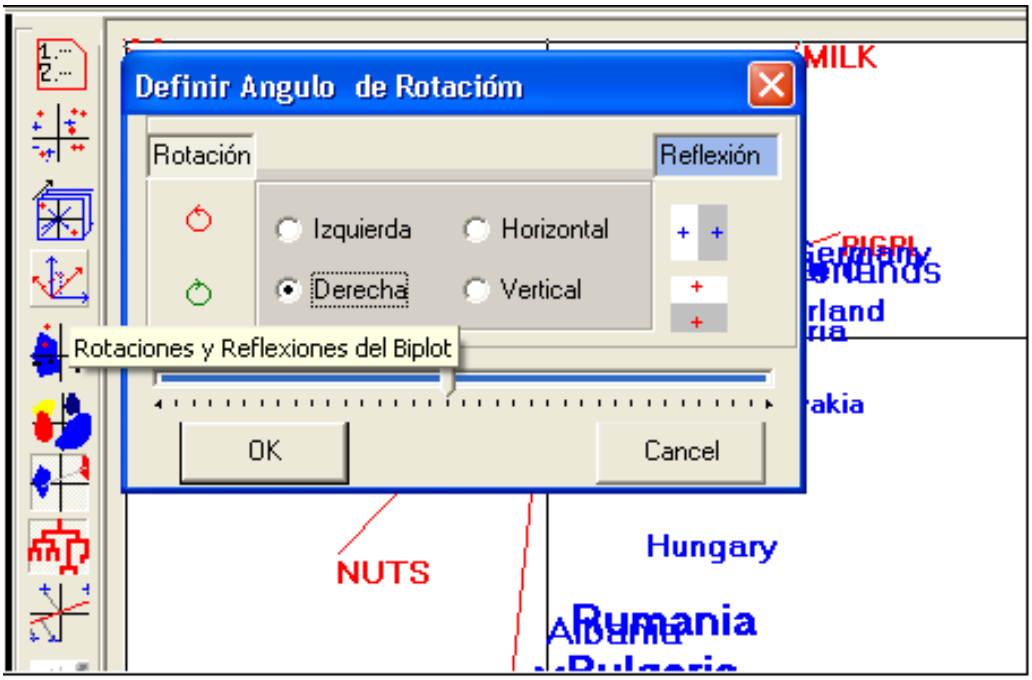


Figura 6.7.2.1. Ventana para definir la transformación geométrica.

Dado que las rotaciones mantienen los ángulos y las distancias entre marcadores - la interpretación del biplot se altera pero la lectura mejora.

Lo mismo ocurre con la reflexión, especificada en la misma ventana de la **figura 6.7.2.1.**

6.8. CREACIÓN, INTERPRETACIÓN Y COMPARACIÓN DE GRUPOS.

En el **capítulo IV**, apartado 4.2. se han definido como operaciones básicas de interpretación de resultados, las operaciones de Identificación, Caracterización, Interpretación y Comparación de Grupos.

En el apartado 4.3 se ha definido un lenguaje de interpretación que se basa en el concepto de grupo.

En los apartados 4.5, 4.6 y 4.7 de ese mismo capítulo se presentan algoritmos para generar expresiones conjuntivas de átomos para caracterizar grupos y particiones.

En el **capítulo V**, apartado 5.5.2 se ha delineado una estrategia para la construcción de síntesis de caracterización y comparación de grupos.

Al implementar el sistema prototipo que ahora se presenta, tuvimos la preocupación de seguir muy de cerca las ideas teóricas desarrolladas en los **capítulos IV** y **V**.

Los grupos (individuos o variables) pueden resultar de decisiones del usuario o expresar resultados de técnicas de análisis, como el análisis cluster.

Así, en la **figura 6.8.1.** se representa un biplot sobre el cual el usuario identificó un grupo y la técnica de análisis cluster identificó los grupos 2, 3 y 4.

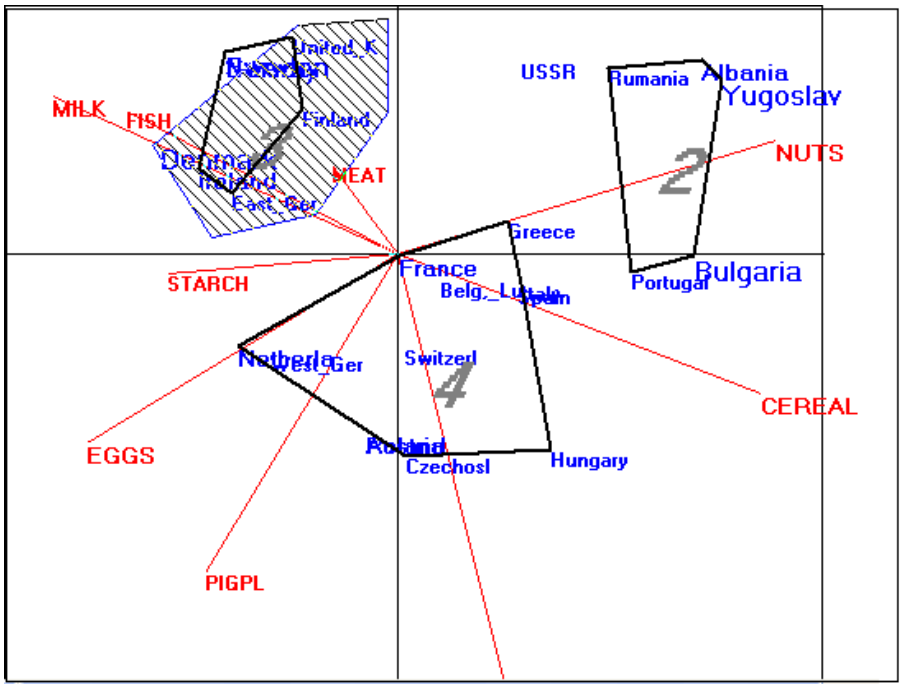


Figura 6.8.1. Un grupo definido por el usuario y 3 grupos generados por análisis cluster.

La caracterización de un grupo se realiza de varios modos:

1º. Presentando las observaciones que lo integran. Ver **figura 6.8.2.**

DefGrupo

Nombre: Grupo nº1

Resumen: 0

Color del Grupo

Ayuda | OK

Composición | Res.VarQuantit | Res.VarCualit | Síntesis | Expr.Conyuntivas

Nº I: Nombre Ind.	Simb.Ind.	CORTHIERR	AREA	MEAT	PIGPL	E
2: Austria	n	n	c	9	14	4
6: Denmark	n	n	n	11	11	4
8: Finland	n	n	n	10	5	3
14: Netherlands	n	n	n	10	14	4
20: Sweden	n	n	n	10	8	4
21: Switzerland	n	n	s	13	10	3
23: United_Kingdor	n	n	n	17	6	5
24: West_Germany	n	n	n	11	13	4

Figura 6.8.2. Composición del grupo nº 1. Para cada una de las variables se indica su valor en los individuos del grupo.

2º. Calculando las estadísticas de las variables cuantitativas y cualitativas, considerando los valores que toman sobre los individuos del grupo y sobre la totalidad de las observaciones. Ver **figura 6.8.3.** y **6.8.4.**

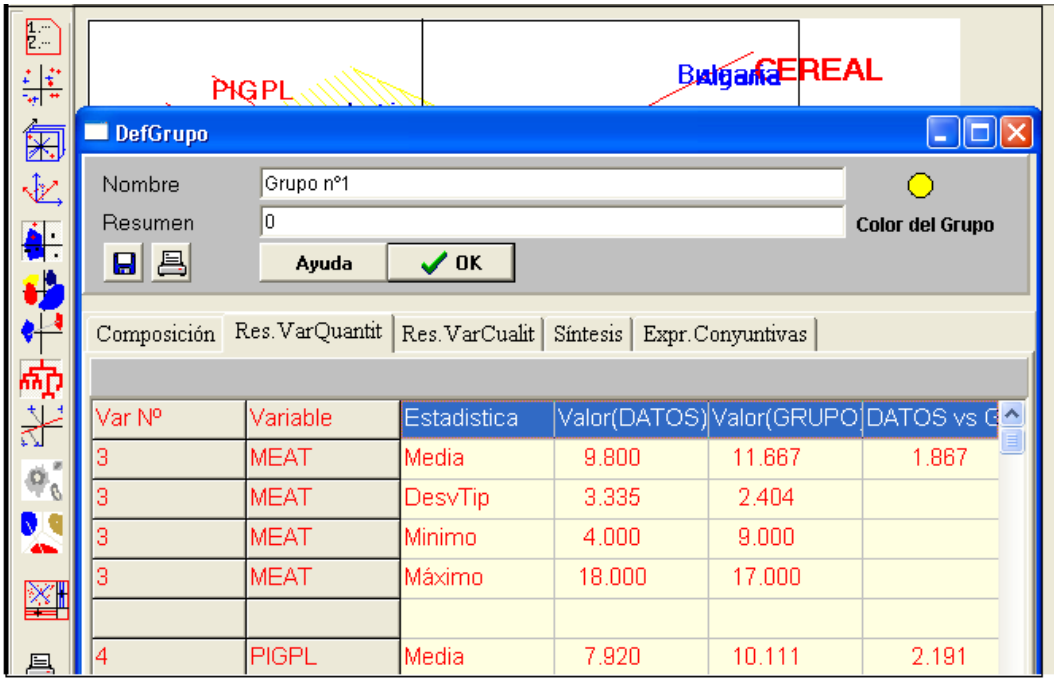


Figura 6.8.3. Resumen de las variables cuantitativas.

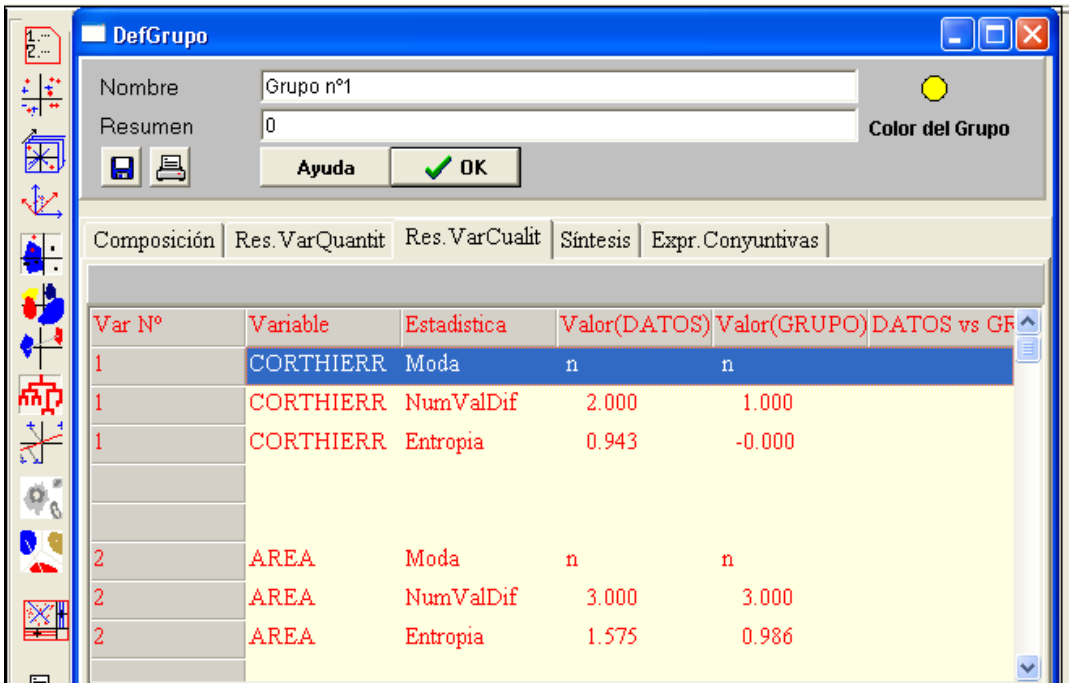


Figura 6.8.4. Comparación de una variable cualitativa: DATOS vs GRUPO.

3º. Sintetizando los hallazgos que resultan de comparar las estadísticas calculadas según los métodos descritos en el apartado 5.5.2. Ver **figura 6.8.5.**

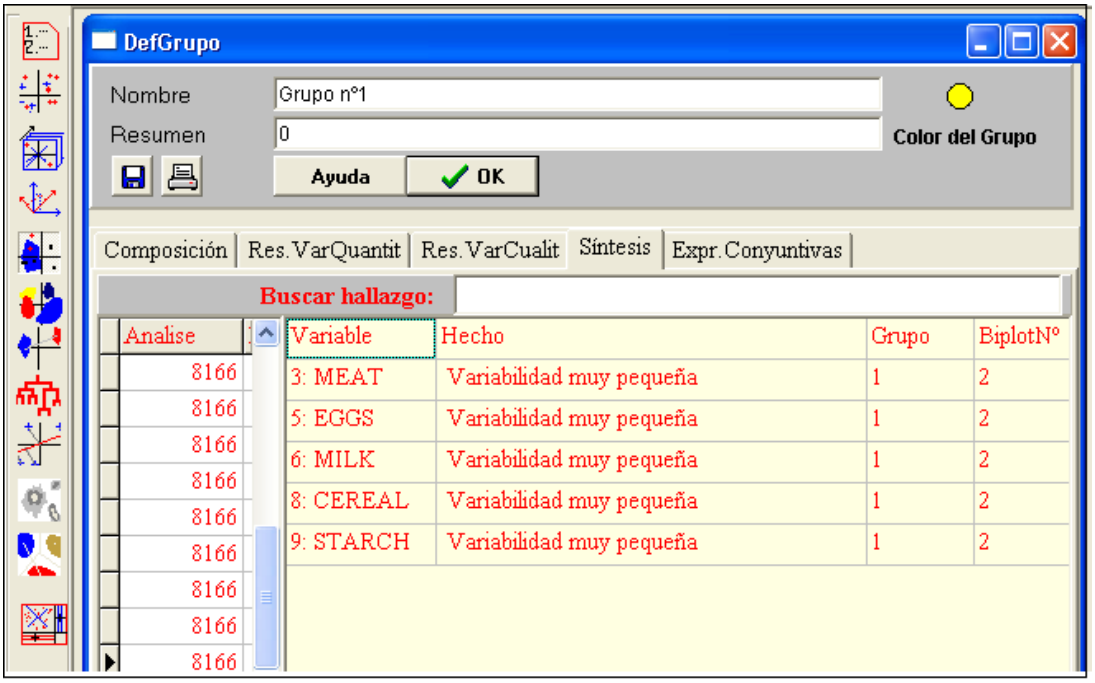


Figura 6.8.5. Síntesis del Grupo nº 1.

4º. Generando - con base en el grafo de intersección y en el concepto de afinidad, expresiones conjuntivas de caracterización de grupos, considerando las variables relevantes identificadas en los pasos anteriores. Ver **figura 6.8.6.**

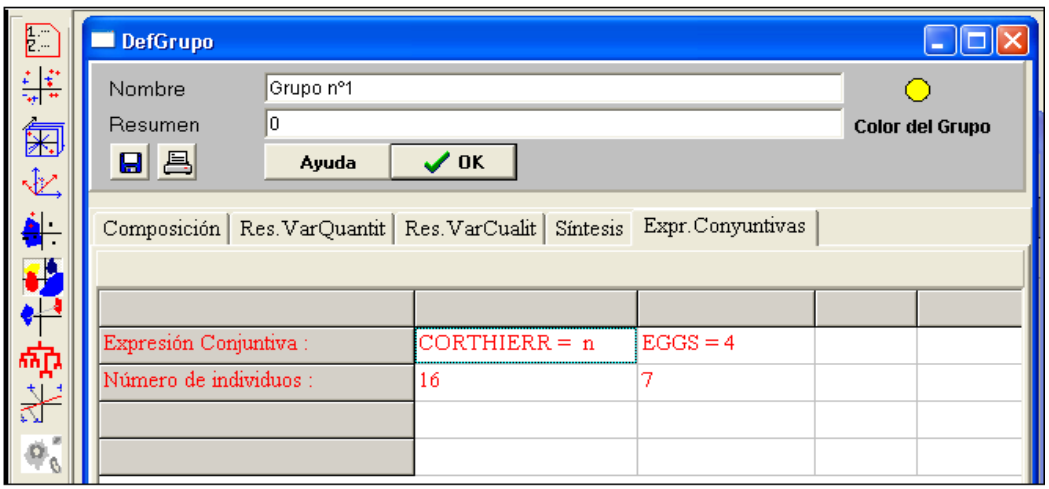


Figura 6.8.6. El sistema sugiere como síntesis de este grupo (CORT HIERR= «*n*») ∧ (EGGS= 4).

En la **figura 6.8.6**, el sistema sugiere como síntesis del grupo Nº 1 la expresión conjuntiva siguiente:

(CORT HIERR= *n*) ∧ (EGGS= 4)

que «traducida» al castellano significa

> «Países que no pertenecen a la cortina de hierro y en que el porcentaje de proteínas proveniente de los huevos es 4».

Para comparar dos grupos que están marcados sobre el biplot actual es necesario informar al sistema cuales son esos grupos. Esto se hace apuntando, sucesivamente, cada uno de los dos grupos por comparar.

El resultado aparece en una ventana que se puede ver en la **figura 6.8.7.**

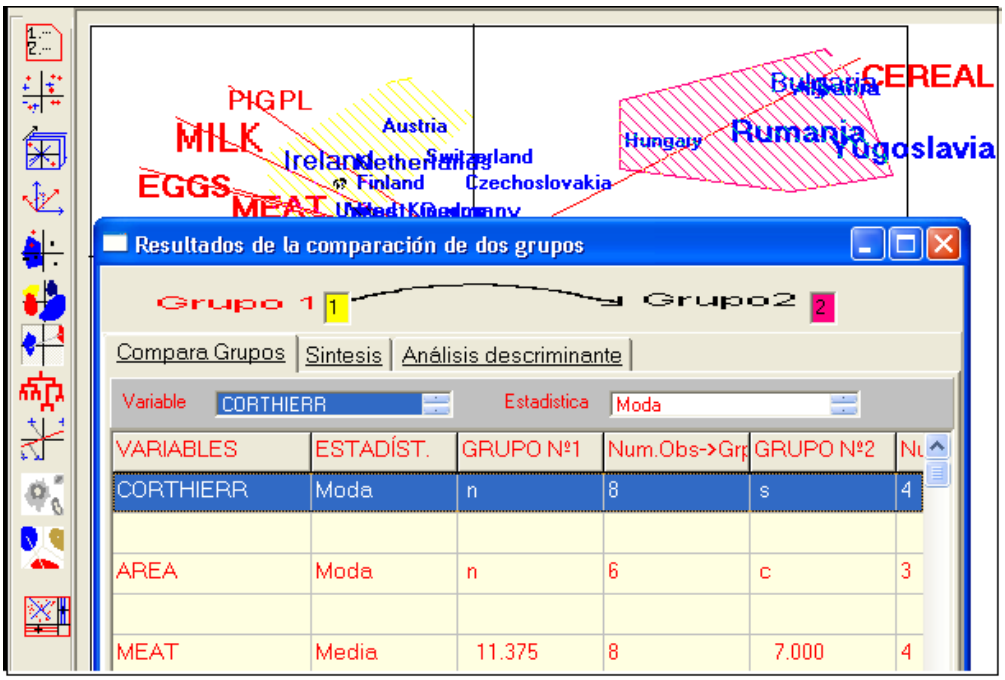


Figura 6.8.7. Comparación de dos grupos Amarillo y Rojo, comparando las estadísticas de las variables cualitativas y cuantitativas sobre los dos grupos.

En esa figura se puede ver que el sistema presenta las estadísticas de las variables calculadas con los valores que toman sobre los dos grupos, de forma que resulta fácil verificar lo que ocurre al pasar del uno al otro.

La **figura 6.8.8.** presenta la síntesis de los hechos relevantes que ocurren al pasar del grupo Nº 1 al grupo Nº 2.

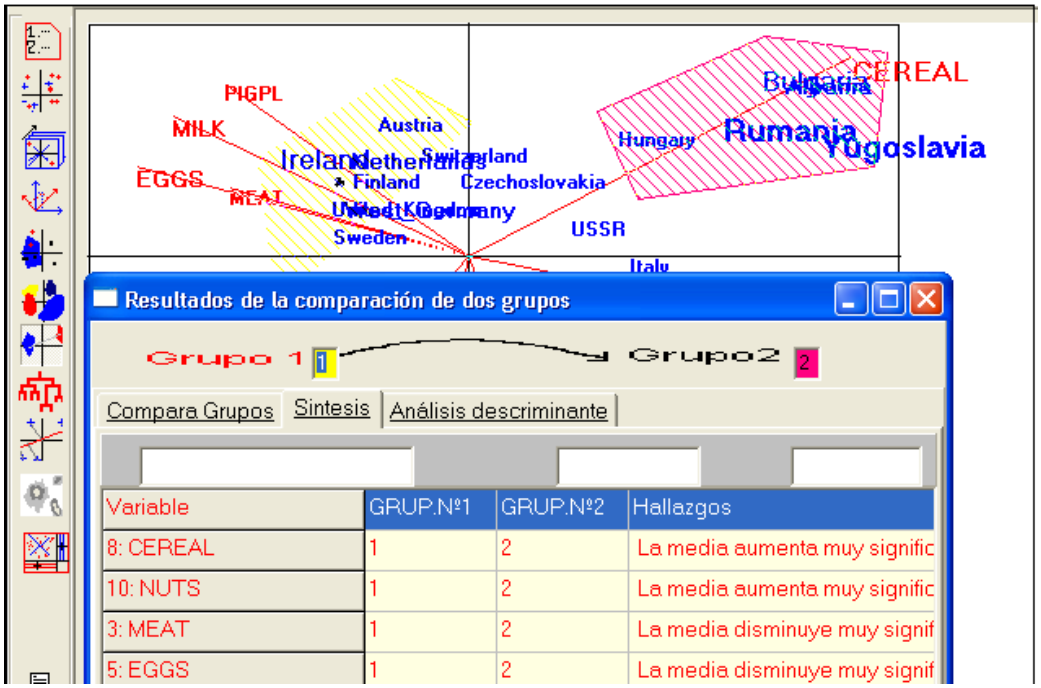


Figura 6.8.8. En síntesis, cuando se pasa del grupo 1 (AMARILLO) para el grupo 2 (ROJO). La media de CEREAL aumenta significativamente.

En el caso específico que la figura presenta, se verifica que las medias de CEREAL y NUTS aumentan de modo muy significativo y que las medias de MEAT y EGGS disminuyen muy significativamente.

6.9. CREACIÓN DE GRUPOS USANDO ANÁLISIS CLUSTER.

En este sistema se usa un programa de análisis cluster para generar jerarquías de partes del conjunto de individuos.

La inclusión de un programa de análisis cluster en un sistema que busca mostrar la adecuación de la técnica de biplots para núcleo de un sistema de

minería de datos se justifica por las consideraciones del apartado 4.1. (ver **figura 4.1.1.**, en particular) y del **capítulo IV**, apartado 4.2.

El programa de análisis cluster genera automáticamente resultados que son grupos de individuos o particiones de grupos que se pueden representar biplots por configuraciones de marcadores.

El objetivo es, principalmente, ilustrar las ideas de interpretación desarrolladas en el **capítulo IV**.

El programa de análisis cluster usa como índice de disimilitud la distancia euclídea entre marcadores de individuos en biplots y como criterio de agregación de clases el criterio de WARD. Ver VICENTE-TAVERA (1992).

Se ha usado, para construir el árbol, un algoritmo de vecinos recíprocos que aparece descrito en RHAM (1980).

En la **figura 6.9.1.** se presenta la ventana con el dendograma resultante de aplicar el programa a los datos de GABRIEL (1981).

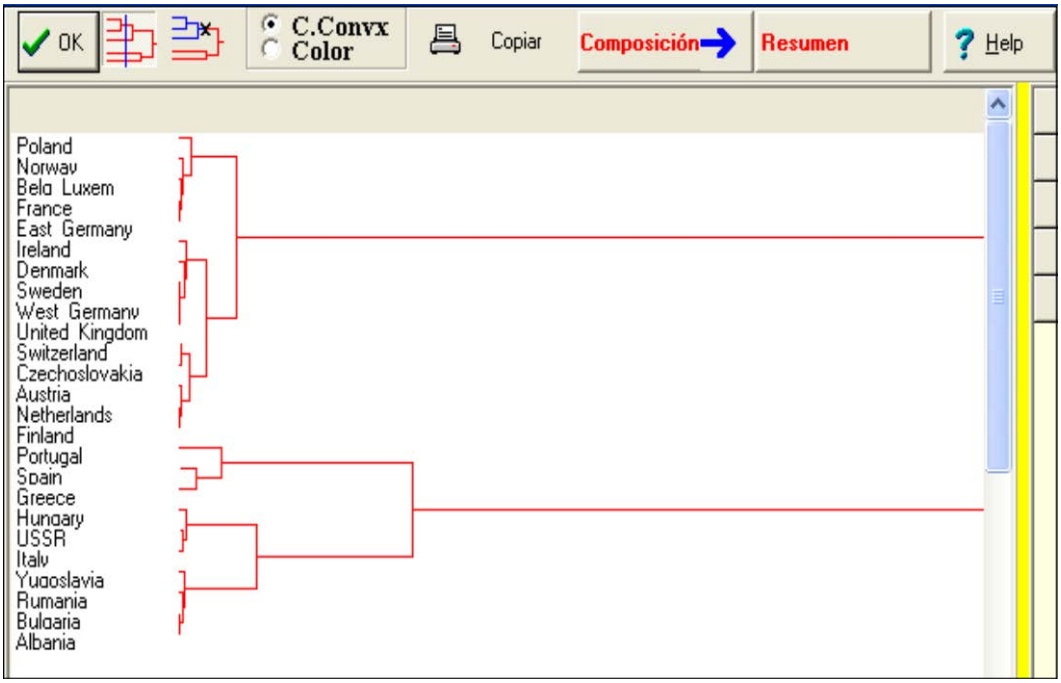


Figura 6.9.1. Árbol de clasificación de los países.

El usuario puede interactuar con ese gráfico de dos modos:

Primer Modo - Cortando el dendograma al nivel que le plazca.

En este caso, el sistema reacciona pintando sobre el biplot actual la partición correspondiente y guardando los grupos que integran a esa partición en la base de datos, para estudio posterior. Ver el resultado en las **figuras 6.9.2.** y **6.9.3.**

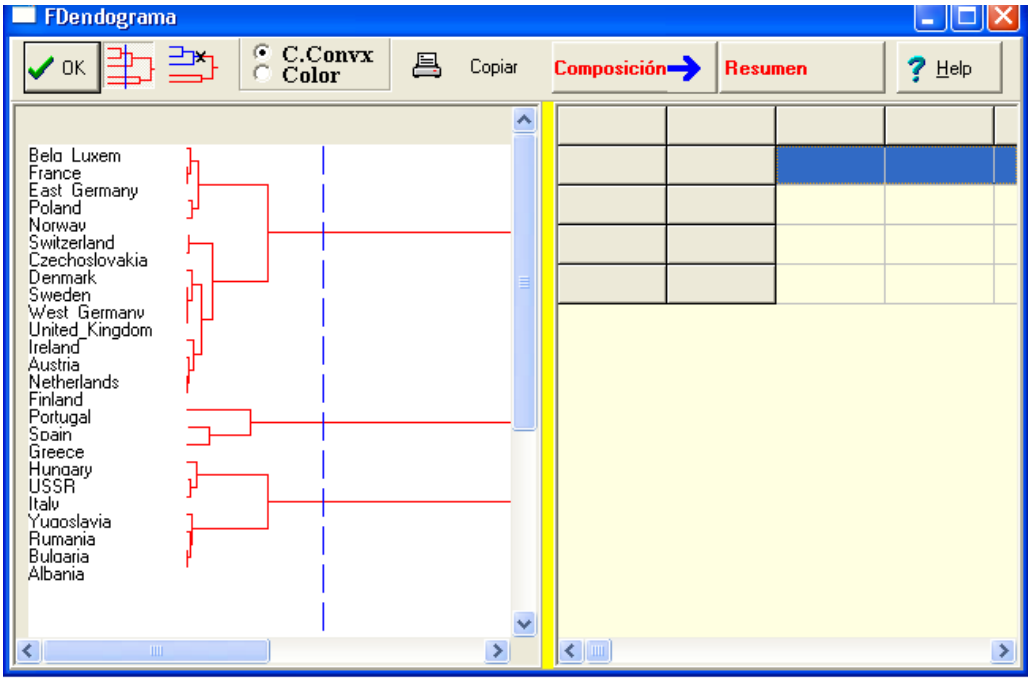


Figura 6.9.2. Definición de una partición por el nivel del corte.

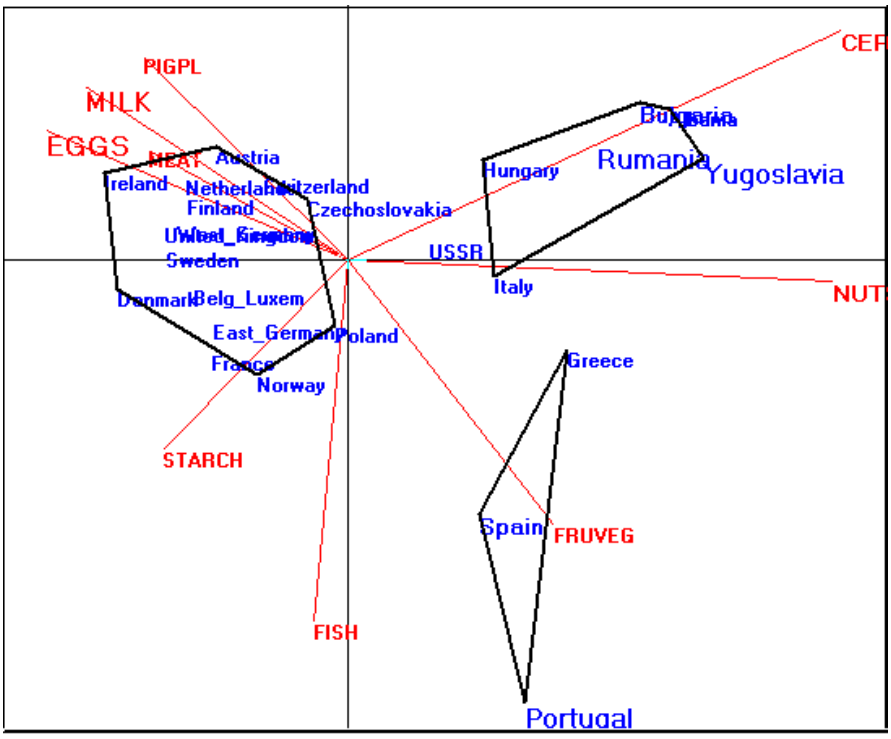


Figura 6.9.3. Definición de una partición por el nivel del corte.

Segundo Modo - Apuntando el vértice a estudiar.

En este caso, el sistema reacciona presentando sobre el biplot actual el cierre convexo del grupo/vértice elegido o pintando en el color deseado por el usuario los marcadores respectivos. El grupo señalado es guardado en la base de datos para estudio posterior. Ver **figura 6.9.4.**

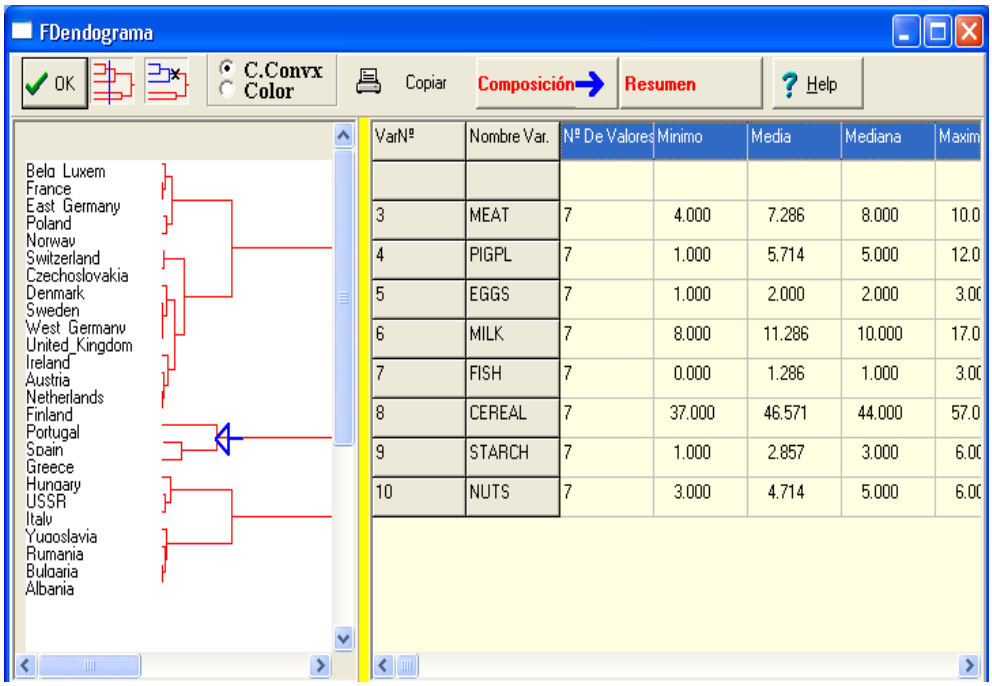


Figura 6.9.4. Estadísticas del vértice apuntado.

Es este segundo modo de interacción también es posible un estudio local (sin salir del programa de análisis cluster) que permite examinar las observaciones que integran el vértice elegido y los resúmenes estadísticos de las variables observadas, usando los valores tomados sobre los individuos del grupo. Ver **figura 6.9.4.**

6.10. ESTUDIO DE GRUPOS.

6.10.1. ORIGEN DE LOS GRUPOS.

En el estudio de los biplots es crucial distinguir entre una configuración de marcadores (de individuos y variables) y los objetos que esos marcadores representan.

Los marcadores están referidos a un sistema de coordenadas. Los objetos (variables e individuos) que representan los marcadores, no.

Una vez que casi todos los métodos de análisis multivariante generan como resultado principal grupos de individuos y de variables, estos grupos pueden ser representados por configuraciones de marcadores en biplots. La individualidad del grupo en el biplot puede ser materializada de modos distintos: por el centro de gravedad de los marcadores del grupo, por un color atribuido al grupo, por el cierre convexo de los marcadores, etc.

De aquí resulta que distintos resultados expresados por grupos- obtenidos por distintos analistas y técnicas - pueden ser incorporados en un mismo biplot. Para eso es necesario y suficiente conocer las coordenadas de esos objetos (individuos y variables) en el biplot actual.

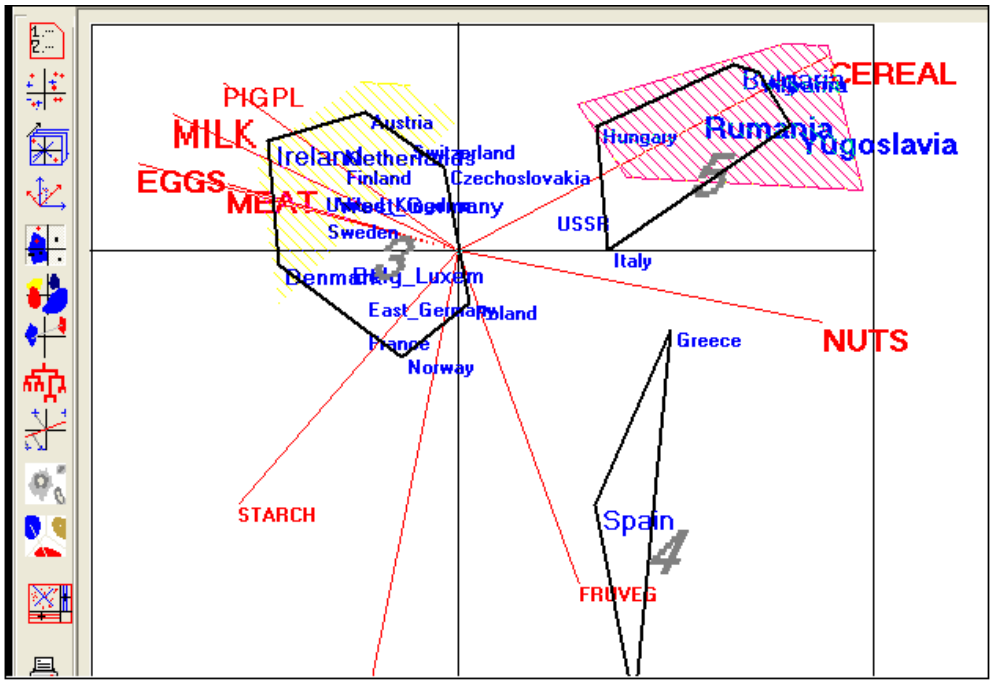


Figura 6.10.1.1. A los grupos AMARILLO y ROJO se añaden los grupos 3, 4, 5, sugeridos por análisis cluster jerárquico.

La figura 6.10.1.1. es un ejemplo. En esa figura se han superpuesto al biplot - en el que han sido definidos por el usuario dos grupos (AMARILLO y ROJO) - tres grupos sugeridos por un método de clasificación jerárquica, representados por sus cierres convexos e identificados por los números 3, 4, 5. Los números 1 y 2 han sido atribuidos por el sistema a los grupos AMARILLO (1) y ROJO (2).

Cuando se elige un nuevo paquete de datos, tiene inicio una nueva sesión de análisis, al que es atribuido un número secuencial (Número de Análisis).

Un nuevo biplot es creado cuando se cambian los parámetros que lo definen: **Tipo** (GABRIEL con $0 \leq \alpha \leq 1$, GALINDO) y **Transformación**.

Los grupos son numerados en secuencia dentro de un mismo biplot: cuando cambia el biplot se inicia una nueva secuencia de grupos.

Esto significa que, al final de una sesión, pueden coexistir muchos grupos: definidos por el usuario, sugeridos por el sistema, importados del exterior (aún no disponible en esta versión), correspondientes a planos factoriales distintos o pertenecientes al mismo plano.

Puede interesar presentar sobre el biplot actual - para comparar grupos, por ejemplo - uno o más de estos grupos.

En la **figura 6.10.1.2.** se pueden ver dos grupos (Amarillo y Rojo) definidos por el usuario sobre el biplot en el que están pintados los cierres convexos de los grupos sugeridos por análisis cluster.

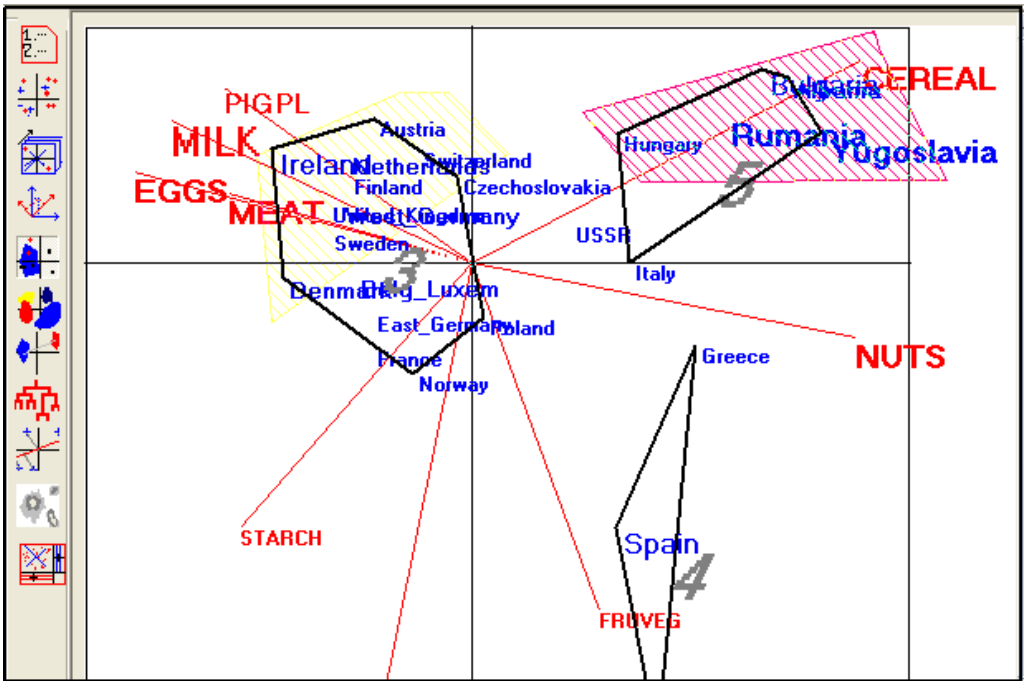


Figura 6.10.1.2. La ventana muestra todos los grupos existentes y permite seleccionar grupos para inspección y estudio.

6.10.2. ELECCIÓN DEL GRUPO O GRUPOS POR ESTUDIAR.

El botón  permite ver cuales son los grupos que están definidos en determinado momento.

El resultado está en la **figura 6.10.2.1.**

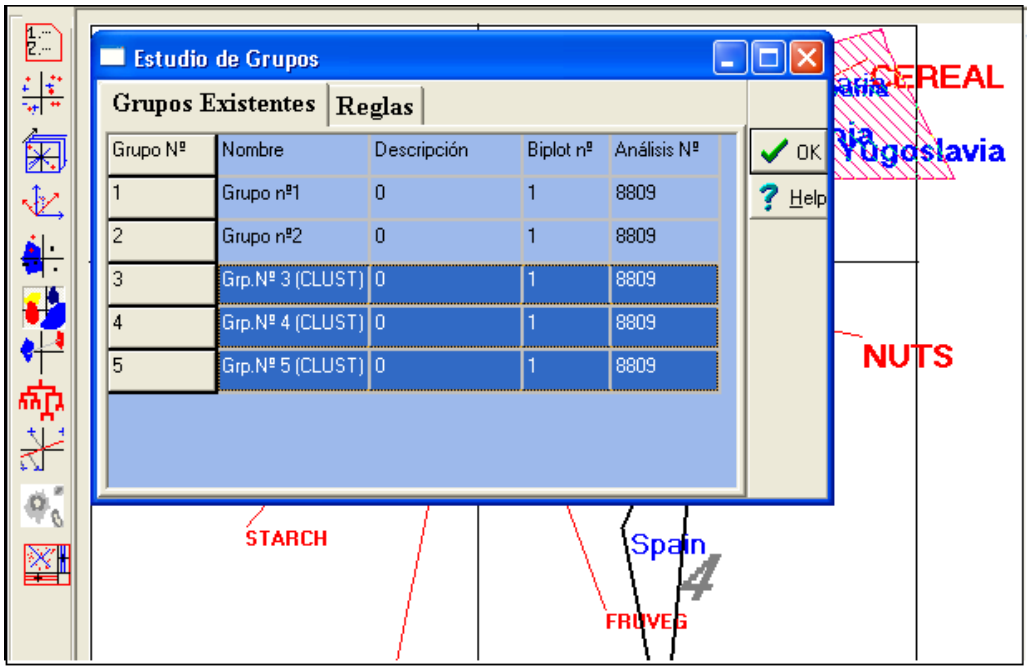


Figura 6.10.2.1. Los grupos existentes en un instante dado pueden pertenecer a biplots distintos, identificados por el usuario o por el sistema.

Por ejemplo: examinando la ventana de la **figura 6.10. 2.1.** se verifica que para el análisis actual (Nº 8809) y para el biplot actual (número 1) existen 4 grupos.

En ese ejemplo se verifica, además, que existen los grupos Nº1 y Nº 2 - definidos por el usuario y los grupos 3, 4, 5 definidos por el programa de análisis cluster.

El nombre de los grupos creados por el usuario puede ser el sugerido por el sistema o bien atribuido por el usuario cuando su percepción del significado del grupo lo permite.

En esta ventana podemos elegir estudiar un solo grupo (del biplot actual o de otro biplot) o un conjunto de grupos. Ver la **figura 6.10.2.2.**

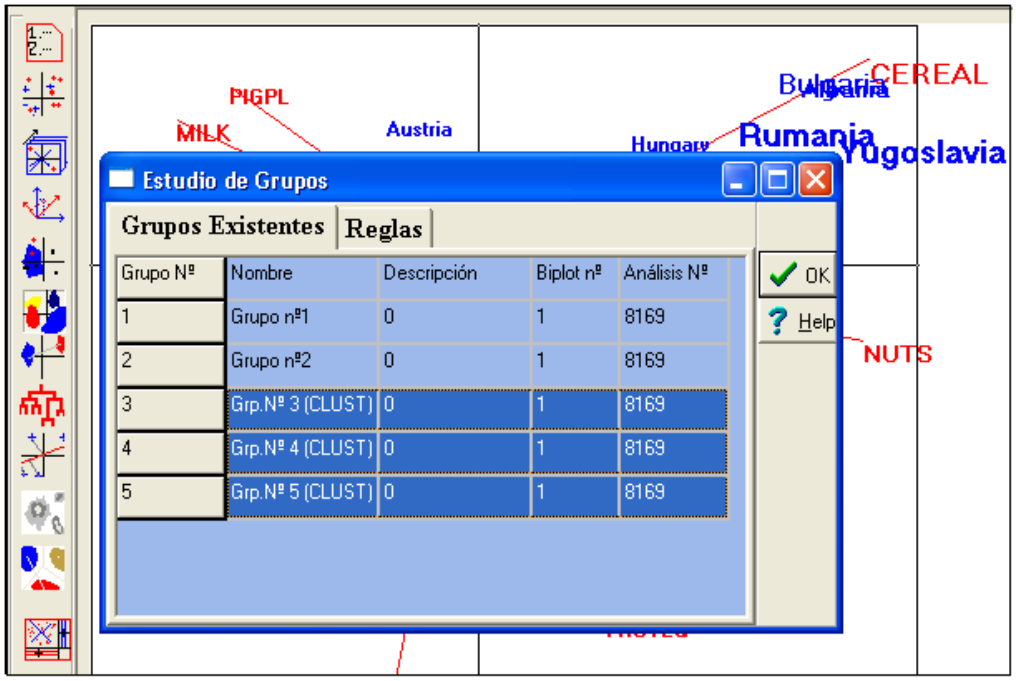


Figura 6.10.2.2. Los grupos 3, 4, 5 están marcados para estudio.

6.10.3. ESTUDIO DE UN GRUPO.

Para cada uno de los grupos marcados, el sistema realiza las tareas siguientes:

1. Pinta el grupo en el biplot actual usando las coordenadas, en el biplot actual, de los marcadores de los individuos y variables del grupo.
2. Pinta sobre cada grupo el número respectivo - si, en Opciones Globales, esa opción está seleccionada.

3. Presenta un resumen del grupo.

4. Al final el sistema obtiene reglas de interpretación de los grupos elegidos - considerando todos los grupos al mismo tiempo - usando la metodología de los árboles de regresión, conforme ha sido explicado en el **capítulo IV**, apartado 4.7.

En nuestro ejemplo, el resultado para el grupo Nº 3 se puede ver en la ventana de la **figura 6.10.3.1.**

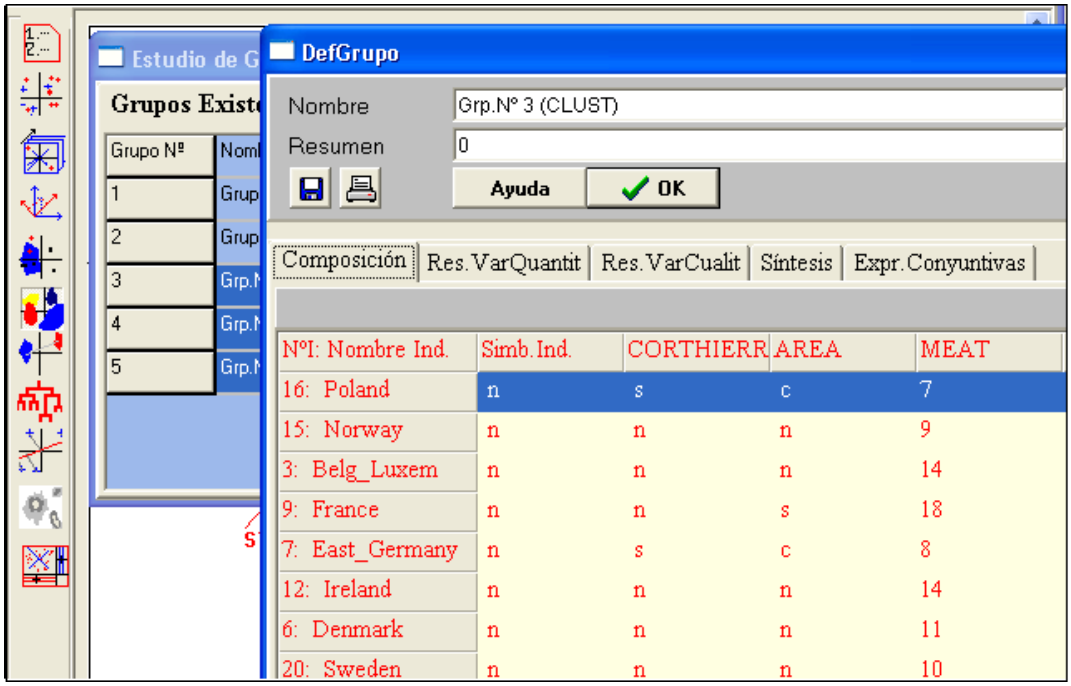


Figura 6.10.3.1. Ventana para estudio del grupo Nº 3.

Repitiendo este procedimiento para cada uno de los grupos seleccionados, al final, la situación está en la **figura 6.10.3.2.**

En esa figura se observa que los grupos 3, 4, 5 han sido pintados sobre el biplot.

Este ejemplo muestra que podemos ver en el biplot actual grupos obtenidos en instantes y por métodos distintos y comparar esos grupos visualmente.

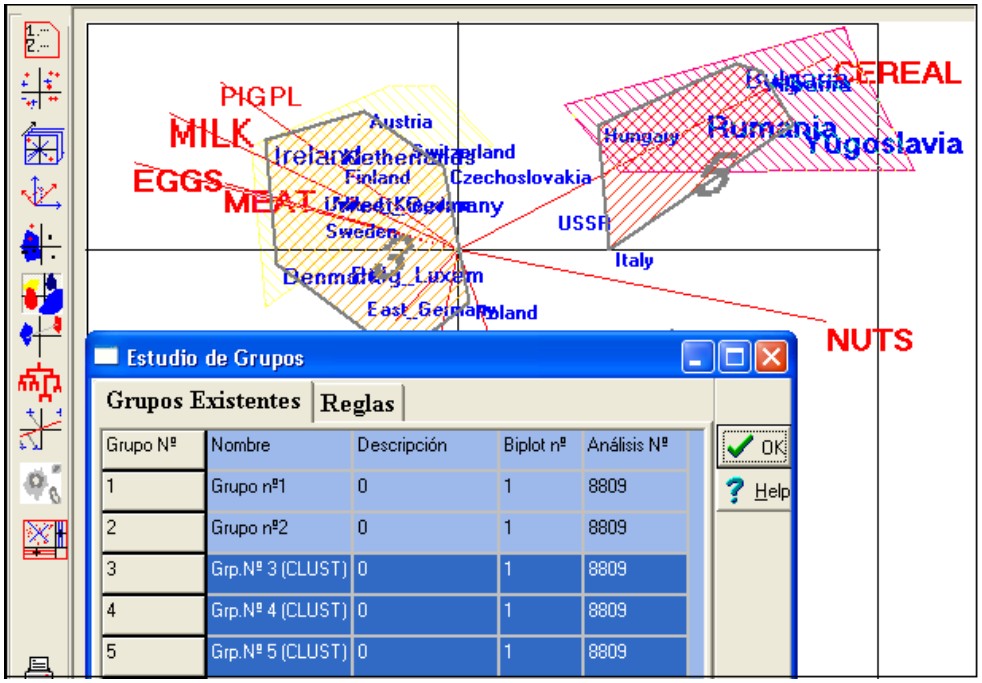


Figura 6.10.3.2. Inspección de la partición generada automáticamente: grupos 3, 4, 5.

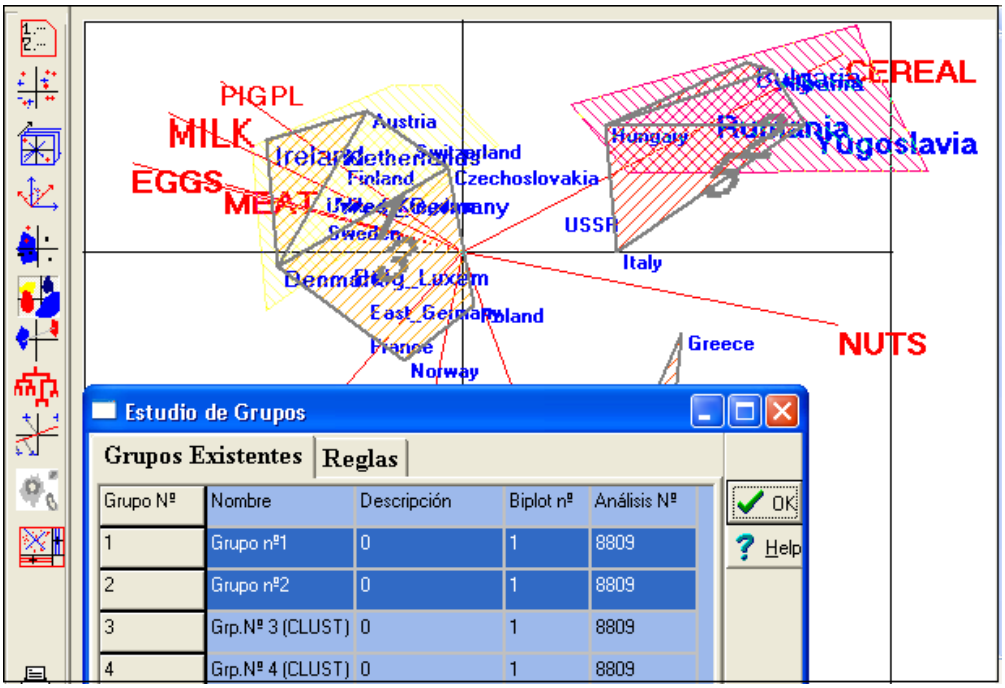


Figura 6.10.3.3. Sobreponer al biplot actual - donde están pintados los grupos 3, 4, 5 - los cierres convexos de los grupos 1 y 2, obtenidos por el usuario.

El sistema permite, también, generar reglas de clasificación, conforme lo que se ha explicado en el apartado 4.7 del **capítulo IV**.

Esas reglas son presentadas como se ilustra en la **figura 6.10.3.4.**

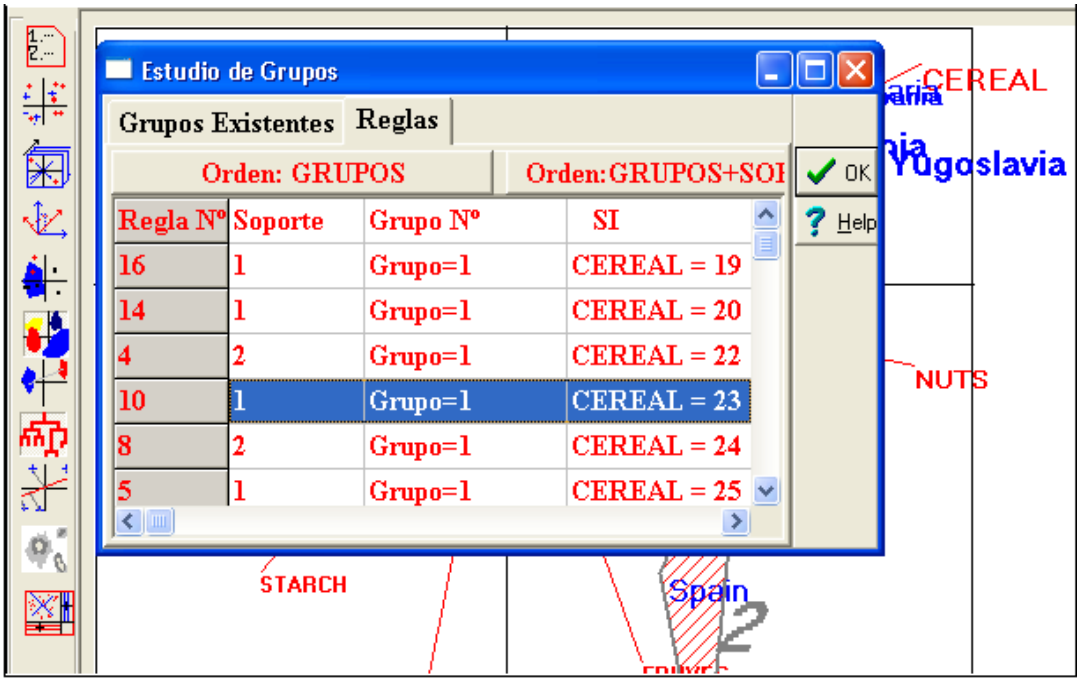


Figura 6.10.3.4. Ventana con las reglas. El ejemplo indicado resulta de datos de tipo continuo, aún no discreteado.

En la **figura 6.10.3.4.**, cada «regla» tiene la estructura siguiente:

Si $(X= x) \wedge (Y= y) \wedge ... \wedge (Z= z)$

Entonces

(GRUPO= *Gg*)

En el ejemplo, la primera fila (Regla Nº 16) "dice":

Si (CEREAL= 19)

Entonces

GRUPO= Nº 1

Cada fila relativa al mismo grupo - grupo 1, por ejemplo - corresponde a una caracterización alternativa de ese grupo.

Entonces, lo que la ventana nos está diciendo es que

Si CEREAL ∈ {19, 20, 22, 23, ..., 36}
Entonces
GRUPO= Nº 1

El ejemplo anterior ha sido obtenido con variables continuas, antes de transformadas en variables discretas lo que produce una multitud de reglas con soportes muy pequeños (1 o 2 individuos), dificultando la interpretación.

6.11. ESTUDIO DE GRUPOS POR PROYECCIÓN.

6.11.1. PROYECCIÓN SEGÚN UNA DIRECCIÓN.

Con este sistema experimental se busca realizar las ideas que han sido explicadas en el **capítulo V**, apartado 5.5.3.

Ejemplo

Consideremos el biplot de los datos GABRIEL (1981) en el que los países han sido divididos en tres grupos. Ver la **figura 6.11.1.1.**

Supongamos que interesa mirar los datos según la perspectiva de lo que separa el grupo 1 del grupo 3.

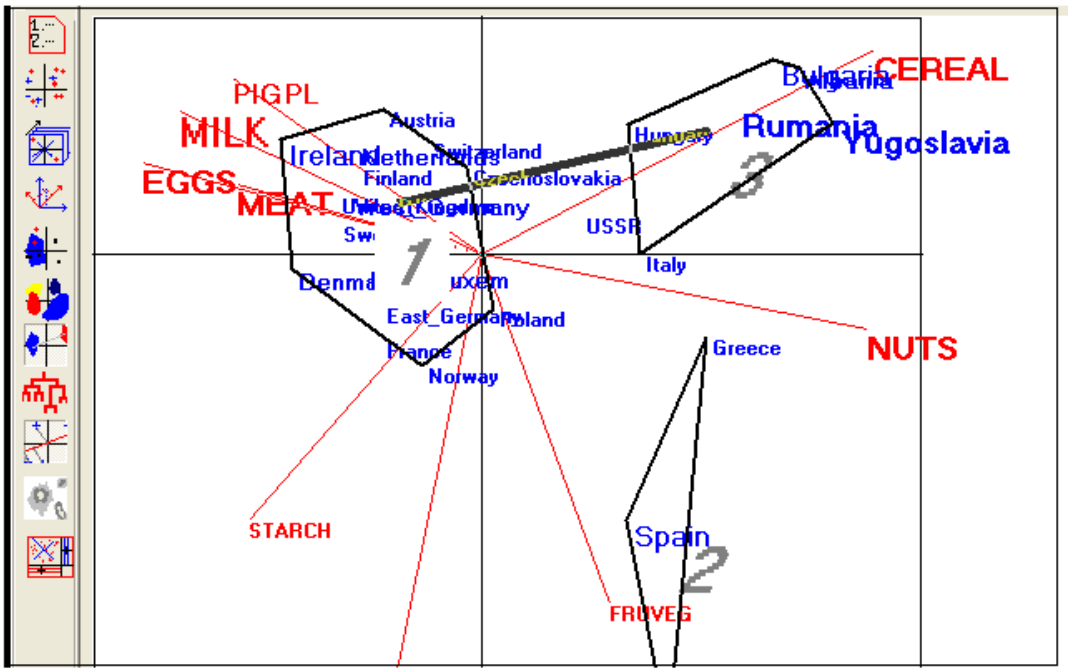


Figura 6.11.1.1. Los centros de los grupos 1 y 3 definen la dirección de proyección.

Interesa considerar la dirección definida por los dos centros de gravedad de los grupos 1 y 3 y seguidamente proyectar las variables y los individuos sobre esa dirección.

En la **figura 6.11.1.1.** se puede ver una dirección aproximada a la definida por los centros de los grupos 1 y 3.

El resultado de proyectar los individuos según esa dirección puede verse en la **figura 6.11.1.2.**

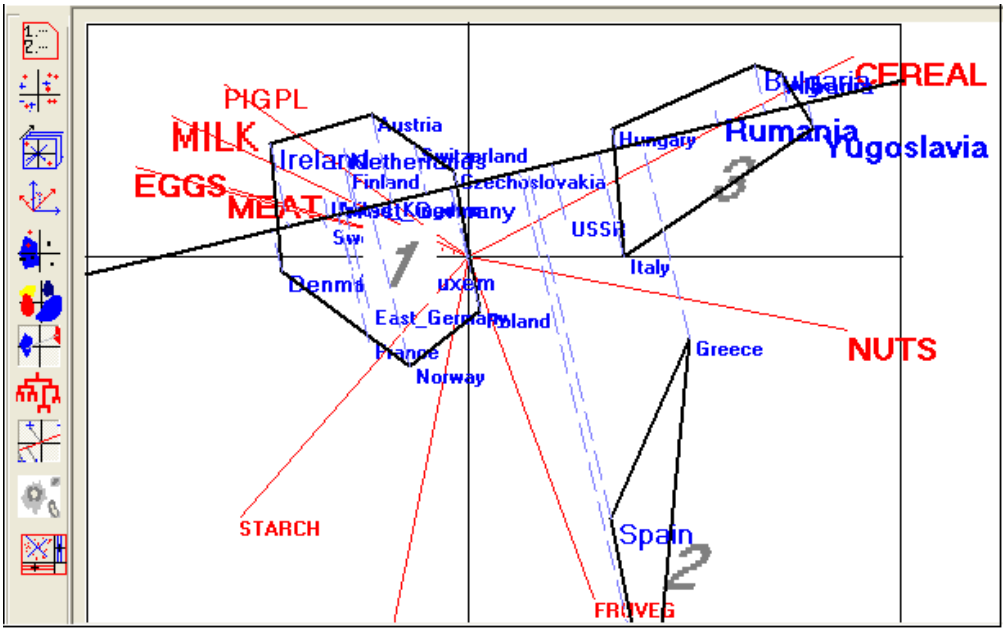


Figura 6.11.1.2. Los individuos (países) han sido proyectados ortogonalmente sobre la dirección definida por los centros de los dos grupos.

En las opciones globales - botón Op. - se puede decidir, en el inicio de la sesión, lo que proyectar: individuos, variables o individuos y variables.

6.11.2. INTERPRETACIÓN DE LAS PROYECCIONES DE LOS INDIVIDUOS.

Al proyectar los individuos según una dirección, estos quedan ordenados según sus proyecciones sobre la dirección. Ver **capítulo V**, apartado 5.5.3.

Mirando la **figura 6.11.1.2.** se puede ver que el orden definido por la dirección de los centros de los grupos 1 y 3 es:

Denmark, Ireland, Sweden, ..., URSS, ..., Italy, ..., Bulgaria, Albania.

Esto significa que hemos obtenido «una perspectiva de los datos - de los individuos - según lo que separa los grupos 1 y 3».

En el sistema esa perspectiva es materializada matemáticamente calculando las distancias de las proyecciones (de los individuos o de las variables) hasta la proyección del centro del gráfico sobre esa misma recta.

Estos valores son incorporados en el fichero de datos en análisis, creando una nueva variable - la variable de trabajo:VarTrab1.

Esta variable puede ser inspeccionada en la parte izquierda de la pantalla. Ver la **figura 6.11.2.1.**

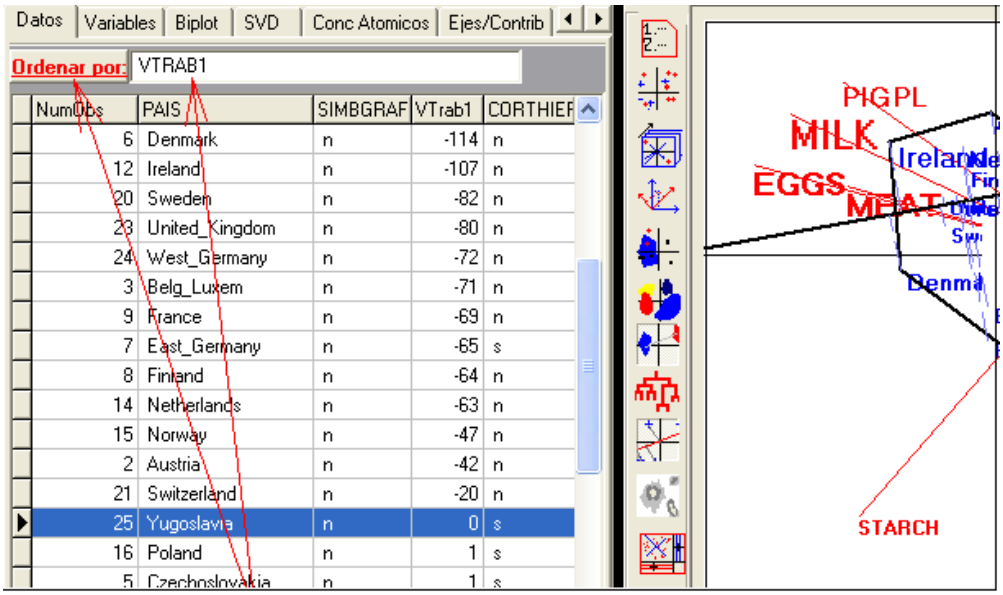
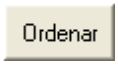


Figura 6.11.2.1. Las flechas apuntan el modo de ordenar los datos según los valores de las proyecciones, almacenados en la variable Vtrab1.

Si en la zona apuntada escribimos el nombre de esa variable, al apretar el botón 

los datos aparecen ordenados según los valores de las proyecciones (los valores negativos y positivos significan que esas proyecciones están de un lado u otro de la proyección del centro del gráfico).

Una vez ordenados los datos según las proyecciones, ese orden se refleja en todas las variables (columnas del paquete de datos).

El examen de las variables ordenadas según ese orden puede revelar patrones interesantes para la interpretación de «lo que separa un grupo de otro».

Por ejemplo, examinando ahora los valores de MEAT y CEREAL - vea la **figura 6.11.2.2.** - se observa que «caminando del grupo 1 al grupo 3 a lo largo de la dirección definida, el porcentaje de proteínas con origen MEAT tiende a disminuir y tiende a aumentar el porcentaje de proteínas con origen en CEREAL».

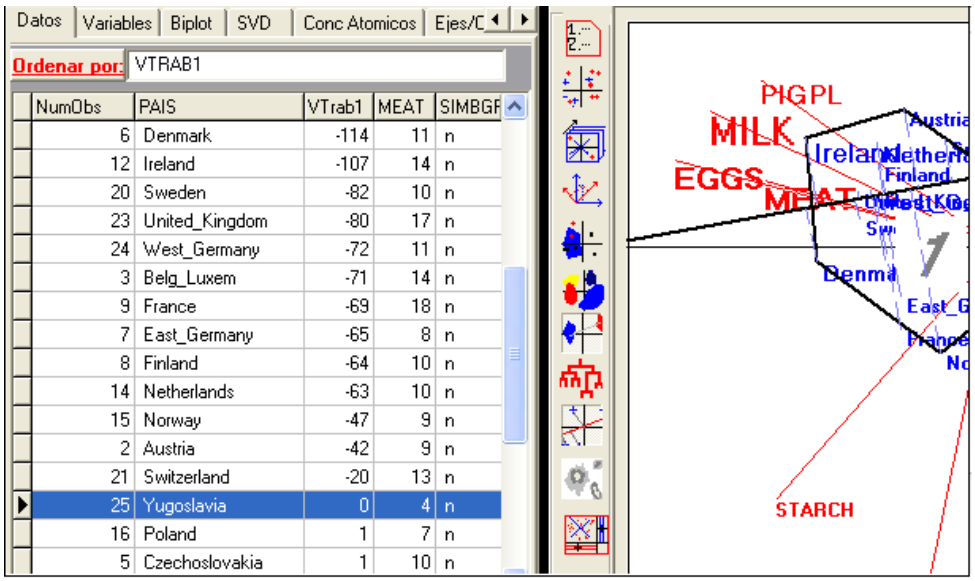


Figura 6.11.2.2. Ordenando los datos por las proyecciones (VTrab1) se observa que los valores MEAT disminuyen y los valores CEREAL aumentan.

Calculando las correlaciones entre estas proyecciones y cada una de las variables cuantitativas podemos investigar cuales son las variables que están más asociadas (linealmente) o son más relevantes para interpretar la dirección.

Esas correlaciones pueden verse consultando la hoja Variables de los datos. Vea la **figura 6.11.2.3.** Los valores de las correlaciones calculadas según este método reflejan las observaciones hechas anteriormente y su interpretación debe atender a las consideraciones del apartado 5.5.3.

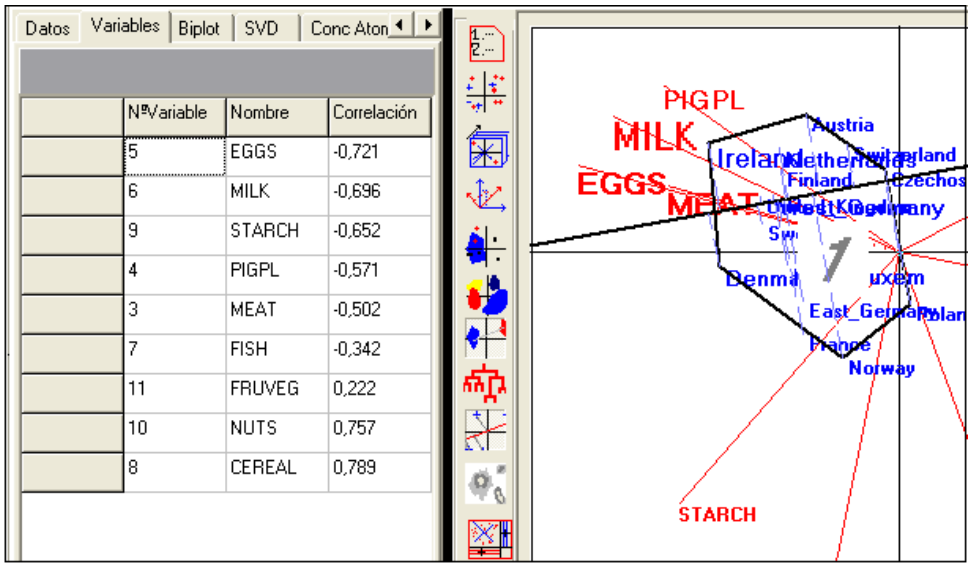


Figura 6.11.2.3. Correlaciones de las variables numéricas con las proyecciones.

6.12. INTERACCIÓN CON LOS DATOS.

6.12.1. VISUALIZACIÓN DE LOS DATOS MARCADOS.

En este sistema prototipo se ha buscado responder a algunos de los problemas identificados en el **capítulo V**, apartado 5.5.1. cuando el gráfico se refiere a un número elevado de individuos.

En particular, se ha buscado responder a dos tipos de problemas:

1. ¿Dónde está, en el gráfico, una observación especifica o un conjunto de observaciones que corresponde, en el conjunto de datos, a un criterio considerado interesante para el usuario?.
2. Si definimos, por una expresión conjuntiva, un conjunto dado de individuos, ¿dónde se ubica en el gráfico, el conjunto de individuos que satisface esa expresión?.

Para responder a estos dos problemas hemos desarrollado un método visual que se presenta en los apartados siguientes. Una vez más, los algoritmos funcionan sobre el grafo de intersección definido en el **capítulo IV**.

Cuando el número de individuos es pequeño, toda la información está «a la vista»; cuando el número de individuos es muy grande resulta difícil identificar un individuo o un conjunto de individuos sobre el gráfico del biplot.

El programa permite - usando el botón  - ubicar sobre el gráfico todas las observaciones que el usuario marque en el conjunto de datos.

En el lado izquierdo de la **figura 6.12.1.1.**, puede verse que han sido marcados los países Czechoslovakia, Greece, Hungary; esas observaciones aparecen representadas sobre el gráfico por el símbolo .

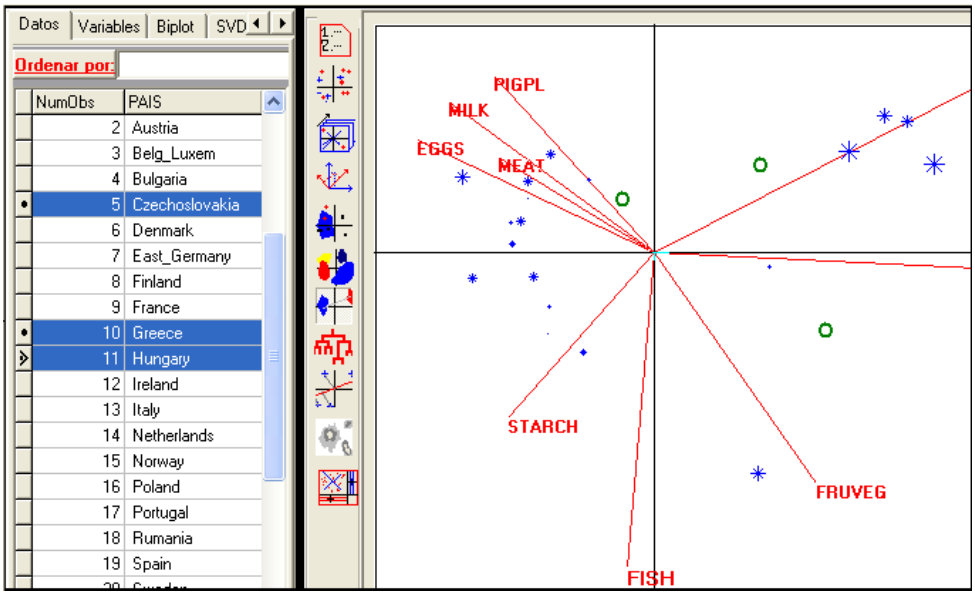


Figura 6.12.1.1. Después de marcar los individuos sobre la tabla de datos a la izquierda, esos individuos quedan representados por círculos verdes sobre el biplot.

6.12.2. VER LOS INDIVIDUOS ORDENADOS SEGÚN UN CRITERIO.

En muchas situaciones interesa verificar si los individuos ordenados según un criterio dado de ordenación están concentrados en una región del biplot o, por el contrario, no existe ninguna tendencia espacial asociada a esa ordenación.

El sistema prototipo permite ordenar los datos por un criterio arbitrario construido usando los nombres de las variables separadas por ";". Ver la parte superior izquierda de la **figura 6.12.2.1.**

Por ejemplo: para ordenar el paquete de datos según el criterio «MILK; MEAT», como en el ejemplo presentado en esa figura.

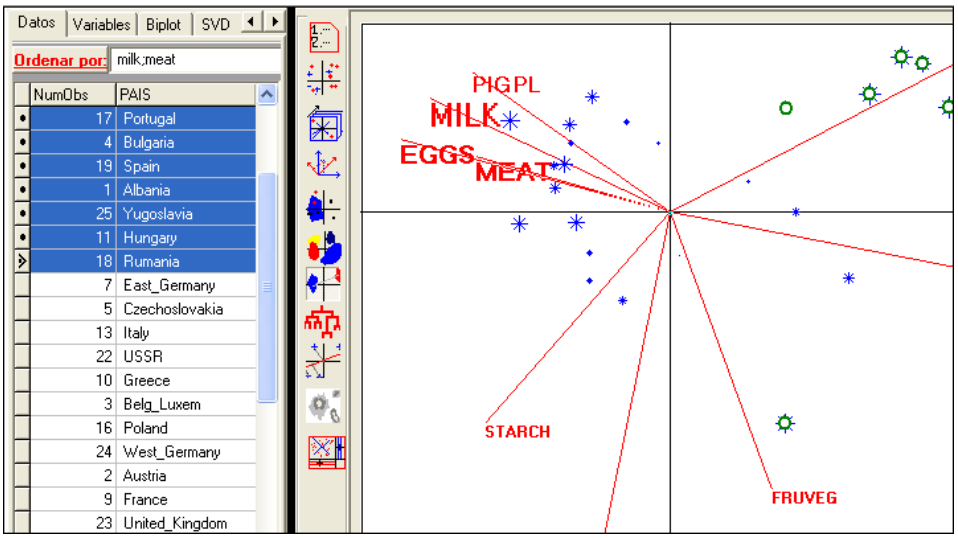


Figura 6.12.2.1. Después de ordenar los datos por MILK y MEAT, se han marcado aquellos países en que estos valores son más pequeños. Esos países pueden pintarse en el biplot círculos verdes.

Combinando los criterios de ordenación con la marcación de los individuos, pueden ser visualizadas relaciones potencialmente «interesantes» sobre el gráfico.

6.12.3. DEFINICIÓN Y VISUALIZACIÓN DE *QUERIES*.

El sistema permite definir expresiones de tipo conjuntivo con las variables y visualizar sobre el biplot el conjunto de individuos que satisfacen esas expresiones.

Para construir las expresiones se usan los átomos actuales – expresiones de tipo (*X*= valor). Ver **capítulo IV**, apartado 4.4.

Los átomos actuales pueden resultar de los átomos iniciales (construidos con los valores observados) después de agrupar esos valores en categorías más apropiadas a la fase de interpretación. Ver en el sistema de Ayuda la descripción del botón  .

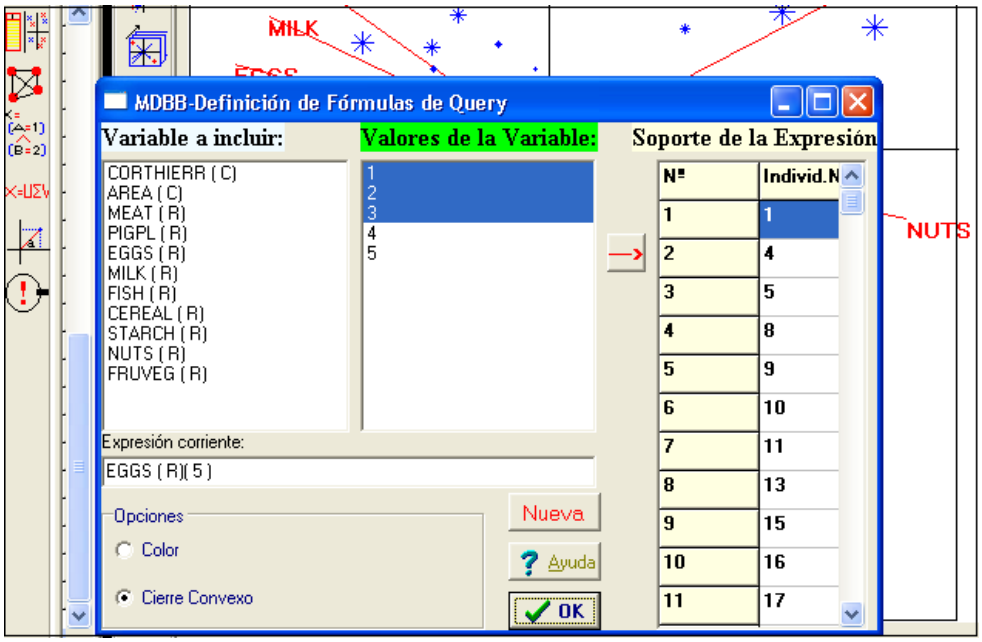


Figura 6.12.3.1. Definir la expresión (EGGS ≤ 3)= (EGGS= 1) ∨ (EGGS= 2) ∨ (EGGS= 3).

En la **figura 6.12.3.1.** se presenta un ejemplo de creación visual de la expresión (EGGS ≤ 3).

Dado que podemos marcar zonas no contiguas del conjunto de valores de una variable, eso significa que podemos construir - usando los átomos del grafo de intersección - expresiones del tipo:

$$X \in C,$$

en donde C es un conjunto arbitrario de valores.

Basta con atender a que:

$$X \in C = \bigcup_{c \in C} (X = c)$$

en que $(X = c)$ es un átomo.

Estas expresiones pueden ser representadas en el biplot por sus cierres convexos o también pintando los marcadores de los individuos que forman su significado, como en la **figura 6.12.3.2.**

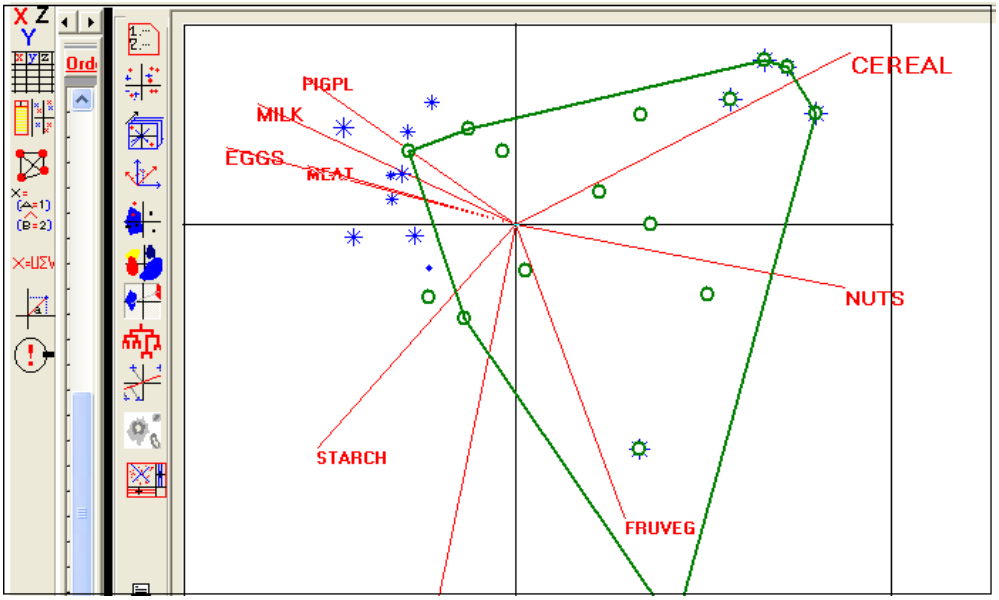


Figura 6.12.3.2. Cierre convexo de la expresión (EGGS ≤ 3).

6.13. ESTUDIO DE LOS ELEMENTOS SUPLEMENTARIOS.

6.13.1. DEFINICIÓN DE LOS ELEMENTOS SUPLEMENTARIOS.

El sistema prototipo implementa las ideas explicadas en el **capítulo V**, apartado 5.5.5.

La definición de elementos suplementarios - individuos y variables - permite introducir información exterior en un biplot.

La información exterior puede ser información nueva acerca de los individuos (variables suplementarias) o información resultante de observar nuevos individuos usando las variables anteriores.

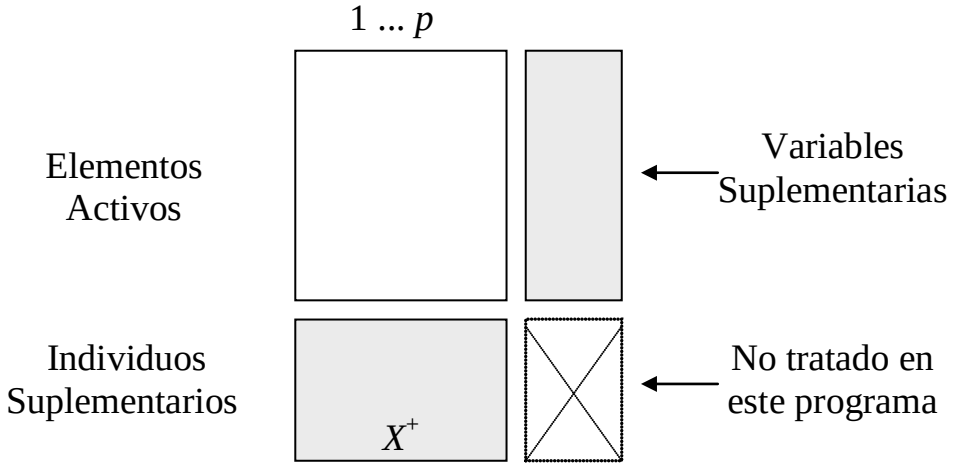
Los individuos y las variables usadas para crear el biplot actual se designan **individuos activos** y **variables activas**.

Se asume que toda la información existente ha sido integrada en el conjunto de datos brutos inicial - antes de editar.

En el inicio de una sesión deben ser indicados los individuos y las variables a usar para definir los biplots que se van a crear en esa sesión.

Los individuos y las variables que han sido elegidos para crear el biplot forman los **elementos activos**; los que no han sido considerados activos son, **potencialmente**, elementos suplementarios.
O sea: es entre esos elementos **no** activos entre los que el usuario debe elegir los individuos y las variables que desea tratar como suplementarios en el biplot actual.

Gráficamente:



La introducción de individuos suplementarios, en un biplot existente, es una forma de predicción: predice en que grupos, con significado fijado por la interpretación del biplot actual, se clasificarían nuevos individuos.

6.13.2. OPCIONES GRÁFICAS PARA ELEMENTOS SUPLEMENTARIOS.

Conviene que los elementos suplementarios (individuos y variables) se distingan de los activos por símbolos y colores especiales.

Estos son elegidos en el lado derecho de la ventana que se presenta en la **figura 6.13.2.1.**

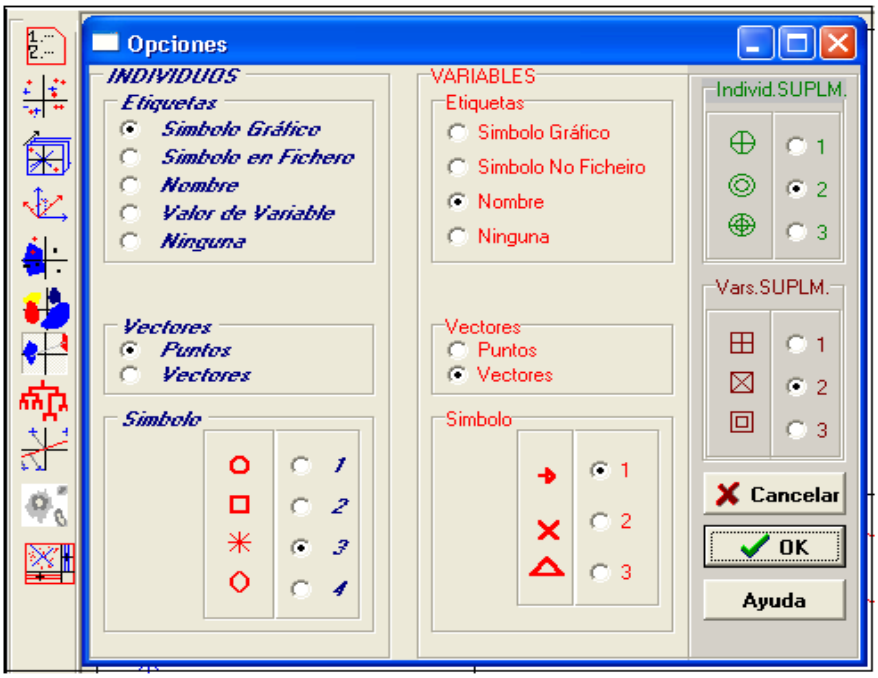


Figura 6.13.2.1. Definir los símbolos gráficos a usar para las variables y los individuos suplementarios.

Los símbolos predefinidos son los que están marcados en la **figura 6.13.2.1.** (y , para individuos y variables, respectivamente).

6.13.3. PINTAR LOS ELEMENTOS SUPLEMENTARIOS.

Todos los elementos no-activos son, potencialmente, elementos suplementarios - individuos o variables.

Una vez construido el biplot, es necesario elegir, entre esos elementos (variables e individuos) no-activos, cuales son los individuos suplementarios y las variables suplementarias que se pretende pintar sobre el biplot actual.

Esto se hace en la ventana que se puede ver en la **figura 6.13.3.1.**

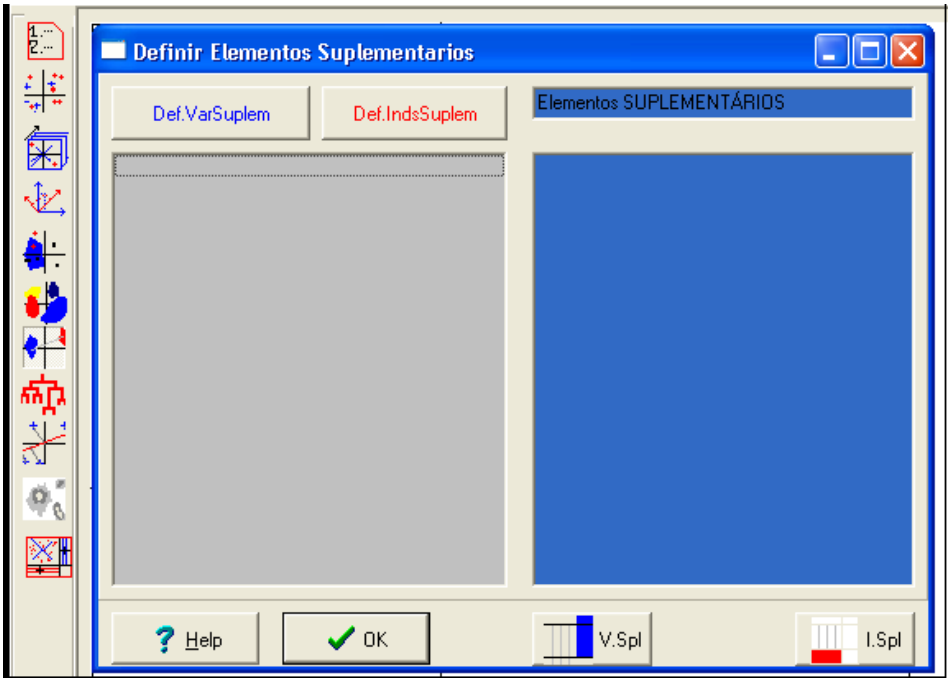


Figura 6.13.3.1. Definir y pintar los elementos suplementarios.

Por ejemplo, si en el fichero PROTEÍNAS elegimos como variables por analizar (activas) las variables:

1: CORTHIERR
2: AREA
3: MEAT
4: PIGPL

5: EGGS
6: MILK
7: FISH

entonces quedan como potencialmente suplementarias las variables

8: CEREAL
9: STRARCH
10: NUTS
11: FRUVEG

Si elegimos como individuos activos los individuos (1-ALBANIA, ..., 20-SWEDEN) quedan como individuos potencialmente suplementarios los siguientes:

21: SWITZERLAND
22: USSR
23: UNITED-KINGDOM
24: WET-GERMANY
25: YUGOSLAVIA

Una vez creado el biplot, el usuario puede elegir entre los elementos (variables e individuos) no-activos, aquellos que desea ver pintados sobre el biplot actual. Esto se realiza en la ventana que se ve en la **figura 6.13.3.2.**

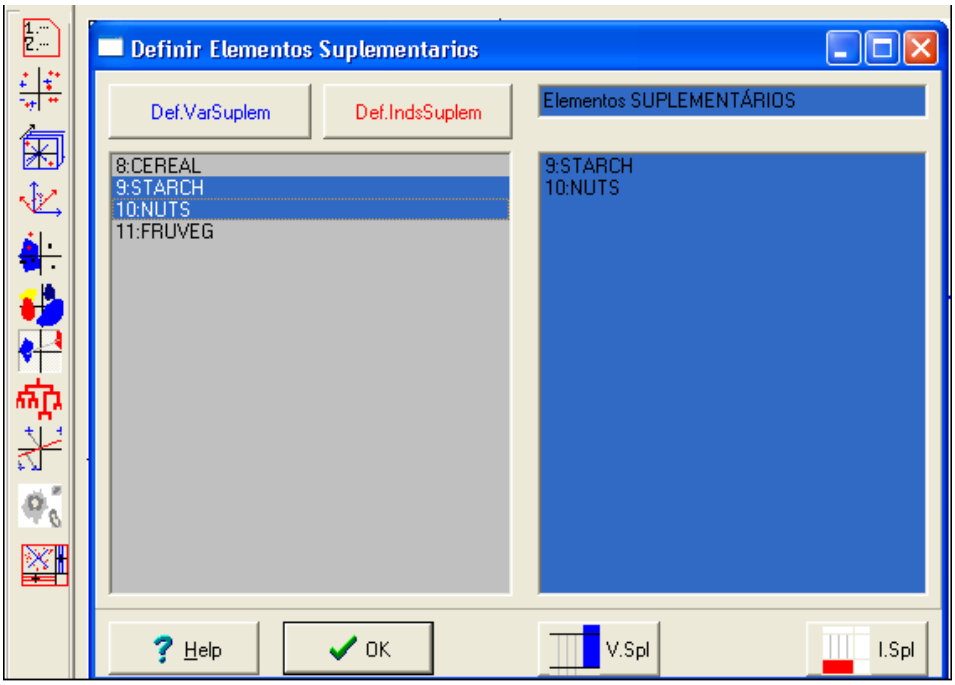


Figura 6.13.3.2. Las variables para pintar como suplementarias son 9: STRACH y 10: NUTS, elegidos entre los no-activos.

6.14. ESTUDIO DE LOS GRÁFICOS DE DENSIDAD.

6.14.1. CREACIÓN DE UN GRÁFICO DE DENSIDAD.

Este gráfico interesa cuando el número de individuos es grande o muy grande. Ver **capítulo V**, apartado 5.4.

El gráfico representa la estimación de la densidad bivariante en cada uno de los 50 × 50 cuadrados de un reticulado en el que se divide la pantalla.

Esta estimación es calculada usando un estimador de tipo KERNEL, con KERNEL NORMAL. Ver SCOTT (1992).

En la estimación de esta densidad se usan solamente las contribuciones de los marcadores de los individuos.

El resultado para los 25 países del paquete de datos PROTEÍNAS está en la **figura 6.14.1.1.**

Figura 6.14.1.1. Gráfico de densidad para los 25 países del paquete PROTEÍNAS.

6.14.2. INTERPRETACIÓN DEL GRÁFICO DE DENSIDAD.

Sobre el gráfico de densidad pueden ser realizadas las operaciones básicas de identificación, caracterización y comparación de grupos, siendo el criterio básico de agrupamiento la estimación de densidad en cada punto, presentada como una tonalidad de gris.

Por ejemplo, en la **figura 6.14.2.1.** han sido identificados tres grupos que el sistema permite caracterizar y comparar.

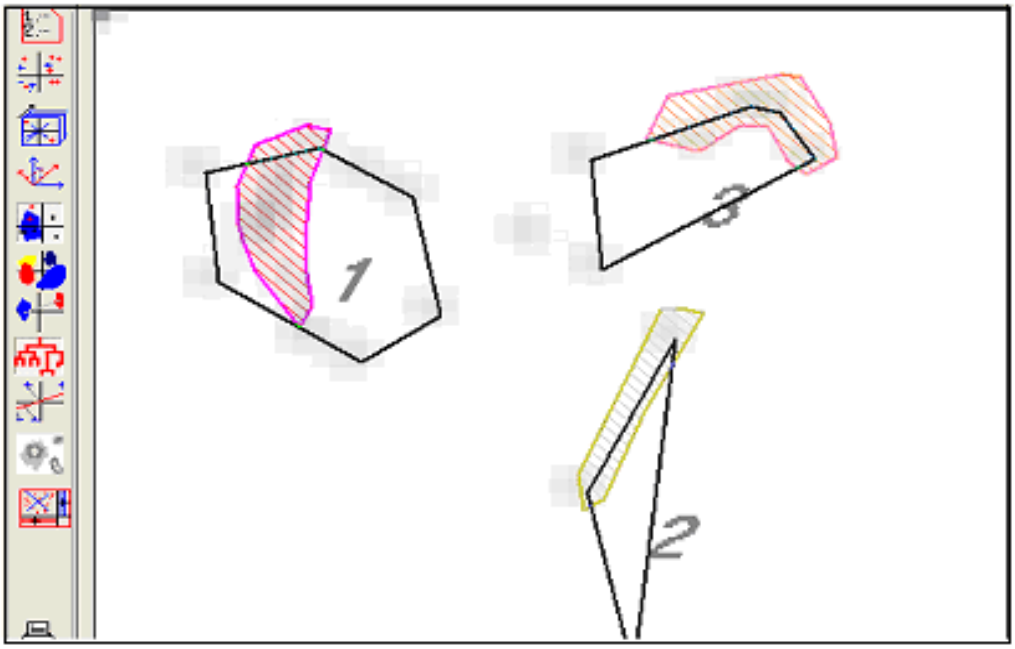


Figura 6.14.2.1. Tres grupos definidos conectando rectángulos con la misma tonalidad de gris.

Un criterio natural para definir grupos en el gráfico de densidad, es marcar el polígono representante del grupo, usando como vértices a puntos de iso-densidad que aquí se presentan con la misma tonalidad de gris.

El grupo así definido contiene todos los individuos que más han contribuido para los valores de la densidad/concentración dentro de ese polígono.

6.15. INTERPRETACIÓN DE LOS EJES FACTORIALES.

6.15.1. INTRODUCCIÓN.

El sistema prototipo suministra varias posibilidades de interpretación de ejes y planos factoriales:

a) Exámen de los resultados de la descomposición del paquete de datos en valores y vectores singulares ($X= U \Sigma V^T$).
b) Caracterización de los grupos de individuos y de variables que se proyectan en las partes extremas de los ejes factoriales, buscando lo que opone esos grupos más extremos.
c) Ordenación de los individuos y variables según sus proyecciones sobre un eje o los dos ejes que definen un plano factorial.
d) Identificación de los individuos y variables que más contribuyen para la formación del eje (contribución relativa de los elementos para los ejes) ($CRE_{(i)}F_{(\alpha)}$).
e) Identificación de los individuos y las variables a que corresponden mayores contribuciones de los ejes para esos individuos (contribuciones relativas de los factores para los elementos ($CRF_{(\alpha)}E_{(i)}$).

En los números **6.15.2.** a **6.15.5.** se explica como en este sistema experimental se obtienen esos resultados.

6.15.2. DECOMPOSICIÓN EN VALORES Y VECTORES SINGULARES.

Apretando el botón X=UΣV el sistema presenta en la hoja SVD el resultado de la descomposición de la matriz de datos (relativa a las variables activas cuantitativas observadas sobre los individuos activos.

El resultado puede verse en la parte izquierda de la **figura 6.15.2.1.**

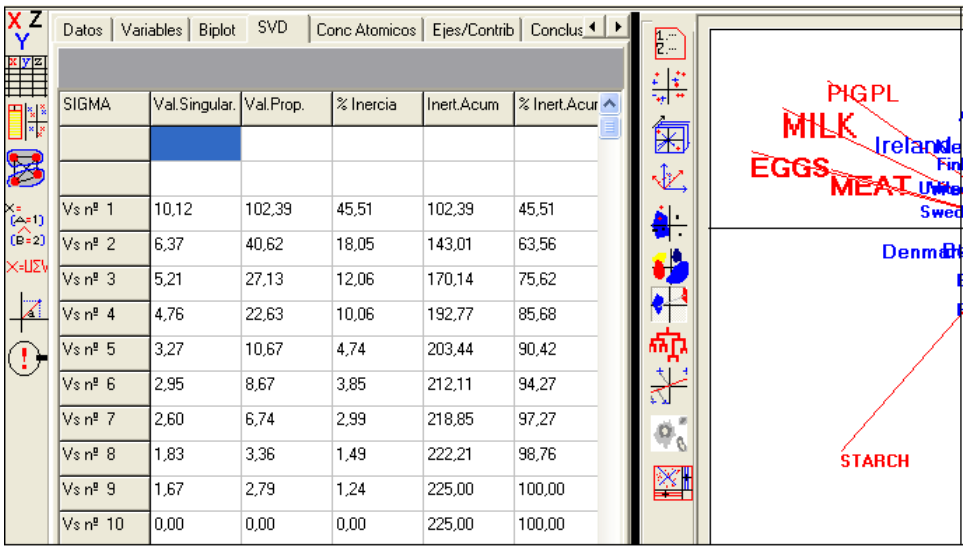


Figura 6.15.2.1. Valores singulares (σ_i) valor propio (λ_i σ_i^2) y porcentaje acumulado de inercia de la descomposición SVD.

En esa figura se puede verificar que los primeros dos ejes factoriales explican 63.5% de la varianza total (información o inercia del paquete de datos).

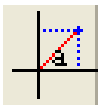
Para obtener el 75.62% de la información serían necesarios los tres primeros ejes factoriales y el 100% de la información exige 8 ejes de un total de 11 posibles.

Examinando la parte restante de esta hoja, podría verificarse que el sistema presenta, también, los vectores propios U y V de la descomposición, correspondientes a los vectores singulares no nulos y las coordenadas de los marcadores en el biplot de GALINDO:

$J = A = U\Sigma$ (Marcadores de los individuos)

$H = B = V\Sigma$ (Marcadores de las variables)

6.15.3. INTERPRETACIÓN DE UN EJE FACTORIAL EXAMINANDO LAS PROYECCIONES DE LOS ELEMENTOS SOBRE EL EJE.

 Marcando una dirección paralela al eje horizontal se obtienen las proyecciones de los individuos (o de las variables) sobre esa dirección. Ver **figura 6.15.3.1 - Proyección de los individuos** y **figura 6.15.3.2. - Proyección de las variables**.

Se puede observar en la **figura 6.15.3.1.** que los países quedan ordenados según el orden:

Ireland, Denmenark, Sweden, Finland, ..., Spain, Italy, Greece, ..., Rumania, ..., Yugoslavia.

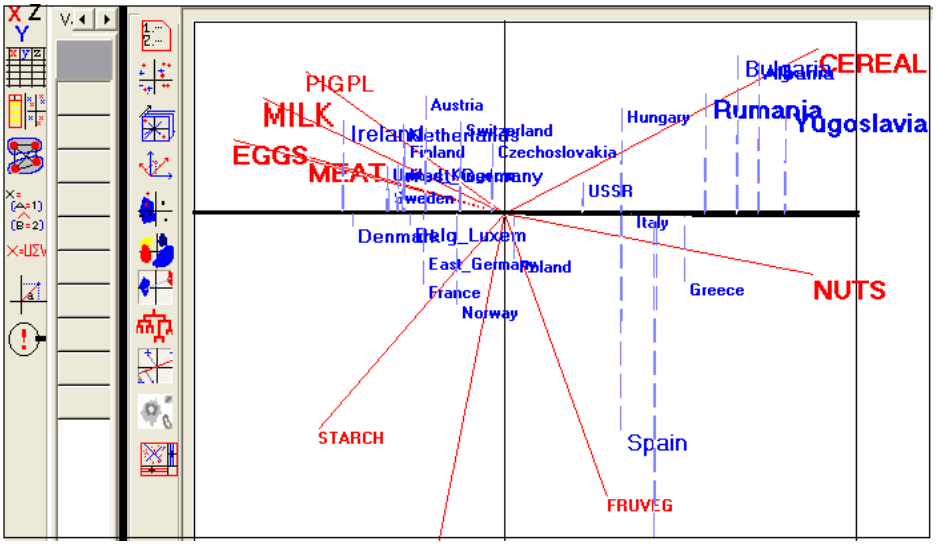


Figura 6.15.3.1. Proyección de los individuos (países) sobre el eje factorial nº 1 - Horizontal (α= 45.5%).

Proyectando ahora las variables (ver **figura 6.15.3.2.**) se verifica (sentido: izquierda - derecha) que el orden de las variables es:

MILK, EGGS, MEAT, ..., FRUVEG, NUTS, CEREAL.

El eje factorial (primer eje factorial) opone, por eso, los países occidentales en donde el origen de las proteínas era predominantemente animal, a los países continentales y del este, en donde el origen de las proteínas era predominantemente vegetal.

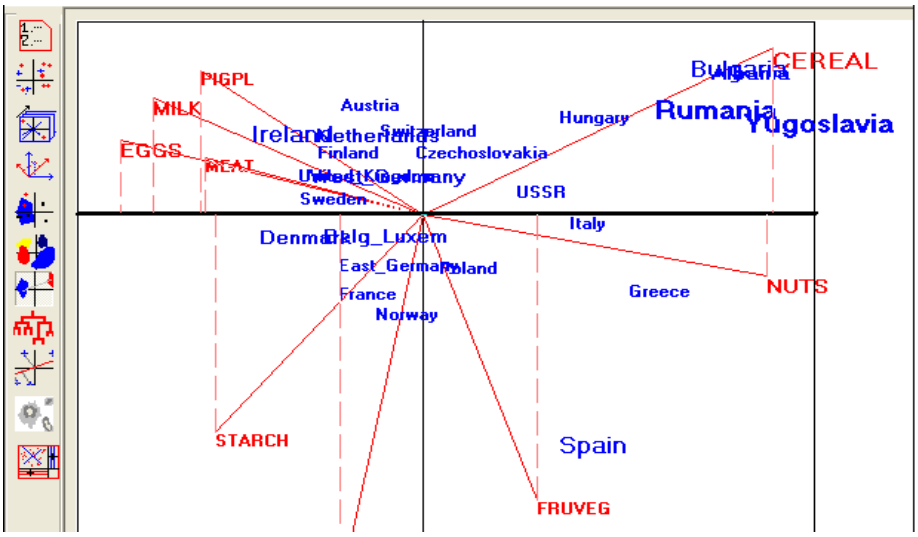


Figura 6.15.3.2. Proyección de las variables sobre el eje factorial nº 2 (λ= 45.5%).

Realizando el mismo ejercicio para el eje vertical (2º eje factorial) se verifica, ahora, la oposición entre una dieta mediterránea (Portugal, Spain, Greece), en donde se verifica que muchas de las proteínas provienen del pescado, de las frutas y vegetales, y de una dieta mixta de cereales y productos animales en donde el pescado y las frutas y vegetales tienen menor expresión.

6.15.4. INTERPRETACIÓN DE LOS EJES EXAMINANDO LAS OPOSICIONES ENTRE LOS PRINCIPALES GRUPOS.

Si realizamos un análisis cluster sobre los marcadores de un biplot de GALINDO, obtenemos la **figura 6.15.4.1.**

En esa figura detectamos tres grandes grupos de países

1. {Ireland, Netherlans, Sweden, ...}
2. {Hungary, Rumania, Bulgaria, Albania, Yugoslavia, ...}

3. {Greece, Spain, Portugal}

Se verifica que el grupo 1 (países Occidentales no pertenecientes a la Cortina de Hierro, en donde el origen de las proteínas era predominantemente animal) se opone, a lo largo del eje nº 1, a los restantes países {Portugal, Spain, Greece, ... , Bulgaria, Albania} en donde las proteínas provenían predominantemente de los vegetales.

El eje vertical opone claramente los países con «dieta mediterránea» {Portugal, Spain, Greece} a los restantes.

Esta dieta mediterránea se caracteriza por el predominio de los vegetales y el pescado en las proteínas consumidas, por oposición a una dieta más continental, asociada a los cereales y productos animales como PigPL, Milk, Eggs.

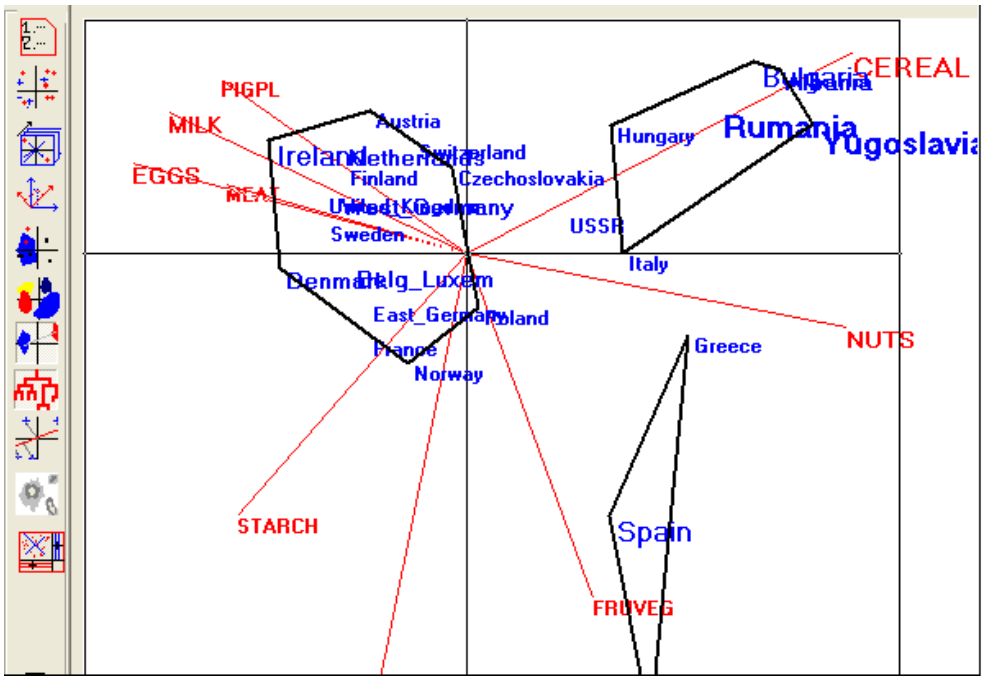


Figura 6.15.4.1. El eje horizontal significa la oposición entre los países consumidores de proteínas de origen animal a los consumidores de proteínas vegetales. El eje vertical destaca la dieta mediterránea (predominancia de frutas, vegetales y pescado) al resto.

6.15.5. INTERPRETACIÓN DE UN EJE USANDO LAS CONTRIBUICIONES RELATIVAS.

El sistema prototipo realiza los cálculos especificados en el apartado 2.4.3. del **capítulo II**.

La ventana de la **figura 6.15.5.1.** da una idea de la interactividad implementada.

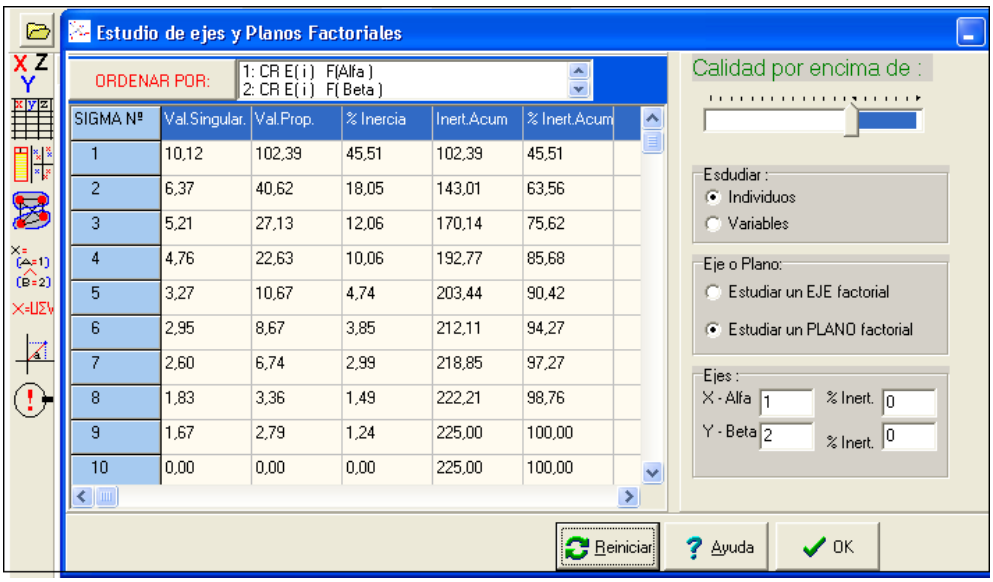


Figura 6.15.5.1. Ventana para estudiar las contribuciones de los ejes para los elementos (variables y individuos) y de los elementos para los ejes.

En esta ventana se puede elegir y estudiar un eje o un plano definido por los ejes *X* (horizontal) e *Y* (vertical).

También se puede elegir entre estudiar las variables y los individuos.

Ver, en la **figura 6.15.5.2.** el estudio de las variables.

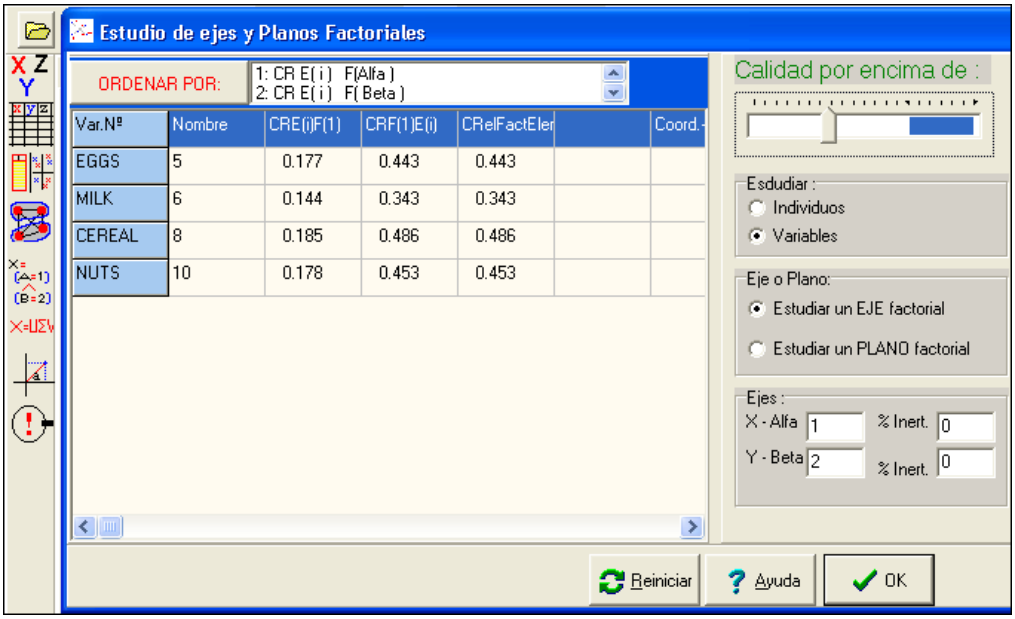


Figura 6.15.5.2. Estudio de las contribuciones relativas de las variables para el eje (factor) nº 1 y del factor nº 1 para las variables.

En el ejemplo de esa figura sólo pueden verse las variables cuya calidad de representación (contribución del factor para las variables) está por encima de 0.5.

Esta condición es regulada dinámicamente por el cursor del margen superior derecho: dislocando ese cursor hacia la derecha o hacia la izquierda altera la lista de variables (o individuos) cuya calidad de representación sobrepasa el límite establecido por el cursor.

El botón Ordenar por: permite ordenar los individuos o las variables presentes en la ventana según el criterio elegido en la lista de opciones a la derecha de ese botón.

Los criterios son, usando la notación del apartado 2.4.3. los siguientes:

1, 2 - $CRE_{(i)}F_{(1)}$ o $CRE_{(i)}F_{(2)}$

Contribuciones relativas de los elementos (individuos y variables) para el factor número 1 ó número 2.

3, 4 - $CRF_{(1)}E_{(i)}$ o $CR\ F_{(2)}E_{(i)}$

Contribuciones relativas del factor 1 ó 2 para el elemento *i* (individuo ó variable).

5 - CALIDAD REP. PLAN.

Calidad de la representación de un elemento (variable o individuo) en el plano definido por los ejes *X* y *Y* a los que están asociados los factores α y β.

En la **figura 6.15.5.3.**, las variables con calidad de representación por encima del valor definido por el cursor, están ordenadas según el criterio CRE (*i*) *F* (Alfa).

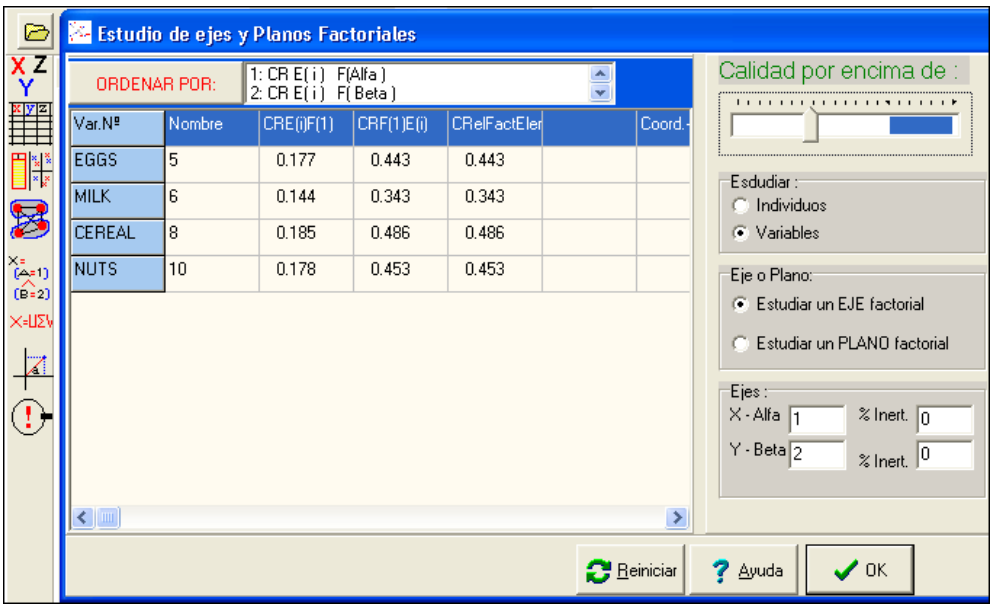


Figura 6.15.5.3. Contribuciones relativas de los elementos para los ejes Alfa y Beta y de los ejes Alfa y Beta para los elementos. Calidad de representación de los elementos en el plano (Alfa, Beta).

C A P Í T U L O VII

APLICACIÓN A DATOS REALES

7.1. INTRODUCCIÓN.

El objetivo de este capítulo es el de ilustrar la utilización del sistema-prototipo descrito en el **capítulo VI** y cuyo programa ejecutable y manual del usuario forman el Anexo de esta tesis.

Recordemos que el núcleo teórico esencial de nuestro trabajo - más allá de mostrar que los métodos de biplot son adecuados a las tareas de minería de datos - es probar que es posible formular matemáticamente, de modo no trivial, el problema de interpretación de resultados de análisis de datos multivariantes (ver **capítulo IV**). En esa perspectiva, se especifica que en el presente capítulo nos proponemos mostrar que las soluciones teóricas y prácticas desarrolladas en el **capítulo IV** e ilustradas con los pequeños ejemplos del texto, funcionan también cuando son aplicadas a datos más cercanos a los que ocurren en el «mundo real».

Mostraremos que, efectivamente, el programa realiza las ideas de interpretación del **capítulo IV** (generación de sugerencias de interpretación, construidas de modo automático) y forma un entorno interactivo adecuado a la minería de datos centrada en los métodos biplot.

No es posible ilustrar de modo estático, usando las páginas de esta tesis, todas las posibilidades del programa. Lo que a continuación se presenta resulta de una selección dictada por la naturaleza de los datos, las necesidades de ilustración de esta tesis y las limitaciones físicas.

7.2. LOS DATOS.

Los datos usados en esta aplicación han sido publicados por el CIS - Centro de Investigación Sociológicas e integran el estudio «Post electoral Generales y Autonómicas Andalucía 1996»., Estudio nº 2210.

El fichero original incluía las 1198 respuestas a un cuestionario, con un total de 141 cuestiones o variables.

De este fichero original hemos excluido las columnas relativas a las elecciones locales, conservando las respuestas relativas a los partidos nacionales mas importantes (PP, PSOE, IU).

Después de excluidas las respuestas con valores faltantes y de *tipo «**NS**- No Sabe*» o «***NR***- *No Responde*» en las variables numéricas, nos concentramos en un fichero con 148 respuestas a 58 cuestiones.

Esas variables y su significado están en la **tabla 7.2.1.** El símbolo, la designación, escalas y valores de esas variables son las usadas en el fichero original del CIS; eso permite relacionar fácilmente nuestras conclusiones con los datos originales.

Hemos codificado como numéricas las variables cuyo significado implica, por lo menos, la posibilidad de un ordenamiento significativo. Esas variables son clasificadas con **N** en la columna «TIPO» de la **tabla 7.2.1.**

Las restantes variables han sido consideradas cualitativas o categóricas y están clasificadas con **C** en la columna «TIPO» de esa tabla.

Si considaremos todos los valores de las 58 variables que han sido conservadas, obtenemos 400 categorías - o variables indicadoras - que forman una tabla en forma disyuntiva completa de 148 filas por 400 columnas.

SÍMBOLO	TIPO (C/N)	DESIGNACIÓN	VALORES
CUES	C	Cuestionario	
MUN	C	Municipio	
TAMUNI	N	Tamaño de habitat	(1, 2, 3)
P401	N	Calificación campaña electoral de J. AZNAR	(1 - Muy Buena ... ... 5 - Muy Mala)
P402	N	Calificación campaña electoral de F. GONZALEZ	(1 - Muy Buena ... ... 5 - Muy Mala)
P403	N	Calificación campaña electoral de J. ANGUITA	(1 - Muy Buena ... ... 5 - Muy Mala)
P11	C	Elecciones Generales 1996: Recuerdo de Voto	(1 - IU; 2 - PP; 3 - PSOE)
P12	C	Cuando decidio votar al partido que voto	(1 - Hace tiempo ... ... 4 - El mismo dia)
P13	C	Volatilidad electoral	(1 - Es la 1ª vez que vota... ... 3 - Suele votar siempre)
P13 A	C	A qué partido voto en anteriores elecciones	(1 - CDS; 2 - IU; 3 - PP; 4 - PSOE)
P14 A01	C	Motivos para votar PP	(1 ... 7)
P14 A02	C	Motivos para votar PP	(1 ... 7)
P14 B01	C	Motivos para votar IU	(1 ... 7)
P14 B02	C	Motivos para votar IU	(1 ... 7)
P14 C01	C	Motivos para votar PSOE	(1 ... 7)
P2201	N	Conocimiento y valoración de J. ANGUITA	(1 - Muy Mal ... ... 10 - Muy Bien
P2202	N	Conocimiento y valoración de J. ARDANZA	(1 - Muy Mal ... ... 10 - Muy Bien
P2203	N	Conocimiento y valoración de J. AZNAR	(1 - Muy Mal ... ... 10 - Muy Bien
P2204	N	Conocimiento y valoración de F. GONZALEZ	(1 - Muy Mal ... ... 10 - Muy Bien
P2205	N	Conocimiento y valoración de J. PUJOL	(1 - Muy Mal ... ... 10 - Muy Bien
P23A01	N	J. AZNAR: Sincero	(1 - Si , 2 - No)
P23A02	N	J. AZNAR: Creíble	(1 - Si , 2 - No)
P23A03	N	J. AZNAR: Dialogante	(1 - Si , 2 - No)

Tabla 7.2.1. Significado de las variables (1/3)

SÍMBOLO	TIPO (C/N)	DESIGNACIÓN	VALORES
P23A04	N	J. AZNAR: Preparado para gobernar	(1 - Si , 2 - No)
P23A05	N	J. AZNAR: con ideas de futuro	(1 - Si , 2 - No)
P23A06	N	J. AZNAR: capaz de llegar a acuerdos	(1 - Si , 2 - No)
P23A07	N	J. AZNAR: Sensible ante los problemas sociales	(1 - Si , 2 - No)
P23B01	N	F. GONZALEZ: Sincero	(1 - Si , 2 - No)
P23B02	N	F. GONZALEZ Creíble	(1 - Si , 2 - No)
P23B03	N	F. GONZALEZ: Dialogante	(1 - Si , 2 - No)
P23B04	N	F. GONZALEZ: Preparado para gobernar	(1 - Si , 2 - No)
P23 B05	N	F. GONZALEZ: con ideas de futuro	(1 - Si , 2 - No)
P23 B06	N	F. GONZALEZ: capaz de llegar a acuerdos	(1 - Si , 2 - No)
P23 B07	N	F. GONZALEZ: Sensible ante los problemas sociales	(1 - Si , 2 - No)
P24	N	Escala ideológica del entrevistado	(1-Izquierda...10- Derecha)
P2501	N	Escala ideológica de IU	(1-Izquierda...10- Derecha)
P2502	N	Escala ideológica del PP	(1-Izquierda...10- Derecha)
P2503	N	Escala ideológica del PSOE	(1-Izquierda...10- Derecha)
P2504	N	Escala ideológica del CIU	(1-Izquierda...10- Derecha)
P2601	C	Por qué partido no votaría nunca	(1 - IU; 2 - PP; 3 - PSOE)
P2602	C	Por qué partido no votaría nunca	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,..)
P2701	C	Partido que mejor representa las ideas de la gente como VD.	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,..)
P2702	C	Partido que le inspira más confianza	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,...)
P2703	C	Partido que tiene mejores líderes	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,…)
P2704	C	Partido más capacitado para gobernar	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,..)
P2705	C	Partido que mejor puede resolver los problemas de la economía	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 - PNV)
P2706	C	Partido que más puede contribuir a la mejora de los servicios	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 - PNV)
P2707	C	Partido que mejor puede resolver los problemas de la seguridad	(1 - IU; 2 - PP; 3 - PSOE; 4 - HB; 5 – PNV,..)
P36A	C	Motivo principal para votar PP	(1 ... 6)
P36B	C	Motivo principal para votar IU	(1 ... 6)
P36C	C	Motivo principal para votar PSOE	(1 ... 6)
P43	N	Sexo del entrevistado	(1 - Hombre , 2 - Mujer)
P44	N	Edad del entrevistado	(1 - [18, 24]; 2 - [25, 34]; 3 - [45, 54]; 5 - [55, 65]; 6 - [65 y más[)
P45	N	Ha ido a la escuela	(1 - Analfabeto ... 3 - Si)
P47	C	Situación laboral del entrevistado	(1- Trabaja ... ... 6- Estudiante ; 7 - Sus labores)
P48	C	Ocupación del entrevistado	

Tabla 7.2.1. Significado de las variables (2/3)

SÍMBOLO	TIPO (C/N)	DESIGNACIÓN	VALORES
P49	C	Situación profesional del entrevistado	(1 - Asalariado fijo ... 6-Miembro de cooperativa)
P49A	C	Tipo de empresa o de organización	(1 - Adm. Pubublica ; 3 - Emp. Privada ; 5 - Serv. Domestico)
P50	C	Ramo de actividad	

Tabla 7.2.1. Significado de las variables (3/3)

7.3. MINERÍA DE DATOS.

7.3.1. INTRODUCCIÓN.

En este tipo de datos, los individuos – identificados por el número del cuestionario (ver columna 1 de la **tabla 7.2.1.**) - son anónimos.

O sea, el analista no tiene, para cada uno de ellos, mas información que la que existe en el paquete de datos y, por lo tanto, el exámen visual del biplot no permite detectar inmediatamente asociaciones «interesantes» entre los individuos y las variables.

Por otro lado, siendo el autor de este trabajo un extranjero con conocimientos muy superficiales acerca de la semántica política en España, importaba verificar hasta que punto podría el sistema desarrollado encontrar estructuras, asociaciones y descripciones no triviales en estos datos.

Hemos fijado el siguiente objetivo: considerando solamente la información relativa a los tres principales partidos nacionales (PP, PSOE, IU) intentar

caracterizar las motivaciones que han llevado a los encuestados a votar en cada uno de esos tres partidos y, además, verificar si las síntesis generadas automáticamente por el sistema son coherentes, no triviales y aproximan de modo aceptable el significado de la estructura de datos identificada por los métodos de biplot y clasificación automática, conforme especificado en el **capítulo IV**.

Como se verifica, en este objetivo se mezclan tanto el deseo de obtener el significado de los datos relativos a las elecciones en España 1996 como el deseo de verificar si los desarrollos teóricos que han motivado esta investigación funcionan con datos no-triviales.

Fijado el objetivo, la secuencia que hemos establecido - entre muchas otras posibles - ha sido la siguiente:

1. Limpieza y preparación de los datos para análisis.
2. Análisis preliminar, buscando localizar, en el biplot, las distribuciones espaciales de las distintas variables.
3. Descubrimiento y caracterización de la estructura de los datos.
4. Construcción automática de sugerencias de interpretación para la estructura descubierta realizada usando análisis de datos multivariantes.
5. Comparación del significado de los grupos y particiones descubiertos con las sugerencias de interpretación obtenidas automáticamente, usando la metodología especificada en el **capítulo IV**.

7.3.2. LIMPIEZA DE LOS DATOS.

Estas operaciones han sido realizadas, en parte, con el editor de datos creado para este sistema y en parte usando recursos exteriores.

Las principales decisiones han sido las de eliminar «variables» cuyo valor era constante para todos dos individuos (no aportando información) o aquellas en las que el número de datos faltantes o respuestas de tipo **NS** - No sabe y **NR** - No responde, era de tal modo elevado que conservarlas tendría efectos inaceptables, dada la necesidad de eliminar o recodificar los valores de otras variables.

Una vez que el biplot es calculado usando solamente las variables cuantitativas, hemos decidido eliminar todos los casos con valores faltantes y de tipo ***NS / NR*** en variables cuantitativas, conservando los valores de ese tipo en variables cualitativas.

Esto no corresponde a ninguna limitación del sistema. Los datos cualitativos y mixtos pueden ser analizados creando variables indicadoras y paquetes en forma disyuntiva completa (identificados por TDC en nuestro sistema).

Los valores faltantes pueden ser recodificados (usando el editor de datos del sistema) cambiando esos valores por otros considerados más convenientes, de acuerdo con distintos criterios para tratar los valores faltantes.

Después de aplicar estas decisiones, el paquete de datos quedó reducido a 58 variables (columnas) y 148 cuestionarios o individuos.

Son esas las variables que integran la **tabla 7.2.1.** El conjunto de datos usados en el estudio integra el CD ROM del sistema, en Anexo a esta tesis.

7.3.3. ANÁLISIS PRELIMINAR DE DATOS.

Hemos decidido usar datos normalizados y, dada la calidad de representación, el HJ - biplot de GALINDO (*RCMP*) como soporte gráfico para todos los estudios que seguidamente se presentan.

La preocupación básica inicial ha sido la de obtener una percepción de las localizaciones, en el biplot, de los valores de las distintas variables observadas, asociando las categorías de esas variables con zonas específicas del biplot

Esto se obtiene usando como etiquetas de los individuos los valores de las variables - numéricas y cualitativas- y por la construcción visual y interactiva de «*queries*», permitida por el sistema.

En la **figura 7.3.3.1.** se puede observar, sucesivamente, el biplot inicial y la distribución espacial de los valores de las variables P11- Partido en que ha votado en 1996, P24- Escala ideológica del entrevistado, P44 – Edad del entrevistado.

Ese biplot corresponde a los dos primeros ejes de la SVD y representa 36% de la varianza total.

Se detectan, inmediatamente, los cuatro grupos de variables siguientes:

SW = {P23 B01, P2 B02, P23 B03, P23 B 04, P23 B05, P23 B06, P23, B07}

NE = {P2203, P24}

NW = {P2204, P403, P2205}

SE = {P23 A01, P23 A0, P23 A03, P23 A04, P23 A05, P23 A06, P23, A07}

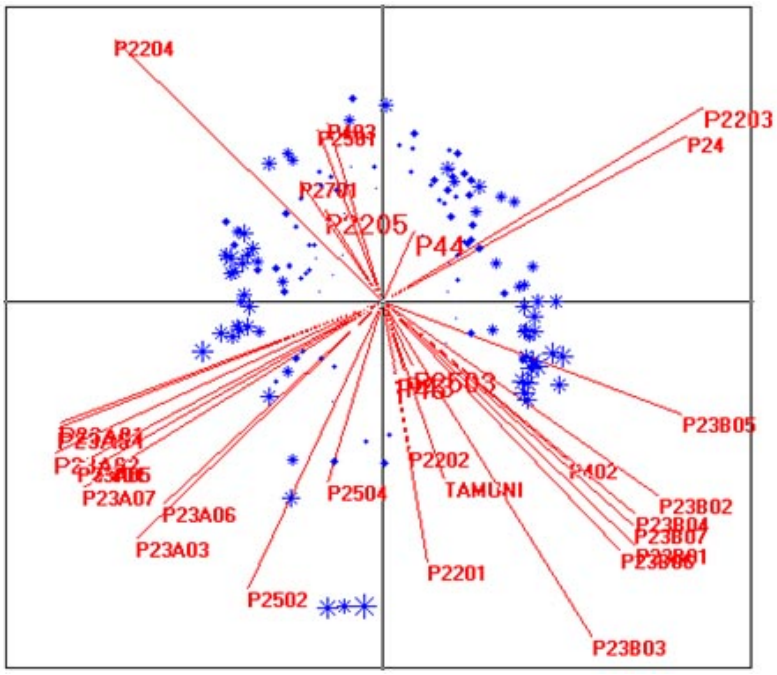


Figura 7.3.3.1. a) - El biplot inicial

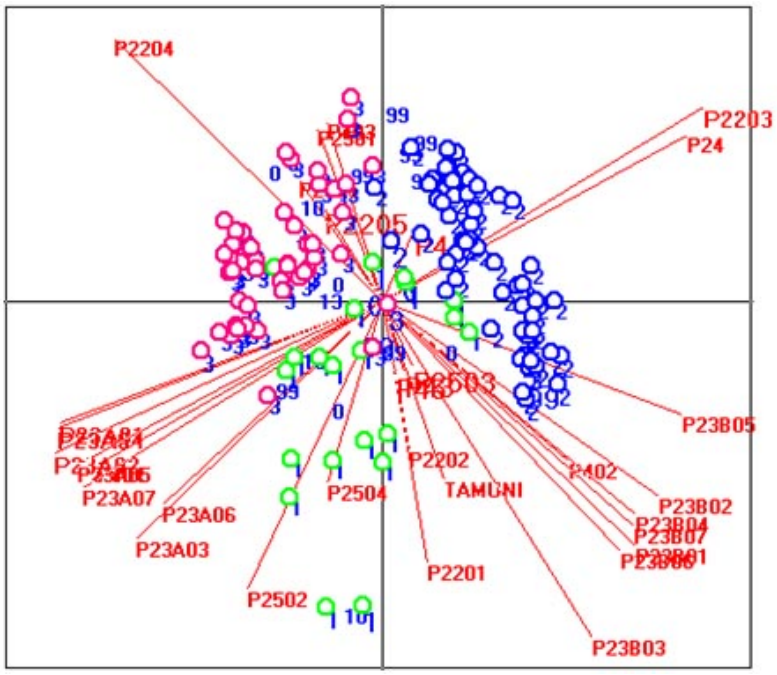


Figura 7.3.3.1. b) - Variable P11-Recuerdo de voto en las elecciones de 1996

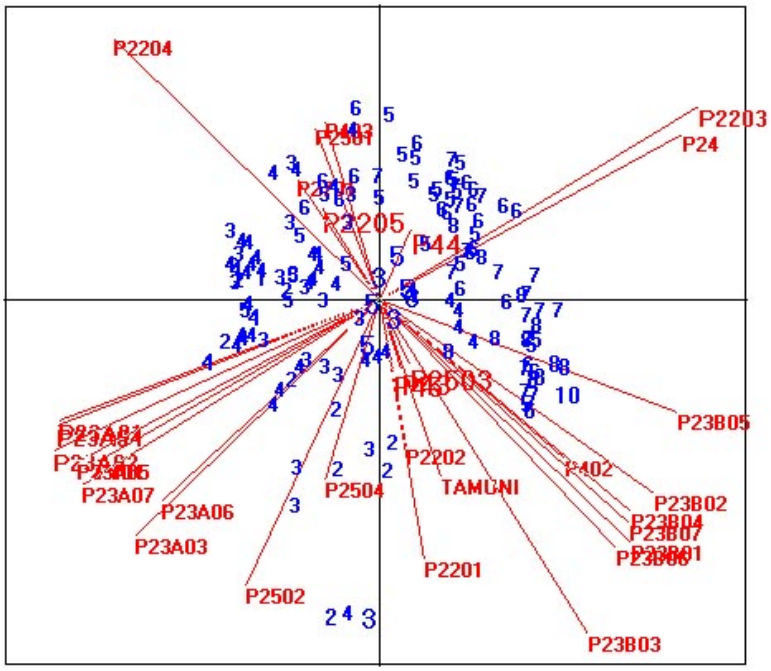


Figura 7.3.3.1. c) - Variable P24-Escala ideológica del entrevistado.

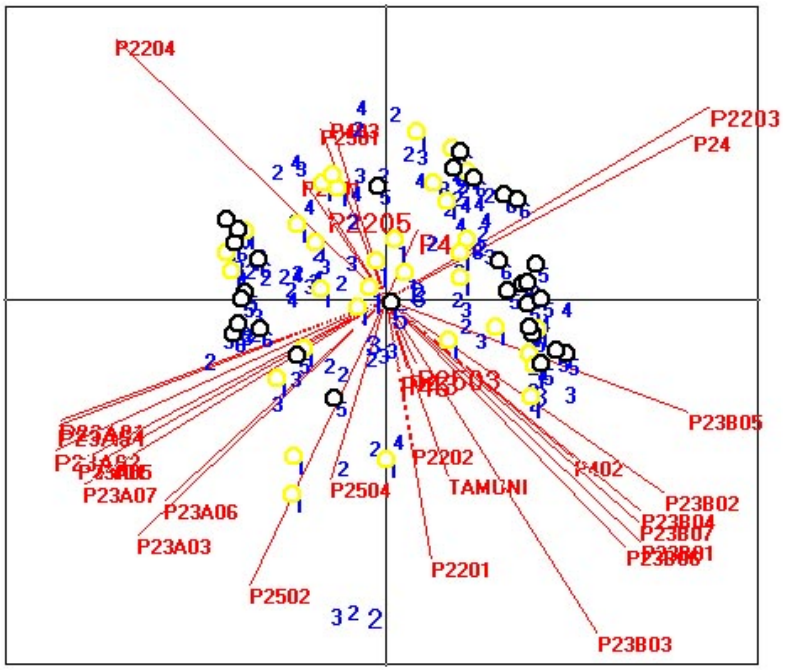


Figura 7.3.3.1 d) - Variable P44-Grupo de edad del entrevistado.

Una cuestión que inmediatamente ocurre es: ¿dónde están, en el biplot, los individuos que han votado PP, PSOE y IU?.

En la **figura 7.3.3.1.b)** se observa la distribución espacial de los valores de la variable P11= «Recuerdo de Voto en las elecciones de 1996» en que: *1- IU, 2 – PP , 3 – PSOE y 99 - NS/NR.*

Usando la construcción interactiva de «*queries*» hemos representado por colores distintos esta misma información.

En esa figura, IU = 1(verde en el original), PP = 2 (azul en el original), PSOE =3 (rojo en el original).

Verificamos que los encuestados que han votado *PSOE* están concentrados en la zona NW del gráfico; los que declaran haber votado PP están en la zona derecha - NE y los que declaran haber votado IU están en la zona SUR del gráfico.

Para confirmar estas localizaciones hemos representado sobre el mismo biplot los valores de las variables P24 - Escala ideológica del entrevistado. Ver **figura 7.3.3.1.** Esta escala ideológica varía de *1 - Izquierda a 10 - Derecha.*

Examinando las **figuras 7.3.3.1.** se obtienen los primeros hallazgos interesantes: se verifica que los que han declarado haber votado PSOE se ven a sí mismos próximos al centro (4, 5, 6) y son localizados aún más al centro por los que declaran haber votado PP.

Recíprocamente, examinando la escala ideológica de los que han declarado haber votado PP se verifica que, los de este grupo no se ven tan a la derecha como son vistos por los que declaran haber votado PSOE.

La **figura 7.3.3.1d)** presenta la distribución espacial de los dos grupos de edad extremos (variable P44): amarillo en el original - los jóvenes *1 = [18 - 24]* y a negro los grupos de edad *5 = [56 - 64] y 6 = [65 y más[.*

Se verifica que, al revés de lo que ocurre con los jóvenes - que están distribuidos de modo homogéneo por todo el espectro político - los mayores están polarizados entre izquierda y derecha. Esto se puede

explicar, quizás, por las crispaciones de la historia de España en los últimos 50 años, que marcan, principalmente los de más edad.

No es posible presentar, por falta de espacio, todos los hallazgos que el estudio, por este método, permite descubrir y que, por si solo, ilustra la adecuación de los biplots para este tipo de estudios.

7.3.4. ESTUDIO Y CARACTERIZACIÓN DE LA ESTRUTURA DE LOS DATOS.

En este punto del estudio habíamos adquirido la percepción de que la información característica de los que declaran haber votado PSOE se situaba en la zona *E - NW* del biplot, la información relativa a los que han declarado haber votado *IU* se situaba entre *SE y SW*.

Convendría ahora que un análisis cluster sugiriera grupos que permitiesen formular una hipótesis de estructura a explotar enseguida.

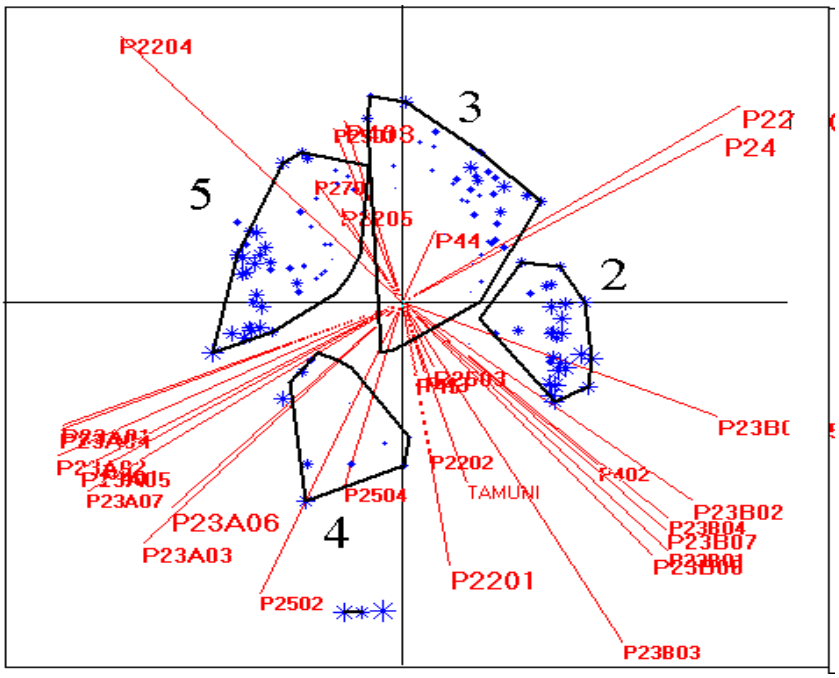


Figura 7.3.4.1 Partición con 4 grupos.

Usando el método de agregación de Ward aplicado a las distancias euclidianas entre los individuos representados sobre los ejes factoriales, se ha obtenido la **figura 7.3.4.1** en donde figura una partición con 4 clases, obtenida «cortando» el dendograma al nivel apropiado.

.

Examinando cada uno de estos 4 grupos usando los instrumentos del sistema (caracterización y comparación de grupos) se verifica que:

Grupo 2

- Corresponde a encuestados que votaron en el PP en las elecciones de 1996 y votan regularmente en el PP.
- Son unánimes -variabilidad nula - en los aspectos siguientes:

  Credibilidad, ideas de futuro y sensibilidad a problemas sociales de AZNAR.

  Ausencia de credibilidad y sinceridad de F. GOZALEZ.

Valoran las capacidades de J. AZNAR muy por encima de la media (sincero, dialogante, preparado para gobernar, capaz de llegar a acuerdos).

Dudan de las capacidades de diálogo, para gobernar y para llegar a acuerdos de F. GONZALEZ.

En una escala 1-10 (1 - Izquierda, 10 - Derecha) se ven cerca de la derecha (7) pero ubican el PP más a la derecha y el PSOE cerca del centro.

(Se consideran menos a la derecha que el PP, partido en el que votan y ven al PSOE como un partido cerca del centro).

- El sistema suministra automáticamente como expresión conjuntiva para este grupo la siguiente:

$$(P23B06 = 2) \wedge (P2704 = 2)$$

Traducida, significa:

«*F. GONZALEZ no es capaz de llegar a acuerdos y el PP es el partido más capacitado para gobernar*»

Grupo 3

- Aunque la moda en este grupo es haber votado PP en las elecciones de 1996, muchos de los encuestados o no han contestado a la cuestión *P11* o entonces han declarado haber votado PSOE, notándose una volatilidad elevada.
- De un modo general valoran positivamente las cualidades de J. AZNAR (sincero, creíble, dialogante, preparado para gobernar, con ideas de futuro, capaz de llegar a acuerdos y sensible a los problemas sociales).

- Con relación a F. GONZALEZ no manifiestan gran oposición y consideran que es capaz de llegar a acuerdos.

Desde el punto de vista ideológico se sitúan un poco por encima del centro (5.4) ubican al PP - partido en que votaron más frecuentemente - entre el centro y la derecha (6.9) y el PSOE prácticamente en el centro (4.8) aunque ligeramente a su izquierda.

El sistema sintetiza este grupo por la expresión conjuntiva siguiente:

(P23 A01= 1)

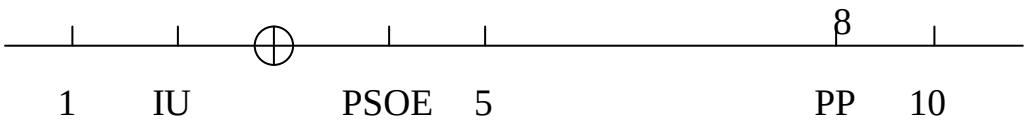
Su significado es : «*AZNAR es sincero*».

En síntesis: se trata de un grupo de centro derecha que vota mayoritariamente PP motivado sobre todo por la personalidad de J. AZNAR.

Grupo 4

- Han votado por IU en las elecciones de 1996 y lo hacen consistentemente.
- Conocen y valoran a Anguita muy por encima de la media.
- Conocen mal y no valoran a J. AZNAR: no es sincero ni dialogante ni preparado para gobernar ni capaz de llegar a acuerdos ni sensible a los problemas sociales.
- F. GONZALEZ - No es sincero.
- En la escala ideológica se sitúan entre la izquierda y el centro (2.9), ven al partido en que han votado (IU) ligeramente más a izquierda

(2.4), localizan al PSOE en el centro y al PP mucho más a derecha que la media.



La expresión conjuntiva que el sistema sugiere como síntesis de este grupo es:

(P2701= 1) ∧ (P2706= 1)

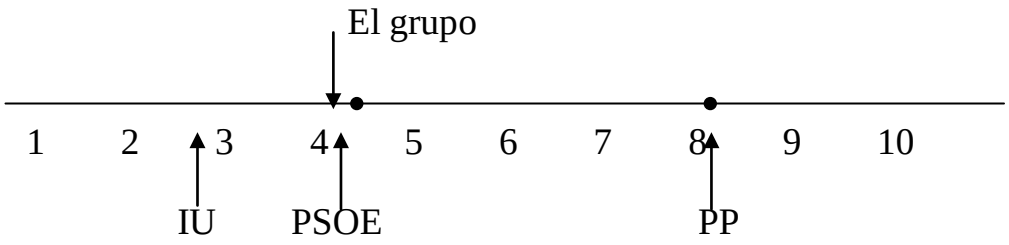
lo que significa:

«*IU es el partido que representa mejor a personas como yo y es el partido que más puede contribuir para mejorar los servicios*».

Grupo 5

- Prácticamente todos los encuestados de este grupo han votado en el PSOE y lo hacen consistentemente hace mucho tiempo.
- Conocen mal y no valoran ni ANGUITA ni J. AZNAR.
- Conocen y valoran F. GONZALEZ muy por encima de la media.
- Son unánimes en la creencia de que GONZALEZ está preparado para gobernar y es capaz de llegar a acuerdos.
- A J. AZNAR no le consideran ni sincero, ni creíble, ni dialogante, ni preparado para gobernar, ni capaz de llegar a acuerdos, ni sensible a problemas sociales.
- A F. GONZALEZ lo consideran sensible, creíble, dialogante, preparado para gobernar, con ideas de futuro y sensible a problemas sociales.

- Desde el punto de vista ideológico la visión de este grupo es la siguiente:



O sea: se consideran de centro izquierda y ven el PSOE bastante a la derecha.

El sistema genera, para este grupo, la expresión conjuntiva siguiente:

(P2706= 3) $\wedge$ ***(P2704= 3)***,

Traducida, significa:

«*El PSOE es el partido que más puede contribuir para mejorar servicios y es el partido más capacitado para gobernar*».

Usando ahora la posibilidad de comparar grupos (botón ) interesa precisar las diferencias entre los grupos **5** (PSOE) y **3** y entre los grupos **2** y **3.**

Recuerde que los grupos 2 y 3 han votado en el PP pero tienen sensibilidades distintas en su visión de la personalidad de F. GONZALEZ.

Comparando el grupo 2 con el grupo 3 se verifica que al caminar de 3 para 2:

- Aumenta el conocimiento y valoración de F. GONZALEZ.
- Aumenta la media de los que consideran a J. AZNAR con ideas de futuro.
- Aumenta la media de los que consideran a F. GONZALEZ creíble, dialogante, preparado para gobernar, con ideas de futuro, sensible a problemas sociales.
- Disminuye la escala ideológica media (aproximación a la izquierda).
- La IU es vista como estando más al centro.

Comparando los grupos 5 y 3 se verifica que, caminando de 5 (PSOE) para 3 (2º Grupo de votantes en el PP) los hechos más relevantes son: aumenta el conocimiento y valoración de la personalidad de J. AZNAR, disminuye el conocimiento y valoración en la personalidad de F. GONZALEZ.

El hecho más relevante es la verificación de que a medida que se camina de 5 (PSOE) para 3: el PSOE es visto, en esta trayectoria, como aproximándose al centro (pasa de 4 a 4.8).

7.3.5. CONSTRUCIÓN AUTOMÁTICA DE SUGESTIONES DE INTERPRETACIÓN DE LA ESTRUTURA DESCUBIERTA.

En el apartado 7.3.5. se han presentado las expresiones conjuntivas sugeridas por el sistema para síntesis del significado de los distintos grupos.

Esas expresiones son construídas con base en el modelo teórico explicado en el apartado 4.6, basado en la medida de afinidad entre átomos, usando el algoritmo presentado en 4.6.2.

Como se ha visto, en este algoritmo no consideramos la eventual existencia simultánea, de otros grupos: nos concentramos en buscar la mejor descripción para un determinado grupo sin que en ese proceso nos preocupemos en contrastar un grupo con otro.

En el apartado 4.7. se propone el uso de los árboles de regresión para buscar sugerencias automáticas de interpretación para un conjunto de grupos - en particular para una partición. Eso se realiza según el algoritmo 4.7.1.

Aplicando, ahora, ese algoritmo a nuestras cuatro clases de la partición descubierta por la combinación Biplot-Analisis Cluster, el sistema sugiere las siguientes reglas para caracterizar esos cuatro grupos al mismo tiempo:

GRUPO 2

Sí *(P2701= 2) ∧ (P23 B06= 2) ∧ (P23 B02= 2)* **Entonces** *(GRUPO= 2)*
Traducido: «*El partido que mejor representa mis ideas es el PP*», *«F. GONZALEZ **no** es capaz de llegar a acuerdos» y* *« F. GONZALEZ **no** es creíble*»

GRUPO 3

Si *(P2701= 2) ∧ (P23 B06= 1)* **Entonces** *(GRUPO= 3)*
Traducido: «*El partido que mejor representa mis ideas es el PP*» *y* *«* F. GONZALEZ ***si** es capaz de llegar a acuerdos*»

GRUPO 4

Si (P2701= 1) **Entonces** *(GRUPO= 4)*
Traducido: «*El partido que mejor representa mis ideas es IU*»

GRUPO 5

Si *(P2701= 3)* $\wedge$ *(P23 A01= 2)* $\wedge$ *(P23 B03= 1)* **Entonces** *(GRUPO= 5)*
Traducido: «*El partido que mejor representa mis ideas es el PSOE»,* *«J. AZNAR* ***no*** *es sincero» y* *« F. GONZALEZ* ***es*** *dialogante*»

Como se observa, *las reglas de interpretación/síntesis - a las que el sistema llega de modo automático - son consistentes con la caracterización de los grupos a que habíamos llegado empíricamente en los apartados 7.3.4. y* ***7.3.5*** *de modo "manual".*

*Obsérvese que el sistema ha detectado y reflejado en las sugerencias de interpretación la sutil distinción de sensibilidades entre los dos grupos (**2** y **3**) de votantes en el PP.*

La **figura 7.3.5.1.** refleja estas observaciones. En esa figura puede observarse la casi coincidencia entre los grupos detectados por los métodos de análisis multivariante Biplot-Análisis Cluster (marcados por 2, 3, 4, 5) y los cierres convexos de las descripciones anteriores, generadas automáticamente.

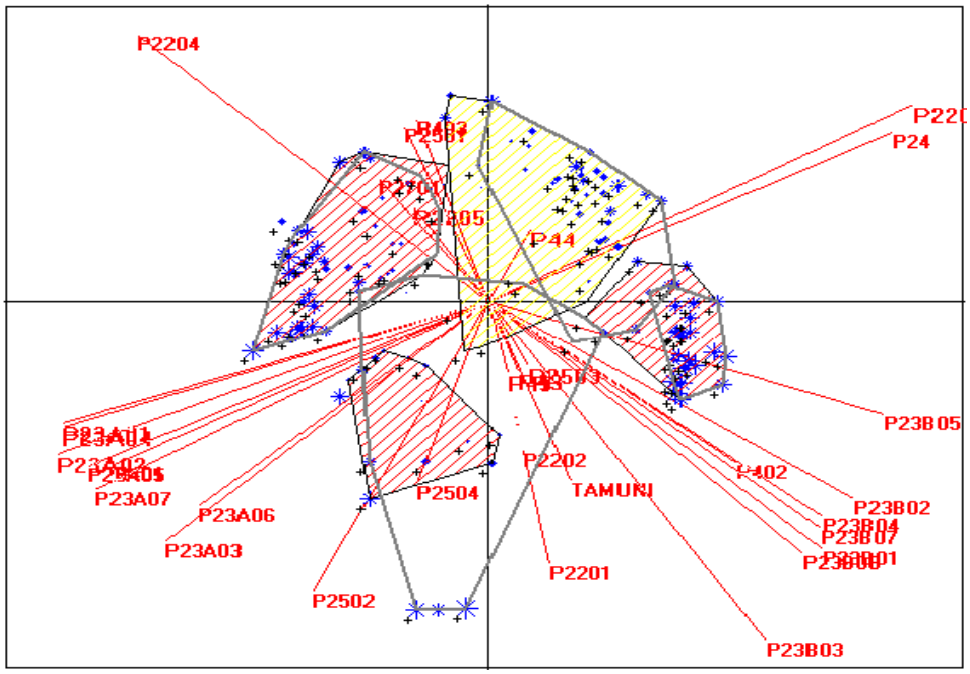


Figura 7.3.5.1. Los cuatro grupos y, a gris, los cierres convexos de las interpretaciones automáticas.

Esta figura sintetiza bien la relevancia de los desarrollos teóricos que hemos propuesto en el **capítulo IV** y su realización por el sistema de minería presentado en el **capítulo VI**.

CONCLUSIONES

1. La exhaustiva revisión bibliográfica realizada pone de manifiesto que, en Análisis de Datos, crece cada vez más el interés por los aspectos relacionados con la riqueza interpretativa de los métodos de análisis, pero no existe una formalización de los mecanismos lógicos que rigen las reglas de interpretación utilizadas por el analista, a pesar de su importancia en la creación de un sistema de minería de datos.

2. Hemos demostrado que la formulación matemática del problema de interpretación de los resultados de un Análisis Multivariante puede ser expressada usando grafos de intersección.

3. El problema de interpretación de los resultados también puede ser formulado usando la teoría de los conjuntos imprecisos.

4. Cuando el resultado sea una partición del conjunto de individuos, el problema de interpretación puede ser formulado como un problema de regresión cualitativa.

5. Hemos demostrado que los métodos BIPLOT son una base idónea para el desarrollo de un sistema de minería de datos ya que estos métodos satisfacen los criterios habituales exigibles. Además, la mayoría de las técnicas de análisis de datos pueden ser expresadas como casos particulares de los BIPLOT.

6. El sistema de minería de datos basado en BIPLOT que hemos desarrollado, y al que hemos denominado BIPLOT PMD, ha sido el laboratorio que nos ha permitido contrastar la validez de la teoría formulada.

7. Los programas que existen en la actualidad para aplicar los métodos BIPLOT producen gráficos factoriales estáticos; la aplicación que se ha desarrollado en este trabajo proporciona gráficos dinámicos que permiten en todo momento la interacción entre el usuario y el gráfico. Esta interacción permite proyectar las nubes sobre cualquier dirección e interpretar las direcciones asociadas a características relevantes de los datos, definir interactivamente expresiones conjuntivas y visualizarlas sobre el gráfico.

BIBLIOGRAFIA

AGRAWALL, R; IMIELINSKI, T. & SWAMI, A. (1993). 'Mining association rules between sets of items in large databases'. *Proceedings of the ACM SIGMOD Conference on Management of Data, **1993***, (May), 207-216.

ALUJA, T.; MORINEAU, A. (1999). *Aprender de los Datos: El Análisis de Componentes Principales. Una Aproximación desde el Data Mining*. EUB - Ediciones Universitarias de Barcelona.

ALUJA, T. (2001). 'La Minería de Datos, entre la Estadística y la Inteligencia Artificial'. Qüestíio, **25** (**3**), 479-478.

AMARO-MARTIN, I. R. (2001). *Manova Biplot Para Diseños con Varios Factores Basado en Modelos Lineales Generales Multivariantes.* Tesis Doctoral. Universidad de Salamanca.

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. (2nd Ed.). John Wiley.

ANDERSON, R. (1990). *Cognitive Psychology and its Implications.* Freeman.

BECKER, R. A.; CLEVELAND, W. S. & SHYU, M. (1996). 'The visual display and control of treillis display'. *Journal of Computational and Graphical Statistics*. **5** (**2**), 123-155.

BENZÉCRI, J. (1973). *L'Analyse des Données.* Dunod.

BENZÉCRI, J. (1992). *Correspondence Analysis Handbook*. Marcel Dekker.

BERGE, C. (1970). *Graphes et Hypergraphes*. Dunod.

BERTIN, J. (1977). *La Graphique et le Traitement Graphique de l'Information.* Flammarion.

BEZDEK, J. & PAL, S. (1992). *Fuzzy Models for Pattern Recognition. Methods that Search for Structures in Data*. IEEE Press.

BLASIUS, J. & GREENACRE, M. (1998). *Visualization of Categorical Data*. Academic Press.

BLÁZQUEZ ZABALLOS, A. (1998). *Análisis Biplot Basado En Modelos Lineales Generalizados*. Tesis Doctoral. Universidad de Salamanca.

BOCK, H. H. & DIDAY, E. (1999). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag.

BORG, I.; GROENEN, P. (1997). *Modern Multidimensional Scaling. Theory and Applications*. Springer Verlag.

BORG, I. & GROENEN, P. (1998). 'Regional Interpretations in Multidimensional Scaling'. In: J. Blasius & M. Greenacre (Eds.). *Visualization of Categorical Data*. Academic Press, pp.: 347-364.

BORGELT, C. & KRUSE, R. (2001). *Graphical Models for Data Analysis and Mining*. John Wiley.

BRADLEY, P. S.; FAYYAD, U. & REINA, C. (1998). *Scaling Clustering Algorithms to Large Data-Bases.* American Association for Artificial Intelligence.

BRADU, D. & GABRIEL, K. (1978). 'The Biplot as a diagnostic tool for models of two-way tables'. *Technometrics*, **20** (**1),** 47-68.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. & STONE,C.J. (1984,1993). *Classification and Regression Trees*. Chapman & Hall.

BUJA, A.; COOK, A. & SWAYNE, F. (1996). 'Interactive high-dimensional data visualization'. *Journal of Computational and Graphical Statistics*. **5** (**1**), 78-99

BURNHAN, K. & ANDERSON, D. (2002). *Model Selection and Multimodel Inference. A Practical Information. Theoretic Approach.* Springer.

CARD, S. K.; MACKINLAY, J. D. & SHNEIDERMAN, B. (Eds.) (1999). *Readings in Information Visualization. Using Vision to Think.* Morgan Kaufman.

CÁRDENAS, O. (2000). *Biplot con Información Externa Basado en Modelos Lineales Generalizados.* Tesis Doctoral. Universidad de Salamanca.

CARLIER, A. & KROONENBERG, P. (1998). 'The case of the french cantons: an application of Three-way Correspondence Analysis'. In: J. Blasius & M. Greenacre (Eds.). Academic Press, pp.: 253-276.

CIPPRA, B. (1999). 'Massive Graphs Pose Big Problems'. *SIAM News 1999* (**04**).

CLEVELAND, W. S. (1993a). 'Research in Statistical Graphics'. *Journal of the American Statistical Association.* 82, 419-423.

CLEVELAND, W. S. (1993b). *Visualizing Data*. Hobart Press.

CLEVELAND, W. S. (2001). *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*. Statistics Research, Bell Laboratories.

CODD, E. (1970). *A Relational Model for Large Shared data Banks. CACM* **13** (**6**).

COOK, B.; BUJA, A & CABRERA (1993). 'Projection pursuit Indexes Based on Orthonormal Functions Expansions'. *Journal of Computational and Graphical Statistics*. **2** (**3**), 225-250

COOK, B.; BUJA, A; CABRERA, J. & HURLEY, C. (1995). 'Grand tour and projection pursuit'. *Journal of Computational and Graphical Statistics*. **4** (**3**), 155-172

COX, T. & COX, M. A. (1994). *Multidimensional Scaling*. Chapman and Hall.

CUADRAS, C. M. (1981). *Métodos de Análisis Multivariante*. Editorial Universitaria de Barcelona.

DIDAY, E.; LEMAIRE, J.; POGET, J. & TESTU, F. (1982). *Éléments dánalyse de données*. Dunod.

DÍAZ-LENO, M. S. (1995). *Los Métodos Biplot como Herramienta de Diagnóstico en la Modelización de Datos Multidimensionales.* Doctoral. Universidad de Salamanca.

DODGE, Y. (1996). 'The Guinea pig of multiple regression'. In: H. Rieder (Ed.). *Robust Statistics, Data Analysis, and Computer Intensive Methods*. Springer, pp.: 91-117.

ECKART, C. & YOUNG, G. (1936). 'The approximation of one matrix by another of lower rank'. *Psychometrika*. **1** (**3**), 212-218.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. & UTHURUSAMY, R. (1996). *Advances in Knowledge Discovery and Data Mining.* Mit Press.

FAYYAD, U. (1997). *Data Mining and Knowledge Discovery*. *Data* Mining and Knowledge Discovery – **1** (**1**), 5-10.

FERNANDEZ-GOMEZ, M.J. (1995). *Contribuciones al Analisis Multivariante Directo del Gradiente Mediante Estudio Combinado de Configuraciones Espaciales.* Tesis Doctoral. Universidad de Salamanca.

FURNAS, G. W. & BUJA, A. (1994). 'Prosection view: dimensional inference through sections and projection. *Journal of Computational and Graphical Statistic.* **3** (**4**), 323-385.

GABRIEL, K. R. (1971). 'The Biplot graphic display of matrices with application to Principal Component Analysis'. *Biometrika.* **58** (**3**), 453-467.

GABRIEL, K. R. (1981). 'Biplot display of multivariate matrices for inspection of data and diagnosis'. In: V. Barnett (Ed.). Interpreting Multivariate Data, *pp:. 147-174. Wiley. Chichester.*

GABRIEL, K. R. (1995a). 'Biplot display of multivariate categorical data with comments on Multiple Correspondence Analysis'. In:W. Krzanowski, (Ed.). *Recent Advances in Descritive Multivariate Analysis, pp:.* 190 – 226. Oxford Science Publications.

GABRIEL, K. R. (1995b). **'**Manova Biplots for Two - Way Contingency Tables'**.** In: W. Krzanowski, (Ed.). *Recent Advances in Descritive Multivariate Analysis, pp:.* 227 – 268. Oxford Science Publications.

GABRIEL, K. R ; GALINDO, M. P. & VICENTE-VILLARDÓN, J.L. (1998). 'Use of Biplots to diagnose independence models in three-way contingency tables. In J. Blasius & M. Greenacre.(Ed.). *Visualization of Categorical Data, pp:* .391 – 404. Academic Press.

GABRIEL, K. R. (1998). 'Generalised bilinear regression'. *Biometrika***. 85** (**3**), 689-700.

GABRIEL, K. R. (2002). **'**Goodness of fit of Biplots and Correspondence Analysis'. *Biometrika.* **89** (**2**), 423-436.

GALINDO, M. P. (1985). *Contribuciones a la Representación Simultánea de Datos Multidimensionales.* Tesis Doctoral. Universidad de Salamanca.

GALINDO, M. P.; GOMEZ-GUTIERREZ, J.M. & VICENTE-VILLARDÓN, J.L. (1986). 'Practica Del Análisis de Correspondencias (Aplicación a Un Problema Biologico)'. *Cuadernos de Bioestadística* **4 (1**), 63-79.

GALINDO, M.P. (1986). Una Alternativa de representación simultanea: HJ-Biplot. *Qüestiió* **10 (1)**, *12 – 23*

GALINDO, M. P. y CUADRAS, C. M. (1986). Una extensión del método Biplot y su Relación con otras Técnicas. *Publicaciones de Bioestadística y Biomatemática. Universidad de Barcelona*, **17.**

GANTER, B. & WILLE, R. (1996). *Formal Concept Analysis. Mathematical Foundations*. Springer.

GAUL, W. & LOCAREK-JUNGE (Ed.) (1998). *Classification in the information Age*. Springer

GAUL, W.; OPITZ, O. & SCHANDER, M. (Eds.) (2000). *Data Analysis. Scientific Modelling and Practical Applications*. Springer.

GOLUB, G.H. & Van LOAN, C.F. (1983). *Matrix Computations.* John Hopkins University Press.
GORDON, A. D. (1999). (2nd Ed.). *Classification*. Chapman & Hall.

GOWER, J. & HAND, D. (1996). *Biplots*. Chapman & Hall

GREENACRE, M. (1984). *Theory and Application of Correspondence Analysis*. Academic Press.

GREENACRE, M. & BLASIUS, J. (1994). *Correspondence Analysis in the Social Sciences*. Academic Press.

GREENACRE, M. (1998). 'Diagnostics for Joint Displays in Correspondence Analysis'. In: BLASIUS et al (1998), *pp:*. 221-238.

GROSS, J. & YELLEN, J. (1999). *Graph Theory and its Applications*. CRC Press.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L. & BLACK, W. C. (1995). (4ª Ed.). *Multivariate Data Analysis with Readings*. Prentice Hall International.

HALL, P.; MARSHALL, D. & MARTIN, R. (2000). 'Merging and splitting eigenspace models'. *IEE Transactions on Pattern Analysis and Machine Intelligence*. **22** (**9**), 1042-1049.

HALL, P.; MARSHALL, D.; MARTIN, R. (2002). 'Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition'. *Image and Vision Computing*. **20**, 1009-1016.

HAN, J. & KAMBER, M. (2001). *Data Mining Concepts and Techniques*. Morgan Kaufmann.

HAND, D. (1998). *Intelligent Data Analysis: Issues and Opportunities.* Elservier Science Inc.

HASTIE, T.; TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. Springer Verlag.

HOSKING, J.; PEDNAULT, E. & SUDAN, M. (1997). *A Statistical perspective on data mining.* IBM Research Publications.

HUANG, C.; McDONALD, J. A. & STUETZLE, W. (1997). 'Variable resolution bivariate plots'. *Journal of Computational and Graphical Statistics*. **6** (**4**), 383-396.

HUBER, T. & NAGEL, M. (1996). 'Data Based Prototyping'. In: R. Helmut (Ed.). *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp.: 197-213. Springer.

JACKSON, J. E. (1991). *A User´s Guide to Principal Components*. John Wiley.

JAMBU, M. (1991). *Exploratory and Multivariate Data Analysis*. Academic Press.

JAROSZEWICZ, S. & SIMOVICI, D. A. (2001). 'A general measure of rule interestingness'. In: L. de Raedt & A. Siebes (Eds.). *PKDD 2001, LNAI 2168*, pp.: 253-265. Springer.

JOHNSON, R. A. & WICHERN, D. W. (1998) (4rd Ed.). *Applied Multivariate Statistical Analysis.* Prentice Hall.

JOLLIFFE, I. T. (2002). (2nd Ed.). *Principal Components Analysis*. Springer.

KLÖSGEN, W. & ZYTKOW, J. (Eds.) (2002). *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

KOLDA, T. & O´LEARY D. (1999). *Computation and Uses of the Semidiscrete Matrix Decomposition*. Tgkolda@sandia.gov; leary@cs.umd.edu.s

KNOBBE, A.; HAAS, M. & SIEBES, A. (2001). 'Proporsitionalisation and aggregates'. In L. de Raedt & A. Siebes (Eds.). *PKDD 2001, LNAI 2168*, pp.: 277-288. Springer.

KROONENBER, P.M. (1983). *Three-Mode Principal Component Analysis*. DSWO Press.

KRZANOWSKI, W. J. (1979). Between groups comparison of Principal Componentes. *Journal of the American Statistical Association*.**74** (**367**), 703-707.

KRZANOWSKI, W. J. (1998, 2000). *Principles of Multivariate Analysis - A User´s Perspective* (Revised Ed.). Oxford Science Publications.

LEBART, L.; MORINEAU, A. & PINON, M. (2002). (3eme Éd.). *Statistique Exploratoire Multidimensionelle.* Dunod.

LIN, T. Y.; CERCONE, N. (1997). *Rough Sets And Data Mining*. Kluwer Academic Publishers.

LOHR, S. (1999). *Sampling: Design and Analysis.* Duxbury Press.

MARTIN-CASADO (1992). *Un Nuevo Procedimiento de Obtención de la Funcción de Utilidad Mediante el Escalamiento Multidimensional. Fundaciones Matemáticas*. Tesis Doctoral. Universidad de Salamanca.

MARTÍN-RODRÍGUEZ, J. (1996). *Contribuciones a la Integración de Subespacios desde una Perspectiva Biplot*. Tesis doctoral. Universidad de Salamanca.

MARTÍN-RODRIGUEZ, J. (2000). *Comparación e Integración de Subespacios Resultantes de Marcadores Biplot*. Universidad de Salamanca Departamento de Estadística.

MARTÍN-RODRIGUEZ, J.; GALINDO-VILLARDÓN; VICENTE-VILLARDÓN, J.L. (2002). 'Comparison and integration of subspaces from a biplot perspective'. Journal of Statistical Planning and Inference 102 (2002), 411- 423.

McKEE, T. & McMORRIS, F. R. (1999*). Topics in Intersection Graph Theory*. SIAM Monograph in Discrete Math. and Appl.

MICHALSKY, R. S; CARBONELL, J. G. & MITCHELL, T. M. (Eds) (1983). *Machine Learning: An Artificial Intelligence Approach*. Tioga Publishing. Palo Alto, CA.

PAWLAK, Z. (1991). *Rough Sets. Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publisher.

PAWLAK, Z. (1998). 'Reasoning about data. A rough set perspective'. In: L. Polkowski & A. Skowron (Eds.). *Rough Sets and Current Trends in Computing*. First International Conference, RSCTC'98 Proceedings, pp.: 25-34.

PEÑA, J. M.; LÉTOURNEAU, S. & FAMILI, F. (1999). 'Application of rough sets algorithms to prediction of aircraft component failure.' In: D. J. Hand; J. N. Kok & M. R. Berthold (Eds.). *IDA'99, LNCS 1642*, pp.: 473-484. Springer-Verlag.

POLKOWSKI, L. & SKOWRON, A. (Eds.) (1998a). 'Rough Sets And Current Trends in Computing'. *First International Conference, RSCTC98 Proceedings*. Springer.

POLKOWSKI, L; SKOWRON, A. (Eds.) (1998b). *Rough Sets. In: Knowledge Discovery -1- Methodology and Applications*. Springer.

POLKOWSKI, L.; TSUMMOTO, S. & LIN, T. Y. (Eds.) (2000). *Rough Sets Methods and Applications. New Developments in Knowledge Discovery in Information Systems*. Springer.

POSSE, C. (1995). ‘Tools for Two Dimensional Exploratory Projection’. *Journal of Computational and Graphical Statistics*. **4** (**2**), 83-100.

PRESS, W. P.; FLENNERY, B. P.; TEULKOSKY, S. A. & VETTERLING, W. T. (1989). *Numerical Recipes in Pascal. The Art of Scientific Computing*. Cambridge University Press.

PROVOST, F. & KOLLURI, V. (2002). ‘Scalability’. In KLÖSGEN et al (2002), 418-433.

QUILAN, J. R. (1993). *Programs for Machine Learning*. Morgan Kaufman.

RENCHER, A. C. (1995). *Methods of Multivariate Analysis*. John Wiley.

RHAM, C. (1980). ‘La classificaction ascendente hierarchique selon la méthode des voisins reciproques’. *Les Cahiers de l’Analyse des Données.* **5** (**3**), 135-144.

RHODES, P. J. (2002). ‘Discovery New Relationships. A Brief Overview of Data Mining and Knowledge Discovery’. In: U. Fayyad, et al. (Ed.) Information Visualization in Data Mining And Knowledge Discovery. Morgan Kaufmann, pp:. 277-290

SCOTT, D. W. (1992). *Density Estimation: Theory, Practice and Visualization.* John Wiley.

SEBER, G. A. F. (1983). *Multivariate Observations*. John Wiley.

SEPÚLVEDA, R. (2003). *Contribuciones al Analisis de Clases Latentes en Presencia de Dependencia Local.* Tesis Doctoral. Universidad de Salamanca.

SOWA, J. F. (2000). *Knowledge Representation, Logical, Philosophical and Computational Foundations*. Books/Coole.

SNEATH, P. H. & SOKAL, R. R. (1973). *Numerical Taxonomy.* Freeman.

SPENCE, R. (2001). *Information Visualization*. Addison-Wesley

STEPANIUK, J. (2000). ‘Knowledge discovery by application of rough set models’. In: L. Polkowski; S. Tsumoto & T. Y. Lin (Eds.). *Studies in Fuzziness and Soft Computing*, pp.: 138- 233. Physica-Verlag.

SZYMON J. & SIMOVICI, D. A. (201). ‘A General Measure of Rule Interestingness’. In: L. de Raedt & A. Siebes (Eds.). *Principles of Data Mining and Knowledge Discovery*, 2001, LNAI 2168, pp 253-265.Springer-Verlag.

THOMPSON, S. (1992). *Sampling*. John Wiley.

THOMPSON, S. K. & SEBER, G. A. F. (1996). *Adaptive Sampling*. John Wiley

TREMMEL, L. (1995). ‘The visual separability of plotting symbols in scatterplots’. *Journal of Computational and Graphical Statistics*. **4** (**2**), 101-112

TUKEY, J. W. (1962). ‘The future of data analysis’. *Ann. Math. Statist*. **33** (**1**), 67 and 812.

TUKEY, J. W. & WILK, M. B. (1966). ‘Data Analysis and Statistics: an expository overview’. *In: L.V. JONES (Ed.). The Collected Works of John W. Tukey . Wadsworth & Brooks/Cole*, pp:. 549-578.

TUKEY, J. W. (1972). Data Analysis, Computation and Mathematics. *In: L.V. JONES (Ed.). The Collected Works of John W. Tukey . Wadsworth & Brooks/Cole*, pp:. 753-775.

TUKEY, J. W. (1984). Data Analysis: History and Prospects. *In: L.V. JONES (Ed.). The Collected Works of John W. Tukey . Wadsworth & Brooks/Cole*, pp:. 985-1001.

UNWIN, A.; HAWKINS, G.; HOFMANN, H. & SIEGL, B. (1996). ‘Interactive Graphics for Data Sets With Missing values- MANET’. *Journal of Computational and Graphical Statistics.* **5** (**2**), 113-122.

VALOIS, J. P. (2000). ‘Approche graphique en analyse des données’. *Journal de la Societé Française de Statistique.* **141** (**4**), 5- 41.

VAPNIK, V. (1998). *Statistical Learning Theory*. John Wiley.

VARELLA-NUALLES, M. (2002). *Los Métodos Biplot como Henamienta de Análisis de Interacción de Orden Superior en un Modelo Lineal/Bilineal.* Tesis Doctoral. Universidad de Salamanca.

VAZQUEZ, M. (1995). *Aportaciones al Análisis Biplot: Un enfoque algebraico.* Tesis Doctoral.Universidad de Salamanca.

VICENTE -TAVERA, S. (1992). *Las Técnicas de Representación de Datos Multidimensionales en el Estudio del Índice de Producción Industrial en la C.E.E.* Tesis Doctoral.Universidad de Salamanca.

VICENTE-VILLARDÓN, J. L. (1992). *Una Alternativa a las Técnicas Factoriales Clásicas Basada en una Generalización de los Métodos BIPLOT*. Tesis Doctoral, Universidad de Salamanca.

WARE, C. (2000). *Information Visualization*. Morgan Kaufman.

WEGMAN, E. J. (1995). 'Huge data sets and the frontiers of computational feasibility'. *Journal of Computational and Graphical Statistics*. **4** (**4**), 281-295.

WILKINSON, L. (1999). *The Grammar of Graphics*. Springer.

WILLS, G. (1999). 'Niche works. Interactive visualization of very large graphs'. *Journal of Computational and Graphical Statistics*. **8** (**2**), 190-212.

YOUNG, G. & HOUSEHOLDER, A. S. (1938). 'Discussion of a set of points in terms of their mutual distances'. *Psychometrika***. 3** (**1**), 19-22.

ZADEH, L. A. (1965). Fuzzy Sets. *Inform. Control*. **8 (1965**), 338-353.