# anuario 1992

## INSTITUTO DE ESTUDIOS ZAMORANOS FLORIAN DE OCAMPO

# ANUARIO 1992

INSTITUTO DE ESTUDIOS ZAMORANOS
"FLORIÁN DE OCAMPO" (C.S.I.C.)

# anuario 1992

# INSTITUTO DE ESTUDIOS ZAMORANOS FLORIAN DE OCAMPO

ISSN.: 0213-82-12
Depósito Legal: ZA - 297 - 1988
Imprime: HERALDO DE ZAMORA. Santa Clara, 25 - ZAMORA
artes gráficas

# ÍNDICE

## ARTICULOS

# ARTÍCULOS

# ENOLOGÍA

# CARACTERIZACIÓN DE VINOS DE TORO MEDIANTE TÉCNICAS QUIMIOMÉTRICAS DE ANÁLISIS MULTIVARIANTE

M.ª CRUZ ORTIZ FERNÁNDEZ
LUIS ANTONIO SARABIA PEINADOR

## INTRODUCCIÓN

En la solicitud de ayuda económica para la realización del presente trabajo se argumentó el interés de caracterizar vinos tintos de la Denominación de Origen de Toro a través de medidas físico-químicas más objetivas y reproducibles que los análisis sensoriales tradicionalmente empleados para mantener la identidad de la producción vinícola de una zona geográfica.

Una herramienta especialmente útil para establecer la tipicidad de un producto alimentario, en este caso los vinos de la D. O. Toro, es la Quimiometría. Esta disciplina de reciente creación tiene como finalidad extraer información útil de las determinaciones químicas y en particular se incluyen como capítulos específicos suyos las técnicas multivariantes para estudiar las relaciones entre las cantidades químicas que describen una muestra alimentaria y su origen.

En la actualidad las técnicas quimiométricas pueden considerarse como un nuevo instrumento en el laboratorio, junto a los instrumentos físicos. El uso correcto de esta metodología es una preocupación constante pues de ello depende en gran medida el que la enorme cantidad de información química proporcionada por el moderno utillaje de los laboratorios sea eficaz para resolver problemas de calidad alimentaria. Más todavía, estamos convencidos de que en Quimiometría alimentaria los análisis de datos conducentes a la determinación del origen, caracterización y evaluación de la calidad no pueden ser realmente efectivos si no se valoran a la par los métodos usados para alcanzar las conclusiones.

Por ello se ha pretendido dar a esta memoria un formato externo discursivo, que permita su lectura y valoración por parte de personas no expertas en la metodología estadística multivariante utilizada que sin embargo son los destinatarios naturales de la misma: todas las personas vinculadas a la elaboración del vino y a la potenciación de la calidad específica de los vinos la D. O. de Toro.

La necesidad de una cuidadosa elaboración diferenciada de los vinos que permita mantener una identidad económica, cultural y geográfica es una tarea digna de los mejores esfuerzos y mucho más en el ámbito castellano-leonés. Hemos querido aportar a esta inquietud nuestros mejores conocimientos científicos para depurar la información contenida en las cuarenta y tres cantidades químicas

medidas sobre vinos procedentes de veinticuatro bodegas y cooperativas adscritas a las D. O. de Toro y Ribera de Duero.

La finalidad del trabajo es construir un modelo para los vinos de Toro que permita no sólo diferenciarlos de otros sino también establecer qué parámetros químicos expresan su personalidad propia. Obviamente no se puede establecer tal modelo para los vinos de la D. O. de Toro si no es por referencia a otros perfectamente identificados como ajenos, es decir pertenecientes a otra D. O. que garantice su procedencia. El modelado es tanto más interesante –y más comprometido– cuanto más próximos sean los vinos, ya que ello puede conducir a una progresiva igualdad facilitada por la similitud del clima, la proximidad geográfica y la variedad genética común de las vides.

La memoria describe no sólo los resultados, sino el argumento lógico-formal que los soporta, incluso en ocasiones se ha apuntado la imposibilidad de usar determinados procedimientos estadísticos de análisis de datos a causa de las propiedades matemáticas que exhiben los aquí analizados.

Desde estos párrafos iniciales proponemos al lector la aventura casi detectivesca de alcanzar formalmente la evidencia objetiva de que los vinos de Toro son diferenciables con gran especificidad, que esta diferenciabilidad es medible, que una vez determinadas las cantidades químicas que la expresan es posible incidir en la personalidad propia de estos vinos de una manera eficaz y controlada ya que en manos del enólogo el instrumento descriptivo construido por la quimiometra se convierte en la razón de modificar un proceso de vinificación.

En la medida en que lo logremos nos habremos sumado a una corriente europea actual y dinámica que propone el uso de todo el potencial de los análisis de datos junto con las modernas técnicas de análisis químico para diferenciar y mantener la especificidad de los productos alimentarios que son compendio de cultura y tradición así como garantía económica de futuro.

La memoria está dividida en tres capítulos y se apoya en veinte gráficos como una manera eficaz de resumir y exponer las ideas que las veinticuatro tablas encierran con excesivo celo en ocasiones. Después de una descripción de las muestras y de las variables usadas en el primero; se pasa a establecer, en el segundo, las propiedades básicas de la tabla de datos que condicionarán todo el análisis posterior: la ausencia de normalidad, altas correlaciones, una estructura latente con sólo cuatro factores y la imposibilidad de obtener una adecuada separación entre ambas clases son los límites que definen la selección de los procedimientos a usar.

El tercero se dedica propiamente a la construcción de los modelos y a su evaluación. Después de una breve introducción dedicada a establecer la diferencia entre un análisis de clasificación y un modelado se comienza aportando evidencia directa de que es posible alcanzar la meta propuesta mediante análisis de

agrupamiento que permiten "ver" la proximidad de unas muestras con otras en el espacio de cuarenta y tres dimensiones en que tenemos planteada la cuestión, además toma cuerpo la idea de la necesidad de usar la estructura latente de los datos en vez de las cantidades originales. El análisis cluster que se muestra en las páginas 23 y siguientes, no sólo es una pieza de convicción sino que en sí mismo alcanza una elevada capacidad de discriminar las muestras, poniendo de relieve interesantes similitudes y diferencias mediante los dendogramas de las páginas 26 y 27.

El análisis prosigue estableciendo la validez de una clasificación de las muestras basada en el origen de las que están próximas –método KNN– para desembocar en un análisis detallado de los grupos de variables y su capacidad para establecer un modelo mediante el procedimiento SIMCA de gran utilidad y eficacia en el ámbito de la Quimiometría alimentaria, es de destacar la inestimable ayuda de los Diagramas de Coomans en esta secuencia de interpretaciones y comparaciones consecutivas para llegar a construir un modelo con elevada capacidad de clasificación y predicción junto con una gran especificidad y sensibilidad. Se demuestra la inutilidad de algunos parámetros químicos en esta tarea esto es de importancia porque en general las determinaciones que no aportan información al problema sólo contribuyen a enmascarar la situación.

Para evitar cualquier concesión a la duda se ha construido otro modelo basado en principios totalmente distintos de SIMCA. Se trata en gran medida de una innovación en el sentido de que se ha usado PLS (Partial Least Squares) como técnica de modelado. Los resultados han superado con creces las expectativas como puede comprobarse el epígrafe correspondiente corroborando la personalidad de los vinos tintos de la D. de Origen Toro a través de las cantidades químicas estudiadas.

Tanto el capítulo dos como el tres finalizan con unas conclusiones parciales a modo de resumen para facilitar la estructuración de los resultados obtenidos.

No podemos cerrar esta introducción sin poner de relieve una idea básica, pero a veces olvidada. El quehacer científico, en especial cuando intenta abordar problemas que interesan al entorno social en el que se desenvuelve, es deudor de otros muchos quehaceres previos a él: transmite conocimientos que otros idearon, utiliza y se fundamenta en resultados empíricos contrastados, se contagia del entusiamo de los emprendedores pero realmente aportará alguna solución práctica en tanto en cuanto sea capaz de considerar el mismo problema, que otros trataron, desde una perspectiva nueva.

De algún modo esta memoria habría sido muy distinta, tal vez imposible, sin la financiación del Instituto de Estudios Zamorano "Florián de Ocampo", sin los análisis realizados en el Departamento de Nutrición y Bromatología de la Universidad de Salamanca bajo la dirección del Dr. Rivas, sin el entusiasmo por todo lo

relacionado con el vino tinto de la enólogo Dña. Lucía García de María que desde el Consejo Regulador de la D. O. de Ribera de Duero y su Bodega Experimental prestó su apoyo y colaboración en la recogida de muestras y organización de los análisis o sin el aliento de personas como el prof. Forina del Instituto de Tecnologie Farmaceutiche ad Alimentari de la Universidad de Génova y expresidente de la Chemometrics Society que puso todo su saber en Quimiometría alimentaria a nuestra disposición. A su vez esperamos que la investigación recogida en esta memoria colabore a mantener y elevar la calidad de los vinos de la D. O. de Toro y motive a personas e instituciones a continuar en el uso y promoción de la Quimiometría alimentaria.

## PROBLEMA

Caracterización de los vinos de Toro mediante técnicas Quimiométricas de clasificación multivariante aplicadas a los datos físico-químicos.

## 1. DATOS

Se dispone de una matriz de datos, cuyas filas son las muestras analizadas a las que nos referiremos como "objetos" y cuyas columnas son las variables físico-químicas medidas en cada muestra.

Los datos recogen un amplio espectro de determinaciones analíticas sobre vinos jóvenes: i) parámetros enológicos convencionales, ii) polifenoles, iii) contenido en compuestos químicos responsables del color: antocianos, taninos y sus combinaciones. La caracterización en base a estos compuestos persigue dilucidar la especificidad del vino, en nuestro caso de la Denominación de Origen "Toro".

La comparación se ha llevado a cabo con vinos próximos en lo geográfico, lo climático y lo varietal con la condición de ser también específicos, son vinos de la Denominación de Origen "Ribera de Duero", de este modo las conclusiones serán realmente significativas.

### 1.1. Objetos

Son muestras de vinos tintos jóvenes de las vendimias de 1985, 1986 y 1987 procedentes de bodegas adscritas a las respectivas Denominaciones de Origen.

Junto a la identificación de cada bodega se ha anotado el número de la muestra y el año de la vendimia.

*19 muestras de la D. O. Toro de las bodegas:*

— Coop. Vino de Toro ($43_{85}$, $44_{85}$, $46_{86}$, $50_{87}$, $64_{85}$, $66_{85}$).
— Bod. José María Fermoselle. Zamora ($48_{86}$, $52_{87}$).

— Bod. Luis Mateos. Toro ($45_{86}$, $47_{86}$, $51_{87}$, $60_{85,87}$, $61_{85,86}$).
— Bod. Fariña. Toro ($49_{86}$).
— Coop. Nuestra Sra. de las Viñas. Morales de Toro ($53_{87}$, $65_{86}$).
— Bod. Porto. Toro ($62_{85}$, $63_{85c}$).

*47 muestras de vino de la D. O. Ribera de Duero de las bodegas:*

— Grandes Bodegas S.A. Roa ($1_{85}$).
— Coop. Virgen de las Viñas. Aranda de Duero ($2_{85}$, $8_{86}$).
— Vinos García S.A. Aranda de Duero. ($3_{85}$).
— Bod. Valduero S.A. Gumiel del Mercado ($4_{85}$, $11_{86}$, $14_{86}$, $18_{86}$).
— Coop. Virgen de la Asunción. La Horra. ($5_{86}$).
— Coop. Santa Eulalia. La Horra ($6_{86}$).
— Coop. San Roque ($7_{86}$).
— Coop. Stma. Trinidad. Fuentespina ($9_{86}$, $17_{86}$, $21_{86}$, $26_{86}$, $26_{87}$, $27_{87}$, $28_{87}$, $31_{87}$, $33_{87}$).
— Coop. Virgen de Fátima. Pedrosa de Duero ($10_{86}$, $22_{86}$, $23_{87}$).
— Coop. San Roque del Encinar. Castrillo de la Vega ($12_{86}$, $19_{86}$, $20_{86}$).
— Coop. Ribera de Duero. Peñafiel ($13_{86}$, $15_{86}$, $37_{87}$, $38_{87}$, $34_{87}$, $36_{87}$).
— Coop. Santísimo Cristo del Consuelo. Baños de Valdeavad ($25_{87}$).
— Coop. Virgen del Rosario. Quintanamanvirgo ($35_{87}$).
— Bod. Experimental. Ribera de Duero ($55_{85}$).
— Pesquera ($56_{85}$).
— Pérez Pascuas ($57_{85}$).
— Viña Pedrosa ($59_{86}$).
— Peñafiel $858_{85}$).

### 1.2. **Variables**

Las siguientes 43 variables han sido determinadas en cada muestra de vino. Se anota seguidamente la codificación usada.

*Parámetros enológicos convencionales:*

GA, grado alcohólico, (% etanol).
AV, acidez volátil (gr. ácido acético/l).
AT, acidez total (gr. ácido tartárico/l).
pH

*Compuestos polifenólicos:*

AP, polifenoles totales, reactivo de Folin-Ciocalteau (mg. de ácido gálico/l).
BP, índice de Somers y Evans de polifenoles totales.
C, catequina, sustancias reactivas a la vainillina (mg. catequina/l).

DP, proantocianidol, método de Peri y Pompei (mg. de cloruro de cianidol/l).
EP, proactocianidol, método de Aubert (mg. de cloruro de cianidol/l).
V/F, relación de contenidos de catequinas y proantocianidoles.
V/LA, relación de contenidos de catequinas y proantocianidoles.
S/V, relación de catequinas e índice de polifenoles totales de Somers y Evans.

*Estructura de los antocianos libres y acilados*

DF, 3-monoglucósido de definidol.
CI, 3-monoglucósido de cianidol.
PT, 3-monoglucósido de petudinol.
PE, 3-monoglucósido de peonidol.
MV, 3-monoglucósido de malvidol.
DFA, acetato del 3-monoglucósido de definidol.
CIA, acetato del 3-monoglucósido de cianidol.
PTA, acetato del 3-monoglucósido de petudinol.
PEA, acetato del 3-monoglucósido de peonidol.
MVA, acetato del 3-monoglucósido de malvidol.
DFC, cumarato del 3-monoglucósido de definidol.
CIC, cumarato del 3-monoglucósido de cianidol.
PTC, cumarato del 3-monoglucósido de petudinol.
PEC, cumarato del 3-monoglucósido de peonidol.
MVC, cumarato del 3-monoglucósido de malvidol.

Estas variables están expresadas en mg. 3-monoglucósido de malvidol/l y se han determinado por cromatografía líquida de alta eficacia (HPLC).

*Antocianos totales y parámetros relacionados con el color y polimerización*

IC, intensidad colorante (determinada con las absorbancias a 420 y 520 nm).
IC2, intensidad colorante (determinada con las absorbancias a 420, 520 y 620 nm).
Tono, tono del color, método de Glories.

— Indices de ionización (porcentaje de antocianos en forma catiónica coloreada).

II, índice de ionización.
IIS, índice de ionización según Somers.
IIMS, índice de ionización modificado de Somers.

— Antocianos totales.
ACG, antocianos totales (método de Glories).
MONB, porcentaje de monómeros (método de Bourzeix).
PRB, porcentaje de polímetros rojos (método de Bourzeix).
PPB, porcentaje de polímetros pardos (método de Bourzeix).
PVP, índice de polivinilpirrolidona.

ACTS, antocianos totales (método de Somers).
ACIS, antocianos ionizados (método de Somers).

— Índices de polimerización de polifenoles:
EQ1, edad química como relación de la absorbancia a 520 nm. con adición de $SO_2$ y con adición de acetaldehído.
EQ1, edad química como relación de la absorbancia a 520 nm. con adición de $SO_2$ y con adición de ácido clorhídrico.
INDP, índice de polimerización.

La tabla de datos a analizar consta de 66 filas (las muestras de vino) y 43 columnas (las variables medidas).

Los valores de estas variables han sido determinadas por Soledad S. Muñoz Hernaz y A. M. Polanco Álvarez (1988, Departamento de Química Analítica, Nutrición y Bromatología, Universidad de Salamanca). En las Memorias de Licenciatura correspondientes se encuentran todas los detalles relativos a las técnicas experimentales utilizadas en la determinación de cada variable.

La utilización de este conjunto de datos para la construcción de un modelo empírico para los vinos de la D. O. Toro, es decir, para su caracterización, se debe a la enorme cantidad de información química que encierran, de modo que aún cuando con ellos se ha procedido a una diferenciación inicial entre ambas denominaciones, la complejidad estadística que presentan y el hecho de que no se han usado conjuntamente todas las variables han sido las razones para abordar el problema objeto de este estudio.

Conviene destacar que en este conjunto de datos se han usado técnicas analíticas de gran reproducibilidad (las determinaciones mediante cromatografía líquida de alta eficacia) al tiempo que se dispone sobre las mismas muestras de los valores de otros parámetros "convencionales". Es una buena oportunidad para determinar qué tipo de parámetros caracterizan a los vinos de Toro, en este caso en relación a los de Ribera de Duero, contribuyendo a definir su "personalidad".

## 2. ANÁLISIS PREVIO DE LOS DATOS

Muchas de las técnicas estadísticas se basan en la hipótesis de normalidad de los datos empíricos. Por ello un paso inicial es el estudio de la normalidad de cada variable.

### 2.1. Normalidad

Cinco test se han aplicado a cada variable porque cada una de ellos explora aspectos distintos del posible fallo de normalidad. Para su referencia se les anotará por la letra que se indica:

D - Test de Kolmogoroff, distribución exacta de Lillefors.
V - Test de Kuiper.

W - Test de Cramer Von Mises.
U - Test de Watson.
A - Test de Anderson-Darling.

Cada uno de los tests contrasta la hipótesis nula: "Los datos proceden de una distribución normal" frente a la alternativa: "Los datos no proceden de una normal".

El nivel de significación es el riesgo que el experimentador está dispuesto a asumir al tomar la decisión de rechazar la hipótesis nula, puesto que por razones aleatorias, unos datos pueden conducir a rechazar la normalidad cuando realmente procedían de una distribución normal. Habitualmente se fija en 0.05.

Los resultados se encuentran codificados en la tabla 1 de doble entrada, cada fila es una variable y cada columna un test, en la casilla en que se cruzan fila y columna se anota el resultado de aplicar el test a la variable. Las anotaciones se leen del siguiente modo:

i) Una "a" indica que no hay evidencia experimental para rechazar la normalidad de los datos a cualquier nivel de significación superior al 0.10.
ii) Una "r" indica que hay que rechazar la hipóteiss de normalidad si el experimentador desea tener un nivel de significación entre el 0.05 y el 0.10.
iii) Finalmente "rr" señala que a un nivel entre 0.05 y el 0.01 habrá de rechazarse.

Sin necesidad de hacer un análisis detallado de la tabla 1, es clara la ausencia de normalidad en las variables objeto de estudio, sólo en 17 de ellas no se tiene evidencia suficiente para rechazar la normalidad.

El mismo análisis de normalidad se ha llevado a cabo en cada categoría. De otro modo, se trata de decidir si las variables siguen una distribución normal para los vinos de Toro (categoría 1) distinta de la de los vinos de Ribera de Duero (categoría 2) en cuyo caso mostrarían ausencia de normalidad en el estudio conjunto.

Es necesario recordara que el tamaño de la muestra incide en el nivel de significación del test, especialmente en aquellos que se basan en agrupar datos en clases. Cuanto menor es el tamaño muestral más conservador se vuelve el test, o lo que es igual hace falta más evidencia experimental para rechazar la hipótesis de normalidad cuando se tienen pocos datos que cuando se tienen muchos. En nuestro caso sólo disponemos de 19 muestras de vinos de Toro frente a las 47 de Ribera de Duero.

El resultado del análisis de normalidad sobre las categorías se recoge en la tabla 2. Una primera lectura señala que las variables medidas se comportan de distinta manera en una categoría que en otra. Sin embargo después de lo dicho en el párrafo anterior no puede concluirse que este efecto sea realmente cierto y no esté causado por los tamaños muestrales.

**Tabla 1. Normalidad de las variables**

| | | D | V | W | U | A |
|---|---|---|---|---|---|---|
| 1 | GA | a | r | a | a | a |
| 2 | AV | rr | rr | rr | rr | rr |
| 3 | AT | rr | rr | rr | rr | rr |
| 4 | pH | a | r | r | r | r |
| 5 | AP | rr | rr | rr | rr | rr |
| 6 | BP | rr | rr | rr | rr | rr |
| 7 | C | r | rr | rr | rr | rr |
| 8 | DP | rr | rr | rr | rr | rr |
| 9 | EP | rr | rr | rr | rr | rr |
| 6 | BP | rr | rr | rr | rr | rr |
| 10 | VF | rr | rr | rr | rr | rr |
| 11 | VLA | rr | rr | rr | rr | rr |
| 12 | SV | rr | rr | rr | rr | rr |
| 13 | DF | a | r | a | a | a |
| 14 | CI | rr | rr | r | r | a |
| 15 | PT | rr | rr | rr | rr | rr |
| 16 | PE | rr | rr | rr | rr | rr |
| 17 | MV | rr | rr | rr | rr | rr |
| 18 | DFA | rr | rr | rr | rr | rr |
| 19 | CIA | r | rr | r | r | rr |
| 20 | PTA | a | a | a | a | a |
| 21 | PEA | rr | rr | rr | rr | rr |
| 22 | MVA | a | a | a | a | a |
| 23 | DFC | a | a | a | a | a |
| 24 | CIC | rr | rr | rr | rr | rr |
| 25 | PTC | a | a | a | a | a |
| 26 | PEC | a | a | r | r | r |
| 27 | MVC | rr | rr | rr | rr | rr |
| 28 | IC | a | a | a | a | a |
| 29 | IC2 | a | a | a | a | a |
| 30 | TONO | a | a | a | a | a |
| 31 | II | a | a | a | a | a |
| 32 | IIS | rr | rr | rr | rr | rr |
| 33 | IIMS | rr | rr | rr | rr | rr |
| 34 | ACG | a | a | a | a | a |
| 35 | MONB | a | a | a | a | a |
| 36 | PRB | a | a | a | a | a |
| 37 | PPB | a | r | r | r | r |
| 38 | PVP | a | a | a | a | a |
| 39 | ACTS | a | a | a | a | a |
| 40 | ACIS | a | a | a | a | a |
| 41 | EQ1 | a | a | a | a | a |
| 42 | EQ2 | r | rr | r | a | r |
| 43 | INDP | r | rr | r | a | r |

Se anota: rr para $p <= 0.01$
r para $0.01 < p <= 0.05$
a para $0.05 < p$

**Tabla 2. Normalidad de las variables por categorías**

| | | CAT. 1 (TORO) | | | | | CAT. 2 (RIBERA) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | D | V | W | U | A | D | V | W | U | A |
| 1 | GA | r | rr | a | a | a | r | rr | r | r | r |
| 2 | AV | a | a | a | a | a | rr | rr | rr | rr | rr |
| 3 | ATa | a | a | a | a | a | rr | rr | rr | rr | rr |
| 4 | pH | a | a | a | a | a | a | r | r | rr | r |
| 5 | AP | a | a | a | a | a | a | r | r | rr | r |
| 6 | BP | a | a | a | a | a | rr | rr | rr | rr | rr |
| 7 | C | a | a | a | a | a | rr | rr | rr | rr | rr |
| 8 | DP | rr | rr | rr | r | rr | rr | rr | rr | rr | rr |
| 9 | EP | a | a | a | a | a | rr | rr | rr | rr | rr |
| 10 | VF | a | a | a | a | a | rr | rr | rr | rr | rr |
| 11 | VLA | a | a | a | a | a | rr | rr | rr | rr | rr |
| 12 | SV | a | a | a | a | a | rr | rr | rr | rr | rr |
| 13 | DF | a | a | a | a | a | r | rr | r | r | r |
| 14 | CI | a | a | a | a | a | rr | rr | rr | rr | r |
| 15 | PT | a | a | a | a | a | r | r | a | a | a |
| 16 | PE | a | r | a | a | a | a | a | a | a | r |
| 17 | MV | a | a | a | a | a | r | rr | rr | rr | rr |
| 18 | DFA | r | r | r | r | r | a | r | r | r | r |
| 19 | CIA | a | a | a | a | r | a | r | a | a | r |
| 20 | PTA | a | a | a | a | r | a | a | a | a | a |
| 21 | PEA | a | r | r | r | rr | a | a | a | a | a |
| 22 | MVA | a | a | a | a | a | r | rr | r | r | r |
| 23 | DFC | a | a | a | a | a | a | a | a | a | r |
| 24 | CIC | a | r | r | r | rr | rr | rr | rr | rr | rr |
| 25 | PTC | a | a | a | a | a | r | r | a | a | a |
| 26 | PEC | a | a | a | a | a | a | a | a | r | r |
| 27 | MVC | a | a | a | a | a | rr | rr | rr | rr | rr |
| 28 | IC | a | a | a | a | a | a | a | a | a | a |
| 29 | IC2 | a | a | a | a | a | a | a | a | a | a |
| 30 | TONO | a | a | a | a | a | a | a | a | a | a |
| 31 | II | a | a | a | a | a | a | a | a | a | a |
| 32 | IIS | a | a | a | a | a | rr | rr | rr | rr | rr |
| 33 | IIMS | a | a | a | a | a | rr | rr | rr | rr | rr |
| 34 | ACG | a | a | a | a | a | a | a | a | a | a |
| 35 | MONB | a | a | a | a | a | a | r | rr | rr | rr |
| 36 | PRB | a | a | a | a | a | a | a | a | a | a |
| 37 | PPB | a | a | a | a | a | r | r | r | r | r |
| 38 | PVP | a | a | a | a | a | a | a | a | a | a |
| 39 | ACTS | a | a | a | a | a | a | a | a | a | a |
| 40 | ACIS | a | a | a | a | a | rr | rr | rr | r | rr |
| 41 | EQ1 | a | a | a | a | a | a | a | a | a | a |
| 42 | EQ2 | a | a | a | a | a | r | r | r | a | r |
| 43 | INDP | a | a | a | a | r | r | rr | r | a | r |

Se anota: rr para $p <= 0.01$
r para $0.01 < p <= 0.05$
a para $0.05 < p$

En todo caso, es indudable que no es posible considerar normales los datos ni globalmente ni por categorías, ya que al menos en la de los vinos de Ribera se tiene un fallo de normalidad mayoritario. Este resultado no ha de entenderse en el sentido de que estas variables presenten algún tipo de anomalía, de hecho en la práctica es una situación más frecuente que la contraria. Sin embargo condiciona completamente la metodología estadística a seguir en el análisis de estos datos, en el sentido de que no deben usarse métodos de clasificación o modelado que usen la hipótesis de normalidad, y en caso de hacerlo habrán de tenerse las lógicas precauciones a la hora de interpretar los resultados.

### 2.2. **Estudio de la capacidad discriminante univariante**

Otro aspecto previo es la relevancia univariante de las variables, ello se hace mediante el peso de Fisher. Las variables con mayor peso de Fisher son aquellas que separan mejor ambas categorías sin tener en cuenta las relaciones internas entre las variables.

Para cada variable el peso de Fisher es proporcional al cociente de la varianza entre las medias de cada categoría y la varianza dentro de cada una de las categorías, mientras que el peso modificado tiene una corrección para tener en cuenta la posibilidad de que la variable en estudio tenga distinta varianza en cada categoría.

Los valores de estos dos índices forman la tabla 3. A excepción del grado alcohólico (2.186) y el índice de Somers y Evans de polifenoles totales (1.303) las restante variable muestran un peso de Fisher por debajo de 0.7 y la mayoría notablemente inferior. Con la corrección debida a la varianza aumentan notablemente los pesos de:

i) V/F, relación de catequinas y polifenoles totales de (0.033 a 1.551).
ii) PEA, acetato del 3-monoglucósido de peonidol de (0.198 a 0.964).
iii) IIMS, índice de ionización de antocianos modificado de Somers (de 0.226 a 0.878).

A la vista de estos resultados se concluye que a excepción del grado alcohólico y del índice de polifenoles totales ninguna de las restantes variables considerada una a una tiene capacidad discriminante suficiente, o lo que es igual: la variabilidad que muestran no se debe al distinto origen de los vinos.

Para visualizar cómo en realidad no es posible clasificar con estas variables se ha procedido a representar los diagramas Box-Whisker de las variables con mayor peso de Fisher y/o modificado.

Estos diagramas constan de:

i) Una "caja" cuyo límite inferior ($Q_1$) es el primer cuartil y el superior ($Q_3$)

**Tabla 3. Peso para clasificación univariante**

| | | P. Fisher | P. Modificado |
|---|---|---|---|
| 1 | GA | **2.186** | **2.187** |
| 2 | AV | 0.399 | 0.545 |
| 3 | AT | 0.110 | 0.156 |
| 4 | pH | 0.000 | 0.000 |
| 5 | AP | 0.695 | 0.703 |
| 6 | BP | **1.303** | **1.304** |
| 7 | C | 0.393 | 0.397 |
| 8 | DP | 0.318 | 0.324 |
| 9 | EP | 0.459 | 0.557 |
| 10 | VF | 0.033 | **1.551** |
| 11 | VLA | 0.162 | 0.647 |
| 12 | SV | 0.164 | 0.169 |
| 13 | DF | 0.080 | 0.084 |
| 14 | CI | 0.022 | 0.074 |
| 15 | PT | 0.000 | 0.002 |
| 16 | PE | 0.293 | 0.374 |
| 17 | MV | 0.001 | 0.003 |
| 18 | DFA | 0.083 | 0.304 |
| 19 | CIA | 0.015 | 0.064 |
| 20 | PTA | 0.050 | 0.087 |
| 21 | PEA | 0.198 | **0.964** |
| 22 | MVA | 0.000 | 0.024 |
| 23 | DFC | 0.005 | 0.166 |
| 24 | CIC | 0.004 | 0.013 |
| 25 | PTC | 0.538 | 0.560 |
| 26 | PEC | 0.002 | 0.002 |
| 27 | MVC | 0.001 | 0.002 |
| 28 | IC | 0.633 | 0.645 |
| 29 | IC2 | 0.633 | 0.643 |
| 30 | TONO | 0.425 | 0.484 |
| 31 | II | 0.028 | 0.038 |
| 32 | IIS | 0.121 | 0.448 |
| 33 | IIMS | 0.226 | 0.878 |
| 34 | ACG | 0.177 | 0.178 |
| 35 | MONN | 0.192 | 0.275 |
| 36 | PRB | 0.089 | 0.217 |
| 37 | PPB | 0.191 | 0.257 |
| 38 | PVP | 0.189 | 0.302 |
| 39 | ACTS | 0.255 | 0.258 |
| 40 | ACIS | 0.016 | 0.052 |
| 41 | EQ1 | 0.345 | 0.414 |
| 42 | EQ2 | 0.028 | 0.153 |
| 43 | INDP | 0.025 | 0.157 |

el tercero; de modo que encierra la mitad central de los datos. La mediana, segundo cuartil ($Q_2$), está representada por la línea horizontal interior a la caja.

ii) Unas líneas verticales que unen el punto central de los lados horizontales de la caja con un valor adyacente. El valor adyacente superior es el mayor dato que es menor o igual de la cota $Q_3 + 1.5\ (Q_3 - Q_1)$. Análogamente el valor adyacente inferior es el menor dato mayor o igual que $Q_1 - 1.5\ (Q_3 - Q_1)$.

iii) Los datos que se encuentran fuera de los valores adyacentes se representan aisladamente porque pueden ser datos anómalos.

La figura 1 muestra que el grado alcohólico tiene un valor mediano (13,8) superior en los vinos de Toro que en los de Ribera (12,4) pero sin embargo no es suficientemente específica ya que al menos un veinticinco por ciento de las muestras de cada origen toman valores entre los dos cuartiles centrales de la otra. Incluso, más del 50% de los valores de grado alcohólico de los vinos de Ribera son valores admisibles para los vinos de Toro; análogamente el 75% de los de Toro serían admisibles como de Ribera.

La figura 2, relativa al índice de Somers y Evans de polifenoles totales (BP) tiene un comportamiento similar. En todo caso cabe significar cómo los valores para los vinos de Ribera están más agrupados, aspecto este que no detecta la varianza por la existencia de cinco objetos muy dispares de los restantes con valores ostensiblemente mayores.

La relación catequinas/polifenoles totales (V/F) está representada en la figura 3. Es evidente que su peso modificado alto se debe a los objetos 1, 10 y 11 de Ribera de Duero singularmente distintos de los demás. Obviamente carece de capacidad discriminante.

La figura 4 evidencia que el peso modificado alto para el 3-monoglucósido de peonidol (PEA) se debe en gran medida al objeto 65 de Toro.

Otro tanto se puede decir al observar la figura 5 respecto del índice de ionización modificado de Somers, en este caso son los objetos 55, 11 y 20 de Ribera los responsables del incremento del peso modificado.

En resumen, ninguna de las variables permite decidir el origen de una muestra entre Toro y Ribera del Duero.

### 2.3. **Correlaciones**

También es necesario un estudio previo de la matriz de las correlaciones entre las variables, por categoría y globalmente. Cada elemento de esta matriz indica el grado de relación lineal entre cada pareja de variables, su estructura define las direcciones de máxima elongación de la nube que forman los objetos en el espacio de 43 dimensiones. Es importante dilucidar dos cuestiones: hasta qué punto las matrices de correlaciones en cada categoría son similares y qué variables aportan información redundante desde el punto de vista estadístico. La segunda cuestión

FIGURA 1

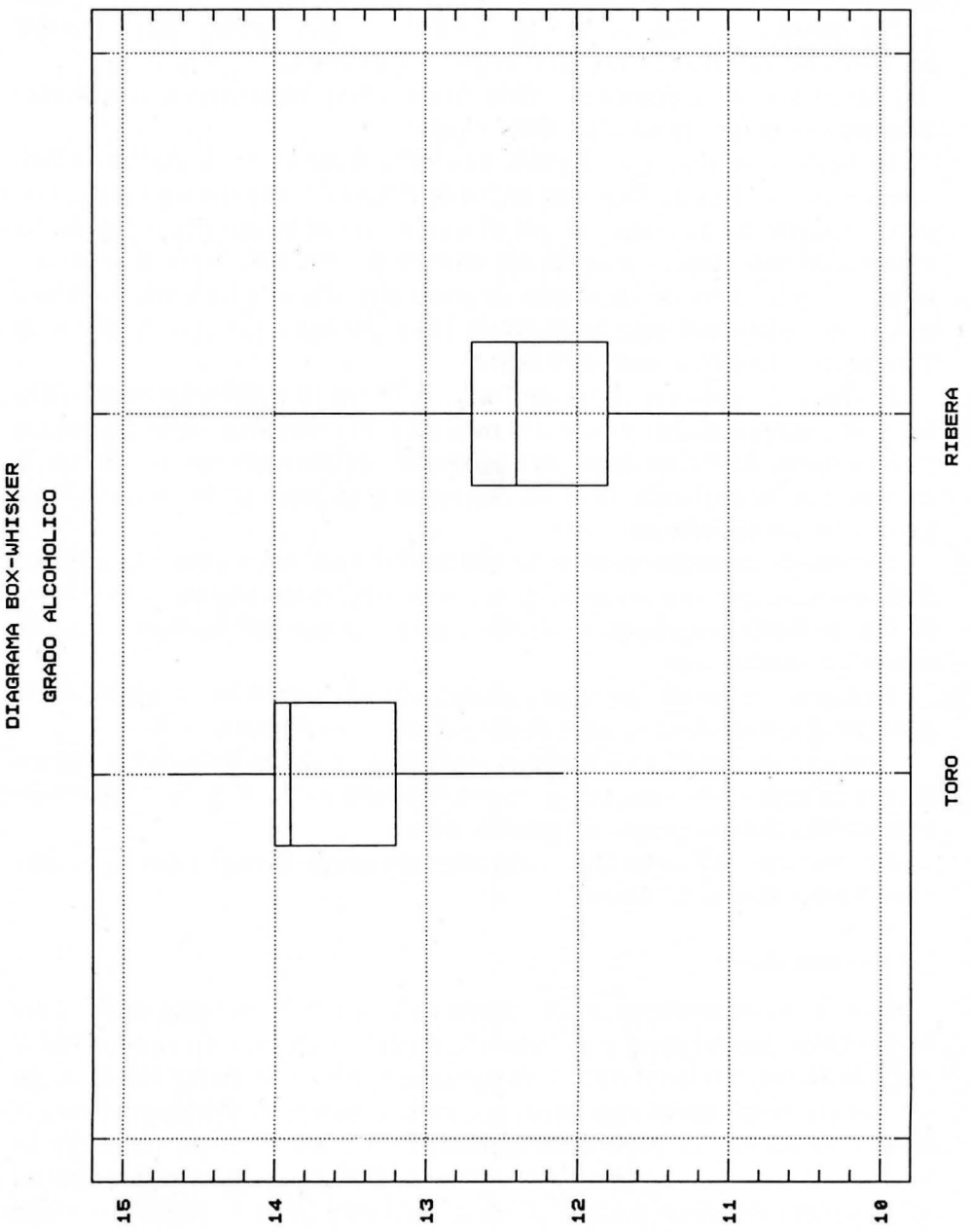


DIAGRAMA BOX-WHISKER
GRADO ALCOHOLICO
15
14
13
12
11
10
TORO
RIBERA

FIGURA 2

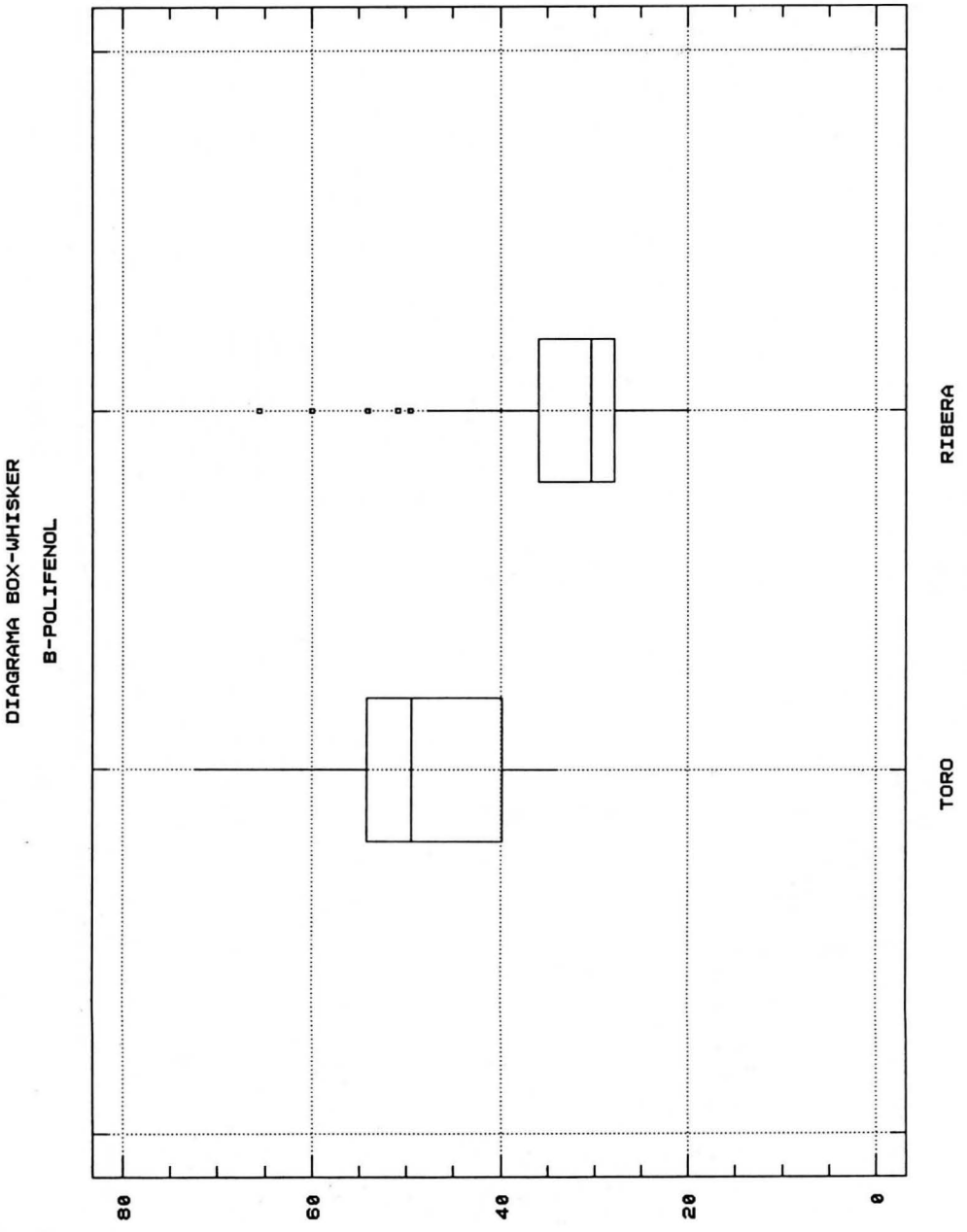


FIGURA 3

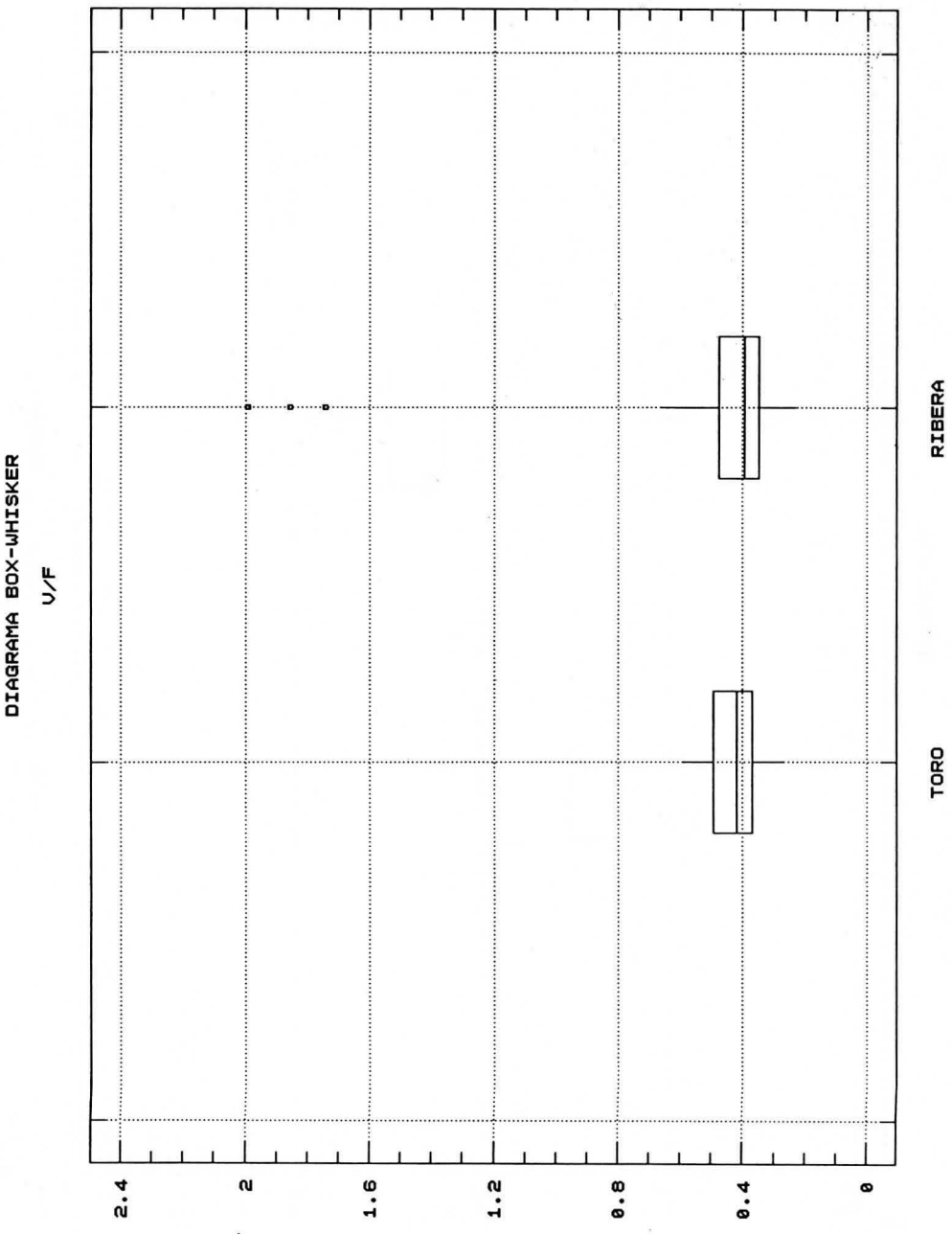


FIGURA 4

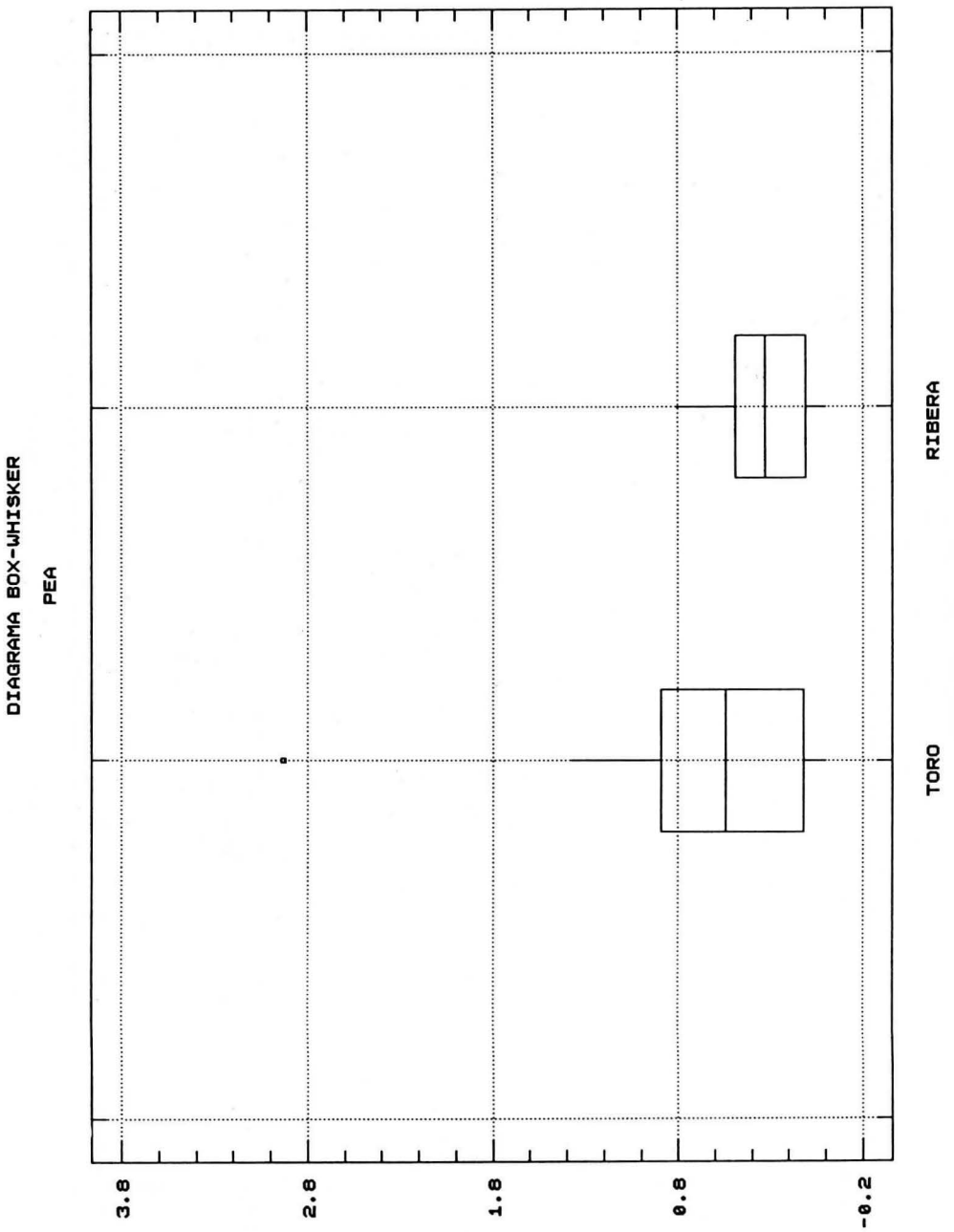
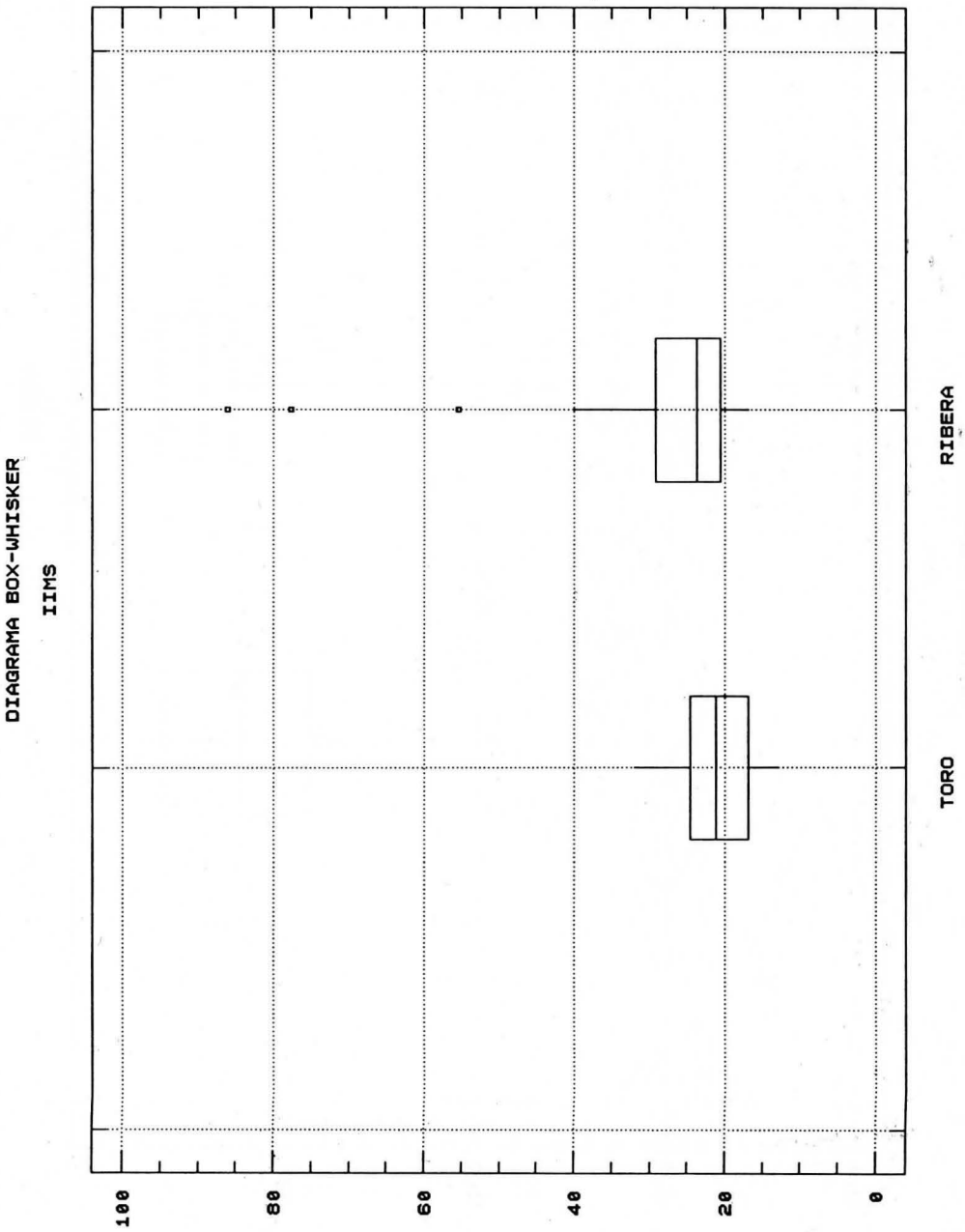


FIGURA 5



es especialmente crítica porque las variables correlacionadas sólo contribuyen a aumentar el rumor de fondo de los datos sin aportar información significativa para la clasificación.

Para analizar todas las posibles correlaciones entre cada par de variables es preciso tener en cuenta 903 coeficientes de correlación en cada una de las clases. Sólo se anotan las parejas de variables con coeficiente de correlación superior en valor absoluto a 0.90.

–En la Categoría 1 (vinos de Toro)
42-43, 28-29, 17-27, 15-17, 17-39, 15-27, 26-27, 37-42, 37-43, 43-39, 17-34, 35-43, 35-42, 35-37.

–En la Categoría 2 (vinos de Ribera)
28-29, 42-43, 17-27, 17-22, 17-23, 23-27, 15-17, 15-39, 27-35, 27-37, 35-37.

–En ambas Categorías conjuntamente
28-29, 42-43, 17-27, 15-17, 26-27, 35-37.

Sin embargo desde un punto de vista global es más significativo determinar las variables que tienen colinealidad alta con las demás, para ello un método eficaz consiste en identificar aquellas variables cuya presencia provoca una disminución significativa del valor del determinante de la matriz de correlaciones.

En la tabla 4, se recogen los resultados del procedimiento. Cuando el determinante aumenta en un factor entre diez y cien al eliminar la variable, se la ha señalado con dos astericos y con tres si el aumento lo es por un factor de cien o más. El valor 10 para este cociente se considera umbral para decidir que la variable presenta colinealidad alta. De nuevo nos encontramos con una notable disimetría entre ambas categorías y con la necesidad de eliminar un gran número de variables si se desea construir un modelo de clasificación que las use directamente.

La alta correlación entre las variables para los vinos de Toro puede explicarse suficientemente por la distribución de las muestras por bodegas, hay que suponer que cada bodega mantiene sus propios métodos de vinificación de modo que a mayor número de bodegas representadas en la muestra ha de esperarse mayor variabilidad y por tanto menos correlación, que es lo que ocurre en la D. O. Ribera frente lo que ocurre en la D. O. de Toro.

Releyendo los datos del epígrafe 1.1 desde este punto de vista se tiene la siguiente clasificación:

| N.º de muestras | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| N.º de bodegas y coop. de Toro | 2 | 3 | – | – | 1 | 1 | – | – |
| N.º de bodegas y coop. de Ribera | 13 | 1 | 2 | 1 | – | 1 | – | 2 |

Evidentemente las relaciones internas entre las variables en los vinos de Ribera de Duero tienen una fuente de variabilidad mayor que los de Toro.

**Tabla 4. Variables correlacionadas con las demás**

| | | CAT. 1 | CAT. 2 | Global |
|---|---|---|---|---|
| 1 | GA | | | |
| 2 | AV | *** | | |
| 3 | AT | | | |
| 4 | pH | | | |
| 5 | AP | *** | ** | ** |
| 6 | BP | *** | *** | ** |
| 7 | C | *** | ** | ** |
| 8 | DP | | ** | |
| 9 | EP | | | |
| 10 | VF | *** | | |
| 11 | VAL | ** | | |
| 12 | SV | *** | | |
| 13 | DF | ** | ** | |
| 14 | CI | *** | | |
| 15 | PT | *** | *** | *** |
| 16 | PE | *** | | |
| 17 | MV | *** | *** | *** |
| 18 | DFA | *** | | |
| 19 | CIA | | | |
| 20 | PTA | *** | | |
| 21 | PEA | | | |
| 22 | MVA | *** | ** | |
| 23 | DFC | | *** | |
| 24 | CIC | *** | | |
| 25 | PTC | *** | | |
| 26 | PEC | *** | ** | ** |
| 27 | MVC | *** | ** | ** |
| 28 | IC | *** | *** | *** |
| 29 | IC2 | ** | ** | |
| 30 | TONO | ** | | |
| 31 | II | | ** | |
| 32 | IIS | | ** | ** |
| 33 | IIMS | ** | | |
| 34 | ACG | *** | | ** |
| 35 | MONB | *** | *** | *** |
| 36 | PRB | *** | | |
| 37 | PPB | *** | ** | ** |
| 38 | PVP | | | |
| 39 | ACTS | *** | ** | |
| 40 | ACIS | | | |
| 41 | EQ1 | *** | ** | ** |
| 42 | EQ2 | *** | *** | *** |
| 43 | INDP | | | |

## 2.4. **Análisis de la estructura interna**

Especialmente adecuado, para el caso de una alta relación entre el número de variables y el de objetos, es la determinación de las componentes principales. Mediante ellas es posible reducir la dimensionalidad del problema sin pérdida de información significativa.

Las componentes principales son combinaciones lineales de las cuarenta y tres variables originales. El cociente $l_{ij}$ es el peso ("loading") de la variable j-ésima sobre la componente i-ésima.

$$CP_i = \sum_{j=1}^{j=43} l_{ij} V_j$$

Las componentes principales tienen dos características interesantes que se siguen del criterio con que se las construye. Recordemos que cada objeto está representado por un vector de cuarenta y tres dimensiones: los cuarenta y tres valores que tiene en las variables medidas. Todos los objetos forman una nube de puntos en el espacio de cuarenta y tres dimensiones que, aun cuando no podamos visualizarla, presenta una estructura propia: es la estructura interna de correlaciones y colinealidades que tienen las variables entre sí.

La primera componente sigue por construcción la dirección de máxima elongación de la nube, es decir la dirección en que más varían los datos químicos de los objetos en estudio. Formalmente se trata de determinar los valores $l_{ij}$ con la condición de que sea máxima la varianza de $CP_1$. La segunda componente se la construye de modo que siga una dirección ortogonal (incorrelada) con la anterior y de máxima varianza, es decir es la variable incorrelada con la primera que explica el mayor porcentaje de varianza no explicada por la primera. Formalmente está caracterizada por Corr ($CP_1$, $CP_2$) = 0 y Var($CP_2$) máxima. De este modo se construyen las sucesivas componentes principales $CP_3$,... $CP_{43}$ incorreladas con las anteriores y de varianza máxima, en número igual a la dimensión del espacio.

Para evitar una nube distorsionada por las diversas escalas de medidas de las variables procederemos a su tipificación: restar la media y dividir por la desviación típica. De este modo la nube de puntos está centrada en el origen de coordenadas y la varianza de cada variable es uno. Este proceso no altera las relaciones internas entre las variables, simplemente garantiza que las elongaciones de la nube de puntos no están causadas por un efecto de escala totalmente irrelevante en nuestro problema. Si las variables están tipificadas cada $l_{ij}$ tiene el significado de coeficiente de correlación, $l_{ij}$ = Corr ($CP_i$, $V_j$). Además los $l_{io}$ son nulos para cada componente.

En la tabla 5 se recogen los trece primeros autovalores de los cuarenta y tres posibles. Mediante ellos se explica el 90% de la varianza que tiene toda la tabla.

**Tabla 5. Componentes principales. Autovalores**

| | V.E. | V.E.(%) | V.A.(%) |
|---|---|---|---|
| $CP_1$ | 15.15 | 35.24 | 35.24 |
| $CP_2$ | 7.51 | 17.47 | 52.71 |
| $CP_3$ | 3.43 | 7.98 | 60.69 |
| $CP_4$ | 2.72 | 6.32 | 67.00 |
| $CP_5$ | 1.92 | 4.46 | 71.47 |
| $CP_6$ | 1.61 | 3.74 | 75.21 |
| $CP_7$ | 1.56 | 3.64 | 78.84 |
| $CP_8$ | 1.16 | 2.71 | 81.55 |
| $CP_9$ | 0.98 | 2.28 | 83.83 |
| $CP_{10}$ | 0.87 | 2.02 | 85.85 |
| $CP_{11}$ | 0.80 | 1.86 | 87.71 |
| $CP_{12}$ | 0.71 | 1.65 | 89.37 |
| $CP_{13}$ | 0.59 | 1.38 | 90.74 |

V.E.: Varianza explicada
V.A.: Varianza acumulada

Para decidir cuantos de estos autovectores encierran información significativa y cuales han de ser atribuidos al rumor de fondo experimental se ha aplicado el procedimiento de doble validación cruzada (double cross-validation). Siguiendo un esquema diagonal de cancelación y cinco grupos de cancelación para la matriz de correlaciones se han obtenido cuatro componentes significativas. Con estas cuatro componentes se explica el 67% de la variabilidad (tabla 5), o lo que es igual el 33% de la variabilidad habrá de atribuirse al rumor de fondo. Lo anterior pone de manifiesto que a pesar del gran número de variables (43), muchas aportan información redundante. Esta idea ya había sido apuntada en el análisis de correlaciones, sin embargo ahora está adecuadamente cuantificada.

## 2.5. **Conclusiones**

- En general, las variables no presentan distribución normal.
- Las variables no son específicas, es decir no es posible con ninguna de ellas alcanzar suficiente diferenciación entre los vinos de Toro y los de Ribera de Duero.
- Existen notables y generalizadas correlaciones entre las variables, lo que exige métodos "flexibles" de modelado (soft-modelling).
- La estructura latente está constituida por cuatro factores validados, que explican el 67% de la varianza de la tabla de datos.

## 3. CLASIFICACIÓN Y MODELADO

La finalidad de este estudio es determinar, en base a la información química, una regla de clasificación y un modelo que permita decidir para cada objeto su pertenencia o no a la clase de vinos de Toro o a los de Ribera.

Las técnicas de clasificación forman parte de la metodología de la Inteligencia Artificial bajo el nombre de Reconocimiento de Pautas (Pattern Recognition, PaRC) y se agrupan en dos niveles. Un método se dice que trabaja al primer nivel de PaRC si clasifica obligatoriamente cada objeto en una de las dos categorías, mientras que lo hace el segundo nivel si además puede clasificar uno o varios objetos fuera de los modelos construidos. Es decir detecta objetos que no pertenecen ni al modelo Toro ni al de Ribera.

Operativamente, en el primer caso se construye una función que divide al espacio de las variables químicas en dos subespacios que se asignan a cada una de las clases. Es el caso del Análisis Discriminante Lineal (ADL), que construye un hiperplano de separación entre las clases. Este método exige a) normalidad en los datos, b) igualdad de las matrices de covarianzas de las categorías, c) un número de objetos mayor que el de variables (se recomienda que al menos sea el doble). En el conjunto de datos en estudio sabemos que las hipótesis a) y b) no se cumplen (epígrafes 2.1. y 2.3).

Los métodos que operan al segundo nivel de PaRC son aquellos que construyen un modelo para cada categoría, o lo que es igual, definen un recinto en el espacio de las variables químicas, de modo que cada objeto es asignado a la categoría de la que dista menos. Esta es la regla de clasificación. Es más, en este caso existe la posibilidad de que un objeto no pertenezca a ninguno de los modelos y esto con independencia de que su asignación a una de las categorías sea correcta o no. Un objeto de este tipo recibe el nombre de anómalo (outlier). También es posible el caso contrario, es decir, un objeto que pertenece a ambos modelos, esto implicaria una falta de "especificidad" por parte del modelo de la categoría a la que pertenece. Por el contrario, la "sensibilidad" del modelo de una categoría es la capacidad de detectar correctamente lo objetos que pertenecen a ella.

Un análisis previo a la tarea de construir un modelo para las categorías en estudio ha de establecer de forma descriptiva las agrupaciones "naturales" entre los objetos. Es un análisis de la estructura de la tabla de los datos complementario al aportado por el análisis de los factores latentes (epígrafe 2.4). Mientras que estos últimos son un análisis de las relaciones mutuas entre las variables, el análisis de agrupaciones describe la proximidad entre las muestras.

### 3.1. Análisis de agrupamiento de los objetos

Para realizar el estudio se establece la similaridad entre cada par de los 66

objetos mediante el cuadrado de la distancia euclídea entre los puntos que los representan en un espacio de 43 dimensiones (los valores de las variables sobre los objetos). De este modo se tiene una matriz cuadrada de dimensión 66.

Los grupos de objetos se han obtenido con el método jerárquico de Ward. En cada etapa del proceso se unen los dos objetos (o grupos) más similares para producir uno sólo, lo distintivo del método de Ward respecto de otros jerárquicos es que se basa en la estructura que se obtendrá después del agrupamiento.

No se entrará en los detalles matemáticos pero sí interesa mostrar la filosofía del procedimiento. Se define la heterogeneidad, $H_g$, de cada grupo, g, existente en una etapa del proceso mediante la suma del cuadrado de la distancia euclídea de cada elemento del grupo al centroide del mismo. Es un índice similar a la varianza dentro del grupo, representa la pérdida de información ocurrida al considerar el centroide como representante del grupo en lugar de los objetos individuales; si en un grupo los objetos son muy similares entre sí su heterogeneidad será pequeña.

Cada etapa consiste en unir dos de los grupos para obtener uno mayor como sigue:

i) Se unen de dos en dos todos los grupos, "p" y "q" de heterogeneidad $H_p$ y $H_q$ respectivamente.

ii) Se evalúa la heterogeneidad de todos los nuevos grupos $H_{(p,q)}$.

iii) Se retiene como nuevo grupo el formado por los dos antiguos tales que $H_{(p,q)} - H_p - H_q$ es mínimo.

En resumen los grupos se amplían sucesivamente de modo que la pérdida de información (el incremento de heterogeneidad) en cada paso debida al crecimiento del grupo sea la mínima posible.

El resultado del proceso se muestra en forma de dendograma en la figura 6. Al nivel de similaridad 1 todos los grupos son singulares, cada objeto sólo es similar a sí mismo, a medida que la similaridad disminuye los objetos se "confunden" en grupos cada vez más amplios, al nivel 0.3 sólo existen dos grupos de objetos similares entre sí. El análisis del dendograma se hace mediante un estudio del significado real de los agrupamientos a medida que decrece la similaridad entre los objetos, sólo cuando los grupos formados son interpretables respecto del problema en estudio tiene sentido el análisis.

En la figura 6 se han marcado con tres colores distintos los tres grupos que se tienen al nivel de similaridad 0.69. La interpretación es la siguiente:

i) En rojo se tienen agrupados 42 objetos de los cuales 37 son de Ribera y 5 de Toro (45, 65, 47, 51 y 52) todos ellos situados en el subgrupo inferior. Por evidente mayoría este es el grupo de los vinos de Ribera de Duero.

ii) En azul se tienen 21 muestras de las que 13 son de Toro y 8 de Ribera (2, 55, 5, 9, 56, 11, 22 y 16). Es el cluster de los vinos de Toro en el sentido de que son mayoría y por tanto quienes definen la característica del grupo.

iii) Finalmente el grupo de color azul oscuro lo forman tres muestras: dos de Ribera (4 y 18) y una de Toro (43).

El aspecto global del dendograma muestra que cada objeto es poco "similar" a otros, sin duda es consecuencia de la relación número de objetos/número de variables en el sentido de que cada objeto fácilmente se diferencia de los demás en el valor de al menos una de las variables. Pero esto puede conducir fácilmente a diferenciaciones falseadas ya que sabemos que el 33% de la variabilidad ha de considerarse como rumor de fondo (ver epígrafe 2.4) y que algunas variables presentan algunos valores muy distintos (véase p.e las figuras 2-5, epígrafe 2.2).

Por ello se ha procedido a hacer un nuevo agrupamiento de los objetos tomando como variables descriptoras de cada uno de ellos sus valores sobre las cuatro primeras componentes principales. El resultado se muestra en la figura 7 y como era previsible los objetos forman grupos más definidos. Los grupos marcados en color se han obtenido al nivel de similaridad 0.5684 y su descripción es:

i) En rojo se han señalado dos grupos formados mayoritariamente por muestras de Ribera de Duero. Globalmente agrupan 46 objetos de los que 6 son de Toro.

El grupo inferior está constituido por 20 muestras todas ellas de Ribera de Duero, más todavía, todas son la vendimia de 1987 a excepción de la 57 que es de 1985; de hecho sólo queda fuera de este cluster una única muestra de 1987: la número 58.

El grupo que aparece en la parte superior del gráfico formado por 26 objetos reúne 20 muestras de Ribera de Duero: de la vendimia de 1985 (1, 4, 3, 58 y 56), de la vendimia de 1986 (12, 6, 21, 17, 19, 22, 7, 8, 9, 13, 15, 10, 11, 20) y la muestra 31 de 1987 junto con 6 muestras de Toro. De estas seis muestras cinco (47, 51, 45, 60 y 61, de diversas vendimias) son todas las que aporta al estudio la Bodega Luis Mateos de Toro indicando quizá procesos de vinificación específicos de esta bodega que asimila el vino producido al de la Ribera de Duero.

Es claro que estos dos grupos marcados en rojo muestran la especificidad de los vinos de Ribera de Duero, en especial la vendimia de 1987.

ii) En azul se han marcado el tercer grupo formado por 20 muestras de las que siete (2, 16, 5, 59, 14, 18 y 55) son de Ribera mayoritariamente de la vendimia de 1986.

Estas nítidas agrupaciones aportan viabilidad al intento de construir un modelo que discrimine ambas categorías basado en la estructura latente. Esta idea ya se sugirió como consecuencia del análisis factorial, pero allí se consideraba únicamente la información aportada por las variables físico-químicas y la redundancia provocada por sus relaciones mutuas. Después del análisis cluster se tienen evidencia de que esa información depurada en forma de estructura factorial es adecuada para el propósito de discriminar y modelar ambos tipos de orígenes.

FIGURA 6

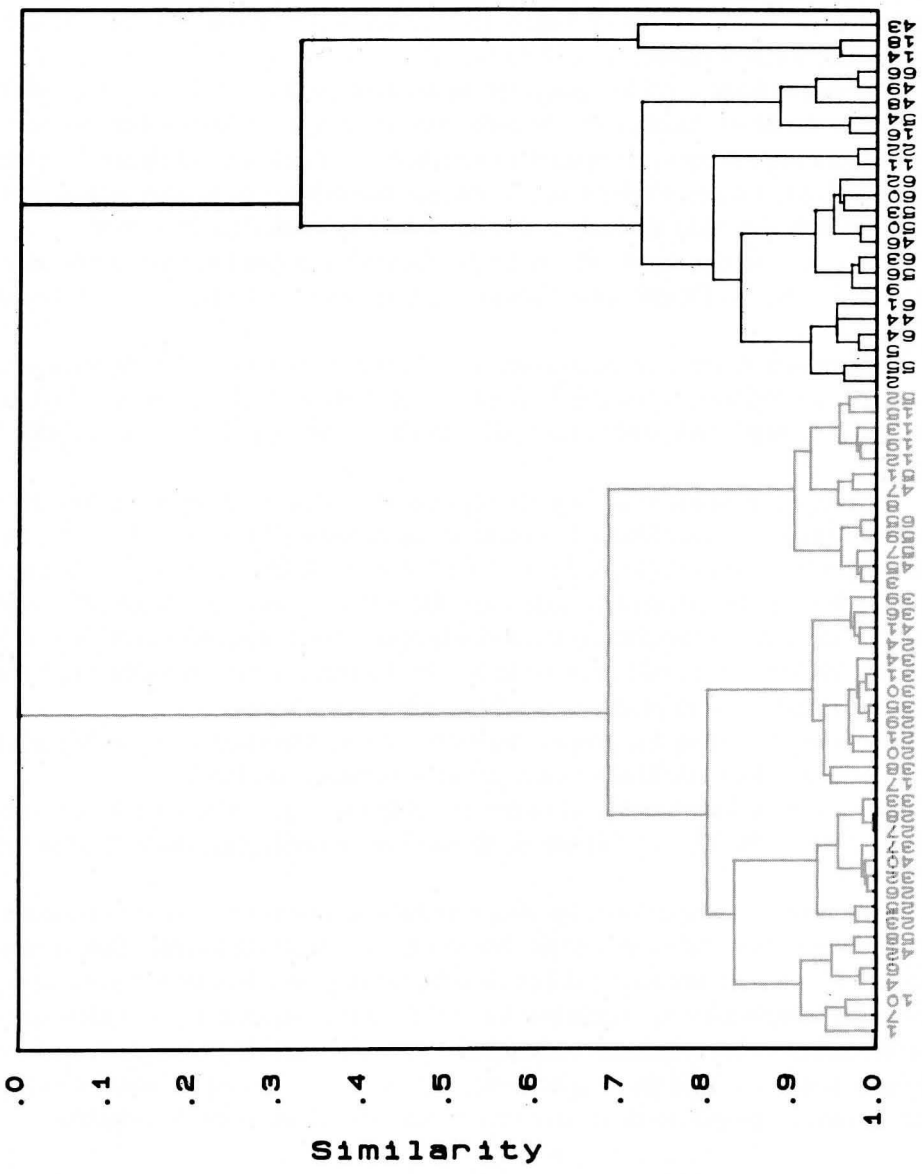
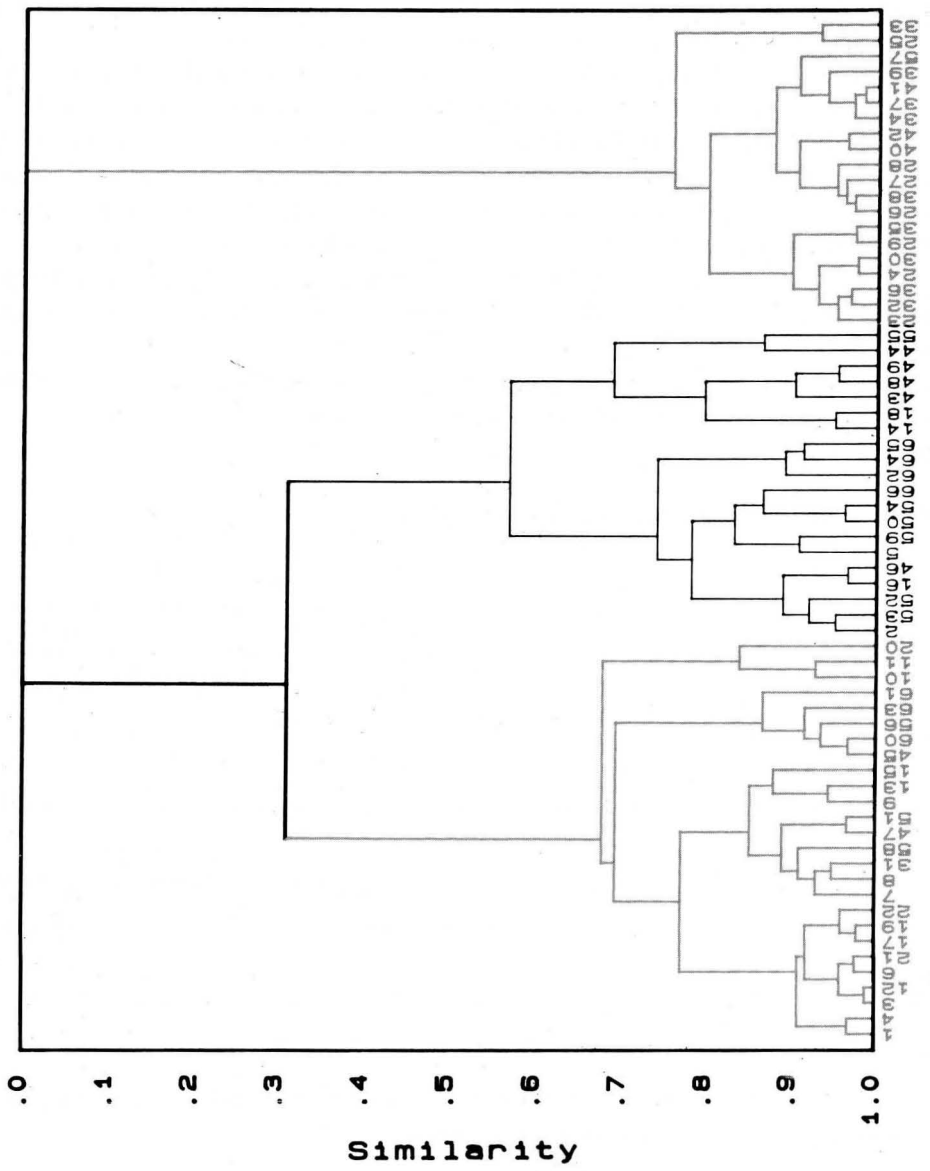


FIGURA 7



### 3.2. **Clasificación. Método del entorno próximo (KNN)**

Puesto que se han de usar métodos robustos frente al fallo de normalidad e independientes de que la estructura latente sea la misma o no en ambas categorías se ha elegido el procedimiento del entorno próximo (KNN). Este método que recibe su nombre de la regla de clasificación (K Nearest Neighbour), es no paramétrico. Más todavía no hace ninguna hipótesis sobre la distribución de las variables con las que se pretende hacer la clasificación. Sin embargo la frontera de las clases no quedará definida explícitamente mediante una fórmula.

En nuestro problema los tamaños muestrales son muy distintos en cada categoría, además después de los análisis cluster es adecuada la distancia euclídea para establecer la proximidad entre objetos. Ambas razones sugieren que para decidir la clase a la que se asigna una muestra no debe usarse el criterio habitual de atribuirla a la que pertenecen la mayoría de los K objetos más próximos. Parece más razonable usar una ponderación por el inverso de la distancia.

La clasificación se ha llevado a cabo con K = 4, 5, 6, 7 y 8 puntos próximos. Se han estandarizado las variables para evitar la influencia de las distintas escalas de medida.

Los resultados se muestran en la tabla 6, la mejor clasificación se tiene con K = 6. Un porcentaje de clasificaciones correctas en torno al 90% en cada categoría es aceptable.

**Tabla 6. Matriz de clasificación, método KNN**

| | Categoría asignada por KNN | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | K = 4 | | K = 5 | | K = 6 | | K = 7 | | K = 8 | |
| | T | R | T | R | T | R | T | R | T | R |
| Toro | 17 | 2 | 15 | 4 | 17 | 2 | 14 | 5 | 13 | 6 |
| Ribera | 4 | 43 | 4 | 43 | 3 | 44 | 2 | 45 | 3 | 44 |

| | Porcentaje de clasificaciones correctas | | | | |
|---|---|---|---|---|---|
| | K = 4 | K = 5 | K = 6 | K = 7 | K = 8 |
| Toro | 89.47 | 78.95 | 89.47 | 73.68 | 68.42 |
| Ribera | 91.49 | 91.49 | 93.62 | 95.74 | 93.62 |
| Global | 90.91 | 87.88 | 92.42 | 89.39 | 86.36 |

A efectos de comparación con los demás análisis, en la tabla 7 se recogen los objetos mal clasificados para cada valor de K.

**Tabla 7. Muestras mal clasificadas con KNN**

| N.º | Cat. real | Cat. asignada | K |
|---|---|---|---|
| 16 | Ribera | Toro | 4, 5, 6, 7, 8 |
| 18 | Ribera | Toro | 4, 5 |
| 45 | Toro | Ribera | 5 |
| 47 | Toro | Ribera | 4, 8 |
| 51 | Toro | Ribera | 5, 6, 7, 8 |
| 52 | Toro | Ribera | 4, 5, 6, 7, 8 |
| 55 | Ribera | Toro | 4, 5, 6, 7, 8 |
| 56 | Ribera | Toro | 4, 5, 5, 8 |
| 61 | Toro | Ribera | 7, 8 |
| 64 | Toro | Ribera | 7, 8 |
| 64 | Toro | Ribera | 7, 8 |

Por su propia construcción el método KNN carece de cualquier tipo de evaluación para la predicción.

### 3.3. **Modelado SIMCA**

El método SIMCA (Soft Independent Modeling of Class Analogy) construye un modelo para cada clase separadamente. Es una "caja" de una o varias dimensiones construída no con las variables químicas originales, sino con los componentes principales de estas variables en cada categoría. De esta forma se supera el problema de la relación número de objetos / número de variables. De hecho SIMCA puede usarse con menos objetos que variables. Como las componentes principales son ortogonales, deja de ser limitante la correlación entre las variables originales y tampoco se hacen hipótesis sobre la distribución probabilística de los datos. El método SIMCA es adecuado al problema que nos ocupa.

En SIMCA los factores latentes en la estructura química definida por las variables químicas se agrupan en dos tipos: a) los factores internos, b) los factores externos. El número de factores internos o componentes determina la dimensión de la "caja" que es el modelo (dimensión uno, si es un segmento, dos si es un rectángulo, tres si es un paralelepipedo, etc.) y estos factores internos recogen las variaciones que muestran los datos en función de la estructura: medias, varianzas y correlaciones. En consecuencia analizando los factores internos, que son combinaciones lineales de las variables originales, es posible interpretarlos e identificar el significado que tienen. Los factores externos (componentes principales minoritarias) reúnen los errores aleatorios, las variaciones no relacionadas con la pertenencia a una clase u otra y el efecto de factores que sólo conciernen a una minoría de vinos.

Es de gran interés determinar la contribución de cada variable en la construc-

ción de factores internos, una medida de ello es el "poder modelante" de las variables en cada una de las categorías. Así pues, las características específicas de los vinos de Toro y de Ribera se expresarán en base a las variables con mayor poder modelante en cada una de las categorías, es decir, en términos químicos y por ende con significado enológico.

Tan importante como el poder modelante, es la "potencia discriminante", es decir, la capacidad que tiene una variable para decidir si un objeto pertenece a una u otra categoría. Es a estas variables las que debemos tener en cuenta cuando queramos poner en evidencia las diferencias entre un vino de Toro y otro de Ribera.

Al tener en cuenta toda esta información respecto las variables químicas medidas y su contribución al modelado, podremos tomar decisiones en el sentido de definir qué vinificaciones, tratamientos, etc. han de ser tenidos en cuenta para aumentar las características propias de la Denominación de Origen. Incluso en el caso de disponer de datos sensoriales de estas muestras de vinos, se podrían relacionar con las variables químicas en orden a incrementar su calidad.

En lo que sigue siempre se ha usado la distancia ampliada de Wold y se ha tomado el 95% como nivel de confianza para construir el recinto.

### 3.3.1. **Modelado con los parámetros enológicos convencionales**

Los datos correspondientes forman una tabla de 64 objetos y cuatro variables: GA, AV, AT y pH. Faltan seis datos, distribuídos en cuatro objetos distintos y cuatro variables, como el porcentaje tanto en términos de variables como de objetos es muy pequeño se les ha reconstruído mediante un análisis factorial que retiene el 95% de la varianza. No se ha observado ninguna influencia específica de las variables y objetos con valores reconstruidos.

Se ha construído un modelo SIMCA con dos componentes por clase. Los resultados numéricos básicos se recogen en la tabla 8. Sólo es destacable la potencia discriminante prácticamente igual para todas las variables. El Grado Alcohólico (es la variable con mayor peso de Fisher de las 43, tabla 3) muestra gran incapacidad para modelar las clases.

A pesar de que sólo se manejan cuatro variables el modelo recoge con dos componentes el 78.7% y el 75.2% de la variabilidad mostrada en cada categoría. Dicho de otro modo la cuarta parte de la variabilidad ha de ser atribuída a los factores externos.

La figura 8 muestra el Diagrama de Coomans relativo a este modelo SIMCA. Cada objeto está representado por una botella esquemática de color azul para los vinos de Toro y rojo para los de Ribera de Duero.

En el eje OX se representa la distancia de cada objeto al modelo de la Categoría 1 (vinos de Toro), la distancia típica que define al modelo de los vinos de Toro

**Tabla 8. SIMCA con dos componentes principales para cada categoría**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 1 GA | 0.1980 | 0.2730 | 2.514 |
| 2 AV | 0.6060 | 0.5854 | 2.413 |
| 3 AT | 0.6428 | 0.5084 | 2.561 |
| 4 pH | 0.8215 | 0.6905 | 2.319 |
| SENSIBILIDAD | 94.7% | 93.6% | |
| ESPECIFICIDAD | 63.8% | 73.7% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 17 | 2 | 89.5% |
| Ribera | 3 | 44 | 93.6% |
| | | Capacidad global de clasificación: | 92.4% |

está representada por la línea vertical azul, de modo que el "recinto vertical" de bordes verdes es una representación plana de la "caja" SIMCA para los vinos de Toro. Los objetos que estén en este recinto están dentro del modelo de la categoría Toro con independencia de su verdadero origen, es la categoría computada para ese objeto por el modelo SIMCA.

En el eje OY se representa la distancia de cada objeto al modelo de categoría 2 (vinos de Ribera de Duero), la distancia típica que define a estos vinos está marcada por la línea horizontal roja, ahora el "recinto horizontal" de bordes rojos es una representación planta del modelo SIMCA construído para los vinos de Ribera.

Ambos modelos se intersecan formando el cuadrado inferior izquierdo, los objetos que se encuentran en este recinto pertenecen a ambos modelos, si bien a efectos de clasificación se les asignará a la clase de la que menos disten. Un gran número de muestras en esta zona indica un modelo poco específico.

Finalmente los objetos anómalos se sitúan fuera de ambos recintos.

Con estas consideraciones respecto de la interpretación de la figura 6 queda clara la poca especificidad de los modelos particularmente la del modelo de Toro, que acepta muchas muestras de Ribera, a pesar de que en su mayoría están correctamente asignados porque distan menos del modelo de Ribera que del de Toro. Esta observación indica el valor relativo de la clasificación sin tener en cuenta otros aspectos del modelo. En resumen estas variables enológicas conven-

FIGURA 8

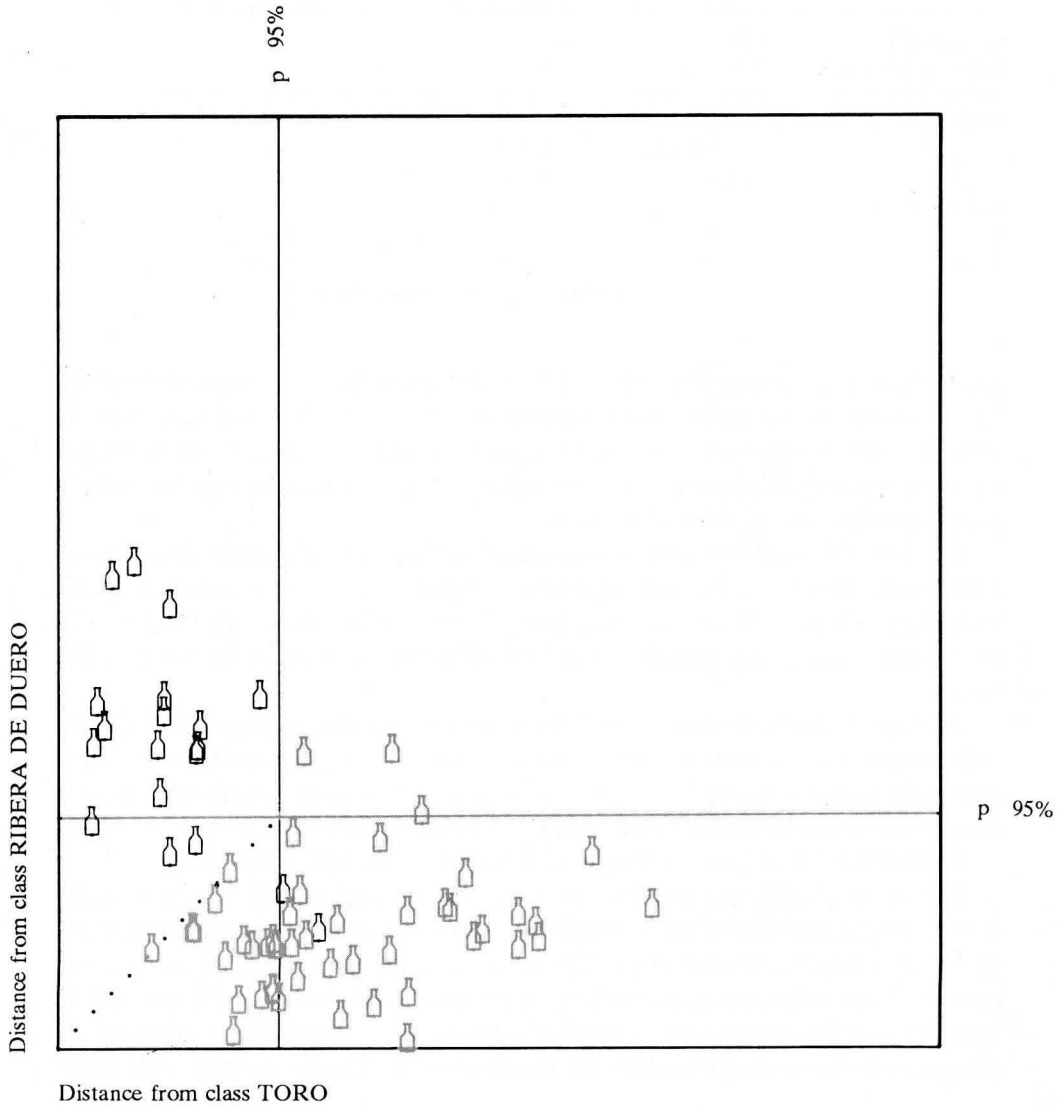
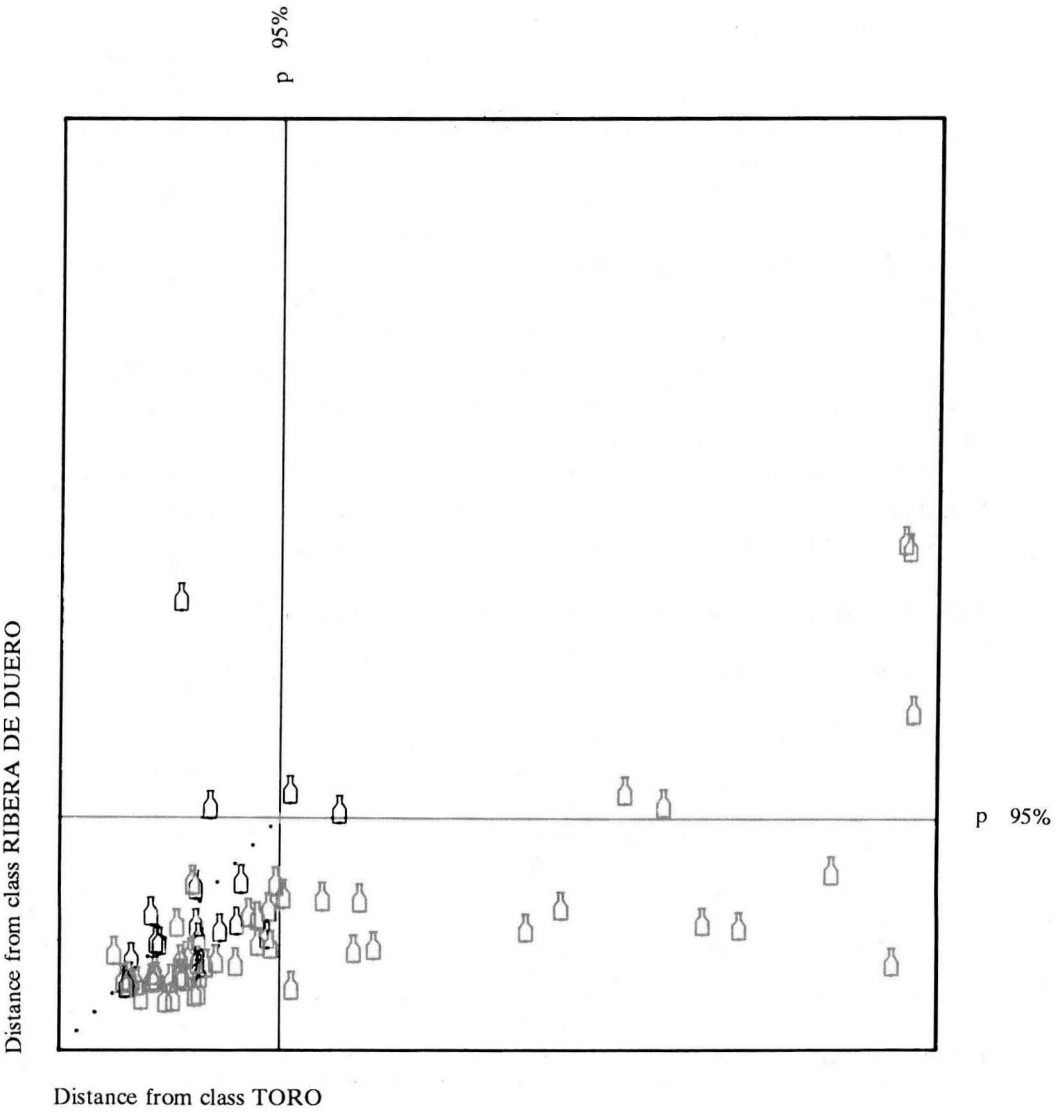

p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

FIGURA 9


p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

cionales permiten una clasificación aceptable fundamentada en el Grado Alcohólico y la Acidez Total, pero no un modelado suficiente.,

### 3.3.2. **Modelado con los compuestos polifenólicos**

En este caso la matriz a estudiar consta de 66 filas, las muestras y 8 columnas, las variables codificadas por AP, BP, C, DP, EP, V/F, V/LA y S/V.

La ausencia de cinco datos (distribuídos entre cuatro objetos distintos y tres variables) ha sido remediada reconstruyendo los valores mediante un análisis factorial con un 95% de la varianza explicada.

Lo más notable es la elevada capacidad de discriminación que muestra la relación entre catequinas y polifenoles totales (V/F) seguida de las de las Catequinas. Precisamente estas dos variables son las que muestran el mayor poder modelante para los vinos de Ribera y los de Toro respectivamente. A excepción de V/L no existen grandes diferencias entre el comportamiento de las variables para modelar ambas clases. De forma global se observa que los índices discriminan, mientras que la capacidad modelante está más ligada al contenido de proantocianidoles en los vinos de Toro y a los polifenoles en los de Ribera. Las catequinas participan tanto para modelar como para discriminar.

En todo caso es notable la escasa especificidad que se puede alcanzar con estas variables en las dos categorías. Incluso la capacidad de clasificar es baja, en especial para los vinos de Toro.

El Diagrama de Coomans, figura 9, de este modelo SIMCA pone en evidencia la inutilidad de estas variables para modelar las categorías. La mayoría de las muestras se encuentran en la zona común de ambos modelos.

**Tabla 9. SIMCA con tres componentes principales para cada categoría**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 5 AP | 0.7127 | **0.7853** | 1.216 |
| 6 BP | 0.6002 | 0.7413 | 1.703 |
| 7 C | **0.8785** | **0.8451** | **4.499** |
| 8 DP | **0.7506** | 0.6783 | 1.756 |
| 9 EP | 0.7223 | 0.6281 | 2.030 |
| 10 V/F | 0.6568 | **0.9061** | **8.265** |
| 11 V/LA | 0.4309 | 0.5881 | 2.655 |
| 12 SV | 0.5929 | 0.6608 | 2.805 |
| SENSIBILIDAD | 89.5% | 91.5% | |
| ESPECIFICIDAD | 30.0% | 26.3% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 9 | 10 | 47.4% |
| Ribera | 7 | 40 | 85.1% |
| | | Capacidad global de clasificación: | 74.2% |

### 3.3.3. **Modelado con la estructura antociánica**

Sin duda es el grupo de variables más homogéneo desde un punto de vista analítico, por ello se describirán con detalle los modelos SIMCA con tres y cuatro componentes.

El modelo con cuatro componentes para cada clase recoge un 89.7% de la varianza que presentan las quince variables en la categoría 1 y el 83.9% en la categoría 2.

De forma global en la tabla 10 se aprecia que los acetatos son las variables con mayor capacidad modelante (moderada) de los vinos de Toro frente a los de Ribera. El caso contrario es el de los monoglucósidos que muestran mayor capacidad modelante en los de Ribera.

Los cumaratos muestran una capacidad modelante similar pero como grupo son quienes aportan la mayor capacidad de discriminación.

De forma individualizada llaman la atención el 3-monoglucósido de malvidol (MV) con alto poder discriminante y capacidad modelante en ambas categorías y el poder discriminante del acetato del 3-monoglucósido de peonidol (PEA).

Sin embargo la capacidad de clasificación es relativamente baja para la clase de Ribera debido a la poca especificidad del modelo construído para los vinos de Toro. Con la finalidad de tener una idea del comportamiento del modelo siempre se realiza un estudio sistemático en cuanto al número de factores internos (componentes). Dado que el modelo con cuatro factores internos incluye mucha varianza existe la posibilidad de que se haya incluido rumor en ellos, posible causa de la mala clasificación.

Al considerar una componente menos en cada categoría se recoge el 83.1% y el 77.6% de la varianza. En la tabla 11 se muestran los resultados numéricos básicos, de los que se deduce que la capacidad modelante y discriminante de las variables es la misma que en el modelo con cuatro componentes (tabla 10) con la notable excepción del acetato del 3-monoglucósido de peonidol (PEA, variable 21). Se puede concluir que la cuarta componente del modelo previo consiste en la inclusión de PEA para la categoría de Toro. En la categoría de Ribera el efecto lo comparte con el cumarato del 3-monoglucósido de cianidol (CIC, variable 24) cuyo poder modelante se incrementa notablemente al pasar de tres a cuatro componentes.

**Tabla 10. SIMCA con cuatro componentes principales para cada categoría**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 13 DF | 0.5086 | 0.6721 | 1.749 |
| 14 CI | 0.3829 | 0.7042 | 1.468 |
| 15 PT | **0.8060** | **0.7471** | 1.532 |
| 16 PE | 0.5937 | 0.6202 | **2.383** |
| 17 MV | **0.7866** | **0.8230** | **2.325** |
| 18 DFA | 0.6967 | 0.3041 | 1.648 |
| 19 CIA | 0.6736 | 0.3456 | 1.532 |
| 20 PTA | 0.6626 | 0.5982 | 1.342 |
| 21 PEA | **0.8033** | 0.5081 | **4.410** |
| 22 MVA | 0.5601 | 0.6078 | 1.742 |
| 23 DFC | 0.7029 | 0.7019 | **2.694** |
| 24 CIC | 0.6347 | 0.4545 | **2.251** |
| 25 PTC | 0.5645 | 0.5407 | **2.467** |
| 26 PEC | 0.6311 | 0.7049 | 1.981 |
| 27 MVC | **0.8177** | **0.7943** | **2.191** |
| SENSIBILIDAD | 89.5% | 87.2% | |
| ESPECIFICIDAD | 55.3% | 73.7% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 18 | 1 | 94.7% |
| Ribera | 10 | 37 | 78.7% |
| | Capacidad global de clasificación: | | 83.3% |

En lo que respecta al porcentaje de clasificaciones correctas se observa que disminuye en cuatro objetos (de 55 con cuatro componentes a 51 con tres). Su análisis indica que las muestras 3, 8, 24, 28 y 32 están correctamente clasificadas con cuatro componentes e incorrectamente con tres; de ellas las 3, 24 y 32 pertenecen también a la clase de los vinos de Toro, pero su distancia SIMCA es menor a la de Ribera. El objeto 20 está incorrectamente clasificado por el modelo con cuatro componentes y correctamente con el de tres aunque en este caso también pertenece a la categoría de Toro. Los valores de estos objetos no presentan ninguna característica especial en las variables PEA y CIC.

En resumen ambos modelos no difieren tanto como podría pensarse con una lectura apresurada de los porcentajes de clasificación.

El Diagrama Coomans del modelo con tres componentes se muestra en la figura 10. La poca especificidad del modelo para los vinos de Toro es evidente:

**Tabla 11. SIMCA con tres componentes principales para cada categoría**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 13 DF | 0.5252 | 0.6724 | 1.858 |
| 14 CI | 0.3917 | 0.6667 | 1.609 |
| 15 PT | **0.8120** | **0.7415** | 1.681 |
| 16 PE | 0.5818 | 0.4562 | **2.069** |
| 17 MV | **0.7393** | **0.8251** | **2.114** |
| 18 DFA | 0.6788 | 0.3106 | 1.726 |
| 19 CIA | 0.5880 | 0.3492 | 1.496 |
| 20 PTA | 0.5126 | 0.5893 | 1.083 |
| 21 PEA | 0.1048 | 0.3341 | **2.731** |
| 22 MVA | 0.4894 | 0.5963 | 1.418 |
| 23 DFC | 0.6490 | 0.6859 | **2.420** |
| 24 CIC | 0.6470 | 0.1234 | 0.943 |
| 25 PTC | 0.5786 | 0.4563 | **2.559** |
| 26 PEC | 0.6371 | 0.6947 | 1.796 |
| 27 MVC | **0.8215** | **0.7737** | **2.357** |
| SENSIBILIDAD | 89.5% | 93.6% | |
| ESPECIFICIDAD | 38.3% | 73.7% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 18 | 1 | 94.7% |
| Ribera | 14 | 33 | 70.2% |
| | Capacidad global de clasificación: | | 77.3% |

un gran número de muestras de vino de Ribera son admitidas en el recinto de los vinos de Toro. Sólo 11 muestras de vinos de Toro y 16 de las de Ribera quedan incluídas en sus respectivas clases de modo específico. Es cierto que se mejora algo la especificidad respecto del modelo con los compuestos polifenólicos, Figura 9, en ambas categorías sin que se alcance un modelo verdaderamente específico para ninguna de las clases. Si se observa comparativamente este diagrama con el de la figura (epígrafe 3.3.1) basado en los parámetros enológicos convencionales se advierte la posibilidad de añadir información selectiva con otras variables químicas que mejoren el modelo.

### 3.3.4. **Modelado con los parámetros ligados al color, antocianos totales e índices de polimerización.**

Este grupo de variables determina características relacionadas con el color y

FIGURA 10

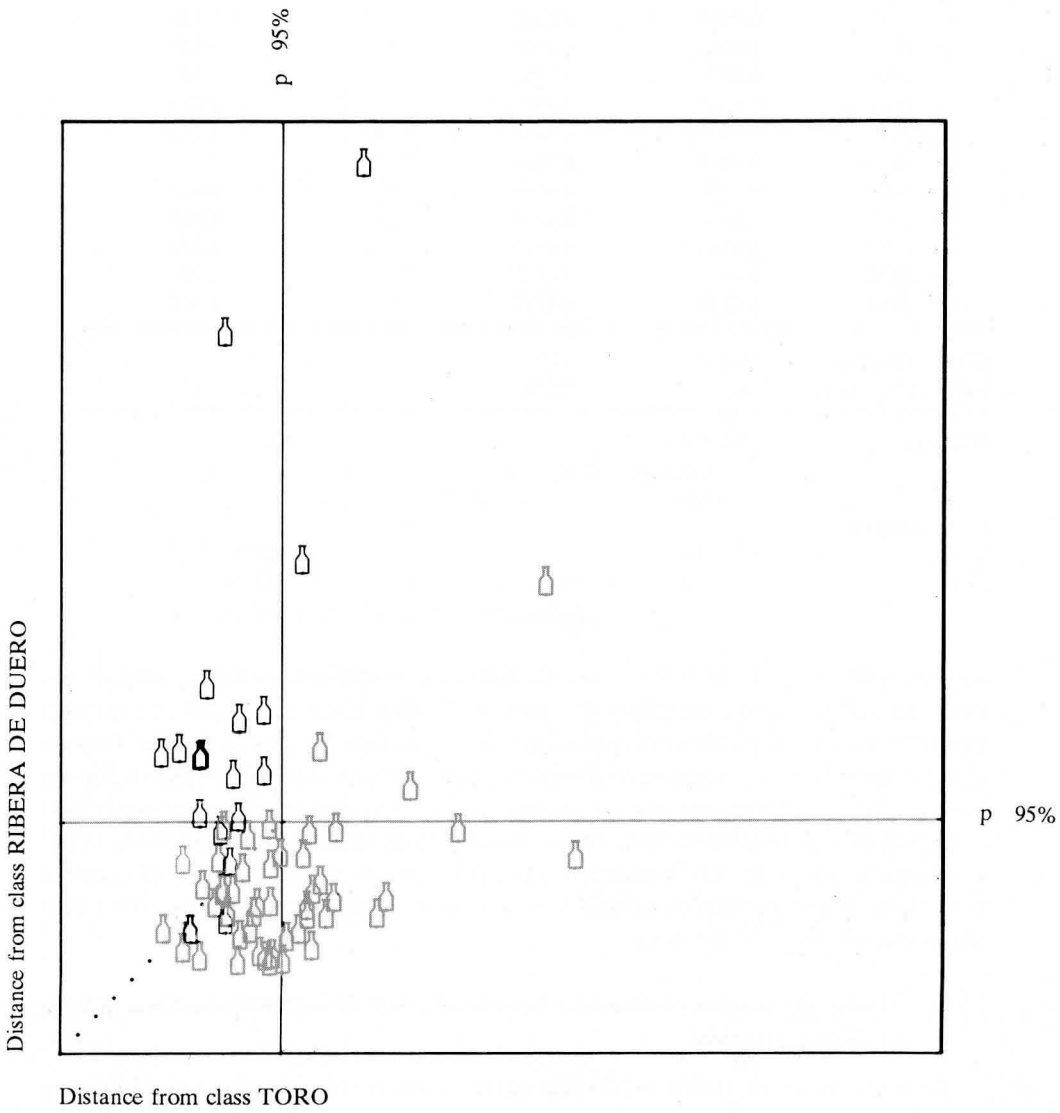

p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

**Tabla 12. SIMCA con tres componentes principales**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 28 IC | **0.9049** | **0.8389** | **2.719** |
| 29 IC2 | **0.8752** | **0.8245** | **2.241** |
| 30 TONO | 0.4644 | 0.4597 | 1.617 |
| 31 II | 0.2226 | 0.5068 | 1.308 |
| 32 IIS | 0.6622 | 0.6129 | **2.570** |
| 33 IIMS | 0.1192 | 0.5204 | **2.748** |
| 34 ACG | **0.7900** | 0.5335 | 1.530 |
| 35 MONB | **0.7726** | 0.7307 | 1.675 |
| 36 PRB | 0.3259 | 0.5079 | 1.187 |
| 37 PPB | 0.7111 | 0.6209 | 1.487 |
| 38 PVP | 0.3496 | 0.2217 | 1.659 |
| 39 ACTS | 0.7015 | 0.7109 | 1.845 |
| 40 ACIS | 0.6741 | 0.5324 | 1.276 |
| 41 EQ1 | 0.4894 | 0.5949 | 1.431 |
| 42 EQ2 | **0.8018** | 0.6678 | 1.551 |
| 43 INDP | **0.8005** | 0.6681 | 1.558 |
| SENSIBILIDAD | 89.1% | 93.6% | |
| ESPECIFICIDAD | 10.5% | 73.7% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 8 | 11 | 42.1% |
| Ribera | 0 | 47 | 100.0% |
| | Capacidad global de clasificación: | | 83.3% |

la proporción de compuestos responsables del mismo. Teóricamente proporcionan un tipo de información en parte complementaria a la aportada por la estructura antociánica en el sentido de que esta última describe las relaciones mutuas entre las sustancias responsables del color, mientras que por ejemplo los índices de ionización determinan la proporción de antocianos de forma coloreada, o las diversas contribuciones según el tipo de polímero y el grado de polimerización. En otras palabras un grupo de variables está relacionado con las proporciones internas de las sustancias responsables del color y el otro con las proporciones respecto de las que no tienen efecto sobre el color.

El resultado de aplicar SIMCA con tres componentes se muestra en la tabla 12. Es destacable el papel que juegan los índices colorantes IC e IC2: modelan ambas clases y sirven para diferenciarlas. En cuanto a la capacidad modelante se

puede resaltar el aspecto diferencial de los antocianos totales (ACG) y la proporción de polímeros rojos (PRB) los primeros contribuyen más al modelo de los vinos de Toro y los segundos a los de Ribera; por contra el porcentaje de los monómeros rojos (MONB) y el de polímetros pardos (PPB) contribuyen al modelado de las dos clases de manera similar. Sin que sea muy acusada también es diversa la capacidad de modelado de EQ2 e INDP. Finalmente señalar que los índices de ionización según Somers (IIS e IIMS) cumplen específicamente un papel discriminante.

Lo llamativo es la nula especificidad para los vinos de Toro del modelo construído, sin embargo de algún modo esta inespecificidad es complementaria a la exhibida por la estructura antociánica ya que en aquel caso (epígrafe 3.3.3, Tabla 11) el porcentaje de clasificación correcto era 94.7% en Toro y 70.2% en Ribera, ahora se tiene la situación inversa 42.1% en Toro y 100% en Ribera.

El Diagrama de Coomans para este modelo constituye la figura 11. Claramente es mejor modelo para las vinos de Ribera pero mucho peor para los de Toro todas cuyas muestras quedan en la zona común de ambos modelos.

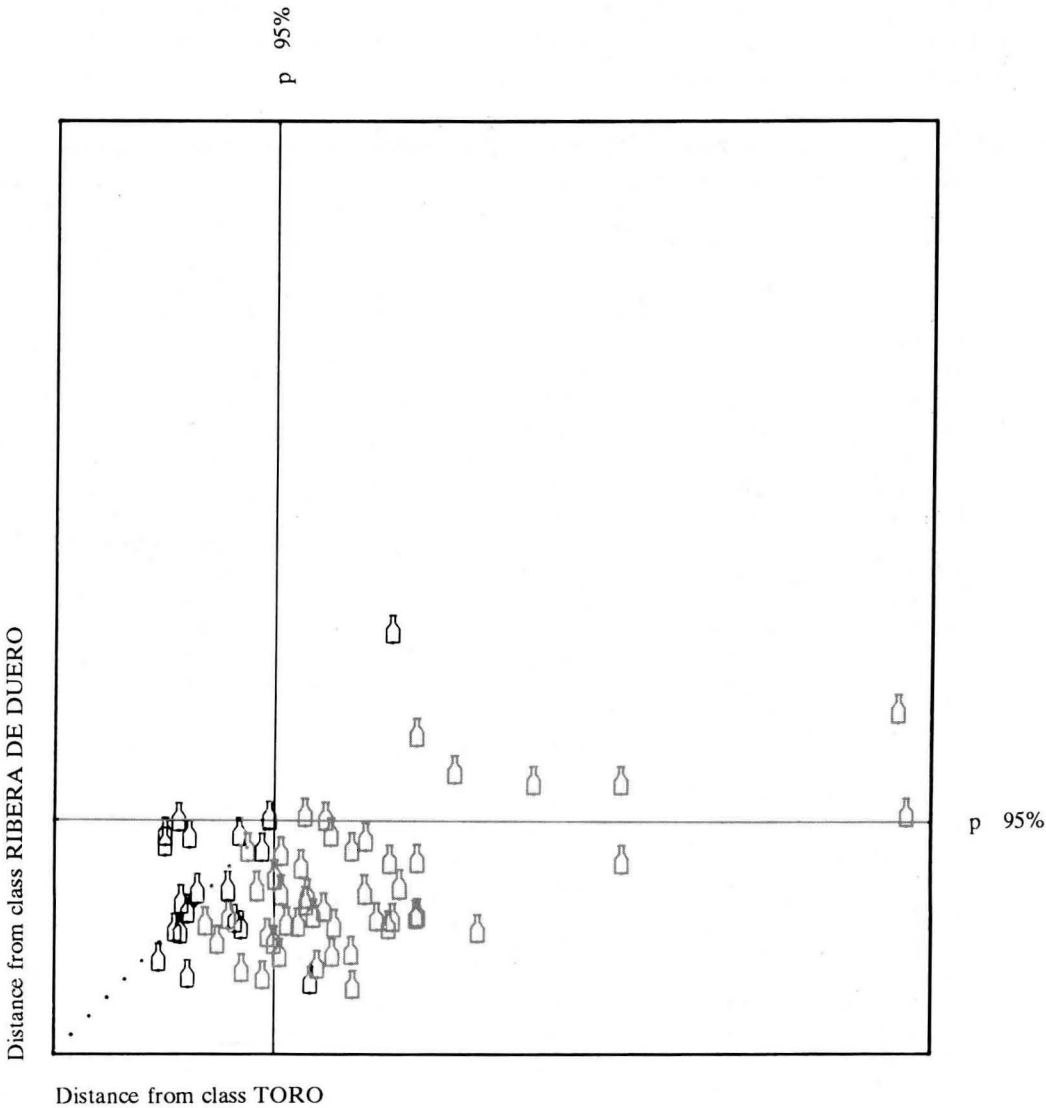
### 3.3.5. **Modelado con la estructura antociánica, parámetros de color antocianos totales e índices de polimerización**

Los dos epígrafes anteriores evidencian que los modelos construídos con la estructura antociánica por un lado y los parámetros ligados al color, índices de ionización y polimerización por otro, son poco específicos pero complementario de algún modo. Es obligado realizar un análisis con las 31 variables conjuntamente porque a priori no es predecible si se mantendrán las características observadas cuando las variables estén en un contexto más amplio.

Inicialmente se ha construído un modelo que recoja al menos el ochenta por ciento de la varianza en cada una de las clases, dado el elevado número de variables esta varianza puede considerarse el límite superior admisible (recuérdese que la varianza cros-validada de toda la tabla de datos es del 67% si bien ahora hay 12 variables menos y por tanto menos rumor). El resultado es que hay que considerar cuatro componentes para los vinos de Toro (81.1% de la varianza) y cinco para los de Ribera (81.6% de la varianza). La Tabla 13 muestra el resumen numérico.

Los antocianos acilados o no (variables 13 a 27) muestran una capacidad modelante similar a la que exhibían cuando intervenían aisladamente en el modelo (Tabla 11, epígrafe 3.3.3). Para los vinos de Toro sólo es reseñable el descenso de la capacidad modelante del acetato de definidol (DFA) a pesar de que el modelo consta de cuatro componentes no tiene influencia como ya habíamos razonado previamente. Lo mismo ocurre con el cumarato de cianidol (CIC) que, a pesar de las cinco componentes del modelo para Ribera, no alcanza el poder

FIGURA 11


p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

modelante de la tabla 10 en que sólo se consideraron cuatro componentes. Por los demás no existen otros cambios apreciables en la capacidad modelante. En lo que respecta a la capacidad discriminante cabe señalar el incremento en el definidol (DF) y la disminución del malvidol (MV) y su cumarato (MVC).

Al comparar los resultados de la tabla 13 con los de la Tabla 9, epígrafe 3.3.4, lo más destacable es la pérdida de influencia de los índices de color (IC, IC2) y los antocianos totales determinados por el método de Glories, sobre todo en capacidad modelante para los vinos de Toro. Por el contrario mantienen su significatividad la edad química (EQ2) y el índice de polimerización (INDP). Es llamativo que todas las determinaciones hechas por el método de Somers (IIS, IIMS, ACTS, ACIS) ha aumentado su capacidad de discriminación.

**Tabla 13. SIMCA con cuatro componentes principales para la categoría 1 y cinco componentes principales para la categoría 2.**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 13 DF | 0.4591 | 0.5899 | **2.056** |
| 14 CI | 0.3454 | 0.5666 | 1.470 |
| 15 PT | **0.8146** | 0.7108 | 1.617 |
| 16 PE | 0.6064 | 0.5148 | **2.033** |
| 17 MV | 0.7231 | **0.8073** | 1.534 |
| 18 DFA | 0.3306 | 0.3295 | 1.425 |
| 19 CIA | 0.4895 | 0.3069 | 1.531 |
| 20 PTA | 0.6218 | 0.5611 | 1.385 |
| 21 PEA | 0.2318 | 0.3514 | **3.056** |
| 22 MVA | 0.4625 | 0.6245 | 1.681 |
| 23 DFC | 0.5089 | 0.7032 | **2.653** |
| 24 CIC | 0.6322 | 0.3151 | 1.777 |
| 25 PTC | 0.4858 | 0.4451 | **2.401** |
| 26 PEC | 0.5571 | 0.6193 | 1.474 |
| 27 MVC | 0.7003 | **0.7482** | 1.533 |
| 28 IC | 0.5932 | 0.7131 | **2.067** |
| 29 IC2 | 0.5744 | 0.6938 | 1.898 |
| 30 TONO | 0.4098 | 0.3558 | 1.565 |
| 31 II | 0.1669 | 0.5860 | 1.378 |
| 32 IIS | 0.6654 | 0.5979 | **3.112** |
| 33 IIMS | 0.2136 | 0.5855 | **2.919** |
| 34 ACG | 0.5700 | 0.6197 | 1.684 |
| 35 MONB | **0.7489** | **0.7439** | 1.775 |
| 36 PRB | 0.3982 | 0.5467 | 1.507 |
| 37 PPB | **0.7722** | 0.6801 | 1.582 |
| 38 PVP | 0.3272 | 0.3608 | 1.610 |

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 39 ACTS | 0.5564 | 0.7105 | **2.208** |
| 40 ACIS | 0.5758 | 0.4250 | 1.423 |
| 41 EQ1 | 0.5588 | 0.5799 | 1.539 |
| 42 EQ2 | **0.8158** | 0.5964 | 1.462 |
| 43 INDP | **0.8139** | 0.6012 | 1.497 |
| SENSIBILIDAD | 89.5% | 89.4% | |
| ESPECIFICIDAD | 91.5% | 89.3% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 17 | 2 | 89.5% |
| Ribera | 0 | 47 | 100.0% |
| | Capacidad global de clasificación: | | 97.0% |

Lo más interesante es que el modelo tiene características homogéneas en ambas categorías y suficientemente elevadas en cuanto a especificidad y clasificación. Esto es lo que se visualiza en el diagrama Coomans de la figura 12. Comparándolo con los de las figuras 10 y 11 es evidente la mejora.

Este modelo es un excelente punto de partida para el proceso de depuración. En lo que sigue mostramos la evolución al reducir el número de componentes (factores internos) en cada categoría para detectar la posible inclusión de información no significativa.

La otra vía posible de afinamiento del modelo: eliminación de los objetos anómalos, ha sido explorada y no ofrece ninguna característica que la haga especialmente interesante, además en el contexto de este trabajo no es razonable diferenciar subgrupos dentro de la D. Origen o por vinificaciones específicas salvo que explicaran la variabilidad de los datos por encima de la debida al origen. En todo caso ese análisis habría de contar ineludiblemente con la colaboración explícita de las Bodegas y Cooperativas para determinar las características comunes al grupo de objetos anómalos.

La tabla 14 muestra los rasgos fundamentales cuando el modelo SIMCA se construye con tres componentes en cada clase explicándose el 75.5% y 73.2% de la varianza respectivamente. Es grande la similitud con la tabla 13, se aprecia una disminución en la capacidad discriminante de los monoglucósidos (variables 13 a 17) y salvo una disminución generalizada en la capacidad modelante no se observa ningún cambio estructural. Los resultados en cuanto a clasificación son igual-

mente excelentes y la especificidad y sensibilidad descienden en ambas categorías de forma equilibrada.

Se prosigue el proceso de reducción de componentes al caso de dos (65.3% y 66.6% de la varianza respectivamente), véase la tabla 15. Se siguen manteniendo los rasgos fundamentales puesto de relieve en los modelos SIMCA de las tablas 13 y 14 con la excepción de la pérdida de capacidad modelante, en los vinos de Toro, de las variables edad química (EQ2) y el índice de polimerización (INDP), a quienes hay que atribuir el descenso en la especificidad.

En este punto del discurso podemos afirmar la posibilidad de un modelado de los vinos de Toro y de Ribera de Duero, en base a su estructura antociánica y parámetros ligados al color. De hecho concurren los tres modelos recogidos en las Tablas 13, 14 y 15 con características suficientemente buenas.

La evolución de un modelo no ha de hacerse solamente en base a su capacidad de clasificación sino que ha de contemplarse su capacidad de predicción. Puede darse el caso de un modelo que discrimine muy bien los vinos que han intervenido en su construcción, pero que sea inestable y no sea capaz de clasificar correctamente otras nuevas muestras, tal modelo es perfectamente inútil.

Piénsese que aun cuando en los modelos construídos sólo intervienen de dos a cinco componentes, que juegan el papel de variables, la relación muestras/componentes no es muy elevada al menos en los vinos de Toro; además existen muchas bodegas que aportan una única muestra. En consecuencia puede ocurrir que cada muestra esté identificada muy específicamente por el valor en alguna componente y al eliminar esta muestra el modelo cambie substancialmente lo que se evidenciaría mediante una pobre capacidad de predicción.

**Tabla 14. SIMCA con tres componentes principales**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 13 DF | 0.4016 | 0.5984 | 1.861 |
| 14 CI | 0.3658 | 0.5701 | 1.500 |
| 15 PT | **0.7789** | 0.7122 | 1.569 |
| 16 PE | 0.6194 | 0.4080 | 1.932 |
| 17 MV | 0.7237 | **0.7703** | 1.339 |
| 18 DFA | 0.2989 | 0.1540 | 1.492 |
| 19 CIA | 0.3128 | 0.0981 | 1.253 |
| 20 PTA | 0.6347 | 0.5219 | 1.690 |
| 21 PEA | 0.0000 | 0.2854 | **2.504** |
| 22 MVA | 0.4779 | 0.5822 | 1.713 |
| 23 DFC | 0.5175 | 0.6997 | **2.748** |
| 24 CIC | 0.5313 | 0.0414 | 1.272 |

| Variable | | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|---|
| | | cat. 1 | cat. 2 | |
| 25 | PTC | 0.5027 | 0.4116 | **2.626** |
| 26 | PEC | 0.5355 | 0.4966 | 1.241 |
| 27 | MVC | 0.6868 | 0.7240 | 1.446 |
| 28 | IC | 0.6069 | 0.7169 | **2.109** |
| 29 | IC2 | 0.5789 | 0.6988 | 1.976 |
| 30 | TONO | 0.1716 | 0.3184 | 1.757 |
| 31 | II | 0.1943 | 0.5645 | 1.295 |
| 32 | IIS | 0.5656 | 0.3998 | **2.425** |
| 33 | IIMS | 0.2064 | 0.3039 | **2.620** |
| 34 | ACG | 0.5846 | 0.6186 | 1.677 |
| 35 | MONB | 0.7195 | 0.7291 | 1.832 |
| 36 | PRB | 0.3638 | 0.5032 | 1.291 |
| 37 | PPB | 0.6140 | 0.6526 | 1.630 |
| 38 | PVP | 0.3278 | 0.1817 | 1.446 |
| 39 | ACTS | 0.5714 | 0.6947 | **2.200** |
| 40 | ACIS | 0.4931 | 0.3938 | 1.396 |
| 41 | EQ1 | 0.4469 | 0.5511 | 1.381 |
| 42 | EQ2 | **0.7703** | 0.5511 | 1.352 |
| 43 | INDP | **0.7693** | 0.5581 | 1.381 |
| SENSIBILIDAD | | 78.9% | 87.2% | |
| ESPECIFICIDAD | | 83.0% | 73.7% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 18 | 1 | 94.7% |
| Ribera | 0 | 47 | 100.0% |
| | Capacidad global de clasificación: | | 98.5% |

**Tabla 15. SIMCA con dos componentes principales**

| Variable | | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|---|
| | | cat. 1 | cat. 2 | |
| 13 | DF | 0.4081 | 0.5039 | 1.623 |
| 14 | CI | 0.2969 | 0.1761 | 1.180 |
| 15 | PT | **0.7666** | 0.7030 | 1.642 |
| 16 | PE | 0.5725 | 0.2908 | 1.917 |
| 17 | MV | 0.7209 | **0.7647** | 1.246 |
| 18 | DFA | 0.2967 | 0.1422 | 1.490 |

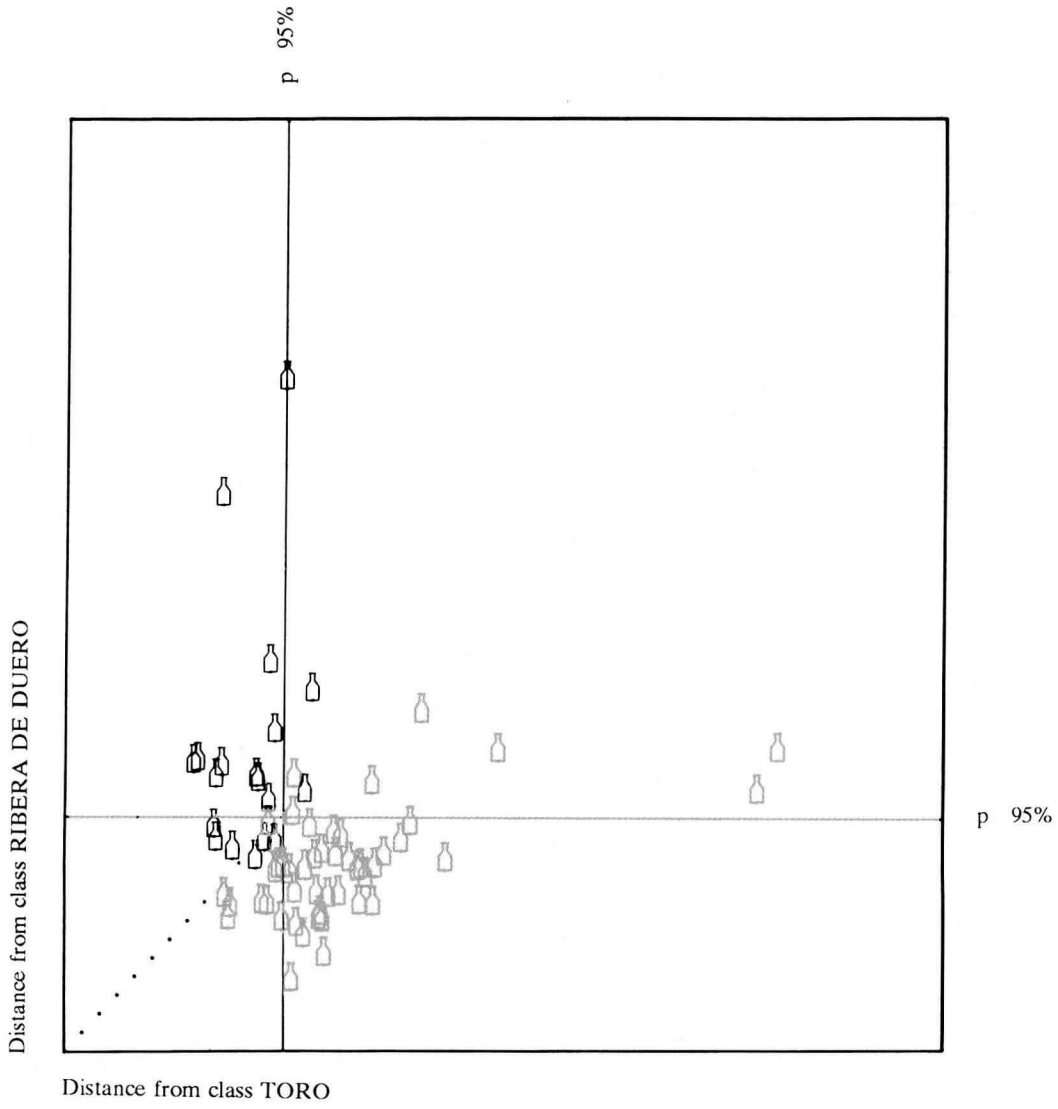
| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 19 CIA | 0.0237 | 0.0523 | 1.060 |
| 20 PTA | 0.5116 | 0.3851 | 1.559 |
| 21 PEA | 0.0000 | 0.2876 | **2.554** |
| 22 MVA | 0.3407 | 0.5870 | 1.353 |
| 23 DFC | 0.4342 | 0.5784 | **2.171** |
| 24 CIC | 0.3268 | 0.0523 | 1.285 |
| 25 PTC | 0.4618 | 0.3904 | **2.448** |
| 26 PEC | 0.5472 | 0.4937 | 1.254 |
| 27 MVC | 0.6836 | 0.6891 | 1.282 |
| 28 IC | 0.5448 | 0.7171 | 1.983 |
| 29 IC2 | 0.5421 | 0.7008 | 1.925 |
| 30 TONO | 0.0481 | 0.0339 | 1.684 |
| 31 II | 0.1378 | 0.4904 | 1.216 |
| 32 IIS | 0.4346 | 0.3634 | **2.140** |
| 33 IIMS | 0.0626 | 0.2586 | **2.362** |
| 34 ACG | 0.5517 | 0.6007 | 1.587 |
| 35 MONB | 0.5919 | 0.7166 | 1.691 |
| 36 PRB | 0.3222 | 0.4979 | 1.283 |
| 37 PPB | 0.5609 | 0.5876 | 1.531 |
| 38 PVP | 0.2643 | 0.1891 | 1.352 |
| 39 ACTS | 0.5780 | 0.6967 | **2.193** |
| 40 ACIS | 0.2084 | 0.3536 | 1.190 |
| 41 EQ1 | 0.4384 | 0.5541 | 1.643 |
| 42 EQ2 | 0.5354 | 0.5450 | 1.152 |
| 43 INDP | 0.5349 | 0.5510 | 1.168 |
| SENSIBILIDAD | 89.5% | 85.1% | |
| ESPECIFICIDAD | 74.5% | 68.4% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 15 | 4 | 78.9% |
| Ribera | 2 | 45 | 95.7% |
| | Capacidad global de clasificación: | | 90.9% |

Para estimar la capacidad de predicción se ha procedido a seleccionar objetos (entre un 10% y un 25% del total) al azar en cada categoría, con los restantes se ha construído el modelo SIMCA con dos, tres y cuatro-cinco componentes. En cada caso se ha aplicado el modelo para predecir la categoría de los objetos que no intervinieron en la construcción. El resultado de la predicción se anota como

FIGURA 12


p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

porcentaje de aciertos por categoría y globalmente, de modo análogo a como se ha anotado la capacidad de clasificación. Además del proceso anterior se ha seguido otro no aleatorio: se construyen tres grupos llamados de cancelación de modo que todos los objetos participen dos ves en el modelado. Estos tres grupos no han intervenido en la construcción de los correspondientes modelos SIMCA y han sido usados para la predicción de este modo se evita se evita la posibilidad de sesgo en el muestreo aleatorio que condicione los resultados. En la tabla 16 se recogen los porcentajes totales de predicción. En cada caso se dispone también de los correspondientes porcentajes de acierto en clasificación cuyos porcentajes totales se recogen así mismo en la tabla 16.

**Tabla 16. Porcentaje de aciertos en clasificación y predicción**

| | Clasificación | Predicción |
|---|---|---|
| MODELO 1 | | |
| Categoría 1 (4 comp.) | 88.9 | 73.9 |
| Categoría 2 (5 comp.) | 100.0 | 98.4 |
| Global | 96.7 | 81.9 |
| MODELO 2 | | |
| Categoría 1 (3 comp.) | 92.0 | 83.3 |
| Categoría 1 (3 comp.) | 100.0 | 97.6 |
| Global | 97.5 | 93.1 |
| MODELO 3 | | |
| Categoría 1 (2 comp.) | 87.5 | 82.6 |
| Categoría 2 (2 comp.) | 97.1 | 93.4 |
| Global | 94.3 | 90.5 |

Estos resultados son muy aceptables y de la misma magnitud que los que habitualmente se encuentran en publicaciones sobre modelado de vinos europeos y norteamericanos. En sí mismos garantizan una estabilidad suficiente al modelo construido. Como era previsible las diferencias no son grandes entre un modelo y otro.

Es posible otra evaluación de las características del modelo construído. Esta valoración es aplicable con independencia de la distribución probabilística subyacente para las variables medidas. Se trata de una estimación en base a los porcentajes de acierto y error de las probabilidades actuales de los errores $\alpha$ y $\beta$ (de primera y segunda especie). Cada modelo construido puede considerarse como la regla de decisión de un test de hipótesis. Como simple estimación de porcentajes estos valores han sido propuestos muy recientemente (M.P. Derde y

otros en Journal of Chemometrics. vol. 3, 1898), bajo los nombres de "selectividad" y "especificidad". En todos los modelos SIMCA construidos se ha hecho mención a estos índices, en este momento es necesaria una detallada explicación dado su mérito para evaluar los modelos construídos.

Dado un modelo para la clase de los vinos de Toro (modelo 1) se define:

— *Sensibilidad del modelo 1:* Proporción de los objetos que perteneciendo a la clase Toro se clasifican correctamente.

— *Especificidad del modelo 1:* Proporción de los objetos que siendo ajenos al modelo de la clase Toro se clasifican como ajenos.

Las definiciones son análogos para el modelo de Ribera (modelo 2).

En términos de test de hipótesis estos índices tienen un significado clave, decidir si un vino es de Toro equivale a construir el siguiente test:

Hipótesis nula, $H_o$: El vino es de Toro.
Hipótesis alternativa: El vino no es de Toro.
Región crítica del test: El conjunto de objetos que están fuera del modelo SIMCA construído.

La probabilidad del error de primer tipo, $\alpha$, es por definición la probabilidad de rechazar la hipótesis nula cuando es cierta, es decir la probabilidad de obtener unas medidas químicas que proporcionen un punto perteneciente a la región crítica (fuera del recinto SIMCA) cuando realmente es un vino de Toro. Si comparamos con la definición de sensibilidad del modelo 1 se tiene como estimación de $\alpha$ para este modelo:

$$\alpha = pr \{ \text{rechazar } H_o / H_o \text{ es cierta} \} = 1 - \text{sensibilidad}$$

La probabilidad del error de segunda especie, $\beta$, es por definición la probabilidad de aceptar la hipótesis nula cuando es falsa, es decir aceptar como vino de Toro un vino que no lo es. Si tenemos en cuenta la definición de especificidad se tiene:

$$\beta = pr \{ \text{aceptar } H_o / H_o \text{ es falsa} \} = 1 - \text{especificidad}$$

Los resultados pueden interpretarse de este modo: Dadas las variables medidas de una muestra de vino, se admitirá que es de Toro si pertenece al modelo SIMCA construído con cuatro y cinco componentes (Tabla 13). Con esta regla de decisión se tiene una probabilidad $\alpha = 1 - 0.895 = 0.105$ de rechazar un vino de Toro cuando en realidad si lo es y una probabilidad $\beta = 1 - 0.915 = 0.085$ de aceptar como vino de Toro una muestra de vino que no lo es.

Igualmente se considera el problema de hacer un modelo para los vinos de Ribera con el mismo planteamiento precedente:

| | |
|---|---|
| Hipótesis nula, $H_o$: | El vino es de Ribera |
| Hipótesis alternativa: | El vino no es de Ribera |
| Región crítica del test (rechazo de $H_o$): | Conjunto de objetos que están fuera del modelo SIMCA construído. |

Conviene destacar que esta región crítica no es la complementaria de la correspondiente al test anterior.

En el caso del modelo de la tabla 13 se tiene $\alpha = 1 - 0.894 = 0.106$, es decir con esta probabilidad rechazaremos un vino Ribera cuando en realidad lo era. Y una probabilidad $\beta = 1 - 0.895 = 0.105$, es decir aceptaremos un vino Ribera cuando era de Toro casi con la misma probabilidad.

**Tabla 17. Sensibilidad y especificidad**

| | $\alpha$ pr { rechazar $H_o$ / / $H_o$ es cierta} | $\beta$ pr { aceptar $H_o$ / / $H_o$ es falsa} |
|---|---|---|
| MODELO 1 | | |
| Categoría 1 (4 comp.) | 0.105 | 0.085 |
| Categoría 2 (5 comp.) | 0.106 | 0.105 |
| MODELO 2 | | |
| Categoría 1 (3 comp.) | 0.211 | 0.170 |
| Categoría 2 (3 comp.) | 0.128 | 0.263 |
| MODELO 3 | | |
| Categoría 1 (2 comp.) | 0.105 | 0.255 |
| Categoría 2 (2 comp.) | 0.149 | 0.316 |

Desde este punto de vista el mejor modelo de los tres es el modelo 1 (Tabla 13) con cuatro componentes para la categoría de los vinos de Toro y cinco para la de los vinos de Ribera del Duero.

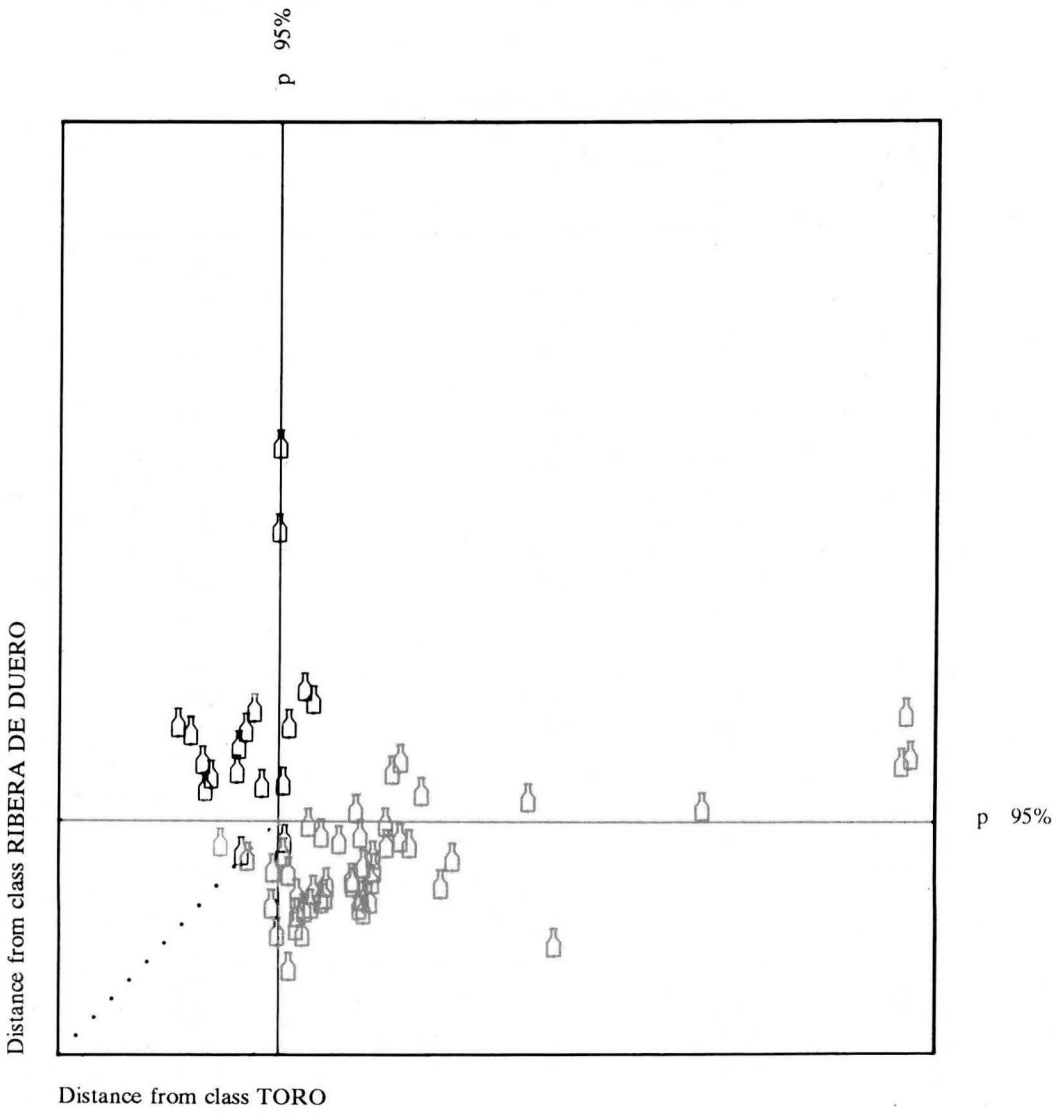
En nuestra opinión, este riesgo de equivocarse con el modelo SIMCA propuesto es mínimo. Debemos considerar que en las muestras de vino existen vinificaciones diferentes e incluso muestras de vino de años diversos (1985, 1986 y 1987).

### 3.3.6. **Clasificación con todos los parámetros**

Puesto que SIMCA no tiene limitación en cuanto al número de variables a utilizar tenemos otro término de comparación para el modelo construído: usar conjuntamente las 43 variables, es decir usar toda la información disponible.

Los resultados numéricos del modelo SIMCA con tres componentes por cate-

FIGURA 13


p 95%
p 95%
Distance from class RIBERA DE DUERO
Distance from class TORO

goría (67.7% y 66.55 de varianza explicada) se recogen en la tabla 18. Respecto a la tabla 13, modelo SIMCA con cuatro y cinco componentes para la estructura antociánica y parámetros ligados al color, la similitud es notable; tanto más cuanto que se observa la pérdida de capacidad modelante de los parámetros convencionales (compárese con el modelo de la tabla 8) y de los compuestos polifenólicos (modelo de la tabla 9).

El Diagrama de Coomans de este modelo se muestra en la figura 13, los modelos son algo más disjuntos que los construídos en el epígrafe anterior figura 12, pero por contra muchas muestras están fuera del recinto de su verdadera clase.

**Tabla 18. SIMCA con tres componentes principales**

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 1 GA | 0.0576 | 0.2170 | **2.174** |
| 2 AV | 0.1666 | 0.2189 | 1.868 |
| 3 AT | 0.2545 | 0.2549 | 1.580 |
| 4 pH | 0.0827 | 0.0219 | 1.094 |
| 5 AP | 0.6186 | 0.5766 | 1.288 |
| 6 BP | 0.5363 | 0.6167 | 1.773 |
| 7 C | 0.4700 | 0.6850 | 1.568 |
| 8 DP | 0.3801 | 0.6439 | 1.163 |
| 9 EP | 0.0000 | 0.5809 | 1.411 |
| 10 V/F | 0.3396 | 0.0597 | **3.66** |
| 11 V/LA | 0.1286 | 0.2317 | 2.103 |
| 12 SV | 0.5730 | 0.3494 | 1.801 |
| 13 DF | 0.4641 | 0.4837 | **2.007** |
| 14 CI | 0.3745 | 0.3351 | 1.436 |
| 15 PT | **0.7939** | 0.6553 | 1.656 |
| 16 PE | 0.5360 | 0.3363 | 1.666 |
| 17 MV | 0.7348 | **0.7572** | 1.396 |
| 18 DFA | 0.1789 | 0.1377 | 1.471 |
| 19 CIA | 0.2694 | 0.0352 | 1.375 |
| 20 PTA | 0.5703 | 0.3936 | 1.890 |
| 21 PEA | 0.0000 | 0.2921 | **2.499** |
| 22 MVA | 0.3087 | 0.5903 | 1.422 |
| 23 DFC | 0.3380 | 0.5924 | **2.183** |
| 24 CIC | 0.4853 | 0.0602 | 1.323 |
| 25 PTC | 0.4889 | 0.3377 | **2.528** |
| 26 PEC | 0.5070 | 0.4897 | 1.290 |
| 27 MVC | 0.6962 | 0.7114 | 1.528 |

| Variable | Capacidad Modelante | | Capacidad discriminante |
|---|---|---|---|
| | cat. 1 | cat. 2 | |
| 28 IC | 0.5846 | 0.6877 | 1.833 |
| 29 IC2 | 0.5687 | 0.6680 | 1.739 |
| 30 TONO | 0.2200 | 0.2610 | 1.534 |
| 31 II | 0.1675 | 0.5856 | 1.484 |
| 32 IIS | 0.4770 | 0.3880 | **2.001** |
| 33 IIMS | 0.1230 | 0.3018 | **2.732** |
| 34 ACG | 0.6208 | 0.5529 | 1.437 |
| 35 MONB | **0.7581** | 0.7123 | 1.367 |
| 36 PRB | 0.3980 | 0.4973 | 1.255 |
| 37 PPB | 0.6477 | 0.6807 | 1.280 |
| 38 PVP | 0.3485 | 0.2139 | 1.617 |
| 39 ACTS | 0.5825 | 0.7315 | **2.334** |
| 40 ACIS | 0.2912 | 0.3764 | 1.713 |
| 41 EQ1 | 0.4779 | 0.5394 | 1.321 |
| 42 EQ2 | 0.7108 | 0.5253 | 1.679 |
| 43 INDP | 0.7101 | 0.5339 | 1.708 |
| SENSIBILIDAD | 84.2% | 89.4% | |
| ESPECIFICIDAD | 80.9% | 84.2% | |

MATRIZ DE CLASIFICACIÓN

| | Categoría computada | | |
|---|---|---|---|
| | Toro | Ribera | |
| Cat. verdadera | | | |
| Toro | 18 | 1 | 94.7% |
| Ribera | 0 | 47 | 100.0% |
| | Capacidad global de clasificación: | | 98.5% |

Se ha realizado un análisis de la capacidad de predicción de este modelo en forma análoga a la del epígrafe anterior. El resultado se muestra en la tabla 19 sin que se observe una mejora significativa respecto del modelo SIMCA construído sin los parámetros enológicos convencionales ni los compuestos polifenólicos.

**Tabla 19. Porcentaje de aciertos en clasificación y predicción**

| | Clasificación | Predicción |
|---|---|---|
| Categoría 1 (3 comp.) | 93.3 | 85.0 |
| Categoría 2 (3 comp.) | 93.3 | 98.1 |
| Global | 96.9 | 94.9 |

En la tabla 20 se muestran las probabilidades de los errores de primer y

segundo tipo asociadas a los modelos de cada categoría. Son algo mayores incluso que el modelo SIMCA con cuatro y cinco componentes construido en el epígrafe anterior (tabla 17). De nuevo se corrobora la inutilidad de las variables añadidas a los efectos de modelado.

**Tabla 20. Sensibilidad y especificidad**

| | $\alpha$<br>pr { rechazar $H_0$ /<br>/ $H_0$ es cierta} | $\beta$<br>pr { aceptar $H_0$ /<br>/ $H_0$ es falsa} |
|---|---|---|
| Categoría 1 (3 comp.) | 0.158 | 0.191 |
| Categoría 2 (3 comp.) | 0.106 | 0.158 |

### 3.4. **Modelado mediante regresión mínimo-cuadrática parcial (PLS)**

Para reafirmar la posibilidad de distinguir las muestras de Toro de las de la Ribera de Duero partiendo de un enfoque totalmente distinto se ha ideado la construcción de un modelo basado en el método de regresión conocido en la literatura como PLS (Partial Least Squares).

PLS es una técnica de regresión que reúne un grupo de buenas propiedades en la tarea de describir una o varias variables respuesta Y mediante un bloque de variables predictoras $X_i$. Dada la relativa novedad del método haremos una sucinta descripción para el caso de una única variable respuesta Y.

Una regresión multilineal por mínimos cuadrados determina la combinación lineal de las variables X que está más correlacionada con la variable Y. El punto débil del método es la tendencia a incluir mucho "rumor de fondo" de las variables descriptoras en el intento de mejorar la correlación con la Y.

En el extremo opuesto se sitúa la regresión sobre las componentes principales de las variables predictoras. En este caso se usan como variables predictoras las primeras componentes para garantizar la eliminación del rumor en los datos, además la ortogonalidad de las componentes principales (epígrafe 2.4) atenúa en gran medida posibles problemas de mal condicionamiento causados por correlaciones y colineidades elevadas entre las variables predictoras. La regresión en componentes principales proporciona, en general, un modelo más estable y con mejores intervalos de confianza para los parámetros que la regresión multilineal. El mayor inconveniente radica en que las componentes se seleccionan sin tomar en consideración la variable respuesta. En concreto pueden existir variables predictoras que varíen en poca cuantía (poca varianza) pero que estén muy relacionadas con la respuesta, en la determinación de las componentes principales estas variables serán omitidas a causa de su pequeña variabilidad que las hace aparecer en el conjunto total como ruido de fondo (últimas componentes principales).

PLS es una técnica que se sitúa entre ambos extremos: usa componentes principales pero se van incorporando de una en una al modelo y en cada paso se la selecciona de modo que se tenga la máxima correlación con la variable respuesta. Utiliza de forma óptima la información aportada por las variables predictoras para generar la combinación lineal más estable y a la vez más correlacionada con la respuesta Y.

En general no se ha usado el método PLS como técnica de modelado, sin embargo en nuestro problema parece sumamente adecuado porque: i) Sabemos que se ha de usar una estructura latente después del análisis previo realizado en el capítulo 2. ii) Hemos construido un modelo estable con estructura latente distinta para cada categoría. iii) Existe la posibilidad de que alguna variable haya sido descartada por su poca variabilidad y por tanto de mejorar el modelo con su inclusión.

El procedimiento consta de tres etapas:

— Se define la variable respuesta binaria
Y (muestra de vino) = 1, si es de Toro
Y (muestra de vino) = 2, si es de Ribera

— Se hace la regresión PLS de la variable. Y sobre las cuarenta y tres variables predictoras que contienen toda la información disponible para construir el modelo.

— Con los valores estimados por PLS se construirá el modelo como un test de hipótesis que permita evaluar la especifidad y sensibilidad en cada categoría.

El avance en el proceso de explicar la variable indicadora de la categoría se recoge en la tabla 21 mediante el porcentaje de varianza explicada. Tan importante como este índice es la varianza "cros-validada" obtenida con tres grupos de cancelación para los que se ha recalculado el modelo y determinado la capacidad de predicción sobre las muestras que no intervinieron en su construcción. Son valores satisfactorios y apuntan a la posibilidad de considerar sólo tres componentes, ya que el incremento al añadir la cuarta es relativamente muy pequeño.

El seguimiento del efecto de añadir cada componente puede hacerse con facilidad a través de los Diagramas Box-Whisker de los valores calculados. La figura 14 es el Diagrama de los valores calculados con una componente. Comparándolo con los Diagramas de las variables con mayor peso de Fisher (figuras 1, 2 del epígrafe 2.2) se aprecia ya una mayor separación de ambas categorías. Esta separación va creciendo con el número de componentes, figuras 15, 16 y 17. Al pasar de tres a cuatro componentes, de la figura 16 a la 17, se observan pocos cambios, de hecho la nueva componente globalmente sólo incorpora un objeto (el 55) hasta este momento anómalo al grupo.

**Tabla 21. Varianza de la variable Y explicada por el modelo PLS**

| | Varianza (%) | Varianza C.V. |
|---|---|---|
| Con 1 componente | 50.3 | 41.5 |
| Con 2 componentes | 62.2 | 46.1 |
| Con 3 componentes | 75.8 | 48.9 |
| Con 4 componentes | 80.8 | 50.9 |

Puesto que se pretende construir un test de hipótesis para los valores calculados es conveniente verificar la normalidad de los mismos para decidir el tipo de test a usar. Esto se muestra en la tabla 21 según la cual se puede aceptar la normalidad para ambas categorías a partir de dos componentes.

**Tabla 22. Normalidad de los valores calculados con PLS**

| | CAT. 1 (TORO) | | | | | CAT. 2 (RIBERA) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | D | V | W | U | A | D | V | W | U | A |
| Con 1 cp. | a | a | a | a | a | r | rr | r | r | r |
| Con 2 cp. | a | a | a | a | a | a | a | a | a | a |
| Con 3 cp. | a | a | a | a | a | a | a | a | a | a |
| Con 4 cp. | a | a | a | a | a | a | a | a | a | a |

Se anota: rr para $p <= 0.01$
r para $0.01 < p <= 0.05$
a para $0.05 < p$

A la vista de la evolución de la regresión parece adecuado tomar tres componentes como modelo definitivo PLS. La normalidad de los valores con tres componentes se corrobora mediante los Diagramas de normalidad para cada categoría, cuya interpretación es sencilla: cuanto más alineados esten los puntos más adecuada es la hipótesis de normalidad.

La figura 18, Diagrama correspondiente a los vinos de Toro, muestra una pauta lineal junto con un valor aislado anómalamente elevado que ya se había apreciado en el Diagrama Box-Whisker (Figura 16). Más evidente es la normalidad de los valores sobre los vinos de Ribera mostrada en la figura 19.

Finalmente en la figura 20 se recogen conjuntamente los histogramas de los valores calculados con PLS y tres componentes.

El modelo de cada categoría es el intervalo de confianza al nivel de significación igual al 95% que se recoge en la tabla 23. Es significativo que los dos modelos son disjuntos, los intervalos que los definen no tienen puntos en común.

FIGURA 14

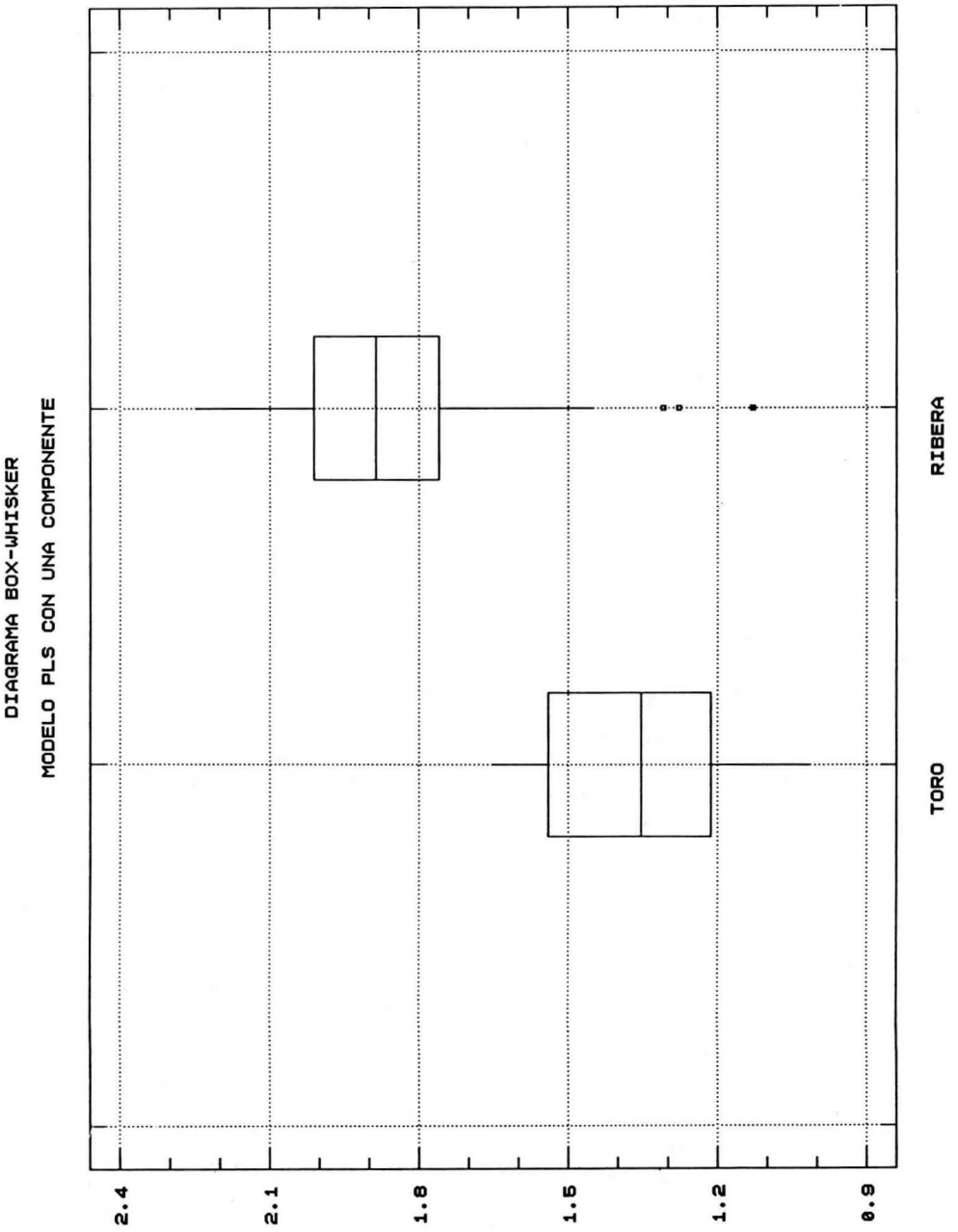


DIAGRAMA BOX-WHISKER
MODELO PLS CON UNA COMPONENTE
2.4
2.1
1.8
1.5
1.2
0.9
TORO
RIBERA

FIGURA 15

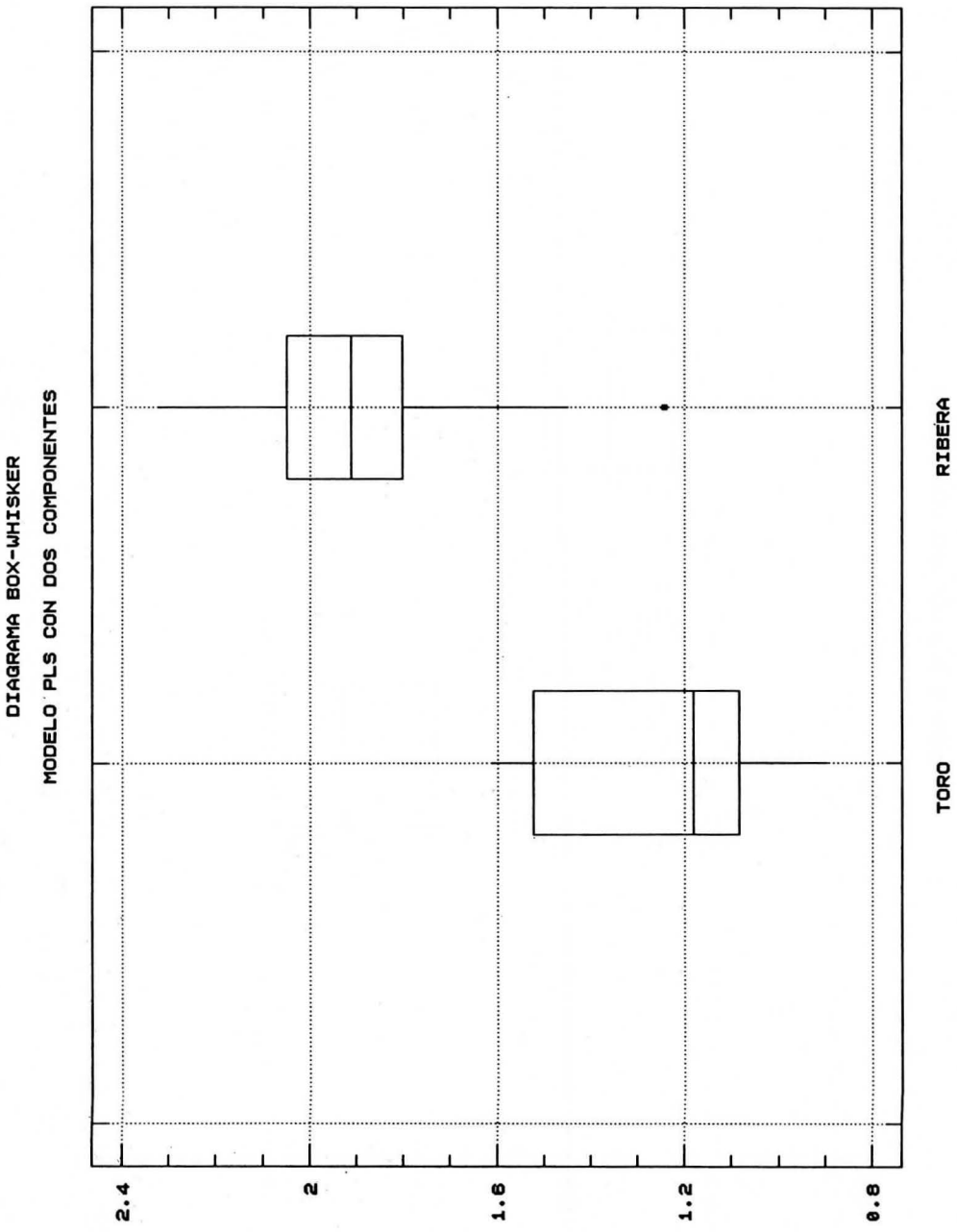


DIAGRAMA BOX-WHISKER
MODELO PLS CON DOS COMPONENTES
2.4
2
1.6
1.2
0.8
TORO
RIBERA

FIGURA 16

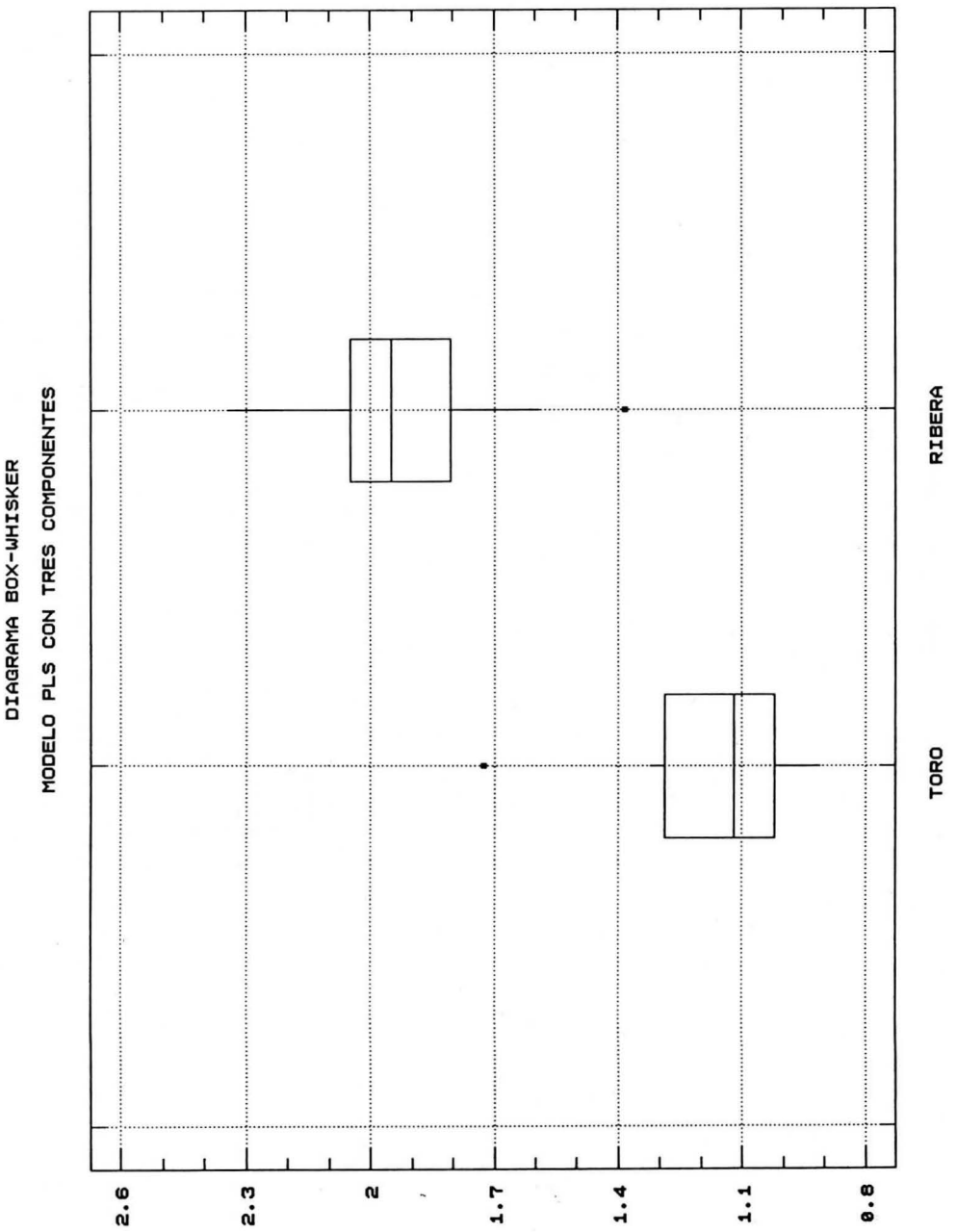


DIAGRAMA BOX-WHISKER
MODELO PLS CON TRES COMPONENTES
2.6
2.3
2
1.7
1.4
1.1
0.8
TORO
RIBERA

FIGURA 17

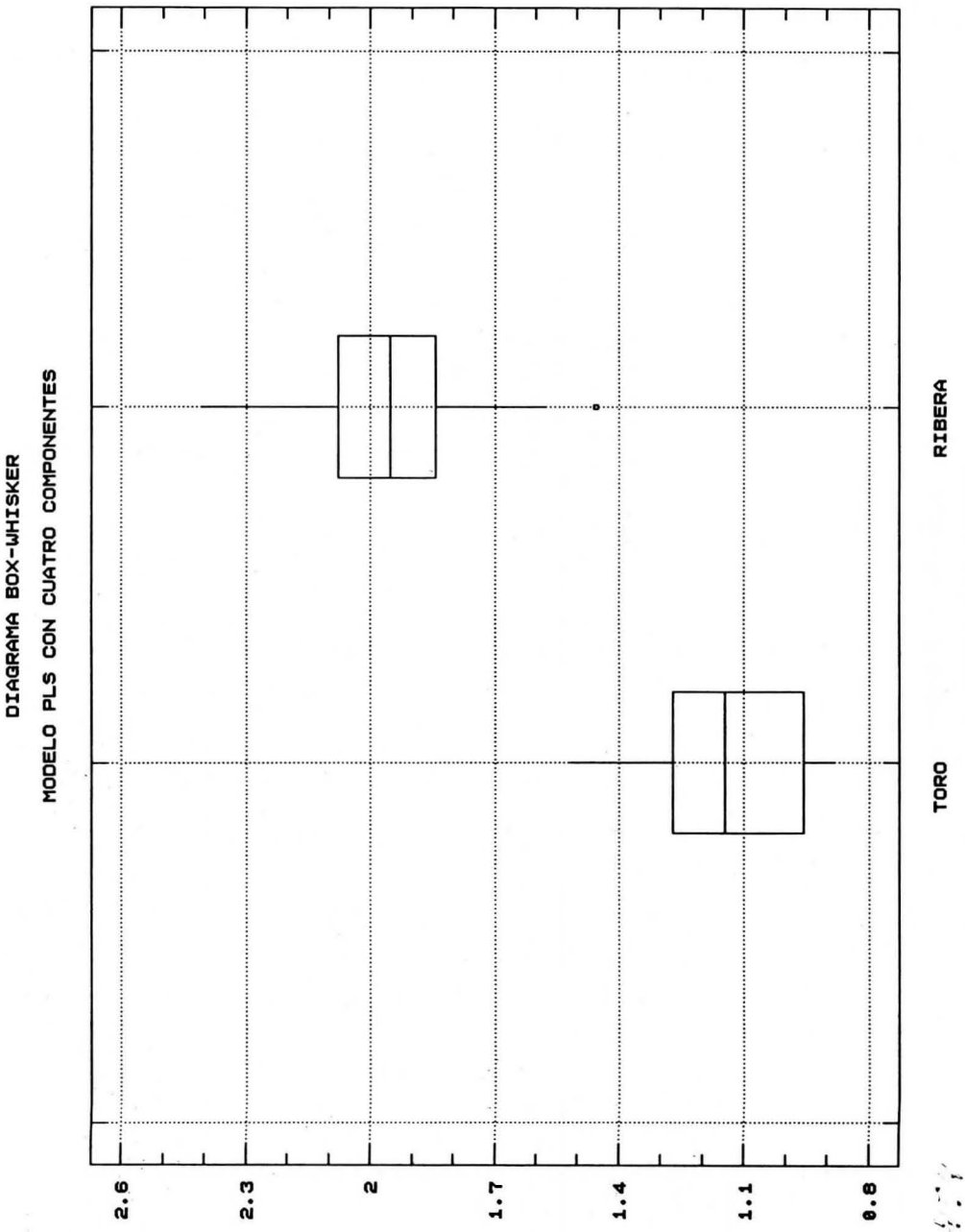


DIAGRAMA BOX-WHISKER
MODELO PLS CON CUATRO COMPONENTES
2.6
2.3
2
1.7
1.4
1.1
0.8
TORO
RIBERA

Otro aspecto notable es la igualdad de desviaciones típicas lo que hace que la regla de decisión sea equilibrada.

**Tabla 23. Modelo obtenido con PLS**

| | Distribución Normal | | | |
|---|---|---|---|---|
| | Media | Des. t. | Modelo al 95% | |
| TORO | 1.165 | 0.187 | 0.797 | 1.532 |
| RIBERA | 1.933 | 0.196 | 1.549 | 2.318 |

Aplicando las definiciones de sensibilidad y especificidad (epígrafe 3.3.5) se tienen lo valores mostrados en la tabla 24 bajo el título "Fr." (frecuencial). Existe otra posibilidad para calcular la sensibilidad y la especificidad que también se recoge en la misma tabla 24 y que incorpora la información adicional sobre la distribución de probabilidad conocida para cada clase. En este caso se puede determinar exactamente los valores teóricos esperados en la siguiente forma:

— *Modelo de Toro,*

$\alpha$ = {rechazar $H_o$ / $H_o$ es cierta} = pr { C(0.797,1.532) / N (1.165,0.187) } = 0.050
$\beta$ = pr {aceptar $H_o$ /$H_o$ es falsa} = pr { (0.797,1.532) / N(1.933,0.196) } = 0.020

— *Modelo de Ribera,*

$\alpha$ = {rechazar $H_o$ / $H_o$ es cierta} = pr { C(1.549,2.138) / N (1.933,0.196) } = 0.050
$\beta$ = pr {aceptar $H_o$ /$H_o$ es falsa} = pr { (1.549,2.318) / N(1.165,0.187) } = 0.020

La tabla 24 evidencia sin lugar a dudas la perfecta separación de ambas categorías, siendo mejor modelo incluso que SIMCA con cuatro y tres componentes, epígrafe 3.3.5.

**Tabla 24. Modelo PLS. Sensibilidad y especificidad**

| | $\alpha$ pr { rechazar $H_o$ / / $H_o$ es cierta} | | $\beta$ pr { aceptar $H_o$ / / $H_o$ es falsa} | |
|---|---|---|---|---|
| | Fr. | Exacta | Fr. | Exacta |
| Categoría 1 (vinos de TORO) | 0.053 | 0.050 | 0.022 | 0.020 |
| Categoría 2 (vinos de RIBERA) | 0.043 | 0.050 | 0.053 | 0.020 |

Un aspecto importante es establecer las variables con mayor capacidad modelante, es decir las que más intervienen en la construcción de la regresión PLS.

FIGURA 18

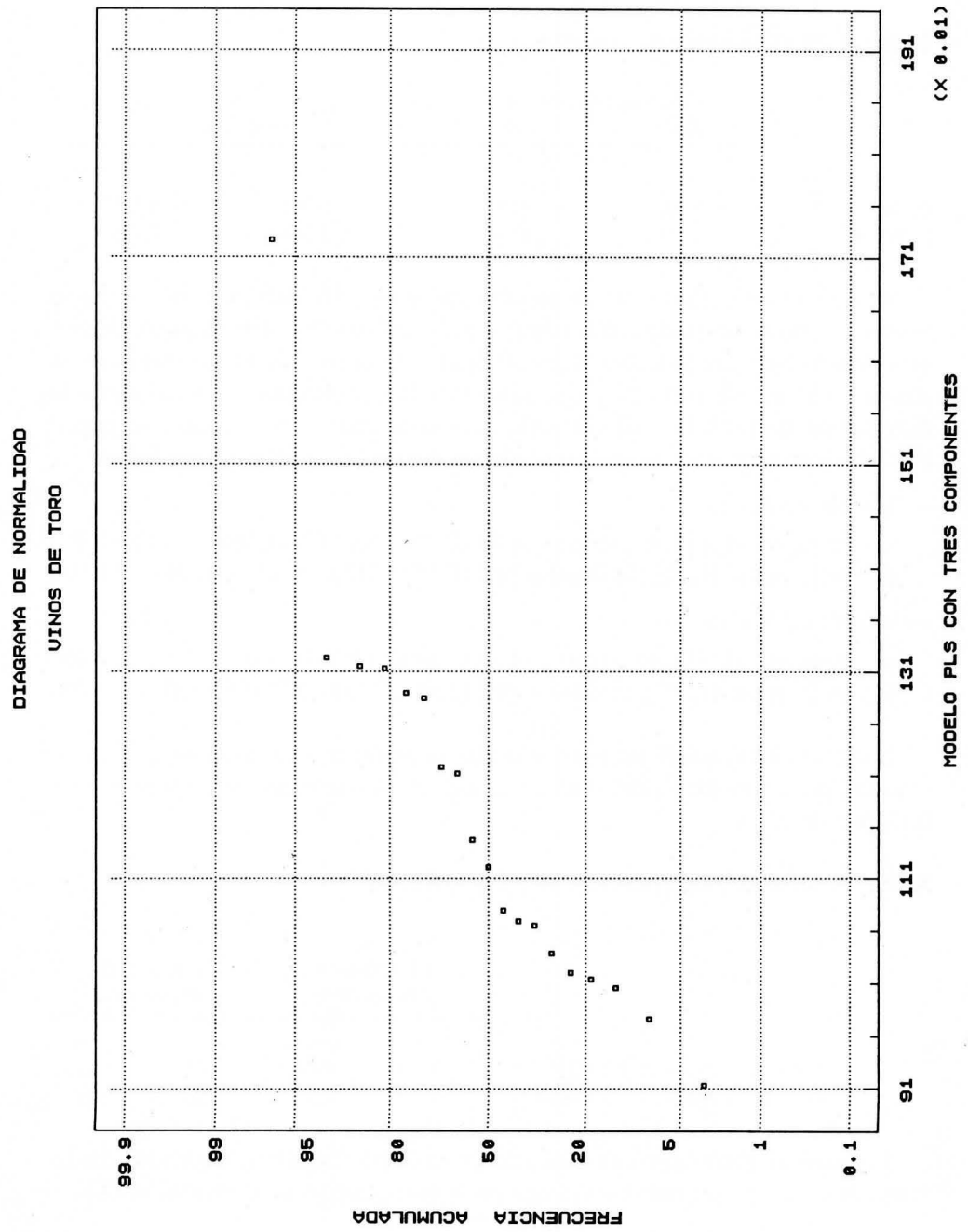


DIAGRAMA DE NORMALIDAD
VINOS DE TORO
FRECUENCIA ACUMULADA
99.9
99
95
80
50
20
5
1
0.1
91
111
131
151
171
191
(X 0.01)
MODELO PLS CON TRES COMPONENTES

FIGURA 19

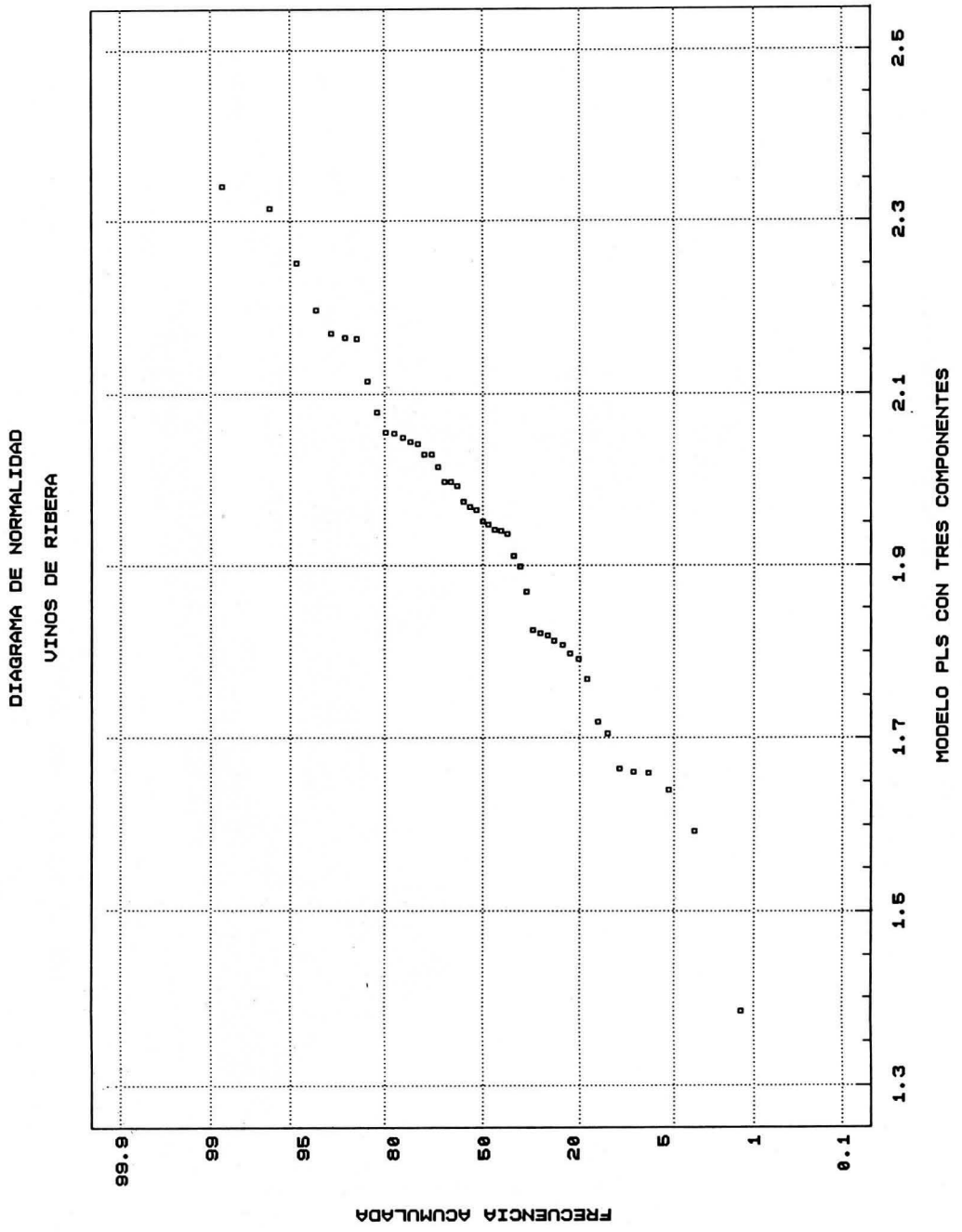


DIAGRAMA DE NORMALIDAD
VINOS DE RIBERA
FRECUENCIA ACUMULADA
99.9
99
95
80
50
20
5
1
0.1
1.3
1.5
1.7
1.9
2.1
2.3
2.5
MODELO PLS CON TRES COMPONENTES

FIGURA 20


HISTOGRAMA DE FRECUENCIAS
RIBERA
TORO
24
20
16
12
8
4
0
0.8
1.1
1.4
1.7
2
2.3
2.6
MODELO PLS CON TRES COMPONENTES

Son notables, valor de la varianza explicada mayor que el 70%:
— el 3-monoglucósido de malvidol (MV, variable n. 17)
— el porcentaje de monómeros (MONB, variable n. 35)

Con valores entre el 60% y el 70% se encuentran:
— el 3-monoglucósido de petudinol (PT, variable n. 15)
— el cumarato de malvidol (MVC, variable n. 27)
— el porcentaje de polímeros pardos (PPB, variable n. 37)
— los antocianos totales (ACTS, variable n. 39)

Entre el 50% y el 60% están:
— los polifenoles totales (AP, variable n. 5 y BP, variable n. 6)
— el 3-monoglucósido de definidol (DF, variable n. 13).
— las intensidades colorantes (IC, variable n. 28 e IC2, variable n. 29)
— los antocianos totales (ACG, variable n. 34)
— las edades químicas (EQ1, variable n. 41 y EQ2, variable n. 2)
— el índice de polimerización (INDP, variable n. 43)

### 3.5. **Conclusiones respecto del modelado**

- El método jerárquico de Ward aplicado tanto sobre las variables originales como sobre componentes principales agrupa por separado los vinos de Toro y los de la Ribera con gran nitidez.
- La clasificación KNN proporciona un porcentaje global de clasificaciones correctas en torno al 90%
- Se ha procedido a la caracterización de los vinos de Toro en relación con los vinos de Ribera de Duero mediante modelos SIMCA (Soft Independent Modeling of Class Analogy). La metodología se ha seleccionado en base a las características que presentan los datos, después de comprobar que cada una de las 43 variables disponibles no puede caracterizar ambas clases.
- Los parámetros enológicos convencionales no permiten un modelado SIMCA satisfactorio de las categorías. Lo mismo cabe decir respecto de los compuestos polifenólicos.
- El modelo SIMCA construído mediante la estructura antociánica no posibilita suficiente especificidad para los vinos de Toro. Lo mismo cabe decir para el grupo de variables resumidas bajo el nombre de parámetros ligados al color, índices de polimerización y de ionización.
- Conjuntamente las variables citadas en el punto anterior permiten un modelado SIMCA satisfactorio. Son necesarios cuatro factores internos para modelar las muestras de Toro y cinco para las de Ribera. La habilidad de clasificación es de 88.9% y 100% para los vinos de Toro y Ribera respectivamente (global del 96.7%). Su capacidad de predicción es del 73.9% y del

98.4% para los vinos de Toro y Ribera respectivamente (global del 81.9%). Con este modelo la probabilidad de rechazar un vino de la D. O. Toro cuando realmente procede de ella es igual 0.105 mientras que la de aceptar un vino como de la D. O. Toro cuando en realidad no procede de ella es 0.085.

- No se mejora el modelo SIMCA por la inclusión de los parámetros enológicos convencionales y los contenidos en polifenoles.
- Considerando todas las variables se ha elaborado un modelo mediante la técnica PLS (Partial Least Squares) que permite fijar a priori la probabilidad de rechazar un vino de la D. O. de Toro. Si esta probabilidad se fija en 0.05 se tiene una probabilidad 0.020 de aceptar un vino como procedente de la D. O. Toro cuando realmente no lo es. La habilidad de clasificación de este modelo es 100% para los vinos de Toro y 95.6% para los de Ribera (global del 97%).
- Para cada modelo se ha establecido en cuánto contribuye cada variable a modelar los vinos de la D. O. Toro y a discriminarlos respecto de los de la D. O. Ribera de Duero.