Análisis factorial de datos cualitativos: un experimento con datos artificiales

POR

L. RAMIREZ DIAZ (1)

Y

F. SANCHO ROYO (2)

SUMMARY

In this work some properties and applications of different factor analysis techniques of qualitative data are discussed. Using an artificially generated set of data we show that the principal component analysis of the total coincidence matrix (rotated solutions) and the analysis of correspondences are the most relevant for studies of descriptive ecology.

RESUMEN

En este trabajo se discuten algunas propiedades y aplicaciones de diferentes técnicas de análisis factorial de datos cualitativos (análisis de componentes principales de la matriz de contingencias múltiples y de coincidencias totales y análisis de correspondencias). Los datos utilizados se han generado de forma artificial, haciendo variar los comportamientos de las variables en sus presencias y ausencias.

Se muestra que el análisis de componentes principales de la matriz de coincidencias totales (soluciones rotadas) y el análisis de correspondencias presentan las mayores ventajas de utilización en estudios de tipificación de comunidades en ecología descriptiva.

(1) Departamento de Ecología. Facultad de Ciencias. Universidad de Murcia.
(2) Departamento de Ecología. Facultad de Ciencias. Universidad de Sevilla.

INTRODUCCION

Frente a los métodos de clasificación, las ordenaciones producidas por las técnicas de análisis factorial (por ejemplo, análisis de componentes principales) presentan, entre otras, las siguientes diferencias en los resultados que se obtienen:

1. La representación (estructuración) de las observaciones no es por medio de grupos discontinuos, sino a través de gradientes. Esta propiedad es muy conocida y se recoge con el nombre de ordenación. Está expresada en los factores o coeficientes de carga de las variables, calculados a partir de los autovectores de la matriz de correlación del análisis de componentes principales.
2. La propiedad de recoger comportamientos opuestos entre grupos de variables, que se traduce en la existencia de una polaridad que afecta a los factores y que corresponde, por ejemplo, en los estudios de tipificación ecológica de comunidades a una distribución de distinto signo de las especies afectadas (González Bernáldez y otros, 1977).

La primera propiedad aludida es muy conocida; examinaremos, sobre todo, la segunda en relación con distintos tipos de tratamiento factorial de datos cualitativos:

a) Análisis de componentes principales de la matriz de correlación basada en tablas de «contingencias múltiples» (Yarranton, 1967; Ivimey-Cook y Proctor, 1967) que denominaremos $ACPC_t$.

b) Análisis factorial de correspondencias (Escofier-Cordier, 1965, 1969; Bénzecri, 1969; González Bernáldez y otros, 1977), $ACPC_r$.

c) Análisis de componentes principales de la matriz de correlación basada en tablas que llamaremos de «coincidencias totales» y que designamos como $ACPC_t$.

Las diferencias entre $ACPC_t$ y $ACPC_i$ se basan, sobre todo, en la manera de tratar las dobles ausencias de los datos.

González Bernáldez y cols. (1973) comentan que cuando el método $ACPC_t$ se empleaba para el análisis de datos obtenidos en la elección a la derecha o a la izquierda de un test de preferencias, los resultados no establecían relación entre la simultaneidad de elección a la izquierda y no elecciones a la derecha, considerando sólo las primeras, que en el estudio se codificaban como presencias. Es decir, dos variables x_1 y x_2, con valores de las observaciones del tipo:

		OBSERVACIONES													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
VARIABLES	x_1	0	0	0	0	1	1	1	1	0	0	0	1	1	1
	x_2	1	1	1	1	0	0	0	0	1	1	1	1	1	1

aparecen independientes en el análisis y relacionadas por tanto con dos componentes distintos e independientes ($r = 0$) cuando su verdadera relación es del tipo $r = -1$.

Ivimey-Cook y Proctor (1967) proponen la utilización de las soluciones rotadas (rotación varimax) de $ACPC_t$, por su claridad al interpretar —en los estudios de tipificación— las direcciones de variación detectadas.

Hill (1973) pone a punto un método, al que reconoce de igual funcionamiento que el usado por Hatheway (1971) y al análisis factorial de correspondencias $ACPC_r$ (Escofier-Cordier, 1965), indicando que todos son variantes del análisis de componentes principales y que debían resumirse con el nombre de «reciprocal-averaging», que él mismo propone.

En el presente trabajo se pretende discutir la comparación de los resultados de los análisis de componentes principales de las matrices de contingencias ($ACPC_t$), coincidencias ($ACPC_i$) y análisis factorial de correspondencias ($ACPC_r$) sobre una tabla original de datos cualitativos generados artificialmente y en los que se investigan las distintas ordenaciones desde el punto de vista de las relaciones abstractas entre las variables, una vez conocido el comportamiento de las mismas en las observaciones. Estos métodos son de enorme utilidad e interés en la

tipificación de las comunidades vegetales y animales dentro del campo de los estudios de ecología cuantitativa, en los que las observaciones son muestras o inventarios y variables las especies estudiadas.

MATERIAL Y METODOS

Para obtener los datos se ha construido una matriz de 50 observaciones. Estas pueden considerarse dentro de dos grupos: A (observaciones 1 a 25) y B (observaciones 26 a 50).

En las observaciones se han registrado las presencias y ausencias de un total de 35 variables, repartidas en siete grupos (cada cinco variables forman un grupo con igual distribución en las 50 observaciones).

Cada grupo de variables puede presentarse en cada uno de los grupos de observaciones (A y B) en tres formas diferentes:

— Estar presente en el 96 por 100 del total de observaciones (variable muy frecuente).
— Estar presente en el 50 por 100 del total de observaciones (variable de frecuencia media y aleatoria).
— Estar presente en un 4 por 100 del total de observaciones (variable de frecuencia rara).

Se utilizan las letras mayúsculas de los dos grupos de observaciones (A y B) para denominar a la frecuencia de ocurrencia del 96 por 100 en dichos grupos; las letras minúsculas, para las frecuencias del 50 por 100, y el símbolo 0 para las frecuencias raras del 4 por 100.

En la tabla 1 se presenta la matriz original de los datos con la siguiente simbología de acuerdo con lo anterior:

Grupos de variables	*% de ocurrencias en las observaciones tipo A*	*% de ocurrencias en las observaciones tipo B*
Ab	96	50
aB	50	96
AO	96	4
OB	4	96
aO	50	4
Ob	4	50
ab	50	50

Así pues, los siete grupos de variables, con cinco repeticiones cada uno, representan las combinaciones posibles de ocurrencias de las variables en los dos grupos de observaciones, eliminando las ausencias y presencias totales.

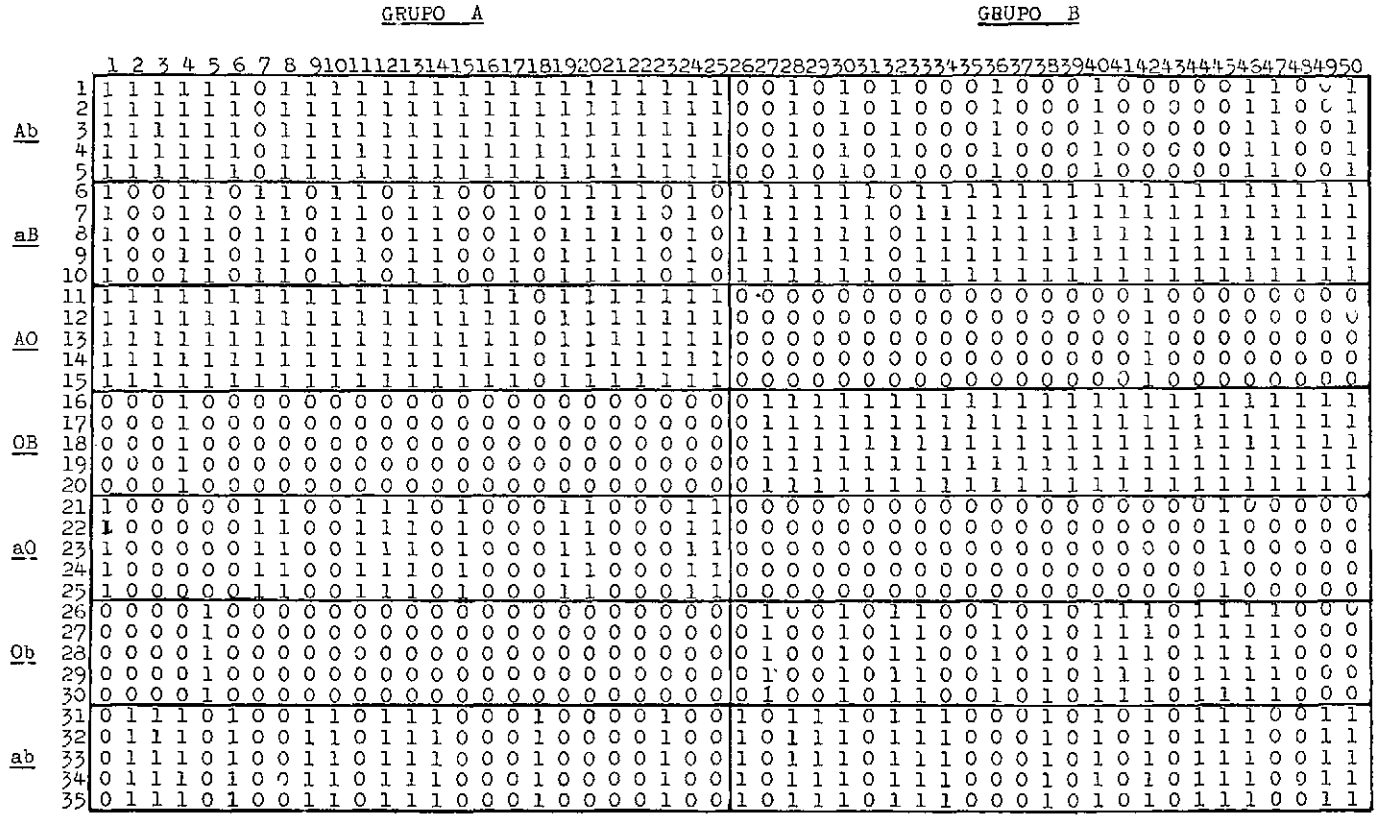

TABLA 1.—Matriz original de datos con la denominación de los grupos de variables Ab, aB, AO, OB, aO, Ob, ab. A y B = 96 % de presencias; a y b = 50 % de presencias; O = 4 % de presencias.

La elaboración de la matriz se ha realizado utilizando en todos los casos extracciones de números aleatorios; por ello, se ha estudiado un número relativamente alto de observaciones que permitieron obtener las proporciones de frecuencias anteriormente citadas.

Por tanto, partimos de una matriz de 35 variables (filas) por 50 observaciones (columnas) (ver tabla 1), cuyos valores x_{ij} son 1 ó 0, según la variable i esté presente o ausente en la observación j.

Para el tratamiento de la matriz se han empleado las técnicas de análisis de componentes principales y la de análisis de correspondencias en la forma siguiente:

Análisis de componentes principales

A partir de la matriz original (tabla 1) se han construido las siguientes matrices:

Contingencias múltiples

De dimensiones 35 por 35, donde cada elemento a_{ij} representa el número de observaciones que coinciden en poseer a la vez las variables i y j. Los valores de la diagonal principal de esta matriz son el número total de presencias de cada variable en las 50 observaciones analizadas.

Coincidencias totales

De dimensiones 35 por 35, donde cada elemento a_{ij} representa el número de observaciones que coinciden en poseer y no poseer, a la vez, a las variables i y j. Los valores de la diagonal principal de esta matriz son todos iguales a 50 (número total de observaciones).

A partir de ambos tipos de matrices se realizaron los análisis de componentes principales correspondientes mediante los siguientes cálculos:

- Matriz de correlación entre filas o columnas (las matrices son simétricas).
- Cálculo de autovalores, autovectores, porcentajes de absorción de varianza y matriz de factores (Harman, 1967; Gittins, 1969).
- Rotación varimax (Harman, 1967) con el cálculo de una nueva matriz de factores y un nuevo reparto del porcentaje inicial de varianza absorbida.

Análisis de correspondencias

Para realizar este análisis sobre la matriz original de 1 y 0 se siguieron los siguientes pasos:

— Cálculo de la matriz de correspondencias (Escofier-Cordier, 1965).
— Obtención de autovalores, autovectores de dicha matriz y porcentaje de inercia (variación) de cada uno de ellos.
— Obtención de los factores de dispersión de las variables y observaciones (coordenadas de variables y observaciones en el mismo subespacio).

RESULTADOS

Los resultados de los distintos análisis se exponen de forma separada para cada uno de ellos, considerando en los análisis de componentes principales la comparación entre soluciones rotadas y sin rotar. Los resultados se presentan resumidos para los siete grupos de variables estudiadas.

Análisis de componentes principales. Matriz de contingencias múltiples

$ACPC_t$ coordenadas no rotadas:

El porcentaje total de varianza absorbida ha sido del 99,5 por 100 para los primeros cinco componentes, repartiéndose entre ellos de la siguiente forma:

$$I = 52,\ II = 35,\ III = 9,\ IV = 3\ y\ V = 0{,}5$$

Los factores de carga (contribuciones de cada grupo de variables a la ordenación) se presentan en la tabla 2.

TABLA 2

Variables	COMPONENTES		
	I	*II*	*III*
Ab	—0,48	0,83	0,15
aB	0,44	0,78	—0,43
AO	—0,76	0,63	0,03
OB	0,90	0,36	—0,19
aO	—0,88	0,26	—0,36
Ob	0,94	0,15	—0,08
ab	0,43	0,70	0,52

En la figura 1 se presentan las ordenaciones de los grupos de variables analizados en los planos definidos por los componentes I-II y I-III. No se presentan otras combinaciones de ejes por ofrecer resultados similares a los anteriores.

$ACPC_t$ coordenadas rotadas:

El nuevo reparto de la varianza con la rotación ha sido el siguiente:

$$I = 20,\ II = 34,\ III = 17,\ IV = 19 \text{ y } V = 0,5$$

Los factores de carga rotados, promediados para cada grupo de variables, son los que aparecen en la tabla 3.

TABLA 3

Variables	COMPONENTES			
	I	*II*	*III*	*IV*
Ab	0,08	0,96	0,23	0,11
aB	0,95	0,24	0,18	— 0,03
AO	— 0,08	0,91	0,04	0,38
OB	0,81	— 0,31	0,31	— 0,38
aO	— 0,08	0,60	— 0,34	0,71
Ob	0,60	— 0,30	0,15	— 0,71
ab	0,33	0,20	0,90	— 0,19

En la figura 2 se presentan las proyecciones de los siete grupos de variables en los planos definidos por los componentes I-II, I-III y I-IV.

Análisis de componentes principales. Matriz de coincidencias totales

$ACPC_i$ coordenadas no rotadas:

El porcentaje total de varianza absorbida fue del 99,5 por 100 y su reparto en los cinco primeros componentes se realizó de la manera siguiente:

$$I = 61,\ II = 17,\ III = 17,\ IV = 4 \text{ y } V = 0,5$$

Los factores de carga para los tres primeros componentes se exponen en la tabla 4.

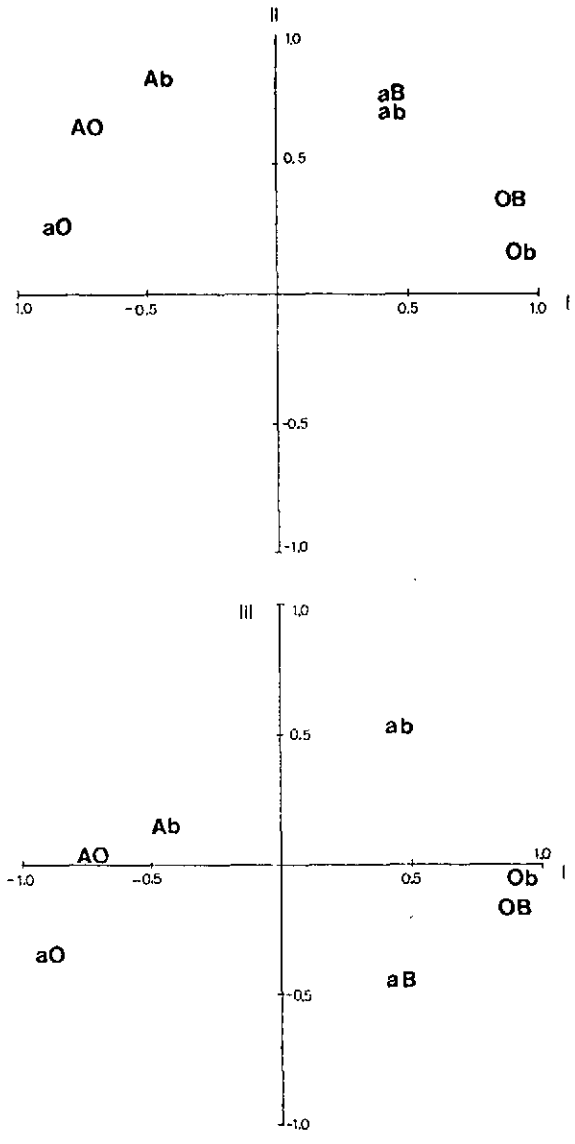

FIGURA 1.—Análisis de componentes principales de la matriz de contingencias múltiples. ($ACPC_t$). Ordenación de los grupos de variables en los planos definidos por los ejes I-II y I-III en soluciones no rotadas.

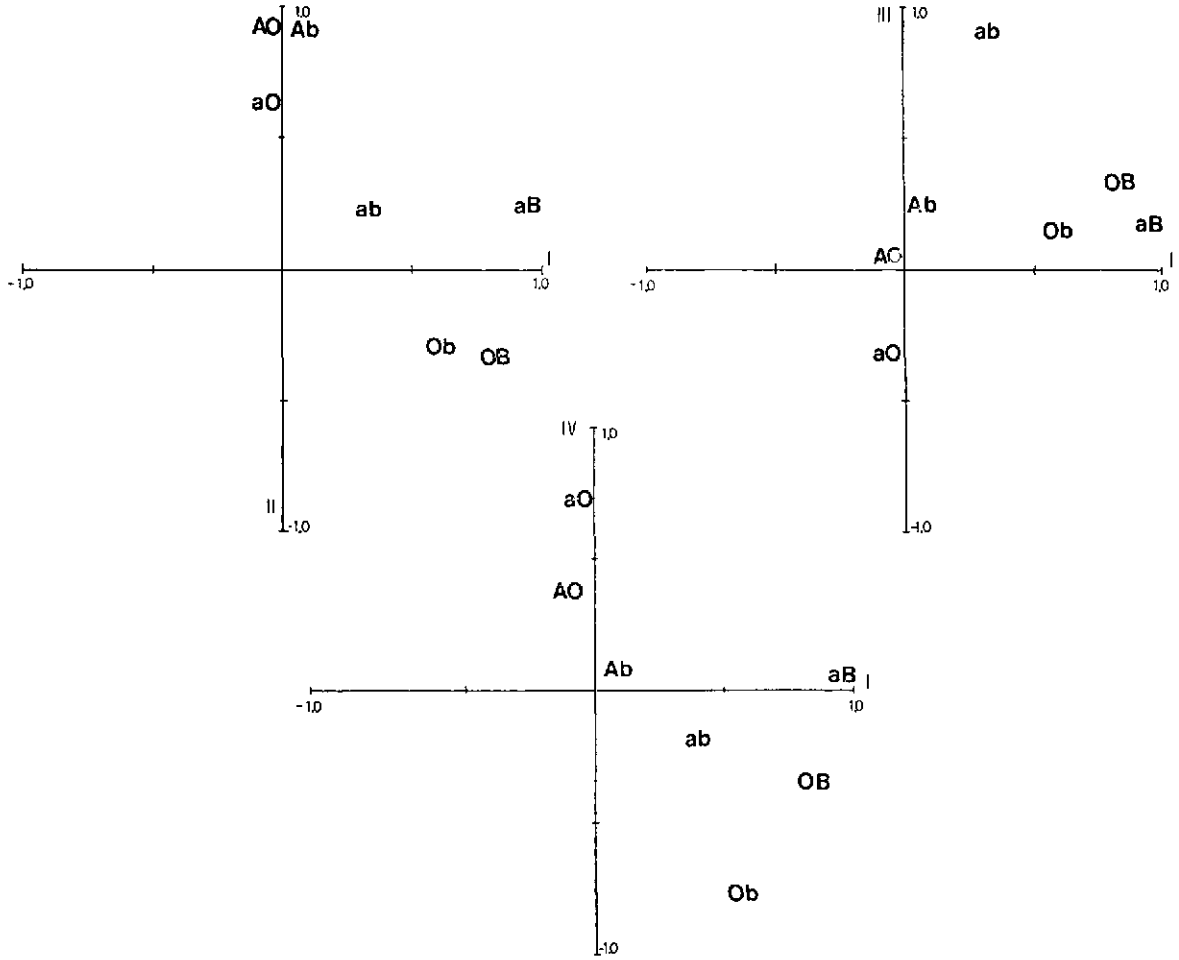

FIGURA 2

Análisis de componentes principales de la matriz de contingencias múltiples. ($ACPC_t$). Ordenación de los grupos de variables en los planos definidos por los ejes I-II, I-III y I-IV en soluciones rotadas.

TABLA 4

Variables	COMPONENTES		
	I	*II*	*III*
Ab	0,86	0,17	0,35
aB	— 0,61	— 0,61	0,48
AO	0,99	0,00	0,06
OB	— 0,99	— 0,04	— 0,06
aO	0,72	— 0,16	— 0,63
Ob	— 0,74	0,17	— 0,58
ab	— 0,38	0,86	0,26

En la figura 3 se presentan las proyecciones de los grupos de variables en los planos definidos por los componentes I-II y I-III.

$ACPC_1$ coordenadas rotadas:

El reparto de la varianza total absorbida en los nuevos componentes rotados fue como sigue:

$$I = 46,\ II = 20,\ III = 28,\ IV = 3\ y\ V = 2{,}5$$

Los nuevos factores de carga pueden verse en la tabla 5.

TABLA 5

Variables	COMPONENTES		
	I	*II*	*III*
Ab	0,93	— 0,02	— 0,24
aB	— 0,23	— 0,17	0,94
AO	0,84	— 0,24	— 0,44
OB	— 0,84	0,20	0,47
aO	0,21	— 0,53	— 0,78
Ob	— 0,91	0,10	— 0,09
ab	— 0,14	0,99	— 0,02

En la figura 4 se presentan las proyecciones de los grupos de variables en los planos definidos por los componentes I-II y I-III.

ANÁLISIS DE CORRESPONDENCIAS. $ACPC_r$

El análisis resultó de alta eficacia al absorber los cinco primeros ejes el 97 por 100 de la inercia total (varianza) de la matriz de correspondencia. El reparto en los ejes fue el siguiente:

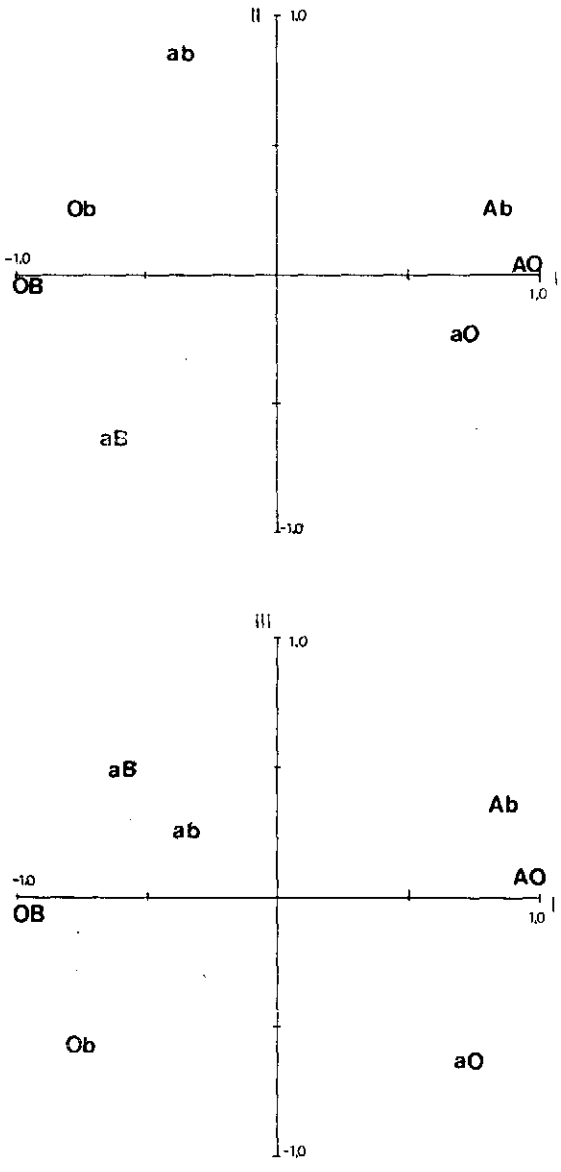

FIGURA 3.—Análisis de componentes principales de la matriz de coincidencias totales. ($ACPC_i$). Ordenación de los grupos de variables en los planos definidos por los ejes I-II y I-III en soluciones no rotadas.

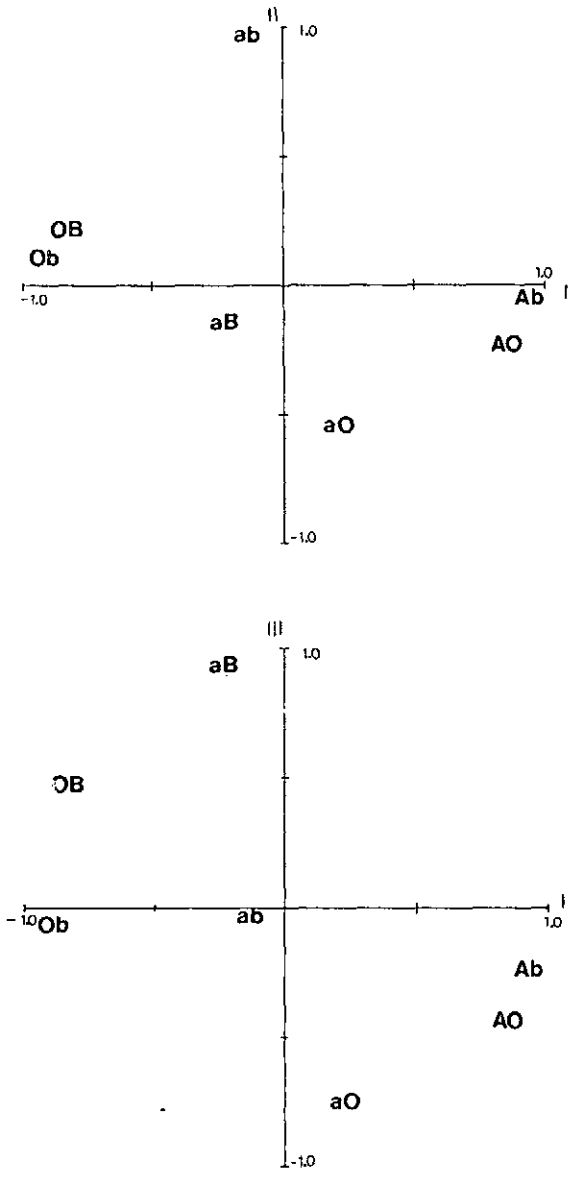

FIGURA 4.—Análisis de componentes principales de la matriz de coincidencias totales. ($ACPC_j$). Ordenación de los grupos de variables en los planos definidos por los ejes I-II, I-III en soluciones rotadas.

$$I = 49,\ II = 21,\ III = 13,\ IV = 10 \text{ y } V = 4$$

Los factores de dispersión de las variables en el nuevo subespacio y referido a los cuatro primeros ejes son como sigue en la tabla 6.

TABLA 6

Variables	EJES			
	I	*II*	*III*	*IV*
Ab	— 0,52	— 0,28	— 0,10	0,24
aB	0,29	0,33	0,26	0,14
AO	— 0,90	— 0,08	— 0,02	0,26
OB	0,98	0,23	0,24	0,02
aO	— 1,11	0,98	— 0,20	— 0,72
Ob	0,88	0,01	— 1,10	0,03
ab	0,17	— 0,76	— 0,17	— 0,47

En la figura 5 se presentan las proyecciones de los siete grupos de variables en los planos definidos por los ejes I-II, I-III y I-IV.

DISCUSION Y CONCLUSIONES

La discusión de los análisis se realizará de forma separada para cada uno de ellos, finalizando con una síntesis o resumen de las principales características de los tres tipos de resultados.

Análisis de componentes principales. $ACPC_t$. Matriz de contingencias múltiples

El primer componente no rotado del análisis (fig. 1) presenta un 50 por 100 de absorción de varianza, y discrimina con polaridad a las variables con presencia en la zona B de las variables con presencia en la zona A. Sin embargo, los valores más altos de los factores los poseen los grupos de variables Ob (positivos) y aO (negativos), con ocurrencias de tipo aleatorio (50 por 100) en cada grupo de observaciones.

El componente II (fig. 1) no posee polaridad y resulta simplemente de una ordenación de los grupos de variables por sus números de presencias (frecuencia de aparición). Los valores de los factores de este componente muestran una correlación de $r = + 0,92$ (altamente significativo para $p \leqslant 0,01$) con el número de presencias de cada grupo de variables.

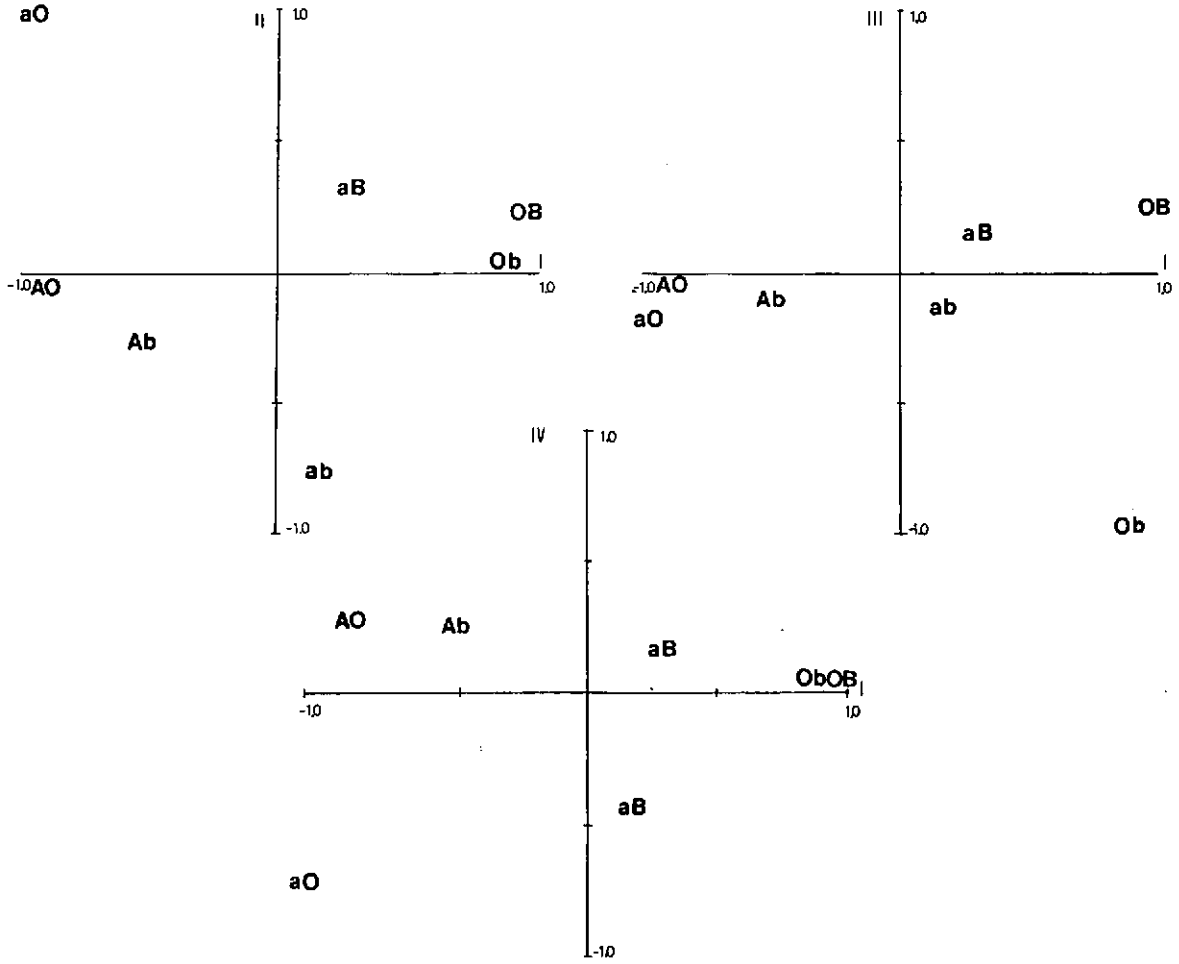

FIGURA 5.—Análisis de correspondencias. ($ACPC_r$). Ordenación de los grupos de variables en los planos definidos por los ejes I-II, I-III y I-IV.

El componente III (fig. 1) no posee una fuerte polaridad y enfrenta al grupo ab con los grupos aB y aO. Los componentes IV y V carecen de interés por su escasa información.

Las soluciones rotadas de este análisis presentan los siguientes puntos de interés: el reparto de la varianza se hace en igual proporción para los ejes I-II y III-IV. El componente I (fig. 2) discrimina con valores altos positivos a los grupos de variables con elevado número de presencias en B (aB y OB) y el componente II (fig. 2) a los de elevado número de presencias en A (Ab y AO). Ninguno de los anteriores componentes posee polaridad, repitiendo el segundo la información del primero, puesto que las presencias en B corresponden a las ausencias en A y viceversa.

El componente III (fig. 2), sin polaridad marcada, establece una separación entre los grupos ab y aO que no tiene un claro sentido, salvo el poner de manifiesto el grupo ab de ocurrencias aleatorias.

El componente IV (fig. 2) enfrenta con polaridad a los grupos aO y Ob que ya estaban discriminados en los componentes I y III.

Análisis de componentes principales. $ACPC_i$. Matriz de coincidencias totales

Las soluciones no rotadas presentan un primer componente (fig. 3) que recoge el 60 por 100 de la varianza de los datos, separando con fuerte polaridad los grupos de variables AO (valores positivos) y OB (valores negativos).

El componente II (fig. 3) discrimina claramente al grupo de variables con distribución aleatoria, ab, del grupo aB, que presenta ocurrencias muy frecuentes; este componente es parecido al tercero del análisis anterior, pero con mayor polaridad.

El componente III (fig. 3), de igual varianza que el anterior, parece responder a una discriminación entre variables con muchas presencias (aB y Ab) y variables con muchas ausencias (Ob y aO).

Las soluciones rotadas de este análisis reducen la importancia numérica del componente I en beneficio del II y III. El componente I (figura 4) discrimina al igual que el no rotado los grupos con presencias en A de los que tienen presencias en B; este componente posee una fuerte polaridad y resume los dos primeros componentes del análisis de la matriz de contingencias múltiples ($ACPC_t$, soluciones rotadas: figura 2).

El componente II (fig. 4) discrimina claramente al grupo de variables con presencias aleatorias, ab.

El tercer componente no es de fácil interpretación; parece reforzar la ordenación con gran polaridad de grupos de variables con presencias en B (aB y OB) frente a grupos con ausencias en B (aO y AO).

Análisis de correspondencias. $ACPC_r$

El eje I, de gran importancia numérica (fig. 5), realiza la separación entre grupos de variables con presencia en B (OB) y grupos de variables con presencias en A (AO), reforzando la polaridad con la oposición de ausencias en A y B (valores positivos Ob, valores negativos aO).

El eje II (fig. 5) separa el grupo de variables de ocurrencias aleatorias, ab, del aO. El eje III, de difícil interpretación (fig. 5), discrimina el grupo Ob, y el eje IV al grupo aO; ambos grupos corresponden a variables de gran número de ausencias y ya estaban discriminados en el eje I.

CONCLUSIONES GENERALES

Dado que el comportamiento de las variables podía referirse a dos tipos de presencias, en A o en B, o a un comportamiento aleatorio de las ocurrencias, podemos obtener, a tenor de los resultados anteriormente expuestos y la discusión realizada, las siguientes conclusiones:

1.ª Los tres tipos de tratamiento realizan la discriminación en un solo eje de los dos tipos más importantes de comportamiento de las variables (el análisis de componentes de la matriz de contingencias $ACPC_t$ en las soluciones no rotadas, en contra de lo expuesto por Ivimey-Cook y Proctor en 1967). Destaca por su fuerte polaridad el análisis de la matriz de coincidencias $ACPC_i$.

2.ª Las soluciones no rotadas de los análisis de las matrices de contingencias y coincidencias $ACPC_t$ y $ACPC_i$ poseen un componente relacionado únicamente con el número de presencias (contingencias) y número de presencias y ausencias (coincidencias). Esto, unido a la importancia numérica del I componente, hace necesaria la utilización de las soluciones rotadas con vistas a la posible interpretación de las direcciones de variación detectadas (en favor de lo expuesto por Ivimey-Cook y Proctor).

3.ª El análisis de correspondencias no presenta ejes distintos del I que posean interés en la ordenación.

4.ª Las soluciones rotadas del análisis de la matriz de contingencias carecen de polaridad y, al tener en cuenta sólo las presencias, funcionan a oscuras del fenómeno de presencias y ausencias opuestas, produciendo dos componentes para una misma discriminación, ignorando que las presencias en A y ausencias en B tienen el mismo significado, pero de signo opuesto, que las ausencias en A y presencias en B.

5.ª El análisis de componentes principales de la matriz de coincidencias totales presenta fuerte polaridad en sus ejes y realiza con menor número de ellos las discriminaciones necesarias para ordenar y poner de manifiesto los tipos más sobresalientes de comportamiento de las variables.

BIBLIOGRAFIA

BENZECRI, J. P. (1969), «Statistical analysis as a tool to make patterns emerge from data». En: S. WATANABE, *Methodologies of Pattern recognition:* 35-60, Academic Press, Nueva York.

ESCOFIER-CORDIER, B. (1965), *L'analyse factorielle des correspondances.* Thèse 3e cycle. Rennes.

ESCOFIER-CORDIER, B. (1969), «L'analyse factorielle des correspondances», *Cah. Bur-Univ. Rech. opér.*, Université de Paris, 13.

GITTINS, R. (1969), «The application of ordination techniques». En: RORISON, I. H., *Ecological aspects of the mineral nutrition of plants,* 37-66, Blackwells, Oxford.

GONZÁLEZ BERNÁLDEZ, F.; GARCÍA NOVO, F., y SANCHO ROYO, F. (1973), «Analyse de réactions face au paysage naturel», *Options méditerranéennes,* 17: 66-81.

GONZÁLEZ BERNÁLDEZ, F.; RAMÍREZ DÍAZ, L.; TORRES MARTÍNEZ, A., y DÍAZ PINEDA, F. (1977), «Estructura de la vegetación de la marisma de la Reserva Biológica de Doñana (Huelva). I, Análisis factorial de datos cualitativos», *Anales de Edafología y Agrobiología,* XXXVI (9-10): 989-1003.

HARMAN, H. H. (1967), *Modern Factor Analysis,* 2nd ed., University of Chicago Press.

HATHEWAY, W. H. (1971), «Contingency-table analysis of rain forest vegetation». En: PATIL, G. P.; PIELOU, E. C., and WATERS, W. E., *Statistical Ecology (3), Many Species Populations, Ecosystems and Systems Analysis,* 271-313, Pennsylvania State University Press, University Park, Pennsylvania.

HILL, M. O. (1973), «Reciprocal averaging: and eigenvector method of ordination», *J. Ecol.,* 61: 237-49.

IVIMEY-COOK, R. B., and PROCTOR, M. C. F. (1967), «Factor analysis of data from an East Devon heath», *J. Ecol.,* 55: 405-13.

SANCHO ROYO, F. (1974), *Actitudes ante el paisaje. Estudio experimental.* Publicaciones de la Universidad de Sevilla, Serie Ciencias, núm. 19.

YARRANTON, G. A. (1967), «Principal component analysis of data from Saxicolous Bryophyte vegetation at steps Bridge, Devon. I. A quantitative assessment of variation in the vegetation», *Canad. J. Botany,* 45: 93-115.