Revista de Historia de la Psicología
1994, vol. 15, nº 3-4, pp. 357-368

ESTUDIO COMPARATIVO DE DOS MÉTODOS ESTADÍSTICOS PARA EL ANÁLISIS DE COFIRMACIONES: PROBLEMAS ASOCIADOS Y PRESENTACIÓN DE UN EJEMPLO DE APLICACIÓN

Tomás Martinez y Fernando Canet
Facultad de Psicología. Universidad de Valencia.

RESUMEN

En este estudio, se hace un comparación entre dos procedimientos estadísticos para el estudio de la colaboración científica, basada en el análisis de la cofirmación de los autores, con el fin de detectar las relaciones científicas existentes entre ellos. Se discute y justifica los parámetros de aplicación de las técnicas de Análisis de homogeneidad, y el Análisis de conglomerados (Cluster). Además, se presentan los resultados de un ejemplo y se discute la adecuación de cada método.

Los resultados obtenidos hacen ver el mejor ajuste de las técnicas de conglomerados, mostrando claramente las agrupaciones entre autores, aunque no es capaz de mostrar claramente a los autores que, siendo los más productivos, actúan de líderes y de conexión entre grupos.

La bibliometría y las nuevas técnicas de tratamiento informático de la documentación son muchas las posibilidades que se nos han abierto para abordar el estudio de la producción científica que existe en un área de investigación. No se puede olvidar que cuando se pretende analizar un área de estudio, el problema no es la carencia de información sino mas bien su exceso. Ningún investigador puede pretender en la actualidad conocer toda la producción científica que se realiza en su área, y menos si pretende incluir los estudios colindantes a su trabajo. Esto ha obligado ha hacer uso de técnicas que permitan obtener una idea clara y rápida del panorama que existe en ese campo: autores más importantes, en que revistas publican, cómo se relacionan entre ellos, etc., facilitando la selección de la documentación a estudiar en profundidad.

Si bien son muchos los problemas que se puede solucionar con las técnicas actualmente existentes, son otros muchos los que quedan por resolver. Problemas que han de ser tratados desde una doble perspectiva estadístico-documental, ya que si bien la estadística es imprescindible en estos nuevos enfoques, no lo es menos conocer las características de las variables a estudiar, y lo que es más importante conocer los tipos de estudios útiles, y su interpretación y aplicación al campo de la documentación. Resultando por lo tanto vital la colaboración entre ambos grupos de científicos, de tal manera que los estudios tengan una aplicabilidad real, a la vez que muestran una eficaz metodología.

En este estudio se ha realizado una primera aproximación al estudio de la colaboración científica. Este problema ha sido especialmente estudiado a partir de que (Price y Beaver,1966) introdujera el concepto de "colegios invisibles" Son muchos los trabajos que podemos encontrar en la actualidad en los que se incluye algún tipo de estudio de la relación entre científicos de un campo. Estos estudios se han llevado a cabo normalmente a través de técnicas sociométricas y/o gráficas. A la vez son muchos los índices empleados en la detección de la organización social de los grupos científicos, cómo son la relación maestro-discipulo, revistas en que publican, el enlace bibliográfico, cocitación, cofirmación, etc.

En esta caso se ha adoptado la cofirmación, ya que además de considerarlo uno de los índices más apropiados y ser fácilmente generalizable a otros índices, presenta la gran ventaja de su tratamiento informático.

El estudio se centra en una comparación de dos técnicas diferentes:

La primera de ellas basada en técnicas de estadística multivariantes de reducción de la dimensionalidad. Estas técnicas no habían sido utilizadas hasta ahora con este tipo de datos, matrices binarias de presencia-ausencia, debido principalmente a problemas metodológicos, pero gracias a las nuevas técnicas estadísticas de escalamiento óptimo, y su inclusión en paquetes estadísticos, han hecho posible este tipo de análisis. El procedimiento en que se basan es una forma particular del ACPC (Análisis de Componentes Principales Categórico), basado en la codificación de variables dispuestas en matrices de indicadores (Gifi 1990). Esta técnica ha recibido diversos nombres, pero nosotros la denominaremos HOMALS, (Análisis de Homogeneidad por Mínimos Cuadrados Alternativos), (Gifi 1990; SPSS Inc, 1990) nominación con la que aparece en el paquete estadístico utilizado. Esta técnica ha sido muy utilizada especialmente con variables nominales. La necesidad de estas técnicas se debe a que el ACP (Análisis de Componentes Principales) no es adecuado para analizar matrices de datos dicotómicas. En términos generales esta ligado con la idea de que diversas variables pueden medir la misma cosa, aunque ello implique alguna perdida de información. Esto se consigue en base a transformaciones no lineales tendentes a minimizar la perdida de información a través de procesos iterativos. Se llega de esta manera a conseguir una seria de soluciones y representaciones gráficas. El número de estas depende del número de dimensiones que se solicite al ejecutar el análisis, dado que las soluciones están anidadas, las n primeras dimensiones siempre son iguales en un nuevo estudio con n+k dimensiones. La mejor opción sería pedir un gran número de dimensiones y cortar a partir de aquella que no muestre una información relevante. La distancia que se obtiene entre los puntos esta relacionada con la similaridad en sus patrones de respuesta. Un punto con baja frecuencia marginal aparece mas hacia la periferia, y los autores más significativos aparecerán más hacia el centro. Esto quedara expuesto más claramente a través del ejemplo mostrado.

La segunda técnica utilizada es una técnica de clasificación o análisis de conglomerados (Cluster Analysis). Estas técnicas intentan clasificar a partir de distintos niveles jerárquicos, de tal forma que los elementos de una clase deben ser razonablemente homogéneos. Es decir, va agrupando progresivamente los autores en función del número de firmas conjuntas obteniendo una estructura arborescente que ofrece una serie de ventajas complementarias a la reducción de la dimensionalidad, ya que permite diversos niveles de agrupamiento, y su forma de representación gráfica, el dendograma, es más comprensiva que en dimensiones. De la misma manera la elección de la técnica concreta a utilizar viene determinada por el tipo de datos, matriz de tipo "matching" con 0-1 .Por otra parte las variables podemos considerarlas homogéneas ya que usan la misma métrica, y además la información relevante aparece más en la presencia que en la ausencia, ya que las ausencias dependen de la amplitud de la muestra. Teniendo en cuenta estas características el índice más adecuado para utilizar es el Jaccard que no considera ausencias conjuntas, y es apto, por tanto, para matrices de tipo "matching".

$$similaridad = \frac{a}{a+b+c}$$

donde,

	presencia	*ausencia*
presencia	*a*	*b*
ausencia	*c*	*d*

o lo que es lo mismo

$$similaridad = \frac{intersección\ de\ firmas}{union\ de\ firmas}$$

Una vez obtenido el índice de similaridad se debe escoger el algoritmo de agrupación más indicado. Existen diversas formas de agrupar: 1) Aglomerativa/Divisiva. En la primera de ellas se parte de elementos sueltos que se van combinado en grupos, mientras en la otra se parte de un grupo que se va partiendo. 2) Jerárquica/No-Jerárquica. En la primera se van realizando agrupaciones con los distintos niveles, mientras que en la no jerárquica no se establece relación entre los grupos. Dada las características de nuestra información es evidente el uso de técnicas aglomerativas y jerárquicas. Esto llevaría a considerar aquellas que parten de agrupamientos en función de distancias promedios. De entre ellas seleccionamos el método basado en el promedio entre grupos UPGMA (unweighted pair-group method using arithmetic averages). Este método tiene la ventaja de que no toma en cuenta la distancias existentes entre los miembros internos del grupo. Esto evita el sesgo que se produciría al estimar la similaridad de los grupos teniendo en cuenta la similaridad que ya existe entre sus elementos internos. Además los datos de Milligan (1990) muestran al UPGMA como uno de los procedimientos más robustos a la hora de soportar distorsiones en los datos.

Se han aplicado ambos modelos a conjunto de artículos de la misma temática, obtenidos a través de Psiclit. Se aislaron los autores, y a través del programa Escalacitas (en elaboración por Valero) se consiguió la matriz de datos original en la que aparecen los autores de cada articulo en filas, y el total de autores en columnas, codificándose la coincidencia con un 1. Una vez obtenida la matriz de datos original se analizó el programa estadístico SPSS.

Comandos para HOMALS:

```
RECODE ALL (0=2).

HOMALS VARIABLES ALL(2) /ANALYSIS ALL /DIMENSION 25
/PLOT NDIM (1,MAX) all(8).
```

Comandos para el CLUSTER:

```
PROXIMITIES ALL /VIEW VARIABLE/ MEASURE JACCARD /MATRIX
OUT(*).

CLUSTER /MATRIX IN(*) METHOD BAVERAGE/PLOT DENDROGRAM
/PRINT NONE.
```

En Homals se obtuvo 25 dimensiones, todos ellos como se puede ver (tabla 1) con valores específicos muy bajos. Esto se debe a que no aparecen grandes grupos muy cohesionados, sino mas bien pequeños grupos dispersos.

TABLA 1

DIMENSION	EIGENVALUE
1	.0147
2	.0139
3	.0128
4	.0120
5	.0114
6	.0111
7	.0111
8	.0110
9	.0107

10	.0107
11	.0106
12	.0105
13	.0105
14	.0105
15	.0104
16	.0103
17	.0103
18	.0102
19	.0101
20	.0101
21	.0098
22	.0094
23	.0092
24	.0090
25	.0089

A continuación se presenta un cuadro resumido con los pasos de aglomeración que va realizando el CLUSTER (tabla 2). En la primera columna se ofrece el número de paso. En la segunda y tercera los números de los objetos (autores) que va aglomerando en cada paso. En la cuarta columna presenta el promedio de similaridad que ofrece esa aglomeración. Este cuadro solo recoge la parte significativa, o sea, hasta que este promedio llega a 0, lo que indica que ya no puede aglomerar más.

TABLA 2

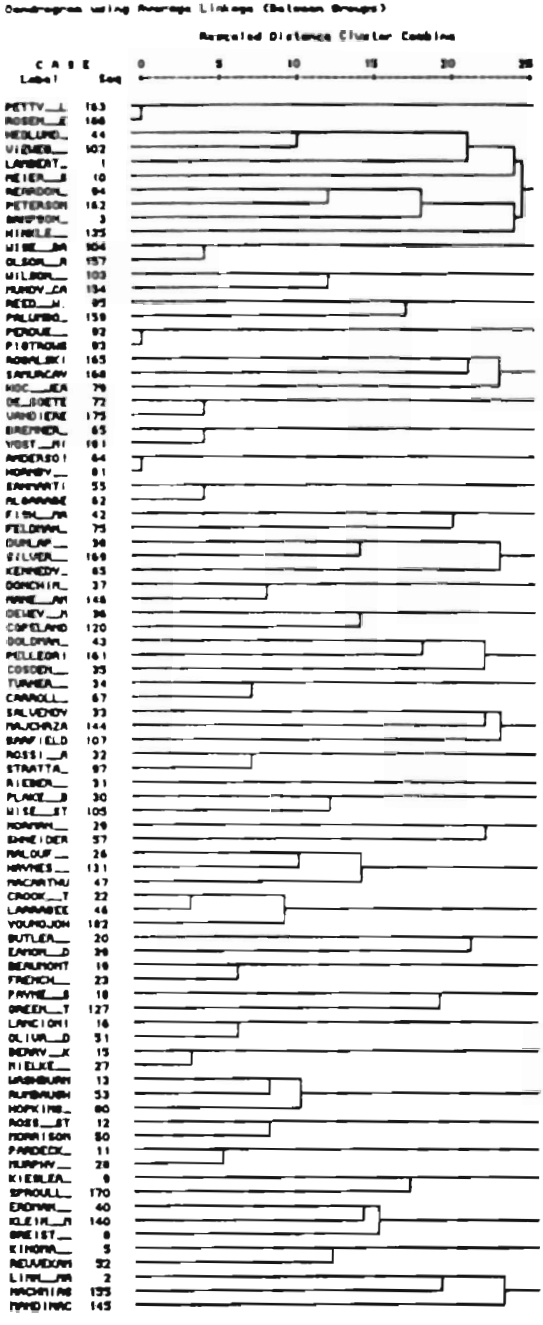
Agglomeration Schedule using Average Linkage (Between Groups)

	Clusters Combined			Stage Cluster	1st Appears	Next
Stage	Cluster 1	Cluster 2	Coefficient	Cluster 1	Cluster 2	Stage
1	163	166	1.000000	0	0	62
2	92	93	1.000000	0	0	113
3	64	81	1.000000	0	0	136
4	15	27	.875000	0	0	170
5	22	46	.857143	0	0	18
6	55	62	.833333	0	0	143
7	65	181	.800000	0	0	135
8	72	175	.800000	0	0	129
9	104	157	.800000	0	0	105
10	11	28	.777778	0	0	174
11	16	51	.750000	0	0	169
12	19	23	.750000	0	0	166
13	32	97	.714286	0	0	156
14	34	67	.714286	0	0	154
15	37	146	.666667	0	0	151
16	13	53	.666667	0	0	21
17	12	50	.666667	0	0	173
18	22	182	.619048	5	0	163
19	26	131	.571429	0	0	26
20	44	102	.571429	0	0	38
21	13	80	.563492	16	0	172
22	94	162	.500000	0	0	34
23	103	154	.500000	0	0	106
24	30	105	.500000	0	0	158
25	5	52	.500000	0	0	179
26	26	47	.436508	19	0	160
27	38	169	.428571	0	0	47

28	40	140	.428571	0	0	30
29	36	120	.428571	0	0	152
30	8	40	.366667	0	28	176
31	9	170	.300000	0	0	175
32	95	159	.285714	0	0	112
33	43	161	.250000	0	0	42
34	3	94	.250000	0	22	49
35	2	155	.214286	0	0	44
36	18	127	.200000	0	0	167
37	42	75	.166667	0	0	148
38	1	44	.158824	0	20	48
39	20	39	.153846	0	0	165
40	165	168	.142857	0	0	45
41	33	144	.100000	0	0	46
42	35	43	.100000	0	33	153
43	29	57	.083333	0	0	159
44	2	145	.066667	35	0	181
45	79	165	.062500	0	40	123
46	33	107	.050000	41	0	155
47	38	85	.050000	27	0	150
48	1	10	.033333	38	0	50
49	3	135	.020833	34	0	50
50	1	3	.002976	48	49	52

A continuación se presenta el dendograma (tabla3) reducido (tabla3) donde aparecen los autores que se ligan con alguien a los distintos niveles posibles.

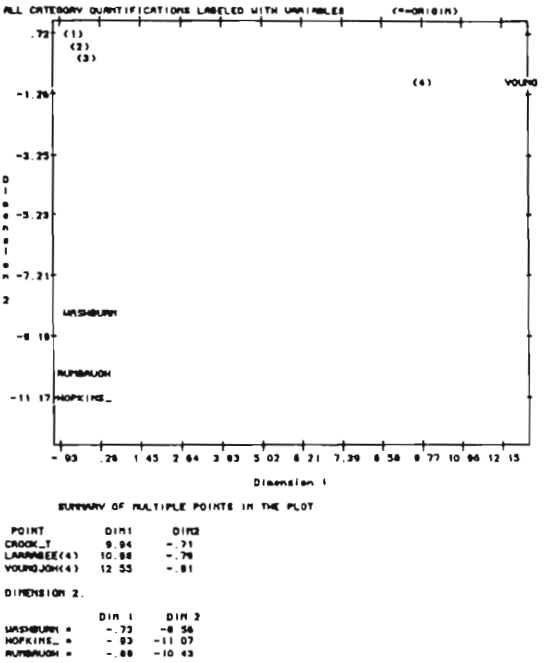
TABLA3


Dendrogram using Average Linkage (Between Groups)
Rescaled Distance Cluster Combine
CASE
Label Seq
0 5 10 15 20 25
PETTY__L 183
ROSEN__E 166
HEDLUND_ 44
VIEWEG__ 102
LAMBERT_ 1
MEIER__E 10
REARDON_ 94
PETERSON 162
BRISTOW_ 3
KIMBLE__ 135
WIBE__BA 104
OLSON__R 157
MILBOR__ 103
MUNOV_CA 134
REED___H 95
PALUMBO_ 159
PERDUE__ 92
PIOTROWS 93
NOWALSKI 165
SAMUNCAN 160
NOC___KA 79
DE_BOOTE 72
VANDIENE 175
BAERMEN_ 65
YOST___RI 101
APGERSDI 64
MORPHY__ 81
LAKKARTI 55
ALBARRAN 82
FISH__RA 42
FIELDMAL 75
DURLAP__ 30
SILVER__ 169
KENNEDY_ 85
DONCHIN_ 37
MANE___AN 148
DEWEY__A 98
COPELAND 120
GOLDMAN_ 43
MELLEGRI 161
COSDEN__ 35
TURNER__ 34
CARROLL_ 67
SALVENOV 33
MALCHAZA 144
BARFIELD 107
ROSSI__A 32
STRATTA_ 97
AIENDA__ 91
PLACE__B 30
WISE___ST 105
MORMAN__ 29
BAMEIDER 57
MALOUF__ 26
HAYNES__ 131
MACARTHU 47
CROOK__T 22
LARRABEE 46
YOUNOJON 102
BUTLER__ 20
EARON__D 39
BEAUMONT 19
FRENCH__ 23
PAYNE__B 18
GREEN__T 127
LANCIONI 16
OLIVA__D 31
BERRY__K 15
NIELSE__ 27
WASHBURN 13
RUMBAUGH 53
HOPKINS_ 80
ROSS___ST 12
MORRISON 50
PARDECK_ 11
MURPHY__ 28
KIERLEA_ 9
SPROULL_ 170
ERDMAN__ 40
KLEIN__A 140
BREIST__ 8
KINDRA__ 5
REVAMONT 52
LIPA___RA 2
MACHRIAS 135
RAMDINAC 145

Para la comparación de los resultados utilizaremos las dimensiones obtenidas por HOMALS, con la porción del CLUSTER en que este representada esa dimensión. Por ultimo, para obtener un criterio objetivo de comparación se han utilizado las tabulaciones cruzadas (crosstab) de las combinaciones de autores obtenidas con el fin de decidir cual de ellas es mas adecuada.

Empezaremos con la dimensión 1 de Homals:

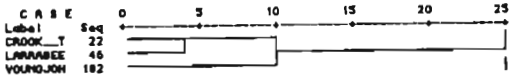
HOMALS: DIMENSION 1 Y 2



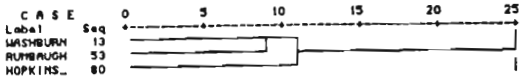
Si observamos la gráfica vemos como en la dimensión 1 aparecen tres autores alejados del 0 en coordenadas, por lo que estos tres autores forman un grupo de colaboración científica. Si vemos la parte del cluster (grupo1) donde aparecen estos autores observamos que se han agrupado de la misma forma. Crook y Larrabee aparecen muy juntos en el Homals, Youngh, sin embargo, puntúa mas, y por lo tanto en el Cluster se une con posterioridad. La misma interpretación podríamos hacer de la Dimensión 2 y el grupo 2 del Cluster.

CLUSTER:

GRUPO1

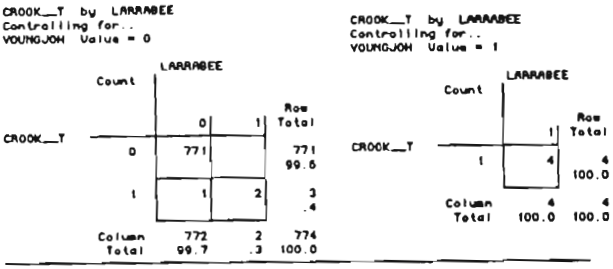


GRUPO2:



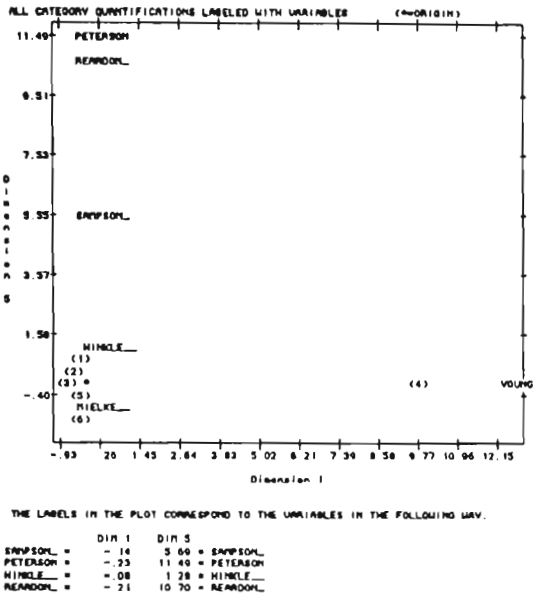
Estas dos agrupaciones de autores aparecen de forma muy clara con ambos tipos técnicas, cosa que nos sucederá en todos los casos, por lo tanto se hizo necesario el ejecutar las tabulaciones cruzadas de los autores implicados con el fin de obtener un criterio objetivo. A continuación mostramos un ejemplo de crosstab realizado con los autores obtenidos en el primer grupo;

CROSSTAB: GRUPO 1.

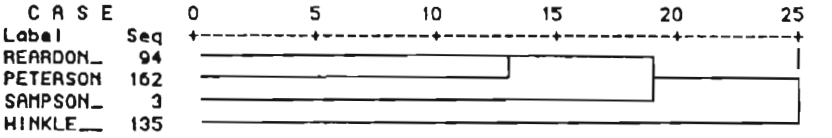


Si miramos el Crosstab entre, los autores, vemos que Youngjoh, ha publicado todo con el grupo, por lo que en Homals aparece en una posición mas extrema, y Crook_T aparece con menor puntuación al tener algo solos, ademas del publicado con Larrabee, estos por tanto tienen un articulo juntos, por lo que han colaborado en 8 ocasiones, esto hace que el cluster los una antes que ha Youngh, que solo ha colaborado con ellos en 7 ocasiones. Repetir esta mecánica entre las distintas agrupaciones de autores resultaría tedioso, por lo que nos limitaremos a comentar las dimensiones que muestren algún nuevo dato de interés.

DIMENSION 5:



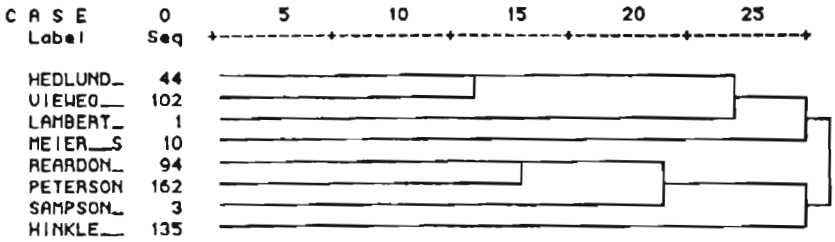
CLUSTER: GRUPO 5



Comentamos este caso ya que al revisar la matriz de cofirmación vemos que junto a Sampson firma otro autor, Meier, que no aparece en el gráfico del dendograma. Este será agrupado mas tarde, en el cluster formado por Lambert, Vieweg, y Hedlund. Repasando el Schedule del Cluster vemos que ambos grupos son unidos al final, pero esto no es apreciable en la impresión, ya que su coeficiente es de 0,002976, este valor representa el promedio de las similaridades entre ambos grupos, según el procedimiento de aglomeración UPGMA. Este grupo también fue comprobado a través de los crosstab.

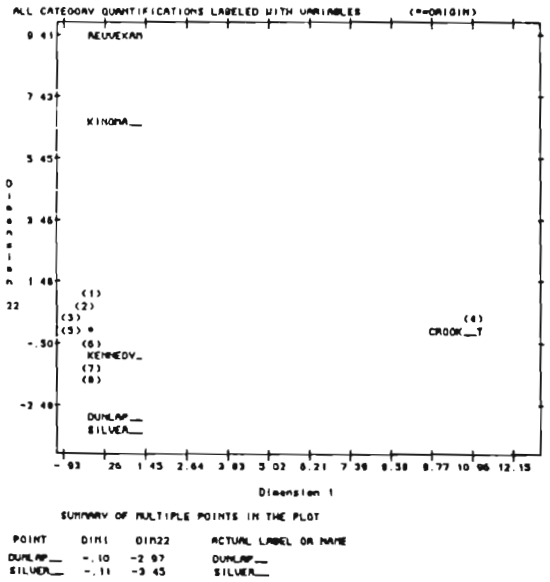
El grupo 5 quedaría de la siguiente forma:

CLUSTER: GRUPO 5:

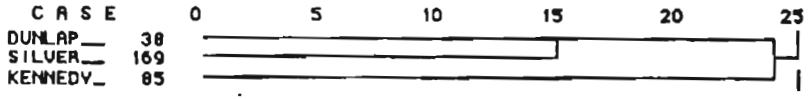


Los autores que no aparecen en el Homals fueron agrupados independientemente en la Dimensión numero 9 a excepción de Meirs, que como se ve en el Cluster hizo la función de puente entre ambos grupos. Se ha revisado la matriz de discriminación, utilizada por Homals, de autores en la variable para Meirs, con el fin de discriminar si el error se debía a un fallo gráfico, el valor encontrado fue de 0.003, con lo cual no podemos concluir que Homals detectara la presencia del autor en el grupo, ya que su peso es irrelevante.

DIMENSION 22:



CLUSTER: GRUPO 21:



Como vemos el dendograma recoge un autor mas que la gráfica del Homals, mirando el crosstab, vemos que tiene varios trabajos con los miembros de este grupo. Por otra parte su valor en la tabla de discriminación del Homals es de 0.007, por lo que no parece significativo. Este autor no presenta ninguna otra característica especial, ya que no aparece con ningún otro autor.

CROSSTAB: GRUPO 21

KENNEDY_ by SILVER_
Controlling for..
DUMLAP_ Value = 0

KENNEDY_	SILVER_ 0	SILVER_ 1	Row Total
0	767	1	768 99.5
1	4		4 .5
Column Total	771 99.9	1 1	772 100.0

KENNEDY_ by SILVER_
Controlling for..
DUMLAP_ Value = 1

KENNEDY_	SILVER_ 0	SILVER_ 1	Row Total
0	2	3	5 83.3
1	1		1 16.7
Column Total	3 50.0	3 50.0	6 100.0

Aparte de estas agrupaciones, el Cluster detecto 11 grupos mas, que no aparecían dentro de ninguna dimensión del Homals, estos nuevos grupos fueron corroborados a través del Crosstab.

CONCLUSIONES:

Una vez visto todos los resultados parece manifestarse una mejor discriminación de los grupos a través del cluster. Este detecto 34 grupos distintos frente a 25 del Homals, y sin contar los dos incompletos de este ultimo. Un primer aspecto que queda patente es que este tipo de CLUSTER, por su forma de operar, es exhaustivo y agrupa a cualquier autor que haya colaborado con otro, aunque sea una sola vez. Naturalmente las agrupaciones se manifiestan a distintos niveles según al proximidad de los miembros del grupo. En este sentido, al HOMALS se le han escapado bastantes relaciones. Otra ventaja que encontramos en el Cluster es su fácil y rápida visualización de los resultados, en Homals, por el contrario, los resultados son bastante difíciles de leer en muchas ocasiones, ya que al tener que representar múltiples distancias entre autores en una dimensión produce una gran distorsión que puede conducir a ciertos errores.

Lo que ya no queda tan claro con ninguna de estas técnicas es la detección de posibles líderes. Por un lado, en el CLUSTER parece que la tendencia del supuesto líder es a situarse en una agrupación tardía (el líder se agrega al grupo en último lugar, o sea en la línea de unión que se une más a la derecha en el dendograma). Pero no siempre es así, puesto que también le ocurre lo mismo, en algunos casos, a autores que han publicado muy poco con el grupo o con éste y otro grupo simultáneamente.

En el HOMALS ocurre algo similar, y a predicho en parte cuando se explicó el funcionamiento de esta técnica. En este caso el líder tiende hacia el menor valor en la dimensión, acercándose a los que no pertenecerían al grupo. Pero este mismo fenómeno le ocurre a algunos autores con escasas relaciones con este grupo.

Por otro lado también hay que señalar los mayores requerimientos del HOMALS en términos de tiempo, cantidad de memoria y tamaño de output frente al CLUSTER.

En Homals encontramos otro problema más, como fue el no haber encontrado autores que trabajaban dentro de un grupo, como son los casos de Kennedy y Meier, y especialmente en este ultimo que servía de nexo entre dos conglomerados de autores.

Los resultados obtenidos nos hacen inclinarnos por el Análisis Cluster como la técnica mas adecuada de las estudiadas. En primer lugar es sencilla y rápida de ejecutar, sus resultados son claros y fácilmente interpretables, y presenta una gran fiabilidad.

Este tipo de estudio no ha de limitarse a los autores, estas técnicas serian igualmente útiles para el análisis de otras características bibliográficas, como son los descriptores, análisis de la cocitación entre autores, que como hemos visto es también un índice que nos indica la colaboración entre los miembros de la comunidad científica.

REFERENCIAS BIBLIOGRÁFICAS

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley and Sons.

Gifi, A. (1990). *Nonlinear Multivariate*. Analysis. Wiley.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychoometrika, 45,* 325-342

Price, D. J. S. y Beaver, D. (1966). Colloboration in an invisible college. *American Psychologist, 21,* 1011.

SPSS Inc. (1990) *SPSS Categories*..

Valero P.; Molina J. G. (1993). Bibcount 1.1 Programa informático al analisis de conjuntos de referencias bibliográficas. *III Simposium de metodología de las ciencias sociales y del comportamiento*. Santiago de Compostela.

Valero P.; Molina J. G. y Sanmartin, J. (1992). Un grupo de herramientas informaticas para el análisis de conjuntos de referencias bibliograficas. *Revista de Historia de Psicología. Vol. 13.* Nº 1.

Valero P.; Molina J. G. y Sanmartin, J. (1993). Un programa inforático para el análisis de conjuntos de referencias bibliograficas (En prensa).