Revista Investigación Educativa - Vol. 14, n.º 1, págs. 115-128

REDTAB: PROGRAMA DE REDUCCIÓN DE GRANDES TABLAS DE CONTINGENCIA BIDIMENSIONALES

por

Yosu Yurramendi Mendizabal; Luis Joaristi Olariaga y Luis Lizasoain Hernández

Universidad del País Vasco–Euskal Herriko Unibertsitatea[1]

SUMARIO

Se trata de examinar la interpretación de una tabla de contingencia bidimensional de grandes dimensiones. Nuestro propósito en el presente artículo consiste en sintetizar la información de la tabla inicial en una más reducida. Para ello, se agregan las filas y las columnas pertenecientes a las mismas clases. Considerando los diferentes niveles de una jerarquía, una jerarquía de clases de un conjunto define un conjunto de particiones. Como ejemplo, se presentan los pasos del proceso de operación del programa REDTAB así como las tablas de resultados.

ABSTRACT

Given a large scale two–way contingency table, the problem of its interpretation is examined. In this paper, an interpretation consists of a reduced table which summarizes the information contained in the initial table. Each set partition allows one to reduce the initial table, by

---

1 *Dirección de los autores:* Yosu Yurramendi, Dpto. de Ciencias de la Computación e Inteligencia Artificial. E–mail: ccpyumej@sisb00.si.ehu.es.

Luis Joaristi, Luis Lizasoain, Dpto. de Métodos de Investigación y Diagnóstico en Educación. E–mail: plpjooll@sf.ehu.es, plplihel@sf.ehu.es (respectivamente).

Este trabajo se ha realizado en el contexto del proyecto de investigación de la UPV–EHU 140226–TA 193/92.

adding the rows or columns relative to the elements which belong to the same class. A class hierarchy of a set defines a set of partitions, by considering the different levels of the hierarchy. As an example, the steps in the REDTAB program running process are presented with the tables of the output file.

1. INTRODUCCIÓN

En este artículo se presenta un programa estadístico que tiene como objetivo reducir el tamaño de grandes tablas de contingencia bidimensionales con el fin de facilitar su interpretación.

La *reducción* que proponemos se lleva a cabo mediante la *agrupación* de *filas* y *columnas* de dicha tabla, y debe realizarse teniendo en cuenta la totalidad de la información contenida en la tabla inicial.

Esta afirmación implica que se debe considerar y evaluar el equilibrio entre la pérdida de la información inherente a la reducción del tamaño de la tabla y la ganancia de interpretabilidad asociada. Tal balance debe efectuarse en cada etapa del proceso, siendo éste el criterio fundamental a la hora de decidir si se prosigue o no con el proceso de reducción.

Este tipo de problema —inicialmente formulado por Fisher (1969)— ha sido estudiado por varios autores: Hartigan (1975) y Bertin (1977) tratan de reorganizar la tabla inicial mediante la adecuada permutación de filas y columnas; Anderberg (1973) apunta un enfoque iterativo en el cual el proceso de agrupamiento se desarrolla alternativamente sobre las filas y las columnas hasta que los grupos finales aparecen como mutuamente armónicos; para ello Govaert (1983, 1984) ha desarrollado un algoritmo.

Greenacre (1988), partiendo de dos clasificaciones jerárquicas, una sobre las filas y otra sobre las columnas, establece las podas de ambos árboles de forma independiente y con un criterio probabilístico basado en las relaciones entre los valores propios del análisis factorial de correspondencias de la tabla inicial y los índices de las clasificaciones jerárquicas usando el criterio de la inercia. Una poda de un árbol define una partición del conjunto correspondiente.

Nuestro planteamiento está también basado en *dos clasificaciones jerárquicas*, pero en vez de considerar un criterio probabilístico usamos un *índice cruzado* (Yurramendi, 1984) definido sobre los pares de clases correspondientes a ambas jerarquías. Además no utilizamos jerarquías *indexadas* y la poda de los árboles se conforma *coordinamente*.

Una clase dada de jerarquías es unida a una partición del conjunto; y un *par de particiones*, una de cada jerarquía, define una *tabla de contingencia reducida*. El problema consiste en seleccionar un par adecuado de particiones.

2. FUNDAMENTACIÓN TEÓRICA

Sean I y J dos conjuntos finitos no vacíos. Se define una tabla de datos sobre I y J tal que a cada par de elementos $(i,j) \in IxJ$ corresponde un número real n_{ij}. Sean $n_i = \sum_{j \in J} n_{ij}$ y $n_j = \sum_{i \in I} n_{ij}$ las marginales de la tabla.

Sean *HI* y *HJ* dos jerarquías definidas sobre los conjuntos *I* y *J* respectivamente.

Para todo par de clases no simples $(PI,PJ) \in HI \times HJ$, se define el índice cruzado como:

$$v(PI,PJ)=\Delta(PI,PJ)-\sum_{PJ' \in suc(PJ)}\Delta(PI,PJ')-\sum_{PI' \in suc(PI)}\Delta(PI',PJ)+\sum_{PI' \in suc(PI)}\sum_{PJ' \in suc(PJ)}\Delta(PI',PJ')$$

siendo *suc(PI)* el conjunto de clases inmediatamente sucesor de *PI* en *HI*, análogamente *suc(PJ)*, y *Δ(PI,PJ)* una función relacionada con lo que se puede considerar una *cantidad de información*. Por ejemplo,

$$\Delta(PI,PJ)=(n\, n_{PI,PJ}-n_{PI}n_{PJ})^2 / n_{PI}n_{PJ}$$

donde $n_{PI,PJ}=\sum_{i \in PI}\sum_{j \in PJ} n_{ij}$, $n_{PI}=\sum_{i \in PI} n_{i'}$, $n_{PJ}=\sum_{j \in PJ} n_j$ and $n=\sum_{i \in PI}\sum_{j \in PJ} n_{ij}$

Se cumple la siguiente propiedad del índice así definido:

$$\sum_{PI \in HI}\sum_{PJ \in HJ} v(PI,PJ = \sum_{i \in I}\sum_{j \in J}\Delta(i,j)$$

En el ejemplo propuesto se obtiene una descomposición del estadístico χ^2 en términos de los valores del índice cruzado; tal estadístico está relacionado con la medida de la entropía de la teoría de la información y puede considerarse como una cantidad de información (Benzécri, 1973).

Las *fórmulas de descomposición* son la base del criterio que permite comparar la ganancia de interpretabilidad y la pérdida de información originada por la reducción de la tabla inicial. Evidentemente, cuanto menor es la dimensión de la tabla, menor es la información considerada y mayor es la interpretabilidad.

3. ETAPAS DEL PROCESO

Tras haber definido los conceptos teóricos básicos, se explican a continuación los pasos del proceso de operación del programa.

Los datos de entrada son la tabla de contingencia y las dos jerarquías.

Con tales datos, el programa calcula los valores del índice cruzado, mostrándolos bajo un formato de tabla en que las filas son las clases no simples de *HI* y las columnas las de *HJ*.

A continuación, el usuario analiza estos valores, que representan porciones de la cantidad de información global (fórmulas de descomposición). Así, se rechazarían los más bajos ya que implican una pequeña pérdida de información; sin embargo, sí se mantienen los más elevados.

Las clases de *I* y *J* correspondientes a los valores elevados se utilizan en la construcción de la nueva tabla reducida. Las sucesoras inmediatas de estas clases constituyen las filas y columnas de la nueva tabla.

El sistema muestra junto a esta tabla reducida el índice de cantidad de la información asociado (por ejemplo, χ^2), así como el porcentaje sobre el valor inicial.

Finalmente se compara la cantidad de pérdida de información con la ganancia de interpretabilidad inherente a la reducción de la tabla.

Tras examinar los resultados, el programa ofrece la posibilidad de generar nuevas tablas reducidas. El conjunto de las tablas reducidas seleccionadas puede considerarse como un «efecto zoom» aplicado a la amplia tabla inicial. Así, el programa que presentamos permite explorar las porciones de información más relevantes.

4. APLICACIÓN

Basado en el índice cuyas características y propiedades acaban de ser descritas, se ha desarrollado el programa REDTAB.

En este momento está disponible un primer prototipo cuyas principales características y procedimientos pasamos a describir.

El sistema requiere de cuatro ficheros de entrada:

1. Fichero de Parámetros:

Contiene los parámetros fundamentales como, entre otros, son los siguientes:
— número de elementos del conjunto *I*
— número de elementos del conjunto *J*
— número de nodos seleccionados en el árbol *I*
— número de nodos seleccionados en el árbol *J*

2. Fichero de Datos

Contiene la tabla de contingencia inicial con los datos originales.

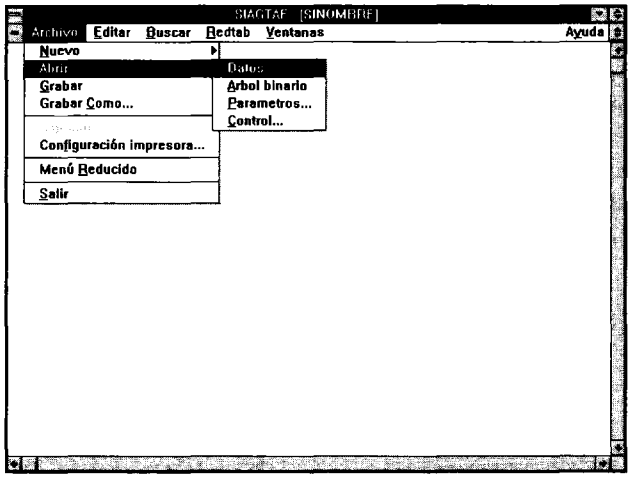
3. Ficheros del árbol *I* y del árbol *J*

Contiene los nodos del árbol binario sobre el conjunto *I* y sobre el conjunto *J*.

4. Fichero de Control:

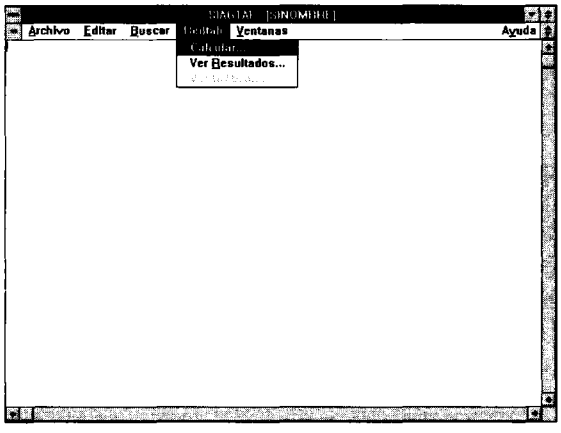
Es el encargado de controlar la ejecución del programa mediante una correcta administración de los ficheros.

Una vez que el programa ha sido arrancado, la ventana principal es la siguiente. Si abrimos el menú de archivo, las opciones son las habituales. Por ejemplo, aquí se ha elegido la opción de **Abrir** y, como puede verse, permite la apertura de los cuatro tipos de ficheros antes citados.



Las opciones relativas a **Editar**, **Buscar** y **Ventanas** ofrecen los procedimientos habituales en este tipo de aplicaciones.

Así que ahora vamos a centrarnos en la opción de REDTAB que agrupa las funciones de ejecución del programa y de acceso a los ficheros de resultados y a los ficheros de gráficos.



Los datos de este ejemplo provienen de un estudio publicado en *Les Cahiers de l'analyse des données* (nº 1, 1985, pp. 53-74) por Adamès, G. Se trata de la explotación del banco de datos estadísticos de la UNESCO, y más en concreto las cifras relativas a las tasas de escolarización de 97 Estados miembros de los que se disponía de información suficientemente fiable.

La matriz original de datos es una tabla de contingencia bidimensional en la que aparecen dos conjuntos: los Estados miembros y los tantos por mil sobre la población de estudiantes escolarizados según sexos y edades.

Esta última variable ha sido tratada según el siguiente esquema:

H01. Estudiantes de sexo masculino (H) menores de 7 años.
M01. Estudiantes de sexo femenino (M) menores de 7 años.
H07–H17. Estudiantes de sexo masculino (H) de la edad correspondiente al valor numérico.
M07–M17. Estudiantes de sexo femenino (M) de la edad correspondiente al valor numérico.
H02. Estudiantes de sexo masculino (H) mayores de 17 años.
M02. Estudiantes de sexo femenino (M) mayores de 17 años.

TABLA I
TABLA DE CONTINGENCIA DE ESTADOS POR EDAD Y SEXO

	H01	H07	H08	.....
ALG (Argelia)	87	81	77	
ARG (Argentina)	53	53	53	
AUL (Australia)	61	48	47	
BAH (Bahrein)	35	47	50	
.				
.				
.				
.				
.				
ZAM (Zambia)				

Esta matriz configura el fichero de entrada UNESCO.DAT.

Una vez que sobre esta matriz se realizan sendas clasificaciones jerárquicas (sobre los Estados y sobre las edades), se dispone también de la descripción de los nodos de las clasificaciones tanto para *I* como para *J*. Estos son los ficheros UNESCO.FIL y UNESCO.COL, respectivamente.

TABLA II

Clasificación jerárquica: nodos de la clasificación sobre los Estados (conjunto I). Está obtenida a través del procedimiento RECIP de SPAD.N. Se ha suprimido el histograma de los índices de nivel, proporcionado también por el mencionado procedimiento, por cuestiones relativas a espacio. Por la misma razón sólo se presentan los 19 últimos nodos.

NÚM	AINE	BENJ	EFF	POIDS	ÍNDICE
175	164	69	9	9012.00	.00085
176	149	163	7	7015.00	.00085
177	158	142	6	6003.00	.00085
178	166	57	7	7008.00	.00090
179	168	157	7	7010.00	.00103
180	160	170	14	14019.00	.00135
181	176	178	14	14023.00	.00138
182	175	167	13	13019.00	.00145
183	165	173	7	7013.00	.00152
184	183	177	13	13016.00	.00195
185	171	159	19	19030.00	.00203
186	174	180	21	21102.00	.00218
187	179	172	13	13014.00	.00350
188	162	186	25	25109.00	.00378
189	181	185	33	33053.00	.00396
190	187	188	38	38123.00	.00888
191	184	189	46	46069.00	.00894
192	182	190	51	51142.00	.01597
193	191	192	97	97211.00	.02833

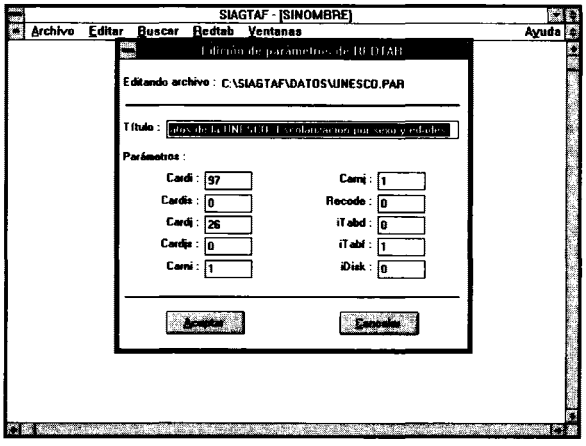
SOMME DES INDICES DE NIVEAU = .11036

TABLA III

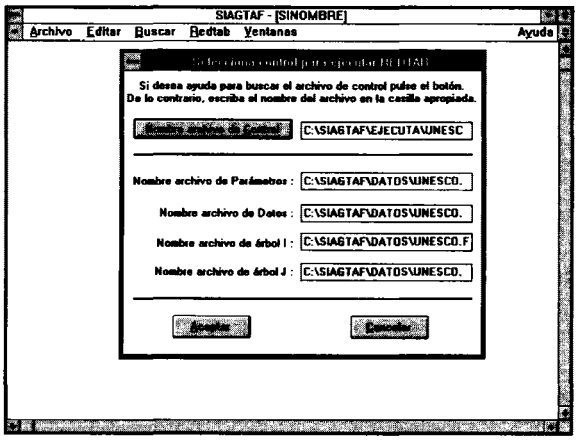
Clasificación jerárquica: nodos de la clasificación sobre las edades (conjunto J). Está obtenida a través del procedimiento RECIP de SPAD.N. Por razones de espacio se suprimen el histograma de los índice de nivel así como los 16 primeros nodos.

NÚM	AINE	BENJ	EFF	POIDS	ÍNDICE
43	13	26	2	2972.00	.00181
44	37	34	4	9423.00	.00223
45	42	32	7	38008.00	.00297
46	44	38	7	13701.00	.00473
47	46	39	10	23334.00	.00539
48	45	40	12	61882.00	.00868
49	47	43	12	26306.00	.00936
50	49	48	24	88188.00	.03104
51	50	41	26	97211.00	.03389

El cuarto fichero es el de parámetros (UNESCO.PAR).

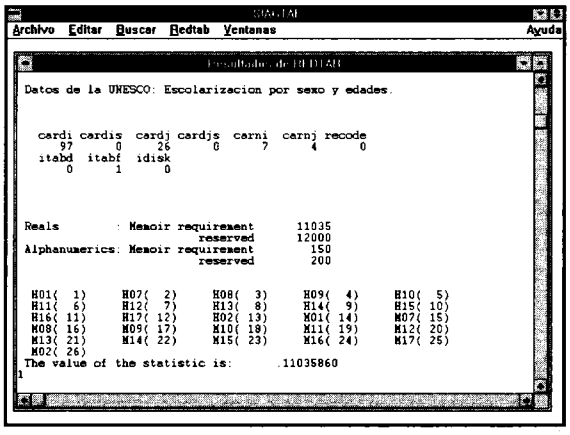


Por último, se debe disponer de un fichero de control que asegure una adecuada administración de los ficheros. En esta figura vemos el contenido de la ventana relativa al fichero UNESCO.CTL.



Una vez que el usuario ya dispone de todos los ficheros de entrada, debe elegirse la opción de **Calcular** para ejecutar el programa. Como resultado de tal elección se abre un cuadro de diálogo en el que el sistema, en primer lugar, solicita al usuario el nombre del fichero de control. Una vez que tal nombre ha sido especificado, el programa completa el resto de los cuadros con los nombres de los ficheros correspondientes, y al pulsar el botón de **Aceptar** el programa inicia el proceso de reducción.

Una vez que el proceso ha finalizado, basta con elegir la opción de **Ver resultados**, para tener acceso al fichero de resultados (output file).



A continuación vamos a examinar con más detalle los diferentes resultados obtenidos. Y estos consisten en las siguientes tablas:

— tabla de valores del índice cruzado (sobre 1.000)
— tabla reducida
— tabla de perfiles de las filas
— tabla de perfiles de las columnas
— la descripción de las clases para cada partición realizada en *I* y en *J*.

Veamos ahora las tablas obtenidas con los datos de nuestro ejemplo. En una primera pasada del programa, éste generaría una tabla inicial del índice cruzado para *todos* los pares de nodos de ambas clasificaciones:

TABLA IV
TABLA DEL ÍNDICE CRUZADO

n_i / n_j	1	2	3	....	....	23	24	25
1								
2								
3								
4								
.								
.								
.								
.								
94								
95								
96								

20

10

Valor inicial del índice cruzado: 0,110358
(100%)

En este caso, esta primera tabla constaría de 96 clases en *I* y de 25 en *J*. Evidentemente, tratar de interpretar una tabla de estas dimensiones (independientemente de la reducción que suponga con respecto a la original) resulta una tarea difícil. Se impone en consecuencia una reducción inicial de la tabla que no comporte una excesiva pérdida de información.

En este momento, el usuario, selecciona —por ejemplo— la porción sombreada que tiene unas dimensiones de 20x10 y en la que presumiblemente van a parecer los mayores valores del índice. Con estos nuevos parámetros, se vuelve a ejecutar el programa, y en este caso la tabla de los valores del índice cruzado es la que sigue.

TABLA V

TABLA DE VALORES DEL ÍNDICE CRUZADO (SOBRE 1.000)

nI– –nJ	43	44	45	46	47	48	49	50	51	
175	0.	0.	1.	0.	0.	0.	7.	0.	1.	9.
176	0.	0.	0.	0.	0.	2.	1.	5.	1.	9.
177	1.	2.	1.	0.	0.	0.	0.	2.	2.	9.
178	0.	0.	1.	5.	0.	1.	1.	0.	1.	8.
179	0.	0.	0.	0.	0.	0.	2.	2.	5.	11.
180	0.	0.	0.	1.	2.	5.	1.	1.	4.	15.
181	0.	2.	0.	1.	7.	1.	0.	2.	3.	15.
182	0.	1.	1.	0.	1.	0.	4.	8.	0.	17.
183	4.	0.	1.	4.	0.	1.	3.	0.	2.	15.
184	0.	0.	0.	0.	0.	0.	10.	0.	10.	21.
185	2.	0.	0.	5.	2.	8.	2.	0.	1.	22.
186	0.	0.	0.	0.	1.	3.	0.	1.	20.	25.
187	0.	0.	0.	0.	0.	**1.**	**0.**	**1.**	**32.**	36.
188	0.	0.	0.	3.	0.	**0.**	**6.**	**10.**	**20.**	40.
189	0.	1.	2.	3.	4.	**2.**	**31.**	**1.**	**0.**	44.
190	1.	1.	0.	6.	2.	**49**	**0.**	**8.**	**31.**	99.
191	2.	0.	2.	1.	3.	**0.**	**11.**	**41.**	**43.**	103.
192	0.	0.	1.	0.	6.	**11.**	**1.**	**1.**	**167.**	187.
193	2.	10.	12.	15.	14.	**3.**	**5.**	**247.**	**10.**	317.
	13.	18.	22.	44.	44.	88.	87.	332.	352.	

Valor del índice cruzado: 0,0827685
(75,9137%)

Si además examinamos la tabla reducida, vemos como, en este caso, esta primera reducción del tamaño de la tabla de contingencia sólo ha implicado una pérdida del valor inicial del estadístico del 24%.

El examen de esta tabla proporciona al usuario el criterio fundamental a la hora de

decidir sobre el proceso. La disminución del valor del estadístico debe interpretarse como un indicador de la pérdida de información producida por la reducción del tamaño.

La decisión que en cada momento debe tomar el usuario es valorar hasta qué punto esta pérdida de información es compensada con la ganancia obtenida en la interpretabilidad de la tabla. Y para ello el programa proporciona el resto de las tablas, y además las propiedades del índice facilitan dicha interpretación.

Veamos estas cuestiones siguiendo con el mismo ejemplo, pero con una tabla aún más reducida.

Como puede verse en la tabla anterior, considerando los valores altos del índice, podríamos definir nuevas particiones en ambos conjuntos. Por ejemplo, la de 8x5, la de 5x5 o la de 3x3. Supongamos que optamos por estudiar la de 8x5. De nuevo ponemos el programa en funcionamiento con estos nuevos parámetros, y las tablas que se obtienen son las siguientes:

TABLA VI
TABLA REDUCIDA

nI—nJ	41	47	43	45	40	
182	256.	2759.	361.	5644.	3999.	13019.
184	570.	4541.	800.	4299.	2806.	13016.
181	1362.	4242.	257.	4759.	3403.	14023.
185	1942.	4920.	871.	6937.	4360.	19030.
162	798.	967.	22.	1271.	949.	4007.
186	2853.	3767.	428.	8400.	5654.	21102.
179	401.	1278.	154.	3610.	1567.	7010.
172	841.	860.	79.	3088.	1136.	6004.
	9023.	23334.	2972.	38008.	23874.	

El valor del estadístico es: .06161948
(55.8357% del valor inicial)

Aquí vemos la tabla reducida. En este caso, el valor del índice representa el 55,53% del valor inicial.

Lógicamente, en cada ciclo el estadístico va perdiendo valor y aumentando en consecuencia el porcentaje de pérdida de información. Compete al usuario el establecer el criterio de parada valorando de forma simultánea la interpretabilidad de las tablas reducidas y la pérdida de información que las mismas comporten.

Para facilitar la interpretación, el programa proporciona además la descripción de las clases tanto para el conjunto *I* como para el *J*. En este caso, se trata de las siguientes:

TABLA VII
LA PARTICIÓN ELEGIDA DEL CONJUNTO J CONSTA DE 5 CLASES

nomJ	card	descripción de las clases
41	2	M01 H01
47	10	H14 H15 H16 H17 M15 M16 M17 M12 M13 M14
43	2	H02 M02
45	7	H11 H10 H09 H08 H07 H12 H13
40	5	M11 M10 M09 M08 M07

TABLA VIII
LA PARTICIÓN ELEGIDA DEL CONJUNTO I CONSTA DE 8 CLASES

nomI	card	descripción de las clases
182	13	gua par ins tur dom col els nic por phi hon rwa zam
184	13	nor bru swe pan som bul yug gdl swi rom lux bah net
181	14	les bot ire jam sin hun kir cyp bel fra aul mat jpn mlw
185	19	sam srl ton gre ita spa arg qat kuw per hok swa rok jor sau irq ira gha lib
162	4	uk nze ber bar
186	21	tha vie cos ken ecu cub guy egy syr tun cmr prc moz tri grn fij bur mar uae mal alg
179	7	gam bdi ivc zai upv mli mor
172	6	chd oma caf tog ben ind

TABLA IX
EL VALOR INICIAL DEL ESTADÍSTICO (ES DECIR, EL DE LA TABLA ORIGINAL, SIN NINGUNA REDUCCIÓN NI AGRUPAMIENTO) ES 0.11036

nI—nJ	10	5	3	2
20	76,97%			
8		55,84%		
5		47,76%		
3			32,24%	
2			19,51%	0,74%

Como vemos, en estas tablas aparecen los elementos constituyentes de cada clase.

Para los datos de este ejemplo, la siguiente tabla muestra de forma resumida los porcentajes del estadístico para los diferentes pares de particiones que, en función de los valores del índice cruzado, se han estimado como más pertinentes.

Pensamos que esto ilustra adecuadamente el modo de proceder con este programa. Como antes se dijo, en función de las características de los datos, el usuario debe valorar conjuntamente la ganancia de interpretabilidad que se deriva al reducir el tamaño de la tabla, y la pérdida de información que inevitablemente ello comporta. Y como resultado, se optará entre unas u otras soluciones.

En conclusión, en nuestra opinión este índice cruzado y el programa desarrollado pueden ser una herramienta útil para facilitar el estudio e interpretación de grandes tablas de contingencia.

De esta manera, el usuario puede explorar diferentes porciones de la tabla original pues el programa se encarga de efectuar dichas reducciones y de proporcionar información sobre los componentes de las mismas para facilitar su interpretación.

REFERENCIAS

ANDERBERG, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

BENZECRI, J.P. et coll. (1973). *L'Analyse des Données*. Dunod, Paris.

BERTIN, J. (1977). *La graphique et le traitement graphique de l'information*. Flammarion, Paris.

FISHER, W. (1969). *Clustering and aggregation in economics*. The John Hopkins Press. Baltimore.

GOVAERT, G. (1983). *Classification Croisée*. Thèse d'Etat. Paris VI.

GOVAERT, G. (1984). *Data analysis and Informatics.* Ed. Diday. North Holland. Amsterdam.

GREENACRE, T. (1988). *Clustering the rows and columns of a contingency table.* Journal of Classification. 5: 39-51.

HARTIGAN, J. (1975). *Clustering algorithms.* John Wiley and sons. New York.

YURRAMENDI, Y. (1984). *Contributions à la recherche des méthodes aidant l'interprétation des classifications hiérarchiques.* Thèse 3ème cycle. Paris VI.