INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS - OEA

Oficina del IICA en Uruguay

MANUAL DE USUARIOS
PROGRAMAS DE CLASIFICACION DEL IICA

Montevideo, 1977

MANUAL DE USUARIOS PARA PROGRAMAS DE CLASIFICACION DEL IICA

Montevideo, 1977

Presentación

El éxito de divulgación alcanzado por el IICA al proponer nuevos enfoques, métodos y técnicas de tipificación de empresas, ha tenido una importante contrapartida en el desarrollo y adopción de programas de cómputo.

La demanda de información proveniente de distintos centros de investigación y de varios países, ha hecho que se agotaran rápidamente publicaciones referentes a bases conceptuales y técnicas de cómputo para tipificar. Por este motivo, resultó necesario reeditar el Manual de Usuarios que se había distribuido a comienzos de 1977.

La presente edición del Manual ha sido revisada y ampliada para recoger los últimos programas que se incorporaron a la pequeña biblioteca que el IICA pone a disposición de los Países Miembros.

Es oportuno destacar que los programas fueron en su casi totalidad generados sobre material proporcionado o sugerido por técnicos del Centro Interamericano de Enseñanza de Estadística (CIENES-OEA) y todos fueron desarrollados en la División Computación de la Universidad de la República. En este proceso de hacer operativas las propuestas fue actor central Alvaro Machado.

MANUAL DE USUARIOS PARA

PROGRAMAS DE CLASIFICACION DEL IICA

INDICE

MANUAL DE USUARIOS PARA

PROGRAMAS DE CLASIFICACION DEL IICA

I. Introducción

En este trabajo se presentan instrucciones para utilizar los programas de conglomeración adaptados o desarrollados para la biblioteca del Instituto Interamericano de Ciencias Agrícolas en Montevideo, durante 1976 y 1977.

Asimismo, se presentan instrucciones para el uso de programas de Análisis Discriminante. Los programas de clustering son la contraparte operativa requerida para explorar técnicas propuestas para tipificar empresas por Pedro Ferreira(*) y Alfredo Alonso(**). Con su adaptación no se ha pretendido el desarrollo completo de una biblioteca de cómputo sobre taxonomía numérica, sino tan sólo permitir en Uruguay el proceso de exploración de nuevas posibilidades recomendado en el Seminario sobre Métodos y Problemas en la Tipificación de Empresas Agropecuarias.

Los programas de clustering que se ajustaron a las posibilidades y necesidades del IICA y de varias instituciones nacionales del Uruguay son básicamente los siguientes tres:

1) Van Rijsbergen, para conformar conglomerados en base a la predefinición de una matriz de distancia entre empresas y de un nivel de distancia tolerable para que dos empresas pertenezcan a un mismo cluster.

2) Sparks, que agrupa previa definición del número de clusters que

---

(*) Véase Ferreira, P. en Vol. I de "Seminario sobre Métodos y Problemas en Tipificación de Empresas Agropecuarias". IICA. Serie de Informes, Cursos y Conferencias Nº 92. Montevideo, 1975.

(**) Véase Alonso, A. en "Reunión Técnica sobre Tipificación de Empresas Agropecuarias". IICA. Serie de Informes de Conferencias, Cursos y Reuniones No. 136. Montevideo, 1978

se desea obtener y de la estimación de cuáles empresas constituyen los centros de cada cluster solicitado.

3) Wishart, que parte de considerar cada unidad como un cluster y arriba a un conglomerado único, agrupando en cada etapa de cálculo de manera de minimizar el error global (suma de distancias al cuadrado).

Además de estos programas, tomados de la numerosa literatura disponible, se desarrolló un programa exploratorio que se consideró útil como previo al uso del mecanismo de Van Rijsbergen. Mediante este programa exploratorio, el usuario puede probar alternativas de definición de distancias (las cuales generarán distancias matrices de distancia) y obtener indicios sobre las tolerancias que puede probar al usar el "Fast Hierarchical Algorithm" de Van Rijsbergen.

Cuando la naturaleza de los algoritmos básicos lo permitió, los programas del IICA incorporaron flexibilidad en términos de opciones para ponderar, standardizar variables originales y definir distancias.

Todos los programas fueron desarrollados en la IBM-340/44 de la Universidad de la República y su demanda de memoria central es aproximadamente como sigue:

Van Rijsbergen	76K
Sparks	28K
Wishart	73K
Previo a V.R.	82K

A estos cuatro algoritmos se agregó en 1977 uno propuesto por Anderberg, M.R., que puede ser utilizado para facilitar la interpretación de los resultados obtenidos mediante la utilización del análisis de conglomeración. Este programa se emplea cuando ya se aplicó un algoritmo jerárquico de conglomeración y se desea ver qué variables fueron jugando en las etapas del proceso que lleva a una partición monotética.

Las tareas de programación y pruebas requeridas para estos programas del IICA fueron hechas por el CP Alvaro Machado. El ajuste final de programas y manual de usuario estuvo a cargo de los CP I. Gallo y J. Caffera, con la cooperación del Ing. Alfredo Alonso.

Los programas cuyo uso se expone en este documento sirven para problemas de taxonomía en general, excediendo así el marco de refe-

rencia que motivó su desarrollo. Se espera, por ende, que los mismos puedan ser aprovechados por investigadores que actúan en muy diversos campos científicos.

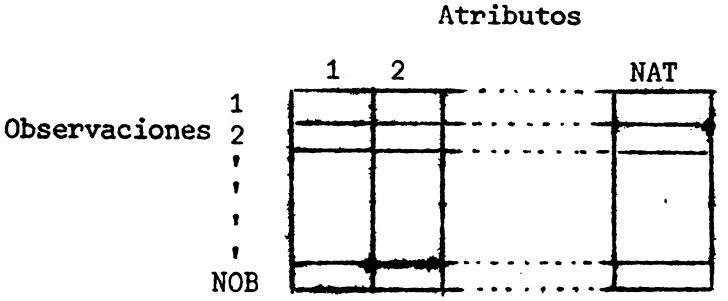
II. Programa de "Van Rijsbergen"

1. Introducción al Programa

Función: Análisis de Conglomeración

Este programa permite realizar Análisis de Conglomeración de acuerdo con el método descripto por C.J. Van Rijsbergen en The Computer Journal, 1970, V. 13, págs. 324-326.

Se supone una cantidad de elementos a clasificarse y que cada elemento tiene asociada una cantidad de atributos. Por una "observación", se entiende el conjunto de los valores numéricos de los atributos asociados a un elemento. Por "juego de datos" o "matriz de datos", se entiende un conjunto de observaciones. En la matriz de datos, cada fila es una observación y cada columna un atributo.



Los programas asumen que la matriz de datos se suministra por filas perforándose los atributos correspondientes a cada elemento ordenadamente en una o más tarjetas, en campos cuya ubicación, tamaño y naturaleza se determinan en el formato.

El programa asigna números correlativos a partir de 1 a cada elemento, según el orden en que se ingresan.

El programa permite realizar varios análisis por ejecución. Un

análisis comprende:

1) una tarjeta de control
2) una o más tarjetas de formato de los datos
3) una o más tarjetas de pesos (ya que pueden usarse ponderaciones)
4) un juego de datos

2. Formato de las tarjetas

a. Tarjeta de control

Col.	Variable	Descripción	
1-48	NOM	título del análisis (alfanumérico)	
49-50	en blanco	NO USADAS	
51-53	NOB	cantidad de observaciones	
54-55	NAT	cantidad de atributos	
56-57	NTAR	cantidad de tarjetas del formato de lectura de los datos	
58-59	NOR	opción de normalización de la matriz de datos	1 - NO 2 - SI
60-62	IDIS	elección de distancia 1 - valor absoluto 2 - euclideana 3 - Mahalanobis	
63-64	IMA	impresión o no de la matriz de datos	1 - NO 2 - SI
65-66	IMD	impresión o no de la matriz de distancia	1 - NO 2 - SI
67-72	CL	nivel inicial (F6.2)	
73-78	FINCR	incremento (F6.2)	
79-80	IMAN	impresión o no de la matriz normalizada	1 - NO 2 - SI

b. Tarjeta de formato

Col.	Descripción
1-80	El formato de lectura de las tarjetas de datos se especifica de acuerdo a las reglas de la sentencia FORMAT del FORTRAN IV, omitiéndose rótulo y la palabra FORMAT, usándose la cantidad de tarjetas que se requiera.

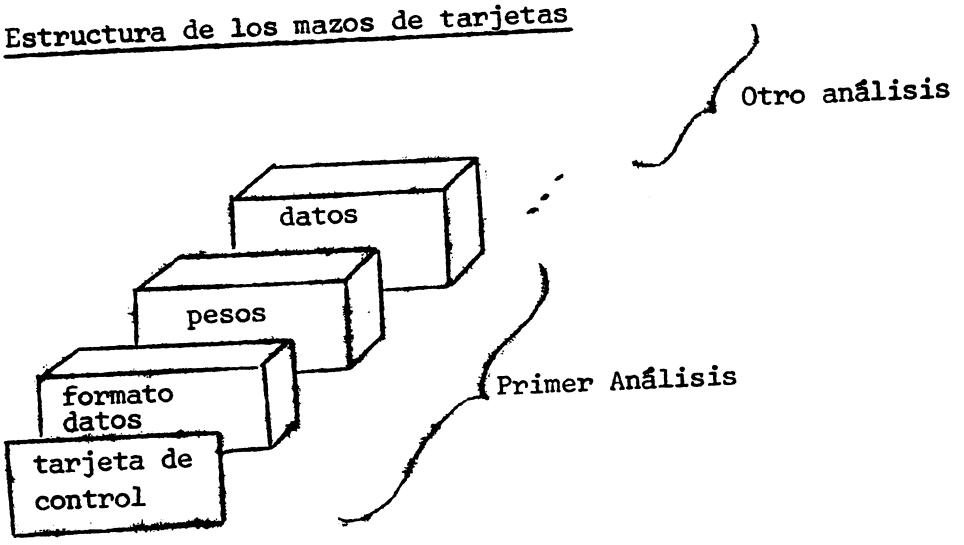
c. Tarjeta de pesos

Col.	
1- 5	Pesos asignados a cada variable. El formato de esta tarjeta es 16F5.2
6-10	
11-15	
16-20	Se usa la cantidad de tarjetas que se requiera
.....	
76-80	

d. Tarjeta de datos

El formato de estas tarjetas se debe corresponder con el especificado.

3. Estructura de los mazos de tarjetas



NOTAS

En la sentencia DIMENSION del programa principal se ha contemplado un máximo de 100 elementos y 20 atributos. No se puede procesar más de 100 elementos, pero sí más de 20 atributos siempre que: NOB * NAT ≤ 2000 siendo NOB: cantidad de elementos, NAT: cantidad de atributos.

Para alterar estas especificaciones debe cambiarse la sentencia DIMENSION y las sentencias con rótulos: 1234, 1235 y 1236.

Ejemplo

Sea: NOB = 120 y NAT = 30
Tenemos que: NOB * NAT = 3600
Las alteraciones son:

```
DIMENSION VO(3600), D(120,120), R(30,30), C(30,30),
          VF1(30), VF2(30), V1(30), V2(30), V3(30), IV1(30),
          IVO(30), W(30)

1234  N100 =  120
1235  N20  =   20
1236  N2000= 3600
```

Sub-rutinas del programa:

Programa Principal
NORMA
DIST1 . . . . . Valor absoluto
DIST2 . . . . . Euclideana
DIST3 . . . . . Mahalanobis
BORRA
MATDIS
LINK
CLOUT
COUNT
DATA
CORRE

III. Programa de "Sparks"

1. Introducción al Programa

Función: Análisis de Conglomeración

Este programa permite realizar Análisis de Conglomeración de acuerdo con el método descripto por D.N. Sparks en el Journal of the Royal Statistical Society, Serie C, Algorithm 58, Applied Statistics 1973, Vol. 22, N° 1.

Valen idénticas consideraciones generales que para II.

En particular: acá se entiende por sub-análisis, la aplicación del método a un juego de datos. A un mismo juego de datos pueden realizársele varios sub-análisis por ejecución. Se llama análisis al conjunto de sub-análisis.

Una aplicación del método (o sea un sub-análisis) exige entre otros datos la especificación de los elementos que se toman como centros iniciales de cluster. A estos efectos, el programa numera correlativamente de 1 a N los elementos según su orden de ingreso.

Un análisis comprende:

1) una tarjeta de control del análisis
2) una o más tarjetas de formato de los datos
3) un juego de datos

Cada sub-análisis comprende:

1) una tarjeta de control del sub-análisis
2) una o más tarjetas de identificación de los objetos que se toman como centros iniciales de clusters.

2. Formato de las tarjetas

Para un análisis:

a. Tarjeta de control del análisis

Col.	Variable	Descripción
1-48	NOM	Título del análisis (alfanumérico)
49-50	en blanco	NO USADAS
51-53	NAN	Cantidad de sub-análisis a efectuar
54-56	NOB	Cantidad de elementos
57-59	NAT	Cantidad de atributos
60	en blanco	
61	IMA	Impresión o no de la matriz de datos 1 - NO 2 - SI
62-63	NTAR	Cantidad de tarjetas del formato de lectura de los datos

b. Tarjeta de formato

Col.	
1-80	El formato de lectura de las tarjetas de datos se especifica de acuerdo a las reglas de la sentencia FORMAT del FORTRAN IV, omitiéndose rótulo y la palabra FORMAT, usándose la cantidad de tarjetas que se requiera.

c. Tarjeta de datos

El formato de estas tarjetas se debe corresponder con el especificado.

Para un sub-análisis:

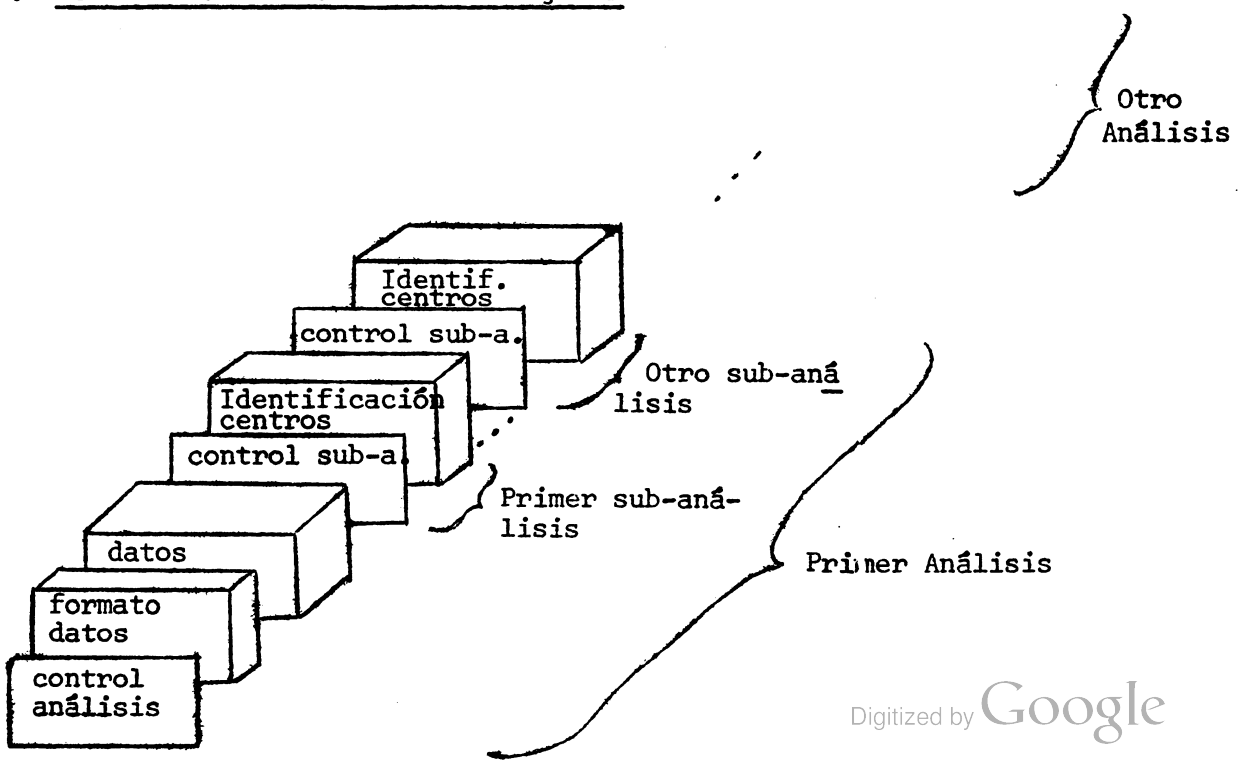
a) Tarjeta de control del sub-análisis

Col.	Variable	Descripción	
1- 2	NCL	Cantidad de clusters	
3- 4	NMO	Cantidad mínima de elementos en cada cluster	
5	IMM	Impresión o no matriz de centros	1 - NO 2 - SI
6	IES	Normalización o no de la matriz de datos	
7	IMCN	Impresión o no de la matriz de datos normalizada	

b) Tarjetas de identificación de los elementos que se toman como centros de cluster iniciales

1- 5 6-10 11-15 16-20 76-80	número de los elementos tomados como centros de cluster iniciales	De no alcanzar una tarjeta se sigue en otras de igual formato.

3. Estructura de los mazos de tarjetas



NOTAS:

En la sentencia DIMENSION del programa principal se ha contemplado un máximo de 100 objetos y 20 atributos. No puede procesarse más de 100 objetos, pero sí más de 20 atributos, siempre que:

NOB X NAT ≤ 2000 donde: NOB - número de objetos
NAT - número de atributos

Para alterar estas especificaciones debe cambiarse la sentencia DIMENSION y las sentencias con rótulos 1234, 1235 y 1236 del programa principal.

Ejemplo

Se desean clasificar 120 objetos con 30 atributos.

NOB = 120
NAT = 30 NOB X NAT = 3600

Las nuevas especificaciones serán:

```
DIMENSION VC(3600), VC(3600), DEV(120), IB(120), IF(120), E(120)
...
...
1234   N100    =  120
1235   N20     =   30
1236   N2000   = 3600
```

Las sub-rutinas que comprende "SPARKS" son las siguientes:

- PROGRAMA PRINCIPAL
- CLUST2
- Subrutina IMPVC
- Subrutina NORMA
- Subrutina DATA

IV. Programa de "Wishart"

1. Introducción al Programa

Función: Análisis de Conglomeración

Este programa permite realizar análisis de Conglomeración por 6 métodos: 1) Vecino más lejano
2) Vecino más cercano
3) Mediana
4) Promedio del grupo
5) Centroide
6) Ward

El detalle de los métodos se expone en un artículo de David Wishart en BIOMETRICS de marzo de 1969.

Valen idénticas consideraciones generales que para II.

Un análisis comprende:

1) una tarjeta de control
2) una o más tarjetas del formato de lectura de los datos
3) un juego de datos

2. Formato de las tarjetas

a. Tarjeta de control

Col.	Variable	Descripción
1-48	NOM	Título del análisis (alfanumérico)
49-50	en blanco	NO USADAS
51-53	NOB	Cantidad de elementos
54-56	NAT	Cantidad de atributos
57-58	NTAR	Cantidad de tarjetas del formato de lectura de los datos
59	NOR	Normalización o no de la matriz de datos 1-NO 2-SI
60	IMA	Impresión o no de la matriz de datos 1-NO 2-SI

61	IMD	Impresión o no de la matriz de distancias	1-NO	2-SI
62	IMAN	Impresión o no de la matriz normalizada	1-NO	2-SI
63	METO	Código de selección del método:		

1 - vecino más cercano
2 - vecino más lejano
3 - intermedio
4 - promedio del grupo
5 - centroide
6 - método de Ward

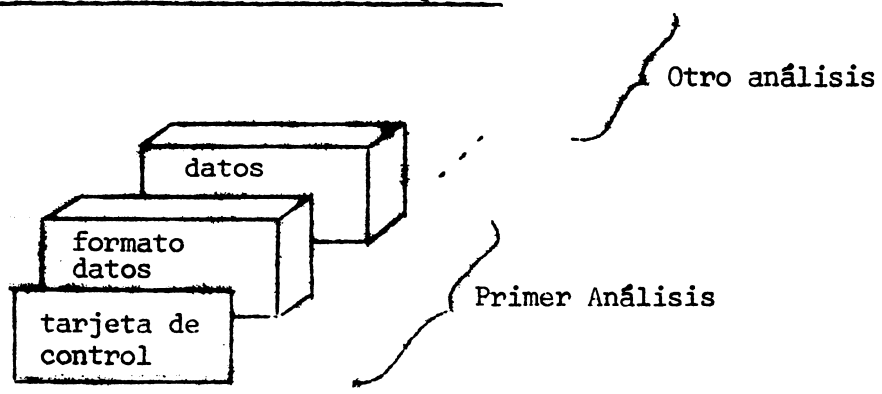
b. Tarjeta de formato

Col.	
1-80	El formato de lectura de las tarjetas de datos se especifica de acuerdo a las reglas de la sentencia FORMAT del FORTRAN IV, omitiéndose rótulo y la palabra FORMAT, usándose la cantidad de tarjetas que se requiera.

c. Tarjeta de datos

El formato de estas tarjetas se debe corresponder con el especificado.

3. Estructura de los mazos de tarjetas



Las sub-rutinas que comprende Wishart son:

- Programa principal
- NORMA
- DIST2
- DAVID
- D2
- DATA

V. Programa "Previo a Van Rijsbergen"

1. Introducción al Programa

Función: Explorar la matriz de distancias de un conjunto de objetos, a efectos de poder iniciar con información útil el proceso de conglomeración previsto por Van Rijsbergen.

Este programa fue preparado por Alvaro Machado.

Valen idénticas consideraciones generales que para II.

Se construye una matriz de distancia (disimilaridad) entre objetos, conforme al criterio que proponga el usuario: euclideana, valor absoluto, Mahalanobis (con opciones a ponderar atributos en todos los casos).

El algoritmo provee al usuario de la siguiente información:

a) Distancias Posibles

El número de objetos (empresas) determina la cantidad de distancias posibles entre ellos, tomados de dos en dos. La fórmula es:

$$\frac{NE\ (NE-1)}{2}$$

donde NE: número de empresas

b) Distancia Media

Es el valor medio de la matriz de distancias

$$D = \frac{\Sigma\ d}{NE}$$

donde d: distancia entre dos objetos

c) Distancia Mínima

Es el valor mínimo hallado en la matriz de distancias

d) Distancia Máxima

Es el valor máximo hallado en la matriz de distancias

Un análisis comprende:

1) una tarjeta de control
2) una o más tarjetas de formato de lectura de los datos
3) una o más tarjetas de pesos
4) un juego de datos

La única diferencia en el manejo de este programa con respecto al manejo del programa II radica en que:

en la tarjeta de control del análisis las columnas 67-78 se deben dejar en blanco.

Las sub-rutinas que componen el previo a Van Rijsbergen son:

- Programa Principal
- H
- BORRA
- NORMA
- DIST1
- DIST2
- DIST3
- CORRE
- DATA

VI. Programa de Análisis Discriminante

1. Introducción al Programa

Función: Cálculo de parámetros de función lineal discriminante e información asociada.

Se calcula un conjunto de funciones lineales en base a los datos de varios grupos, con el objetivo usual de clasificar un nuevo elemento individual en alguno de esos grupos. La clasificación de un nuevo elemento individual en algún grupo es hecha evaluando cada una de las funciones y tomando el grupo para el cual el valor es el máximo.

El programa disponible es parte del paquete científico de IBM y se utiliza en tipificación como prueba de la calidad de un agrupamiento hecho por cualquier método.

El análisis de discriminante consiste del programa principal:

MDISC

y tres sub-rutinas:

DMATX

MINV

DISCR

Se puede hacer análisis de discriminante con este paquete de sub-rutinas con las siguientes restricciones:

1) los grupos no deben ser más de 5. Esto puede alterarse fácilmente y las tarjetas de lectura de datos ya prevén más de 5 grupos.

2) las variables no pueden ser más de 10
3) las observaciones totales no deben ser mayores de 250
4) por tarjeta deben venir 12 observaciones cada una de ellas en 6 columnas, comenzando en columna 1.

Si no se cumpliera alguna de esas condiciones, se debe modificar la sentencia DIMENSION correspondiente de MDISC (MAIN PROGRAM).

Si las observaciones están perforadas en forma distinta, es necesario modificar el formato de entrada. Las normas para modificar el programa se describirán más adelante.

2. Formato de las tarjetas

a. Tarjeta de control

La primera tarjeta es leída por el programa principal MDISC.

El diseño standard es el siguiente:

Col.	Descripción
1- 6	Nombre del trabajo
7- 8	Cantidad de grupos
9-10	Cantidad de variables
11-15	Cantidad de observaciones en el primer grupo
16-20	Cantidad de observaciones en el segundo grupo
21-25	Cantidad de observaciones en el tercer grupo
	.....
	.....
	.....
65-70	Cantidad de observaciones en el 12° grupo

Si hay más de 12 grupos, se continúa en una segunda tarjeta.

1- 5	Número de observaciones en el 13° grupo
6-10	Número de observaciones en el 14° grupo

b. Tarjeta de Datos

La primera variable de cada observación debe comenzar en columna 1. Después de la última tarjeta de datos si se desea otro análisis viene otro grupo de tarjetas en el cual la 1ra. es otra tarjeta de control segunda del grupo de tarjeta de datos, y así reiteradamente.

c. Impresiones standard de salida (output)

El programa proporciona la siguiente salida:

1) Medias de variables en cada grupo
2) Matriz de dipersión total
3) Medias comunes
4) D^2 (Mahalanobis generalizada)
5) Constante y coeficientes de cada función discriminante
6) Probabilidad asociada con el mayor valor de la función evaluada para cada observación y función que mejor clasifica a la observación.

La capacidad del programa puede ser incrementada o decrementada haciendo cambios en la sentencia DIMENSION. De venir ya perforados los datos, puede alterarse la lectura standard modificándose la sentencia 5 FORMAT (12F6.0).

VII. Programa "ERROR"

1. Introducción al Programa

Función: Desagregación de la suma de cuadrados total

Este programa sirve para analizar la varianza dentro de grupos para algunas variables seleccionadas y fue tomado de Anderberg, M.R. "Cluster Analysis for Applications", Academic Press, 1973.

Se emplea una vez completado un proceso de clustering. El programa

toma la suma de cuadrados dentro de grupos como la varianza no explicada y la suma de cuadrados entre grupos como la varianza explicada por los agrupamientos. Luego calcula el porcentaje de varianza no explicada para cada una de las variables que se seleccionen y para cada una de las etapas de clustering.

Mediante su utilización se puede explicar la influencia relativa de cada una de las variables en la definición de los agrupamientos resultantes del proceso.

2. Formato de las tarjetas

Tarjeta	Col.	Variable	Descripción
1	-	-	Título
2	1- 5	NTDAT	Unidad de entrada de los datos
	6-10	NTMRG	Unidad de entrada de las especificaciones de unión de clusters
	11-15	NS	Número de objetos
	16-20	NV	Número de variables en la corrida (máximo 14)
3	-	FMTD	Formato de lectura de los datos
4	(Deberán haber NV tarjetas)		
	1- 4	LABEL(I)=4	Nombre de la i-ésima variable
5	-	DATA	Datos originales, de acuerdo con el formato FMTD
6	(Deberán haber tantas de estas tarjetas como etapas se generaron de acuerdo con cualquier procedimiento jerárquico de clustering)		
	1-10	K	Etapa del proceso de clustering
	11-20	II	Cluster que se une en la etapa K (número de orden del menor)
	21-30	JJ	Cluster que se une en la etapa K (número de orden del mayor)
	31-46	CR	Valor de la función objetivo, si el algoritmo empleado generó esta información que es sólo para impresión.

3. Impresión standard de salida (output)

La impresión de salida se realiza en forma de cuadro, especificando:

a. Etapa del proceso de clustering (de 1 a NS-1)
b. Número de clusters (de NS-1 a 1)
c. Número de los clusters que se unen: II y JJ (II < JJ)
d. Valor de la función objetivo: CR
e. Proporción de varianza no explicada por los agrupamientos de cada una de las variables seleccionadas.

* * * * *

Mim.No.129/78
/mmc