PROYECTO COOPERATIVO DE INVESTIGACION SOBRE TECNOLOGIA AGROPECUARIA EN AMERICA LATINA (PROTAAL)

Documento Nº 8

TIPIFICACION DE CONGLOMERADOS Y SU ANALIS DE ESTABILIDAD

ALFREDO ALONSO
HUGO COHAN

IICA
INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS - OEA

OFICINA EN COLOMBIA

Publicación miscelánea Nº 166

Bogotá, Septiembre de 1977

El Proyecto Cooperativo de Investigación sobre Tecnología Agropecuaria (PROTAAL) representa un esfuerzo que tiene como fin desarrollar un conjunto de investigaciones referidas a la naturaleza del proceso tecnológico agropecuario en la región. Este esfuerzo es llevado a cabo con la cooperación del Instituto Interamericano de Ciencias Agrícolas (IICA), quien actúa como agencia ejecutora; el Instituto Colombiano Agropecuario (ICA); la Fundación Ford; el Programa de las Naciones Unidas para el Desarrollo (PNUD), y el Centro Internacional de Investigaciones para el Desarrollo del Canadá (CIID).

El Proyecto plantea el análisis de dicho proceso desde una perspectiva integradora, que toma al proceso tecnológico como un fenómeno endógeno al funcionamiento de la sociedad en que el mismo se desarrolla. Este análisis intenta proveer información útil para el mejor entendimiento del problema tecnológico, y consecuentemente a la definición de políticas, modelos organizacionales y acciones que contribuyan al progreso tecnológico y al desarrollo del sector agropecuario.

Las actividades del Proyecto se iniciaron el 1° de enero de 1977 y desde el punto de vista organizativo las mismas se materializan principalmente a través de la participación de un número de equipos de investigación pertenecientes a instituciones oficiales y privadas de diversos países del continente.

A fin de hacer conocer los resultados de estas investigaciones y favorecer el intercambio de información en un sentido más amplio, el Proyecto se propone editar una serie de trabajos y monografías de los siguientes tres tipos:

1. Trabajos metodológicos y resultados de investigaciones empíricas que resultan de las actividades centrales del Proyecto.

2. Trabajos que surgen de actividades vinculadas al Proyecto.

3. Trabajos preparados por los integrantes del Proyecto y eventualmente por otros autores, que estén relacionados a las actividades del Proyecto y que sean útiles al desarrollo del mismo.

Los trabajos serán publicados, en general, en versiones no definitivas y por lo tanto, los comentarios críticos son solicitados.

Martín E. Piñeiro

Eduardo J. Trigo

Raúl Fiorentino

PUBLICACIONES DEL PROYECTO

Documento No. 1 : Martín Piñeiro, Eduardo Trigo y Raul Fiorentino. "El Proceso de Generación Difusión-Adopción de Tecnología Agropecuaria en América Latina". IICA Oficina en Colombia, Enero de 1977. Mimeografiado.

Documento No. 2 : Martín Piñeiro y Eduardo Trigo. "La Transferencia de Tecnología y la Educación Superior". Seminario sobre la Educación Agrícola para el Desarrollo Rural y Económico. IICA Oficina en Colombia, Abril 25-27 de 1977.

Documento No. 3 : Martín Piñeiro y Eduardo Trigo. "Un Marco General para el Análisis del Progreso Tecnológico Agropecuario: Las Situaciones de Cambio Tecnológico". IICA Oficina en Colombia, Abril de 1977. Publicación Miscelánea No. 149.

Publicado también como: Primer Seminario de Modernizaçao de Empresa Rural. Ministerio de Agricultura SUPLAN y Fundaçao Getulio Vargas EIAP, Río de Janeiro, Mayo de 1977.

Documento No. 4 : Martín Piñeiro y Eduardo Trigo. "La Planificación de la Investigación a partir de Programas por Producto: Algunos comentarios críticos". IICA Oficina en Colombia, Agosto de 1977. Publicación Miscelánea No. 150.

Publicada también como: (a) Informe Técnico No. 39 Programa de Estudios Agroeconómicos. División de Estudios Socioeconómicos. Instituto Colombiano Agropecuario. Bogotá, Julio de 1977. (b) Seminario sobre Producción Animal en Areas de Agricultura Tradicional. Facultad Agronómica, Universidad de Nariño. IICA Oficina en Colombia, Pasto, Mayo de 1977. Mimeografiado.

Documento No. 5 : Eduardo Trigo y Martín Piñeiro. "Análisis de los Modelos Institucionales de Generación de Tecnología Agropecuaria: Algunas ideas metodológicas." IICA Oficina en Colombia, Agosto de 1977. Publicación Miscelanea No. 151.

Documento No. 6 : Martín Piñeiro, Eduardo Trigo y Raúl Fiorentino. "La Generación y Transferencia de Tecnología Agropecuaria. Notas sobre la Funcionalidad de los Centros Nacionales de Investigación." IICA - Oficina en Colombia, Agosto de 1977.

Documento No. 7 : Martín Piñeiro, Eduardo Trigo y Raúl Fiorentino. "Notas sobre la Metodología para el Estudio de la Naturaleza y Efectos de las Innovaciones Tecnológicas en el Sector Agropecuario". IICA - Oficina en Colombia. Agosto de 1977.

Documento No. 8 : Alonso Alfredo y Hugo Cohan. "Tipificación de Conglomerados y su Análisis de Estabilidad". IICA-Oficina en Colombia. Septiembre, 1977.

Documento No. 9 : Reunión Preparatoria de la Conferencia internacional sobre Potencial para la Cooperación entre Sistemas Nacionales de Investigación Agrícola. Bellagio, Italia, Octubre 17-21 de 1977."
"Sistemas Nacionales de Investigación Agrícola en América Latina". IICA- Oficina en Colombia, Septiembre, 1977.

Documento No. 8

TIPIFICACION POR CONGLOMERADOS Y SU ANALISIS DE ESTABILIDAD*

Alfredo Alonso
Hugo Cohan

* Documento preparado como parte de las acciones desarrolladas por la Oficina del IICA en Uruguay.

Colombia, Septiembre de 1977

CONTENIDO

TIPIFICACION POR CONGLOMERADOS Y SU ANALISIS DE ESTABILIDAD

(versión preliminar para discusión)

I. INTRODUCCION

El presente trabajo se basa en un estudio en proceso para el "Proyecto de Desarrollo Regional Fondo Simón Bolívar", en Uruguay.(*) Se trata de exploraciones para identificar un conjunto de sectores censales con representatividad suficiente, en términos de características generales del área, como para iniciar en ellos un proyecto piloto de desarrollo. En función de algunas hipótesis teóricas o información sobre la zona, se propusieron a-priori las características medias del sector censal que representaría al conjunto en el cual podría realizarse ese esfuerzo de desarrollo.

El planteo es en realidad parte de un conjunto de exploraciones metodológicas más amplias. La forma específica que asume esta esta presentación es un intento de aproximación a propuestas del PROTAAL.(**) En efecto, el proyecto PROTAAL prevé agrupar empresas en función de atributos teóricos; estimación a-priori de características definitorias de empresas predominantes. Por ello se estima que este trabajo, puede contribuir al desarrollo de la metodología para dicho proyecto.

Para hacerlo más facilmente compatible con el PROTAAL, en este documento metodológico de discusión se identifican como "empresas" a cada uno de los 47 sectores censales con los que en realidad se trabajó.

Conviene advertir que si se comienza por una tipología apriorística es difícil llegar más allá que a docimar el planteo apriorístico sobre las características de esa "unidad representativa"..Pero aún así, con ese limitado enfoque, la potencial inestabilidad de los conglomerados podría inducir a serios errores, de no aprovecharse las sugerencias metodológicas recientemente formuladas por Ferreira y Kaminsky.(***)

(*) Convenio IICA-MAP

(**) Proyecto Cooperativo de Investigación sobre Tecnología Agropecuaria en América Latina

(***) Trabajos presentados a la reciente Reunión Técnica Sobre Tipificación de Empresas Agrarias. En publicación (3) (4)

Ese es el mensaje fundamental que se plantea en el presente aporte.

Las técnicas empleadas en el proceso han sido ya resumidas en un documento de difusión (1).

II. DESCRIPCION GENERAL DEL AREA Y VARIABLES UTILIZADAS

La información utilizada corresponde al departamento de Cerro Largo y proviene del Censo General Agropecuario de 1970, tabulada a nivel de sector censal.

El departamento citado se encuentra situado al Noreste de la República Oriental del Uruguay. En 1970 fueron censadas en él 1:311.860 hectáreas dedicadas a la actividad agropecuaria; esto representa aproximadamente el 8% de la superficie total del país.

Se trata de un departamento fundamentalmente ganadero, donde las tierras dedicadas a agricultura representan sólo el 2.32% del total.

Para cada uno de los 47 sectores censales en que se divide, se obtuvo información sobre nueve variables. Ellas son:

1. Tamaño promedio de las explotaciones: superficie censada en cada sector dividida entre el número de explotaciones correspondiente.

2. Hectáreas por trabajador: hectáreas censadas divididas entre el número de trabajadores de 14 años y más.

3. Mano de obra familiar: número de trabajadores de 14 años y más, comprendidos en la categoría "productor y miembros de su familia", dividido entre el número total de trabajadores de 14 años y más.

4. Hectáreas por tractor: hectáreas censadas divididas entre el número total de tractores.

5. Tamaño de los potreros: superficie total censada, dividida entre el número total de potreros.

6. Relación Agrícola-Ganadera: superficie dedicada a agricultura dividida entre la superficie dedicada a ganadería (en porcentaje).

7. Superficie mejorada: superficie de pasturas mejoradas dividida entre la superficie dedicada a ganadería (en porcentaje).

8. Relación Ovino-Bovino: número total de ovinos dividido entre el total de bovinos.

9. Ganado lechero: total de ganado lechero sobre el total de vacunos (en porcentaje).

La matriz original de observaciones (47 x 9) y su versión normalizada se presentan en el Anexo 1.

Conviene insistir en que las observaciones son ya medias de distintas empresas que se encuentran en cada sector censal, aunque en este trabajo se las interpreta y discute como si surgieran de haber relevado un reducido universo de 47 predios.

III. HIPOTESIS PLANTEADA

Se trató de explorar los problemas metodológicos y operativos que podría traer el proponer "empresas representativas" a-priori para luego conglomerar con las técnicas disponibles en el IICA (2), a efectos de docimar la hipótesis planteada. La "representatividad" fue definida como porcentaje de vacunos sobre el total del área.

En forma paralela y buscando una mayor similitud con el planteo de PROTAAL, aunque sea sólo en apariencia, se identificaron dos grupos de variables. Un primer grupo de variables consideradas como "estructurales", se constituyó por las 4 primeras (Tamaño medio, Has/trabajador, mano de obra familiar/Ha , Has/tractor). El grupo de variables "de comportamiento", estuvo formado por las 4 últimas (Has Agricultura/Has Ganadería, Has Mejoradas/Has Ganadería, Ovinos/Bovinos, Ganado Lechero/Ganado Total). El PROTAAL, en realidad, aún no ha definido claramente si empleará esta diferenciación. Pero parece tener una tendencia a, eventualmente, intentarla.

Aún dentro de los límites de un trabajo exploratorio de metodología, y con las restricciones de tiempo e información, con las que se operó, puede intentarse una justificación para esta propuesta. Las unidades estructurales sugerirían características relativamente estables de los predios, definitorias del comportamiento (presumiblemente diferenciado) ante una dada situación de política económica o institucional.

Se plantea entonces, la posibilidad de agrupar de acuerdo con las variables "estructurales", para luego probar la calidad de los agrupamientos obtenidos mediante las variables "de comportamiento". Debido al enfoque dado al problema, se dejó como caso especial la variable número 5 "tamaño promedio de los potreros", porque ella depende fundamentalmente del tamaño de las empresas, estando así altamente correlacionada con la variable número 1 (superficie media).

Se supone la existencia de una "empresa representativa". Ella contribuiría con un 50% o más del total de vacunos del área ("representatividad") y, en términos de las variables "estructurales", se podría definir de la siguiente forma(*):

1. Tamaño _____ 500 hectáreas
2. Mano de obra _____ 3 trabajadores
   (166,67 hectáreas por trabajador)
3. Mano de obra familiar _____ 60 por ciento
4. Mecanización _____ 1 tractor cada 1000 hectáreas

La hipótesis planteada, en realidad, propone la existencia de tres grupos de empresas que definirían las tres "empresas tipo" de la zona, de modo que los "tipos teóricos" serían los del Cuadro No.1

(*) En alguna similitud con los atributos (elasticidades precio de demanda, régimen de tenencia, tamaño, tipo de mano de obra) propuestos en el PROTAAL.

CUADRO No 1: ATRIBUTOS PREVISTOS PARA LAS EMPRESAS

ATRIBUTOS	EMPRESAS TIPO			PROMEDIO DEPARTAMENTAL SEGUN CENSO
	I	II	III	
1. Tamaño (hectáreas)	150	500	1.000	300
2. Mano de obra (hectáreas por trabajador)	75	166	250	120
3. Mano de obra familiar (porcentaje)	90	60	20	70
4. Mecanización (hectáreas por tractor)	1.500	1.000	2.500	1.800
5. Representatividad (porcentaje de vacunos)	20	50	30	100

Es decir que se propone la existencia de una distribución tri modal, en un espacio de empresas definido por cuatro variables. Una de ellas (la II) sería la "más representativa".

IV. METODOS DE CLASIFICACION UTILIZADOS

Una vez predefinidos los tipos de empresa y la "empresa representativa" de la región, se procedió a agrupar las 47 empresas mediante dos de los algoritmos disponibles en el IICA. Se emplearon el algoritmo de Sparks y el de Ward.

Ambos métodos de análisis de conglomeración utilizados proceden a clasificar los elementos de un conjunto de modo que la varianza dentro de los conglomerados sea mínima y, por consiguiente, la varianza entre grupos máxima.

1. Algoritmo de Ward

El algoritmo de Ward es jerárquico, a partir de la partición politética (cada elemento es un conglomerado) va agrupando observaciones o conglomerados hasta llegar a la partición monotética (todos los elementos en un mismo conglomerado).

El método opera a partir de una matriz de distancias euclideanas al cuadrado calculada entre los elementos y los va agrupando de modo que la varianza dentro de clusters (función objetivo) sea mínima.

En un principio, el algoritmo elige la menor distancia al cuadrado y une a los dos elementos para formar un conglomerado. A partir de allí procede a corregir las distancias de los demás elementos con respecto al cluster recién formado, de modo que las nuevas distancias van a expresar el doble del incremento que se producirá en la función objetivo al unir los dos grupos considerados. El proceso continúa en la misma forma, eligiendo en cada iteración la menor $d^2(i,j)$ para ver qué conglomerados se deben fusionar, incrementando cada vez la función objetivo en $\frac{1}{2}$ de $d^2(i,j)$. Al realizar (n-1) iteraciones se llega a la partición monotética.

Al finalizar el análisis, el valor de la función objetivo será igual a la suma de cuadrados total, ya que en este paso final la varianza entre clusters es nula.

2. Algoritmo de Sparks

El algoritmo de Sparks no es jerárquico, de modo que al iniciar el análisis el usuario debe fijar el número de clusters y los centros iniciales de los conglomerados.

Para cada observación, el programa calcula las distancias a los centros iniciales y las asigna al más cercano. Luego calcula la media de los agrupamientos formados y estos valores pasan a ser los nuevos centros.

Se examinan todos los elementos, calculando las distancias a los nuevos centros. Se reasigna una observación que pertenece al cluster S_k si la suma de los desvíos al cuadrado con respecto al centro del cluster St, es menor que con respecto al centro de S_k, aún cuando los centros cambien simultáneamente.

El mecanismo sigue repitiendo el proceso, hasta que no se produce ninguna reasignación y calcula el valor de la suma de cuadrados dentro de clusters.

V. APLICACION DE LOS METODOS DE CONGLOMERACION

1. Resultados de conglomerar con cuatro variables

Partiendo de los tipos de empresas propuestos en la sección III, se procedió a agrupar las explotaciones utilizando el método de Sparks. El procedimiento, conviene recordar, requiere proponer centros iniciales. Estos centros iniciales fueron los tres tipos previstos por hipótesis. Para evitar que las unidades de medida influyeran en la determinación de las distancias entre empresas se procedió a normalizar las variables.

El resultado que se obtuvo se presenta en el Cuadro No 2.

CUADRO No 2: PARTICION DE SPARKS EN TRES GRUPOS

CLUSTER	NUMERO DE ELEMENTOS	EMPRESAS EN CADA CLUSTER
I	26	1-2-3-4-5-6-7-13-14-15-16-17-19-28-32-33-35-37-38-41-42-43-44-45-46-47
II	14	8-9-20-21-23-24-26-27-29-30-34-36-39-40
III	7	10-11-12-18-22-25-31

La suma de cuadrados dentro de clusters fue de 68.02, de modo que la varianza no explicada representa el 36.18% del total.

Los promedios de los agrupamientos resultantes que podrían usarse como definidores de las "empresas tipo" y su "representatividad" son los del Cuadro No 3 (a compararse con lo previsto, presentado en el Cuadro No 1).

CUADRO No 3: VALORES MEDIOS EN CADA CLUSTER
SPARKS - CUATRO VARIABLES - TRES GRUPOS

ATRIBUTOS	EMPRESAS TIPO		
	I	II	III
1. Tamaño (hectáreas)	218.47	408.10	1735.55
2. Mano de obra (hectáreas por trabajador)	92.83	190.68	341.63
3. Mano de obra familiar (porcentaje)	77	73	31
4. Mecanización (hectáreas por tractor)	1322.41	4295.54	2875.79
5. Representatividad (porcentaje de vacunos)	38.17	37.87	23.96

En términos generales, la empresa tipo II se asemeja bastante a la "empresa representativa" que se había planteado como hipótesis. Sin embargo su "representatividad" no llega al nivel esperado, en la medida en que no aparece como predominante en la producción de vacunos de la zona (criterio que se había preestablecido).

Los resultados no permiten considerar la hipótesis planteada como totalmente correcta. No obstante ello, pese al problema de representatividad, pudo habérsela retenido. O, en un esfuerzo por salvar otros aspectos del proyecto en el que la tipificación se inserta, pudo haberse hecho algún ajuste ad-hoc en los atributos. Y, sin embargo, aún cuando esto hubiera pasado el requisito de representatividad, puede haber error en la tipificación. Para obtener conclusiones

más acertadas, se deberían haber analizado los datos con mayor detalle. En lo que sigue, se prestará atención a esto.

En este sentido se trató de estudiar la estabilidad de las clasificaciones obtenidas. Como señalan Ling y Killough (5) las técnicas conformadoras de grupos proceden a agrupar los elementos en clusters, existan estos en la población original o no.

Se debería, entonces, analizar la estabilidad de los grupos formados tratando de encontrar los agrupamientos "naturales" que puedan existir en la población. En este caso, además, se tiene el problema de que el método de Sparks suele ser sumamente sensible a la elección de los centros iniciales, obteniéndose mínimos locales en lugar del mínimo global cuando no existen grupos perfectamente definidos en la población analizada.

Para estudiar cómo se agrupaban los elementos al cambiarse el número de clusters, se obtuvieron las particiones en 2,4 y 5 conglomerados, eligiéndose como centros iniciales a las empresas en base a su tamaño. Los resultados obtenidos se resumen en el Cuadro No 4.

CUADRO No 4: PARTICIONES ALTERNATIVAS CON SPARKS (CUATRO VARIABLES)

PARTICION	NUMERO DE ELEMENTOS EN CADA CONGLOMERADO					SUMA DE CUADRADOS DENTRO DE CLUSTERS
	I	II	III	IV	V	
2 clusters	40	7	-	-	-	105.88
3 clusters	26	14	7	-	-	68.02
4 clusters	20	13	10	4	-	48.04
5 clusters	17	11	8	8	3	37.92

Los agrupamientos aparecen como poco estables. Al pasar a particiones de fineza en aumento, se producen cambios importantes de empresas entre grupos.

Se procedió, entonces, a agrupar a las empresas mediante el algoritmo de Ward, de modo de comparar los resultados obtenidos por dos métodos de clustering.

Un resumen de los resultados obtenidos aparecen en el Cuadro No 5.

CUADRO No 5: RESUMEN DE LA EVOLUCION DE PARTICIONES CON WARD
(CUATRO VARIABLES)

PARTICION	NUMERO DE ELEMENTOS EN CADA CONGLOMERADO					SUMA DE CUADRADOS DENTRO DE CLUSTERS
	I	II	III	IV	V	
2 clusters	40	7	-	-	-	105.88
3 clusters	27	13	7	-	-	68.73
4 clusters	17	13	10	7	-	50.19
5 clusters	17	13	10	4	3	37.17

Comparando los cuadros No 4 y No 5, se advierte que la concordancia de las particiones en 2 conglomerados obtenidas por los dos métodos es perfecta. Las particiones en 3 conglomerados difieren en clasificar a un solo elemento, resultando mejor la asignación que hace Sparks porque genera una menor suma de cuadrados dentro de clusters. En las otras particiones se encuentran diferencias apreciables en el número de elementos en cada cluster y en la asignación de las empresas. Quedan dudas así, sobre si debe mantenerse la propuesta de tres grupos.

Al aparecer las clasificaciones como poco estables, se volvió a aplicar el método de Sparks. En esta ocasión se cambiaron los centros iniciales, produciéndose numerosos cambios en los agrupamientos resultantes. Se confirma la inestabilidad y se impone un estudio más detallado del problema.

De las particiones que se obtuvieron al cambiarse para Sparks los centros iniciales se eligieron los más eficientes, o sea las que mostraban una menor intravarianza. Ellas se compararon con los agrupamientos obtenidos con Ward. El detalle de los agrupamientos aparece en el Anexo 2, pero un resumen se tiene en el Cuadro No 6.

CUADRO No 6: COMPARACION DE AGRUPAMIENTOS ALTERNATIVOS CON WARD Y SPARKS * (CUATRO VARIABLES)

PARTICIPACION	METODO	NUMERO DE ELEMENTOS EN CADA CONGLOMERADO					SUMA DE CUADRADOS DENTRO DE CLUSTERS
		I	II	III	IV	V	
2 CLUSTERS	SPARKS	40	7	-	-	-	105.88
	WARD	40	7	-	-	-	105.88
3 CLUSTERS	SPARKS	26	14	7	-	-	68.02
	WARD	27	13	7	-	-	68.73
4 CLUSTERS	SPARKS	17	10	14	6	-	46.31
	WARD	17	13	10	7	-	50.19
5 CLUSTERS	SPARKS	17	13	10	4	3	35.85
	WARD	17	10	13	4	3	37.17

* Adviértase que como "Sparks" figuran los resultados más eficientes obtenidos de explorar agrupamientos generados con distintos centros iniciales.

Antes de realizar un análisis más profundo se podría pensar en que existen en la población 2 ó 3 grupos "naturales". Ello lo sugieren los resultados obtenidos con los dos métodos y el hecho de que para esas dos particiones no se produjeron cambios al cambiarse los centros iniciales en Sparks.

Para evaluar la similaridad entre las particiones, se utilizaron tablas de contingencia o índices basados en los pares similares.

2. Décimas de resultados alternativos con cuatro variables

a) Tablas de contingencia

Se plantearon tablas de contingencia para docimar la existencia de asociación entre las particiones obtenidas con los dos métodos.

A partir del valor de X^2 calculado se midió el grado de asociación, mediante el estadígrafo de contingencia C de Pearson. Por ejemplo para las particiones en 3 clusters se tienen los elementos del Cuadro No 7.

CUADRO No 7: TABLA DE CONTINGENCIA SPARKS-WARD, TRES CLUSTERS (CUATRO VARIABLES)

SPARKS \ WARD	I	II	III	f.m.
I	26	0	0	26
II	1	13	0	14
III	0	0	7	7
f.m.	27	13	7	47

Para todas las particiones analizadas se debe rechazar la hipótesis nula de ausencia de asociación a nivel 1%, y el estadígrafo C calculado revela en todos los casos que el grado de asociación es muy alto. Los resultados se resumen en el Cuadro No 8.

CUADRO No 8: PRUEBAS DE X^2 y C PARA AGRUPAMIENTOS ALTERNATIVOS (WARD Y SPARKS - CUATRO VARIABLES)

PARTICION	TABLAS DE CONTINGENCIA		ESTADIGRAFO DE CONTINGENCIA	
	X^2 CALCULADO	X^2 TABLAS	C CALCULADO	C MAXIMO
2 clusters	47.12	6.63	0.71	0.71
3 clusters	89.16	13.3	0.81	0.82
4 clusters	112.91	21.7	0.84	0.87
5 clusters	169.17	32.0	0.88	0.89

En todos los casos, los resultados se deben tomar con las reservas que surgen de considerar el elevado número de casillas con ceros que aparecen en las tablas de contingencia (ver anexo 3).

b) Pares similares

Este método se basa en la consideración de los elementos de a pares, registrando como similares (P.S.) a las unidades que son clasificadas en un mismo cluster por los dos métodos.. Dos observaciones que pertenecen a un mismo cluster en las dos particiones forman un "par similar".

Para calcular el número de pares similares, se aprovechó la información resumida en las tablas de contingencia. Por ejemplo, en la que aparece en el Cuadro No. 7 se tiene:

- 26 observaciones son clasificadas en el cluster I por ambos métodos, de modo que se tienen C_2^{26} pares similares.
- 13 observaciones que son clasificadas en el cluster II por ambos métodos (C_2^{13}).
- 7 observaciones son clasificadas en el cluster III por los dos métodos (C_2^7).

En total, se tienen 424 pares similares, porque resta solamente una empresa que es clasificada en el cluster I por Ward y en el II por Sparks y una sola observación no puede determinar ningún par similar.

A partir del número de pares similares se calculó la medida de similaridad entre particiones de Rand (Ra) que es igual al número de pares similares sobre el total de pares posibles o sea:

$$Ra : \frac{P.S.}{C_2^n}$$

Sin embargo, como señala Kaminsky, M.(4) el Ra exhibe limitaciones como medida de similaridad, debido a que al aumentar el número de clusters el número de pares similares posibles va disminuyendo. El Ra se va haciendo así cada vez menor, no posibilitando la comparación de diferentes particiones.

Kaminsky propone un nuevo índice (Mak). Este parece más adecuado que el Ra, debido a que compara los PS con el máximo posible de pares similares dado un agrupamiento y luego dado el otro, tomando como medida de la similaridad un promedio de los dos resultados.

Por ejemplo, para la partición en tres grupos se tienen clusters formados por 26, 14 y 7 elementos según Sparks y por 27, 13,y 7 elementos según Ward. Entonces se tiene:

$$\frac{424}{C_2^{27} + C_2^{13} + C_2^7} = \frac{424}{450} = 0.9422$$

$$\frac{424}{C_2^{26} + C_2^{14} + C_2^7} = \frac{424}{437} = 0.9703$$

$$Mak = \frac{0.9422 + 0.9703}{2} = 0.956$$

Los resultados obtenidos fueron los del Cuadro No 9.

CUADRO No 9: INDICES DE SIMILARIDAD PARA PARTICIONES ALTERNATIVAS (WARD Y SPARKS CON 4 VARIABLES)

PARTICION	Ra	Mak
2 clusters	0.74	1.00
3 clusters	0.39	0.96
4 clusters	0.23	0.86
5 clusters	0.22	0.89

Se puede apreciar la ventaja del Mak para evaluar la similaridad entre varias particiones. En la partición en 2 clusters, los dos métodos dan la misma clasificación y el Mak es igual a 1, máximo valor posible, mientras que el Ra es igual a 0.74 que no expresa la similaridad existente.

A partir de los resultados obtenidos, se puede concluir que la partición en 2 conglomerados es la más conveniente, aunque la partición en 3 clusters presenta también una similaridad elevada.

La hipótesis de "tres tipos" podría mantenerse, pero también podría pasarse, con alguna ventaja, a una de dos agrupamientos. Esto en cuanto a la estabilidad de grupos generados por métodos alternativos.

Las pruebas efectuadas sólo tienden a confirmar en este caso, la impresión que podía deducirse de inspeccionar la información contenida en el Cuadro No 6.

Ahora bien, si estos grupos son realmente "particiones naturales" de nuestro pequeño universo, tampoco debieran cambiar al agregarse variables tipificatorias a las cuatro originales.

Y disponemos de una hipótesis que sugiere que hay cuatro variables "de comportamiento" que se asocian a las estructurales.

En lo que sigue, se presentan nuevas exploraciones de estabilidad mediante el agregado de las cinco variables adicionales disponibles.(*)

En laguna medida, en lo que sigue jugarán simultáneamente la estabilidad relativa de los conglomerados y la hipótesis de asociación entre las variables identificadas como "estructurales" y las consideradas "de comportamiento".

3. Resultados de conglomerar con nueve variables

Habiendo quedado algunas dudas sobre la existencia de particiones naturales, se intentó obtener conclusiones más definitivas incrementando el número de atributos empleados para clasificar.

Los agrupamientos en función del total (nueve) variables disponibles se generaron por medio del algoritmo de Ward. Los resultados se compararon con los obtenidos empleando Sparks con cuatro variables. Para el caso de cuatro variables, analizado en las secciones precedentes de este capítulo, se disponía de muchos resultados alternativos. Se eligieron los de Sparks, porque (para un número dado de particiones) generaron no mayor intravarianza que los de Ward, como se comprueba en el Cuadro No 6.

Se dispuso así de la posibilidad de comparar agrupamientos diferentes entre sí, tanto con respecto al algoritmo que los generó como con referencia al número de variables empleadas (aunque cuatro de ellas fueron comunes a ambas pruebas).

Los resultados que se obtuvieron con Ward sobre las nueve variables, comparados con los agrupamientos resultantes de Sparks se resumen en el Cuadro No 10.

(*) Recuérdese que existe una alta correlación entre tamaño de potreros y superficie, lo que de alguna manera implica ponderar a "tamaño" el doble de lo que se pondera a otras variables.

CUADRO No 10: COMPARACION DE CLUSTERS ALTERNATIVOS WARD (9 VARIABLES) Y SPARKS (4 VARIABLES)

PARTICION	METODO	NUMERO DE ELEMENTOS EN CADA CLUSTER			
		I	II	III	IV
2 CLUSTERS	SPARKS	40	7	-	-
	WARD	40	7	-	-
3 CLUSTERS	SPARKS	26	14	7	-
	WARD	28	12	7	-
4 CLUSTERS	SPARKS	17	14	10	6
	WARD	19	12	9	7

Aparentemente las clasificaciones son similares porque los grupos tienen un número parecido de elementos. En realidad, para las particiones en 3 y 4 clusters los resultados son muy diferentes por la asignación que hacen de las empresas. (Ver Anexo 2)

El análisis de la similaridad de las particiones se realizó mediante tablas de contingencia e índices de similaridad. Un resumen de los resultados obtenidos con las tablas de contingencia se precenta en el Cuadro No 11.

CUADRO No 11: PRUEBAS DE X^2 y C PARA AGRUPAMIENTOS ALTERNATIVOS WARD (9 VARIABLES) Y SPARKS (4 VARIABLES)

PARTICION	TABLAS DE CONTINGENCIA		ESTADIGRAFOS DE CONTINGENCIA	
	X^2 CALCULADO	X^2 TABLAS	C CALCULADO	C MAXIMO
2 CLUSTERS	47.12	6.63	0.71	0.71
3 CLUSTERS	57.96	13.27	0.74	0.82
4 CLUSTERS	77.07	21.66	0.79	0.87

Analizando los resultados, se debe rechazar la hipótesis nula de ausencia de asociación para las tres particiones. El estadígrafo de contingencia revela asociación perfecta entre las clasificaciones obtenidas por los dos métodos en la partición en dos conglomerados. Para las otras particiones, el grado de asociación es alto, pero en niveles inferiores a los obtenidos cuando se compararon los grupos basados exclusivamente en las cuatro variables estructurales (compárese con Cuadro No 8).

Los resultados obtenidos con los índices de similaridad se presentan en el Cuadro No. 12.

CUADRO No 12: INDICES DE PARES SIMILARES PARA PARTICIONES ALTERNATIVAS (WARD, 9 VARIABLES Y SPARKS, 4 VARIABLES)

PARTICION	Ra	Mak
2 clusters	0.74	1.00
3 clusters	0.25	0.60
4 clusters	0.15	0.57

La similaridad de las particiones en dos conglomerados es perfecta. Pero ella baja bruscamente para las particiones en 3 y 4 clusters, denotando un grado de similaridad bajo.

De acuerdo con los análisis realizados, se puede concluir que en la población analizada existen 2 grupos "naturales" de empresas. Esto da idea de los riesgos que se podrían correr si se hubiera aceptado la hipótesis planteada a-priori luego de la primer corrida, sin hacer un análisis bastante más detallado del problema.

A partir de la partición en 2 conglomerados tal vez sí puedan definirse las "empresas tipo" de la región. Los valores promedio de los grupos formados se presentan en el Cuadro No 12.

CUADRO No 12: VALORES MEDIOS DE LAS 9 VARIABLES PARA UNA PARTICION EN DOS GRUPOS

A T R I B U T O S	EMPRESAS TIPO I	EMPRESAS TIPO II
Tamaño (hectáreas)	287.11	1736.07
Mano de obra (hectáreas por trabajador)	127.48	342.15
Mano de obra familiar (porcentaje)	76	32
Mecanización (hectáreas por tractor)	2366.23	2871.63
Tamaño de los potreros (hectáreas)	78.08	201.26
Relación Agrícola Ganadera (porcentaje)	3.62	0.83
Superficie mejorada (porcentaje)	5.88	3.96
Relación ovino-bovino	1.91	1.45
Ganado lechero (porcentaje)	2.84	0.68

Con los valores promedios de cada grupo, se podría intentar definir las "empresas tipo" de la siguiente forma:

Tipo I: Empresas chicas, con abundante mano de obra familiar, dedicadas a la explotación pecuaria, con algo de agricultura y lechería.

Tipo II: Empresas muy grandes, con mano de obra asalariada, extensivas, de producción ganadera de carne y lana.

La "representatividad" de las "empresas tipo" medida en porcentaje del total de vacunos del área es del 76% para la Tipo I y 24% para la Tipo II.

Las empresas tipo I surgen entonces como representativas de la zona; además, ocupan el 77% del área y poseen el 81% de los ovinos.

Al considerar los resultados obtenidos se debe tener presente que los dos grupos retenidos no son muy homogéneos, en la medida en que para la partición elegida la varianza no explicada llega al 56%, ya que en la distribución de la varianza total sólo el 44% corresponde a la intervarianza.

Es decir que, aún después de todo este proceso, no tenemos resultados muy definitivos en cuanto a la posibilidad de elegir el caso representativo que se buscaba..Pero algunos errores serios, aunque más no sea por un mejor conocimiento del universo, podrían evitarse a esta altura del proceso. Si la estabilidad es crucial, este universo tiene dos grupos.

Es difícil ser más concluyente que esto en propuesta de conclusiones, porque no existe una idea muy clara del problema. La advertencia metodológica está ejemplificada con un caso real. Y es aún más difícil de obtener de un toy-problem como éste, planteado sólo para obtener similaridad con requisitos del PROTAAL.

Alguna información adicional sobre la utilidad de las diversas clasificaciones producidas puede obtenerse mediante aplicaciones de análisis discriminante. Ese será el tema del próximo capítulo.

VI. ANALISIS A POSTERIORI DE LAS CLASIFICACIONES OBTENIDAS

Habiéndose evaluado la similaridad de las particiones obtenidas y elegido la partición en 2 conglomerados para definir las "empresas tipo" de la región, se procedió luego a probar la calidad de la clasificación, mediante la utilización de la técnica de análisis discriminante.

1. Análisis discriminante

El análisis discriminante es una técnica que permite describir y clasificar elementos representados por un número elevado de variables.

Como indican Lebart y Fenelon (6) "el problema que se propone resolver el A.D. es el de buscar entre todas las combinaciones lineales de las variables, las que tengan una varianza entre ellas máxima (a fin de exaltar las diferencias entre clases) y una intravarianza mínima (de modo que las clases estén bien delimitadas). Estas combinaciones lineales son las funciones discriminantes".

Existe evidentemente una semejanza entre el análisis discriminante y el análisis de conglomeración, pero se debe tener presente que el primero se debe aplicar sobre clases previamente definidas mientras que el segundo se utiliza para construir dichas clases.

Para docimar la hipótesis de que las funciones discriminantes pueden haber surgido al azar, se computa el estadígrafo D^2 o distancia generalizada de Mahalanobis. Este puede ser usado, asumiendo normalidad en la distribución de las variables, como ji cuadrado con m(g-1) grados de libertad, donde m es el número de variables y g es el número de clases.

2. Aplicaciones y Resultados del Análisis Discriminante

Se efectuaron seis aplicaciones de Análisis Discriminante. Su naturaleza, en términos de número de grupos y variables utilizadas, se codifica en el Cuadro No 13.

CUADRO No 13: CODIGO IDENTIFICATORIO DE PRUEBAS DE ANALISIS DISCRIMINANTE

No DE CLUSTERS DE SPARKS (con cuatro variables estructurales)	No Y TIPO DE VARIABLES EN ANALISIS DISCRIMINANTE		
	4 de Comportamiento	4 Estructurales	9 Total de Variables
2 *	A	B	C
3	D	-	-
4	E	-	-
5	F	-	-

* Los dos grupos son idénticos a los constituídos con Ward, empleando el total de nueve variedades.

El resultado de estas aplicaciones, indicado por D^2, su significación estadística o falta de ella y el porcentaje de predios mal clasificados, se presenta en el Cuadro No. 14.

CUADRO No 14: RESULTADOS DE LAS SEIS PRUEBAS DE ANALISIS DISCRIMINANTE

CODIGO IDENTIFICATIVO DE LA PRUEBA	VALOR DEL D^2	SIGNIFICATIVO?	PORCENTAJE DE PREDIOS MAL CLASIFICADOS
A	13.36	SI	29
B	119.61	SI	0
C	169.73	SI	0
D	35.42	SI	36
E	57.50	SI	34
F	59.25	SI	40

Las cuatro variables estructurales se emplearon para agrupar y se ha visto que el agrupamiento en dos clusters es estable aunque el número de variables se amplíe al total de las nueve disponibles. Por lo tanto, los casos B y C son sólo pruebas de que los algoritmos de clustering hicieron lo que se suponía debían hacer.

Las pruebas A, D, E y F son más interesantes. Todas ellas emplean sólo las 4 variables de comportamiento, no usadas para constituir los grupos (al menos, no al inicio). Estas cuatro aplicaciones permitirían juzgar la bondad de los agrupamientos, de ser válida la hipótesis relacional entre variables estructurales y de comportamiento.

En todos los casos, la prueba de D^2 obligaría a rechazar la hipótesis de que los grupos surgieron al azar. Pero, también en todos los casos, surgen porcentajes de mala clasificación no inferiores al 29%. El caso A (con los dos clusters naturales) es el que mejor resulta, aunque no puede aceptarse calurosamente como no rechazando las hipótesis involucradas.

Aparentemente, la hipótesis de asociación entre los observables elegidos para estructura y comportamiento no es buena. O, lo que es lo mismo, de haberse constituído los clusters con las variables 6 a 9, se hubieran generado distintos grupos. Este tipo

de situaciones requiere hacer pesar la teoría que debe guiar el proceso de tipificación, para que ella no sea una operación sin guía. En el caso que estamos presentando la poca teoría que se intentó (estructurales y de comportamiento) no se presentó más que como ejemplo. Por ende, conviene no intentar conclusiones definitivas.

VII. RESUMEN Y CONCLUSIONES

En base a un trabajo en proceso para un Proyecto de Desarrollo Regional en Uruguay, se presentan algunos ejemplos que pueden aportar a la Metodología del PROTAAL en la identificación de empresas representativas.

El trabajo original, cuyos datos se han empleado en este documento de discusión, se refiere a agrupación de sectores censales y tiene objetivos distintos a los del PROTAAL. A efectos de facilitar la comprensión del presente escrito, se consideran "empresas" a cada uno de estos sectores y se suponen objetivos similares a los de PROTAAL para la clasificación de empresas. Esta investigación tiene así mucho de forzada, pese a lo cual se espera transmita el mensaje de advertencia metodológica que pretende contener.

Dicho mensaje metodológico se refiere a dos aspectos fundamentales, vinculados entre sí:

1. partir con definiciones a-priorísticas de tipos que se autojustifiquen con cualquier evidencia, que sean irrefutables, no parece ser una aplicación científica válida, y

2. los algoritmos de que disponemos en IICA para constituir conglomerados, y de allí definir "empresas representativas", pueden generar grupos sólo aparentes y, por lo tanto, inestables ante cualquier cambio razonable de algoritmo o variables.

En este documento se definen variables estructurales y de comportamiento para un reducido universo de 47 predios. Con las estructurales primero, y con el total de variables disponibles más adelante, se van probando distintos algoritmos, docimándose la estabilidad de los agrupamientos que van resultando. Distintas pruebas van llevando a la conclusión de que, cualquiera sea el juicio inicial del que se parta,

este pequeño universo tiene sólo dos grupos "naturales", porque ellos son estables ante diversos cambios posibles. Todos los otros agrupamientos que puedan constituirse son arbitrios, tal vez operativamente útiles, pero no fáciles de justificar. La estabilidad en este sentido tiene una definición precisa. Otras propiedades (tal vez competitivas) pueden exigirse, y ellas pueden ser muy razonables, pero habrá que definirlas con precisión antes de efectuarse el trabajo..No parece haber regla más sencilla para evitar inconscientes trampas en el manejo de datos.

Algunas pruebas de Análisis Discriminante, que se efectuaron para ayudar a definir la calidad de los agrupamientos, sólo sirvieron claramente para confirmaciones triviales (del tipo "las técnicas de conglomeración hicieron lo que tenían que hacer"). Cuando se buscaron resultados más útiles, por ejemplo aplicando A.D.. con sólo variables de comportamiento (no usadas para conglomerar), apenas se verificó una leve tendencia a confirmar que existen efectivamente dos grupos naturales. De más importancia, podría deducirse que no se verifica la hipótesis de asociación entre las variables elegidas para estructura y comportamiento.

Siendo esta una mera advertencia, con teoría y objetivos que no permiten precisar mucho sobre las tipologías que se analizan, no hay interpretación sustancial que valga la pena extraer sobre el pequeño universo analizado.

El proyecto PROTAAL sí tiene teoría y objetivos, lo que es requisito obvio para extraer conclusiones sobre las agrupaciones de empresas que intentará. Los métodos de conglomeración disponibles, y tal vez las pruebas y sugerencias de este documento, pueden ser útiles para ese proceso. Pero lo serán sólo en la medida en que las relaciones que postula entre conceptos no observables (tales como dominación, influencia, etc.) y los tipos de empresa que plantea, no surjan de definiciones incontrastables por evidencia alguna. Es decir, en la medida en que se empleen algunas reglas mínimas de análisis científico.

B I B L I O G R A F I A

(1) ALONSO, A. "Algunas técnicas de conglomeración..Su naturaleza y sus posibilidades en tipificación de empresas". Reunión Técnica sobre Tipificación de Empresas Agropecuarias. IICA- DIEA. Montevideo, mayo 1977.

(2) COHAN, H.E..y ALONSO, A. "Aplicación de técnicas estadísticas para tipificación de empresas agropecuarias". Coordinación del Plan de Acción del IICA en Uruguay. Montevideo, junio 1977.

(3) FERREIRA, P. "Algunos comentarios sobre evaluación de clusterings" Reunión Técnica sobre Tipificación de Empresas Agropecuarias. IICA - DIEA. Montevideo, mayo 1977.

(4) KAMINSKY, M. "Comentarios sobre procesos de tipificación y su validación". Reunión Técnica sobre Tipificación de Empresas Agropecuarias. IICA - DIEA. Montevideo, mayo 1977.

(5) LING, R.F. y KILIOUGH, G.G. "Probability Tables for Cluster Analysis Based on a Theory of Random Graphs", Journal of the American Statistical Association, Vol.71, No 354, 1976.

(6) LEBART y FENELON "Statistique et informatique apliqueés", Dunod, 1973.

Mimeog.. 121
HEC/lbf

ANEXO 2

DETALLE DE LOS AGRUPAMIENTOS GENERADOS

Nota: El ANEXO 1 fue eliminado por dificultades para reproducción.

2.1 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES. PARTICION EN 2 CONGLOMERADOS

CLUSTER	NUMERO DE ELEMENTOS	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	$n_I = 40$	1-2-3-4-5-6-7-8-9-13-14-15-16-17-19-20-21-23-24 26-27-28-29-30-32-33-34-35-36-37-38-39-40-41-42 43-44-45-46-47
II	$n_{II} = 7$	10-11-12-18-22-25-31

2.2 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES. PARTICION EN 3 CONGLOMERADOS

CLUSTER	NUMERO DE ELEMENTOS	METODO	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	$n_I = 26$	SPARKS	1-2-3-4-5-6-7-13-14-15-16-17-19-28-32 33-35-37-38-41-42-43-44-45-46-47
	$n_I = 27$	WARD	1-2-3-4-5-6-7-8-13-14-15-16-17-19-28-32 33-35-38-41-42-43-44-45-46-47-37
II	$n_{II} = 14$	SPARKS	8-9-20-21-23-24-26-27-30-34-36-39-40-29
	$n_{II} = 13$	WARD	9-20-21-23-24-26-27-29-30-34-36-39-40
III	$n_{III} = 7$	SPARKS	10-11-12-18-22-25-31
	$n_{III} = 7$	WARD	10-11-12-18-22-25-31

2.3 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES. PARTICION EN 4 CONGLOMERADOS

CLUSTER	NUMERO DE ELEMENTOS	METODO	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	n_I = 17	SPARKS	1-2-3-4-14-15-17-28-33-35-37-38-41-42 43-44-45
	n_I = 17	WARD	1-2-3-4-14-15-17-28-33-35-37-38-41-42 43-44-45
II	n_{II} = 10	SPARKS	9-21-23-24-26-27-34-36-39-40
	n_{II} = 13	WARD	9-20-21-23-24-26-27-29-30-34-36-40-39
III	n_{III}= 14	SPARKS	5-6-7-8-13-16-19-20-25-29-30-32-46-47
	n_{III}= 10	WARD	5-6-7-8-13-16-19-32-46-47
IV	n_{IV} = 6	SPARKS	10-11-12-18-22-31
	n_{IV} = 7	WARD	10-11-12-18-22-25-31

2.4 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES. PARTICION EN 5 CONGLOMERADOS

CLUSTER	NUMERO DE ELEMENTOS	METODO	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	$n_I = 17$	SPARKS	1-2-3-4-14-15-17-28-33-35-37-38-41-42 43-44-45
	$n_I = 17$	WARD	1-2-3-4-14-15-17-28-33-35-37-38-41-42 43-44-45
II	$n_{II} = 13$	SPARKS	5-6-7-8-13-16-19-20-29-30-32-46-47
	$n_{II} = 10$	WARD	5-6-7-8-13-16-19-32-46-47
III	$n_{III} = 10$	SPARKS	9-21-23-24-26-27-34-36-39-40
	$n_{III} = 13$	WARD	9-20-21-23-24-26-29-30-34-36-39-40-27
IV	$n_{IV} = 4$	SPARKS	10-11-25-31
	$n_{IV} = 4$	WARD	10-11-25-31
V	$n_V = 3$	SPARKS	12-18-22
	$n_V = 3$	WARD	12-18-22

2.5 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR EL METODO DE WARD CON 9 VARIABLES. PARTICION EN 3 CONGLOMERADOS.

CLUSTER	NUMERO DE ELEMENTOS	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	$n_I = 12$	4-5-14-15-16-17-32-33-43-44-45-46
II	$n_{II} = 28$	1-2-3-6-7-8-9-13-19-20-21-23-24-26-27-28 29-30-34-35-36-37-38-39-40-41-42-47
III	$n_{III} = 7$	10-11-12-18-22-25-31

2.6 RESULTADOS DE LAS CLASIFICACIONES OBTENIDAS POR EL METODO DE WARD CON 9 VARIABLES. PARTICION EN 4 CONGLOMERADOS.

CLUSTER	NUMERO DE ELEMENTOS	ELEMENTOS QUE COMPONEN CADA CLUSTER
I	$n_I = 12$	4-5-14-15-16-17-32-33-43-44-45-46
II	$n_{II} = 9$	1-2-3-28-35-37-38-41-42
III	$n_{III} = 19$	6-7-8-9-13-19-20-21-23-24-26-27-29-30-34 36-39-40-47
IV	$n_{IV} = 7$	10-11-12-18-22-25-31

A N E X O 3

ANALISIS DE SIMILARIDAD ENTRE PARTICIONES

* TABLAS DE CONTINGENCIA
* INDICES DE SIMILARIDAD

3.1 COMPARACION DE LAS PARTICIONES EN DOS CONGLOMERADOS OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES

3.1.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	f.m.
I	40 (34.04)	0 (5.96)	40
II	0 (5.96)	7 (1.04)	7
f.m.	40	7	47

$T = 47.12$

X^2 (1 g.l., 0.01) = 6.63

$$C = \sqrt{\frac{47.12}{94.12}} \quad \therefore \quad C = 0.71$$

$$\max C = \sqrt{\frac{1}{2}} \quad \therefore \quad \max C = 0.71$$

3.1.2. INDICES DE SIMILARIDAD

$$Ra = \frac{801}{1081} = 0.74$$

$$Mak = \left(\frac{801}{801} + \frac{801}{801} \right) \cdot / \cdot 2 = 1.00$$

3.2 COMPARACION DE LAS PARTICIONES EN TRES CONGLOMERADOS OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES

3.2.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	III	f.m.
I	26 (14.94	0 (7.19)	0 (3.87)	26
II	1 (8.04)	13 (3.87)	0 (2.09)	14
III	0 (4.02)	0 (1.94)	7 (1.04)	7
f.m.	27	13	7	47

$T = 89.16$

X^2 (4 g.l.; 0.01) = 13.27

$$C = \sqrt{\frac{89.16}{136.16}} \quad \therefore \quad C = 0.81$$

$$\max C = \sqrt{\frac{2}{3}} \quad \therefore \quad \max C = 0.82$$

3.2.2. INDICES DE SIMILARIDAD

$$Ra = \frac{424}{1081} = 0.39$$

$$Mak = \left(\frac{424}{437} + \frac{424}{450} \right) \cdot / \cdot 2 = 0.96$$

3.3. COMPARACION DE LAS PARTICIONES EN CUATRO CONGLOMERADOS OBTENIDAS POR LOS METODOS DE SPARKS Y DE WARD CON 4 VARIABLES.

3.3.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	III	IV	f.m.
I	17 (6.15)	0 (4.70)	0 (3.62)	0 (2.53)	17
II	0 (3.62)	10 (2.77)	0 (2.13)	0 (1.49)	10
III	0 (5.06)	3 (3.87)	10 (2.98)	1 (2.09)	14
IV	0 (2.17)	0 (1.66)	0 (1.28)	6 (0.89)	6
f.m.	17	13	10	7	47

$T = 112.91$

X^2 (9 g.l.; 0.01) = 21.66

$$C = \sqrt{\frac{112.91}{159.91}} \quad \therefore \quad C = 0.84$$

$$\max C = \sqrt{\frac{3}{4}} \quad \therefore \quad \max C = 0.87$$

3.3.2. INDICES DE SIMILARIDAD

$$Ra = \frac{244}{1081} = 0.23$$

$$Mak = \left(\frac{244}{287} + \frac{244}{280} \right) \text{./.} \ 2 = 0.86$$

3.4 COMPARACION DE LAS PARTICIONES EN CINCO CONGLOMERADOS OBTENIDAS POR LOS METODOS DE SPARKS Y WARD CON 4 VARIABLES.

3.4.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	III	IV	V	f.m.
I	17 (6.15)	0 (3.62)	0 (4.70)	0 (1.45)	0 (1.09)	17
II	0 (4.[illegible])	10 (2.77)	3 (3.60)	0 (1.11)	0 (0.83)	13
III	0 (3.62)	0 (2.13)	10 (2.77)	0 (0.85)	0 (0.64)	10
IV	0 (1.45)	0 (0.85)	0 (1.11)	4 (0.34)	0 (0.26)	4
V	0 (1.09)	0 (0.64)	0 (0.83)	0 (0.26)	3 (0.19)	3
f.m.	17	10	13	4	3	47

$T = 169.17$

X^2 (16 g.l.; 0.01) = 32.00

$$C = \sqrt{\frac{169.17}{216.17}} \quad \therefore \quad C = 0.88$$

$$\max C = \sqrt{\frac{4}{5}} \quad \therefore \quad \max C = 0.89$$

3.4.2. INDICES DE SIMILARIDAD

$$Ra = \frac{238}{1081} = 0.22$$

$$Mak = \left(\frac{238}{268} + \frac{238}{268} \right) \div 2 = 0.89$$

3.5 COMPARACION DE LAS PARTICIONES EN 3 CONGLOMERADOS OBTENIDAS POR SPARKS CON 4 VARIABLES Y WARD CON 9 VARIABLES.

3.5.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	III	f.m.
I	14 (15.49)	12 (6.64)	0 (3.87)	26
II	14 (8.34)	0 (3.57)	0 (2.09)	14
III	0 (4.17)	0 (1.79)	7 (1.04)	7
f.m.	28	12	7	47

$T = 57.96$

X^2 (4 g.l.; 0.01) = 13.27

$$C = \sqrt{\frac{57.96}{104.96}} \quad \therefore \quad C = 0.74$$

$$\max C = \sqrt{\frac{2}{3}} \quad \therefore \quad \max C = 0.82$$

3.5.2. INDICES DE SIMILARIDAD

$$Ra = \frac{269}{1081} = 0.25$$

$$Mak = \left(\frac{269}{437} + \frac{269}{465} \right) \cdot / \cdot 2 = 0.60$$

3.6 COMPARACION DE LAS PARTICIONES EN 4 CONGLOMERADOS OBTENIDAS POR SPARKS CON 4 VARIABLES Y WARD CON 9 VARIABLES

3.6.1. TABLA DE CONTINGENCIA

SPARKS \ WARD	I	II	III	IV	f.m.
I	8 (4.34)	9 (3.26)	0 (6.87)	0 (2.53)	17
II	0 (2.56)	0 (1.91)	10 (4.04)	0 (1.49)	10
III	4 (3.57)	0 (2.68)	9 (5.66)	1 (2.09)	14
IV	0 (1.53)	0 (1.15)	0 (2.43)	6 (0.89)	6
f.m.	12	9	19	7	47

$T = 77.07$

X^2 (9 g.l.; 0.01) = 21.66

$$C = \sqrt{\frac{77.07}{124.07}} \quad \therefore \quad C = 0.79$$

$$\max C = \sqrt{\frac{3}{4}} \quad \therefore \quad \max C = 0.87$$

3.6.2 INDICES DE SIMILARIDAD

$$Ra = \frac{166}{1081} = 0.15$$

$$Mak = \left(\frac{166}{287} + \frac{166}{294} \right) ./. \ 2 = 0.57$$