UNIVERSIDAD NACIONAL DE SALTA
Facultad de Ciencias Económicas,
Jurídicas y Sociales
Instituto de Investigaciones Económicas

Reunión de Discusión Nº 61
Fecha: 11/12/1991
Hs.: 17

TECNICAS DE CLUSTERING: UN EJERCICIO DE APLICACION

Juan Carlos Cid

1. INTRODUCCION

Los métodos de clustering sirven para clasificar un conjunto de elementos en grupos de acuerdo a su semejanza.

Distintos procedimientos de clustering se incluyen en un paquete de utilitarios de estadística que se denomina SAS y que está disponible en la Dirección de Estadísticas y Censos, en el Centro de Cómputos de la Universidad y en la Estación Experimental de Cerrillos del INTA.

El autor del presente informe se interesó en el tema por la posible aplicación de esta técnica a datos individuales del reciente censo agropecuario, de modo de agrupar a las distintas explotaciones por la semejanza que mostraran en ciertas variables (tamaño, uso del suelo, régimen de tenencia, etc.) . El resultado del agrupamiento permitiría entonces diseñar una tipología de empresas del sector.

Pero luego se pensó que estas técnicas resultarían también aplicables y útiles en otros problemas en los que, hasta donde se tuvo conocimiento, no se habían utilizado anteriormente en nuestro medio.

El objetivo de nuestro trabajo fue plantear y discutir esas posibles aplicaciones. Los temas investigados fueron: i) la definición de pequeña y mediana empresa industrial (PYME) ; ii) la delimitación de estratos de población y viviendas para una encuesta de hogares; y iii) el análisis del rendimiento de egresados de nivel secundario en una carrera universitaria.

Señalemos como una importante limitación que hemos preferido enfatizar el esquema general y su aplicación a cada caso particular, antes que arribar a resultados y conclusiones definitivos. Y como el tiempo de computación aumenta considerablemente al incrementarse el número de observaciones, se ensayó el método con muestras de datos cuya selección se explicará en cada caso y no con poblaciones completas.

Aun así, el volumen del material a procesar y la cantidad de cuadros de resultados obtenidos, tornaron conveniente dividir nuestro escrito para adecuarlo a una extensión recomendable para una reunión de discusión. Es por eso que en este informe, luego de referirnos resumidamente al esquema conceptual del análisis cluster (sección 2) , se discute exclusivamente la aplicación del enfoque al primero de los

temas (sección 3), incluyendo en un anexo algunos resultados que parecieron interesantes. Además, en la sección 4 se adelanta la metodología que se estuvo aplicando en los otros dos problemas, en el convencimiento de que los comentarios y observaciones que se nos hagan en la reunión de discusión servirán para mejorar la calidad del trabajo. Queda para un próximo informe la presentación de los resultados en detalle para esos otros casos.

2. TECNICAS DE CLUSTERING(1)

El problema consiste en que, dado un lote de **n** individuos caracterizados cada uno de ellos por los valores que asumen **m** características o variables, se pretende clasificarlos en grupos (clusters) en base a su semejanza o similaridad. Los grupos no están previamente definidos, sino que precisamente se quiere establecerlos a partir de los datos.

La finalidad principal de un clustering sería mostrar y resumir cierta información en una forma organizada. Pero cuando uno halla que un objeto de un cluster tiene determinada propiedad, espera que otros objetos del mismo cluster posean la misma característica. De allí al intento de una explicación de su existencia y la formulación de una teoría hay un corto trecho.

Generalmente los distintos métodos existentes requieren la construcción de una matriz **nxn** de distancias entre los individuos. El segundo paso consiste en la formación de los grupos o clusters por medio de un procedimiento algorítmico.

Frecuentemente las variables se estandarizan o se aplica algún otro procedimiento para que todas tengan el mismo orden de magnitud (consiguientemente, el mismo peso en la determinación de las distancias entre individuos)

A partir de esa base común, puede establecerse una primera gran división, entre clustering jerárquico y no jerárquico.

a) Agrupamientos jerárquicos: son aquéllos conformados de tal manera que un cluster puede estar completamente incluido dentro de otro (el primero es entonces un subconjunto del segundo) y ninguna otra clase de unión o conjunción es admitida. En otros términos, dados dos clusters, o no tienen ningún elemento en común (son disjuntos) o necesariamente uno de ellos es una parte del otro. Un ejemplo de ordenamiento jerárquico lo dan las ciencias naturales, al agrupar distintos individuos en especies, a éstas en géneros, y luego sucesivamente se conforman familias, órdenes y clases.

Otro ejemplo que puede resultar interesante se plantea en el manual de usuario de SAS. Consiste en efectuar un agrupamiento clustering de 10 ciudades estadounidenses teniendo en cuenta las distancias entre ellas (las celdillas indican la distancia en millas)

	AT	CH	DE	HO	LA	MI	NY	SF	SE	WA
ATLANTA	0	587	1212	701	1936	604	748	2139	2182	543
CHICAGO		0	920	940	1745	1188	713	1858	1737	597
DENVER			0	879	831	1726	1631	949	1021	1494
HOUSTON				0	1374	968	1420	1645	1891	1220
LOS ANGELES					0	2339	2451	347	959	2300
MIAMI						0	1092	2594	2734	923
NEW YORK							0	2571	2408	205
SAN FRANCISCO								0	678	2442
SEATTLE									0	2329
WASHINGTON										0

---

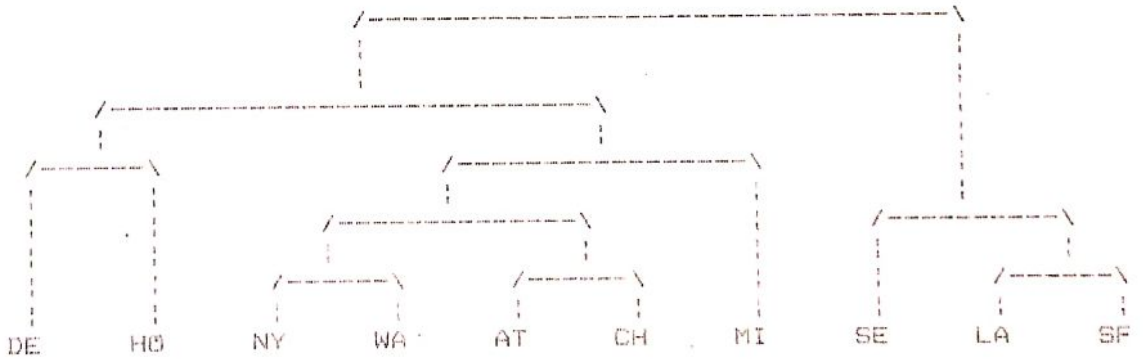
(1) Esta sección se limita a resumir los capítulos pertinentes de Romero Villafranca y SAS/STAT.

Uno de los posibles métodos de clustering efectuó este agrupamiento:

Nombre del cluster	Clusters (o individuos) unidos		Frecuencia del nuevo cluster
9	New York	Washington	2
8	Los Angeles	San Francisco	2
7	Atlanta	Chicago	2
6	cluster 7	cluster 9	4
5	cluster 8	Seattle	3
4	Denver	Houston	2
3	cluster 6	Miami	5
2	cluster 3	cluster 4	7
1	cluster 2	cluster 5	10

El procedimiento arrancó de los 10 individuos -las 10 ciudades- y fue agrupando sucesivamente a individuos y/o conjuntos de individuos en nuevos conjuntos mayores, tomando en cuenta su relativa cercanía. El último cluster agrupa a la totalidad de los individuos.

El ordenamiento jerárquico puede representarse con un esquema llamado dendograma (también denominado por algunos autores fenograma o diagrama de árbol, y tree en los programas de SAS) . En nuestro ejemplo, resulta:



Observesé que en este ejemplo se partió de una matriz de distancias entre los individuos a agrupar, lo que en otras circunstancias puede ser el resultado de una primera etapa a cumplirse. Es decir que la base de datos originales habría correspondido al conjunto de los pares de coordenadas geográficas de cada una de las ciudades.

Queremos insistir antes de continuar, en que existen diferentes criterios de clustering. El agrupamiento de las 10 localidades transcripto es el efectuado por una de esas técnicas y el resultado podrá diferir en caso de aplicarse otra (2) . Precisamente, una forma de testeo para comprobar si los clusters son "reales" y están bien definidos consiste en aplicar métodos alternativos para comparar los resultados. En el ejemplo de las 10 ciudades los distintos procedimientos coinciden en sugerir una división en dos clusters a lo largo de un eje este-oeste (es decir que se reúnen por un lado las ciudades

---

(2) Los métodos de clustering jerárquico disponibles en SAS son once. La diferencia entre ellos estriba en la definición que adoptan de la distancia entre dos clusters.

ubicadas sobre la costa del Atlántico y por el otro las que están cerca del Pacífico) . Algunos de los métodos indican un posible tercer grupo conteniendo a Denver y Houston.

Los algoritmos de clasificación jerárquica pueden ser aglomerativos (van fusionando clusters para obtener particiones cada vez menos finas, hasta llegar a un único cluster final que engloba la totalidad de los individuos, es el caso del ejemplo y también de los agrupamientos que se muestran en el anexo) o divisivos (arrancan del conjunto universal y van particionándolo progresivamente) .

Además, de acuerdo a las variables o características consideradas, los algoritmos pueden ser politéticos (donde las fusiones o particiones toman en cuenta todas las variables) o monotéticos (cada fusión o partición se basa en una sola variable) .

b) Agrupamientos no jerárquicos: no admiten ningún tipo de solape o superposición entre clusters, de modo que cada individuo se ubica o clasifica en uno y sólo uno de los clusters (es por ello que a esta clase de agrupamiento tambien se la llama de clusters disjuntos) . Los métodos de clustering no jerárquico requieren que uno decida previamente el número de grupos a formar y efectúan la partición mediante un algoritmo de modo de satisfacer cierto criterio de optimalidad (por ejemplo, minimizar la varianza intra clases)

Los procedimientos de clustering en grupos disjuntos son recomendados cuando se trabaja con un gran número de individuos, digamos 10.000 observaciones. Se establece entonces un agrupamiento en un número fácilmente manejable de clusters, por ejemplo 50. Luego se pueden tomar esos 50 clusters como si fueran observaciones o individuos -en vez del archivo entero- y agruparlos en un clustering jerárquico. El objetivo principal de esta secuencia es economizar el tiempo de computación, que resulta aproximadamente proporcional al número de observaciones en el caso de clustering no jerárquico y al cuadrado o al cubo del número de observaciones en el caso de jerárquico.

El paquete SAS contiene un método de agrupamiento en clusters disjuntos que puede usarse en combinación con un procedimiento jerárquico.

Por último, en el manual de usuario del SAS también se mencionan los clustering con solape y los fuzzy clusters, como otras dos posibilidades aparte de los jerárquicos y disjuntos. No se hallaron mayores referencias a estos ordenamientos en la bibliografía consultada.

3. La pequeña y mediana empresa en la industria manufacturera.

La definición de establecimiento pequeño y mediano en la industria manufacturera (PYME) se basó tradicionalmente en el número de ocupados. Aunque los autores de distintos países difieren significativamente en el nivel propuesto de corte entre las PYME y los establecimientos considerados grandes, existió siempre relativo consenso sobre las características que distinguirían a aquel tipo de empresas (3). Las PYME se asocian con un reducido tamaño de planta, baja dotación de capital por hombre ocupado, escasa eficiencia, utilización de procesos productivos no complejos, independencia respecto a grandes conglomerados económicos, etc.

Pero como consecuencia de los avances tecnológicos registrados en la actividad manufacturera, se observó en los últimos años la aparición de ciertos casos en los que la aplicación del criterio del número de ocupados conduciría a error: se trata de establecimientos de reducido tamaño ocupacional pero con alta dotación de capital por hombre ocupado y dedicados a procesos productivos complejos.

Algunos estudiosos plantearon entonces definiciones de PYME que atiendan a otras características -algunas de ellas ya mencionadas más arriba-. Por ejemplo, se ha hecho hincapié en la gestión personalizada y el escaso peso en el mercado. Pero posiblemente por la dificultad que ofrecen la cuantificación de estas variables y la escasez de datos estadísticos, frecuentemente esos mismos autores terminaron proponiendo definiciones operativas de tipo ocupacional.

Una objeción adicional es que aun cuando el nivel ocupacional fuera el criterio a utilizar, éste suele diferir según la actividad manufacturera de que se trate.

Yoguel y Gatto proponen para la industria argentina una combinación de valor de producción y ocupación para definir los estratos de tamaño económico. En base a resultados censales, definen como PYMEs a los establecimientos con un valor de producción entre 50.000 dólares y 3,5 millones y que tienen más de 5 personas ocupadas:

	Número de ocupados	Valor bruto de producción
Grandes	-	mayor que 3,5 MUS$
PYMEs	mayor que 5	entre 50.000 US$ y 3,5 MUS$
Microindustrias	menor o igual que 5	
	(o mayor pero con...	menos de 50.000 US$)

Dentro de este rango tan amplio quedan comprendidas unidades productivas muy heterogéneas. Por ello los autores proponen intervalos de

---

(3) A lo largo de esta discusión no diferenciaremos entre establecimiento y empresa o firma. Esta pérdida de precisión se justifica en la ampliamente difundida terminología de "PYME" por un lado y la disponibilidad de datos censales referidos exclusivamente a establecimientos individuales por el otro.

menor amplitud para dividir las PYMEs en los siguientes subestratos:

i) establecimientos **medianos:** su valor de producción supera los 700.000 US$ y al menos el 75 % de ocupados son asalariados.
ii) establecimientos **intermedios:** el valor de producción está entre 120 y 700 mil dólares anuales.
iii) establecimientos **pequeños:** su producción está comprendida entre US$ 50.000 y 120.000 anuales.

Las firmas del primer subestrato tienen una organización formal parecida a las de las grandes, con ellas comparten mercados y suelen estar vinculadas por encadenamientos productivos, se dedican a bienes intermedios y operan en una dimensión espacial relativamente importante. Las plantas del subestrato inferior poseen menor nivel organizativo (son generalmente empresas unipersonales o sociedades de hecho) y producen bienes-salario en procesos de escasa complejidad, para mercados limitados geográficamente. Subsiste un alto grado de heterogeneidad en los establecimientos intermedios: algunas plantas se aproximan por sus características a las medianas, y otras a las pequeñas.

Yoguel y Gatto probaron contraponer esta clasificación que llamaron económica con otra que diera prioridad al tamaño ocupacional. La intersección de ambos criterios permite comprobar la existencia y relevancia de plantas grandes desde el punto de vista económico (es decir, con un valor de producción superior a US$ 3.500.000) y que poseen menos de 150 ocupados. Este tipo de establecimientos representa aproximadamente el 20 % del valor de producción total del estrato de establecimientos grandes y si se incluyen los que tienen hasta 200 ocupados, se llega casi al 28 %. Teniendo en cuenta que en nuestro país es habitual considerar PYMEs a las firmas con hasta 200 personas ocupadas, concluyen que una política estatal de promoción de ese tipo de establecimientos que se fundamente en el criterio ocupacional habitual, puede acabar beneficiando a empresas en verdad grandes (al menos, en términos de la clasificación económica de Yoguel y Gatto) .

Esta clasificación corresponde al conjunto de los establecimientos industriales del país. Para observar el resultado de la aplicación del doble criterio, económico y ocupacional, con mayor desagregación, los autores citados analizaron a nivel de rama o subgrupo (código CIIU a 5 dígitos) el límite de tamaño. El corte se fijó eliminando en cada rama y dentro del subgrupo de firmas grandes (más de US$ 3,5 millones de producción) a aquéllas con mayor ocupación, hasta alcanzar el 80 % del valor de producción del subgrupo. El límite así obtenido variaba según la actividad. A modo de ejemplo, el 80 % del valor de producción de las panaderías grandes se genera en plantas con más de 75 ocupados y en las refinerías de azúcar el mismo porcentaje se alcanza en fábricas con más de 200 ocupados. Esto equivale a decir que el máximo de tamaño ocupacional hasta el cual una planta debiera considerarse una PYME variará de rama en rama. Un problema similar puede presentarse en la cota inferior, entre las PYMEs pequeñas y las microindustrias.

Llegados a este punto, surge una pregunta: ¿por qué el límite arbitrario de 3,5 millones y no un millón o cualquier otro valor para discriminar entre grandes y medianas empresas ? ¿cuál es el fundamento de la fijación de tramos ocupacionales del tipo "hasta 200 ocupados" ? Aparte del atractivo que suelen ejercer los números redondos en cientos y miles cuando se quieren establecer cortes, nada nos asegura que la situación real no sea que la diferencia esencial en lo que hace a procesos productivos, organización, productividad, etc, existe entre las plantas por abajo de 2.950.422 dólares de producción y otras que superan esa cifra.

Como vimos antes, la ventaja del procedimiento clustering es precisamente que permite organizar la información sin categorías o clasificaciones apriorísticas.

Los datos empleados para intentar un ejercicio de agrupamiento estuvieron referidos a 80 establecimientos industriales de la provincia encuestados en 1987. Del archivo disponible se seleccionaron las siguientes características:

a) grupo de actividad (código CIIU a 5 dígitos) (4)
b) cantidad de ocupados
c) potencia instalada en HP
d) cantidad de personal remunerado, medida en meses-hombre
e) valor total de remuneraciones
f) cantidad de energía eléctrica utilizada en MWH
g) valor del consumo de combustibles y lubricantes
h) valor de la producción

La cantidad de ocupados en el establecimiento y la potencia instalada están medidas al 30 de abril de 1987, mientras que el resto de las variables desde d) hasta h) se refiere al año 1986. El agrupamiento pretendió tomar en cuenta no solamente el número de ocupados o el valor de la producción. Está claro que la elección de las variables que se proponen para el clustering es un tema crucial. El investigador debiera disponer de una hipótesis con respecto a las características o variables con mayor poder discriminatorio en la población que se está manejando (5)

Los resultados obtenidos así como la base de datos utilizada se detallan en el anexo A. Se ensayó la aplicación de los métodos Ward, que minimiza la varianza intra-clusters en cada uno de los pasos del agrupamiento; unión por promedios (average linkage) , que define a la distancia entre clusters como un promedio de distancias entre pares de observaciones, una de cada grupo; y density linkage, que efectúa un encadenamiento no paramétrico, basándose en estimaciones de densidad. Además se trabajó con todas las variables y -alternativamente- con sólo dos, cantidad de ocupados y valor del producto. También se probó con estandarizar los valores de esas variables.

El cuadro 1 presenta los datos individuales originales del archivo usado en el trabajo. Por razones de espacio, varios de los nombres que encabezan las columnas están resumidos: remunera se refiere al total de remunerados durante el año; consumwh al consumo de energía eléctri-

---

(4) El código de actividad no constituye una variable para clustering sino una característica de cada observación que puede considerarse para comprobar si los individuos de un mismo cluster tienden a pertenecer mayoritariamente a cierta rama industrial.

(5) Una característica a menudo mencionada como relevante para diferenciar económicamente a las empresas es el grado de poder monopólico que tienen. Pero intentar medir con algún índice la concentración industrial a partir de resultados censales enfrenta múltiples limitaciones. Señalemos cuatro muy importantes: i) la escasa correspondencia entre el concepto teórico de mercado de un bien y la clasificación del código CIIU de las actividades productivas; ii) la falta de información sobre el alcance geográfico de cada mercado así como de la posible oferta de bienes importados; iii) las plantas multiproducto usualmente son ignoradas en los censos, adjudicándose la actividad a la producción principal; y iv) los relevamientos censales generalmente están referidos a plantas y no a empresas.

ca ,medido en MWh; combusti al valor de lubricantes y combustibles consumidos durante el año.

El primer clustering se realizó con solamente dos variables previamente estandarizadas, los ocupados y el valor del producto. Observesé en el cuadro 2 que en este caso son frecuentes los "ties", que indican situaciones en que existen dos o más pares de clusters con la misma distancia mínima, por lo que debe obviarse la indeterminación con algún procedimiento arbitrario.

Los cuadros 3 a 5 presentan los resultados con agrupamientos efectuados utilizando las 7 variables propuestas y con los métodos Ward y Average, en este último, con la variante de estandarizar los valores. En general puede observarse que, con algunas pequeñas diferencias, en los tres clusterings existen ciertos establecimientos que se incorporan al agrupamiento al final del proceso. Corresponden a outliers, es decir, individuos con datos fuera de rango. Se trata en todos los casos de los establecimientos industriales más grandes de la provincia: las dos refinerías de azúcar, la destilería de YPF, los acopios de tabaco y una fábrica de cigarrillos. Por ejemplo, comparando con el procesamiento por Ward para las 7 variables, el método average conserva a las observaciones 9, 21, 22, 23, 47 y 53, y reaparece dentro de la categoría de grandes la 27, que había detectado el Ward con sólo 2 variables, y pierde a las observaciones 24 y 54 al agruparlas junto con establecimientos más pequeños. En el procedimiento con average y variables estandarizadas, siguen figurando 9, 22, 23, 24 47 y 53 y además se recupera a la observación 24.

Pero lo más llamativo de todos estos procesos es que el agrupamiento no guarda similitud con el que presentáramos como ejemplo, a partir de las distancias entre 10 ciudades de EEUU. En vez de formarse grupos inicialmente pequeños que se van fundiendo en otros más grandes, obviamente no tan homogéneos, para arribar a una instancia previa al cluster raíz en que se tengan digamos 3 o 4 grupos, aquí surge un núcleo inicial de establecimientos relativamente pequeños que a medida que aumenta la distancia admitida (representada por Semipartial R-Squared en el Ward y por Normalized RMS Distance en el Average) va "capturando" poco a poco, prácticamente de a uno, a establecimientos no tan pequeños. En otros términos, imaginemos en un espacio de n dimensiones una nube de observaciones que posee una alta densidad en la región cercana al origen (en nuestro ejemplo de 7 variables descriptivas de las industrias, tendremos un sistema de 7 semiejes ortogonales; . A medida que nos alejamos de ese origen, la nube va perdiendo densidad, las observaciones aparecen cada vez más espaciadas y distantes entre sí. Cuando se pretende agruparlas por su homogeneidad, lo primero que surge es la semejanza de las que se situaban cerca del origen, a medida que el cluster inicial va creciendo, van cayendo dentro de él las observaciones más alejadas, que en rigor están más cerca de ese cluster que de la próxima observación "grande".

Cabe ahora señalar que uno de los supuestos establecidos en la mayoría de los procedimientos de clustering es que las poblaciones que forman los grupos se distribuyen normalmente (por eso se las llama multinormales) . Se estaría entonces violando ese supuesto en el caso de los 80 establecimientos industriales. Los cuadros 6 a 9 apuntaron a tratar de evitar ese inconveniente. El método de encadenamiento por densidad es especialmente recomendado para poblaciones con variables no multinormales, y los clustering con el Ward pero con trim del 10 y del 15 % omiten en el análisis esas proporciones de las observaciones con más baja estimación de densidad. Sin embargo admitamos que los resultados no mejoraron mucho. Observesé por ejemplo que con density

linkage en el cuadro 7, desde el nivel 56 de la jerarquía y hasta llegar al clúster raiz en el 1, se van incorporando las observaciones de una en una al grupo que se había formado.

En el gráfico 1 del Anexo A se incluyó un tree o dendograma, que corresponde al agrupamiento del cuadro 9, hecho por el método Ward con variables estandarizadas y dejando afuera a 12 puntos muy alejados de la nube (son los que se listan en la parte superior del gráfico) . Si bien los establecimientos grandes que quedaban remanentes se agrupan convenientemente (observaciones 1, 3, 6, 12, 13, 14, 16, 17, 19, 26, 48, 74 y 76) lo hacen al final del proceso, en el cuadro 9 puede verificarse que las distancias entre clusters aumentaron significativamente en esa etapa.

Otra de las técnicas disponibles en el paquete SAS es el análisis de componentes principales. El propósito es derivar un pequeño número de combinaciones lineales a partir de un grupo de variables, que retenga tanta información de las variables originales como sea posible. A menudo ese pequeño número de componentes principales puede usarse en lugar de las variables originales para graficar, en clustering, etc. El cuadro 10 del anexo A muestra resultados de dicho análisis para las observaciones de establecimientos industriales. Las componentes principales se listan en función de su importancia en el sentido de explicar la varianza de la población. El resultado parecería alentador porque las dos primeras (PRIN1 y PRIN2) explican más del 99 % de la varianza total. Pero los coeficientes de PRIN1, que denotan cómo se construye la primera combinación lineal, dan un peso casi excluyente a la variable original potencia. En el mismo sentido, los valores de la última fila de la matriz de covarianza son mucho más altos que el resto. Como las variables originales no se habían estandarizado, la potencia instalada adquiere demasiado peso.

Por último, se ensayó el análisis dejando de lado la variable potencia instalada y practicando el agrupamiento no con las variables originales sino con las dos primeras componentes principales (cuadro 11) . El método utilizado fue el average. Dado que el programa permite seleccionar una cantidad de clusters deseada como salida, es decir, definir un corte en determinada jerarquía del dendograma, se pidió listar la composición y características para 7 clusters. Este número es arbitrario, se eligió a partir de la historia del clustering del cuadro 12, haciendo un corte entre las distancias normalizadas de 0,429290 y 0,651473. El cuadro 13 presenta los resultados de este enfoque (6) . En un grupo se incluyen 65 establecimientos con medias de 29 ocupados (el valor máximo es 117) y de 786 miles de australes de producto, otro conjunto se compone de 8 industrias, con entre 125 y 302 ocupados, un promedio de 190 trabajadores, y 4,621 millones de australes de valor de producción. Obsérvese que en esta última variable no se da una separación tajante como ocurría con el empleo: el máximo dentro del grupo de los más pequeños asciende a 5,649 millones y el mínimo en el segundo es 1,518. Los 5 restantes clusters se forman con solamente uno o dos establecimientos grandes en cada uno.

---

(6) El cuadro tiene en realidad dos partes, en la primera se listan las observaciones que componen cada cluster y en la segunda están detallados por cluster el número de observaciones, el valor mínimo, el máximo, la media y el desvío estándar de cada variable.

4. Otras dos aplicaciones de la técnica de clustering

4.1 Estratificación de la población para la Encuesta Permanente de Hogares

En la ciudad de Salta se viene efectuando la Encuesta Permanente de Hogares (EPH) desde hace más de 10 años. Esta encuesta intenta captar algunas características ocupacionales de la población y además, conocer la estructura demográfica, situación habitacional, educación, migraciones e ingresos.

Por la naturaleza de las características que investiga, la encuesta toma como unidad de información al hogar, pero la unidad física que se lista para luego hacer el muestreo es la vivienda.

De acuerdo al tamaño del aglomerado urbano INDEC diseñó diferentes muestras. Para los aglomerados que en 1970 superaban las 50.000 viviendas se hizo un muestreo probabilístico en dos etapas de selección, eligiendo primero áreas llamadas unidades de primera etapa (UPE) con probabilidad proporcional a su tamaño, y dentro de ellas, una cantidad de viviendas dispersas. En las ciudades de menor tamaño, el muestreo fue en una sola etapa.

Independientemente del tipo de muestreo, se procedió a estratificar el marco muestral, utilizando como indicador el aspecto edilicio y urbanístico predominante.

En Salta existen 6 estratos, que fueron definidos en el inicio del trabajo, es decir en 1978:

alto
comercial
medio alto
medio
medio bajo
bajo

La estratificación es importante porque al delimitarse las unidades primarias UPE se trató de asimilarlas a radios censales, pero uno de los criterios fue que correspondieran a un solo estrato, de manera que cuando no resultaban homogéneas, se subdividieron. Además, para hacer la selección, las UPE se ordenaron según una serpentina geográfica dentro de cada estrato. En tercer lugar, la estratificación puede utilizarse para presentar los resultados de la encuesta desagregados por estrato. No ocurre así en Salta, claro que seguramente el tamaño de la muestra en nuestra ciudad no permitiría realizar estimaciones confiables para pequeñas subclases.

La delimitación de los estratos en la ciudad capital fue resultado de una serie de recorridas en vehículo que permitieron apreciar, obviamente que sin entrar en las viviendas, las características exteriores de la edificación, lo que la metodología de INDEC denominó el aspecto edilicio y urbanístico predominante.

Creemos que el reciente censo poblacional brinda una excelente oportunidad para diseñar una estratificación sobre fundamentos más objetivos y, en lo posible, cuantificables (además servirá para actualizar el marco muestral de la EPH)

Los censos poblacionales en nuestro país tradicionalmente no interrogan a los individuos respecto a sus ingresos. Pero áreas temáticas

tales como nivel de confort de la vivienda (materiales predominantes, existencia de sanitarios, antigüedad, etc) , educación, rol ocupacional y calificación, pueden servir para estratificar a los hogares en niveles socioeconómicos con un razonable margen de aproximación.

En el relevamiento del corriente año, al igual que ocurriera en 1980, se utilizaron dos cuestionarios distintos en Salta Capital: el 80 % de los hogares fue censado con un cuestionario básico B que contiene 20 preguntas (12 para el hogar y 8 individuales para cada uno de los integrantes) , mientras que un 20 % de la población respondió al cuestionario ampliado A : 16 preguntas para el hogar y hasta 28 para las personas (7), preguntas entre las que obviamente, también se encontraban las del básico.

Una primera cuestión a dilucidar es si conviene seleccionar de entre las preguntas contenidas en el cuestionario básico, aquéllas que permitan la caracterización socioeconómica de la población o si se justificará enriquecer el análisis con los temas contemplados en el ampliado. Por ejemplo: rama de actividad del establecimiento donde se desempeñan los activos, calificación laboral, cobertura con obra social, condición de alfabetismo.

Esta cuestión está en parte relacionada con la decisión acerca del individuo a agrupar en los clusters. En este sentido existen tres alternativas interesantes: hogares (aproximadamente 80.000 - 85.000 en Salta Capital) ; manzanas (unas 5.000) y radios censales (280)

En el caso de inclinarnos por los hogares, una ventaja es que la base de datos se obtiene directamente del archivo censal, sin demasiadas elaboraciones intermedias. Pero el elevado número de observaciones tornaría muy engorroso el clustering, demandando un considerable tiempo de procesamiento. Seguramente debería recurrirse al procedimiento de SAS que genera grupos disjuntos. Señalemos además que las variables a utilizar serían en su mayoría dicotómicas, por ejemplo: existencia o no de materiales precarios -piso de tierra, tabiques de chapa, etc.- en la vivienda.

Una variante de esta alternativa es considerar exclusivamente al 20 % de los hogares censados con el cuestionario ampliado: las observaciones a agrupar se reducen a menos de 20.000 y se pueden agregar otras variables (como ya se mencionó más arriba, preguntas del cuestionario A que no están en el B) . Subsiste el carácter de dicotómicas en la mayor parte de las variables a incorporar.

Pero volvamos a nuestro clustering: Una vez determinados los clusters que agrupan a viviendas relativamente homogéneas, resta la labor nada sencilla de volcar esos agrupamientos a la dimensión espacial. En efecto, nuestro interés apunta a establecer áreas geográficas identificadas con determinados niveles socioeconómicos (más o menos extensas en función de la escala adoptada, es decir del margen de error admitido)

El considerar como individuos u observaciones para el clustering a los radios u otras áreas semejantes (del tipo de las unidades de primera etapa de la EPH) requiere una mayor elaboración de los microdatos. Una ventaja radica en que obtenemos variables para 280 radios

---

(7) "Hasta" porque algunos grupos de preguntas se formulan sólo a cierto subconjunto de la población, por ejemplo, las referidas a fertilidad, a mujeres de 14 años o más.

que ya no tienen por qué ser dicotómicas. Resultaría probablemente conveniente utilizar porcentajes de hogares en cada área:

1) en viviendas precarias sobre el total del área.
2) en viviendas sin agua por cañería y/o sin retrete.
3) en viviendas con pisos construidos con materiales precarios.
4) hogares con hacinamiento (3 o más personas por cuarto)
5) hogares cuyo jefe no completó instrucción primaria.
6) hogares con uno o más niños de 6 a 12 años que no asisten a escuela.

Nuevamente, éstas son variables contempladas en el cuestionario básico del censo. Como la muestra del 20 % se determina dentro de los radios (cada radio se divide en 8 a 12 segmentos de unas 35 viviendas y se selecciona aleatoriamente uno de cada cinco segmentos) , es factible incluir también variables del cuestionario ampliado (8) .

Para ensayar un clustering como éste no se dispuso, como habría sido conveniente, de los microdatos del censo de población y vivienda de 1980 puesto que los del censo de este año recién se están procesando. Recurrimos entonces a un sucedáneo: la onda de la EPH de octubre de 1989. Se tuvo en cuenta que esta encuesta releva muchas de las variables que veníamos mencionando. Como individuos para el agrupamiento se consideró a las 80 áreas UPE, que comprenden generalmente entre 10 y 20 viviendas cada uno. Las variables finalmente elegidas fueron los siguientes porcentajes de hogares:

1) en viviendas de tipo precario (inquilinato, vivienda en villa y vivienda no destinada a fines habitacionales)
2) en viviendas sin agua corriente y/o sin electricidad y/o sin baño de uso exclusivo.
3) con 3 o más habitantes por cuarto de uso exclusivo.
4) con jefe sin escolaridad primaria completa y simultáneamente 3 o más inactivos por cada activo ocupado.
5) analfabetos sobre el total de la población de 6 años o más del área (éste no es un porcentaje de hogares sino de población)

Otra variable que se había pensado tomar en consideración fue la cantidad de niños que no concurrían a un establecimiento educativo como proporción del total de población en edad escolar, pero en la etapa de recopilación de la información se comprobó que no discriminaba en absoluto porque tenía valor nulo en prácticamente todas las áreas.

Los ensayos en esta aplicación de la técnica de clustering, aplicación que será tema de un próximo informe, parecen contradecir la estratificación que se viene aplicando en la Encuesta Permanente de Hogares. Si bien las UPE del estrato definido por la EPH como el más alto caen dentro de un cluster que se caracteriza por los promedios menores en las 5 variables, y esto ocurre independientemente del método empleado, en ese mismo cluster descubrimos algunas áreas definidas por la EPH como pertenecientes al estrato medio e incluso medio bajo. Si la división de la ciudad que hizo la EPH hace unos años era lógica y los equivocados son nuestros resultados, cabrían dos alternativas: i) que haya que considerar en nuestro clustering otras características del hogar que aproximen más adecuadamente al nivel socioeconómico (por ejemplo, las que agregue el cuestionario censal y que no

---

(8) Hay un aumento de la variancia al pasar del área urbana total a la más reducida del radio censal. Este fenómeno, llamado efecto conglomerado, aumenta el error.

existen en el de la EPH) ; y ii) que haya que incluir características "externas" al propio hogar pero que hacen también a ese estatus que se quiere determinar (disponibilidad de servicios como pavimento, luz de mercurio, etc. o inclusive indicadores como cercanía a la plaza principal o ubicación en una zona exclusiva como la que rodea al monumento a Güemes) . La otra explicación plausible, solucionable también en la medida en que se disponga de los resultados del censo poblacional, es que las observaciones dentro de cada área son pocas y el error muestral demasiado alto.

4.2 Los alumnos de la carrera de contador: su clasificación en base a sus calificaciones

Como es sabido, en el archivo de alumnos de una Facultad se tiene información no solamente de las notas obtenidas en cada asignatura, sino también otros datos que pueden servirnos para caracterizarlos. En particular, nos interesaba testear si es posible establecer algún nexo entre el tipo de estudios cursados en el nivel medio y el rendimiento alcanzado en la etapa universitaria.

El procedimiento de clustering se aplicará a los alumnos de la carrera de Contador. Para eliminar el problema de las observaciones "perdidas", se seleccionó un subconjunto que cumplía la condición de poseer por lo menos alguna calificación en las primeras 21 materias (equivalen a 3 años de la carrera en el plan 73). Al obtenerse finalmente este archivo del Centro de Cómputos pudo comprobarse que incluía a 121 alumnos de ese plan y 20 del plan 85, de modo que se tuvo que reclasificar la muestra para trabajar con cada plan por separado (los códigos de materia cambian entre uno y otro plan) .

Limitándonos al análisis de los alumnos del plan 73 y en base a los ensayos realizados hasta ahora, nos permitiremos las siguientes consideraciones:

En primer lugar, puede parecer desproporcionado el número de variables tomadas en cuenta para el clustering: la matriz de distancias, que es de 121 filas por 121 columnas, tiene que calcularse con distancias euclídeas para pares de vectores de 21 elementos cada uno. A este respecto, digamos que en los agrupamientos ensayados el computador efectuó los cálculos en fracciones de minuto. Sin embargo, y a pesar de que no encontramos mayores referencias en la bibliografía, existiría otra objeción y es que nos estamos alejando de cierta relación deseable entre el número de individuos y el número de variables o características. En la medida que se introduzcan demasiadas variables, se correría el riesgo de crear diferencias "artificiales" entre los individuos.

Los recursos que se nos ocurrieron para solucionar este posible problema fueron dos: i) el análisis de los componentes principales a que se hizo mención más arriba, que permite seleccionar un reducido número de combinaciones lineales armadas a partir del conjunto inicial de variables, con la propiedad de explicar la mayor proporción posible de su variación; y ii) trabajar directamente con promedios de calificaciones por área. Los resultados con la primera de las alternativas no fueron demasiado alentadores: las tres primeras componentes explicaron apenas el 39 % de la varianza total y el 80 % recién se alcanzó al considerar la duodécima. En cambio, parece conveniente redefinir las variables para el clustering como promedios. Nuestros resultados fueron positivos haciendo: área contable (promedio de Introducción a la Contabilidad Superior, Contabilidad Superior, Costos, Contabilidad

Pública y Contabilidad Mecanizada) ; matemática (Algebra Superior, Introducción al Análisis Matemático, Análisis Matemático, Estadística y Matemática Financiera) ; derecho (Derecho Civil, Derecho Constitucional, Derecho Comercial y Derecho Societario) ; humanidades (Introducción a la Problemática Filosófica, Fuentes de Producción e Historia de las Instituciones) ; economía (Economías I, II y III) y administración (Principios de Administración es la única asignatura) .

En segundo lugar, con relación al colegio secundario de origen del alumno, una limitación muy importante que presentó la información proporcionada por el Centro de Cómputos radicó en que casi la mitad de las 121 observaciones no contaban con ese dato. Como los 65 registros que sí lo tenían arrojaban una variedad de 22 colegios secundarios distintos, se decidió agrupar a éstos por orientación (comercial, normal y bachiller, técnica) , y por jurisdicción (nacional, provincial, privada) .

De la revisión de los datos surge que existe un grupo relativamente reducido de alumnos que equivaldrían a los "outliers" que habíamos hallado en la industria manufacturera. Pero aquí se trata de casos en que se tiene un promedio excelente en un área junto con uno muy bajo en otra. Se está investigando la conveniencia de sacar a estas observaciones del clustering, lo que a primera vista podría parecer contrario al objetivo del estudio (9).

Por último, creemos que en este caso y en el de las áreas cubiertas por la EPH no haría falta estandarizar las variables, como sí se requería con las observaciones de industria. Más aun, resultaría contraproducente al reducir el poder discriminatorio de ciertas características o variables.

El procedimiento que hemos estado aplicando en este problema pasa por las siguientes etapas:

i) Realización de agrupamientos aplicando distintos métodos de clustering. El programa, como vimos en industria manufacturera, permite detenerse en una jerarquía arbitraria del árbol; por ejemplo, al arribarse a un número de 4 clusters o al superarse determinada distancia entre los clusters a unir.

ii) Cálculo de los parámetros descriptivos de cada cluster (media, desvío estándar, etc.) . Esto permite caracterizar a un grupo de alumnos como los destacados en contabilidad, a otro como el de los aventajados en matemáticas, etc.

iii) Testeo de la hipótesis nula de que las 121 observaciones se distribuyen entre los clusters independientemente del colegio donde se cursó el secundario.

---

(9) Ya se mencionó que el parámetro "trim" se usa en ese cometido.

ANEXO A

En este Anexo se incluyen cuadros generados por el programa SAS. Debido a que aún no se maneja adecuadamente el editor del programa, la única adaptación que se les practicó fue insertarles al comienzo un número correlativo de cuadro para facilitar la referencia a ellos dentro del texto principal. Para facilitar la interpretación, a continuación se detallan las definiciones y fórmulas utilizadas en los cuadros.

Definiciones básicas

n	número de observaciones
m	número de variables
G	cantidad de clusters formados en un nivel dado de la jerarquía
x_i	i-ésima observación (vector fila)
C_K	K-ésimo cluster
N_K	cantidad de observaciones incluidas en C_K
$\bar{x}$	vector de medias para el total de la población
$\bar{x}_K$	vector de medias para el cluster C_K
$\|x\|$	extensión euclídea o módulo del vector x (raíz cuadrada de la suma de cuadrados de los elementos del vector)
T	suma de cuadrados de desvíos totales: $\Sigma_{i=1}^{n} \|x_i - \bar{x}\|^2$
W_K	suma de cuadrados de desvíos dentro de un cluster: $W_K = \Sigma_{i \in CK} \| x_i - \bar{x}_K \|^2$
P_G	suma de los W de cada cluster, para todos los clusters en determinado nivel de la jerarquía: $P_G = \Sigma W_J$, sobre los G clusters existentes en el nivel G
B_{KL}	dada la unión $C_M = C_K \cup C_L$, $B_{KL} = W_M - W_K - W_L$
$d(x,y)$	distancia o medición de disimilaridad entre los vectores x e y
D_{KL}	distancia o medición de disimilaridad entre los clusters C_K y C_L

Métodos de clustering

La distancia entre dos clusters se puede definir directamente y también en forma combinatoria. En este último caso, se plantea una ecuación que actualiza la matriz de distancias cada vez que dos clus-

ters se unen. En las fórmulas combinatorias que siguen se supone que los clusters C_K y C_L se unen para formar C_M y la fórmula calcula la distancia entre el nuevo cluster C_M y otro cualquiera C_K.

a) Single linkage (unión simple)

distancia entre dos clusters: $D_{KL} = \min_{i \in CK} \min_{j \in CL} d(x_i, x_j)$

fórmula combinatoria: $D_{JM} = \min (D_{JK}, D_{JL})$

En este método la distancia entre dos clusters es la mínima distancia entre una observación de un cluster y una observación del otro. Como distancia entre el nuevo cluster C_M y un tercero C_K, se toma a la menor de las dos que existían entre este último y los dos clusters que se fundieron en C_M.

b) Average linkage (unión por promedios)

distancia entre dos clusters: $D_{KL} = \Sigma_{i \in CK} \Sigma_{j \in CL} d(x_i, x_j) / (N_K \cdot N_L)$

$$D'_{KL} = \|\bar{x}_K - \bar{x}_L\|^2 + W_K/N_K + W_L/N_L$$

fórmula combinatoria: $D_{JM} = (N_K D_{JK} + N_L D_{JL}) / N_M$

La distancia entre dos clusters es la distancia promedio entre pares de observaciones, una de cada cluster.

c) Ward: minimum variance

distancia entre dos clusters: $D_{KL} = B_{KL} =$

$$= \|\bar{x}_K - \bar{x}_L\|^2 / (1/N_K + 1/N_L)$$

fórmula combinatoria:

$$D_{JM} = [(N_J + N_K) D_{JK} + (N_J + N_L) D_{JL} - N_J D_{KL}] / (N_J + N_M)$$

Con el método Ward, en cada generación se minimiza la suma de cuadrados de los desvíos intra-cluster para todas las particiones obtenibles a partir de la unión de dos clusters de la generación previa.

d) Density linkage

Dado un punto x y un entero positivo k, se llama $r_k(x)$ a la distancia desde x hasta la k-ésima observación más cercana. La densidad estimada en x se llama f(x) , es el cociente entre el número de observaciones contenidas dentro de la esfera, que determinan el centro x y el radio $r_k(x)$, y el volumen de esa misma esfera.

Se considera que dos observaciones son adyacentes cuando:

$$d(x_i x_j) \leq \max [r_k(x_i), r_k(x_j)]$$

En ese caso, se define para el par $x_i ; x_j$ una nueva distancia

$d^*(x_i x_j)$ que es la recíproca de un estimado de la densidad media entre ambos puntos:

$$d^*(x_i x_j) = \tfrac{1}{2} \; [1/f(x_i) + 1/f(x_j)]$$

En caso de que los puntos x_i y x_j no sean adyacentes, d^* se considera infinita.

Una vez calculadas las distancias d^*, se recurre a un clustering de unión simple (single linkage)

Coeficientes que aparecen en los clustering

Con el metodo Ward se imprime para la unión de dos clusters C_K y C_L el parámetro semipartial R^2, definido como la disminución en la proporción de varianza a causa de la última unión realizada:

$$\text{semipartial } R^2 = B_{KL} \; / \; T$$

Además se calcula, para cada nivel de la jerarquía, el estadístico R^2:

$$R^2 = 1 - (P_G \; / \; T)$$

En el método Average Linkage se incluye Normalized RMS Distance, que es la distancia euclídea (no cuadrada) media entre pares de observaciones, tomando una observación de cada cluster.

La variable fusion density en el método Density Linkage es el valor de la distancia d^* tal como se la definió más arriba.

Medidas de la forma de la distribución aparecen en los cuadros de Ward con corte o trim. Recordemos que las formulas de la medida de asimetría y de curtosis son:

$$m_3 = \Sigma \; (x_i - \bar{x})^3 \; / \; n.s^3 \qquad \text{(skewness)}$$

$$m_4 = \Sigma \; (x_i - \bar{x})^4 \; / \; n.s^4 \qquad \text{(kurtosis)}$$

Además, se incluye un coeficiente de bimodalidad, calculado como:

$$b = (m_3^2 + 1) \; / \; [m_4 + 3.(n-1)^2 \; / \; (n-2).(n-3)]$$

El coeficiente b tiene un máximo de 1 cuando una población se compone de solamente dos valores distintos entre sí.

CUADRO 1. DATOS DE 80 ESTABLECIMIENTOS INDUSTRIALES DE SALTA
(Fuente: Encuesta Industrial 1986)

OBS	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO	POTENCIA
1	65	796	383	420	29	5649	19200
2	50	476	202	159	1	31	260
3	57	706	271	428	11	2257	34100
4	3	324	89	197	0	2142	31000
5	12	144	48	129	30	383	30000
6	69	816	194	144	11	883	9700
7	19	199	37	38	8	217	1600
8	6	12	2	2	1	39	650
9	354	5423	2410	5006	734	15738	477500
10	13	156	40	5	0	922	160
11	8	161	30	10	0	85	600
12	36	459	190	561	37	3473	80800
13	22	301	97	523	128	5199	385
14	47	535	242	175	6	4652	62000
15	18	392	46	27	8	317	3400
16	125	1500	716	747	47	2943	61100
17	142	1467	777	539	36	4224	77300
18	189	2263	1017	597	176	6271	60860
19	84	1015	302	173	21	1936	14550
20	29	238	101	17	16	250	1200
21	302	1892	1264	1362	54	10742	237770
22	515	4024	1827	2114	72	52999	177600
23	760	5702	3302	205	195	16420	328700
24	211	2566	1866	1725	0	117361	75915
25	26	422	62	27	1	1017	11400
26	134	1621	418	43	5	1518	14400
27	219	1910	566	1817	68	2720	166000
28	12	120	26	3	0	80	1200
29	11	114	33	4	0	162	2700
30	50	598	136	259	22	156	20300
31	33	466	104	101	9	380	31959
32	13	150	34	13	6	65	10500
33	9	44	6	17	2	40	13200
34	14	144	25	18	1	151	17500
35	2	0	0	7	0	7	1700
36	1	0	0	1	0	1	1100
37	20	192	53	7	1	214	3400
38	8	60	15	4	0	135	1800
39	37	428	98	74	3	219	19500
40	42	482	90	82	26	248	220
41	21	253	49	12	1	279	6000
42	27	268	49	22	0	74	13200
43	9	74	16	11	0	60	1750
44	261	3064	752	471	18	7010	42500
45	45	534	236	89	0	581	21240
46	22	233	84	42	0	350	5500
47	262	3141	1920	3713	1071	11062	191540
48	100	945	352	568	115	2839	43000
49	33	281	66	28	0	957	5250
50	27	273	55	44	10	482	3200
51	13	126	31	21	1	58	3825
52	26	220	42	228	2	845	14325
53	176	2208	2122	16542	2680	257655	29129
54	277	3637	1824	1946	250	2784	118000
55	8	60	14	6	5	82	1500
56	9	90	18	9	7	39	13044
57	74	477	117	52	0	760	7500
58	50	635	156	53	0	601	26500
59	9	72	16	10	1	79	1500
60	23	267	56	41	0	174	23900

CUADRO 1 (CONT)

OBS	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTU	POTENCIA
61	15	147	28	13	0	134	5000
62	23	258	53	20	3	260	8700
63	8	79	17	10	0	39	9550
64	39	458	137	89	10	372	12660
65	22	415	48	180	6	1456	8750
66	30	322	82	41	6	582	3750
67	17	156	51	102	12	595	9720
68	2	0	0	1	2	20	100
69	21	209	44	13	1	54	150
70	25	252	76	27	8	186	5200
71	150	1812	802	788	26	1537	72000
72	74	271	65	20	0	178	145
73	26	288	103	14	0	139	8300
74	117	1031	156	788	15	5244	509
75	3	16	4	3	1	44	175
76	94	1099	400	81	0	623	18800
77	30	296	94	68	2	509	5600
78	38	404	107	30	0	698	3100
79	7	45	12	7	2	56	0
80	1	0	0	0	0	4	0

CUADRO 2. CLUSTER CON METODO WARD (2 VARIABLES)

Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.28884	0.577681	0.644420	0.64442
2	0.71116	.	0.355580	1.00000

The data have been standardized to mean 0 and variance 1

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
79	OB33	OB56	2	0.000000	1.000000	T
78	OB11	OB55	2	0.000000	1.000000	T
77	OB36	OB80	2	0.000000	1.000000	T
76	OB32	OB51	2	0.000000	1.000000	T
75	OB35	OB68	2	0.000000	1.000000	
74	OB43	OB59	2	0.000000	1.000000	
73	CL79	CL74	4	0.000000	1.000000	
72	CL78	OB63	3	0.000000	1.000000	
71	OB66	OB77	2	0.000000	1.000000	
70	CL72	OB38	4	0.000000	1.000000	
69	OB60	OB62	2	0.000000	1.000000	
68	OB2	OB30	2	0.000000	1.000000	
67	OB25	OB52	2	0.000000	1.000000	
66	OB41	OB69	2	0.000000	1.000000	
65	OB7	OB37	2	0.000000	0.999999	
64	OB8	OB79	2	0.000000	0.999999	T
63	OB34	OB61	2	0.000000	0.999999	
62	OB70	OB73	2	0.000000	0.999999	
61	OB28	OB29	2	0.000000	0.999999	
60	CL75	OB75	3	0.000000	0.999998	
59	OB5	CL61	3	0.000000	0.999998	
58	OB46	CL69	3	0.000000	0.999998	

CUADRO 2 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
57	OB15	OB67	2	0.000000	0.999997	
56	OB42	OB50	2	0.000001	0.999997	
55	OB64	OB78	2	0.000001	0.999996	
54	OB20	CL71	3	0.000001	0.999995	
53	CL70	CL73	8	0.000001	0.999994	
52	CL60	CL77	5	0.000001	0.999994	
51	CL65	CL66	4	0.000001	0.999993	
50	CL76	CL63	4	0.000001	0.999992	
49	OB31	OB49	2	0.000001	0.999990	
48	CL56	CL62	4	0.000001	0.999989	
47	OB57	OB72	2	0.000001	0.999988	
46	OB39	CL55	3	0.000001	0.999987	
45	CL68	OB58	3	0.000001	0.999986	
44	CL64	CL53	10	0.000003	0.999983	
43	CL59	OB10	4	0.000003	0.999981	
42	OB40	OB45	2	0.000003	0.999977	
41	CL43	CL50	8	0.000004	0.999974	
40	CL67	CL48	6	0.000004	0.999969	
39	CL51	CL58	7	0.000004	0.999965	
38	CL54	CL49	5	0.000006	0.999959	
37	OB6	CL47	3	0.000008	0.999951	
36	CL39	OB65	8	0.000009	0.999943	
35	CL36	CL57	10	0.000010	0.999932	
34	CL46	CL42	5	0.000016	0.999916	
33	OB48	OB76	2	0.000023	0.999893	
32	OB16	OB26	2	0.000024	0.999869	
31	OB4	CL52	6	0.000024	0.999845	
30	CL38	CL40	11	0.000028	0.999818	
29	OB12	OB14	2	0.000030	0.999787	
28	CL45	OB3	4	0.000034	0.999753	
27	OB17	OB71	2	0.000036	0.999716	
26	CL41	CL44	18	0.000045	0.999672	
25	OB19	CL33	3	0.000049	0.999623	
24	OB44	OB47	2	0.000051	0.999572	
23	CL32	OB74	3	0.000083	0.999489	
22	CL29	OB13	3	0.000115	0.999374	
21	CL28	CL34	9	0.000129	0.999245	
20	CL31	CL26	24	0.000133	0.999112	
19	OB1	CL37	4	0.000136	0.998976	
18	CL35	CL30	21	0.000137	0.998839	
17	CL23	CL27	5	0.000222	0.998617	
16	CL24	OB54	3	0.000232	0.998385	
15	OB18	OB27	2	0.000234	0.998152	
14	CL21	CL22	12	0.000310	0.997842	
13	CL19	CL25	7	0.000364	0.997478	
12	OB21	CL16	4	0.000471	0.997007	
11	CL14	CL18	33	0.001137	0.995869	
10	OB9	CL12	5	0.002434	0.993435	
9	CL17	CL15	7	0.003073	0.990362	
8	CL11	CL20	57	0.003229	0.987133	
7	CL13	CL9	14	0.008266	0.978866	
6	OB22	OB23	2	0.017129	0.961738	
5	CL7	CL8	71	0.044382	0.917356	
4	OB24	OB53	2	0.061723	0.855633	
3	CL10	CL6	7	0.079610	0.776023	
2	CL3	CL4	9	0.309451	0.466572	
1	CL5	CL2	80	0.466572	0.000000	

CUADRO 3. CLUSTER CON METODO WARD (7 VARIABLES)

Ward's Minimum Variance Cluster Analysis

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.7237E9	4.7314E9	0.852124	0.85212
2	9.9226E8	9.9158E8	0.147724	0.99985
3	681961.2	365465.6	0.000102	0.99995
4	316495.6	298561.6	0.000047	1.00000
5	17934.01	15381.74	0.000003	1.00000
6	2552.273	1427.645	0.000000	1.00000
7	1124.628	.	0.000000	1.00000

Root-Mean-Square Total-Sample Standard Deviation = 30976.91
Root-Mean-Square Distance Between Observations = 115905

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
79	OB55	OB59	2	0.000000	1.000000	
78	OB79	OB80	2	0.000000	1.000000	
77	OB68	OB75	2	0.000000	1.000000	
76	OB38	OB43	2	0.000000	1.000000	
75	OB69	OB72	2	0.000000	1.000000	
74	OB35	CL76	3	0.000000	1.000000	
73	OB33	OB56	2	0.000000	1.000000	
72	OB8	OB11	2	0.000000	1.000000	
71	OB28	OB36	2	0.000000	1.000000	
70	CL77	CL78	4	0.000000	1.000000	
69	OB46	OB77	2	0.000000	1.000000	
68	OB15	OB37	2	0.000000	1.000000	
67	OB61	OB70	2	0.000000	1.000000	
66	OB7	CL79	3	0.000000	1.000000	
65	CL73	OB42	3	0.000000	1.000000	
64	OB2	OB40	2	0.000000	0.999999	
63	OB50	OB78	2	0.000000	0.999999	
62	OB20	CL71	3	0.000000	0.999999	
61	CL66	CL74	6	0.000000	0.999999	
60	CL64	CL75	4	0.000000	0.999999	
59	OB62	OB73	2	0.000000	0.999999	
58	CL68	OB66	3	0.000000	0.999999	
57	OB41	CL69	3	0.000000	0.999998	
56	CL58	OB51	4	0.000000	0.999998	
55	OB63	OB67	2	0.000000	0.999998	
54	OB29	CL63	3	0.000001	0.999997	
53	OB19	OB26	2	0.000001	0.999996	
52	CL60	CL70	8	0.000001	0.999996	
51	OB13	OB74	2	0.000001	0.999995	
50	CL65	OB64	4	0.000001	0.999995	
49	OB30	OB39	2	0.000001	0.999994	
48	CL72	CL62	5	0.000001	0.999993	
47	CL57	OB49	4	0.000001	0.999993	
46	CL47	CL67	6	0.000001	0.999992	
45	OB32	CL55	3	0.000001	0.999991	
44	CL52	OB10	9	0.000001	0.999989	
43	CL56	CL54	7	0.000001	0.999988	
42	OB6	OB65	2	0.000001	0.999987	
41	OB57	CL59	3	0.000002	0.999985	
40	CL49	OB76	3	0.000002	0.999983	

CUADRO 3 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
39	CL61	CL48	11	0.000002	0.999980	
38	CL53	OB52	3	0.000003	0.999978	
37	CL42	CL45	5	0.000004	0.999974	
36	OB5	OB31	2	0.000004	0.999970	
35	OB4	CL36	3	0.000004	0.999966	
34	CL40	OB45	4	0.000005	0.999962	
33	OB14	OB16	2	0.000005	0.999957	
32	CL37	OB25	6	0.000005	0.999952	
31	OB58	OB60	2	0.000007	0.999945	
30	CL34	OB34	5	0.000010	0.999935	
29	CL33	OB18	3	0.000011	0.999925	
28	CL32	CL41	9	0.000012	0.999912	
27	OB12	OB17	2	0.000013	0.999899	
26	CL44	CL39	20	0.000013	0.999886	
25	CL38	CL50	7	0.000014	0.999873	
24	OB3	CL35	4	0.000016	0.999856	
23	OB44	OB48	2	0.000021	0.999834	
22	CL43	CL46	13	0.000026	0.999808	
21	OB1	CL30	6	0.000044	0.999764	
20	CL27	OB71	3	0.000070	0.999694	
19	CL26	CL51	22	0.000092	0.999602	
18	CL21	CL31	8	0.000097	0.999506	
17	CL28	CL25	16	0.000136	0.999369	
16	CL19	CL22	35	0.000194	0.999175	
15	CL24	CL23	6	0.000344	0.998832	
14	CL20	CL29	6	0.000676	0.998156	
13	OB27	OB47	2	0.000688	0.997469	
12	CL18	CL17	24	0.000938	0.996531	
11	OB22	CL13	3	0.002677	0.993854	
10	CL14	OB54	7	0.003893	0.989961	
9	CL12	CL15	30	0.004016	0.985946	
8	OB21	CL11	4	0.005177	0.980769	
7	CL9	CL16	65	0.008362	0.972407	
6	OB24	OB33	2	0.020822	0.951585	
5	OB9	OB23	2	0.020886	0.930698	
4	CL7	CL10	72	0.052467	0.878232	
3	CL5	CL8	6	0.110728	0.767504	
2	CL4	CL6	74	0.132636	0.634867	
1	CL2	CL3	80	0.634867	0.000000	

CUADRO 4. CLUSTER CON METODO AVERAGE (VARIABLES ORIGINALES)

Average Linkage Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
79	OB55	OB59	2	0.000119	
78	OB79	OB80	2	0.000608	
77	OB68	OB75	2	0.000694	
76	OB38	OB43	2	0.000789	
75	OB35	CL76	3	0.001241	
74	CL77	CL78	4	0.001280	
73	OB69	OB72	2	0.001295	
72	OB33	OB56	2	0.001409	
71	OB8	OB11	2	0.001435	
70	OB28	OB36	2	0.001530	
69	OB46	OB77	2	0.001728	
68	OB7	CL79	3	0.001888	
67	OB15	OB37	2	0.001951	
66	CL72	OB42	3	0.002044	
65	OB61	OB70	2	0.002049	
64	OB2	OB40	2	0.002248	
63	CL68	CL75	6	0.002302	
62	OB50	OB78	2	0.002393	
61	CL74	CL73	6	0.002446	
60	OB20	CL70	3	0.002655	
59	OB62	OB73	2	0.003641	
58	CL67	CL62	4	0.003893	
57	OB41	CL69	3	0.004216	
56	CL64	CL61	8	0.004282	
55	CL63	CL60	9	0.004522	
54	CL58	OB66	5	0.004795	
53	CL56	CL71	10	0.004903	
52	OB63	OB67	2	0.005132	
51	CL66	OB64	4	0.005979	
50	CL57	CL65	5	0.006005	
49	CL54	OB51	6	0.006056	
48	OB19	OB26	2	0.006669	
47	CL50	OB49	6	0.006838	
46	OB13	OB74	2	0.006933	
45	OB30	OB39	2	0.007265	
44	CL49	OB29	7	0.007565	
43	CL53	OB10	11	0.007772	
42	OB32	CL52	3	0.008199	
41	OB6	CL42	4	0.009547	
40	OB57	CL59	3	0.010229	
39	CL43	CL55	20	0.011609	
38	CL40	OB65	4	0.011726	
37	CL45	OB76	3	0.012233	
36	CL48	OB52	3	0.013022	
35	CL41	OB25	5	0.015230	
34	CL37	OB45	4	0.016758	
33	OB5	OB31	2	0.017139	
32	OB4	CL33	3	0.017458	
31	CL44	CL47	13	0.018755	
30	CL35	CL38	9	0.018803	
29	CL36	CL51	7	0.019144	
28	OB14	OB16	2	0.019715	
27	OB58	OB60	2	0.022971	
26	CL34	OB34	5	0.023593	
25	CL28	OB18	3	0.026998	
24	OB3	CL32	4	0.031083	
23	CL39	CL31	33	0.032491	

CUADRO 4 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
22	OB12	OB17	2	0.032496	
21	CL30	CL29	16	0.039734	
20	OB44	OB48	2	0.040781	
19	OB1	CL26	6	0.046938	
18	CL23	CL46	35	0.049184	
17	CL19	CL27	8	0.055494	
16	CL22	OB71	3	0.066537	
15	CL18	CL21	51	0.083970	
14	CL17	CL24	12	0.099474	
13	CL16	CL25	6	0.138188	
12	CL14	CL20	14	0.170446	
11	CL12	CL15	65	0.209486	
10	OB27	OB47	2	0.233082	
9	OB22	CL10	3	0.414950	
8	CL13	OB54	7	0.429517	
7	OB21	CL9	4	0.562977	
6	CL11	CL8	72	0.602514	
5	CL6	OB24	73	1.145567	
4	CL7	OB23	5	1.205154	
3	CL5	CL4	78	1.854844	
2	CL3	OB53	79	2.253427	
1	CL2	OB9	80	3.902473	

CUADRO 5. CLUSTER CON METODO AVERAGE (VARIABLES ESTANDARIZADAS)

Average Linkage Cluster Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.60642	2.69741	0.658060	0.65806
2	1.90901	1.62539	0.272716	0.93078
3	0.28362	0.17344	0.040518	0.97129
4	0.11019	0.06429	0.015741	0.98704
5	0.04589	0.00762	0.006556	0.99359
6	0.03827	0.03168	0.005467	0.99906
7	0.00659	.	0.000941	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Distance Between Observations = 3.741657

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
79	OB43	OB59	2	0.001302	
78	OB68	OB80	2	0.002767	
77	OB35	OB36	2	0.003169	
76	OB38	CL79	3	0.004005	
75	CL76	OB55	4	0.005020	
74	CL77	CL78	4	0.005306	
73	CL74	OB75	5	0.006286	
72	OB28	OB29	2	0.006638	
71	OB51	OB61	2	0.007997	
70	CL75	OB79	5	0.008831	

CUADRO 5 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
69	OB8	CL73	6	0.010097	
68	OB7	OB69	2	0.010254	
67	OB41	OB62	2	0.010900	
66	OB46	OB70	2	0.011103	
65	OB66	OB77	2	0.011318	
64	OB33	OB56	2	0.012455	
63	CL68	OB37	3	0.012487	
62	CL72	CL71	4	0.012535	
61	OB10	OB11	2	0.013887	
60	CL64	OB63	3	0.015135	
59	OB49	CL65	3	0.015542	
58	CL67	CL66	4	0.016934	
57	CL61	CL62	6	0.016945	
56	OB32	OB67	2	0.018058	
55	CL59	OB50	4	0.019284	
54	CL69	CL70	11	0.020303	
53	CL58	OB73	5	0.022851	
52	CL53	CL55	9	0.024844	
51	OB25	OB65	2	0.025695	
50	CL63	CL57	9	0.026440	
49	OB20	CL52	10	0.028563	
48	CL56	OB34	3	0.030028	
47	CL48	CL60	6	0.030634	
46	OB39	OB64	2	0.030951	
45	OB42	OB52	2	0.031155	
44	OB40	OB78	2	0.032586	
43	OB15	CL51	3	0.033616	
42	OB30	OB58	2	0.041602	
41	CL49	CL45	12	0.042014	
40	CL50	CL54	20	0.042203	
39	CL43	CL41	15	0.045561	
38	CL40	CL47	26	0.048597	
37	CL42	OB45	3	0.050707	
36	OB5	OB60	2	0.051019	
35	OB2	CL44	3	0.054890	
34	OB19	OB76	2	0.057746	
33	OB57	OB72	2	0.058662	
32	OB4	CL36	3	0.059033	
31	OB31	CL46	3	0.059947	
30	CL37	CL31	6	0.064353	
29	CL38	CL39	41	0.072287	
28	OB16	OB17	2	0.079949	
27	OB1	OB3	2	0.082131	
26	CL35	CL33	5	0.084087	
25	OB12	OB14	2	0.096168	
24	CL26	CL30	11	0.100078	
23	CL28	OB71	3	0.100558	
22	CL32	CL29	44	0.100798	
21	OB6	CL34	3	0.103926	
20	CL27	CL21	5	0.116484	
19	CL24	CL22	55	0.132597	
18	CL19	OB13	56	0.146289	
17	CL20	OB48	6	0.154712	
16	CL17	OB74	7	0.175624	
15	CL16	OB26	8	0.229655	
14	CL23	OB18	4	0.254602	
13	CL18	CL25	58	0.265075	
12	CL15	CL13	66	0.291430	
11	CL14	OB44	5	0.417785	
10	OB21	OB27	2	0.438085	

CUADRO 5 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
9	CL11	CL10	7	0.619855	
8	CL9	OB54	8	0.700779	
7	CL12	CL8	74	0.862557	
6	OB22	OB24	2	0.995300	
5	OB9	OB47	2	1.220948	
4	CL5	CL6	4	1.486621	
3	CL4	OB23	5	1.573717	
2	CL7	CL3	79	2.010148	
1	CL2	OB53	80	3.865437	

CUADRO 6. CLUSTER CON METODO DENSITY (VARIABLES ORIGINALES)

Density Linkage Cluster Analysis

(K = 5)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster Lesser	Greater	
79	OB68	OB75	2	6.56E-19	4.81E-19	1.03E-18	
78	CL79	OB69	3	6.29E-19	4.52E-19	1.03E-18	
77	CL78	OB79	4	6.29E-19	4.52E-19	1.03E-18	
76	OB35	OB59	2	2.89E-19	2.12E-19	4.51E-19	
75	CL76	OB55	3	2.85E-19	2.08E-19	4.51E-19	
74	CL75	OB43	4	2.85E-19	2.08E-19	4.51E-19	
73	CL77	OB60	5	2.43E-19	1.37E-19	1.03E-18	
72	CL74	OB7	5	1.84E-19	1.62E-19	4.51E-19	
71	CL73	OB72	6	9.73E-20	5.11E-20	1.03E-18	
70	CL72	OB38	6	9.69E-20	5.43E-20	4.51E-19	
69	CL70	OB28	7	8.4E-20	5.23E-20	4.51E-19	
68	CL69	OB20	8	1.62E-20	8.41E-21	4.51E-19	
67	CL68	OB36	9	1.23E-20	6.31E-21	4.31E-19	
66	CL71	OB11	7	4.29E-21	2.15E-21	1.03E-18	
65	CL66	OB2	8	2.54E-21	1.27E-21	1.03E-18	
64	CL65	OB40	9	2.37E-21	1.19E-21	1.03E-18	
63	CL64	OB8	10	1.98E-21	9.93E-22	1.03E-18	
62	CL63	CL67	19	1.72E-21	4.51E-19	1.03E-18	*
61	OB15	OB37	2	1.48E-21	1.18E-21	1.97E-21	
60	CL61	OB50	3	1.08E-21	7.49E-22	1.97E-21	
59	CL60	OB66	4	1.03E-21	6.99E-22	1.97E-21	
58	OB46	OB77	2	7.35E-22	6.7E-22	8.15E-22	
57	CL59	OB78	5	4.08E-22	2.28E-22	1.97E-21	
56	CL57	OB29	6	1.59E-22	8.28E-23	1.97E-21	
55	CL56	OB51	7	1.58E-22	8.22E-23	1.97E-21	
54	CL58	OB70	3	1.46E-22	8E-23	8.15E-22	
53	CL54	OB61	4	6.64E-23	3.46E-23	8.15E-22	
52	CL53	OB49	5	6.64E-23	3.46E-23	8.15E-22	
51	CL62	OB10	20	6.07E-23	3.04E-23	1.03E-18	
50	CL52	OB41	6	2.4E-23	1.22E-23	8.15E-22	
49	OB63	OB67	2	5.94E-24	4.86E-24	7.64E-24	
48	CL49	OB62	3	4.64E-24	3.33E-24	7.64E-24	
47	CL48	OB6	4	3.06E-24	1.92E-24	7.64E-24	
46	CL47	OB32	5	2.77E-24	1.7E-24	7.64E-24	

CUADRO 6 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster: Lesser	Greater	
45	CL46	OB65	6	2.08E-24	1.2E-24	7.64E-24	
44	CL45	OB73	7	1.93E-24	1.2E-24	7.64E-24	
43	OB33	OB42	2	1.28E-24	1.19E-24	1.39E-24	
42	CL43	OB64	3	1.25E-24	1.14E-24	1.39E-24	
41	CL50	OB57	7	9.72E-25	5.06E-25	8.15E-22	
40	CL42	OB52	4	8.93E-25	6.59E-25	1.39E-24	
39	CL40	OB56	5	8.93E-25	6.59E-25	1.39E-24	
38	CL41	CL44	14	8.79E-25	7.64E-24	8.15E-22	*
37	CL38	OB25	15	4.9E-25	2.53E-25	8.15E-22	
36	CL37	CL39	20	4.14E-25	1.39E-24	8.15E-22	*
35	CL36	OB26	21	5.57E-26	2.84E-26	8.15E-22	
34	CL35	OB19	22	4.01E-26	2.03E-26	8.15E-22	
33	OB39	OB76	2	3.35E-26	2.03E-26	9.53E-26	
32	CL33	OB45	3	1.9E-26	1.05E-26	9.53E-26	
31	CL32	OB30	4	1.59E-26	8.68E-27	9.53E-26	
30	CL34	OB34	23	6.76E-27	3.4E-27	8.15E-22	
29	CL30	CL31	27	6.56E-27	9.53E-26	8.15E-22	
28	CL29	OB60	28	8.2E-28	4.12E-28	8.15E-22	
27	CL51	OB13	21	3.19E-28	1.6E-28	1.03E-18	
26	OB4	OB5	2	2.73E-28	2.36E-28	3.23E-28	
25	CL27	OB74	22	2.71E-28	1.35E-28	1.03E-18	
24	CL28	OB58	29	2.32E-28	1.17E-28	8.15E-22	
23	CL24	CL26	31	1.72E-28	3.23E-28	8.15E-22	
22	CL23	OB1	32	1.51E-28	7.53E-29	8.15E-22	
21	CL22	OB31	33	1.42E-28	9.07E-29	8.15E-22	
20	CL21	OB3	34	1.49E-29	7.61E-30	8.15E-22	
19	CL20	OB48	35	7.18E-31	3.59E-31	8.15E-22	
18	CL19	OB44	36	4.68E-31	2.34E-31	8.15E-22	
17	OB14	OB71	2	1.2E-31	6.59E-32	6.81E-31	
16	CL17	OB16	3	8.3E-32	4.42E-32	6.81E-31	
15	CL16	OB17	4	8.3E-32	4.42E-32	6.81E-31	
14	CL15	OB18	5	7.25E-32	3.83E-32	6.81E-31	
13	CL14	OB12	6	2.23E-32	1.13E-32	6.81E-31	
12	CL13	OB54	7	4.47E-35	2.24E-35	6.81E-31	
11	CL12	OB27	8	2.44E-36	1.29E-36	6.81E-31	
10	CL11	OB47	9	2.07E-36	1.09E-36	6.81E-31	
9	CL10	OB22	10	1.47E-36	7.57E-37	6.81E-31	
8	CL9	OB21	11	4.21E-37	2.52E-37	6.81E-31	
7	CL8	OB24	12	1.06E-37	5.28E-38	6.81E-31	
6	CL7	OB23	13	1.2E-38	6.02E-39	6.81E-31	
5	CL18	OB53	37	3.96E-40	1.98E-40	8.15E-22	
4	CL5	CL6	50	3.96E-40	6.81E-31	8.15E-22	*
3	CL4	OB9	51	1.15E-40	5.74E-41	8.15E-22	

* indicates fusion of two modal clusters

7 modal clusters have been formed.

CUADRO 7. CLUSTER CON METODO DENSITY (VARIABLES ESTANDARIZADAS)

Density Linkage Cluster Analysis

K = 5

Root-Mean-Square Total-Sample Standard Deviation = 1

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster Lesser	Greater
79	OB68	OB80	2	3.4656E9	2.6845E9	4.8878E9
78	CL79	OB36	3	2.582E9	1.7543E9	4.8878E9
77	CL78	OB75	4	2.4719E9	1.6542E9	4.8878E9
76	CL77	OB35	5	2.0799E9	1.321E9	4.8878E9
75	OB38	OB55	2	8.9178E8	6.7865E8	1.3001E9
74	CL75	OB79	3	3.265E8	1.867E8	1.3001E9
73	CL74	OB59	4	3.265E8	1.867E8	1.3001E9
72	CL76	OB8	6	2.4065E8	1.2336E8	4.8878E9
71	CL73	OB43	5	1.8232E8	98032342	1.3001E9
70	CL72	CL71	11	1.4856E8	1.3001E9	4.8878E9 *
69	CL70	OB29	12	28517071	15437549	4.8878E9
68	CL69	OB28	13	16673266	18124072	4.8878E9
67	CL68	OB10	14	5398316	3171475	4.8878E9
66	CL67	OB51	15	5398316	3171475	4.8878E9
65	CL66	OB11	16	3917313	2195974	4.8878E9
64	OB41	OB70	2	3187148	3095076	3284867
63	CL65	OB61	17	2415648	1294063	4.8878E9
62	CL64	OB77	3	2235989	1694822	3284867
61	CL62	OB62	4	2157963	1606752	3284867
60	CL61	OB46	5	2066407	1507302	3284867
59	CL60	OB37	6	2027306	1507302	3284867
58	CL59	OB50	7	1812889	1251901	3284867
57	CL63	CL58	24	1392567	3284867	4.8878E9 *
56	CL57	OB66	25	1253425	774473	4.8878E9
55	CL56	OB7	26	1220999	756034	4.8878E9
54	CL55	OB49	27	1054841	628301	4.8878E9
53	CL54	OB32	28	775815	553963	4.8878E9
52	CL53	OB73	29	688121	384314	4.8878E9
51	CL52	OB69	30	635713	353262	4.8878E9
50	CL51	OB20	31	418058	223235	4.8878E9
49	CL50	OB56	32	398149	310745	4.8878E9
48	CL49	OB63	33	289634	151746	4.8878E9
47	CL48	OB67	34	130834	66794.9	4.8878E9
46	CL47	OB33	35	90296.6	49154.4	4.8878E9
45	CL46	OB42	36	89151.2	45227.0	4.8878E9
44	CL45	OB34	37	47698.5	24922.2	4.8878E9
43	CL44	OB15	38	42702.1	21499.4	4.8878E9
42	CL43	OB78	39	34002.2	17173.4	4.8878E9
41	CL42	OB25	40	22958.7	11832.8	4.8878E9
40	CL41	OB65	41	9273.6	4649.5	4.8878E9
39	CL40	OB52	42	8120.4	4065.5	4.8878E9
38	CL39	OB60	43	2775.8	1431.9	4.8878E9
37	CL38	OB64	44	2750.1	1494.7	4.8878E9
36	CL37	OB39	45	2120.1	1085.5	4.8878E9
35	CL36	OB40	46	1676.4	839.1	4.8878E9
34	CL35	OB30	47	771.8	520.2	4.8878E9
33	CL34	OB45	48	417.9	242.9	4.8878E9
32	CL33	OB58	49	397.4	243.2	4.8878E9
31	CL32	OB5	50	396.5	198.4	4.8878E9
30	CL31	OB31	51	304.2	170.2	4.8878E9

CUADRO 7 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster Lesser	Greater
29	CL30	OB2	52	294.2	148.4	4.8878E9
28	CL29	OB4	53	139.1	69.7547	4.8878E9
27	CL28	OB57	54	112.6	58.5128	4.8878E9
26	CL27	OB5	55	70.6337	37.8889	4.8878E9
25	CL26	OB6	56	64.8912	34.6037	4.8878E9
24	CL25	OB72	57	44.3598	22.1807	4.8878E9
23	CL24	OB1	58	8.1694	4.5783	4.8878E9
22	CL23	OB19	59	5.9820	3.2474	4.8878E9
21	CL22	OB13	60	5.2014	2.6007	4.8878E9
20	CL21	OB14	61	2.6517	1.3331	4.8878E9
19	CL20	OB48	62	1.4719	0.7505	4.8878E9
18	CL19	OB76	63	1.4692	0.7505	4.8878E9
17	CL18	OB74	64	0.4316	0.2172	4.8878E9
16	CL17	OB12	65	0.1616	0.0808	4.8878E9
15	CL16	OB26	66	0.0699	0.0354	4.8878E9
14	CL15	OB16	67	0.0656	0.0343	4.8878E9
13	CL14	OB17	68	0.0234	0.0119	4.8878E9
12	CL13	OB71	69	0.0137	0.00851	4.8878E9
11	CL12	OB18	70	0.00927	0.00536	4.8878E9
10	CL11	OB27	71	0.000554	0.00028	4.8878E9
9	CL10	OB44	72	0.000483	0.000243	4.8878E9
8	CL9	OB54	73	0.000029	0.000014	4.8878E9
7	CL8	OB21	74	0.000028	0.000014	4.8878E9
6	CL7	OB22	75	2.613E-6	1.314E-6	4.8878E9
5	CL6	OB24	76	1.61E-6	8.049E-7	4.8878E9
4	CL5	OB47	77	1.584E-6	7.92E-7	4.8878E9
3	CL4	OB9	78	1.698E-7	8.54E-8	4.8878E9
2	CL3	OB23	79	8.359E-8	4.192E-8	4.8878E9
1	CL2	OB53	80	3.98E-10	1.99E-10	4.8878E9

* indicates fusion of two modal clusters

3 modal clusters have been formed.

CUADRO 8. CLUSTER CON METODO WARD (TRIM = 10 %)

Ward's Minimum Variance Cluster Analysis

Simple Statistics Before Trimming

	Mean	Std Dev	Skewness	Kurtosis	Bimodality
OCUPADOS	74.8	121.1	3.4	14.2	0.7
REMUNERA	784.2	1161.5	2.5	6.7	0.8
SALARIOS	342.2	638.6	2.7	7.4	0.8
CONSUMWH	549.8	1977.9	7.1	55.7	0.9
COMBUSTI	75.2	329.9	6.8	51.3	0.9
PRODUCTO	7025.6	31834.8	6.9	51.4	0.9
POTENCIA	35598.0	75483.4	3.9	17.3	0.8

0 observation(s) trimmed with estimated density 0.0000143331 or less.

CUADRO 8 (CONT)

Simple Statistics After Trimming

	Mean	Std Dev	Skewness	Kurtosis	Bimodality
OCUPADOS	43.4	51.7	2.3	5.6	0.7
REMUNERA	474.2	573.3	2.4	6.5	0.7
SALARIOS	150.5	215.8	2.4	5.2	0.8
CONSUMWH	157.9	286.9	3.4	15.4	0.7
COMBUSTI	13.4	29.8	3.8	16.1	0.8
PRODUCTO	1073.4	1666.8	2.1	3.6	0.8
POTENCIA	16829.0	26472.9	3.3	14.0	0.7

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.15644	4.47684	0.736634	0.73663
2	0.67959	0.06134	0.097085	0.83372
3	0.61825	0.31398	0.088322	0.92204
4	0.30427	0.14290	0.043465	0.96551
5	0.16137	0.09811	0.023053	0.98856
6	0.06326	0.04645	0.009037	0.99760
7	0.01681	.	0.002401	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Distance Between Observations = 3.741657

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
71	OB35	OB36	2	0.000001	0.999999	
70	OB43	OB59	2	0.000001	0.999997	
69	OB68	OB75	2	0.000003	0.999994	
68	CL71	OB80	3	0.000004	0.999990	
67	OB38	CL70	3	0.000004	0.999986	
66	OB8	CL69	3	0.000007	0.999979	
65	OB28	OB29	2	0.000007	0.999971	
64	OB51	OB61	2	0.000009	0.999962	
63	OB55	OB77	2	0.000015	0.999948	
62	CL66	CL68	6	0.000016	0.999932	
61	OB41	OB62	2	0.000018	0.999914	
60	OB11	CL65	3	0.000019	0.999895	
59	OB37	OB69	2	0.000028	0.999867	
58	OB33	OB63	2	0.000031	0.999836	
57	OB32	OB56	2	0.000033	0.999803	
56	CL67	CL63	5	0.000034	0.999768	
55	OB66	OB77	2	0.000039	0.999729	
54	CL60	CL64	5	0.000046	0.999683	
53	OB50	OB70	2	0.000058	0.999625	
52	OB46	OB73	2	0.000060	0.999565	
51	OB7	CL53	3	0.000086	0.999480	
50	CL61	CL52	4	0.000095	0.999385	
49	CL57	CL58	4	0.000104	0.999281	
48	OB49	OB78	2	0.000123	0.999158	
47	CL49	OB34	5	0.000128	0.999031	
46	CL51	OB15	4	0.000128	0.998902	
45	CL50	OB42	5	0.000135	0.998768	
44	CL48	CL55	4	0.000140	0.998628	

CUADRO 8 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
43	CL54	CL59	7	0.000149	0.998479	
42	CL62	CL56	11	0.000162	0.998316	
41	OB39	OB64	2	0.000171	0.998145	
40	CL46	OB20	5	0.000233	0.997912	
39	OB45	OB58	2	0.000263	0.997649	
38	OB25	CL44	5	0.000276	0.997373	
37	OB52	OB65	2	0.000349	0.997024	
36	CL40	OB67	6	0.000378	0.996646	
35	OB57	OB72	2	0.000401	0.996245	
34	OB10	CL43	8	0.000431	0.995814	
33	OB31	OB60	2	0.000453	0.995361	
32	CL33	CL41	4	0.000587	0.994775	
31	CL36	CL45	11	0.000771	0.994003	
30	CL42	CL34	19	0.000989	0.993014	
29	OB2	OB6	2	0.000998	0.992017	
28	CL31	CL38	16	0.001078	0.990939	
27	OB30	OB40	2	0.001094	0.989845	
26	CL32	CL39	6	0.001324	0.988520	
25	CL29	CL27	4	0.001387	0.987134	
24	OB4	CL37	3	0.001501	0.985632	
23	OB19	OB76	2	0.001519	0.984113	
22	CL30	CL47	24	0.001538	0.982575	
21	OB16	OB17	2	0.001829	0.980746	
20	CL25	CL35	6	0.002346	0.978400	
19	CL24	OB5	4	0.002488	0.975912	
18	CL23	OB26	3	0.002640	0.973272	
17	CL21	OB71	3	0.003408	0.969864	
16	CL20	CL26	12	0.003837	0.966007	
15	OB12	OB14	2	0.004041	0.961966	
14	OB1	OB74	2	0.004736	0.957230	
13	OB3	CL15	3	0.004848	0.952382	
12	CL19	CL28	20	0.005634	0.946748	
11	CL16	CL12	32	0.008032	0.938716	
10	OB13	OB48	2	0.009800	0.928916	
9	CL14	CL13	5	0.015870	0.913046	
8	CL11	CL22	56	0.021347	0.891698	
7	CL9	CL18	8	0.028123	0.863576	
6	CL17	OB44	4	0.030604	0.832972	
5	CL6	OB18	5	0.043809	0.789163	
4	CL7	CL10	10	0.045164	0.744000	
3	CL5	OB27	6	0.059651	0.684349	
2	CL4	CL3	16	0.154581	0.529767	
1	CL2	CL8	72	0.529767	0.000000	

CUADRO 9. CLUSTER CON METODO WARD (TRIM = 15 %)

Ward's Minimum Variance Cluster Analysis

Simple Statistics Before Trimming

	Mean	Std Dev	Skewness	Kurtosis	Bimodality
OCUPADOS	74.8	121.1	3.4	14.2	0.7
REMUNERA	784.2	1161.5	2.5	6.7	0.8
SALARIOS	342.2	638.6	2.7	7.4	0.8
CONSUMWH	549.8	1977.9	7.1	55.7	0.9
COMBUSTI	75.2	329.9	6.8	51.3	0.9
PRODUCTO	7025.6	31834.8	6.9	51.4	0.9
POTENCIA	35598.0	75483.4	3.9	17.3	0.8

12 observation(s) trimmed with estimated density 0.00423533722 or less.

Simple Statistics After Trimming

	Mean	Std Dev	Skewness	Kurtosis	Bimodality
OCUPADOS	33.9	33.1	1.7	2.5	0.7
REMUNERA	369.0	362.9	1.8	3.1	0.7
SALARIOS	113.2	149.7	2.7	8.4	0.7
CONSUMWH	113.2	183.9	2.3	4.5	0.8
COMBUSTI	9.9	22.1	4.1	18.7	0.8
PRODUCTO	878.6	1390.2	2.2	4.1	0.8
POTENCIA	12798.9	17602.5	2.3	5.7	0.7

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.76939	3.73255	0.681342	0.68134
2	1.03685	0.46162	0.148121	0.82946
3	0.57523	0.22498	0.082176	0.91164
4	0.35026	0.20415	0.050037	0.96168
5	0.14611	0.06371	0.020873	0.98255
6	0.08240	0.04264	0.011771	0.99432
7	0.03976	.	0.005680	1.00000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Distance Between Observations = 3.741657

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
67	OB43	OB59	2	0.000003	0.999997	
66	OB35	OB36	2	0.000003	0.999994	
65	OB68	OB75	2	0.000006	0.999988	
64	OB38	CL67	3	0.000009	0.999979	
63	CL66	OB80	3	0.000010	0.999969	
62	OB28	OB29	2	0.000015	0.999953	
61	OB8	CL65	3	0.000018	0.999935	
60	OB51	OB61	2	0.000020	0.999915	
59	OB55	OB79	2	0.000031	0.999885	
58	CL61	CL63	6	0.000036	0.999849	
57	OB41	OB62	2	0.000041	0.999808	

CUADRO 9 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Semipartial R-Squared	R-Squared	Tie
56	OB11	CL62	3	0.000048	0.999760	
55	OB37	OB69	2	0.000059	0.999701	
54	CL64	CL59	5	0.000070	0.999631	
53	OB33	OB63	2	0.000073	0.999559	
52	OB32	OB56	2	0.000082	0.999476	
51	OB66	OB77	2	0.000085	0.999392	
50	OB50	OB70	2	0.000108	0.999283	
49	CL56	CL60	5	0.000113	0.999170	
48	OB46	OB73	2	0.000134	0.999036	
47	OB7	CL55	3	0.000192	0.998844	
46	CL57	CL48	4	0.000209	0.998636	
45	CL52	CL53	4	0.000221	0.998414	
44	OB49	CL51	3	0.000253	0.998161	
43	OB15	CL50	3	0.000297	0.997864	
42	CL45	OB34	5	0.000299	0.997565	
41	CL46	OB42	5	0.000304	0.997261	
40	CL44	OB78	4	0.000305	0.996957	
39	OB39	OB64	2	0.000371	0.996585	
38	CL49	CL54	10	0.000378	0.996207	
37	CL43	OB20	4	0.000460	0.995747	
36	OB45	OB58	2	0.000591	0.995157	
35	OB25	CL40	5	0.000618	0.994538	
34	CL47	OB10	4	0.000634	0.993905	
33	OB52	OB65	2	0.000746	0.993159	
32	OB57	OB72	2	0.000877	0.992282	
31	CL37	OB67	5	0.000929	0.991352	
30	OB31	OB60	2	0.001065	0.990288	
29	CL34	CL38	14	0.001148	0.989140	
28	CL30	CL39	4	0.001415	0.987725	
27	CL31	CL41	10	0.001472	0.986253	
26	CL27	CL35	15	0.001888	0.984366	
25	OB6	OB19	2	0.002032	0.982334	
24	CL29	CL58	20	0.002229	0.980105	
23	OB2	OB40	2	0.002236	0.977869	
22	CL28	CL36	6	0.002958	0.974911	
21	CL23	OB30	3	0.003043	0.971868	
20	OB4	CL33	3	0.003206	0.968662	
19	CL24	CL42	25	0.003639	0.965023	
18	CL25	OB76	3	0.003883	0.961140	
17	OB16	OB17	2	0.003904	0.957236	
16	CL20	OB5	4	0.004854	0.952382	
15	CL21	CL32	5	0.005283	0.947099	
14	OB3	OB14	2	0.008289	0.938810	
13	CL15	CL22	11	0.008374	0.930436	
12	CL18	OB26	4	0.010562	0.919874	
11	OB1	CL14	3	0.010740	0.909134	
10	CL16	CL26	19	0.012260	0.896873	
9	CL13	CL10	30	0.015277	0.881597	
8	CL11	OB12	4	0.015544	0.866053	
7	OB13	OB48	2	0.022137	0.843915	
6	CL8	OB74	5	0.032080	0.811836	
5	CL9	CL19	55	0.048015	0.763821	
4	CL6	CL12	9	0.063638	0.700182	
3	CL4	CL17	11	0.088168	0.612014	
2	CL3	CL7	13	0.093362	0.518652	
1	CL2	CL5	68	0.518652	0.000000	

CUADRO 10. ANALISIS DE COMPONENTES PRINCIPALES

Principal Component Analysis, 80 Observations 7 Variables

Simple Statistics

	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH
Mean	74.7625000	784.175000	342.1625000	549.787500
Std	121.1093495	1161.487531	638.5971065	1977.946075

	COMBUSTI	PRODUCTO	POTENCIA
Mean	75.2500000	7025.58750	35598.01250
Std	329.9381031	31834.79902	75483.35406

Covariance Matrix

	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH
OCUPADOS	14667	132934	70868	78508
REMUNERA	132934	1349053	704070	987856
SALARIOS	70868	704070	407806	721633
CONSUMWH	78508	987856	721633	3912271
COMBUSTI	11768	152221	116070	632219
PRODUCTO	1113622	12311498	10532794	56268697
POTENCIA	7489574	75446971	39737652	48000574

	COMBUSTI	PRODUCTO	POTENCIA
OCUPADOS	11768	1113622	7489574
REMUNERA	152221	12311498	75446971
SALARIOS	116070	10532794	39737652
CONSUMWH	632219	56268697	48000574
COMBUSTI	108859	8648140	7356108
PRODUCTO	8648140	1013454429	337391708
POTENCIA	7356108	337391708	5697736740

Total Variance = 6716983825.9

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	5.7237E9	4.7314E9	0.852124	0.85212
PRIN2	9.9226E8	9.9158E8	0.147724	0.99985
PRIN3	681961.2	365465.6	0.000102	0.99995
PRIN4	316495.6	298561.6	0.000047	1.00000
PRIN5	17934.01	15381.74	0.000003	1.00000
PRIN6	2552.273	1427.645	0.000000	1.00000
PRIN7	1124.628	.	0.000000	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
OCUPADOS	0.001319	0.000579	-.041121	0.080596	0.042152	-.107316	0.989201
REMUNERA	0.013305	0.006941	-.211245	0.898036	-.355168	0.141003	-.051540
SALARIOS	0.007059	0.007730	-.122197	0.355878	0.841036	-.372547	-.110345
CONSUMWH	0.009077	0.053212	0.948546	0.234129	-.050310	-.199998	0.000758
COMBUSTI	0.001391	0.008182	0.191100	0.069285	0.402719	0.888696	0.081544
PRODUCTO	0.071596	0.995886	-.049922	-.023245	-.004773	0.005435	-.000067
POTENCIA	0.997277	-.072140	-.001578	-.015166	-.001032	0.001088	0.000044

CUADRO 10 (CONT)

Correlation Matrix

	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO	POTENCIA
OCUPADOS	1.0000	0.9450	0.9163	0.3277	0.2945	0.2888	0.8193
REMUNERA	0.9450	1.0000	0.9492	0.4300	0.3972	0.3330	0.8605
SALARIOS	0.9163	0.9492	1.0000	0.5713	0.5509	0.5181	0.8244
CONSUMWH	0.3277	0.4300	0.5713	1.0000	0.9688	0.8936	0.3215
COMBUSTI	0.2945	0.3972	0.5509	0.9688	1.0000	0.8234	0.2954
PRODUCTO	0.2888	0.3330	0.5181	0.8936	0.8234	1.0000	0.1404
POTENCIA	0.8193	0.8605	0.8244	0.3215	0.2954	0.1404	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.60642	2.69741	0.658060	0.65806
PRIN2	1.90901	1.62539	0.272716	0.93078
PRIN3	0.28362	0.17344	0.040518	0.97129
PRIN4	0.11019	0.06429	0.015741	0.98704
PRIN5	0.04589	0.00762	0.006556	0.99359
PRIN6	0.03827	0.03166	0.005467	0.99906
PRIN7	0.00659	.	0.000941	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7
OCUPADOS	0.390180	-.346815	-.370280	-.220455	0.735652	0.023969	-.003716
REMUNERA	0.415536	-.296955	-.112725	-.273073	-.515567	0.553429	0.282451
SALARIOS	0.443696	-.165715	-.176523	-.040260	-.397226	-.626462	-.436886
CONSUMWH	0.358166	0.448377	0.225100	-.008189	0.114503	0.446074	-.638628
COMBUSTI	0.343462	0.448733	0.381887	-.479235	0.090567	-.317468	0.442922
PRODUCTO	0.314897	0.477295	-.509580	0.557582	-.029292	0.000421	0.320342
POTENCIA	0.361989	-.365871	0.603211	0.576427	0.114225	-.030313	0.150647

CUADRO 11. ANALISIS DE COMPONENTES PRINCIPALES (SIN VARIABLE POTENCIA)

Principal Component Analysis, 80 Observations 6 Variables

Simple Statistics

	OCUPADOS	REMUNERA	SALARIOS
Mean	74.7625000	784.175000	342.1625000
Std	121.1093495	1161.487531	638.5971065

	CONSUMWH	COMBUSTI	PRODUCTO
Mean	549.787500	75.2500000	7023.58750
Std	1977.946075	329.9381031	31834.79902

Correlation Matrix

	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
OCUPADOS	1.0000	0.9450	0.9163	0.3277	0.2945	0.2888
REMUNERA	0.9450	1.0000	0.9492	0.4300	0.3972	0.3330
SALARIOS	0.9163	0.9492	1.0000	0.5713	0.5509	0.5181
CONSUMWH	0.3277	0.4300	0.5713	1.0000	0.9688	0.8936
COMBUSTI	0.2945	0.3972	0.5509	0.9688	1.0000	0.8234
PRODUCTO	0.2888	0.3330	0.5181	0.8936	0.8234	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	4.08344	2.45857	0.680573	0.680573
PRIN2	1.62487	1.43045	0.270811	0.951384
PRIN3	0.19442	.	0.032403	0.983787

Eigenvectors

	PRIN1	PRIN2	PRIN3
OCUPADOS	0.378308	-.485253	0.148573
REMUNERA	0.407149	-.427163	-.153292
SALARIOS	0.452188	-.294199	0.032039
CONSUMWH	0.419107	0.401156	-.208409
COMBUSTI	0.403763	0.405651	-.548030
PRODUCTO	0.384660	0.412346	0.780788

CUADRO 12. CLUSTER CON METODO AVERAGE (VARIABLES PRIN1 Y PRIN2)

Average Linkage Cluster Analysis

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.08344	2.43857	0.715350	0.71535
2	1.62487	.	0.284650	1.00000

Root-Mean-Square Total-Sample Standard Deviation = 1.689423
Root-Mean-Square Distance Between Observations = 3.378847

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
79	OB36	OB80	2	0.000071	
78	OB43	OB63	2	0.000495	
77	OB35	OB68	2	0.000574	T
76	CL78	OB59	3	0.000581	
75	OB66	OB77	2	0.001097	
74	OB28	OB27	2	0.001289	
73	OB33	OB79	2	0.001393	
72	CL77	CL79	4	0.001672	
71	OB11	CL74	3	0.001886	
70	OB60	OB62	2	0.002146	
69	OB37	OB69	2	0.002150	
68	OB49	OB73	2	0.002327	
67	OB38	OB55	2	0.002423	T
66	OB10	OB32	2	0.002424	
65	OB39	OB78	2	0.002483	
64	OB50	OB70	2	0.002705	
63	OB34	OB51	2	0.002728	
62	OB46	CL70	3	0.003215	
61	OB15	CL64	3	0.003342	
60	OB8	OB75	2	0.003469	
59	CL73	CL67	4	0.003914	
58	CL66	OB61	3	0.004075	
57	CL76	OB56	4	0.004483	
56	CL58	CL63	5	0.004946	
55	OB40	OB64	2	0.005175	
54	OB25	CL75	3	0.005184	
53	OB57	OB58	2	0.005711	
52	OB42	CL62	4	0.006775	
51	CL59	CL57	8	0.006947	
50	CL60	CL72	6	0.007560	
49	CL56	CL71	8	0.008060	
48	OB7	CL69	3	0.008152	
47	OB41	CL52	5	0.008243	
46	OB20	CL68	3	0.008361	
45	OB31	CL65	3	0.008605	
44	OB2	OB45	2	0.009294	
43	CL46	CL54	6	0.009497	
42	CL61	CL47	8	0.010310	
41	OB4	OB52	2	0.012005	
40	OB5	OB67	2	0.015033	
39	CL44	CL53	4	0.016251	
38	CL45	CL55	5	0.017459	
37	OB14	OB30	2	0.018071	
36	CL43	OB65	7	0.018432	

CUADRO 12 (CONT)

Number of Clusters	Clusters Joined		Frequency of New Cluster	Normalized RMS Distance	Tie
35	CL48	CL42	11	0.020670	
34	CL50	CL51	14	0.021816	
33	CL41	CL40	4	0.024562	
32	CL38	OB72	6	0.027108	
31	CL35	CL36	18	0.028392	
30	CL39	CL37	6	0.032008	
29	CL34	CL49	22	0.032510	
28	OB46	OB74	2	0.034589	
27	OB16	OB17	2	0.038587	
26	CL33	CL31	22	0.039019	
25	OB3	OB6	2	0.041387	
24	OB19	OB76	2	0.056290	
23	CL30	CL32	12	0.056701	
22	OB1	CL28	3	0.067479	
21	CL26	CL29	44	0.067708	
20	CL23	OB12	13	0.081271	
19	CL27	OB71	3	0.090637	
18	CL22	CL24	5	0.091646	
17	CL20	CL25	15	0.098291	
16	OB18	OB27	2	0.115596	
15	CL19	OB26	4	0.138866	
14	OB21	OB44	2	0.138973	
13	CL17	CL21	59	0.144126	
12	CL13	OB13	60	0.147056	
11	CL15	CL16	6	0.245218	
10	OB24	OB47	2	0.252209	
9	CL18	CL12	65	0.289133	
8	CL11	CL14	8	0.388140	
7	OB22	OB54	2	0.429290	
6	OB9	CL7	3	0.651473	
5	CL6	CL10	5	0.676147	
4	CL9	CL8	73	0.699506	
3	CL5	OB23	6	1.416715	
2	CL4	CL3	79	1.793287	
1	CL2	OB53	80	4.265830	

CUADRO 13. DETALLE DE LOS CLUSTERS FORMADOS CON AVERAGE

CLUSTER=1

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
36	-1.0408	0.44677	1	0	0	1	0	1
80	-1.0410	0.44661	1	0	0	0	0	4
43	-0.9757	0.38292	9	74	16	11	0	60
63	-0.9769	0.38415	8	79	17	10	0	39
35	-1.0364	0.44406	2	0	0	7	0	7
68	-1.0350	0.44547	2	0	0	1	2	20
59	-0.9752	0.38493	9	72	16	10	1	79
66	-0.7565	0.19739	30	322	82	41	6	582
77	-0.7571	0.20104	30	296	94	68	2	509
28	-0.9446	0.34801	12	120	26	3	0	80
29	-0.9437	0.35227	11	114	33	4	0	162
33	-0.9899	0.40198	9	44	6	17	2	40
79	-0.9934	0.40504	7	45	12	7	2	56
11	-0.9384	0.34860	8	161	30	10	0	85
60	-0.8283	0.24498	23	267	56	41	0	174
62	-0.8333	0.25022	23	258	53	20	3	260
37	-0.8716	0.28082	20	192	53	7	1	214
69	-0.8695	0.27385	21	209	44	13	1	54
49	-0.7784	0.20266	33	281	66	28	0	957
73	-0.7844	0.19766	26	288	103	14	0	139
38	-0.9851	0.39209	8	60	15	4	0	135
55	-0.9799	0.39842	8	60	14	6	5	82
10	-0.9084	0.33563	13	156	40	5	0	922
32	-0.9160	0.33850	13	150	34	13	6	65
39	-0.6872	0.12129	37	428	98	74	3	219
78	-0.6933	0.11556	38	404	107	30	0	698
50	-0.7978	0.24410	27	273	55	44	10	482
70	-0.8062	0.24042	25	252	76	27	8	186
34	-0.9254	0.33683	14	144	25	18	1	151
51	-0.9311	0.34409	13	126	31	21	1	58
46	-0.8212	0.25107	22	233	84	42	0	350
15	-0.7986	0.23250	18	392	46	27	8	317
6	-1.0177	0.42333	6	12	2	2	1	39
75	-1.0240	0.43322	3	16	4	3	1	44
61	-0.9216	0.32787	15	147	28	13	0	134
56	-0.9608	0.38404	9	90	18	9	7	39
40	-0.6281	0.11536	42	482	90	82	26	248
64	-0.6292	0.09791	39	458	137	89	10	372
25	-0.7519	0.18250	26	422	62	27	1	1017
57	-0.5428	-0.05488	74	477	117	52	0	760
58	-0.5364	-0.03665	50	635	156	53	0	601
42	-0.8256	0.22666	27	268	49	22	0	74
7	-0.8684	0.30455	19	199	37	38	8	217
41	-0.8481	0.25808	21	253	49	12	1	279
20	-0.7724	0.22666	29	238	101	17	16	250
31	-0.6671	0.13552	33	466	104	101	9	380
2	-0.5428	0.01598	50	476	202	159	1	31
45	-0.5198	-0.00540	45	534	236	89	0	881
4	-0.7906	0.34608	3	324	89	197	0	2142
52	-0.7951	0.30577	26	220	42	228	2	845
5	-0.8536	0.39542	12	144	48	129	30	385
67	-0.8568	0.34473	17	156	51	102	12	595
14	-0.4378	0.05712	47	535	242	175	6	4650
30	-0.4984	0.04924	50	598	136	259	22	156
65	-0.7329	0.25042	22	415	48	180	6	1496
72	-0.6656	0.03081	74	271	65	20	0	178
48	0.1441	-0.16646	100	945	352	588	115	2839
74	0.0419	-0.22308	117	1031	156	785	15	5244
3	-0.2953	-0.03276	57	706	271	428	11	2247
6	-0.3506	-0.16121	69	816	194	144	11	683

CUADRO 13 (CONT)

CLUSTER=1 (continued)

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
19	-0.1264	-0.31244	84	1015	302	173	21	1936
76	-0.0574	-0.49003	94	1099	400	81	0	623
1	-0.0982	-0.08506	65	796	383	420	29	5649
12	-0.4302	0.25423	36	459	190	561	37	3473
13	-0.4710	0.53781	22	301	97	523	128	5199

CLUSTER=2

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
16	0.6305	-0.68439	125	1500	716	747	47	2943
17	0.6731	-0.80759	142	1467	777	539	36	4224
71	0.8448	-0.97464	150	1812	802	788	26	1537
18	1.4773	-1.18881	189	2263	1017	597	176	6271
27	1.2113	-0.90276	219	1910	566	1817	68	2720
26	0.2722	-0.84054	134	1621	418	43	5	1518
21	1.9419	-1.55585	302	1892	1264	1362	54	10742
44	1.5842	-1.86004	261	3064	752	471	18	7010

CLUSTER=3

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
24	3.6193	-0.32623	211	2566	1866	1725	0	117361
47	4.4659	-0.42581	262	3141	1920	3713	1071	11062

CLUSTER=4

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
22	4.4482	-2.73076	515	4024	1827	2114	72	52999
54	3.1395	-2.09909	277	3637	1824	1946	250	2784

CLUSTER=5

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
9	5.8182	-1.95095	354	5423	2410	5006	734	15738

CLUSTER=6

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
23	6.1472	-5.71880	760	5702	3302	205	195	16420

CLUSTER=7

OBS	PRIN1	PRIN2	OCUPADOS	REMUNERA	SALARIOS	CONSUMWH	COMBUSTI	PRODUCTO
53	11.6802	7.94302	176	2206	2122	16542	2680	257655

CUADRO 13 (CONT)

CLUSTER=1

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
65	PRIN1	65	-1.041	0.144	-0.738	0.277
	PRIN2	65	-0.490	0.538	0.217	0.204
	OCUPADOS	65	1.000	117.000	29.292	25.571
	REMUNERA	65	0.000	1099.000	315.446	267.630
	SALARIOS	65	0.000	400.000	89.077	94.153
	CONSUMWH	65	0.000	788.000	97.969	161.196
	COMBUSTI	65	0.000	126.000	9.046	21.869
	PRODUCTO	65	1.000	5649.000	785.585	1329.088

CLUSTER=2

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
8	PRIN1	8	0.272	1.942	1.079	0.567
	PRIN2	8	-1.860	-0.684	-1.102	0.409
	OCUPADOS	8	125.000	302.000	190.250	65.128
	REMUNERA	8	1467.000	3064.000	1941.125	521.661
	SALARIOS	8	418.000	1264.000	789.000	259.667
	CONSUMWH	8	43.000	1817.000	795.500	553.372
	COMBUSTI	8	5.000	176.000	53.750	53.363
	PRODUCTO	8	1518.000	10742.000	4620.625	3198.732

CLUSTER=3

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
2	PRIN1	2	3.619	4.466	4.043	0.599
	PRIN2	2	-0.426	-0.328	-0.377	0.069
	OCUPADOS	2	211.000	262.000	236.500	36.062
	REMUNERA	2	2566.000	3141.000	2853.500	406.586
	SALARIOS	2	1866.000	1920.000	1893.000	38.184
	CONSUMWH	2	1725.000	3713.000	2719.000	1405.728
	COMBUSTI	2	0.000	1071.000	535.500	757.311
	PRODUCTO	2	11062.000	117361.000	64211.500	75164.744

CLUSTER=4

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
2	PRIN1	2	3.139	4.445	3.792	0.923
	PRIN2	2	-2.731	-2.099	-2.415	0.447
	OCUPADOS	2	277.000	515.000	396.000	168.291
	REMUNERA	2	3637.000	4024.000	3830.500	273.650
	SALARIOS	2	1824.000	1827.000	1825.500	2.121
	CONSUMWH	2	1946.000	2114.000	2030.000	118.794
	COMBUSTI	2	72.000	250.000	161.000	125.865
	PRODUCTO	2	2784.000	52999.000	27891.500	35507.367

CUADRO 13 (CONT)

CLUSTER=5

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
1	PRIN1	1	5.818	5.818	5.818	.
	PRIN2	1	-1.951	-1.951	-1.951	.
	OCUPADOS	1	354.000	354.000	354.000	.
	REMUNERA	1	5423.000	5423.000	5423.000	.
	SALARIOS	1	2410.000	2410.000	2410.000	.
	CONSUMWH	1	5006.000	5006.000	5006.000	.
	COMBUSTI	1	734.000	734.000	734.000	.
	PRODUCTO	1	15738.000	15738.000	15738.000	.

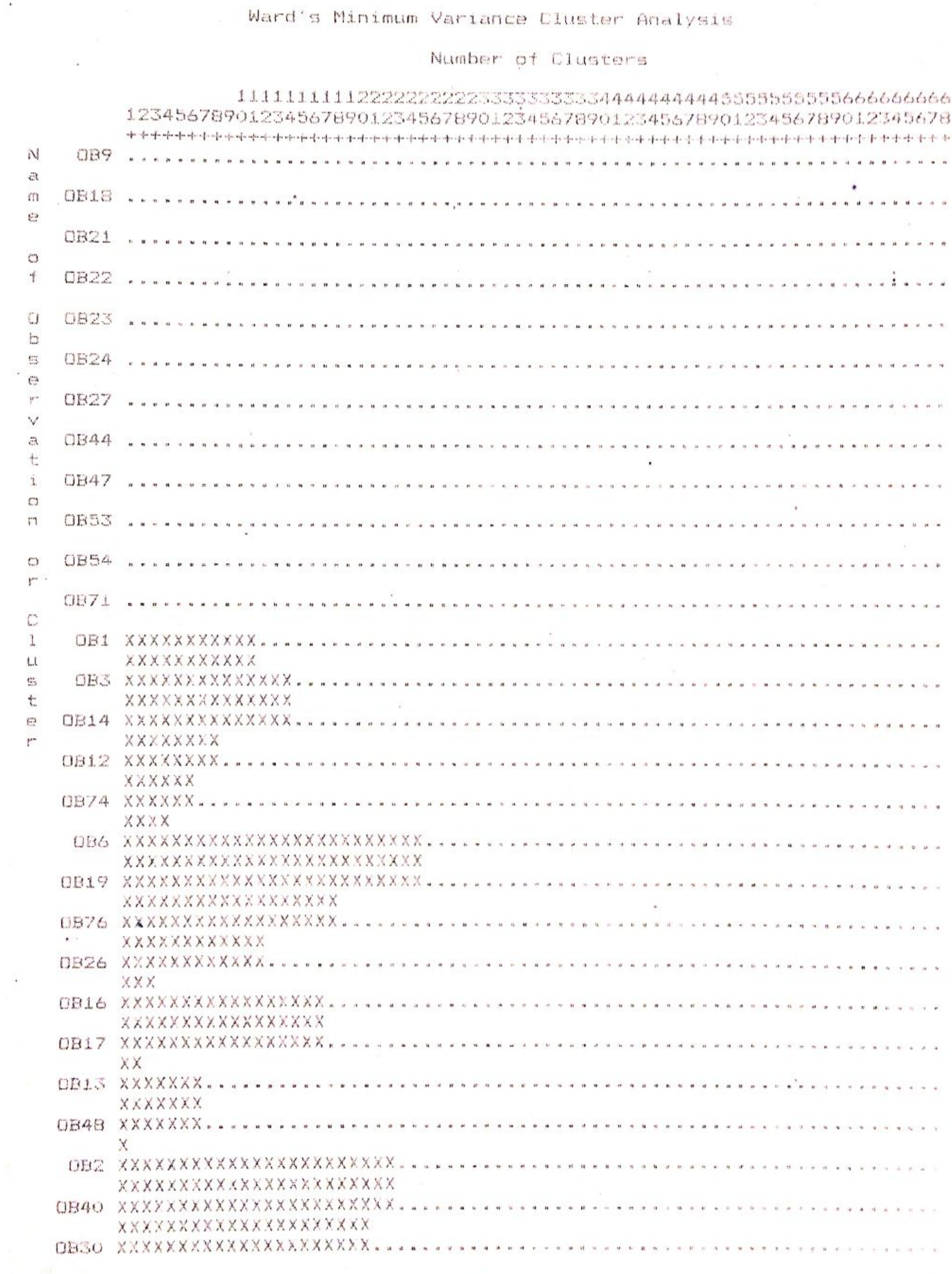
CLUSTER=6

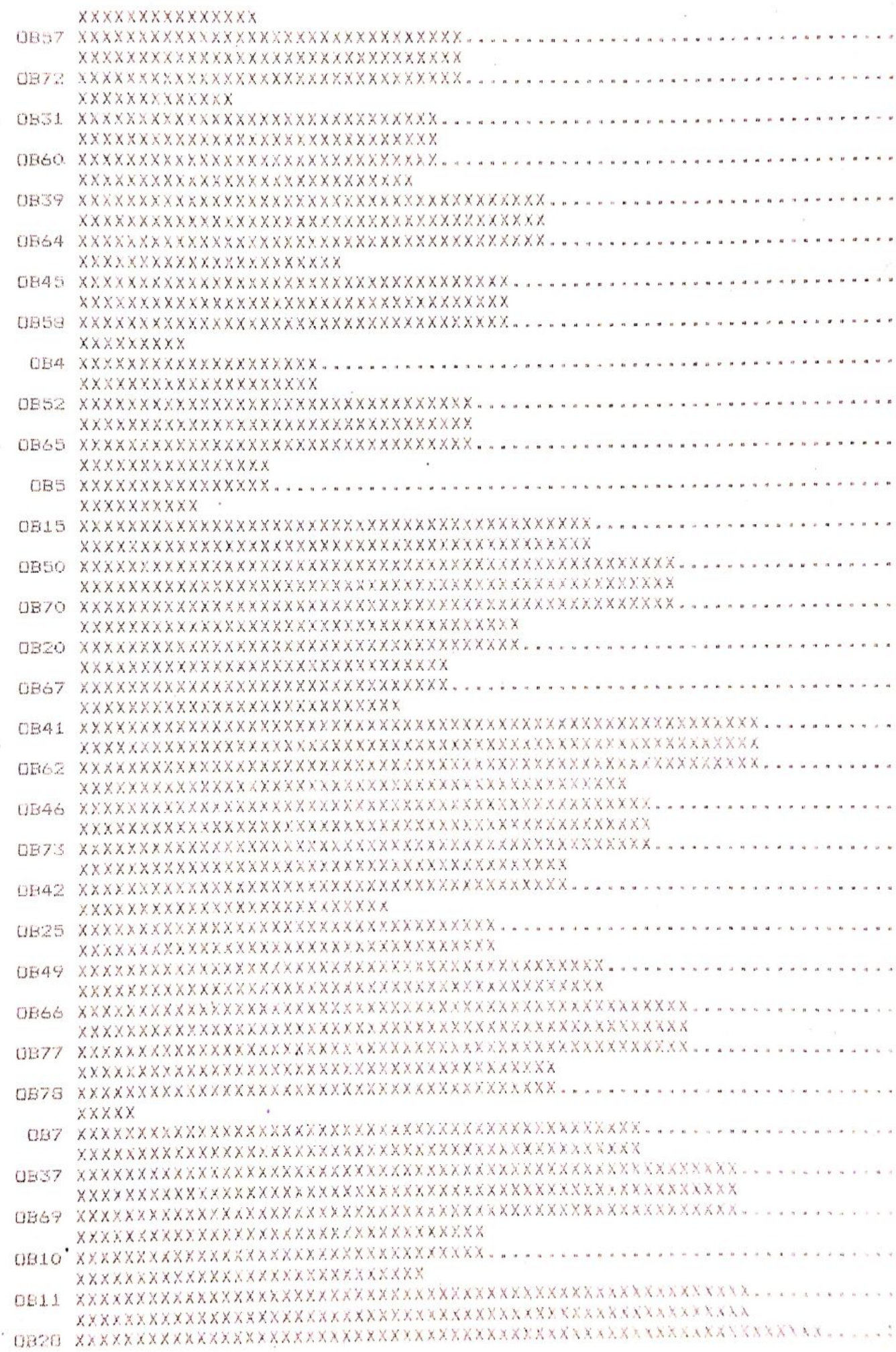
N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
1	PRIN1	1	6.147	6.147	6.147	.
	PRIN2	1	-5.719	-5.719	-5.719	.
	OCUPADOS	1	760.000	760.000	760.000	.
	REMUNERA	1	5702.000	5702.000	5702.000	.
	SALARIOS	1	3302.000	3302.000	3302.000	.
	CONSUMWH	1	205.000	205.000	205.000	.
	COMBUSTI	1	195.000	195.000	195.000	.
	PRODUCTO	1	16420.000	16420.000	16420.000	.

CLUSTER=7

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
1	PRIN1	1	11.680	11.680	11.680	.
	PRIN2	1	7.943	7.943	7.943	.
	OCUPADOS	1	176.000	176.000	176.000	.
	REMUNERA	1	2208.000	2208.000	2208.000	.
	SALARIOS	1	2122.000	2122.000	2122.000	.
	CONSUMWH	1	16542.000	16542.000	16542.000	.
	COMBUSTI	1	2680.000	2680.000	2680.000	.
	PRODUCTO	1	257655.000	257655.000	257655.000	.

GRAFICO 1. TREE A PARTIR DE RESULTADOS CON WARD





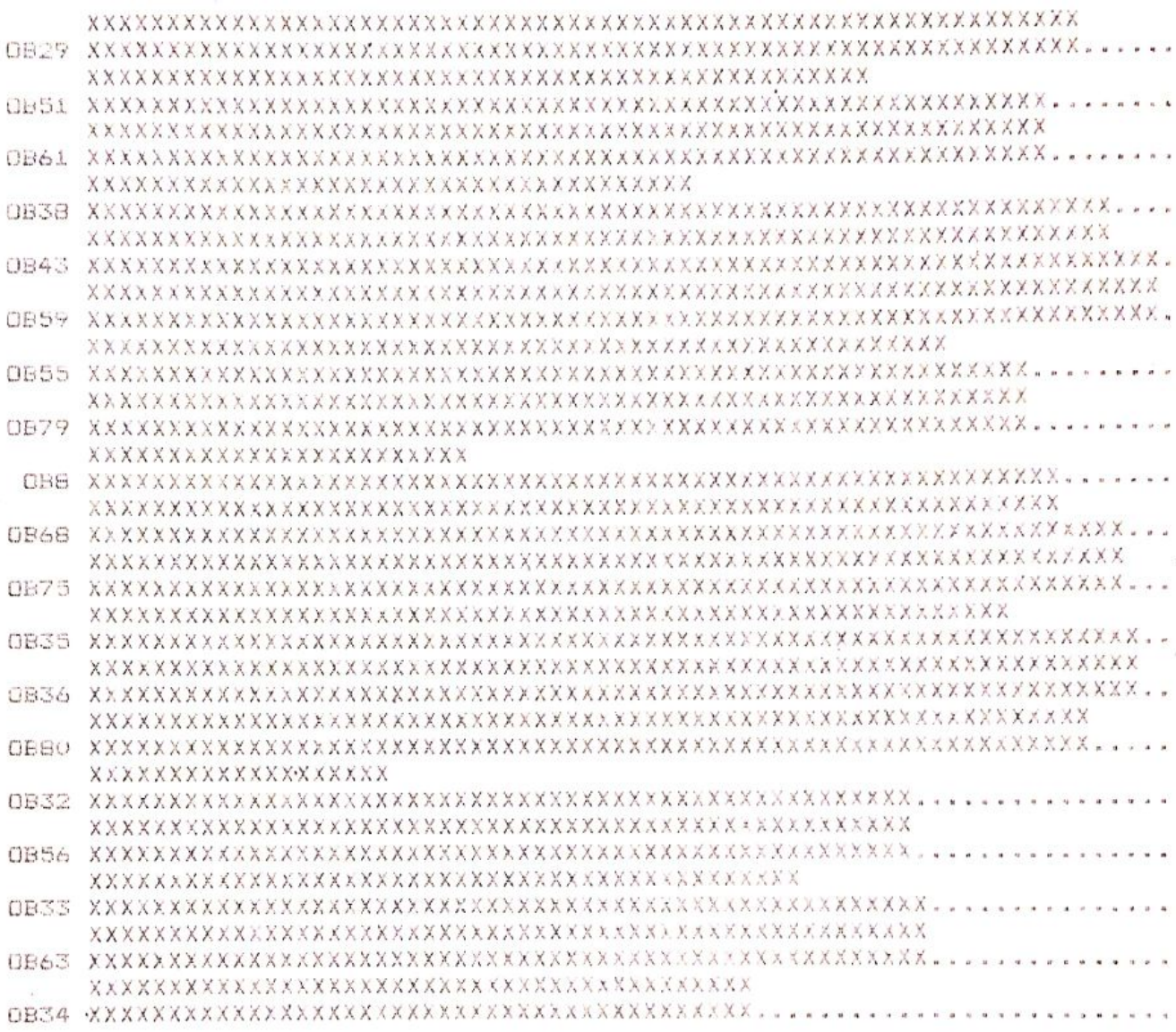


GRAFICO 2. REPRESENTACION DE LOS DATOS SEGUN COMPONENTES PRINCIPALES

Plot of PRIN2*PRIN1. Legend: A = 1 obs, B = 2 obs, etc.

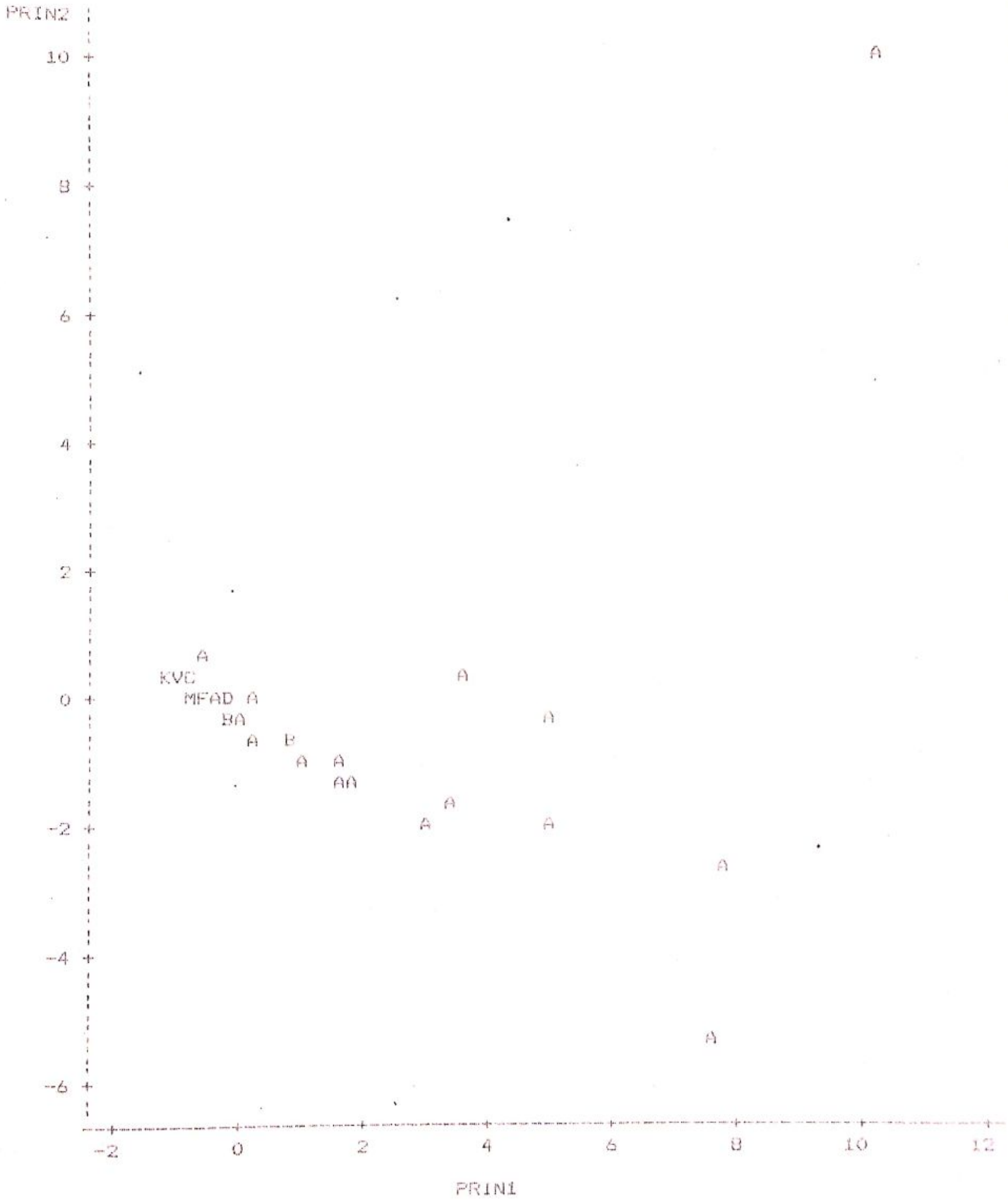
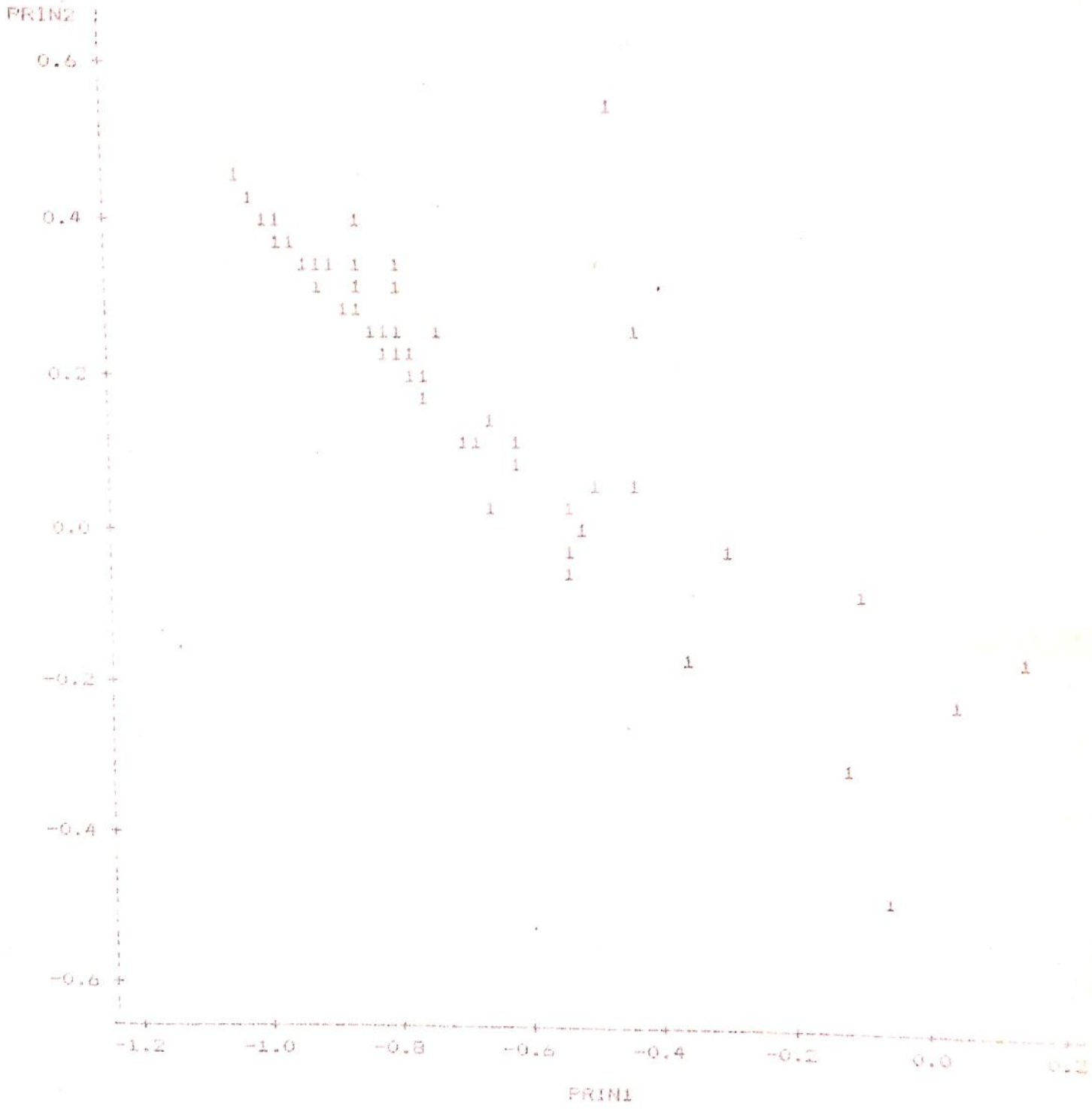


GRAFICO 3. REPRESENTACION DE LAS OBSERVACIONES DE CADA CLUSTER

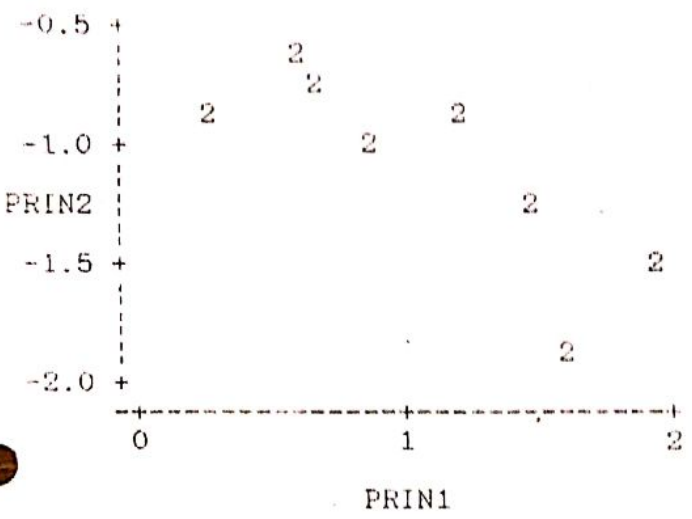
CLUSTER=1

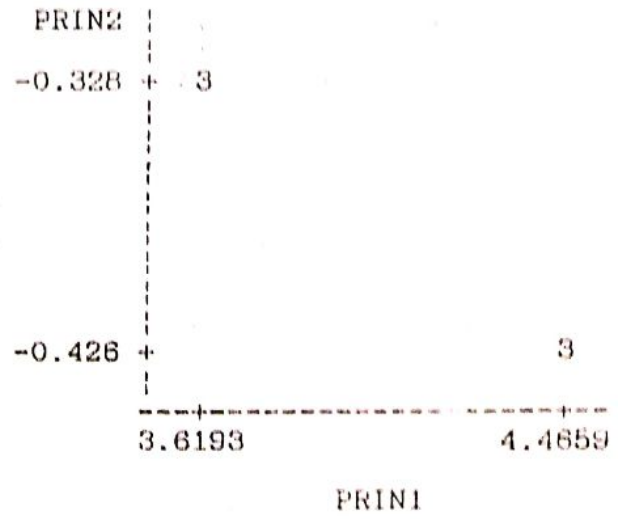


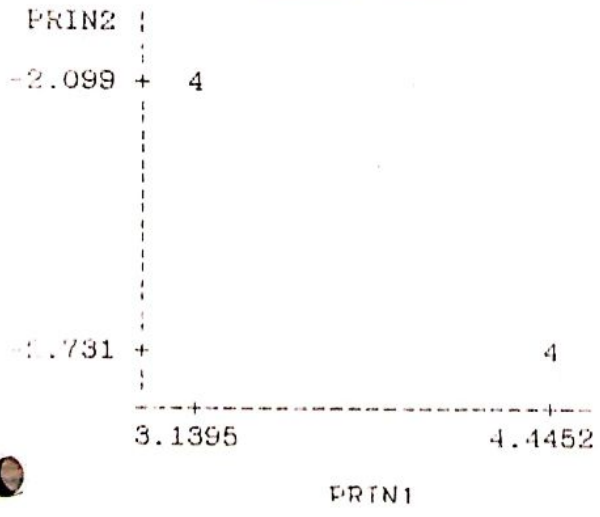
NOTE: 17 obs hidden.

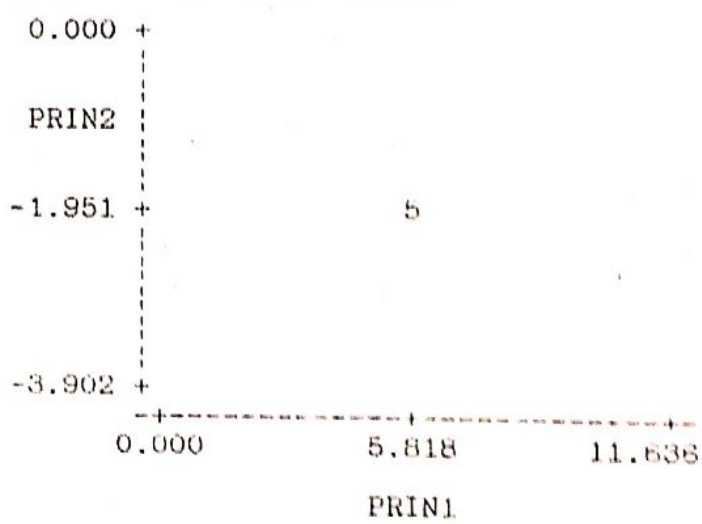
GRAFICO 3 (CONT)

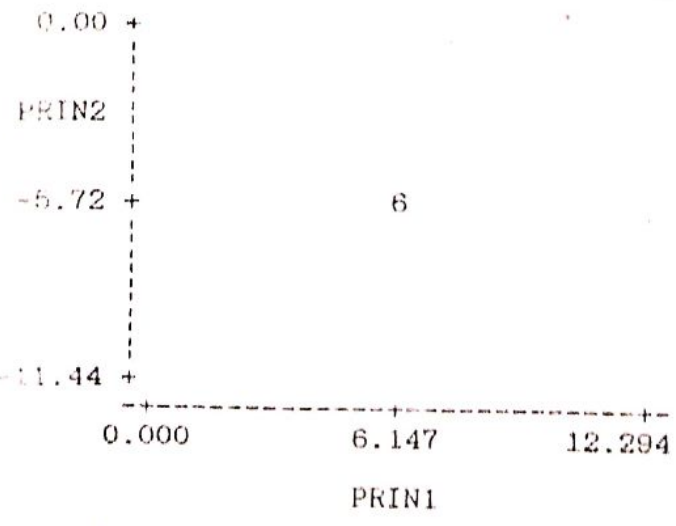
Plot of PRIN2*PRIN1=CLUSTER.











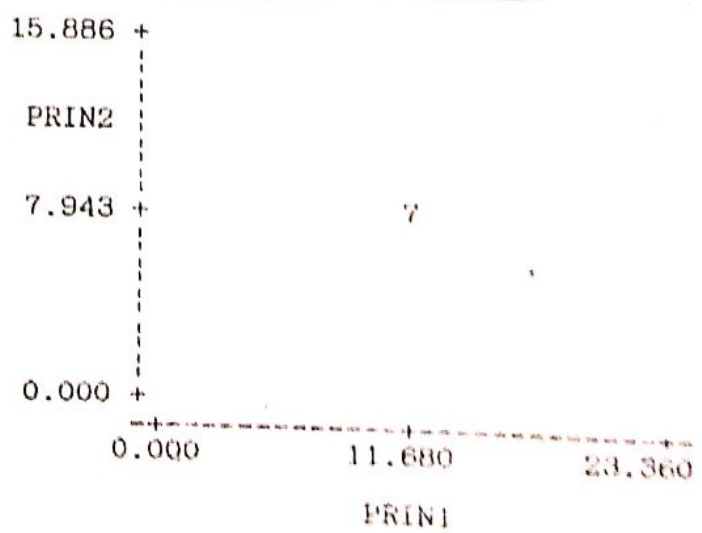
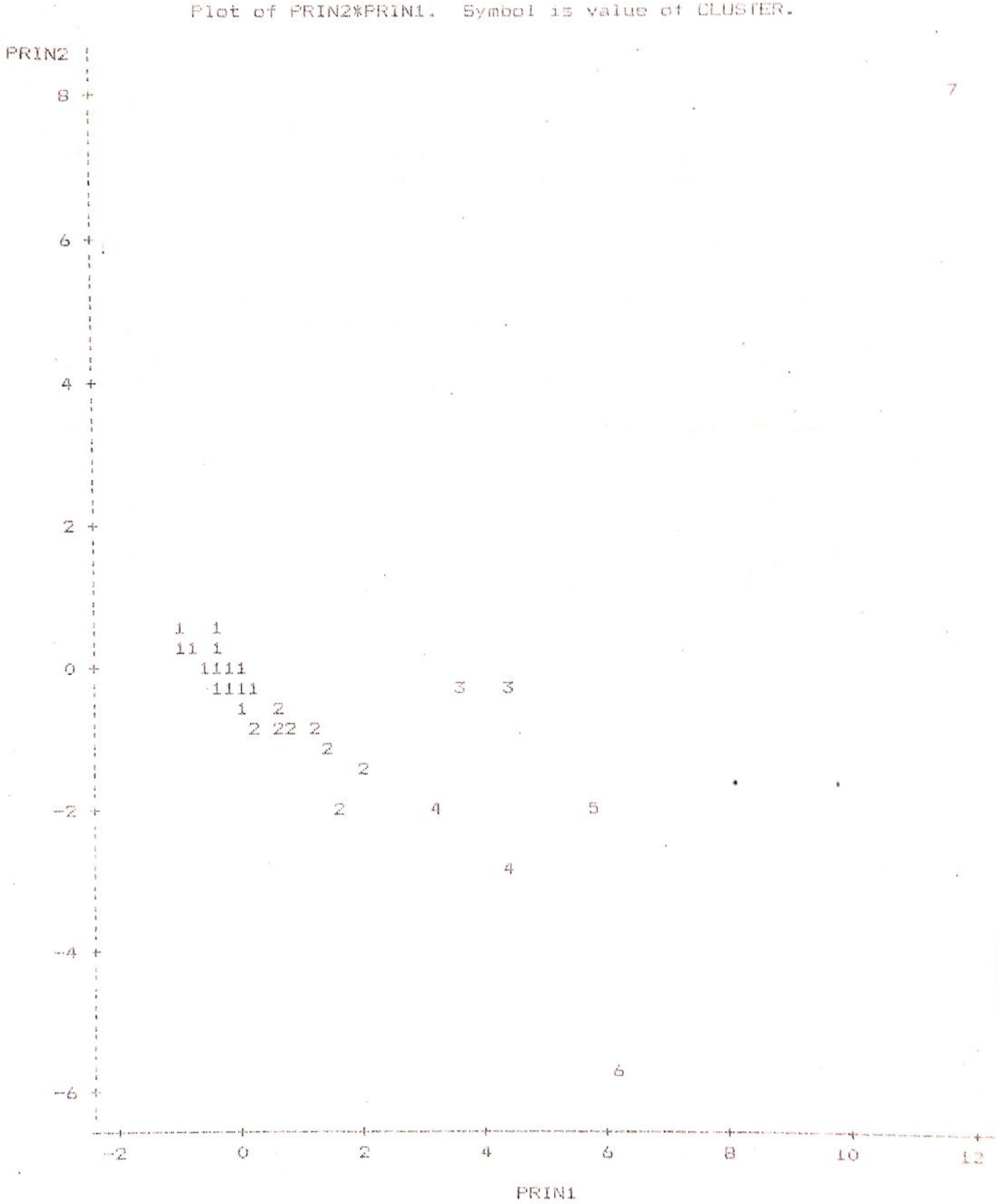


GRAFICO 4. REPRESENTACION CONJUNTA DE LAS OBSERVACIONES



REFERENCIAS BIBLIOGRAFICAS

CHARRE DE TRABUCHI, C. y otras "Encuesta Permanente de Hogares. Diseño de las muestras" INDEC. Buenos Aires, sin mención del año de edición.

ALTIMIR, O. y otros "La pobreza en la Argentina" INDEC. Buenos Aires, 1984.

ROMERO VILLAFRANCA, R. "Introducción a los métodos de análisis estadistico multivariante" IRICE. Rosario, 1982.

SAS INSTITUTE "SAS Introductory Guide for Personal Computers, Version 6 Edition" Cary, USA, 1985.

SAS INSTITUTE "SAS/STAT User's Guide, Release 6.03 Edition" Cary, USA, 1988.

TATSUOKA, M. "Multivariate Analysis: Techniques for Educational and Psychological Research" J. Wiley and Sons. New York, 1971.

YOGUEL , G. y GATTO, F. "La problemática de las pequeñas y medianas empresas industriales: algunos aspectos metodológicos aplicados al caso argentino" Convenio de Cooparación Técnica CFI-CEPAL, Doc. de Trabajo No 18. Buenos Aires, 1989.

Universidad Nacional de Salta
Facultad de Ciencias Económicas
Jurídicas y Sociales
Instituto de Investigaciones Económicas

REUNIONES DE DISCUSION

Nro.	Fecha	Autor	Titulo
52	21/05/90	Eduardo Antonelli	"Un Modelo Postkeynesiano Dinámico II"
53	28/05/90	Jorge Paz	"Contenido Directo de Factores y Exportaciones Industriales: Algunas Evidencias sobre el Caso Argentino"
54	19/06/90	Norma Cecilia Mena de Méndez	"La Distribución del Ingreso: Algunas Reflexiones Teóricas"
55	11/07/90	Eduardo Antonelli	"Desequilibrios Externo y Fiscal e Inflación: Un Enfoque Postkeynesiano"
56	20/12/90	Eduardo Antonelli	"Nivel de Precios, Distribución del Ingreso e Inflación"
57	7/ 3/91	Guillermo J. Lloret	"Ensayo sobre un Modelo de Política Fiscal Antiinflacionaria mediante un Impuesto a Tasa Progresiva sobre el Incremento Acumulado del Precio"
58	23/ 5/91	Eusebio C. del Rey	"Erradicación del Mal de Chagas: Análisis de los Costos"
59	20/ 6/91	Eduardo Antonelli	"Inflación: Análisis y Evidencia Empirica" (Versión Preliminar)
60	19/ 8/91	Jorge Augusto Paz	"Variables Asociadas al Crecimiento Económico: Una Evaluación Empírica"
61	11/12/91	Juan Carlos Cid	"Técnicas de Clustering: Un Ejercicio de Aplicación"