PROYECTO FINAL DE INGENIERÍA INDUSTRIAL

"Segmentación Estadística de una Base de Datos Financiera"

Autor: Martín Decia

Director de Proyecto: Ing. Mariano F. Bonoli

Año 2012

Dedicatoria

A todos aquellos que ayudaron directa o indirectamente a mi desarrollo como profesional.

A mi familia.

Resumen ejecutivo

El presente documento consiste en un análisis estadístico de una cartera de clientes de una entidad financiera en Argentina, mediante el proceso de segmentación estadística de una base de datos.

El objetivo principal del trabajo reside en reconocer la importancia y riqueza de segmentar un mercado a través de herramientas y métodos estadísticos, para poder identificar y satisfacer la existencia de diferentes necesidades y comportamientos en los grupos de clientes, permitiendo así establecer propuestas comerciales y de marketing, como estrategias para atender las necesidades de los conglomerados de manera homogénea, aumentando su rentabilidad y reduciendo costos y riesgos para la entidad.

La motivación para abordar esta temática resulta en el verdadero desafío que surge de tratar a los clientes como grupos homogéneos, debido a la diversidad y complejidad de factores que hacen a la calidad consistente en la prestación de un servicio como los niveles de demanda, la influencia entre usuarios y la percepción de cada cliente.

En una primera parte se busca introducir al lector en el tema mediante la explicación de conceptos y métodos asociados al proceso de segmentación. A continuación se prosigue con mencionar el marco de normativas en el país asociadas al manejo de la información de los usuarios y algunos antecedentes internacionales en la aplicación de procesos estadísticos con objetivos financieros.

Luego, se procede a explicar y desarrollar cada una de las etapas de la segmentación estadística de una base de datos entre los cuales se pueden destacar procesos como “data mining”, el armado de la vista minable y la selección y aplicación de un software de segmentación estadística, el SPSS.

Por otro lado, para la determinación de la metodología a aplicar se decide combinar dos tipos de métodos de segmentación el jerárquico y el no-jerárquico, con la intención de combinar las fortalezas de ambos algoritmos para lograr una interpretación más valiosa de la cartera de clientes.

Finalmente, se busca explicar y comprender las características de los conglomerados generados, buscando posicionarlas de manera cualitativa en ejes de variables de comportamiento financiero. Así también, surge poder dimensionar un sector de posición del cliente financieramente atractivo para la entidad y que acciones y estrategias implementar para poder fortalecer su fidelidad con la institución y ajustar los productos a su capacidad de pago, reduciendo así el nivel de riesgo de incobrabilidad.

Descriptor bibliográfico

Este trabajo tiene por finalidad un análisis estadístico de una cartera de clientes de una entidad financiera en Argentina.

A partir del relevamiento sobre el marco conceptual, legal y antecedentes en mercados financieros, se procede a desarrollar cada una de las etapas del proceso estadístico a través de un software y una metodología de combinar algoritmos de segmentación. Se destacan las variables de comportamiento y posiciones de cada uno de los conglomerados resultantes y sus respectivos beneficios. Luego, considerando también modelos de experiencias nacionales e internacionales, se establecen los lineamientos estratégicos de desarrollo futuro.

Palabras clave: conglomerado, segmentación jerárquica, K-medias, componentes principales.

Executive brief

The following paper will consist in a statistical analysis of a client portfolio from an Argentinean financial entity, through the process of statistical segmentation of a database.

The main purpose is to recognize the importance and value in the segmentation of a market through statistical tools and methods, in order to identify and satisfy the existence of different needs and behaviours in the groups of clients, establishing commercial and marketing proposals, like strategies to attend the clusters needs in an homogeneous way, raising profitability and reducing costs and risks to the entity.

The motivation to address this issue resides on the true challenge that arises from treating clients as homogeneous groups due to the diversity and complexity of the factors that make the quality of providing a service like the demand levels, the influence between users and the perception of the customer.

The objective of the first stage is to introduce the reader in the current theme through the explanation of concepts and methods related to the segmentation process. The following point continues introducing legal framework related with the user's information management and some international backgrounds on the application of statistical process with financial purposes.

Then, the paper proceeds to explain and develop every stage of the statistic segmentation of a data base where takes place process like data mining, the assembly of the minable view and the selection and application of a statistical segmentation software, SPSS.

Furthermore, to determine the methodology to be applied is decided to combine two types of segmentation methods on hierarchical and nonhierarchical, with the purpose of combining the strengths of both algorithms to achieve a more valuable interpretation of the customer portfolio.

Finally, it seeks to explain and understand the characteristics of generated clusters, intending to position them on a qualitative way in lines of performance variables. Also, comes to sizing a sector client's position financially attractive to the entity and which actions and strategies to implement in order to strengthen their loyalty to the institution and adjusting the products to their ability to pay, thus reducing the level of non-collection.

Abstract

This paper aims to conduct a statistical analysis of a client portfolio from an Argentinean financial entity.

From conceptual and legal framework and financial market backgrounds, proceeds to develop each stage of the process through a statistical software and a methodology to combine segmentation algorithms. It highlights the behavior variables and positions of each of the resulting clusters and their respective benefits. Then, considering also models of national and international experiences sets the strategic guidelines for future development.

Key words: cluster, hierarchical segmentation, K-means, main components

Tabla de contenidos

Pág.

1. Introducción

1.1 Caracterización y Objetivos del Proyecto Final

El presente trabajo de tesis consiste en un análisis estadístico de una cartera de clientes de una entidad financiera mediante el proceso de segmentación estadística de una base de datos.

La mayor dificultad que se plantea en términos de atención al cliente cuando se trabaja con grandes masas de datos, es el no poder identificar sus características principales, por ende es importante comprender el concepto de segmentación.

Segmentar un mercado implica reconocer la existencia de diferentes individuos en la composición de una muestra y que estos reaccionan de forma diferente a las propuestas comerciales y de marketing.

Cuando una entidad reconoce que los servicios que ofrece no sirven de la misma forma ni con la misma eficacia las expectativas de todos sus clientes, nace la necesidad de la segmentación.

Aquí se presenta la primera dificultad de segmentar para una entidad financiera ya que debe poder establecer grupos homogéneos de clientes, sin embargo, la naturaleza propia de los servicios es la heterogeneidad. Debido a que la calidad consistente en la prestación de un servicio depende de factores no totalmente controlables por la empresa como los niveles de demanda, la influencia entre usuarios y la percepción de cada cliente, el encontrar una manera diferente de acercarse a ellos tratándolas como grupos homogéneos plantea un verdadero desafío.

Es por ello que resulta fundamental que las organizaciones deben ser capaces de clasificar correctamente a sus potenciales clientes, pudiendo así determinar los servicios que se adapten mejor a sus características y a sus necesidades económicas.

En resumen, es importante que los diferentes segmentos cumplan con los siguientes requisitos:

- Ser internamente homogéneos.
- Ser diferente de los demás en cuanto a su reacción ante acciones comerciales.
- Ser fácilmente identificables.
- Ser accesible en cuanto a grupo.

Esta segmentación proporciona diferentes ventajas a las entidades financieras:

- Identificar segmentos de población en los que el presupuesto del sector de marketing tenga un mayor rendimiento.

- Adaptar la oferta de servicios según las características de la población que se quiere alcanzar.

- Aislar los grupos en los que la acción de ventas va a tener mayor incidencia e impacto.

- Reconocer los segmentos satisfechos y mercados saturados en la actualidad (importante para el lanzamiento de nuevos productos)

1.2 Etapas del informe

En esta sección se mencionan los capítulos que conforman el trabajo de tesis y se brinda un breve resumen de su contenido.

En primer lugar, la sección Marco conceptual se describe brevemente una introducción a los métodos estadísticos de segmentación.

Luego, en Marco legal se enumera y describe toda aquella normativa asociada al tema de estudio.

A continuación los antecedentes que describen precedentes nacionales e internacionales de aplicar métodos estadísticos de segmentación en entidades financieras.

Luego, en la descripción de la investigación se procederá a cada una de las etapas de la segmentación estadística de una base de datos.

Tras las etapas, se proponen cuales serían las estrategias a llevar a cabo indicando su forma de aplicación y los resultados esperados, según el cluster a apuntar.

A continuación se detallan las conclusiones rescatadas del análisis realizado (Sección Conclusiones).

Por último, se incluyen las secciones correspondientes a la Bibliografía y Anexos.

2. Marco conceptual

Mediante la correcta aplicación de los métodos “clustering” se permite comprender el comportamiento, características de los clientes dentro de una base de datos agrupándolos en segmentos.

El análisis Cluster en la investigación de mercados es usado para la segmentación de mercados; comprensión del comportamiento del comprador (identificación de grupos de compradores homogéneos para analizar el comportamiento de cada grupo por separado); identificar oportunidades para nuevos productos, seleccionar mercados de prueba, reducción de datos con el fin de facilitar el manejo de la información.

Por ello, es también conocido como análisis de clasificación o taxonomía numérica. La clusterización es un conjunto de técnicas estadísticas utilizadas para clasificar los objetos o casos en grupos homogéneos llamados conglomerados (clusters) con respecto a algún criterio de selección predeterminado. Los objetos dentro de cada grupo (conglomerado), son similares entre sí (alta homogeneidad interna) y diferentes a los objetos de los otros conglomerados o clusters (alta heterogeneidad externa). Es decir, que si la clasificación hecha es óptima, los objetos dentro de cada cluster estarán cercanos unos de otros y los cluster diferentes estarán muy apartados

Los algoritmos de formación de conglomerados se agrupan en dos categorías:

- Segmentación K-medias: Método de dividir el conjunto de observaciones en k conglomerados, en que k lo define inicialmente el usuario.
- Segmentación Jerárquica: Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados

3. Marco Legal

NORMA	TEMA
C.N. ARTICULO 43	Toda persona podrá tomar conocimiento de los datos a ella referidos y de su finalidad, que consten en registros o banco de datos públicos, o los privados destinados a proveer informes, y en caso de falsedad o discriminación, para exigir la supresión, rectificación, confidencialidad o actualización de aquéllos
LEY Nº 17.622 - 1968	Creación del Sistema Estadístico Nacional.
LEY 25.326 – Año 2000	Protección de los Datos Personales
LEY 26.343 – Año 2008	Modificación. Incorporación del Art. 47 a la Ley 25.326 de Protección de los Datos Personales.
LEY 26.388 – Año 2008	Reforma del Código Penal en materia de Delitos Informáticos.
Decreto nacional 1.558/2001	Principios generales relativos a la protección de datos. Derechos de los titulares de los datos. Usuarios y responsables de archivos, registros y bancos de datos. Control. Sanciones.
Disposición Nº 2/2003	Registro Nacional de Bases de Datos. Habilitase el mencionado Registro y dispónese la realización del Primer Censo Nacional de Bases de Datos.
Disposición Nº 2/2005	Implementación del Registro Nacional de Bases de Datos Privadas alcanzadas por la Ley Nº 25.326. Formularios de inscripción.
Constituciones Provinciales - Artículo 16.	Ciudad Autónoma de Buenos Aires- Libre acceso de la persona a su archivo en banco de datos. También actualización, rectificación, confidencialidad o supresión, cuando esa información lesione o restrinja algún derecho
Ley Provincial Nº 1.845.	Ley de Protección de Datos Personales.
Documentación Internacional	Comisión de las Comunidades Europeas. Decisión C (2003) 1731. Decisión sobre la adecuación de la protección de los datos personales en Argentina.

Tabla 1. Normativa asociada como marco legal

- **LEY 25.326 Incluye artículos vetados por Decreto Nº 955/2000 y las modificaciones introducidas por las Leyes 26.343 y 26.388**

ARTÍCULO 11.- (Cesión).

3. El consentimiento no es exigido cuando:

e) Se hubiera aplicado un procedimiento de disociación de la información, de modo que los titulares de los datos sean inidentificables.

ARTICULO 26.- (Prestación de servicios de información crediticia).

1. En la prestación de servicios de información crediticia sólo pueden tratarse datos personales de carácter patrimonial relativos a la solvencia económica y al crédito, obtenidos de fuentes accesibles al público o procedentes de informaciones facilitadas por el interesado con su consentimiento.

ARTICULO 28.- (Archivos, registros o bancos de datos relativos a encuestas).

2. Si en el proceso de recolección de datos no resultara posible mantener el anonimato, se deberá utilizar una técnica de disociación, de modo que no permita identificar a persona alguna.

Esto quiere decir que el proyecto cumple con la normativa vigente dado que el procedimiento de segmentación que se llevará a cabo es meramente estadístico y dichos datos personales no pueden asociarse al titular ni permitir, por su estructura, contenido o grado de desagregación, la identificación de aquél.

4. Antecedentes

- Federación Latinoamericana de Bancos – FELABAN

Introduce el concepto de segmentación de mercado de clientes con el fin de detectar actitudes sospechosas relacionadas con diversos factores de riesgo como el financiamiento del terrorismo, el lavado de dinero y potenciales fraudes

Esta aplicación de la segmentación se pone en manifiesto en la definición del concepto para FELABAN

"El objetivo principal de la Segmentación del Mercado es analizar las operaciones de un cliente para definir si son o no son sospechosas. El sistema de detección de operaciones mediante segmentación del mercado se basa en los siguientes principios:"

Cada segmento o grupo de operaciones debe corresponder a un grupo de clientes que tiene características comunes.

Los clientes que realizan normalmente operaciones en un determinado segmento, no tienen justificación financiera para realizar operaciones en ciertos segmentos (segmentos superiores).

Si un cliente cambia a un segmento diferente, esto necesariamente se debe a un cambio en su actividad económica.

Ciertos cambios de segmento, por ejemplo cuando disminuye el volumen de las operaciones (segmentos inferiores), no son inusuales.

En términos sencillos, segmentar el mercado significa definir criterios relevantes mediante los cuales se pueden agrupar las operaciones activas, pasivas y neutras del sujeto obligado. Existen muchas formas diferentes de realizar este control, cada entidad, de acuerdo con sus características propias debe definir la forma más efectiva de segmentar el mercado para establecer controles a las operaciones inusuales."

- Articulo de Luis Bernardo Quevedo, Director Unidad de Control de Cumplimiento del Banco de Bogotá – Colombia. Presidente del COPLAFT1

[1] COMITÉ LATINOAMERICANO PARA LA PREVENCIÓN DEL LAVADO DE ACTIVOS Y EL FINANCIAMIENTO DEL TERRORISMO (COPLAFT)

“Es indispensable desarrollar procesos de segmentación de cada una de las fuentes de riesgo principales, con el propósito de dividir el comportamiento de cada uno de ellos en grupos que tengan características similares. De esta manera, es más fácil identificar y asociar los riesgos a un solo grupo de individuos dentro del universo del factor.”

5. Descripción del proceso de Segmentación Estadística

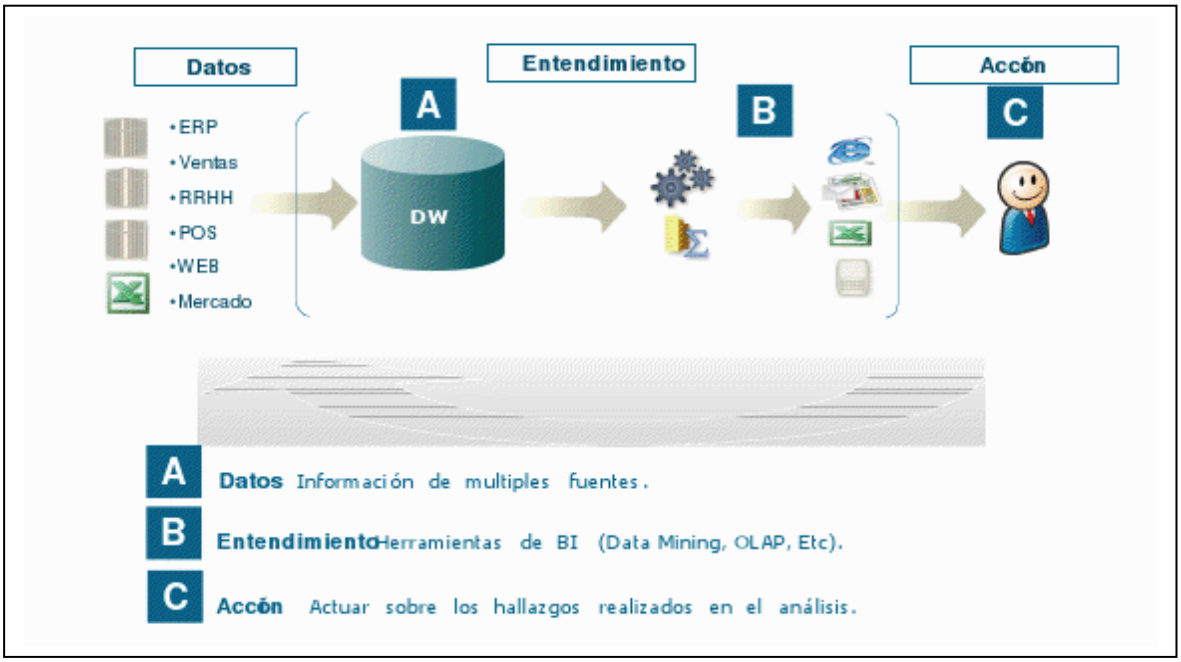

Figura 1. Proceso de segmentación estadística

A. Datos:

- Relevamiento y agrupamiento de la base de datos

B. Entendimiento

- Data mining y selección de variables de interés
- Armado de vista minable y reducción de la dimensionalidad
- Selección del software de Segmentación Estadística
- Selección del método de Segmentación Estadística.
- Trabajo con el Software de Segmentación Estadística.

C. Acción

- Interpretación de los conglomerados

5.1 Relevamiento y agrupamiento de la base de datos.

Es fundamental el relevamiento y agrupamiento de la base de datos dado que debe ser compactada en una única tabla así poder trabajar con mayor flexibilidad y simplicidad.
Se depura la base de datos a través de correctas consultas en MS Access de manera de poder limpiar la misma y allí seleccionar los registros que verdaderamente puedan agregarle valor al proyecto, eliminando registros duplicados, nulos o inválidos.

1 Base de datos única en MS Access

Se contaba con dos bases en formato .doc las cuales fueron importadas al Access conformando así un sistema único como base de datos mediante una consulta de creación de tabla primero y luego de datos anexados para incorporar ambas bases
Se ajustaron parámetros como los nombres de los campos y el tipo de datos (texto, número, fecha/hora), las tabulaciones, los puntos y comas para variables como ingresos, evitando así errores de importación.

2 Eliminación de registros duplicados, nulos o inválidos

La base en principio contaba con 72920 registros

Al correr la consulta para eliminar duplicados así la información al momento de la segmentación fuese lo mas certera posible se redujo la base de datos a 56159 clientes.

Mas allá de los fines estadísticos esta reducción tiene sentido dado que la base en bruto se extrajo de un reporte que no depura duplicados sino extrae por operaciones de los clientes. Es normal que un cliente realice más de una operación así como también reciba productos como renovaciones o extensiones en sus prestaciones.

Por otro lado se excluyeron aquellos clientes que recibieron exclusivamente visas por la flexibilización en los requerimientos y solicitud de datos para los mismos.

3 Consulta para Base final de trabajo

Se vincularon ambas tablas la tabla inicial con el registro único de clientes para poder conformar una base de clientes sin registros duplicados con las variables actualizadas al momento de la última operación que hayan realizado.

Para ello se utilizo:
Un Joint entre ambas tablas
Se escogió que la consulta traiga el máximo en el campo fecha es decir la última operación vigente
Se

4 Características de la Base

a. Estructura demográfica

Se presenta la estructura de la base de datos discriminada según sexo y rangos de edad como descripción de la identidad de la base.

EDAD	VARONES	MUJERES	AMBOS
18-24	1833	559	2392
24-35	7628	3567	11195
35-50	10943	7338	18281
50-65	7866	7701	15567
65-80	3478	4609	8087
80 y mas	205	432	637
TOTAL	31953	24206	56159

Tabla 2. Población por sexo y grupos de edad

Los datos de la tabla coinciden con una mayor actividad crediticia de los varones durante los dos tercios de su vida laboral. Luego en los últimos tramos de expectativa de vida predominan las mujeres coincidiendo así con la estructura demográfica de nuestra población.

b. Estructura socioeconómica

Para identificar el comportamiento de la base se calculo el nivel de ingresos promedio según edad y sexo, luego se califico al promedio de ambos en una pirámide de referencia según la clase social donde quedara ubicada.

EDAD	PROMEDIO M	PROMEDIO F	PROMEDIO AMBOS	CLASE[2]
18-24	$ 2.200	$ 1.783	$ 2.102	BAJA

[2] Parámetro de referencia generalmente aceptado a la hora de la pirámide socioeconómica (www.iprofesional.com)

24-35	$ 2.518	$ 2.061	$ 2.372	BAJA
35-50	$ 2.841	$ 2.324	$ 2.634	BAJA
50-65	$ 2.739	$ 2.376	$ 2.560	BAJA
65-80	$ 2.428	$ 1.904	$ 2.129	BAJA
80 y mas	$ 5.226	$ 4.525	$ 4.751	MEDIA BAJA
TOTAL	$ 2.673	$ 2.249	$ 2.490	BAJA

Tabla 3. Nivel de ingresos promedio por sexo y grupos de edad

Luego para complementar el análisis del comportamiento crediticio se presenta una tabla similar pero con el compromiso en el mercado de los integrantes de la base. Por compromiso en el mercado se entiende por la sumatoria mensual de[3]:

Compromiso mercado $= \sum$ **ctas de pmos vtes + pagos min de tarjet act + int de cta ctes**

EDAD	COMPROMISO PROMEDIO M	COMPROMISO PROMEDIO F	COMPROMISO PROMEDIO AMBOS
18-24	$ 300	$ 258	$ 290
24-35	$ 330	$ 259	$ 307
35-50	$ 412	$ 299	$ 367
50-65	$ 399	$ 273	$ 336
65-80	$ 129	$ 102	$ 114
80 y mas	$ 55	$ 37	$ 43
TOTAL	$ 350	$ 242	$ 303

Tabla 4. Nivel de compromiso promedio por grupos de edad

Podemos ver como el varón presenta una mayor tolerancia al endeudamiento que la mujer lo cual puede explicarse posiblemente por su mayor participación en la actividad laboral y en los gastos como el principal sostén del hogar.

5.2 Data mining y selección de las variables de interés

La base inicialmente contaba con 122 variables las cuales fueron reducidas a variables que resulten significativas para la investigación

A continuación se presenta el detalle de algunas de ellas:

Nombre Variable	Tipo de dato	Tipo de Variable	Explicación	Unidad

[3] Dictamen de clasificación de la entidad financiera
http://d1010140.mydomainwebhost.com/evaluadora/download_calificaciones.php?f=34&PHPSESSID=8b8943550f965977f8b8511cc76f7b7c

Edad	Numero	Descripción	Edad	# años
Hijos	Numero	Comportamiento	Numero de Hijos	# hijos
Atraso_Max	Texto	Comportamiento	Atraso en Cumplimiento Máximo	# meses
AntiguedadLab	Numero	Comportamiento	Vigencia en actividad laboral	# meses
AntiguedadCli	Numero	Comportamiento	Vigencia como cliente	# meses
Prestamo/Solicitud	Numero	Descripción	Nro de Operación	#######
Fecha	Fecha/Hora	Descripción	Fecha Registro Operación	##-mes-##
MaxCuota	Numero	Comportamiento	Máxima Cuota	$
MaxLimitePR	Numero	Comportamiento	Máximo Otorgamiento	$
MaxLimiteTC	Numero	Comportamiento	Máxima Tarjeta	$
Ingresos	Numero	Comportamiento	Monto de Ingresos Mensual	$
BCRA	Numero	Comportamiento	Situación en BCRA	1 a 6
NivelRiesgoInforme	Texto	Comportamiento	Nivel de Riesgo Crediticio	Bajo/Medio/Alto/MuyAlto
NivelRiesgoCL	Texto	Comportamiento	Nivel de Riesgo entidad	Bajo/Medio/Alto/MuyAlto
RiesgoInformeRN	Texto	Comportamiento	Informe Riesgo según Riesgo Net	Bajo/Medio/Alto/MuyAlto
scFraude_Nivel	Texto	Comportamiento	Nivel de Fraude	Bajo/Medio/Alto/MuyAlto
TipoCliente	Texto	Descripción	Tipo de Cliente	Cliente Conocido/Nuevo
Sexo	Texto	Descripción	Sexo	M / F
DomTipoTel	Texto	Descripción	Tipo de Teléfono	PARTICULAR/FAMILIAR/MENSAJES
FamiliaProd	Texto	Descripción	Agrupación del producto	PRIMERA OP/ DEALERS/RENOV/MUTUAL
Actividad	Texto	Comportamiento	Rubro de	PRIVADA/PUBLICA/FZAS

			Actividad Laboral	EGURIDAD
Vivienda	Texto	Comportamiento	Tipo de Vivienda	PPIA/CLOSPADRES/COTROFAM/ALQ/OTRO
Indeseable	Texto	Comportamiento	Marca de Indeseable por Incumplimiento	SI / NO
EstadoCivil	Texto	Descripción	Estado Civil	SOLTERO/CASADO/VIUDO/SEPARADO
Alta_Laboral	Numero	Descripción	Fecha Alta Laboral	yyyymmdd
Alta_Cliente	Texto	Descripción	Fecha de Alta de Cliente en Entidad	yyyymmdd
Alta_Credito	Numero	Descripción	Fecha Alta Primer Crédito	yyyymmdd
Provincia	Numero	Descripción	Código de Provincia	1 a 23
ConyIngresos	Numero	Comportamiento	Monto de Ingresos Mensual Conyugue	$
compromiso_mercado	Numero	Comportamiento	Compromiso en Mercado	$

Tabla 5. Características de las variables de la base

Una variable importante de destacar es la de BCRA que explica la clasificación de deudores establecida por el Banco Central de la República Argentina

Situación	Clasificación	Características
1	En situación normal	- Situación financiera capaz de atender adecuadamente todos sus compromisos financieros - Situación financiera líquida, con bajo nivel y adecuada estructura de endeudamiento en relación con su capacidad de ganancia - Alta capacidad de pago de las deudas (capital e intereses) en las condiciones pactadas - Adecuado sistema de información que permite conocer en forma permanente su situación financiera
2	Con seguimiento Especial	- Capaz de atender adecuadamente todos sus compromisos financieros - Incurra en atrasos de hasta 90 días en los pagos de sus obligaciones - En negociación o con acuerdos de refinanciación.

3	Con Problemas	- Presenta una situación financiera ilíquida y un nivel de fondos insuficiente para atender la totalidad de su compromiso financieros - Solo puede hacerse cargo del pago de intereses - Incurra en atrasos de hasta 180 días - Mantenga convenios de pago resultantes de concordatos judiciales o extrajudiciales homologados - Pertenezca a un sector de la actividad económica o ramo de negocios cuya tendencia futura no sea firme - Haya sido demandado judicialmente por la entidad para el cobro de su acreencia
4	Con alto riesgo de insolvencia	- El flujo de fondos es insuficiente para cubrir ni siquiera el pago de intereses - Dificultades para cumplir acuerdos de refinanciación - Incurra en atrasos de hasta 360 días - Sistema de información inadecuado sobre su real situación financiera y económica - Haya solicitado el concurso preventivo, celebrado un acuerdo preventivo extrajudicial o requerido quiebra aun no declarada - Pertenezca a un sector de la actividad económica o ramo de negocios cuya tendencia futura no sea firme
5	Irrecuperable	- Situación financiera mala con suspensión de pagos, quiebra decretada o pedido de su propia quiebra - Obligación de vender a pérdida activos de importancia para la actividad desarrollada - Incurra en atrasos superiores a un año, cuente con refinanciación del capital y con financiación de perdidas de explotación - Sistema de información inadecuado sobre su real situación financiera y económica - Pertenezca a un sector de la actividad económica o ramo de negocios en extinción
6	Irrecuperable por disposición técnica.	i) Entidades liquidadas por el Banco Central. ii) Entes residuales de entidades financieras públicas privatizadas o en proceso de privatización o disolución iii) Entidades financieras cuya autorización para funcionar haya sido revocada

		por el Banco Central y se encuentren en estado de liquidación judicial o quiebra. iv) Fideicomisos en los que SEDESA sea beneficiario.

Tabla 6. Características de la situación en BCRA

5.3 Armado vista minable y Reducción Dimensionalidad

Ante una base con más de cien variables plantea la necesidad de reducir la dimensionalidad, es decir, el número de columnas que representan a mis variables.

Una manera de para alcanzar dicha reducción es transformar un conjunto Xi de variables de interés en Zi componente principal

Ejemplo:

$$\textbf{Poder adquisitivo}_{\textbf{Grupo familiar}} = \frac{\sum \text{Ingresos grupo familiar} - \text{Compromiso en mercado}}{\text{\# Integrantes grupo familiar}}$$

Donde:

Zi = Poder adquisitivo Grupo Familiar (Componente principal)

X1 = Ingresos
X2 = ConyIngresos
"Ingresos grupo familiar" = X1 + X2

X3 = Hijos
X4 = EstadoCivil
"# Integrantes grupo familiar" = X3 + 1 (titular) + 1 (si EstadoCivil = "CASADO" o "SEPARADO")

"Compromiso en mercado" = X5 = compromiso_mercado

5.4 Selección del software de Segmentación Estadística.[4]

Se realizó un análisis comparativo entre tres software disponibles en el mercado: SPSS, SAS y R, a través de una matriz de selección de software ponderada.

Variable	Aspecto		SPSS		SAS		R	
Funcionalidad		15	10	150	5	75	5	75
	Amigabilidad para Usuario		Excelente		Baja - Regular		Baja - Regular	
Soporte		15	7	105	7	105	3	45
	Soporte Tecnico/Foros Web		Bueno		Bueno		Bajo	
Capacidades		10	7	70	7	70	10	100
	Variedad Analisis Estadisticos		Buena		Buena		Excelente	
		5	6	30	7	35	10	50
	Sistema Operativo		Windows		Windows/Mac/Linux		Windows/Mac/Linux	
Input / Output		10	10	100	7	70	8	80
	Documentacion		Excelente		Buena		Buena- Excelente	
		10	3	30	7	70	7	70
	Manipulacion de datos		Baja		Buena		Buena	
		10	6	60	8	80	10	100
	Calidad de Graficos		Regular		Buena- Excelente		Excelente	
		10						
	Control de Procesos		Baja		Excelente		Excelente	
Acceso		15	8	120	7	105	9	135
	Disponibilidad Licencia		Alta		Media		Alta	
	Total	100		665		610		655

Tabla 7. Matriz de selección de software

La comparación se centró en cinco variables (Funcionalidad, Soporte, Capacidades, Input/Output, Acceso) y sus respectivos aspectos de evaluación, con diferentes puntuaciones. A cada aspecto se le asigna un factor de ponderación correlacionado con los objetivos del proyecto, por el cual se multiplica la calificación de la variable para determinar el software seleccionado de trabajo.

A continuación se detalla el análisis de cada uno de los aspectos considerados:

- Amigabilidad para el usuario:

SPSS presenta una clara ventaja respecto a sus competidores dado que es posible acceder a todas las opciones mediante un menú de funciones ubicado en la parte superior.

[4] Scientific Electronic Library Online Argentina (SCIELO)
http://www.scielo.org.ar/scielo.php?pid=S1667-782X2008000200007&script=sci_arttext

Mientras que SAS y R requieren un conocimiento previo para poder operar con ellos tanto a nivel sintaxis como comandos lo cual dificulta su rápida ejecución para usuarios poco familiarizados con lenguajes de programación.

a. Soporte

- Asistencia Técnica / Foros Web / Help: On site

Otra ventaja que presenta SPSS y SAS es el soporte el cual también permite a través del servicio al cliente obtener respaldo técnico frente a posibles problemas de ejecución. Es verdad que R presenta a su vez un respaldo por parte de la comunidad científica mediante el numeroso aporte de diversos documentos gratuitos pero no ofrece ninguna garantía real ni una empresa como responsable legal.

Por ultimo cabe destacar que existen foros en Internet donde se brinda asistencia a los respectivos problemas para los tres programas constituyendo la principal alternativa de ayuda gratuita.

b. Capacidades

- Variedad de Análisis Estadísticos

SPSS ofrece un gran espectro de procesos estadísticos que alcanzan a cubrir los aplicados en el campo de la ingeniería. Sin embargo, cuando se requieren un alto grado de especificación en los mismos como por ejemplo un ajuste de convergencia para un modelo no lineal, no cumple con la versatilidad que si presenta SAS. Por ultimo R por su desarrollo abierto a la comunidad científica presenta un modelo más sólido con algoritmos robustos tanto para ajustes en modelos lineales y no lineales.

- Sistemas Operativos

La ventaja aquí es de R el cual presenta la versatilidad de ser instalado con facilidad y de manera estable en las tres plataformas más populares Windows, Mac y Linux.

c. Input / Output

- Documentación

Los tres programas presentan documentación como manuales y libros aplicados. La ventaja de SPSS es que ofrece una documentación fácil de entender y aplicar, debido a que su aplicación fue originalmente destinada a ciencias sociales más allá de la comunidad científica.

- Manipulación de Datos

Los tres softwares permiten una flexible interacción con datos en formatos estándares como lo son .txt, ASCII, dat así como también el atractivo de trabajar con datos en formato .MExcel.

- Calidad de Gráficos

La ventaja de una representación grafica permite la representación de resultados y comprender mejor su entendimiento. Aquí es importante la posibilidad de importar gráficos e imágenes, la personalización de reportes/output/gráficos aspectos que resultan difíciles de modificar en SPSS contrario a sus dos competidores. Por ultimo R, presenta una amplia gama de formatos para exportar y una calidad superior en la textura de sus gráficos.

- Control de Procesos

Los proceso estadísticos utilizan una serie de algoritmos con diferentes variantes cuando el usuario desconoce los mismos por lo general los programas optan por aquellos predefinidos. En SPSS no resulta sencillo cambiar estas especificaciones y no presenta mucha versatilidad al respecto. R presenta mayor flexibilidad al ser un código abierto donde el usuario tiene la libertad de crear sus propias funciones.

SPSS y SAS ofrecen generalmente un mayor detalle en sus outputs para un procedimiento estadístico cualquiera. En cambio, R ofrece como salidas sólo aspectos básicos y el usuario es el responsable de solicitar mas detalles. Esto puede considerarse de manera ambigua dado que para usuarios con escasos conocimientos estadísticos contar con demasiadas salidas puede inducir en errores de análisis e interpretación.

d. Acceso

- Disponibilidad de Licencia

En un primer análisis de costos R presenta la gran ventaja de ser gratuito tanto el software como sus actualizaciones. Sin embargo pueden obtenerse fácilmente versiones anteriores tanto de SPSS y SAS en la web, cuyos precios para licencias corporativas legitimas se estiman en USD1500 y USD 7200 respectivamente.

Selección:

Se decide optar por el software SPSS no solo por la amigabilidad que presenta para usuarios inexpertos sino también por la gran variedad de documentación y asistencia que existe en la web.

Por ultimo, mas allá que concluyera con la mejor puntuación se cuenta con experiencia previa en su aplicación de funciones y en el uso de sus herramientas estadísticas razones que favorecen su selección.

5.5 Selección del método de Segmentación Estadística.

En el marco conceptual se presentaron los dos algoritmos de formación de conglomerados:

	TIPO DE SEGMENTACION	
PROCEDIMIENTO	SEGMENTACION K-MEDIAS 1) Determina k semillas aleatorias 2) A cada pto. se le asigna un cluster 3) Recalcula para cada cluster el baricentro 4) Repite iterativamente 2 y 3 hasta que no haya cambios	SEGMENTACION JERARQUICA 1) Calcula todas las distancias 2) Busca la distancia mas chica 3) Agrupa los dos puntos mas cercanos que pasan a representarse únicamente por su baricentro
VENTAJA	Permite trabajar con grandes masas de datos	Visualizo numero de segmentos Segmenta lo distintivo
DESVENTAJA	Obliga a indicar la cantidad de clusters requeridos	Requiere mucho computo

Tabla 8. Características de los tipos de segmentación

En una primera aproximación por la gran cantidad de datos que contiene la base de estudio se debiera optar por el tipo de segmentación "k – medias", sin embargo, la experiencia de computo y calculo con el software a continuación posiblemente aclare cual es el método mas optimo para realizar la segmentación.

5.6 Trabajo con el Software de Segmentación Estadística.

Se descargo la licencia gratuita del SPSS PASW STATISTICS 18

Luego se importa la base para trabajar en la cual el programa nos indica si queremos trabajar con una muestra representativa aleatoria el cual dependerá del algoritmo de segmentación a utilizar y de la capacidad del software de trabajo.

Tipo de Segmentación	# casos	% muestra	Motivo
Jerárquica	1668 / 2230	3% / 4%	Menor tiempo de procesamiento

			Permite descubrir "k" clusters
K - medias	11194	20%	Trabajar con grandes masas de datos

Tabla 9. Metodología en los tipos de segmentación

Metodología

Se procederá a trabajar con ambos tipos de segmentación para lograr resultados del estudio de un mayor valor y precisión.

En primer lugar se procede a la "Estandarización de la matriz de datos", consiste en que normalmente las variables se encuentran medidas en diferentes escalas e unidades. El proceso de estandarización convierte a las variables medidas en bruto para la base en unidades adimensionales y, por ende, comparables.

A continuación, se procede al primer tipo de segmentación jerárquico y la interpretación del grafico dendograma como resultado en donde se determina el número de clusters representativo de la base.

Luego, se vuelve a repetir el proceso de segmentación jerárquica con una nueva muestra de la base de datos para confirmar que los resultados de la segmentación correspondientes sean similares.

Finalmente se recurre al método no jerárquico para trabajar con una mayor cantidad de casos conociendo de antemano el conjunto de clusters logrando así combinar ambos métodos de segmentación para una mejor interpretación de la base.

Estandarización de la matriz de datos

Algunos ejemplos dentro del proceso de estandarización fueron:

TipoCliente	Total	Conversión
Cliente Conocido	27380	0
Cliente Nuevo	28779	1

Tabla 10.1 Estandarización de la variable "TipoCliente"

Vivienda	Total	Conversión
ALQUILADA	1023	0
CON LOS PADRES	4012	0,25
CON OTRO FLIAR	3507	0,5

OTRO	1012	0,75
PROPIA	46605	1

Tabla 10.2 Estandarización de la variable "Vivienda"

NivelRiesgo	Total	Conversión
Alto	4631	0,8
Bajo	10861	0
Medio	13656	0,4
MedioAlto	8636	0,6
MedioBajo	13153	0,2
MuyAlto	5222	1

Tabla 10.3 Estandarización de la variable "NivelRiesgo"

El objetivo consistió en traducir variables de información tanto descriptiva como de comportamiento en variables binarias medibles y escalables para el proceso de segmentación

Segmentación Jerárquica

- ### Primera Muestra

1) Análisis Factorial

El objetivo del Análisis de Componentes Principales es el de identificar a partir de un conjunto de n variables observables, un nuevo conjunto k tal que sea menor a n y este constituido por variables no observables, denominadas factores tal que:

- Se pierda la menor cantidad de información posibles
- La solución obtenida tenga interpretación valida
- "n" sea un numero pequeño

Los pasos a seguir serán los siguientes:

1. Evaluación de la correspondencia de realizar el análisis: numero suficiente de variables, precisión buscada
2. Extracción de los factores
3. Calculo de las puntuaciones factoriales según cada caso

Será importante que para la formación de componentes principales se disponga de la mayor cantidad de variables de comportamiento posibles así la futura segmentación refleje no solo la identidad de los clusters sino también patrones de comportamiento y hábitos de sus integrantes.

Las variables seleccionadas para el análisis fueron:

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
NivelRiesgoInforme	,4051	,28038	1668
scoreCL	844,59	123,092	1668
NivelRiesgoCL	,391	,2988	1668
MaxCuota	757,89	591,862	1668
MaxLimitePR	11059,11	9145,655	1668
ScoreVrz	463,19	363,343	1668
Ingresos	2266,21	1623,111	1668
BCRA	1,16	,655	1668
AntiguedadLab	127,48	116,541	1668
RCI	38,09	15,515	1668
RiesgoInformeRN	,1238	,28133	1668
compromiso_mercado	292,39	491,329	1668
scFraude_Nivel	,540	,3441	1668
Poder	1429,08	1305,118	1668

Tabla 11. Estadísticos descriptivos del proceso de factores

Estadísticos descriptivos

El programa a través de la matriz de correlaciones crea las nuevas variables que serán representativas según dos indicadores fundamentales:

- Las componentes deben tener todas varianzas (autovalores) mayores que 1
- Deben representar en conjunto entre ellos el 70% o mas de la varianza de las variables originales

Como podemos observar a continuación el método de extracción del SPSS extrajo cinco componentes que representan un porcentaje acumulado de varianza del 73,855% y sus respectivos autovalores son mayores que 1.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,840	27,430	27,430	3,840	27,430	27,430
2	2,388	17,054	44,484	2,388	17,054	44,484
3	1,671	11,933	56,417	1,671	11,933	56,417
4	1,367	9,766	66,182	1,367	9,766	66,182
5	1,074	7,672	73,855	1,074	7,672	73,855
6	,835	5,964	79,819			
7	,733	5,235	85,054			
8	,591	4,220	89,274			
9	,454	3,246	92,520			
10	,297	2,124	94,644			
11	,288	2,055	96,699			
12	,215	1,537	98,237			
13	,149	1,065	99,302			
14	,098	,698	100,000			

Tabla 12. Método de extracción: Análisis de Componentes principales.

Estas 5 componentes principales son guardadas como variables en la base de datos

2) Segmentación Jerárquica

Como hemos dicho el sentido de una buena segmentación es el de encontrar y establecer segmentos cuyo cuyos elementos internos sean lo mas homogéneo posible y lo mas heterogéneo entre dichos segmentos.

Se procedió a utilizar la segmentación del tipo jerárquico dada la imposibilidad de establecer correctamente de antemano el número de clusters por el tamaño y complejidad propia de la base.

Entonces para determinar el número de conglomerados y sus centroides procedemos a la segmentación.

- Primera Corrida

1) El primer paso es el de incorporar las variables que queremos que formen parte de la segmentación:

Fueron seleccionadas las variables que representaran comportamiento y aquellas que hayan sido sometidas al proceso de factores por lo tanto se seleccionaron las siguientes:

- Las 5 componentes principales

2) Se escoge como Método de conglomeración la "Vinculacion inter-grupos" el cual realiza comparaciones de un grupo con otro grupo, y de ésta forma se van obteniendo los conglomerados respectivos de forma mas o menos esférica.

3) Como medida se selecciona la "Distancia Euclídea al cuadrado[5]"

4) Se solicita que en resultados la información estadística sea el "Historial de conglomeración" y la "Matriz de distancias" con un "rango de soluciones entre 2 y 10".
Aquí se indica el rango de posibles clusters que detectará la segmentación.

5) En gráficos se solicita únicamente el dendograma, la representación grafica que mejor ayuda a comprender el resultado de un análisis cluster

6) En la opción guardar pedimos al software que nos genere 9 nuevas variables conteniendo en ellas el conglomerado que se le asigna a los casos/sujetos.
7) Resultados:

Primer dendograma

[5] Distancia Euclidea al cuadrado: La suma de los cuadrados de las diferencias entre los valores de los elementos.

$$d_E(P,Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

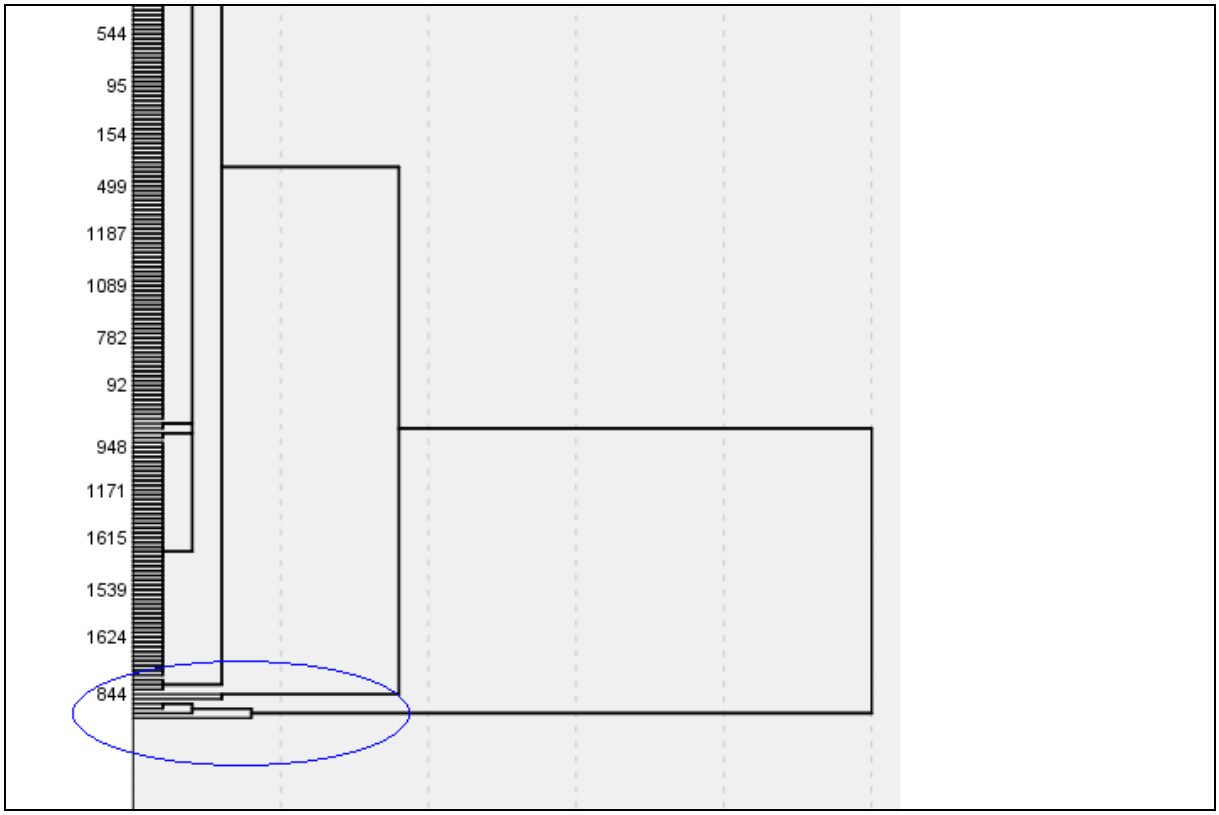

Figura 2. Método de extracción: Primer Dendograma

Como se observa en el gráfico existen una serie de casos los cuales contradicen con el comportamiento común de la base ubicándose en una ramificación externa a la principal. Entonces para determinar el número de conglomerados y sus centroides procedemos a la segmentación.

1) Detección de casos excepcionales

Para detectar los mismos en el momento en que el software pide un rango de soluciones se solicita un rango de dos clusters. Es decir, quedarán agrupados la mayoría en el primer grupo y los casos atípicos en el segundo grupo.

A continuación se presenta la depuración de los casos y algunas observaciones:

Total de casos	Excluidos	Casos Remanentes
1668	11	1657

Tabla 13: Casos excluidos

- Tres casos presentan un otorgamiento máximo de $100000 sin contar con los requisitos para semejante monto considerando además que el monto promedio de otorgamiento de la base es de $11595.5, se llega a la conclusión que se tratan de clientes VIP o excepciones especiales aprobadas sin los requisitos necesarios dado que su nivel de ingreso medio es solamente de $2683.
- Otros casos presentan una relación monto máximo / ingreso relativamente baja es decir que por algún motivo el monto otorgado es menor o casi igual a su nivel de ingresos mensual tratándose de casos donde por el elevado nivel de ingresos (promedio medio de $15364) comparado con la media de la base ($2489) o sospechas de riesgo crediticio se les otorga un monto menor (promedio $16250) al máximo acorde a su nivel salarial.

Segundo Dendograma (sin casos excepcionales)

- Determinación del número de clusters

Es subjetiva la decisión sobre el número de clusters dado que esta sujeta en gran medida a la interpretación y criterio del encargado de la segmentación. Especialmente, cuando se considera el factor tamaño en el número de casos, ya que si se seleccionan demasiado pocos, los clusters que resultaran muy posiblemente serán heterogéneos y artificiales, por otro lado si se seleccionan demasiados, la interpretación de los mismos suele ser complicada.

En general, el procedimiento a seguir consiste en cortar el dendrograma mediante una línea horizontal como en el gráfico siguiente, determinamos el número de clusters en que dividimos el conjunto de objetos. Aquí entra en criterio personal a que altura de la ramificación realizar el corte para poder visualizar los grupos, teniendo en cuenta la distancia entre los mismos y la altura de la ramificación.

A continuación veremos que se presenta una distribución de ramificaciones mas regular que permite interpretar que el número de clusters presentes en la muestra de la base se encuentra en un rango entre 5 y 7, aquí entra el criterio de considerar o no como representativos individualmente o en conjunto los últimos 3 grupos de clusters por su proximidad y tamaño.

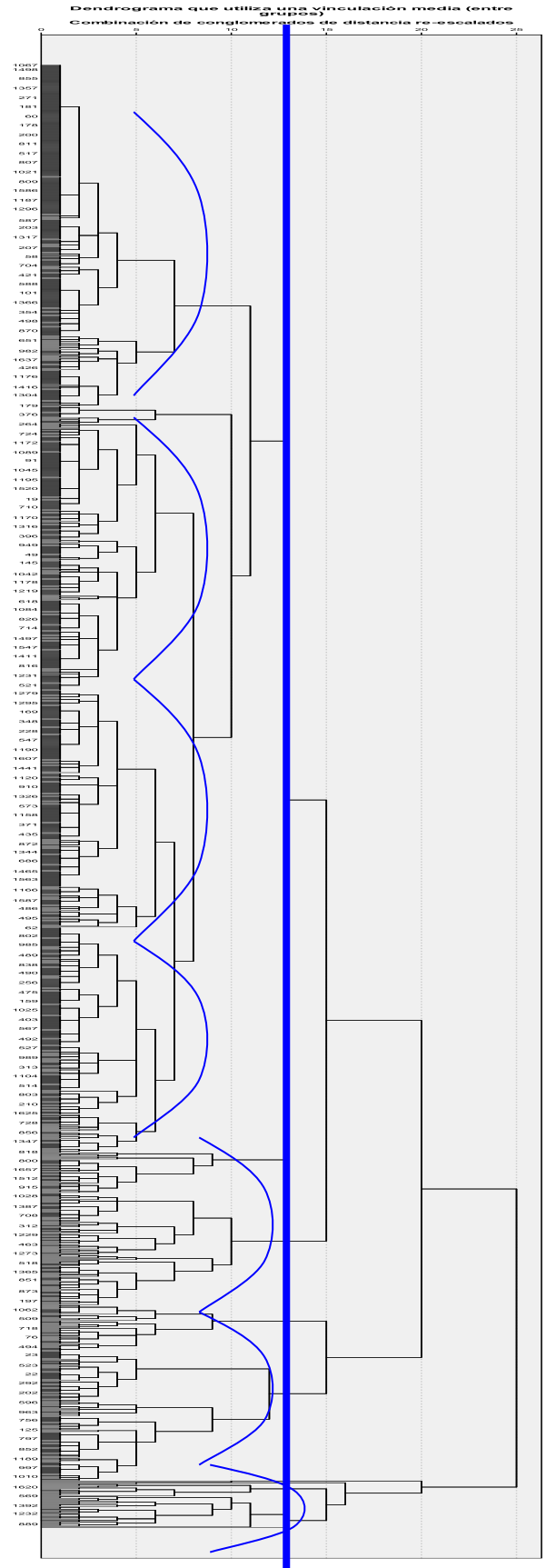

Figura 3. Dendograma de Primera Muestra

- Segunda Muestra

1) Análisis Factorial

Las variables seleccionadas para el análisis fueron las mismas que para la primera muestra:

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
NivelRiesgoInforme	,4030	,28296	2230
scoreCL	845,60	121,104	2230
NivelRiesgoCL	,389	,2998	2230
MaxCuota	754,81	590,531	2230
MaxLimitePR	11325,69	9161,557	2230
ScoreVrz	458,45	365,174	2230
Ingresos	2384,58	5010,035	2230
BCRA	1,16	,658	2230
AntiguedadLab	127,52	115,609	2230
RCI	38,20	15,099	2230
RiesgoInformeRN	,1201	,27963	2230
compromiso_mercado	304,19	500,184	2230
Poder	1487,24	3139,641	2230
scFraude_Nivel	,541	,3464	2230

Tabla 14. Estadísticos descriptivos del proceso de factores

Como podemos observar a continuación el método de extracción del SPSS extrajo cinco componentes que representan un porcentaje acumulado de varianza del 73,801% y sus respectivos autovalores son mayores que 1.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,845	27,463	27,463	3,845	27,463	27,463
2	2,201	15,724	43,187	2,201	15,724	43,187
3	1,769	12,636	55,823	1,769	12,636	55,823
4	1,443	10,308	66,131	1,443	10,308	66,131
5	1,074	7,670	73,801	1,074	7,670	73,801
6	,849	6,064	79,865			

7	,756	5,398	85,263			
8	,568	4,059	89,321			
9	,506	3,615	92,936			
10	,321	2,291	95,227			
11	,291	2,078	97,305			
12	,212	1,514	98,818			
13	,095	,677	99,495			
14	,071	,505	100,000			

Tabla 15. Método de extracción: Análisis de Componentes principales.

Estas 5 componentes principales son guardadas como variables en la base de datos

2) Segmentación Jerárquica

Entonces para determinar el número de conglomerados y sus centroides procedemos a la segmentación.

Primera Corrida

Nuevamente se realiza la exclusión de casos excepcionales solicitando al programa que guarde las posibles soluciones en un rango de dos.

Es decir, quedarán agrupados la mayoría en el primer grupo y los casos atípicos en el segundo grupo.

A continuación se presenta la depuración de los casos y algunas observaciones:

Total de casos	**Excluidos**	**Casos Remanentes**
2230	7	2223

Tabla 16: Casos excluidos – Primera corrida

- Nuevamente, aparecen casos (4) con un nivel de monto excepcional de $100000 correspondiente a clientes VIP o excepciones extraordinarias.
- Aparece un caso también que posiblemente corresponda a un error en la carga de datos dado que figura con un ingreso de $ 208892.

Segunda Corrida

Se realiza nuevamente anomalías en el proceso de segmentación a través del dendograma y se decide excluir casos nuevamente debido al elevado número de casos atípicos detectados (215) al repetir el procedimiento de exclusión.

Total de casos	Excluidos	Casos Remanentes
2223	215	2008

Tabla 17. Casos excluidos – Segunda corrida

- Se observa nuevamente que el monto de otorgamiento máximo de estos casos ($28880) es considerablemente elevado respecto de la media de la base$11595.5)

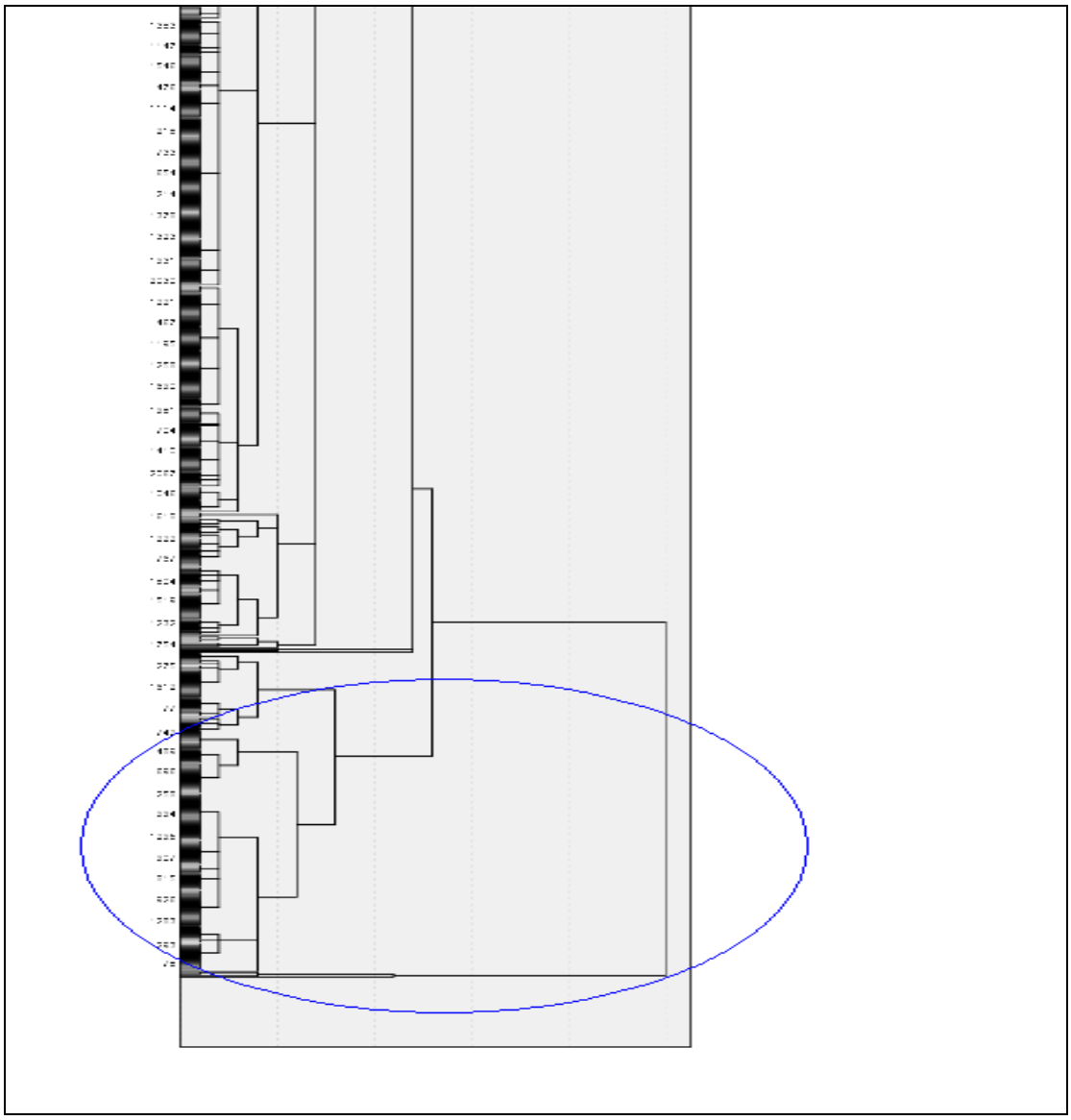

Figura 4. Dendograma de Segunda Muestra – Segunda Corrida

Tercera Corrida

- Determinación del número de clusters

Nuevamente hay que considerar la importancia de la subjetividad en el criterio e interpretación del método para seleccionar los clusters.

En esta segunda muestra se presenta un mayor espectro en la visualización de los subgrupos lo cual pudiera ser producto de una mejor depuración resultando en una mayor riqueza en el proceso de segmentación.

En el dendograma se observa la aparente presencia de 4 a 5 grupos de clusters representados con curvas convexas si se considera o no el grupo de menor tamaño presente en la parte inferior del dendograma.

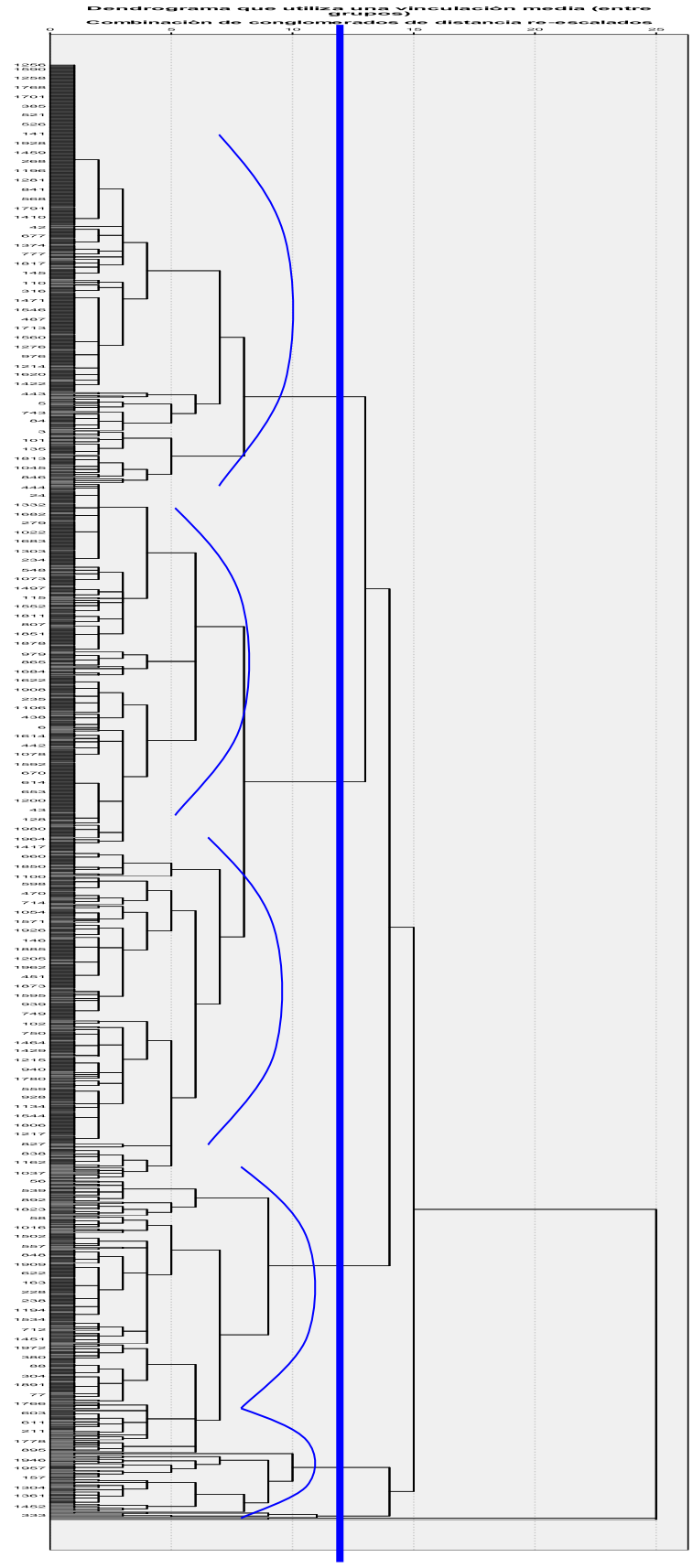

Figura 5. Dendograma de Segunda Muestra – Tercera Corrida

Segmentación K - medias

Gracias al tipo de segmentación previa se puede interpretar que el número de clusters representativos presentes en la base es aproximadamente 5, lo cual permite realizar el nuevo tipo de segmentación
La ventaja principal reside en que al saber con anticipación el número de clusters representativo, podremos trabajar con mayores masas de datos, muestras estadísticamente representativas de la base original.

Para que valorar el número de variables a tener en cuenta para esta segmentación se vuelve a someter la nueva muestra al análisis de componentes principales.

- Muestra

1) Análisis Factorial

Las variables seleccionadas para el análisis fueron las mismas que para el procedimiento de segmentación jerárquica:

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
NivelRiesgoInforme	,4052	,29180	11153
scoreCL	843,40	125,204	11153
NivelRiesgoCL	,397	,3055	11153
MaxCuota	789,65	1919,168	11153
MaxLimitePR	11515,25	9283,025	11153
ScoreVrz	459,86	365,676	11153
Ingresos	2390,56	4556,530	11153
BCRA	1,18	,703	11153
AntiguedadLab	126,47	117,284	11153
RCI	39,27	95,393	11153
RiesgoInformeRN	,1284	,29405	11153
Poder	1479,63	3253,930	11153
scFraude_Nivel	,535	,3473	11153

Tabla 18. Estadísticos descriptivos del proceso de factores

Como podemos observar a continuación el método de extracción del SPSS extrajo cinco componentes que representan un porcentaje acumulado de varianza del 79,013% y sus respectivos autovalores son mayores que 1.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,935	30,272	30,272	3,935	30,272	30,272
2	2,050	15,766	46,039	2,050	15,766	46,039
3	1,858	14,289	60,328	1,858	14,289	60,328
4	1,413	10,869	71,197	1,413	10,869	71,197
5	1,016	7,817	79,013	1,016	7,817	79,013
6	,772	5,936	84,950			
7	,589	4,527	89,477			
8	,481	3,704	93,181			
9	,366	2,814	95,995			
10	,219	1,686	97,681			
11	,175	1,350	99,031			
12	,096	,737	99,768			
13	,030	,232	100,000			

Tabla 19. Método de extracción: Análisis de Componentes principales.

Estas 5 componentes principales son guardadas como variables en la base de datos

2) Segmentación Jerárquica

Se procede a una segmentación sin representación gráfica de este tipo para poder excluir casos excepcionales evitando arrastrar errores o casos anormales.

Primera Corrida

Nuevamente se realiza la exclusión de casos excepcionales solicitando al programa que guarde las posibles soluciones en un rango de dos.

Es decir, quedarán agrupados la mayoría en el primer grupo y los casos atípicos en el segundo grupo.

A continuación se presenta la depuración de los casos y algunas observaciones:

Total de casos	Excluidos	Casos Remanentes
11153	10	11143

Tabla 20. Casos excluidos – Primera corrida

- Un caso particularmente presentaba características de cliente VIP o algún producto especial como para un empleado al poseer como monto máximo $66.000 y no presentar ingresos coherentes con el mismo.

- Los otros 9 casos poseían un nivel de ingresos demasiado elevado en promedio arriba de $20.000 en relación a los montos máximos otorgados por solamente $15.000, lo cual indicaría un error de algún tipo.

3) Segmentación k – medias

Se seleccionan como variables las cinco componentes principales

En Nº de conglomerados: 5

Clusters de Segmentación k – medias	
Cuenta de Clusters	Total
1	1025
2	4093
3	3417
4	20
5	2588
Total general	11143

Tabla 21. Distribución de casos según conjunto de cluster.

	Conglomerado				
	1	2	3	4	5
REGR factor score 1 for analysis 1	3,99676	,20933	1,93775	2,38192	-,76342
REGR factor score 2 for analysis 1	-,33884	-,32432	3,57615	7,25875	,43336
REGR factor score 3 for analysis 1	,12221	-,12578	4,78121	-3,07687	,01717
REGR factor score 4 for analysis 1	-1,93871	,71745	1,24089	2,06406	2,87835
REGR factor score 5 for analysis 1	2,31932	-1,97852	1,05408	-,75220	3,87476

Tabla 22. Centros iniciales de los conglomerados

5.7 Interpretación de los conglomerados

En primer lugar, se busco posicionar a cada uno de los conglomerados de manera cualitativa según dos ejes o características principales:

- Capacidad de pago: esta característica es mayor para clientes que presentan una adecuada estructura de endeudamiento respecto de sus ingresos, un nulo atraso en sus compromisos de mercado, una situación ideal en el BCRA, un poder adquisitivo elevado.

- Fidelidad: vendría a ser el poder de retención que tienen los productos sobre el cliente. Es decir, la tasa de renovación, el vínculo o antigüedad con la entidad, la cantidad de productos que solicita.

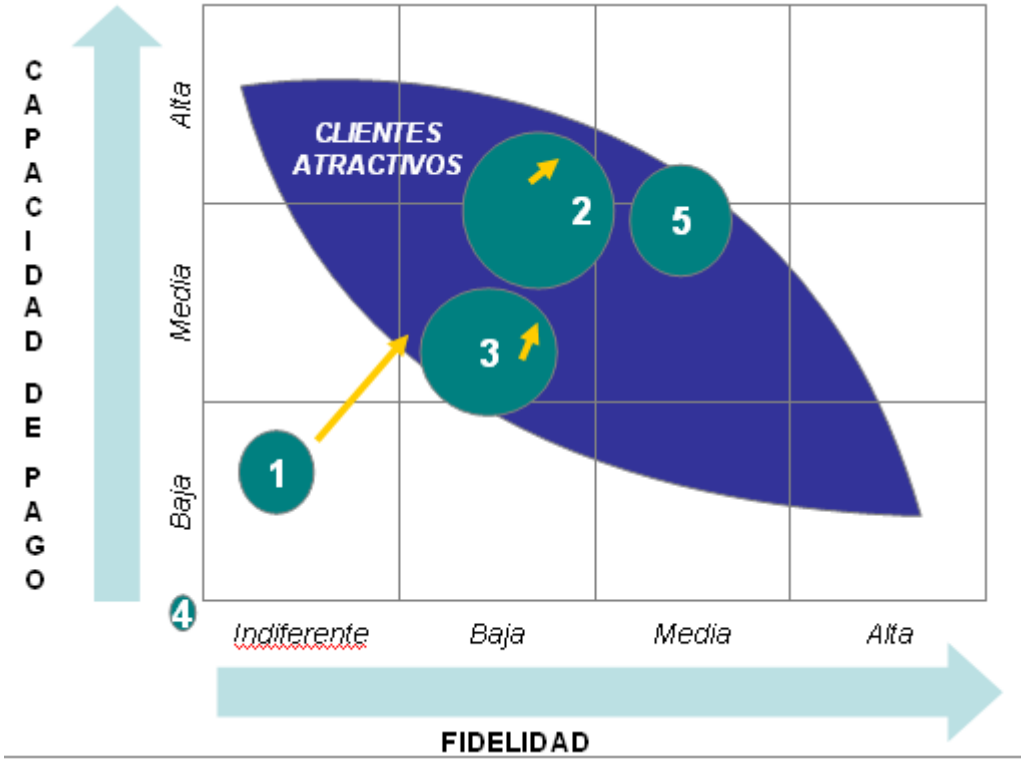

Figura 6. Disposición de los clusters

Aquellos clientes con una alta capacidad de pago y una media-alta fidelidad son clientes considerados muy atractivos, dado que reúnen las condiciones para poder brindarles y renovarles servicios o productos, sin exponer la entidad a un riesgo elevado.

Sin embargo, es necesario comprender como la dualidad entre el atraso de pago y la alta voluntad de pago puede jugar a favor de una entidad financiera y de que forma uno debe ajustarse al tipo de cliente para el beneficio mutuo, esto explica el amplio espectro de "clientes atractivos".

Es decir, por ejemplo, un cliente con alta efectividad de pago deberá ser continuamente incluido en campañas para activar una alta rotación en sus productos adquiridos. Por otro

lado, un cliente con dificultades financieras en sus obligaciones debe ser tratado en forma diferencial, el mismo posiblemente recurra a una refinanciación de su plan de deuda o una flexibilización en sus tasas de interés antes de perjudicar su situación patrimonial.

Se describen a continuación algunas características de los conjuntos surgidos a partir del proceso de segmentación estadística, así como también algunas acciones a implementar para poder fortalecer su fidelidad con la institución y ajustar los productos a su capacidad de pago, reduciendo así el nivel de riesgo de incobrabilidad o evitar llegar a una instancia judicial.

5.7.1 Cluster 1

Una particularidad de este conjunto que representa casi el 10 % de la muestra es su situación promedio en el Banco Central de la República Argentina (BCRA) de 2,76, la mas alta de todos los conjuntos.

La misma situación refleja el primer comportamiento del grupo la dificultad de afrontar sus compromisos de pago ya sea por:

- Una inadecuada estructura de ingresos – compromisos
- Un elevado nivel de atrasos cercano o superior a los 180 días
- Posible pertenencia a un sector empleador inestable
- Posibilidad de verse involucrados en obligaciones, denuncias o instancias judiciales.

Otra característica que también explica la alta situación media en BCRA de este cluster en particular es que únicamente presenta clientes bajo la agrupación Refinanciación y Mutuales.

FamiliaProd	% sobre total
REFINANCIACION	95,8%
MUTUALES	4,2%

Tabla 23. Distribución por Agrupación - Cluster 1

Es necesario comprender la naturaleza del cliente en este conglomerado el cual por diversos motivos no puede responder a su compromiso de pago, es decir, presenta una inestabilidad económica la cual se refleja en recurrir a "salva vidas" como pueden ser los planes de Refinanciación de deuda o vinculaciones con sus recibos de haberes a través de convenios con Mutuales.

En conclusión, son clientes propensos a la falta de voluntad de pago por lo cual pueden ser atraídos por productos o contratos flexibles que no les hagan sentir hostigamiento a la hora de concretar sus obligaciones como lo pueden ser:

- Ofertas de campañas de Refinanciación: en donde se "rearma" su compromiso total en la entidad con una nueva deuda la cual puede presentar un alto plazo de pago bajo una tasa de interés flexible que le brinde una cierta comodidad de pago al cliente.

- Productos vinculados con sus recibos de haberes bajo la modalidad de una Mutual en donde al directamente debitarle la cuota como un monto de su acreditación mensual ya sea por jubilación o pensión. Es importante mencionar que a través de dichos productos, el cliente no debe cumplir estricta y físicamente con el acto de "ir a pagar a la entidad".

5.7.2 Cluster 2

El primer punto a destacar en este conglomerado es que en contraposición con el primer cluster, presentan una situación normal bajo la clasificación como deudores en el BCRA dado que presentan una media exacta de 1.

Esta situación esta caracterizada por deudores con una sólida situación financiera frente a sus compromisos de pago sostenido por:

- Buena liquidez y estructura estable de endeudamiento en relación a sus ingresos.

- Alta capacidad y voluntad de pago en las condiciones establecidas: tiempo, intereses.

- Acceso a información confiable y certificada respecto de su situación económica financiera.

Respecto a los productos adquiridos sus agrupaciones también representan en cierta medida el comportamiento de este conjunto:

FamiliaProd	% Acumulado
DEALERS VIP	39,2%
RENOVACION	71,2%
PRIMERA OPERACION	95,8%
REFINANCIACION	98,3%
CAMPAÑA NVOS CTES	99,1%

MUTUALES	99,5%
GRANDES DEALERS	99,7%
DEALERS	99,9%
RENV TARJETAS	100,0%

Tabla 24. Distribución por Agrupación - Cluster 2

Como se puede observar entre las agrupaciones “Dealers Vip”, “Renovación” y “Primera Operación” representan casi el total del cluster: 96%

Estas tres agrupaciones representa clientes que no tienen problemas en cumplir sus obligaciones con el mercado esto se observa en el comportamiento de la entidad financiera hacia los mismos:

- Darles productos especiales a través de comercios a través de “Dealers VIP” es decir satisfacer sus necesidades de consumo.
- Renovarles sus obligaciones con la entidad demostrando su fidelidad y buen comportamiento de pago.
- Confiar es su comportamiento de pago en el mercado para otorgarles como nuevos clientes un producto como primera operación.

Se puede destacar que este conjunto de clientes son ideales para ofrecer nuevos productos dado que presentan una elevada predisposición a cumplir con sus obligaciones y una alta fidelidad a continuar operando con la entidad a futuro. Esta retención puede ser incluso mayor fortaleciendo el lazo con los mismos incluyéndolos continuamente en la cartera activa de clientes:

- Campañas de tratamiento preferencial bajo la denominación de Clientes VIP con políticas atractivas ya sea para productos a través de comercios, tarjetas con beneficios y limites extraordinarios, montos excepcionales para sus propios negocios o emprendimientos y la posibilidad de brindarles préstamos hipotecarios.
- Campañas de Renovación aprovechando su buen comportamiento de pago y renovando su fidelidad con la entidad

5.7.3 Cluster 3

En relación a su clasificación como deudor por el BCRA presenta un comportamiento normal la media de este conjunto es muy cercana a 1.

Esto indica una situación financiera capaz de afrontar sus compromisos de pago en el mercado en tiempo y forma con una estructura de ingresos certificada, sólida y confiable.

Respecto a las agrupaciones presentes en este conglomerado:

CAMPAÑA NVOS CTES	% sobre total
MUTUALES	71,6%
DEALERS VIP	10,7%
REFINANCIACION	8,3%
RENOVACION	8,2%
PRIMERA OPERACION	0,8%
CAMPAÑA NVOS CTES	0,3%
DEALERS	0,1%
RENV TARJETAS	0,0%

Tabla 25. Distribución por Agrupación - Cluster3

La agrupación con más presencia son Mutuales con más del 70% del total, las cuales apuntan en general a un público jubilado o pensionado representado en el promedio de edad de esta agrupación 60 años. Es necesario comprender la naturaleza de las personas mayores[6] que tienen una alta tendencia al ahorro y no suelen recurrir a la solicitud de prestamos salvo por circunstancias excepcionales en donde su grado de endeudamiento suele ser menor. Estas personas suelen hacen foco en la confianza y seguridad a la hora de tratar con entidades financieras.

- Un enfoque posible es el ofrecimiento de productos a tasas muy accesibles en donde la persona mayor no se vea afectada buscando tal vez posibles vinculaciones con el estado para aumentar su seguridad.

- Otro aspecto a considerar es la necesidad de cajas de ahorro para el cobro de pensiones o haberes, tarjetas de debito o crédito a través de las mismas y la oferta de cajas de seguridad para satisfacer esta necesidad de ahorro y sensación de seguridad.

[6] Políticas Sociales para Las Personas Mayores en el Próximo Siglo: Actas Del Congreso, Murcia, 10-12 de Noviembre de 1999

5.7.4 Cluster 4

Este conglomerado pertenece a casos excepcionales por varios motivos:

- Son solo 20 casos representando en tamaño una minoría de un tamaño muy inferior al resto de los clusters.
- Presentan un monto promedio máximo de otorgamiento de casi $96000, no solo extraordinario respecto del resto de la muestra sino también para la oferta común de la entidad financiera.
- Su nivel de ingresos es de aproximadamente $3300, presentando una estructura futura de ingresos-compromisos muy difícil de afrontar.

Las agrupaciones presentes son:

FamiliaProd	Total
MUTUALES	1
RENOVACION	19
Total general	20

Tabla 26. Distribución por Agrupación - Cluster4

Es decir todos los clientes resultan estar familiarizados con la entidad lo cual probablemente justifique estas condiciones preferenciales.

Estas excepciones suelen ser aprobadas para empleados de confianza dentro de la entidad, o clientes VIP relacionados con directores y accionistas que exceptúan un cálculo racional de otorgamiento.

5.7.5 Cluster 5

El último cluster presenta una situación media normal en BCRA (cercana a 1), lo cual implica que los miembros presentan una relación ingreso-compromiso en el mercado estable.

Como se puede ver en los productos obtenidos el porcentaje más alto es Renovación, es decir, clientes con buen comportamiento como deudores, que recurren nuevamente a una prestación. Luego, Dealers VIP, clientes que obtienen productos en comercios de confianza accediendo a ellos a través de "canales Premium".

FamiliaProd	% del Total
RENOVACION	73,9%

DEALERS VIP	14,2%
PRIMERA OPERACION	7,9%
MUTUALES	3,7%
REFINANCIACION	0,2%
CAMPAÑA NVOS CTES	0,1%
DEALERS	0,0%
GRANDES DEALERS	0,0%
RENV TARJETAS	0,0%

Tabla 27. Distribución por Agrupación - Cluster4

Respecto a las posibles características de este grupo además de su buen comportamiento en sus obligaciones financieras presenta una cierta fidelidad al escoger la gran mayoría la entidad para volver a adquirir una prestación.

El foco en la atención personalizada para dichos clientes es importante para aumentar aún el nivel de renovación de préstamos, por lo tanto, el seguimiento por parte de "atención al cliente" dentro de la entidad juega un papel fundamental, algunas sugerencias pueden ser:

- Productos bonificados: Al cliente volver a solicitar un préstamo en la entidad contar con beneficios especiales, como cancelación anticipada, tasas preferenciales, última cuota bonificada.

- Segmento de bienvenida y seguimiento: Asistir al cliente, es decir, recordarle telefónicamente o vía e-mail de sus vencimientos, cercanía con las sucursales, canales disponibles y formas de pago: física en sucursal, rapipagos, pago fácil o en forma online.

6. Conclusiones

Completado el análisis de detalle, es el momento de recopilar y resaltar las conclusiones que se infieren del presente trabajo.

En primer lugar hay que destacar la subjetividad como un factor fundamental en la segmentación. El criterio del investigador y la interpretación en cada una de las etapas de análisis es un riesgo a asumir, sin embargo, la profundidad del estudio y la dedicación de tiempo aplicado llevará a un procedimiento y resultados lo mas certeros posibles acorde con los objetivos propuestos.

El recurrir e interpretar a variables de comportamiento para realizar tanto el análisis de factores y de segmentación, resulta mucho mas valioso a la hora de la conformación de los conglomerados o clusters. El mismo consiste un desafío mucho mayor que agrupar a una población en forma demográfica como puede ser la edad o el sexo, implica:

- Escoger que factores representan idóneamente a un cliente financiero y ver la correlación de los mismos.

- Interpretar y trabajar criteriosamente con los conglomerados encontrados, buscando patrones de comportamiento que al fin y al cabo los identifiquen en forma homogénea.

- Encontrar dichos perfiles heterogéneos a simple vista y atacarlos con estrategias que puedan ser aplicadas en conjunto buscando una mayor efectividad y atracción en la oferta de productos.

- Comprender como la dualidad entre el atraso de pago y la alta voluntad de pago puede jugar a favor de una entidad financiera y de que forma uno debe ajustarse al tipo de cliente para el beneficio mutuo. Es decir, un cliente con alta efectividad de pago deberá ser continuamente incluido en campañas para activar una alta rotación en sus productos adquiridos. Por otro lado, un cliente con dificultades financieras en sus obligaciones debe ser tratado en forma diferencial posiblemente recurra a una refinanciación de su plan de deuda o una flexibilización en sus tasas de interés.

- Entender que un mercado esta conformado por personas mayores resistentes a la idea del endeudamiento y mas propensos al ahorro pero que continuamente esta involucrados en operaciones en entidades bancarias por el sistema de haberes y pensiones.

- Atender un mercado joven y emprendedor que ven con buenos ojos el endeudamiento en pos de un apalancamiento positivo en sus proyectos, entendiendo el beneficio que puede generarse en constituirse como acreedores de los mismos pero sin recurrir al hostigamiento y la puja por cobrar.

- Aprovechar el furor por el consumo que existe en nuestra sociedad mediante vinculaciones con locales, comercios, supermercados e incluso agencias de viajes o aerolíneas.

Finalmente, es necesario tener conciencia de las múltiples variables que entran en juego a la hora de modelizar a un cliente y que el riesgo de una potencial incobrabilidad siempre esta presente. Aquí es donde la investigación apunta a intentar comprender en forma estadística la naturaleza de una cartera de clientes y consumidores que en forma complementaria con un

análisis exhaustivo de riesgos puede acotar esta brecha de distancia con la percepción de los clientes.

7. Planteo de futuras líneas de investigación

- Actualización y seguimiento:

La segmentación debe ser una práctica continua para captar información de los clientes transformándola de forma inteligente y al final usando la misma para mejorar la ventaja competitiva.

Por ende, es fundamental ver la evolución de los miembros de cada uno de los clusters y observar cambios de integrantes de un cluster a otro, variaciones en la magnitud de cada uno de los clusters y sus componentes para anticiparse a variaciones del mercado y necesidades de los clientes.

- Recurrir a otros métodos:

Redes neuronales para analizar datos financieros[7]: el SPSS, software con el que se trabajó en la investigación, cuenta con este procedimiento como herramienta.

Se utiliza para el reconocimiento de patrones a partir de cierta información. Puede aplicarse con fines de encontrar comportamientos dentro de clientes de entidades bancarias a partir de información contable.

Existen antecedentes de detección de fraudes con este tipo de procedimientos:

- En 1995 el Instituto de Ingeniería del Conocimiento de la Universidad Austral[8] realizo un proyecto de detección de fraude en tarjetas de crédito. Se realizó en forma conjunta con IBM España para la Sociedad Española de Medios de Pago (SEMP) con excelentes resultados y vigencia en la actualidad.

- Predicciones de quiebra bancarias[9] por la Revista Española de Financiación y Contabilidad

- Evaluación asimétrica de una red neuronal artificial: Aplicación al caso de la inflación en Colombia[10]

[7] http://ciberconta.unizar.es/leccion/introduc/490.HTM
[8] **http://www.iic.uam.es -** José Ramón Dorronsoro Ibero, Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid
[9] Española de Financiación y Contabilidad http://www.aeca.es/pub/refc/articulos.php?id=0473
[10] http://www.banrep.gov.co/docum/ftp/borra377.pdf

8. Bibliografía

8.1 *Documentos*

Kazmier, Leonard. Mc Graw-Hill (1999), "Estadística aplicada a la administración y la economía. "

Jay L. Devore; Thompson Editores (1998) "Probabilidad y estadística para ingeniería y ciencias."

Rincon M. (1984), "Conciliación censal y determinación de la población base"

Pagés, E. (1992), "Análisis factoriales simples y múltiples"

Lebart, L., Morineau, A. & Piron, M. (1995), "Statisitique exploratoire multidimensionnelle"

Serrano Cinca C. (2012): "La Contabilidad en la Era del Conocimiento"

Manuel E. Medina Ternero (1999): "Políticas Sociales para las personas mayores en el próximo siglo"

Philip Kotler-Gary Armstrong (2003): "Fundamentos de Marketing"

Revista Española de Innovación, Calidad e Ingeniería del Software (2007 Vol.3, No. 1)

Manual del usuario del sistema básico de PASW® Statistics 18

8.2 Sitios Web

Sitio Web Federación Latinoamericana de Bancos: www.felaban.com

Sitio Web Banco Central de la Republica Argentina: www.bcra.gov.ar

Sitio Web SCIELO (Scientific Electronic Library Online) www.scielo.org.ar/scielo.php

Sitio Web Universidad de Salamanca – Grupo de Estadística Aplicada: www.biplot.usal.es

Sitio Web Universidad Complutense de Madrid: www.ucm.es

Sitio Web Universidad de Murcia: www.um.es

Sitio Web Universidad Autónoma de Madrid: www.uam.es

Sitio Web Universidad Carlos III de Madrid- Departamento de Estadística: www.halweb.uc3m.es

Sitio Web Escuela de Ingería de Sistemas y Computación de la Universidad del Valle - Cali-Colombia: www.eisc.univalle.edu.co

Sitio Web Universidad Nacional de Colombia: www.unal.edu.co

Sitio Web Universidad de San Carlos de Guatemala: www.usac.edu.gt

Sitio Web CETYS (Centro de Enseñanza Técnica y Superior) www.cetys.mx

Sitio Web Protección de Datos Personales: www.protecciondedatos.com.ar

Sitio Web Marketing Bancario www.cruls6.blogspot.com.ar

Sitio Web Banco de la Republica de Colombia: www.banrep.gov.co

Sitio Web Asociación de Técnicos de Informática: www.ati.es

9. Anexo

9.1 Descripción de los métodos de segmentación estadística[11]:

Los algoritmos de formación de conglomerados se agrupan en dos categorías:

- Segmentación K-medias: Método de dividir el conjunto de observaciones en k conglomerados, en que k lo define inicialmente el usuario, en donde se sigue un criterio de optimización, de manera que cada observación/dato pertenezca únicamente a un conglomerado.

Este algoritmo, busca a través de un método iterativo:

- Los centroides (medias, medianas) de los k clusters
- Asignar a cada individuo a un cluster

El criterio de optimización buscado es el de "maximizar la homogeneidad dentro de los grupos y heterogeneidad entre los conglomerados".

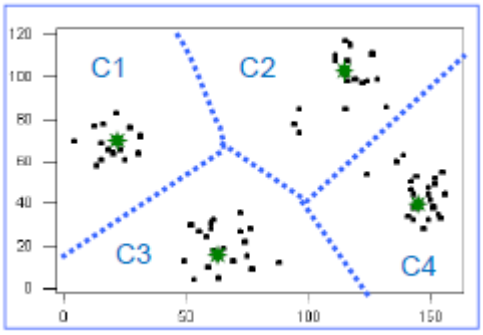

Figura 7. Segmentación k-medias

Una manera de cuantificar este criterio de optimización es minimizar la suma de los cuadrados de las diferencias entre cada dato y la media de su grupo

$$SSW = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\| x_i^k - \bar{x}^k \right\|^2$$

[11] http://www.uam.es/personal_pdi/ciencias/ajustel/docencia/ad/AD10_11_Cluster.pdf (Sitio Web Universidad Autónoma de Madrid)

Pasos del algoritmo “k – medias”:

1. Escoger un número “k” grupos iniciales en forma arbitraria y aleatoria, calculando, sus correspondientes centroides.

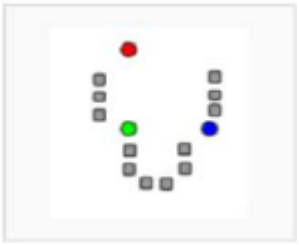

Figura 8. Paso 1

2. Se asocia cada “n” observación con la media mas cercana quedando comprendido en alguno de los “k” clusters.

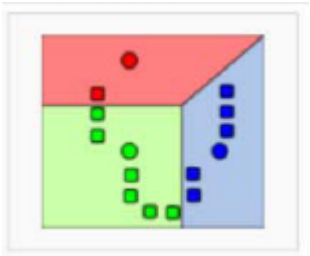

Figura 9. Paso 2

3. Se procede a recalcular, el centroide de cada “k” cluster convirtiéndose en las nuevas medias/centroides de los grupos.

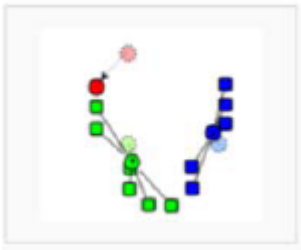

Figura 10. Paso 3

4. Se repita en forma iterativa los pasos 2 y 3, es decir, se recorren todas las observaciones re asignándolas al conglomerado cuyo centroide este a menor distancia y luego

se recalculan los centroides. Este proceso continúa hasta lograr alcanzar la convergencia, es decir, no se produzcan más reasignaciones.

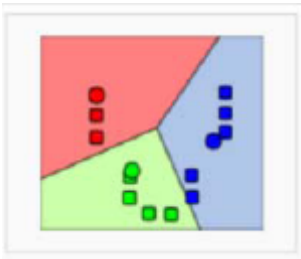

Figura 11. Paso 4

- Segmentación Jerárquica: Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados, siguiendo algún criterio especificado.

Dentro de los criterios para unir en métodos jerárquicos se destacan los siguiente “métodos de enlace” que utilizan la proximidad entre pares de individuos para unirlos como grupos de individuos.

1. Enlace sencillo: procede utilizan la mínima distancia/disimilitud entre dos individuos de cada grupo. Resulta muy útil para identificar casos atípicos.

2. Enlace completo: utiliza como metodología la máxima distancia/disimilitud entre dos individuos de cada grupo.

3. Enlace promedio: utiliza la media (mediana) de las distancias/disimilitud entre todos los individuos de los dos grupos.

4. Enlace de centroides: utiliza la disimilitud/distancia entre los “centros” de los grupos.

5. Método de Ward: utiliza la suma de las distancias al cuadrado a los centros de los grupos.

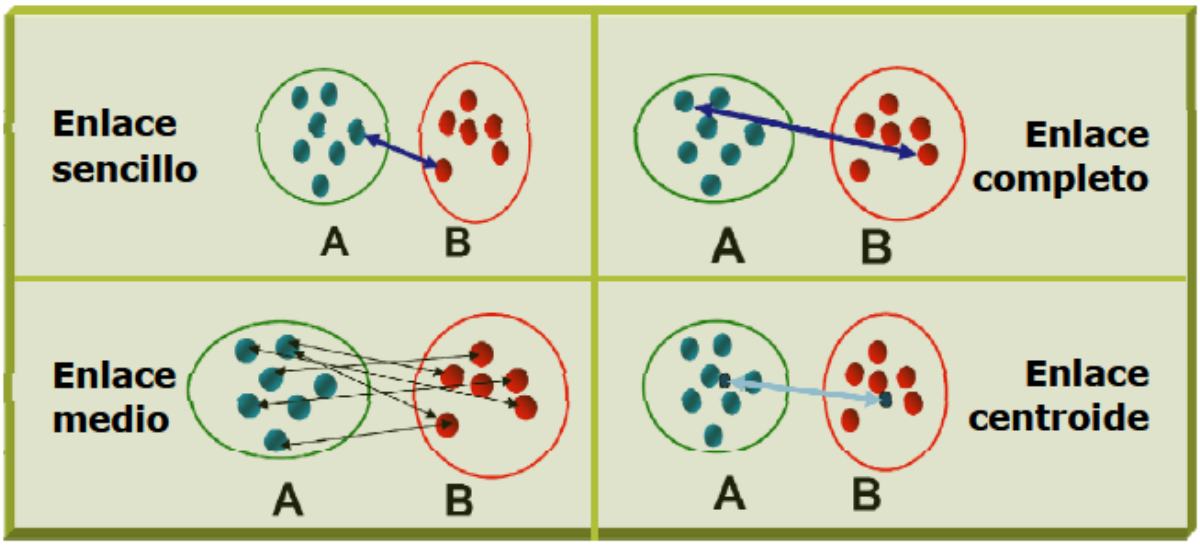

Figura 12. Métodos de enlace

Una representación que permite interpretar la metodología de la segmentación jerárquica es el dendograma.

Esta representación gráfica en forma de árbol esta constituida por:

- Trazos horizontales (verticales) que representan los clusters
- Trazos verticales (horizontales) que representan las etapas de la fusión. La separación entre las etapas de la fusión es proporcional a la distancia que se encuentran los grupos que se funden en esa etapa.

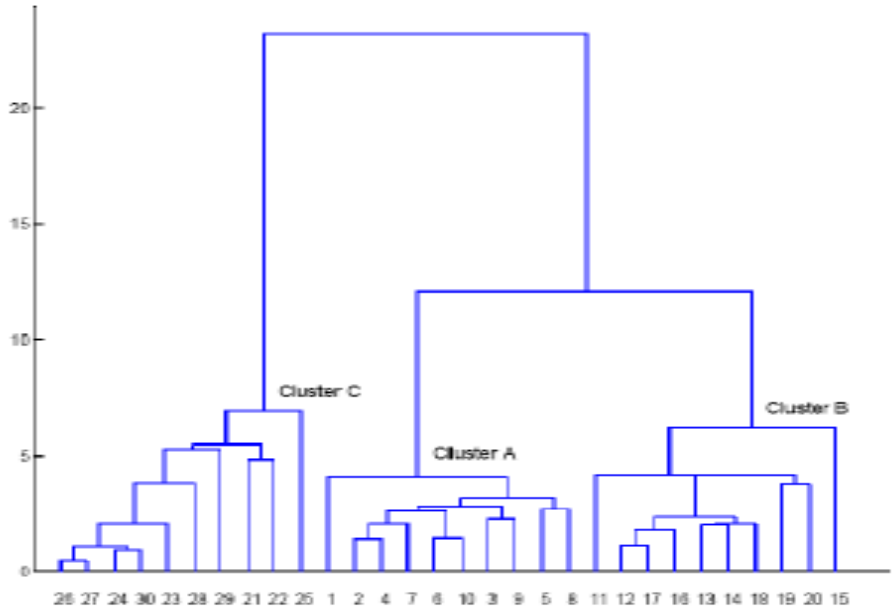

Figura 13. Ejemplo de Dendograma

Ejemplo método K-MEDIAS:[12]

Se tienen cuatro observaciones cuya matriz de datos se presenta a continuación:

$$\begin{bmatrix} 0 & 3 & 9 & 12 \\ 4 & 1 & 6 & 10 \\ 10 & 7 & 3 & 4 \\ 10 & 10 & 3 & 1 \end{bmatrix}$$

Tabla 28. Matriz de Datos

Se utiliza el método de las k-medidas para formar dos conglomerados. También se usan como metodología las distancias euclídeas.

Presentadas en forma de vectores, las cuatro observaciones (filas) son:

$$\underline{x}_1 = \begin{bmatrix} 0 \\ 3 \\ 9 \\ 12 \end{bmatrix} \qquad \underline{x}_2 = \begin{bmatrix} 4 \\ 1 \\ 6 \\ 10 \end{bmatrix} \qquad \underline{x}_3 = \begin{bmatrix} 10 \\ 7 \\ 3 \\ 4 \end{bmatrix} \qquad \underline{x}_4 = \begin{bmatrix} 10 \\ 10 \\ 3 \\ 1 \end{bmatrix}$$

[12] http://www.jorgegalbiati.cl/ejercicios_7/Conglomerados.pdf – Vigente al 05/07/11

Se comienza definiendo en forma arbitraria dos conglomerados iniciales:

$$A = \{\underline{x}_1\} \quad \text{y} \quad B = \{\underline{x}_2, \underline{x}_3, \underline{x}_4\}$$

Los dos centroides respectivos son:

$$\overline{\underline{x}}_A = \begin{bmatrix} 0 \\ 3 \\ 9 \\ 12 \end{bmatrix} \quad y \quad \overline{\underline{x}}_B = \begin{bmatrix} 8 \\ 6 \\ 4 \\ 5 \end{bmatrix}$$

Algoritmo Iterativo:

En primer lugar, se calculan las distancias de cada observación a los centroides de cada conglomerado.
En caso que una observación este a menor distancia del conglomerado vecino, se cambia de conglomerado y procede a recalcular los centroides pasando así a la siguiente iteración.

Iteración 1

La tabla a continuación muestra las distancias euclídeas (al cuadrado) de las observaciones a los centroides, arrancando por x1

centroide		$\overline{\underline{x}}_A$	$\overline{\underline{x}}_B$
observación	$\underline{x}_1$	0	147
	$\underline{x}_2$	33	70

Tabla 29. Cuadro de distancias al cuadrado i

Se produce un cambio de x2 del conglomerado B a A y termina la iteración 1 (no es necesario seguir probando con x3 ni x4).

Iteración 2

Se presentan los nuevos centroides, recalculados. Ahora A = {x1, x2 } y B = {x3, x4 }

$$\overline{\underline{x}}_A = \begin{bmatrix} 2 \\ 2 \\ 7{,}5 \\ 11 \end{bmatrix} \quad \overline{\underline{x}}_B = \begin{bmatrix} 10 \\ 8{,}5 \\ 3 \\ 2{,}5 \end{bmatrix}$$

Cuadro de distancias al cuadrado, partiendo de x3:

centroide		$\overline{x}_A$	$\overline{x}_B$
observación	$\underline{x}_3$	158.25	4.5
	$\underline{x}_4$	248.25	4.5
	$\underline{x}_1$	8.25	256.5
	$\underline{x}_2$	8.25	157.5

Tabla 30. Cuadro de distancias al cuadrado ii

Se observa que las cuatro observaciones quedaron bien clasificadas, entonces no hay necesidad de más cambios, por lo tanto los dos conglomerados resultantes son:

$$A = \{\underline{x}_1, \underline{x}_2\} \qquad \text{y} \qquad B = \{\underline{x}_3, \underline{x}_4\}$$

Ejemplo método JERARQUICO:

Se parte de la información de una muestra de un cuestionario, el mismo consiste en siete entrevistados que responden a una encuesta de diez preguntas, cada una de las respuestas contenía las siguiente alternativas para contestar: *a*, *b*, *c*, *d* y *e*

La matriz de datos de respuestas sería la siguiente:

	pregunta	1	2	3	4	5	6	7	8	9	10
encuestado	1	*a*	*b*	*b*	*c*	*a*	*b*	*b*	*a*	*a*	*d*
	2	*a*	*c*	*b*	*c*	*d*	*e*	*e*	*a*	*b*	*c*
	3	*c*	*b*	*b*	*c*	*d*	*a*	*b*	*c*	*a*	*d*
	4	*a*	*b*	*e*	*c*	*a*	*d*	*b*	*a*	*a*	*c*
	5	*c*	*c*	*b*	*b*	*d*	*a*	*b*	*c*	*d*	*d*
	6	*a*	*c*	*e*	*c*	*d*	*c*	*e*	*a*	*e*	*d*
	7	*b*	*b*	*c*	*a*	*a*	*a*	*b*	*c*	*a*	*b*

Tabla 31. Matriz de datos

Como distancia entre casos se usará el número (o la fracción, dividiendo el número por 10) de respuestas diferentes, y el método como de calculo para la distancia entre conglomerados, la del vecino más próximo.

Iteración 1

Se presenta la matriz de distancias entre los siete encuestados, siendo cada caso un conglomerado:

$D_1 =$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	0	6	4	3	7	6	6
(2)	6	0	7	6	7	4	10
(3)	4	7	0	6	3	7	5
(4)	3	6	6	0	9	6	6
(5)	7	7	3	9	0	7	7
(6)	6	4	7	6	7	0	10
(7)	6	10	5	6	7	10	0

Tabla 32. Matriz de distancias i

Inicialmente, se unen 1 con 4 y 3 con 5 a la distancia 3.

Iteración 2.

Entonces, la nueva matriz de distancias entre conglomerados queda dispuesta de la siguiente manera:

$D_2 =$

	(1, 4)	(2)	(3, 5)	(6)	(7)
(1, 4)	0	6	4	6	6
(2)	6	0	7	4	10
(3, 5)	4	7	0	7	5
(6)	6	4	7	0	10
(7)	6	10	5	10	0

Tabla 33. Matriz de distancias ii

Ahora, se unen (1, 4) con (3, 5) y (2) con (6) a la distancia 4.

Iteración 3.

La matriz de las distancias entre conglomerados queda:

$D_3 =$

	(1, 3, 4, 5)	(2, 6)	(7)
(1, 3, 4, 5)	0	6	5
(2, 6)	6	0	10
(7)	5	10	0

Tabla 34. Matriz de distancias iii

Luego se procede a unir (1, 3, 4, 5) con (7) a la distancia 5. Nótese que las distancias de unión van aumentando con cada paso. Es decir, cada vez se unen observaciones más disímiles.

Ultima matriz de distancias entre conglomerados:

$D_4 =$

	$(1,3,4,5,7)$	$(2,6)$
$(1,3,4,5,7)$	0	6
$(2,6)$	6	0

Tabla 35. Matriz de distancias iv

Finalmente, se unen todos en un sólo conglomerado, a la distancia 6.

El siguiente gráfico, el dendongrama, ilustra la forma en como se fueron uniendo los conglomerados hasta formar uno solo. La escala dispuesta en forma horizontal corresponde a la distancia en que produjeron las uniones correspondientes según cada caso.

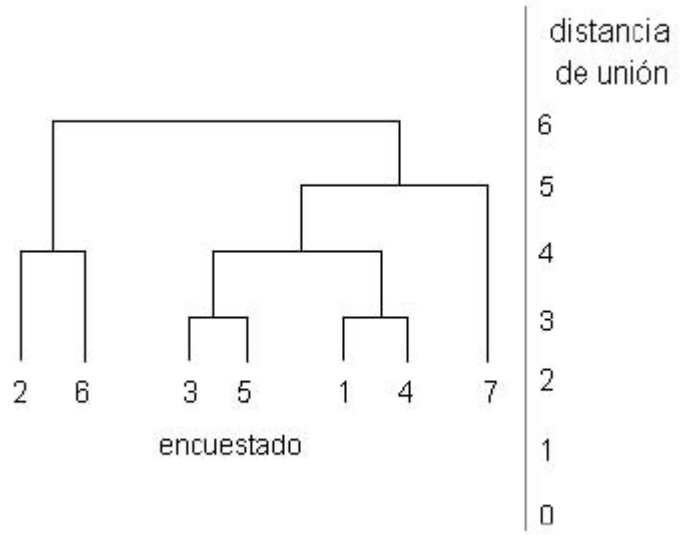

Figura 14. Ejemplo de Dendograma ii

Interpretando el grafico se desprende que:

- Si deseamos tener dos conglomerados, estos serían (1,3,4,5,7) y (2,6).
- Si deseamos tener tres, serían (7), (1,3,4,5) y (2,6).
- Si queremos 5, estos serían (1,4), (3,5), (2), (6) y (7).

9.2 Aplicación de Herramienta SPSS [13]

Análisis de Componentes Principales – Análisis Factorial

En la barra de menú principal se presiona: Analizar- Reducción de dimensiones- Factor

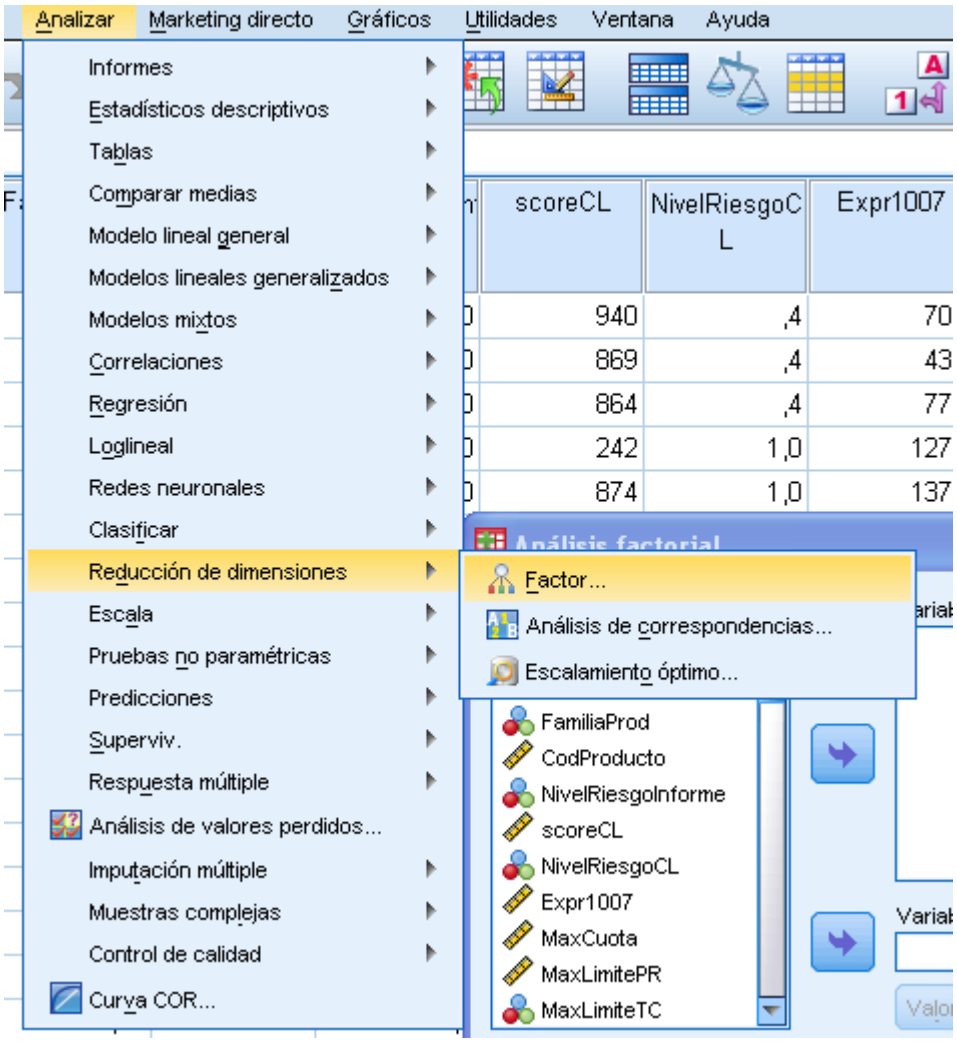

Figura 15. Ejemplo Herramientas SPSS

[13] Manual del usuario del sistema básico de IBM SPSS Statistics 18
ftp://ftp.soc.uoc.gr/Psycho/spss/spss18.0/Documentation/Spanish/Manuals/PASW%20Statistics%2018%20Core%20System%20User%27s%20Guide.pdf

En variables arrastramos aquellas que queramos someterlas al análisis factorial.

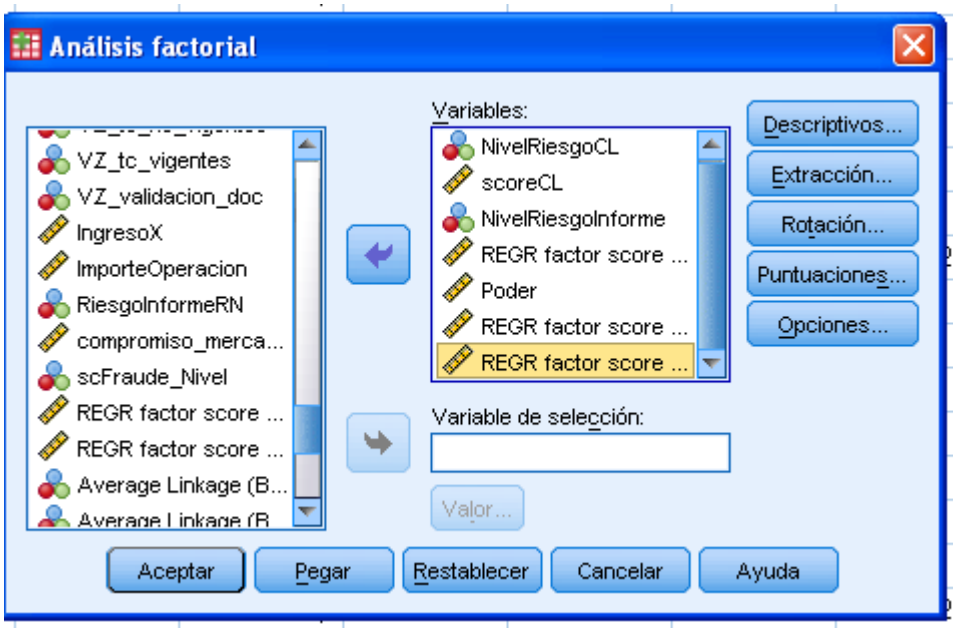

Figura 16. Análisis factorial – Selección de Variables

En descriptivos podemos seleccionar Estadísticos y detalles para la Matriz de correlaciones.

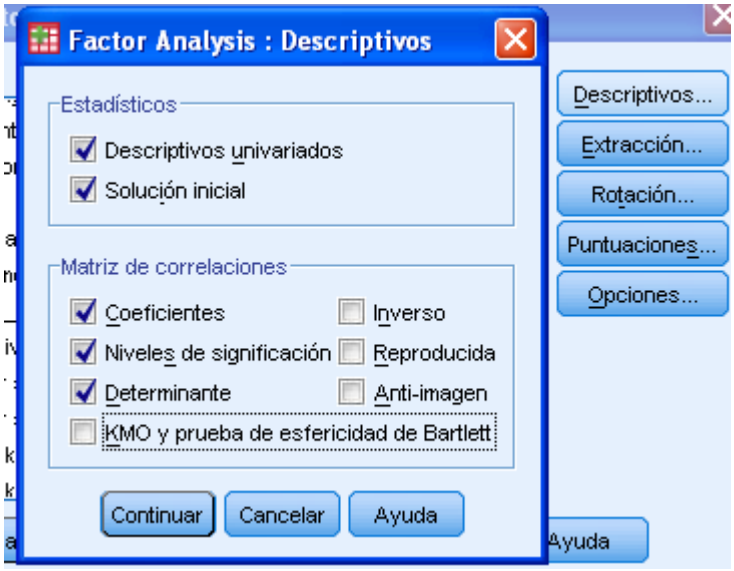

Figura 17. Análisis factorial – Estadísticos descriptivos

En la opción Extracción: escogemos como método componentes principales, que analizar y extraer autovalores mayores que 1. También fijamos el máximo de iteraciones para la convergencia que por default es 25.

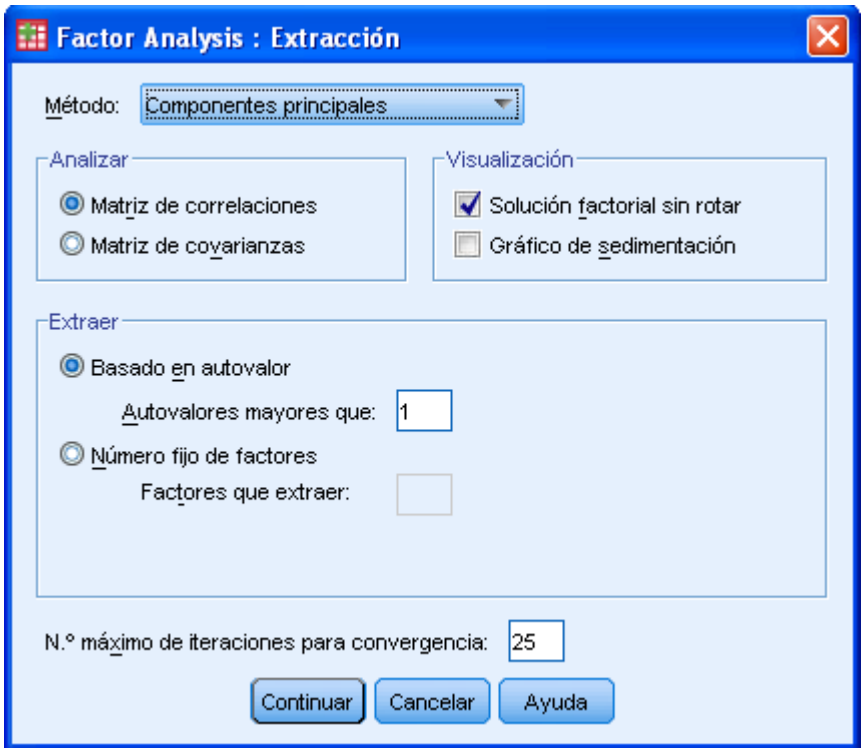

Figura 18. Análisis factorial – Extracción

Método jerárquico

En la barra de menú principal se presiona: Analizar- Clasificar- Conglomerados Jerárquicos

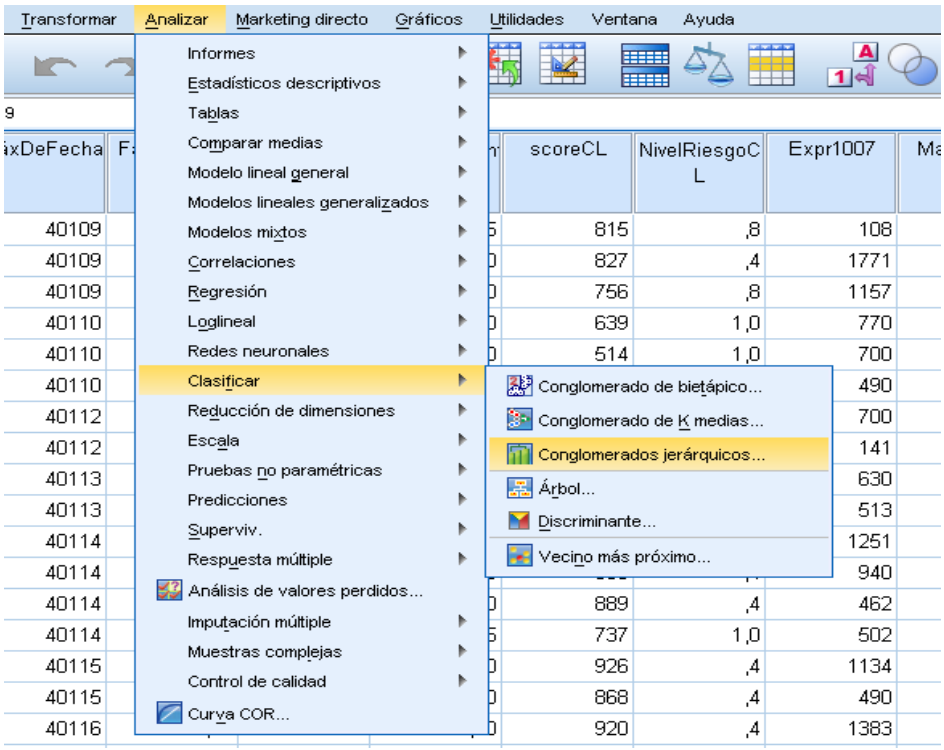

Figura 19. Clasificar – Conglomerados jerárquicos

Luego podemos escoger las variables que deseamos incorporar al método:

Figura 20. Análisis de conglomerados jerárquicos - Variables

En "Estadísticos" solicitamos "Historial de conglomeración" y "Matriz de distancias" con un rango de soluciones de 2 a 10.

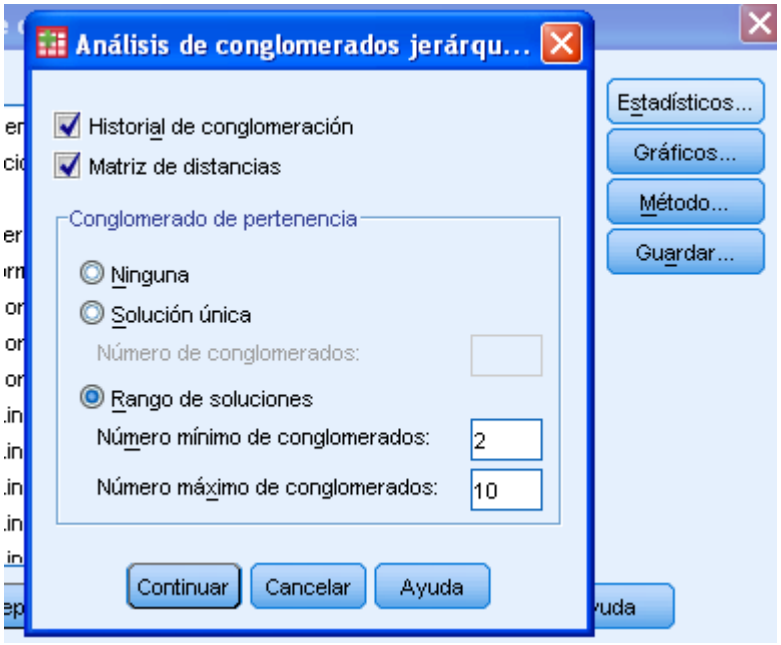

Figura 21. Análisis de conglomerados jerárquicos - Estadísticos

En la opción de "Gráficos" solicitamos únicamente el Dendograma, indicando ninguna en la opción "Témpanos".

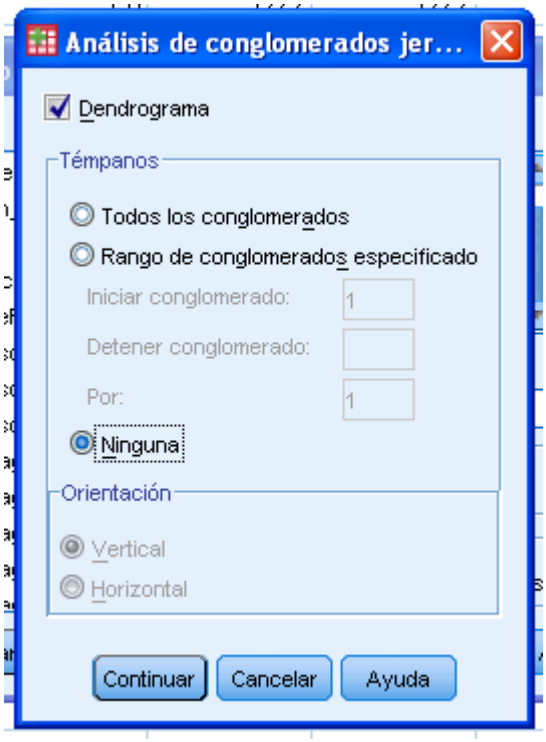

Figura 22. Análisis de conglomerados jerárquicos - Gráficos

Finalmente, en la opción guardar se solicita que se generen nueve nuevas variables, donde cada una contiene al conglomerado al que se le asigna a cada uno de los individuos.
Por eso solicitamos un rango de soluciones de 2 a 10.

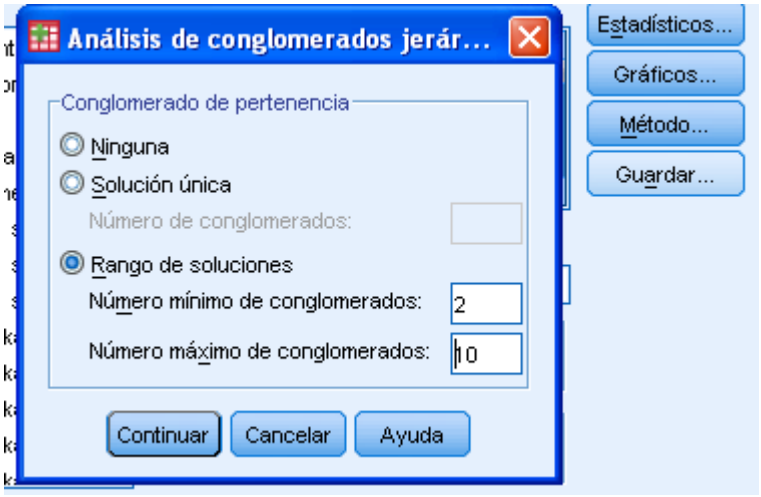

Figura 23. Análisis de conglomerados jerárquicos - Guardar

Método K-medias

En la barra de menú principal se presiona: Analizar- Clasificar- Conglomerado de K medias

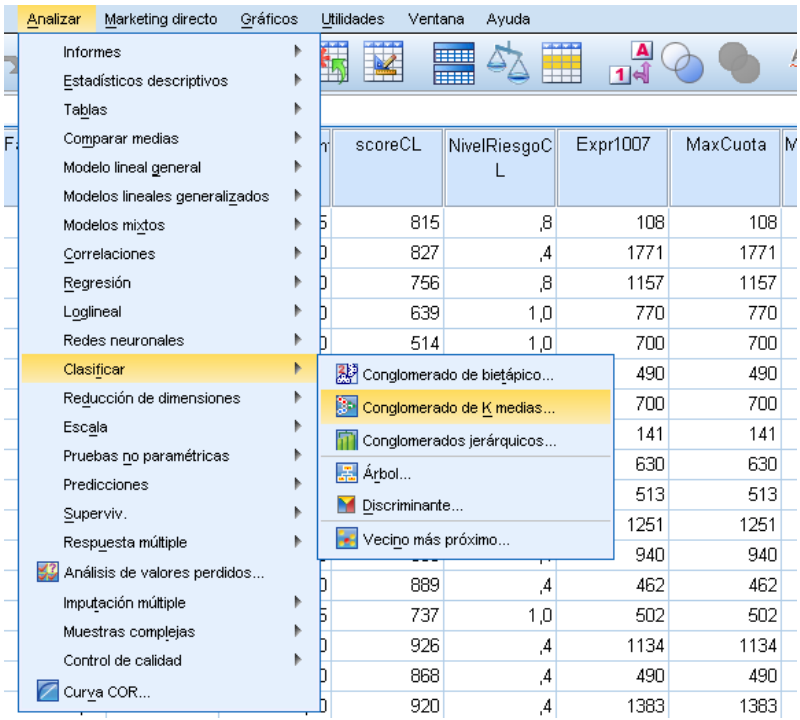

Figura 24. Clasificar – Conglomerado de K medias

Se señalan las variables para incluirlas como tales a la opción de "Variables"

Se indica el número de clusters que deseamos en la opción Número de Clusters

En la opción Método se elige Iterar y clasificar, donde determina las sucesivas iteraciones e indica cómo llevar a cabo la clasificación final.

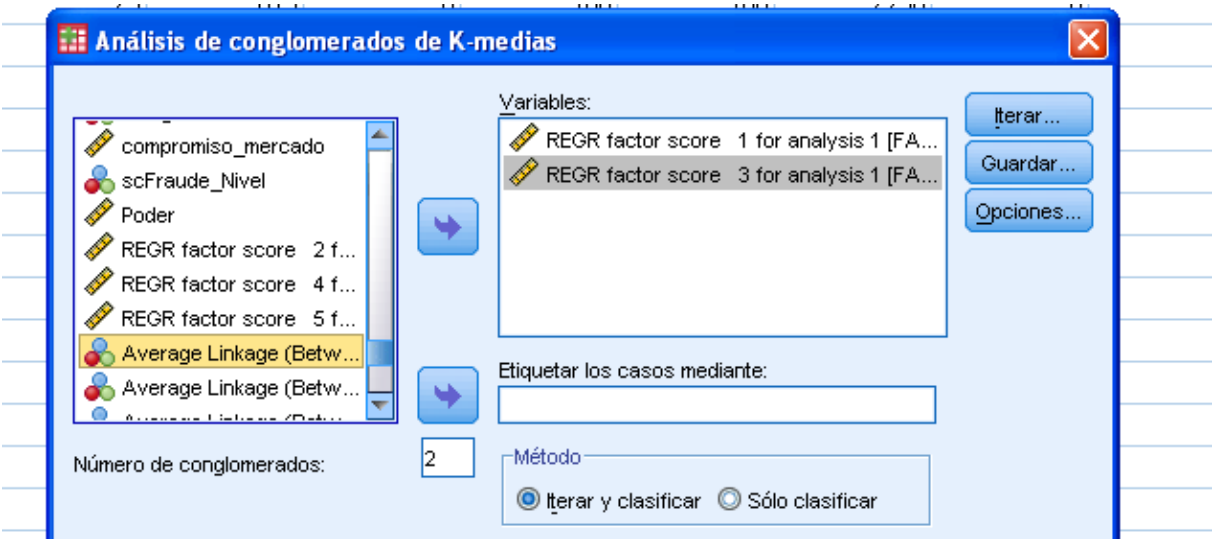

Figura 25. Análisis de conglomerados de K-medias - Variables

En la opción Centros de conglomerados se puede optar por especificar la carpeta con los centro de cluster iniciales.

En la opción "iterar" podremos determinar un máximo de iteraciones y un criterio de convergencia (Por defecto el proceso se realiza con 10 iteraciones y 0 convergencias.)

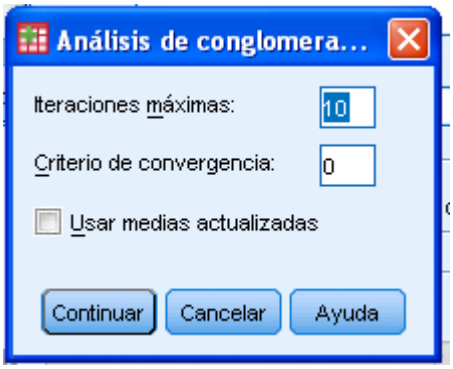

Figura 26. Análisis de conglomerados de K-medias - Iterar

También es posible utilizar la opción "usar medias actualizadas" así los centros de cada cluster cambiarán tras la suma de cada uno de los elementos. En cambio, en caso de no seleccionarla, calculará los centros de cluster después de que todos los elementos hayan sido ubicados en cada cluster.

Se avanza seleccionando Continuar.

Pulsando la opción "Guardar" se podrán guardar las nuevas variables en archivos de datos en los que podremos determinar la clase de cluster de cada objeto (Conglomerado de pertenencia) y Distancia desde centro del conglomerado.

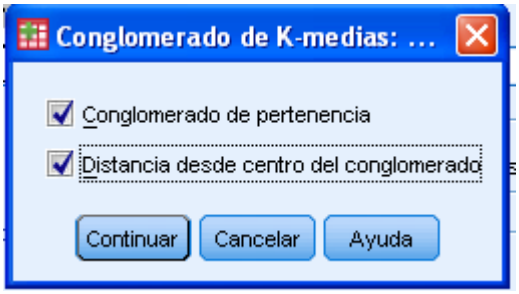

Figura 27. Análisis de conglomerados de K-medias - Guardar

El botón de opciones en estadísticos adicionales se opta por la información adicional que se desee: Centro de conglomerados iniciales, Tabla de Anova, Información del conglomerado para cada caso.

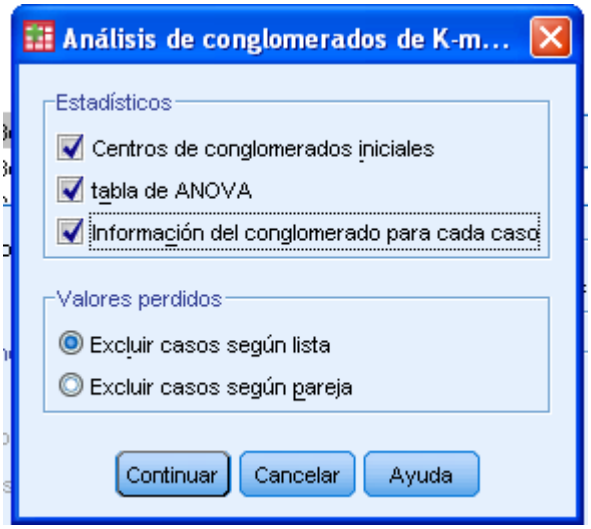

Figura 28. Análisis de conglomerados de K-medias – Estadísticos