ESCUELA POLITÉCNICA NACIONAL

ESCUELA DE CIENCIAS

MODELO DE SEGMENTACIÓN DE CLIENTES EN EL ÁREA BANCARIA: UN ESTUDIO DESDE LA DEPURACIÓN DE DATOS ATÍPICOS HASTA LA OBTENCIÓN DEL SCORE DE CLIENTES RENTABLES

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

EDGAR FABIÁN VARGAS FERNÁNDEZ
edgarmhfa@yahoo.com

DIRECTOR: MAT. JAIME ANDRADE
jaime.andrade@epn.edu.ec

QUITO, SEPTIEMBRE 2009

DECLARACIÓN

Yo, EDGAR FABIÁN VARGAS FERNÁNDEZ, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

Edgar Fabián Vargas Fernández

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por EDGAR FABIÁN VARGAS FERNÁNDEZ bajo mi supervisión.

_____

Mat. Jaime Andrade
DIRECTOR DE PROYECTO

AGRADECIMIENTOS

Mi sincero agradecimiento:

Al Mat. Jaime Andrade por su apoyo en la elaboración de esta tesis.

Al Dr. Juan Carlos De los Reyes por inculcar en mi la pasión por el Análisis Numérico y el Cálculo Científico.

A Raúl Muñoz por su sincera amistad y apoyo permanente a los largo de mis estudios.

DEDICATORIA

A mi madre.

Edgar Vargas

RESUMEN

En la presente tesis se desarrolla la segmentación de clientes de una institución bancaria, partiendo desde la búsqueda y depuración de datos atípicos multivariantes sobre la base de estudio para después realizar los modelamientos de: conglomerados, supervivencia, tiempo de vida, valor del tiempo de vida; hasta llegar a un score sobre este último indicador.

Palabras claves: Segmentación de mercados, datos atípicos multivariantes, análisis de conglomerados, análisis de supervivencia, tiempo de vida, valor del tiempo de vida.

ABSTRACT

The present thesis develops customer segmentation of a banking institution, starting from the search and purification of multivariate outliers on the basis of study and then perform models: clustering, survival, lifetime, lifetime value, until reach a final score on this indicator.

Keywords: Market segmentation, multivariate outliers, cluster analysis, survival analysis, lifetime, life time value.

TABLA DE CONTENIDO

AGRADECIMIENTOS iv

RESUMEN vi

Capítulo 1: CONCEPTOS BÁSICOS 1
1.1 Clasificación del mercado según el tipo de demanda 2
1.2 Segmentación de mercados 3
1.2.1 Ventajas de una correcta segmentación 3
1.2.2 Características de un segmento de mercado 4
1.2.3 Variables de segmentación 5

Capítulo 2: ANÁLISIS DE DATOS ATÍPICOS 8
2.1 Datos atípicos 8
2.2 Efecto de datos atípicos 11
2.3 Identificación de grupos atípicos método multivariante 14
2.4 Algoritmo de búsqueda de datos atípicos 16

Capítulo 3: ANÁLISIS DE CONGLOMERADOS 20
3.1 Procedimiento de un análisis conglomerado 20
3.1.1 Planteamiento del problema 20
3.1.2 Selección de una medida de distancia o semejanza 21
3.1.3 Selección de un procedimiento conglomerado 21
3.2 Elección del número de conglomerado 24
3.3 Interpretación y perfiles del conglomerado 24
3.4 Evaluación de la confiabilidad y validez 24
3.5 Análisis conglomerado en dos fases 24
3.5.1 Número de conglomerados 26
3.5.2 Validez de la solución por conglomerados en dos fases . . . 27
3.6 Interpretación de los resultados 27
3.7 Verificación de la solución 27

Capítulo 4: ANÁLISIS DE SUPERVIVENCIA 29
4.1 Definición 29
4.2 Tipo de modelos 30
4.2.1 Modelos paramétricos 30
4.2.2 Modelos No Paramétricos 31
4.3 Modelo de riesgo proporcional de Cox 31

4.4 Estimación de los coeficientes y contrastes de hipótesis 31
4.5 Estrategias de modelización 33
4.6 Relación entre la supervivencia individual y de la muestra 35

Capítulo 5: ANÁLISIS: LT Y LTV 37
5.1 Lifetime 37
5.2 Ecuación del Life Time (LT) 38
5.3 Life Time Value 42

Capítulo 6: SCORE COMO FUNCIÓN DEL LTV 45
6.1 Criterio de scoring para los clientes 45
6.2 Resultados numéricos del modelo propuesto 45
6.2.1 Depuración de datos atípicos 46
6.2.2 Análisis de conglomerados 47
6.2.3 Resultados del Análisis de Cluster 47
6.3 Resultados Análisis de Supervivencia 54
6.3.1 Cálculo Supervivencia 54
6.3.2 Resultados Supervivencia 55
6.3.3 Análisis LT y LTV 67
6.3.4 Score 78

CONCLUSIONES 79

Apéndice A: Códigos fuente de implementación 80
A.1 Algoritmo de datos atípicos multivariantes 80
A.2 Implementación del LT 99
A.2.1 Conglomerado 1 99
A.2.2 Conglomerado 2 101
A.2.3 Conglomerado 3 103
A.2.4 Conglomerado 4 106
A.3 Implementación del LTV 108
A.3.1 Conglomerado 1 108
A.3.2 Conglomerado 2 112
A.3.3 Conglomerado 3 115
A.3.4 Conglomerado 4 118
A.3.5 Tabla Rentabilidad promedio vs. edad 121

Bibliografía 124

ÍNDICE DE FIGURAS

2.1 Dato atípico (ejemplo gráfico) 9
2.2 Ejemplo univariante de dato atípico 14

3.1 Análisis por conglomerado 21
3.2 Procedimiento por conglomerado 22

5.1 Función S_0 dada por (5.5) 39
5.2 Ejemplo: Ajuste la función S_0 40
5.3 Ejemplo: Función S_0 extendida en el eje de antigüedad 41

6.1 Matriz de gráfico de dispersión para la base de datos a depurar de datos atípicos 46
6.2 Tabla de Porcentajes Cluster 50
6.3 Ciclo de vida por cada cluster 51
6.4 Distribución de hombres y mujeres dentro de cada cluster 51
6.5 Código de Negocio del cliente por cada cluster 52
6.6 Estado civil por cada cluster 52
6.7 Provincia 52
6.8 Crédito Ventanilla 53
6.9 Débito en Ventanilla 53
6.10 Débito ATM 53
6.11 Resumen de la Regresión 55
6.12 Pruebas de la Regresión 55
6.13 Variables de la Ecuación de Supervivencia 56
6.14 Variables excluidas en la Ecuación de Supervivencia 57
6.15 Función de supervivencia 58
6.16 Función de Hazard 59
6.17 Resumen de la Regresión 60
6.18 Pruebas de la Regresión 60
6.19 Variables de la Ecuación de Supervivencia 61
6.20 Variables excluidas en la Ecuación de Supervivencia 61
6.21 Función de supervivencia 62
6.22 Función de Hazard 62
6.23 Resumen de la Regresión 63
6.24 Pruebas de la Regresión 63
6.25 Variables de la Ecuación de Supervivencia 63
6.26 Función de supervivencia 64

6.27 Función de Hazard 64
6.28 Resumen de la Regresión 64
6.29 Pruebas de la Regresión 65
6.30 Variables de la Ecuación de Supervivencia 65
6.31 Función de supervivencia 65
6.32 Función de supervivencia 66
6.33 Función de Hazard 66
6.34 Aproximación de S_0, Conglomerado 1 67
6.35 Aproximación de S_0, Conglomerado 2 68
6.36 Aproximación de S_0, Conglomerado 3 69
6.37 Aproximación de S_0, Conglomerado 4 70
6.38 Función de rentabilidad vs. edad, Conglomerado 1 72
6.39 Función de rentabilidad vs. edad, Conglomerado 2 74
6.40 Función de rentabilidad vs. edad, Conglomerado 3 76

ÍNDICE DE TABLAS

2.1 Valores simulados de β_p para el algoritmo (2.2) 18

3.1 Notaciones análisis de conglomerados en dos etapas 25

6.1 Variables a incluir en el análisis por conglomerados 48
6.2 Resultados corrida cluster (Criterios) 49
6.3 Distribución de Clusters 50
6.4 Centroides de variables continuas en el modelo de cluster 51
6.5 Cuartiles de LTV por conglomerado 79

ÍNDICE DE ALGORITMOS

2.1 Algoritmo de búsqueda de proyecciones (*projection pursuit*) 16
2.2 Algoritmo de detección de datos atípicos multivariantes 17
2.3 Algoritmo para el cálculo de la dirección $\mathbf{d}$ del algoritmo 2.2 18
5.1 Cálculo de LT 39
5.2 Cálculo de LTV 44

Capítulo 1

CONCEPTOS BÁSICOS

La necesidad primordial en una institución bancaria es identificar sus mejores clientes y sus necesidades a través de la relación cliente–institución bancaria, esto se consigue con una correcta segmentación. Segmentar un mercado es una estrategia de negocios para construir relaciones rentables y duraderas con el cliente, esto implica: levantar conocimiento acerca de los clientes, crearles propuestas valiosas de acuerdo a sus necesidades y mejorar su experiencia como cliente dentro de la institución bancaria. Los modelos clásicos de segmentación se basan en: variables demográficas, económicas, culturales, políticas, legales y tecnológicas de los clientes, se propone una alternativa diferente a las propuestas clásicas de segmentación utilizando modelos estadísticos robustos como la detección de datos atípicos multivariantes, ajuste de la función de supervivencia con una función del tipo no lineal, asi cómo también el uso de funciones de Gausianas para el ajuste de curvas Rentabilidad vs. edad, entre otros.

Las técnicas de segmentación no sólo se limitan a clientes de una empresa o un negocio, modelos de segmentación han sido usados en áreas como: Astronomía, Biología, Psicología, etc.. Como ejemplo del primero se tiene una segmentación de estrellas utilizando las variables: luminosidad y temperatura de las estrellas; de hecho desde que estamos en primaria tenemos una noción básica de lo que es segmentación, por ejemplo cuando creamos equipos para jugar fútbol separando compañeros por tamaño de cada jugador, habilidad de juego, edad, etc. Con el ejemplo antes mencionado se ve que dividir, clasificar en grupos ideales de jugadores garantiza el éxito o no en el juego, con la misma idea esto se torna indispensable en una institución bancaria en la que se busca las características de los clientes más rentables para poder hacer que los clientes que aún no lo son lleguen a convertirse en uno de ellos.

En la Administración de la relación del cliente, o Customer Relationship Management (CRM), la segmentación es usada para clasificar clientes de acuerdo a criterios ya sean de negocios o bajo un fundamento teórico de fondo, consúltese [5].

En este capítulo se revisarán conceptos básicos provenientes de Marketing, empezando por la definición de mercado hasta las variables utilizadas comúnmente para realizar una segmentación de mercado.

1.1 Clasificación del mercado según el tipo de demanda

El concepto de mercado ha tenido a lo largo del tiempo diversas acepciones; cada una de ellas ha sido el resultado de la percepción de diversas disciplinas. Sin embargo, dicho concepto se tiene que abordar desde el punto de vista de mercadotecnia, sobre todo si quiere hacerse un estudio de segmentación de mercados.

Existen varias definiciones de mercado, cada una de ellas con un enfoque disciplinario distinto; por ejemplo para los economistas, el mercado será el lugar geográfico en donde coinciden oferentes y demandantes para realizar una transacción comercial. En el comportamiento del mercado influyen muchos factores, uno de ellos es la demanda; a partir de su comportamiento se pueden generar una clasificación de la misma.

La demanda de mercado se puede definir como la cantidad de productos que están dispuestos a adquirir los consumidores en un período determinado, bajo condiciones de precio, atención y servicio.

Bajo la perspectiva de la demanda, el mercado puede definirse como el conjunto de consumidores y/o compradores que ejercen una demanda específica sobre un producto y se clasifican como:

1. **Mercado disponible**. Son todos aquellos consumidores que tienen una necesidad específica y cuentan con las características necesarias para consumir un producto. Un ejemplo de mercado disponible, son todos aquellos consumidores que tienen la necesidad de un jabón líquido de manos, que además cuentan con las características del poder adquisitivo, costumbre, etc., para comprarlo.

2. **Mercado real**. Son todos aquellos consumidores del mercado disponible que compran un producto específico. El ejemplo sería todos los consumidores que compran el jabón líquido. Es importante notar que no todos los consumidores que tienen una necesidad y las características específicas para comprar un producto son consumidores reales, pero si consumidores disponibles.

3. **Mercado potencial**. Es el conjunto de consumidores que no forman parte del mercado real, pero en ocasiones si pueden formar parte del mercado disponible. Estas personas no consumen el producto específico, debido a que no tienen las características del segmento o porque consumen otro producto. Como ejemplo se puede citar a los niños, que aún no tienen las car-

acterísticas de poder adquisitivo para comprar y consumir un determinado producto, pero en un futuro podrán hacerlo.

4. **Mercado meta o mercado objetivo**. Puede definirse como el conjunto de consumidores pertenecientes al mercado disponible, que pueden formar parte del mercado real y potencial, al cual se dirigen todos los esfuerzos y acciones de mercadotecnia de la institución, con la finalidad de que todos ellos se conviertan en consumidores reales de un determinado producto.

1.2 Segmentación de mercados

La definición y reconocimiento del tipo de mercado donde se comercializan un determinado producto nos brinda una visión generalizada del tipo de actividades comerciales que debemos realizar. Sin embargo, los mercados son diversos y cada uno de ellos está integrado de tipos de persona de edad, sexo, costumbres, y personalidades diferentes.

Esto nos lleva a pensar que si cada ser humano es único, sus gustos y preferencias así como los productos que compra serán también distintos. Lógicamente es imposible tener productos que hayan sido diseñados para el uso de un solo individuo, pero sí se pueden hacer grupos de personas, es decir, segmentos que tengan gustos y preferencias parecidos. Éste es el origen de la segmentación de mercados, que puede definirse como la división de un universo heterogéneo en grupos con al menos una característica homogénea.

Se puede decir también que la segmentación de mercado es el proceso de separar clientes, o clientes potenciales en diferentes grupos o segmentos con clientes que comparten un conjunto de necesidades que se pueden satisfacer mediante una oferta diferente de Marketing.

1.2.1 Ventajas de una correcta segmentación

La segmentación de mercados es una actividad que brinda certeza a la institución en el desarrollo de sus actividades, en forma particular brinda algunas ventajas como:

- **Certidumbre en el tamaño del mercado**. Al conocer el grupo, podrá calcularse en casi todos los casos el tamaño del mercado, es decir, el número aproximado de personas que conforman el mercado disponible, o sea, el número aproximado de personas que pueden comprar un determinado producto.

- **Claridad al establecer planes de acción**. Al conocer a los integrantes del mercado meta se tendrá claridad en los planes de acción a desarrollar.
- **Identificación de los consumidores integrantes del mercado**. Conocer a nuestros consumidores nos dará certeza en las decisiones de mercado que se tomen.
- **Reconocimiento de actividades y deseos del consumidor**. Las costumbres de los consumidores nos sirven para saber cómo satisfacer sus necesidades en forma oportuna.
- **Simplificación en la estructura de marcas**. Es posible evitar una gran variedad de marcas no productivas al conocer a nuestro mercado.
- **Facilidad para la realización de actividades promocionales**. Las actividades promocionales estarán dirigidas únicamente al grupo de interés, de modo tal que se cuidarán los recursos de la institución bancaria y se tendrán resultados más efectivos.
- **Simplicidad para planear**. La planeación se simplifica al conceptualizar las actividades para un grupo específico del mercado.

1.2.2 Características de un segmento de mercado

Para que un segmento de mercado sea realmente eficaz debe ser:

- **Medible**, es decir, debemos conocer el número aproximado de elementos que lo conforman.
- **Susceptible a la diferenciación**, es decir, debe responder a un programa de mercadotecnia distinto a otros segmentos y con otros productos.
- **Accesible**, o sea, se debe llegar al público integrante de manera sencilla.
- **Susceptible a las acciones planeadas**, esto se referie a la capacidad que se debe tener para satisfacer al mercado identificado con las acciones que son posibles para la institución bancaria.
- **Rentable**, es decir, debe representar un ingreso que justifique la inversión.

1.2.3 Variables de segmentación

Para segmentar un mercado es necesario considerar una serie de variables que nos brinden la posibilidad de hacerlo de manera clara y precisa. Las variables de segmentación que deben ser consideradas en la segmentación del mercado son:

Demográficas

Tienen la característica particular que se pueden medir en forma estadística, dentro de este grupo encontramos variables como:

- Edad
- Sexo
- Nivel socio–económico
- Estado civil
- Nivel de instrucción

Variables como el sexo, edad, estilo de vida, nivel de instrucción, estado civil, etc., cuando se utilizan para definir los segmentos están implícitamente afirmando, por ejemplo, de que cada persona de 30 a 35 años, responderá a la misma oferta. Tomando en cuenta estudiantes de una escuela, con este tipo de segmentación se esperaría que todos ellos usen la misma ropa, tengan los mismos intereses y que todos ellos prefieran una marca determinada de auto para conducir. Sin embargo este enfoque bien podría estar ignorando uno o más de los siguientes puntos:

- diferentes divisiones y departamentos existentes en un negocio pueden tener diversas aplicaciones para el producto o servicio que ofertan. Por ejemplo se puede dar dos opciones de oferta a las mismas personas dentro de un rango de edad dependiendo de un factor adicional que vendrá dado más por criterio de negocio.
- incluso dentro de una sola división de una empresa, habrían diferentes aplicaciones para el producto o servicio que se brinda, a su vez, pueden tener diferentes especificaciones relacionadas con ellas.

Aunque las variables demográficas por sí solas no pueden definir un segmento porque no define la propuesta de un segmento, tienen un papel importante que desempeñar en una segmentación. Esta información básica sobre los clientes

puede ser utilizada para identificar las características de perfiles de particulares asociados con los clientes en cada segmento. En otras palabras, las variables demográficas ayudan a identificar quien está en cada segmento, a su vez que ayuda a determinar cómo llegar a ellos.

Geográficas

Son las variables ambientales que dan origen a las diferencias en la personalidad de comunidades por su estructura geográfica. En este grupo se encuentran variables como:

- Unidad geográfica
- Condiciones geográficas
- Raza
- Tipo de población

Así como las variables demográficas, segmentos basados en áreas geográficas, asumen que cada individuo en un área predeterminada debe reaccionar de la misma manera a una oferta particular, basta con darnos cuenta de nuestros vecinos en nuestro vecindario, cada uno tiene diferentes preferencias en cuanto a lugar de compra, de comida, vestido, etc. Nuevamente considerando únicamente estas variables no pueden definir segmentos de mercado.

Psicográficas

En la actualidad las variables psicográficas han tenido una influencia total en los motivos y decisiones de compra del consumidor, no son claramente perceptibles y no siempre pueden medirse, sin embargo, representan un excelente medio para posicionar y comercializar los productos de una empresa, por lo que merecen ser estudiadas con detenimiento. Las variables psicográficas se integran como sigue:

- Grupos de referencia
- Clase social
- Personalidad
- Cultura
- Ciclo de vida familiar
- Motivos de compra

De posición del usuario o de uso

Estas variables se refieren a la disposición que tiene el consumidor ante la posible compra de un producto, es decir, a la posición que juega dentro de nuestro segmento de mercado. Las variables de posición del usuario o de uso son:

- Frecuencia de uso
- Ocasión de uso
- Tasa de uso
- Lealtad
- Disposición de compra

Lo que podemos concluir, es que, si bien cada uno de los anteriores enfoques para la formación de segmentos de clientes pueden ser administrativamente convenientes, por si mismos no definen segmentos, más bien nos dicen que son un útiles en conjunto para llevar a cabo una correcta segmentación, de aquí que, surge la necesidad de utilizar un modelo de segmentación basado no sólo en un tipo de variables, sino más bien utilizarlas en conjunto mediante un modelo de clasificación que utilize tanto variables continuas como categóricas (detallado en el Capítulo 3, pág 20), para las variables continuas se realiza primero una depuración de datos atípicos (detallado en el Capítulo 2, pág 2). Sin embargo como el objetivo primordial de la institución bancaria es identificar sus mejores clientes, rentablemente hablando, se parte primero de un análisis de supervivencia aplicado a los clientes (detallado en el Capítulo 4, pág 4) para posteriormente calcular el tiempo esperado de vida del cliente o *Lifetime* (LT) y su valor presente neto o *Lifetime Value* (LTV) (detallado en el Capítulo 5,pág 5) y finalmente tener un *score* sobre este último indicador (detallado en el Capítulo 6,pág 6).

Capítulo 2

ANÁLISIS DE DATOS ATÍPICOS

Ante cualquier estudio estadístico, es de vital importancia realizar un estudio - minucioso de la calidad de la información, pues al no realizalo cualquier análisis posterior podría contendrá resultados distorsionados y realizar conclusiones utilizando esta información nos llevaría a una toma de decisión errónea.

Está es la principal razón de realizar un estudio de la calidad de la información, se tratará entonces de identificar aquellos datos que se encuentran en una base de datos que puedan destruir cualquier otro estudio posterior a este, y en vista que se utilizan siempre bases de datos con más de una variable continua, determinar datos atípicos multivariantes sobre la misma de forma multivariante, permitirá utilizar una técnica robusta en búsqueda de estos datos.

2.1 Datos atípicos

Definición 2.1.1 (Dato atípico)**.** *Llamaremos datos atípicos a aquellas observaciones que parecen haberse generado de forma distinta al resto de los datos. Los datos atípicos pueden ser causadas por errores de medición o transcripción, cambios en el instrumento de medición o heterogeneidad intrínseca de los elementos observados.*

Como ejemplo de la definición imaginemos que estamos estudiando las características de los clientes de un Banco, donde la gran mayoría tiene una rentabilidad promedio de $\$500$, pero se ha incluido en la muestra un cliente cuya rentabilidad es exageradamente alta $\$10,000$. Esta observación será atípica y corresponde a una heterogeneidad real de los datos. Es importante detectarla ya que obtendremos una mejor descripción de los datos separando ambos tipos de rentabilidad.

De los análisis efectuados sobre datos recogidos en condiciones de estrecho control, revelan que es frecuente que aparezcan entre 1% y 3% de observaciones atípicas respecto al resto de la muestra, cuando los datos se han recogido sin un cuidado especial, la proporción de datos atípicos puede llegar al 5% de observaciones y ser incluso mayor. La característica de un sólo valor atípico es simple ya que, por definición, debe estar alejada del resto, con lo que la distancia entre el dato atípico y el resto de la observaciones será grande.

Definición 2.1.2 (Dato atípico (*definición alternativa*))**.** *Se puede definir un atípico como un punto que se encuentra lejos del centro de los datos.*

En este punto es oportuno realizar un gráfico para lograr un mejor entendimiento de lo que es un dato atípico[1], el ejemplo muestra una observación atípica denotada como a en el gráfico (2.1).

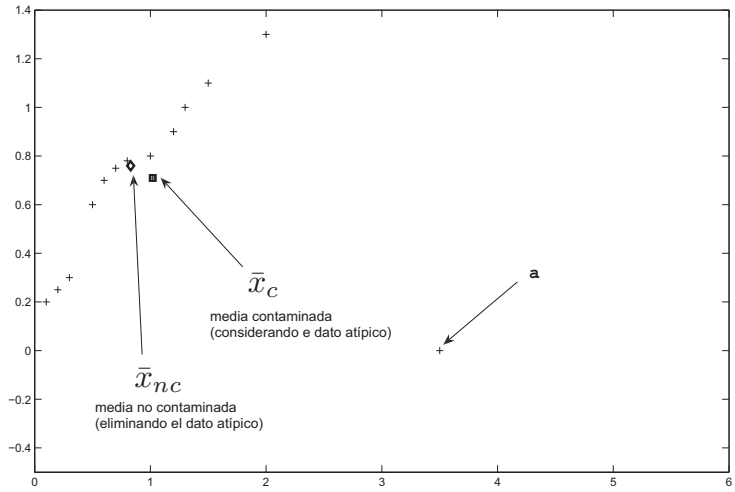

Figura 2.1: Ejemplo de una muestra que posee un dato atípico

Breve discusión

El gráfico (2.1) ayuda a comprender (gráficamente) la importancia de realizar un estudio de datos atípicos, pues al no hacerlo estaremos cometiendo un gravísimo error, ya que la presencia de un sólo dato atípico provoca grandes distorsiones al momento de realizar cualquier análisis. Del gráfico podemos decir:

- El vector de medias $\bar{\mathrm{x}}$ cambia de ubicación dependiendo de la ubicación del punto a en el plano, por ejemplo en el gráfico que estamos analizando tenemos un ligero desplazamiento del vector de medias $\bar{\mathrm{x}}$, se debe notar que para este ejemplo se han utilizado valores pequeños comparados con la cantidad de datos que se utilizan en la vida real, así que uno no se debe dejar engañar, pues si consideramos por un instante que se trata de miles de dólares, entonces es obvio que tal media contaminada influye totalmente para cualquier análisis posterior y más aún para una toma de decisiones.

- En este ejemplo gráfico únicamente se ha utilizado un dato atípico, imaginemos lo mismo pero con varios datos atípicos y no sólo con dos variables sino

[1]Este ejemplo gráfico explica la definición (gráficamente) de un dato atípico únicamente utilizando dos variables (x, y).

ahora con 3, 4 o más variables, hasta 3 variables podríamos (si el número de datos no es grande) visualizarlo pero a medida que el número de datos crece la visualización se vuelve imposible. He aquí la importancia de realizar un estudio multivariante de datos atípicos.

De aquí en adelante se discutirá los datos atípicos, pero desde el punto de vista matemático, es decir, se justificará matemáticamente cuando un dato es considerado como atípico, así como también se estudiará un algoritmo de búsqueda de datos atípicos multivariantes en un base de datos dada propuesta en [13].

Según la definición (2.1.2), un dato atípico será considerado como tal cuando su distancia respecto a la media sea grande, utilizando como medida de distancia la distancia euclídea, una observación será atípica en esta métrica si:

$$\mathrm{d_E}(\mathbf{x}_i, \bar{\mathbf{x}}) = \left[(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_i - \bar{\mathbf{x}})\right]^{1/2} \tag{2.1}$$

es grande. Para identificar observaciones atípicas podríamos hacer un histograma de estas distancias y ver si existen puntos mucho más alejados que los demás. Sin embargo, esta medida de distancia no es razonable cuando exista dependencia entre las observaciones. El problema es que la distancia euclídea no tiene en cuenta la estructura de correlación de los datos y una posibilidad mejor es estandarizar previamente los datos de forma multivariante.

La estandarización univariante utiliza la transformación:

$$\mathbf{y} = \mathbf{D}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}) \tag{2.2}$$

donde la matriz $\mathbf{D}^{-1/2}$ es cuadrada y diagonal con términos:

$$\mathbf{D}^{-1/2} = \begin{bmatrix} s_1^{-1} & 0 & \dots & 0 \\ 0 & s_2^{-1} & \dots & 0 \\ 0 & 0 & \dots & s_3^{-1} \end{bmatrix}$$

$\mathbf{y}$ tiene media cero y varianza unidad, en cuanto a la matriz de varianzas y covarianzas será la matriz de correlación de las variables primitivas.

Mientras que la estandarización multivariante utiliza la transformación:

$$\mathbf{y} = \mathbf{S}_x^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}) \tag{2.3}$$

con esta transformación la nueva variable tiene media cero y matriz de varianzas identidad.

La raíz cuadrada de la matriz $\mathbf{S}_x$ utilizada en (2.3) no es única, ya que si $\mathbf{S}_x = \mathbf{S}_x^{1/2}\mathbf{S}_x^{1/2'}$ también lo será $\mathbf{S}_x = \mathbf{S}_x^{*1/2}\mathbf{S}_x^{*1/2'}$. Una forma de definir la raíz cuadrada

de una forma única es exigir que la matriz $\mathbf{S}_x^{1/2}$ sea simétrica para ello se puede utilizar la descomposición de Cholesky que se encuentra con más detalle en [10].

De esta manera, los datos transformados tienen media cero y matriz de covarianza identidad y podemos buscar atípicos con la distancia euclídea, eliminando el problema de la correlación entre las variables, se define las variables estandarizadas multivariantemente por:

$$y = \mathrm{S}_x^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}) \tag{2.4}$$

La distancia euclídea al cuadrado entre una observación y_i y su media (cero), será:

$$\begin{aligned} \mathrm{d}_{\mathrm{E}}^2(y_i, 0) &= y_i' \, y_i \\ &= (\mathbf{x}_i - \bar{\mathbf{x}})' \, \mathrm{S}_x^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= \mathrm{d}_{\mathrm{M}}^2(\mathbf{x}_i, \bar{\mathbf{x}}) \end{aligned}$$

Y la distancia euclídea entre las variables incorrelacionadas equivale a la distancia de Mahalanobis entre las variables originales. Podríamos entonces identificar datos atípicos calculando la distancia de Mahalanobis para todo ellos y viendo si existe algún punto con una distancia mucho mayor que el resto.

2.2 Efecto de datos atípicos

Las consecuencias de una sola observación atípica pueden ser graves: distorsionar las medias, desviaciones típicas de las variables y destruir las relaciones existentes entre ellas.

Supongamos que en una muestra multivariante de tamaño n se introduce un valor atípico, a, que es un vector de falsas observaciones. Llamando $\bar{\mathbf{x}}$ y S al vector de medias y matriz de covarianzas sin el dato atípico y $\bar{\mathbf{x}}_{\mathsf{c}}$ y S_{c} al vector de medias y matriz de covarianzas de la muestra contaminada con este dato atípico, luego se tiene que:

$$\bar{\mathbf{x}}_{\mathsf{c}} = \bar{\mathbf{x}} + \frac{a - \bar{\mathbf{x}}}{n+1} \tag{2.5}$$

$$\mathrm{S}_{\mathsf{c}} = \frac{n}{n+1}\,\mathrm{S} + \frac{(a - \bar{\mathbf{x}})(a - \bar{\mathbf{x}})'}{n+1}\left(\frac{n}{(n+1)}\right) \tag{2.6}$$

Demostración. Para las demostraciones de (2.5) y (2.6), llamemos n al número total de observaciones que no son atípicos; con esto en mente:

1. Para (2.5) se tiene:

$$\begin{aligned}
\bar{\mathbf{x}}_{\mathrm{c}} &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i \\
&= \frac{1}{n+1}\Big(\sum_{i=1}^{n} x_i + a\Big) \\
&= \frac{1}{n+1}(n\bar{\mathbf{x}} + a) \\
&= \frac{1}{n+1}(n\bar{\mathbf{x}} + a + \bar{\mathbf{x}} - \bar{\mathbf{x}}) \\
&= \bar{\mathbf{x}} + \frac{a-\bar{\mathbf{x}}}{n+1}
\end{aligned}$$

2. Para (2.6) se tiene:

$$\begin{aligned}
\mathrm{S}_{\mathrm{c}} &= \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{\mathbf{x}}_{\mathrm{c}})(x_i - \bar{\mathbf{x}}_{\mathrm{c}})' \\
&= \frac{1}{n+1}\sum_{i=1}^{n}\Big(x_i - \bar{\mathbf{x}} + \frac{a-\bar{\mathbf{x}}}{n+1}\Big)\Big(x_i - \bar{\mathbf{x}} + \frac{a-\bar{\mathbf{x}}}{n+1}\Big)' \text{(Utilizando (2.5))} \\
&= \frac{1}{n+1}\sum_{i}^{n}(x_i - \bar{\mathbf{x}}_{\mathrm{c}})(x_i - \bar{\mathbf{x}}_{\mathrm{c}})' - (x_i - \bar{\mathbf{x}}_{\mathrm{c}})\big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)' - \big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)(x_i - \bar{\mathbf{x}}_{\mathrm{c}})' \\
&= +\frac{(a-\bar{\mathbf{x}})(a-\bar{\mathbf{x}})'}{(n+1)(n+1)}
\end{aligned}$$

Para concluir la demostración basta con demostrar que:

$$\sum_{i}^{n} -(x_i - \bar{\mathbf{x}}_{\mathrm{c}})\big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)' - \big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)(x_i - \bar{\mathbf{x}}_{\mathrm{c}})' = 0$$

en efecto:

$$\begin{aligned}
\sum_{i}^{n} -(x_i - \bar{\mathbf{x}}_{\mathrm{c}})\big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)' - \big(\frac{a-\bar{\mathbf{x}}}{n+1}\big)(x_i - \bar{\mathbf{x}}_{\mathrm{c}})' &= -2\Big(\frac{a-\bar{\mathbf{x}}}{n+1}\Big)\sum_{i}^{n}(x_i - \bar{\mathbf{x}}) \\
&= -2\Big(\frac{a-\bar{\mathbf{x}}}{n+1}\Big)(n\bar{\mathbf{x}} - n\bar{\mathbf{x}}) \\
&= 0
\end{aligned}$$

Luego:

$$\mathrm{S_c} = \frac{n}{n+1}\,\mathrm{S} + \frac{(a-\bar{\mathbf{x}})(a-\bar{\mathbf{x}})'}{n+1}\left(\frac{n}{(n+1)}\right)$$

□

Estas fórmulas indican que un solo dato atípico puede afectar mucho al vector de medias y a todas las varianzas y covarianzas entre las variables. El efecto del atípico depende de su tamaño y su posición, ya que los términos más afectados de la matriz dependen de la posición del atípico en el espacio. En general, si el tamaño del atípico es grande, lo que se supone grande, la media, varianza y covarianzas de las variables pueden estar muy distorsionadas.

Para analizar con más detalle la distorsión de los coeficientes de correlación, se considera el caso más simple de $p = 2$ y supongamos $\bar{\mathbf{x}} = 0$, $\mathrm{S} = \mathbf{I}$ y n no muy pequeño de manera que, para simplificar $\bar{\mathbf{x}}_\mathsf{c} \simeq \bar{\mathbf{x}} = 0$. Llamando $\mathrm{S}_\mathsf{c}^{ij}$ a los elementos de $\mathrm{S_c}$ y tomando $n \simeq n+1$, en (2.6) tendremos que:

$$\mathrm{S}_{ii}^{c} \simeq 1 + \frac{a_i^2}{n}, \quad i = 1, 2 \tag{2.7}$$

$$\mathrm{S}_{ij}^{c} \simeq \frac{a_1 a_2}{n}, \quad i \neq j \tag{2.8}$$

Con lo que el coeficiente de correlación entre las dos variables será:

$$r_\mathsf{c} = \frac{a_1\, a_2}{(n + a_1^2)^{1/2}\,(n + a_2^2)^{1/2}}$$

Esta expresión muestra que si las coordenadas del valor atípico a_1 y a_2 tienen el mismo signo y son grandes con relación a $\sqrt{n}$, el coeficiente de corrección tiende a uno, mientras que si tienen signos opuestos, el coeficiente tiende hacia menos uno. Vemos que la distorsión que produce el atípico depende no sólo de su tamaño sino también de su posición en el espacio.

Cuando existe más de un atípico en los datos, puede producirse el efecto conocido como *enmascaramiento*, que consiste en que observaciones atípicas similares se ocultan entre si.

2.3 Identificación de grupos atípicos método multivariante

El procedimiento para detectar grupos atípicos es eliminar de la muestra todos los puntos sospechosos, de manera que evitemos el enmascaramiento (discutido en la sección anterior) y podamos calcular el vector de medias y la matriz de covarianzas sin distorsiones. El primer paso para identificar las observaciones sospechosas es detectar aquellas que lo sean claramente respecto a una variable. Para ello, podemos utilizar el histograma o los diagramas de caja. Una regla simple y automática es considerar sospechosas aquellas observaciones tales que:

$$\frac{|x^i - \text{med}(x)|}{\text{MEDA}(x)} > 4,5 \tag{2.9}$$

Donde med(x) es la mediana de las observaciones, que es un estimador robusto del centro de los datos, y MEDA(x) es la mediana de las desviaciones absolutas $|x_i - \text{med}(x)|$, que es una medida robusta de dispersión. Este método puede verse como una estandarización robusta de los datos.

Esta detección univariante no identificará muchos atípicos multivariantes. Por ejemplo consideremos el caso en que tomamos una muestra con 4 variables y únicamente analizamos dos de ellas como en el gráfico 2.2, el punto $(-1, 0.58)$ (marcado con **a** es claramente atípico, pero no aparecerá como tal en los análisis univariantes.

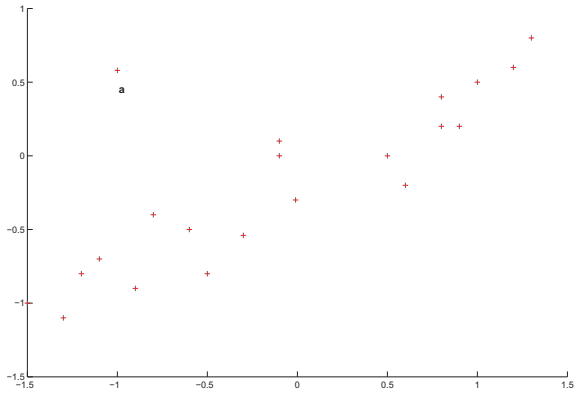

Figura 2.2: Ejemplo univariante de dato atípico

Con frecuencia los atípicos multivariantes corresponde a situaciones con efectos pequeños sobre todas las variables, como un error sistemático de observación en todas, en lugar de un efecto importante sobre una variable. Si el número de variables no es muy grande, los diagramas de dispersión o matrices de de gráficos

de dispersión[2]pueden ayudar visualmente a determinar datos atípicos en dos dimensiones. Para dimensiones mayores no es recomendable utilizar la distancia de Mahalanobis, ya que si existen grupos de atípicos, pueden distorsionar la estimación del centro y la dispersión de los datos enmascarando los atípicos y quizás señalando como atípicos a puntos que no lo son.

Para evitar este problema podemos buscar proyecciones de los datos que muestren las observaciones atípicas. Cualquier observación atípica multivariante debe aparecer como tal al menos en una dirección de proyección. En base a esta idea, Stahel (1981) y Donoho (1982) propusieron generar muchas direcciones al azar, proyectar los puntos sobre estas direcciones y marcar como datos atípicos a aquellas observaciones que aparecen como extremas en estas proyecciones. Para generar direcciones al azar pueden tomarse muestras al azar de p puntos, calcular el plano que las contiene y tomar como dirección el vector ortogonal al plano.

Este método funciona bien con pocas variables, pero al aumentar la dimensión del problema el número de direcciones que necesitamos generar para cubrir razonablemente el espacio y tener garantías de éxito aumenta exponencialmente. Una solución propuesta por Peña et al. es proyectar los datos sobre ciertas direcciones específicas, pero estas direcciones deben ser escogidas de tal manera que tengan alta probabilidad de mostrar atípicos cuando existan. Para el caso univariante el coeficiente de kurtosis aumenta cuando existe una pequeña proporción de atípicos, lo que sugiere investigar las direcciones donde los puntos proyectados tenga máxima kurtosis univariante. Por otro lado, un grupo grande de atípicos puede producir bimodalidad y baja kurtosis, por lo que conviene también explorar las direcciones donde los puntos proyectados tengan mínima kurtosis.

La idea del procedimiento es buscar p direcciones ortogonales de máxima kurtosis y p direcciones ortogonales de mínima kurtosis, eliminar provisionalmente los datos extremos en estas direcciones, calcular la media y la matriz de covarianzas con los datos no sospechosos y después identificar los datos atípicos como aquellos que son extremos con la distancia de Mahalanobis calculada con las estimaciones no contaminadas (sin datos atípicos). En la sección (2.2) se describe el método para hallar datos atípicos en muestras multivariantes.

[2]Véase la figura (6.1) en la página 46.

2.4 Algoritmo de búsqueda de datos atípicos

Las técnicas conocidas como búsqueda de proyecciones (*projection pursuit*) se utilizan para construir un indicador a partir de variables multivariantes que utilizan criterios de proyección y búsqueda de direcciones donde cierto criterio se maximiza o minimiza.

La idea de Peña et al es justamente encontrar direcciones en las que el coeficiente de kurtosis se maximiza o minimiza para lo cual el algoritmo (2.1) forma parte del algoritmo (2.2), el primero es como sigue:

1: Escoger la dimensión del espacio sobre el que se va a proyectar (normalmente 2) y el criterio que se desea maximizar.

2: Encontrar la dirección que maximiza el criterio. Si no es posible obtener la dirección de forma analítica, se hace de manera aproximada. por ejemplo, se selecciona un número grande de direcciones $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$, se evalúa el criterio en cada una y se toma la dirección de este conjunto donde el criterio tome el valor máximo.

3: Encontrar una dirección ortogonal a la primera que maximice el criterio. Esto puede hacerse, por ejemplo, proyectando los datos sobre el espacio ortogonal a la primera dirección, $\mathbf{a}$, lo que se realiza mediante $\mathbf{Y} = (\mathbf{I} - \mathbf{a}\mathbf{a}')\,\mathbf{X}$ y aplicar el algoritmo del punto 2 a los nuevos datos $\mathbf{Y}$.

4: Representar los datos sobre el plano definido por las direcciones de proyección.

Algoritmo 2.1: Algoritmo de búsqueda de proyecciones

Para determinar grupos de datos atípicos, en muestras multivariantes la idea de fondo es analizar la función de kurtosis, y en base a ella determinar si existen grupos de datos atípicos o no. Es decir se busca aquellos puntos para las cuales la función de kurtosis incrementa su valor o disminuye (la idea detrás de todo esto es maximizar y minimizar dicha función).

Entrada: Matriz de datos $\mathbf{x}$, α porcentaje máximo de extracción de datos atípicos.

Salida: Vector $\vec{a}$ cuya posición i–ésima indica un dato atípico

1: $y_i = \mathrm{S}^{-1/2}(\mathbf{x} - \bar{\mathrm{x}}), \quad i = 1, \ldots, n$

2: Calcular p direcciones ortogonales y p proyecciones que maximizan el coeficiente de kurtosis.

a. Asignar $y_i^{(1)} = y_i$, y $j = 1$.

b. La dirección que maximiza el coeficiente de kurtosis se lo obtiene como solución de:

$$\begin{cases} \mathbf{d}_j = \arg\max\limits_{\mathbf{d}} \quad \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(\mathbf{d}'\ y_i^{(j)}\right)^4 \\ s.a. \\ \qquad \mathbf{d}'\,\mathbf{d} = 1 \end{cases} \tag{2.10}$$

c. Los puntos de la muestra son proyectados sobre un subespacio de dimensión $p - j$, ortogonal a la dirección $\mathbf{d}_j$. Se define:

$$v_j = \mathbf{d}_j - \mathrm{e}_1, \qquad Q_j = \begin{cases} \mathbf{I} - \frac{v_j\, v_j'}{v_j'\ \mathbf{d}_j}, & \text{si} \quad v_j'\, d_j \neq 0 \\ \mathbf{I}, & \text{caso contrario.} \end{cases}$$

Donde $\mathrm{e}_1 = (1, 0, 0, \ldots)$. Con la matriz resultante Q_j se define:

$$u_i^{(j)} \equiv \begin{pmatrix} z_i^{(j)} \\ \\ y_i^{(j+1)} \end{pmatrix} = Q_j\, y_i^{(j)}, \quad i = 1, \ldots, n$$

donde $z_i^{(j)}$ es la primera componente de $u_i^{(j)}$, t.q.: $z_i^{(j)} = d_j'\, y_i^{(j)}$ $(v.p.u.)$ (valores de proyección univariante), y $y_i^{(j+1)}$ corresponde a las componentes restantes $p - j$ componentes de $u_i^{(j)}$. Asignar $j = j + 1$, y si $j < p$ ir a 2b. Caso contrario $z_i^{(p)} = y_i^{(p)}$.

3: Calcular otro conjunto de direcciones p (ortogonales) y p proyecciones minimizando del coeficiente de kurtosis.

a. Asignar $y_i^{(p+1)} = y_i$, iteración $j = p + 1$.

b. Ejecutar 2.b. y 2.c, con las modificaciones respectivas.

4: Calcular:

$$r_i = \max_{1 \leq j \leq 2p} \frac{\left| z_i^{(j)} - \text{median}\left(z_i^{(j)}\right)\right|}{\text{MEDA}\left(z_i^{(j)}\right)} \tag{2.11}$$

5: Determinar si i es atípico.

a. Si $r_i > \beta_p$, i es sospechosos de ser atípico. Asignar $\bar{\mathbf{x}} = \mathbf{x} \setminus i$ Repetir el procedimiento con $\bar{\mathbf{x}}$. Iterar hasta que el número de observaciones atípicas sea menor igual a α.

b. Caso contrario el algoritmo termina.

6: Sea U el conjunto de observaciones sin atípicos.

$$\tilde{m} = \frac{1}{|U|}\sum_{i \in U} x_i, \quad \tilde{\mathrm{S}} = \frac{1}{|U| - 1}\sum_{i \in U}(x_i - \tilde{m})(x_i - \tilde{m})', \quad v_i = (x_i - \tilde{m})'\, \tilde{\mathrm{S}}^{-1}\, (x_i - \tilde{m}), \quad \forall i \notin U$$

a. Si $v_i < \mathcal{X}^2_{p,\,0.99}$, con $i \notin U$ incluirlas en U.

b. Caso contrario i es declaro como dato atípico.

Algoritmo 2.2: Algoritmo de búsqueda de datos atípicos

En la ecuación (2.11) r_i es comparado con β_p (tabla 2.1), en la que este último son valores obtenidos mediante simulación con muestras sin datos atípicos, consúltese [13].

p	5	10	20
β_p	4.1	6.9	10.8

Tabla 2.1: Valores simulados de β_p para el algoritmo (2.2)

Para dimensiones mayores se realiza una interpolación de log β_p linealmen te en log$p$.

Para el cálculo de la direcciones de máxima y mínima kurtosis se utiliza el siguiente algoritmo:

1: Seleccionar una dirección inicial $\bar{\mathbf{d}}_0$ t.q. $\|\bar{\mathbf{d}}_0\|^2 = 1$.

2: En la iteración $l+1$, calcular $\bar{\mathbf{d}}_{l+1}$, como el valor propio asociado con el valor propio más grandes (más pequeño) de $M(\bar{\mathbf{d}}_l)$.

3: Terminar cuando $\|\bar{\mathbf{d}}_{l+1} - \bar{\mathbf{d}}_l\| < \epsilon$, y $\mathbf{d}_j = \bar{\mathbf{d}}_{l+1}$

Algoritmo 2.3: Algoritmo para el cálculo de la dirección d del algoritmo 2.2

en el algoritmo (2.3), $M(\bar{\mathbf{d}}_l)$ proviene de aplicar a (2.10) las condiciones de optimalidad, así aplicándolas se tiene:

$$\begin{aligned} 4\sum_{i=1}^{n} \left(\mathbf{d}' y_i^{(j)} \right)^3 y_i^{(j)} - 2\lambda \mathbf{d} &= 0 \\ \mathbf{d}' \mathbf{d} &= 1 \end{aligned}$$

Multiplicando la primera ecuación por $\mathbf{d}$ y reemplazando la restricción, se obtiene el valor de λ. Luego la condición resultante es

$$\left(\sum_{i=1}^{n} (\mathbf{d}' y_i^{(j)})^2 y_i^{(j)} y_i^{(j)'} \right) \mathbf{d} = \sum_{i=1}^{n} (\mathbf{d}' y_i^{(j)})^4 \mathbf{d} \tag{2.12}$$

La ecuación (2.12) indica que el óptimo $\mathbf{d}$ será un vector propio de la matriz:

$$M(\mathbf{d}) \equiv \sum_{i=1}^{n} (\mathbf{d}' y_i^{(j)})^2 y_i^{(j)} y_i^{(j)'}$$

La sección $6.a$ del algoritmo (2.2) toma como referencia la distribución de variable escalar:

$$T^2 = (\mathbf{x} - \mu)' \hat{\mathbf{S}}^{-1} (\mathbf{x} - \mu) \tag{2.13}$$

donde:

$$\begin{aligned} \hat{\mathbf{S}} &:= \text{estimación de } \mathbf{V} \text{ de } \mathbf{x} \sim N_p(\mu, \mathbf{V}) \\ \mu &:= \text{El vector de medias} \end{aligned}$$

La ecuación (2.13) representa la distancia de Mahalanobis entre una variable y su media poblacional, pero calculada con la matriz de varianzas estimada y se denomina T^2 de Hotelling con p y $n-1$ grados de libertad. En vista que $\hat{\mathbf{S}} \rightarrow \mathbf{V}$, T^2 converge a la distancia de Mahalanobis y la distribución de Hotelling a la distribución χ^2_p, consúltese [12].

Capítulo 3

ANÁLISIS DE CONGLOMERADOS

El análisis de conglomerados es una técnica para clasificar objetos o casos en grupos relativamente homogéneos llamados conglomerados o *clusters*. Los objetos de cada conglomerado son semejantes entre ellos y disímiles a los objetos de otro conglomerado. El análisis por conglomerado también se llama análisis de clasificación o taxonomía numérica.

Tanto el análisis por conglomerado como el análisis discriminante se relacionan con la clasificación. Sin embargo, el análisis discriminante requiere conocer de antemano el grupo a que pertenece cada objeto. En contraste, en el análisis por conglomerado no hay información a priori sobre el número de grupos de los objetos. Los conglomerados se desprenden de los datos, para luego ser caracterizados.

3.1 Procedimiento de un análisis conglomerado

El gráfico (3.1) presenta los pasos para realizar un análisis por conglomerados. El primer paso es plantear el problema de conglomerado, para lo cual se definen las variables en las que se basará el conglomerado. Luego debe elegirse una medida apropiada de la distancia. Esta medida determina cuán semejantes o diferentes son los objetos agrupados. La decisión del número de conglomerado depende del criterio del investigador, decisión importante ya que un valor demasiado bajo no permite obtener criterios pertinentes para la clasificación, mientras que un valor elevado no sería manejable desde punto de vista práctico. Por ende, el investigador debe evaluar la validez del proceso.

3.1.1 Planteamiento del problema

La parte más importante del análisis conglomerado es la elección de las variables en las que se va a basar el estudio. La inclusión de apenas una o dos variables irrelevantes puede distorsionar una solución de agrupamiento que de otra manera sería útil. Básicamente, el conjunto de variables elegidas debe describir la semejanza entre objetos en términos de que sean pertinentes para el problema de investigación de mercados. Las variables deben elegirse de acuerdo con las investigaciones pasadas, una teoría, o la consideración de las hipótesis que se comprueban.

Figura 3.1: Análisis por conglomerado

3.1.2 Selección de una medida de distancia o semejanza

Para los datos cuantitativos, la medida generalmente utilizada es la distancia euclidiana. Existen una multitud de otras medidas, como las de Minkowski y de Tchebychev. En caso de tener variables cualitativas, se puede determinar una distancia según los criterios del investigador, que atribuya una distancia menor entre elementos más “parecidos”. Dependiendo de la medida utilizada, las unidades en que se mide cada valor y la importancia de cada criterio, se puede efectuar previamente una homogeneización para suprimir la influencia de la variable en la solución del conglomerado. Es decir un cambio de escala, comúnmente para que tengan media de cero y desviación estándar de uno.

En este punto se ve la importancia de eliminar valores atípicos, pues la distancia seleccionada se verá distorsionada si en la base de estudio existen datos atípicos.

3.1.3 Selección de un procedimiento conglomerado

En la figura (3.2) se encuentra una clasificación de los procedimientos por conglomerado, que pueden ser jerárquico y no jerárquico. El conglomerado jerárquico se caracteriza por el desarrollo de una jerarquía o estructura de árbol (dendograma). Los métodos jerárquicos pueden ser por aglomeración o por división. El conglomerado por aglomeración comienza con cada objeto en un grupo separado. Los conglomerados se forman reuniendo objetos en grupos cada vez mayores. El proceso continúa hasta que todos los objetos son miembros de un único conglo-

merado. El conglomerado por división comienza con todos los objetos reunidos en un solo conglomerado, que se divide hasta que cada objeto está en un grupo separado.

Figura 3.2: Procedimiento por conglomerado

Los métodos por aglomeración son los que más se usan en la investigación de mercados. Comprenden los métodos de enlace, métodos de sumas de errores cuadrados o varianza y métodos de centroides. Los métodos de enlace abarcan los de enlace único, enlace completo y enlace promedio.

- **Método de enlace único.** Se basa en la regla de la mínima distancia o del vecino más cercano. Los primeros dos objetos agrupados son los que tienen la menor distancia entre ellos. Se identifica la siguiente distancia menor o

bien se conglomera un tercer objeto con los primeros dos o se forma un nuevo conglomerado de dos objetos. En cada etapa, la distancia entre dos conglomerados es la de sus puntos más cercanos. El método de enlace simple no funciona bien si los conglomerados están mal definidos.

- **Método de enlace completo.** Es semejante al de enlace único, salvo que se basa en la regla de la distancia máxima o el vecino más alejado. En el enlace completo, la distancia entre dos conglomerados se calcula como la distancia entre sus dos puntos más lejanos.

- **Método de enlace promedio.** Opera de la misma manera, pero la distancia entre dos conglomerados se define como el promedio de la distancia entre todos los pares de objetos, tomando cada elemento de los pares de un conglomerado.

- **Método de Ward.** Con los métodos de varianza se pretende formar conglomerados para reducir al mínimo la variación dentro de cada uno. Un método de varianza muy empleado es el procedimiento de Ward. Se calculan las medias de todas las variables de cada conglomerados. Enseguida se calcula la distancia cuadrada euclidiana de cada objeto a la media del conglomerado. Estas distancias se suman para todos los objetos. En cada etapa se combinan los dos conglomerados con el menor incremento en la suma general de cuadrados de las distancias.

- **Método de centroides.** La distancia entre dos conglomerado es la distancia entre sus centroides (que son las medias de las variables de un conglomerado). De todos los métodos jerárquicos, los de enlace promedio y Ward dan resultados mejores en general.

- El segundo tipo de procedimientos, los métodos de conglomerado no jerárquico, se llaman a menudo conglomerado de k medias.

- **Método del umbral secuencial.** Se elige un centro del conglomerado y se reúnen todos los objetos que se encuentren dentro de un umbral especificado con respecto al centro. Luego se elige otro centro (o semilla) y se repite el proceso con los puntos que no están agrupados. Cuando un objeto está conglomerado con una semilla, ya no se considera para un conglomerado con otras semillas.

- **Método del umbral paralelo.** Funciona de la misma manera, salvo que se escogen al mismo tiempo varios centros de conglomerado y los objetos dentro del umbral se agrupan con el centro más próximo.

- **Método de repartición óptima.** Difiere de los dos procedimientos de umbral en que los objetos pueden volver a asignarse a otros conglomerado para perfeccionar el criterio general, por ejemplo, la distancia promedio interna de cierto número de conglomerado.
- Dos desventajas importantes de los procedimientos no jerárquicos son que el número de conglomerado debe especificarse de antemano, y la selección de centros es arbitraria. Más aún, los resultados del conglomerado dependen de cómo se elijan los centros.

3.2 Elección del número de conglomerado

No existen reglas exactas y rápidas para definir un número óptimo de conglomerados. Se escogerá un valor a partir del cual se puedan caracterizar los conglomerados, manteniendo un tamaño relativo de los conglomerados que sea significativo.

3.3 Interpretación y perfiles del conglomerado

Interpretar y trazar perfiles de conglomerado se hará mediante los diferentes cálculos estadísticos básicos, los más representativos siendo el centroide, y la moda.

3.4 Evaluación de la confiabilidad y validez

Esta es una etapa imprescindible del análisis de conglomerados, puesto que se ha determinado por lo menos un valor empírico. Los procedimientos formales para evaluar la confiabilidad y validez de las soluciones de conglomerado son complejos y no se defienden del todo. Lo que se aplicará son cambios de distancia o de métodos de conglomerado, para determinar la estabilidad de las soluciones.

3.5 Análisis conglomerado en dos fases

Se utilizará el método de conglomerados en dos fases incorporado en el SPSS, una ventaja de este sobre los métodos tradicionales es el manejo tanto de variables continuas como categóricas, así como también la determinación del número

de conglomerados, la notación utilizada para este modelamiento es la siguiente, consúltese [17]:

Notación	Definición
K^A	Número total de variables continuas usadas.
K^B	Número de variables categóricas usadas.
L_K	Número de categorías para la k–ésima variable categórica.
R^k	Rango de la k-ésima variable continua.
N	Número total de datos.
N_K	Número de datos en el conglomerado k
$\widetilde{\sigma}_k^2$	Varianza estimada para la k-ésima variable continua, tomada sobre todos los datos.
$\widetilde{\sigma}_{Jk}^2$	Varianza estimada de la k-ésima variable continua, tomada en el conglomerado j.
N_{jkl}	Número de datos en el conglomerado j, para los cuales la k-ésima variable categórica toma la l-ésima categoría
$\mathrm{d}(j,s)$	Distancia entre los conglomerados j y s
(j,s)	Índice que representa al conglomerado que se obtiene juntando los conglomerados j y s.

Tabla 3.1: Notaciones análisis de conglomerados en dos etapas

Fue diseñado para clasificar datos del tipo: cualitativo y cuantitativo al mismo tiempo, usando una medida de log-verosimilitud para los primeros, y la distancia euclidiana para los cuantitativos.

El proceso se divide en dos etapas:

- Una etapa en la que se define un margen de pertinencia.
- A continuación se analizan los datos uno por uno para ver si se lo añade a un pre–conglomerado ya existente, o si este elemento va a formar un nuevo preconglomerado.

Este método es no jerárquico, uno puede notar también que depende del orden

en el que estén clasificados los datos, pero esta consideración pierde importancia a medida que el tamaño de la base de datos aumenta.

Una vez que este proceso termine, ya no se consideran los datos en su totalidad, sino que a su vez se vuelve a efectuar un proceso de conglomerado a los pre–conglomerado, asumiendo que todo elemento de la base de datos se identifica con su pre conglomerado. Para este proceso se utiliza un método jerárquico.

El programa también incluye algoritmos para determinar un número óptimo de conglomerados, mediante diferentes criterios de información.

Se define la distancia de log-verosimilitud (tanto para variables continuas como categóricas) de la siguiente manera, consúltese [17]:

$$\mathrm{d}_{j,s} = \xi_j + \xi_s + \xi_{<j,s>} \tag{3.1}$$

donde:

$$\begin{aligned} \xi_v &= N_v \left(\sum_{k=1}^{K^A} \frac{1}{2} \log(\hat{\sigma}^2 + \hat{\sigma}_{vk}^2) + \sum_{k=1}^{K^B} \hat{E}_{vk} \right) \\ \hat{E}_{vk} &= \sum_{l=1}^{l^k} \frac{N_{vkl}}{N_v} \log \frac{N_{vkl}}{N_v} \end{aligned}$$

3.5.1 Número de conglomerados

El programa SPSS tiene dos métodos para determinar un valor óptimo del número de conglomerados. En una primera etapa, se calcula el BIC (Criterio de Información Bayesiano) o el AIC (Criterio de Información de Akaike), dados por las ecuaciones:

$$\mathsf{BIC(J)} = -2 \sum_{j=1}^{J} \xi_j + m_j \log \mathsf{N} \tag{3.2}$$

$$\mathsf{AIC(J)} = -2 \sum_{j=1}^{J} \xi_j + 2m_j \tag{3.3}$$

donde:

$$m_j = J\{2K^A + \sum_{K=1}^{K^B} (L_k - 1)\}$$

El criterio Bayesiano se adapta mejor a datos con distribución exponencial, mientras que el de Akaike es indicado para datos con distribución normal. Este

último es muy preciso en caso de escoger las buenas variables, pero varía mucho si se añadieron algunos criterios para el proceso de conglomerado.

Se consideran las variaciones de esos criterios para dar una primera estimación del número de conglomerados, y luego se refina buscando el mayor incremento el la distancia mínima entre dos conglomerados en cada paso del proceso de conglomerados jerárquicos.

3.5.2 Validez de la solución por conglomerados en dos fases

El proceso de conglomerado exige teóricamente que las variables continuas estén distribuidas normalmente, las variables categóricas una distribución multinomial y que todas las variables que se usan sean independientes. Se han hecho muchas comprobaciones empíricas que indican que este procedimiento es en realidad bastante robusto frente a violaciones tanto del supuesto de independencia como de las distribuciones, pero aún así es preciso tener en cuenta hasta qué punto se cumplen estos supuestos.

Para la independencia de variables continuas, se estudiará las correlaciones bivariadas, para variables categóricas, se usan las tablas de contingencia. Para estudiar la independencia entre una variable continua y una categórica, el estudio de las medias basta. Para verificar la distribución de las variables continuas, hace falta graficarlas, y para la de las categóricas, efectuar una prueba de chi-cuadrado.

3.6 Interpretación de los resultados

El programa SPSS también tiene incluido una opción de gráficos por variables, que produce un gráfico para cada conglomerado. Las líneas verticales con guiones muestran los valores críticos para determinar la significancia de cada variable en la formación del conglomerado. Una variable es significativa si el estadístico T de Student excede la línea de guiones positiva o negativa. Las variables que resulten significativas contribuyen a la formación del conglomerado. Un valor negativo en la T indica que la variable toma valor en el conglomerado inferior a su media y un valor positivo indica lo contrario.

3.7 Verificación de la solución

También se obtienen intervalos de confianza al 95% para las medias de las variables cuantitativas en cada uno de los conglomerados, divididos por una línea que

indica la pertenencia o no a cada una de las clases de las variables categóricas.

Capítulo 4

ANÁLISIS DE SUPERVIVENCIA

Análisis de supervivencia son métodos estadísticos para estudiar la ocurrencia y el número de veces de ocurrencia de un evento. Estos métodos comúnmente son utilizados en el estudio de muertes en pacientes de hospitales. De hecho, estos modelos originalmente fueron desarrollados para este propósito, por eso el nombre de análisis de supervivencia para estos modelos, consúltese [2].

El análisis de supervivencia es extremadamente útil para estudiar diferentes clases de eventos como: decesos, fallas en equipos, terremotos, accidentes de tránsito, nacimientos, retiros, arrestos, etc. El análisis de supervivencia a través del tiempo ha sido adaptado en diferentes áreas y cada una de ellas ha preferido llamarlo de diferentes manera, así tenemos: análisis de eventos históricos en Sociología, análisis de duración, transición en Economía, pero el análisis en sí es el mismo realizado en cada una de las áreas mencionadas anteriormente.

El análisis de supervivencia fue diseñado para datos longitudinales sobre la ocurrencia de eventos. Por evento se define un cambio cualitativo que puede estar situado en el tiempo, por ejemplo una transición de estado de no estar casado a estar casado; para el análisis lo importante es determinar cuando el cambio ocurrió, esto es situar el evento en el tiempo. Utilizando regresión logística, la misma ignora la información respecto al tiempo de ocurrencia del evento.

4.1 Definición

Definición 4.1.1 (Supervivencia)**.** *Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable "tiempo hasta que ocurre un evento" y su dependencia de otras posibles variables explicativas.*

En el análisis de supervivencia la variable de interés es el tiempo T (en nuestro caso la antigüedad del cliente) hasta que ocurre un suceso. Sea f la función de densidad de probabilidad de T. Se define la función de supervivencia de la siguiente manera, consúltese [14]:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du \tag{4.1}$$

Si se define la función λ de Hazard por:

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(t < T < t + \Delta t \quad | \quad T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{4.2}$$

Esta función se interpreta generalmente como la velocidad a la cual la supervivencia decrece, o fuerza de mortalidad. Y es la probabilidad de que a un individuo que está siendo observado en el tiempo t le suceda el evento en ese momento.

Y luego se obtiene que

$$S(t) = e^{-\int_0^t \lambda(u) d(u)} \tag{4.3}$$

Demostración. Puesto que $f(u) - S'(u)$, esto permite tener $\lambda(u) = -\frac{d}{du} \log S(u)$, e integrando esta ecuación entre 0 y t se tiene:

$$\log S(u)|_0^t = -\int_0^y \lambda(u) du$$

El resulta se sigue debido a que $S(0) = 1$ □

Existen distintos modelos sobre probabilidad o riesgo de ocurrencia de un evento considerando la existencia de un conjunto de variables explicativas e independientes.

- Un aspecto central es la existencia de una variable temporal.
- Dos componentes estimados en los modelos: parámetro para el tiempo (ciertas particularidades y asunciones según modelos) y parámetros para las variables independientes.

4.2 Tipo de modelos

4.2.1 Modelos paramétricos

Los modelos paramétricos son una alternativa para estimar funciones de probabilidad para el tiempo de vida y realizar contrastes de hipótesis sobre las mismas; la parte fundamental de estos modelos consiste en asumir ciertas funciones de probabilidad para la variable tiempo de espera. Si esta asunción es razonable; además se realizaran pruebas de bondad de ajuste de dichas funciones con respecto a los datos para aceptarlas como válidas.

4.2.2 Modelos No Paramétricos

Los modelos no Paramétricos son una alternativa cuando los datos no siguen una distribución especifica (no asume ninguna función de probabilidad); esta distribución se la estima por máxima verosimilitud, es decir se basa en maximizar la función de verosimilitud de la muestra. Una muestra aleatoria de tamaño n, extraída de una población, estará formada por k $(k \leq n)$ tiempos $t_1 < t_2 < \ldots < t_k$ en los que se observan eventos. En cada tiempo t_i existen n_i "individuos en riesgo" (elementos de la muestra para los que el evento puede ocurrir, o que $T \geq t_i$) y se observan di eventos. Además en el intervalo $[t_i, t_{i+1}]$ se producen m_i pérdidas.

4.3 Modelo de riesgo proporcional de Cox

El objetivo del modelo de riesgo proporcional de Cox (propuesto por Cox en 1972) es un modelo de regresión para el riesgo, o la supervivencia, en función de variables "explicativas" que permita comparar dichas estimaciones, teniendo en cuenta el efecto de otras variables distintas de la que se utiliza para definir los grupos (Activo e Inactivo)

Nuestro interés es modelar no sólo la relación entre la tasa de supervivencia y el tiempo, sino también la posible relación con diferentes variables registradas para cada cliente. Se trata por tanto de calcular la tasa de mortalidad como una función del tiempo y de las variables pronóstico.

La idea fundamental es la misma que en cualquier modelo de regresión, el sistema más utilizado es el denominado de riesgos proporcionales o modelo de Cox, en el que la tasa de mortalidad se calcula como:

$$\lambda(t, x_1, x_2, \ldots, x_n) = \lambda_0(t) e^{b_1 x_1 + b_2 x_2 + \ldots + b_n x_n} = \lambda_0(t) e^{\beta X} \tag{4.4}$$

Donde t es el tiempo, $x_1, x_2, \ldots, x_n$ son las variables pronóstico.

Por tanto la función de supervivencia para el modelo de Cox es la siguiente:

$$S(t) = e^{-\int_0^t \lambda(u, x_1, x_2, \ldots, x_n) d(u)} \tag{4.5}$$

4.4 Estimación de los coeficientes y contrastes de hipótesis

Para estimar los coeficientes por el método de máxima verosimilitud es necesario construir la función de P verosimilitud. Para ello hay varios métodos, pero el primero históricamente y el más general es el debido a Cox y se basa en

la función de verosimilitud parcial y parte de que dado que el modelo no hace ninguna asunción sobre $h_0(t)$, la única contribución de los datos a la verosimilitud es en los tiempos en que se observan eventos.

Los cálculos necesarios para la estimación son muy largos, sobre todo cuando para algún tiempo hay más de un evento. Los paquetes estadísticos suelen usar una aproximación (aproximación de Peto) que es buena cuando, para cada tiempo, el número de eventos es pequeño comparado con los individuos en riesgo.

Teniendo en cuenta que los estimadores lo son por máxima verosimilitud, los contrastes de hipótesis sobre cada coeficiente se hacen usando su normalidad asintótica (prueba de Wald) y los contrastes de hipótesis sobre el modelo completo o sobre un conjunto de coeficientes con el logaritmo del cociente de verosimilitudes (exactamente igual que en regresión logística).

Es decir y resumiendo un poco, un intervalo de confianza al (1- α)% para el coeficiente β_i es:

$$\beta_i \pm Z_{\alpha/2} ee(\beta_i) \tag{4.6}$$

En regresión de Cox los estimadores de la asociación no son los coeficientes β_i sino los riesgos relativos, por lo tanto los intervalos de confianza que interesan calcular son los de los riesgos relativos. Dichos intervalos están dados por:

$$e^{\beta_i \pm Z_{\alpha/2} ee(\beta_i)} \tag{4.7}$$

El estadístico para el contraste:

H0: $\beta_i = a$

H1: $\beta_i \neq a$ siendo a una constante, es:

$$w = \frac{(\beta_i - a)}{ee(\beta_i)} \tag{4.8}$$

y la región crítica: $|z| > z/2_{\alpha^2}$ o equivalentemente:

$$w = \frac{(\beta_i - a)^2}{\text{Var}(\beta_i)} \tag{4.9}$$

Que se distribuye como una ji-cuadrado con 1 grado de libertad, y por tanto, la región crítica para el contraste es $W > \chi^2_\alpha$. A estos contrastes se les denominan contrastes de Wald.

Un contraste que interesa realizar es $\beta_1 = 0$; el que no se pueda rechazar esta hipótesis indica que el riesgo no depende de la variable X.

4.5 Estrategias de modelización

Debido a los dos objetivos distintos que un análisis de regresión puede tener es difícil establecer una estrategia general para encontrar el mejor modelo de regresión, es más, el mejor modelo significa cosas distintas con cada objetivo.

En un análisis predictivo el mejor modelo es el que produce predicciones más fiables para una nueva observación, mientras que en un análisis estimativo el mejor modelo es el que produce estimaciones más precisas para el coeficiente de la variable de interés.

En ambos casos se prefiere el modelo más sencillo posible (a este modo de seleccionar modelos se le denomina parsimonia), de modo que en un análisis estimativo, se puede excluir del modelo una variable que tenga un coeficiente significativamente distinto de cero y que su contribución a la predicción de la variable dependiente sea importante, porque no sea variable de confusión para la variable de interés (el coeficiente de dicha variable no cambia), en un análisis predictivo esa variable no se excluiría.

Sin embargo, hay una serie de pasos que deben realizarse siempre:

1. **Especificación del modelo máximo.**

2. **Especificación de un criterio de comparación de modelos y definición de una estrategia para realizarla.**

3. Evaluación de la fiabilidad del modelo.

4. **Especificación del modelo máximo.** Se trata de establecer todas las variables que van a ser consideradas. Recuérdese que el modelo saturado (el máximo que se puede considerar) tiene $n - 1$ variables pero que, en general, el modelo saturado no tiene interés y el modelo máximo deberá tener menos variables independientes que el modelo saturado (un criterio habitual es incluir como máximo una variable cada 10 eventos).

   El criterio para decidir qué variables forman el modelo máximo lo establece el investigador en función de sus objetivos y del conocimiento teórico que tenga sobre el problema, evidentemente cuanto menor sea el conocimiento previo mayor tenderá a ser el modelo máximo.

   Un modelo máximo grande minimiza la probabilidad de error tipo II o infraajuste, que en un análisis de regresión consiste en no considerar una variable que realmente tiene un coeficiente de regresión distinto de cero.

Un modelo máximo pequeño minimiza la probabilidad de error tipo I o sobreajuste (incluir en el modelo una variable independiente cuyo coeficiente de regresión realmente sea cero).

Debe tenerse en cuenta también que un sobreajuste, en general, no introduce sesgos en la estimación de los coeficientes (los coeficientes de las otras variables no cambian), mientras que un infraajuste puede producirlos, pero que un modelo máximo grande aumenta la probabilidad de problemas de colinealidad.

En el modelo máximo deben considerarse también los términos de interacción que se van a introducir (en un modelo estimativo sólo interesan interacciones entre la variable de interés y las otras)

5. **Comparación de modelos**. Debe establecerse cómo y con qué se comparan los modelos. Si bien hay varios estadísticos sugeridos para comparar modelos, el más frecuentemente usado es el logaritmo del cociente de verosimilitudes, recordando que cuando los dos modelos sólo difieren en una variable, el contraste con el logaritmo del cociente de verosimilitudes es equivalente al contraste de Wald, pero a veces interesa contrastar varias variables conjuntamente mejor que una a una (por ejemplo todos los términos no lineales) o, incluso, es necesario hacerlo (por ejemplo para variables indicadoras).

   Hay que hacer notar que en un análisis estimativo el criterio para incluir o excluir variables distintas a las de interés, es sobre todo los cambios en los coeficientes y no los cambios en la significación del modelo. Los distintos modelos a comparar se pueden construir de dos formas: por eliminación o hacia atrás ("**backward**") y por inclusión o hacia adelante ("**forward**").

   Con la primera estrategia, se ajusta el modelo máximo y se calcula el logaritmo del cociente de verosimilitudes para cada variable como si fuera la última introducida (que es equivalente al contraste de Wald para esa variable), se elige el menor de ellos y se contrasta con el nivel de significación elegido. Si es mayor o igual que el valor crítico se adopta este modelo como resultado del análisis y si es menor se elimina esa variable y se vuelve a repetir todo el proceso hasta que no se pueda eliminar ninguna variable.

   Con la estrategia hacia adelante, se empieza con un modelo de una variable, aquella que presente el mejor logaritmo del cociente de verosimilitudes. Se calcula el logaritmo del cociente de verosimilitudes para la inclusión de todas las demás, se elige el menor de ellos y se contrasta con el nivel de signifi-

cación elegido. Si es menor que el valor crítico, se para el proceso y se elige el modelo simple como mejor modelo, y si es mayor o igual que dicho valor crítico, esa variable se incluye en el modelo y se vuelve a calcular el logaritmo del cociente de verosimilitudes para la inclusión de cada una de todas las restantes, y así sucesivamente hasta que no se pueda incluir ninguna más.

Una modificación de esta última estrategia es la denominada **"stepwise"** que consiste en que, cada vez que con el criterio anterior se incluye una variable, se calculan los logaritmos del cociente de verosimilitudes de todas las incluidas hasta ese momento como si fueran las últimas y la variable con menor logaritmo del cociente de verosimilitudes no significativo, si la hubiera, se elimina. Se vuelven a calcular los logaritmos del cociente de verosimilitudes y se continua añadiendo y eliminando variables hasta que el modelo sea estable.

Las variaciones a estas estrategias consisten en que, con cualquiera de ellas, se puede contrastar varias variables en lugar de una sola y que, en aplicación del principio jerárquico, cuando se contrasta un término de interacción, el modelo debe incluir todos los términos de orden inferior y, si como resultado del contraste, dicho término permanece en el modelo, también ellos deben permanecer en el mismo, aunque no se pueda rechazar que los coeficientes correspondientes no son distintos de cero.

6. **Evaluación de la fiabilidad del modelo.** Una vez encontrado el mejor modelo hay que evaluar su fiabilidad, es decir, evaluar si se comporta igual en otras muestras extraídas de la misma población

4.6 Relación entre la supervivencia individual y de la muestra

Sea $S(t) = e^{-\int_0^t \lambda(U,x_1,x_2,...,x_n)d(u)}$ la función de supervivencia, reemplazando (4.4) en (4.5) tenemos:

$$S(t) = e^{-\int_0^t \lambda_0 e^{\beta X} d(u)} \tag{4.10}$$

Aplicando logaritmo natural a ambos lados de la ecuación anterior, se sigue que

$$\ln(S(t)) = e_{\beta X} \int_0^t \lambda_0 d(u) = e^{\beta X} \ln(S_0(t)) \tag{4.11}$$

Por tanto:

$$S(t) = S_0(t)^{e^{\beta X}} \tag{4.12}$$

Donde $S_0(t)$ es la supervivencia que se obtiene cuando los parámetros son nulos.

Dado que $0 \leq S_0(t) \leq 1$, si βX aumenta, $S(t)$ disminuye.

Capítulo 5

ANÁLISIS: LT Y LTV

Luego de realizar un análisis de conglomerados (Capítulo 3) y sobre cada conglomerado un análisis de supervivencia (Capítulo 4) los pasos finales son calcular para cada cliente y por cada conglomerado hallado el tiempo medio de vida o Lifetime y su valor actual neto Lifetime Value durante su Lifetime.

5.1 Lifetime

Se desea estimar el tiempo medio al que va a ocurrir el suceso o muerte (tiempo medio de muerte), consúltese [14]:

$$TMM = E(T) = \int_0^\infty T(u)f(u)du \tag{5.1}$$

Con una integración por partes, se obtiene que

$$TMM = \int_0^\infty S(u)du \tag{5.2}$$

Esto se cumple, suponiendo TMM finito, dado que $t \to tf(t)$ entonces es integrable en intervalos no acotados superiormente, y se deduce que cuando $t >> 1$, $tf(t) = \mathcal{O}\left(\frac{1}{t}\right)$ resultando en $f(t) = \mathcal{O}\left(\frac{1}{t^2}\right)$,$S(t) = \mathcal{O}\left(\frac{1}{t}\right)$ y finalmente, $\lim_{t\to\infty} tf(t) = 0$ Para un cliente i , de antigüedad α_i , nos interesaremos en

$$TMM_i = E(T|T > \alpha_i) = \alpha_i + \frac{\int_0^\infty S(u)du}{S(\alpha_i)} \tag{5.3}$$

Es decir:

$$TMM_i = E(T|T > \alpha_i) = \alpha_i + \frac{\int_0^\infty S_0(\alpha_i)^{e^{\beta X}}du}{S_i(\alpha_i)} \tag{5.4}$$

Este valor nos da una previsión acerca del tiempo que le queda al cliente por "vivir" dentro de la empresa. Para calcular la parte integral, se aproxima por alguna función con la cual el cálculo integral será realizable.

En este tipo de casos se puede obtener una idea del tiempo que un cliente va a permanecer en el banco. Cabe destacar que los coeficientes r^2de estas regresiones son muy importantes y tienen que estar cercanos a 1 para que el modelo corresponda a la realidad. El error que se comete al elevar a $e^{\beta X}$, y luego al integrar, es pequeño dado que la media de los βX_i vale 0 (esto es un requerimiento para definir $S_0(t)$, y menor aún si su desviación estándar es pequeña).

5.2 Ecuación del Life Time (LT)

Finalmente llamaremos tiempo de vida (lifetime) y notaremos LT, al valor esperado de tiempo restante de vida como cliente. Para el cálculo del **LT** se utilizará la ecuación:

$$LT_i = \frac{\int_{\alpha_i}^{\infty} S_0(u)^{e^{\beta x_i}} d\,u}{S_i(\alpha_i)} \tag{5.5}$$

donde:

$$\begin{aligned} LT_i &:= \text{Lifetime para el cliente } i \\ \alpha_i &:= \text{antigüedad del cliente } i \text{ (en meses)} \\ S_0 &:= \text{función de supervivencia (por cluster)} \\ \beta x_i &:= \text{Dado por el SPSS, denotado por XBE_1} \\ S_i(\alpha_i) &:= \text{Dado por el SPSS, denotado por SUR_1} \end{aligned}$$

Para la aproximación de S_0 se utiliza funciones del tipo:

$$S(x) = \begin{cases} \left(\dfrac{\sin(\frac{x}{a}) - \frac{x}{a} + 2\,\pi}{2\,\pi}\right)^b, & \text{si } x \in [0,\, 2\,\pi\,a] \\ 0 & \text{si } x > 2\,\pi\,a \end{cases} \tag{5.6}$$

La ecuación (5.6), satisface:

- $S(0) = 1$
- $S'(0) = 0$
- $S(2\,\pi\,a) = 0$
- $S'(2\,\pi\,a) = 0$

Que son las características de una función de supervivencia, curva que no satisfaga con estas características no puede ser considerada como función de supervivencia.

Observación 5.2.1. *La aproximación de S_0 es de vital importancia, ya que la misma, si bien es cierto se la podría aproximar con una curva del tipo cuadrática, y de hecho se tienen aproximaciones relativamente altas, esta última no cumple con las características antes mencionadas.*

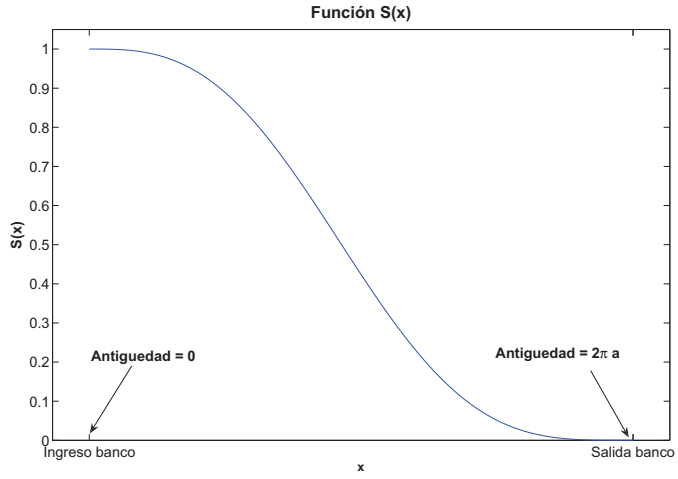

Figura 5.1: Función S_0 dada por (5.5)

Entrada: $\alpha_i, S_0^j, BX_i(\text{XBE_1}), S_i(\alpha_i)(\text{SUR_1})$

Salida: LT_i

1: **para** $j = 1$ **to** nc **hacer**

2: **para** $i = 1$ **to** ncc_j **hacer**

3:
$$LT_i = \frac{\int_{\alpha_i}^{2\,\pi\,a} S_0^j(u)^{\mathrm{E}^{BX_i}}\, d\,u}{S_i(\alpha_i)}$$

4: **fin para**

5: **fin para**

Algoritmo 5.1: Cálculo de LT por cliente

Gráficamente S_0 dada por la ecuación (5.5) tiene la forma:

Donde nc es el número de conglomerados y ncc_j el número de clientes en el conglomerado nc.

Ejemplo 5.2.1. *A continuación se muestra un ejemplo de este proceso*[1] *de manera muy general.*

Como ya mencionó anteriormente después de realizar un análisis de conglomerados a la base de datos, se tomará en consideración para este ejemplo un mercado dado después de dicho proceso.

Para este ejemplo primero se presenta el ajuste de la función S_0 con los resultados obtenidos del SPSS, el ajuste para esta curva se la presenta en el gráfico (5.2).

[1]Este es un ejemplo real, tomado de la base de datos.

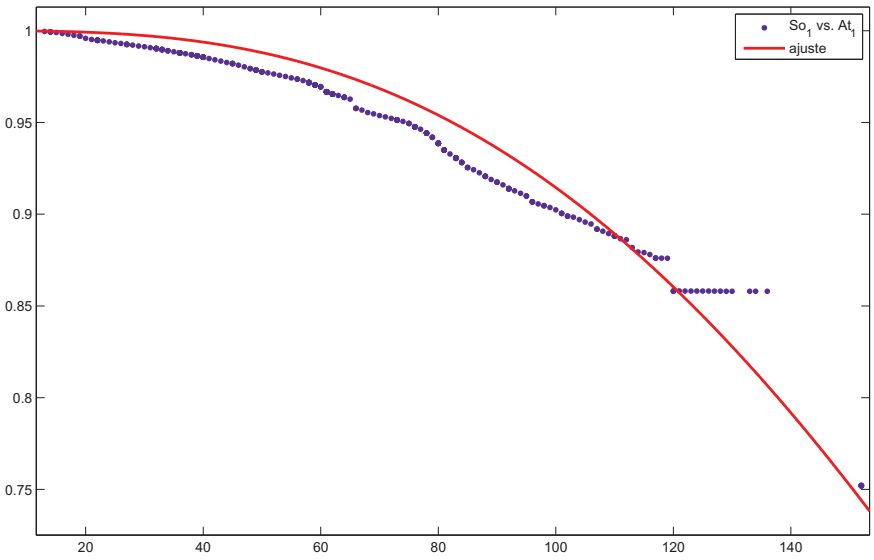

Figura 5.2: Ejemplo: Ajuste la función S_0

Para el ajuste de la función S_0 se utilizó la ecuación (5.5) dada en la página (38), en el gráfico (5.3) muestra, no sólo el ajuste de la función S_0 con su respectivos coeficientes y sus intervalos de confianza al 95%, sino también se determina el punto en el cual se espera (probabilísticamente) que un determinado cliente deserte de la institución bancaria.

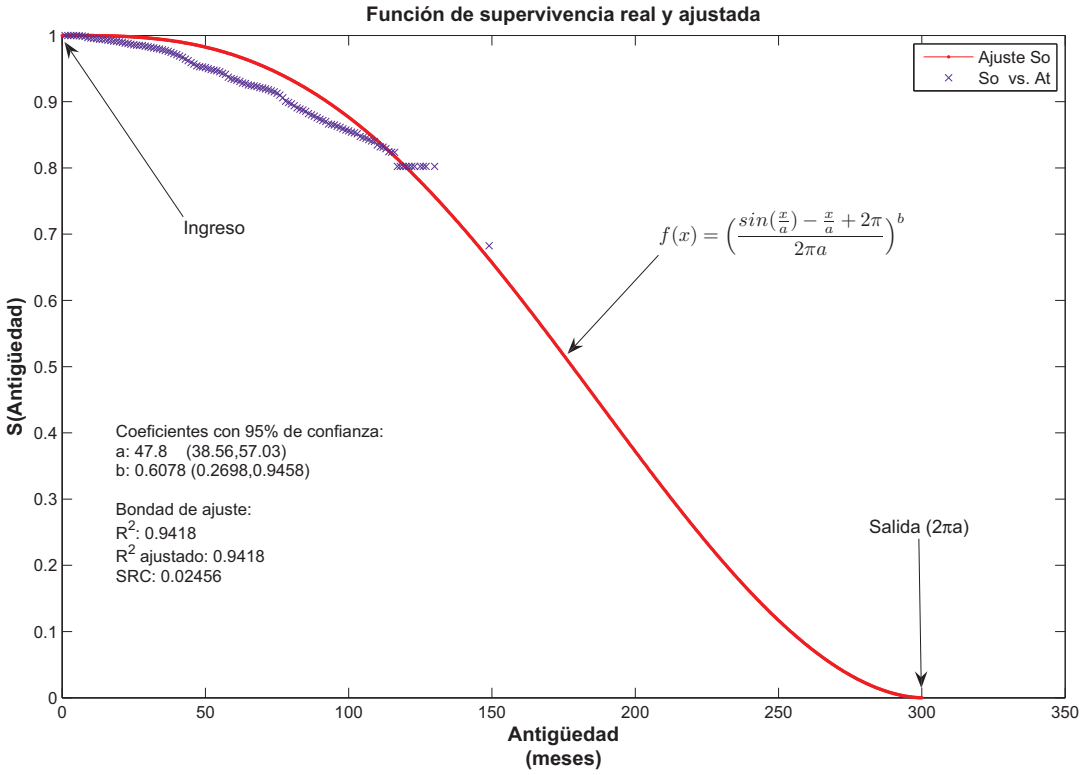

Figura 5.3: Ejemplo: Función S_0 extendida en el eje de antigüedad

Interpretación del gráfico (5.3)

El gráfico (5.3) nos dice que, un cliente que apenas ingresa a la institución bancaria su probabilidad de salir del mismo es nula (o dicho en otras palabras la probabilidad de permanencia en el Banco es del 100%, probabilidad 1 en el gráfico), y a medida que avanza su trayectoria en el Banco su probabilidad de salida se incrementa (su probabilidad de permanencia decrece).

Adicionalmente gráficamente podemos determinar (probabilísticamente) la antigüedad en la cual un determinado cliente salga del Banco.

Observación 5.2.2. *El ajuste para la función S_0 dada por la ecuación (5.5) ofrece para cada ajuste de aproximación un 97.5% en promedio para todos los ajustes realizadas en cada uno de los mercados analizados.*

5.3 Life Time Value

Valor de tiempo de vida (LTV): Valor esperado de un prospecto o cliente durante un período de tiempo dado, descontado a dólares actuales. También se conoce como CLV (Customer Life Value) y tiene las siguientes características:

- El cálculo de LTV es complejo, requiere de la estimación de diversos componentes según la aplicación:
- Ejemplo en marketing:
  - Duración esperada de la relación con el cliente
  - Período de tiempo: medida del incremento de LTV
  - Ingresos: Valor presente de los ingresos producidos por un producto o servicio
  - Costos diversos
  - Tasa de retención
  - Factor de riesgo
  - Ganancias incrementales (Cross-sell/Upsell, etc.)

Un modelo de LTV intenta predecir el valor total de un cliente (persona o negocio) durante una longitud de tiempo predeterminado. De manera similar al NPV, se calcula sobre un cierto número de años y se descuenta a dólares actuales. Los métodos para calcular el valor de tiempo de vida también varían a través de productos e industrias. A medida que los mercados se achican y la competencia aumenta, las compañías buscan oportunidades para obtener ganancias de su base de clientes. Como resultado, muchas compañías expanden sus ofertas de productos y/o servicios en un esfuerzo por hacer cross-sell y up-sell a sus clientes existentes. Este enfoque crea la necesidad de ir más allá del valor neto actual de un producto, a un enfoque de valor de tiempo de vida del cliente (LTV). Esta valuación permite a las compañías asignar recursos en base al valor potencial del cliente. Las mediciones de LTV son útiles para adquirir clientes, manejar sus relaciones con los mismos e incluso cuantificar la salud financiera de largo plazo de una empresa en base a la calidad de su portfolio de clientes. Una vez que se asigna un LTV a cada cliente, la base de clientes puede segmentarse de diversas maneras. Con esta información, una compañía puede tomar acciones o evitarlas en base al beneficio a largo plazo para la compañía. Para el cálculo del LTV se necesita conocer:

Para el cálculo de LTV, se utilizará la ecuación (5.3), la cual tiene como valor agregado haber aumentado un coeficiente para evitar los casos en los que algunas clientes coincidan con el LT de un cliente que actualmente está generando una alta rentabilidad.

$$LTV_i^{c_p}(LT_i) = \frac{R_i}{\overline{R}_{\xi_i}^{C_p}} \sum_{k=\xi_i+1}^{\xi_i+LT_i} \frac{\overline{R}_k^{C_p}}{(1+j)^{k-\xi_i}} \tag{5.7}$$

En la ecuación () el término $\frac{R_i}{\overline{R}_{\xi_i}^{C_p}}$ que puede ser visto como un factor de corrección o una ponderación por cada cliente, esto se debe entender de la siguiente manera: supongamos que existen 100 clientes y que todos ellos tienen la edad de 30 años, no todos generan la misma rentabilidad, existirán algunos clientes que generan más rentabilidad que otros y al colocar este factor se da una diferenciación cliente por cliente pero dentro de su conglomerado que son clientes con características homogéneas a él.

La expresión $\overline{R}_k^{C_p}$ se refiere a la rentabilidad promedio de los clientes con la edad k en el cluster C_p dentro del conglomerado analizado.

De manera similar la expresión $\overline{R}_{\xi_i}^{C_p}$, se refiere al promedio de rentabilidad de todos los clientes con la edad ξ_i, dentro del cluster C_p, en los conglomerados analizados.

En el algoritmo (5.2) se detallan los pasos a seguir para el cálculo del LTV por cada cliente y dentro de cada conglomerado de su pertenencia.

donde:

η_k $=$ número de conglomerados

τ $=$ 14.33%

LT_i $:=$ Lifetime del cliente i en años

ncc_j $:=$ número de clientes en el conglomerado j

ξ_i $:=$ edad del cliente i en años

$\overline{R}_{\xi_i}^{C_j}$ $:=$ promedio de rentabilidad de los clientes con edad ξ_i en el cluster C_j

$\overline{\mathfrak{R}}_{\xi_i}^{C_j}$ $:=$ promedio de rentabilidad de los clientes con edad ξ_i en el cluster C_j aproximada,

Observación 5.3.1. *Para los casos en los que no se tenga la rentabilidad promedio de una cierta edad dentro de un determinado nicho, se tiene nuevamente un ajuste de curvas, en este caso para la rentabilidad promedio por edad y por nicho, así que, como en el caso del LT de la sección (5.1), se tendrán nuevamente curvas que aproximen a la rentabilidad promedio por edad y por conglomerado.*

Entrada: $R_i,\ \tau,\ LT_i,\ \xi_i$

Salida: LTV_i

1: **para** $j = 1$ **to** η_k **hacer**

2: **para** $i = 1$ **to** ncc_j **hacer**

3: **Determinación de** $\overline{R}_{\psi}^{C^j}$

4: **para** $\eta = 1$ **to** LT_i **hacer**

5: $\psi \leftarrow \xi_i + \eta$

6: **si** $\overline{R}_{\psi}^{C^j} \neq \emptyset$ **entonces**

7: $\overline{R}_{\psi}^{C^j} \leftarrow \overline{R}_{\psi}^{C^j}$

8: **caso contrario**

9: $\overline{R}_{\psi}^{C^j} \leftarrow \overline{\mathfrak{R}}_{\psi}^{C^j}$

10: **fin si**

11: **fin para**

12:

$$LTV_i^j(LT_i) = \frac{R_i}{\overline{R}_{\xi_i}^{C_j}} \sum_{\psi=\xi_i+1}^{\xi_i+LT_i} \frac{\overline{R}_{\psi}^{C_j}}{(1+\tau)^{\psi-\xi_i}}$$

13: donde la asignación de $\overline{R}_{\psi}^{C^j}$ se realiza en los pasos (4) . . . (11)

14: **fin para**

15: **fin para**

Algoritmo 5.2: LTV

Capítulo 6

SCORE COMO FUNCIÓN DEL LTV

Finalmente sobre cada conglomerado hallado se calcula un *score* utilizando como variable principal el LTV de cada cliente, el valor calculado sobre cada cliente refleja la rentabilidad que un cliente va a dejar al Banco durante su tiempo de permanencia como cliente del mismo dado por su LT, valor que se calcula previo a un análisis de supervivencia y un análisis de conglomerados.

6.1 Criterio de scoring para los clientes

El score que se asigna a cada cliente viene dad por el cuartil al cual pertenezca su LTV, de esta manera en cada conglomerado hallado se tiene 4 grupos de clientes diferenciados por su nivel de LTV alto, medio, pequeño y bajo.

Los resultados obtenidos de esta propuesta de segmentación se detallan en la siguiente sección.

6.2 Resultados numéricos del modelo propuesto

El modelo de segmentación propuesto sigue los siguientes pasos:

1. Recolección de datos en un período determinado.
2. Análisis y calidad de la Información: depuración de datos atípicos.
3. Análisis de conglomerados.
4. Cálculo de Supervivencia sobre cada conglomerado.
5. En cada conglomerado se calcula el LT de cada cliente.
6. En cada conglomerado se calcula el LTV de cada cliente.
7. Con el LTV calculado de cada cliente se asigna dentro de cada conglomerado un score dependiendo del cuartil al cual pertenezca su LTV.

En cuanto a la recolección de datos se tomo clientes que en el período 2007 generaron rentabilidad positiva.

6.2.1 Depuración de datos atípicos

La base de datos sobre la cual se trabajará de aquí en adelante[1] es de $155\,995$ registros, en la que cada registro contiene la información de cada cliente, sobre las variables:

- edad
- suma rentabilidad en el ultimo año
- Total transacciones realizadas: en cajero, ventanilla tanto las de crédito como las de débito. Para cajero se tomó sólo las de débito.

Utilizando la matriz de gráficos de dispersión, podemos realizar rápidamente un análisis para saber si existe la presencia de datos atípicos destacados.

La figura (6.1) nos revela la existencia de datos atípicos muy destacados en prácticamente cada gráfico de dispersión.

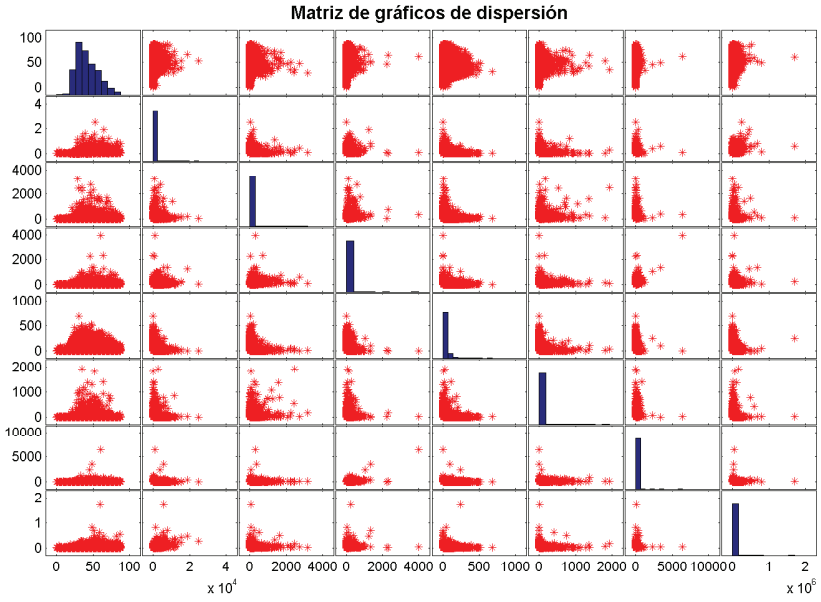

Figura 6.1: Matriz de gráfico de dispersión para la base de datos a depurar de datos atípicos

se corre el algoritmo 2.2 obteniéndose como resultado a $8,941$datos atípicos correspondiente al 5.73%,estos datos son excluidos del resto del análisis y por

[1]Por motivos de confidencialidad en la misma no consta información delicada para el cliente como cédula, nombre, dirección número de teléfono, etc., así como también no representa el total de clientes que posee dicha institución.

sí sólo corresponden a clientes especiales. La matriz de datos resultante de las variables antes mencionadas permiten que la matriz de varianzas y covarianzas sea no singular y por consiguiente permite calcular no sólo su inversa sino también su raíz cuadrada, para la estandarización multivariante, mediante el método de Cholesky usado en descomposición espectral de matrices, consúltese [12].

6.2.2 Análisis de conglomerados

Después de haber detectado datos atípicos de cada uno de los archivos, se procedió a clasificar a los clientes de acuerdo a cada una de las características que presentan estos en el Banco (demográficas y transaccionales).

El método de conglomerados en dos fases que incorpora el paquete estadístico SPSS resulta el más adecuado para nuestro análisis por distintas razones. En primer lugar, el tipo de variables incluidas, que son una combinación de variables continuas, ordinales y nominales. En segundo lugar, el elevado tamaño de la muestra, que en los algoritmos de clusters jerárquicos puede conducir a una sobre estimación del número de clusters. Por último, este método permite identificar y excluir casos extremos (outliers) que podrían interferir en la determinación de los centroides de los cluster.

Para el análisis de Cluster se introdujo las variables que constan en la tabla (6.1) **Variables**

6.2.3 Resultados del Análisis de Cluster

En la tabla (6.2) mediante el criterio BIC (Bayesian Information Criteria, Criterio bayesiano de información), se observa que el valor del criterio Bayesiano tiene valores decrecientes, entonces se toma como cluster aquellos cuyo valores sean los más altos, para este caso los clusters más significativos son 4.

En la tabla (6.3) podemos observar el número de cluster generados, el número de clientes y el porcentaje que representan en la base, un resumen de esta tabla se presenta a continuación en el gráfico (6.2).

En el gráfico (6.2) se observa el porcentaje de clientes por cada cluster.

Observación 6.2.1. *Del gráfico (6.2) se observa:*

- *Para el cluster* 1 *hay* 26.83% *de clientes.*
- *Para el cluster* 2 *hay* 34.89% *de clientes.*
- *Para el cluster* 3 *hay* 13.76% *de clientes.*

Notación	**Definición**
Provincia	Variable nominal
Estado Civil	Variable nominal
Género	Variable nominal
Ciclo de antigüedad	Variable ordinal
Código de negocio	Variable nominal
Ciclo de vida	Variable ordinal
Numero de transacciones de crédito que el cliente realizó por el canal VENTANILLA	Variable continua
Número de transacciones de débito que el cliente realizó por el canal VENTANILLA	Variable continua
Numero de transacciones de crédito que el cliente realizó por el canal ATM	Variable continua

Tabla 6.1: Variables a incluir en el análisis por conglomerados

- *Para el cluster* 4 *hay* 24.52% *de clientes.*

En la tabla (6.4) se observan las medias y desviaciones estándar de cada una de las variables continuas que se incluyeron en el estudio de cluster.

En el gráfico (6.3) se observa para cada cluster los diferentes ciclos de vida para los clientes institución bancaria.

Observación 6.2.2. *Del gráfico (6.3) se concluye lo siguiente:*

- ***Para el conglomerado 1***
  Este cluster está conformado por clientes dentro del ciclo de vida Maduros (personas con edades comprendidas entre 41 y 50 años), así como también clientes dentro del ciclo de vida (personas con edades comprendidas entre 31 y 40 años).

- ***Para el conglomerado 2***
  Este cluster está conformado por clientes dentro del ciclo de vida Jóvenes (personas con edades comprendidas entre 21 y 30 años).

- ***Para el conglomerado 3***
  Este cluster está conformado tanto por clientes dentro del ciclo de vida Adultos (personas con edades comprendidas entre 31 y 40 años), así como también clientes dentro

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	3431590.953			
2	3070875.822	-360715.131	1.000	1.425
3	2817912.839	-252962.982	.701	1.306
4	2624359.965	-193552.874	.537	3.101
5	2562419.808	-61940.158	.172	1.263
6	2513512.908	-48906.899	.136	1.530
7	2481800.191	-31712.718	.088	1.076
8	2452379.221	-29420.970	.082	1.042
9	2424169.122	-28210.099	.078	1.003
10	2396038.403	-28130.719	.078	1.038
11	2368952.782	-27085.621	.075	1.056
12	2343334.223	-25618.559	.071	1.133
13	2320812.903	-22521.319	.062	1.052
14	2299449.905	-21362.998	.059	1.134
15	2280687.923	-18761.983	.052	1.015

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Tabla 6.2: Resultados corrida cluster (Criterios)

del ciclo de vida Maduros (personas con edades comprendidas entre 41 y 50 años), si bien el cluster 1 esta conformado por clientes que se encuentra dentro del mismo ciclo de vida, el cluster 1 esta conformado casi en su totalidad por clientes del sexo masculino y este cluster está conformado por clientes de ambos sexos[2].

- ***Para el conglomerado 4***

  Este cluster está conformado por clientes dentro del ciclo de vida Maduros (personas con edades comprendidas entre 41 y 50 años), también clientes dentro del ciclo de vida Mayores (personas con edades comprendidas entre 51 y 60 años) y en misma proporción con clientes dentro del ciclo de vida Ancianos (personas con edad superior a 60 años).

El gráfico (6.4) muestra la distribución de hombres y mujeres que existe dentro de cada cluster.

En el gráfico (6.5) se observan los diferentes tipos de negocios en los cuales el cliente de institución bancaria se desarrolla económicamente.

Para analizar el estado civil preponderante en cada cluster a continuación se presenta el gráfico (6.6).

[2]Obsérvese el gráfico (6.4) en la página (51)

Cluster Distribution

		N	% of Combined	% of Total
Cluster	1	39450	26.8%	26.8%
	2	51313	34.9%	34.9%
	3	20233	13.8%	13.8%
	4	36058	24.5%	24.5%
	Combined	147054	100.0%	100.0%
Total		147054		100.0%

Tabla 6.3: Distribución de Clusters

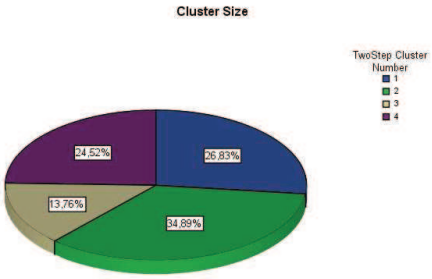

Figura 6.2: Tabla de Porcentajes Cluster

En el gráfico (6.7) se observan las diferentes provincias a los que pertenecen los cliente de institución bancaria.

En el gráfico (6.8) se observan en promedio el número de crédito que los clientes realizan por el canal ventanilla.

En el gráfico (6.9) se observan en promedio el número de débito que los clientes realizan por el canal ventanilla.

En el gráfico (6.10) se observan en promedio el número de débito que los clientes realizan por el canal ATM.

		Número de transacciones de crédito que el cliente ha realizado por el canal VENTANILLA en el período de análisis.		Número de transacciones de débito que el cliente ha realizado por el canal VENTANILLA en el período de análisis.		Número de transacciones de débito que el cliente ha realizado por el canal ATM en el período de análisis.	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	7.387	11.5404	3.678	7.7034	13.062	27.6241
	2	9.367	12.3064	2.380	5.6988	22.762	35.1253
	3	34.108	29.2906	27.486	24.3689	53.523	60.2728
	4	5.308	9.7487	3.340	7.4918	5.481	16.5581
	Combined	11.245	17.7680	6.418	13.9216	20.155	37.7376

Tabla 6.4: Centroides de variables continuas en el modelo de cluster

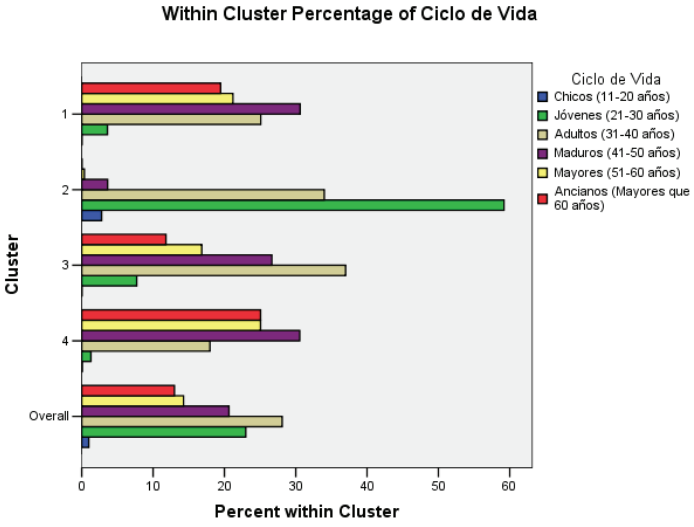

Figura 6.3: Ciclo de vida por cada cluster

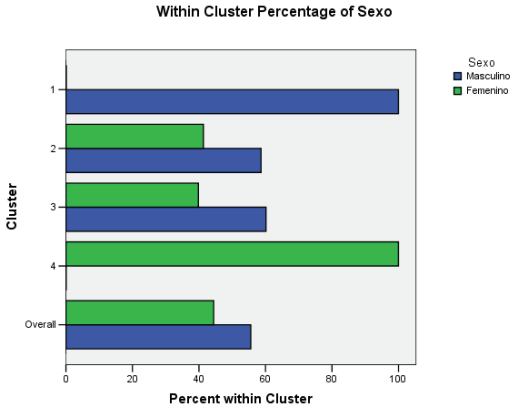

Figura 6.4: Distribución de hombres y mujeres dentro de cada cluster

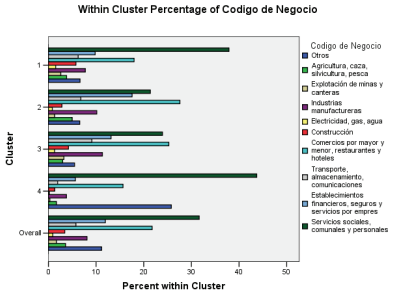

Figura 6.5: Código de Negocio del cliente por cada cluster

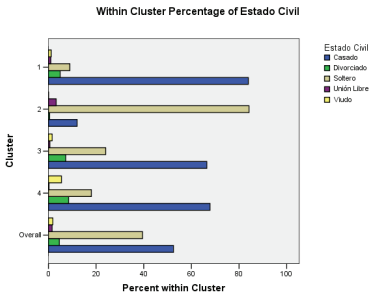

Figura 6.6: Estado civil por cada cluster

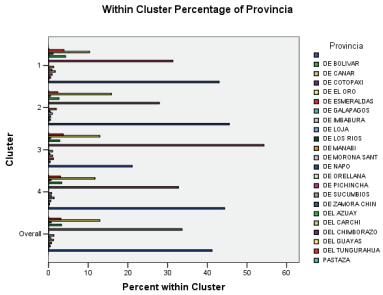

Figura 6.7: Provincia

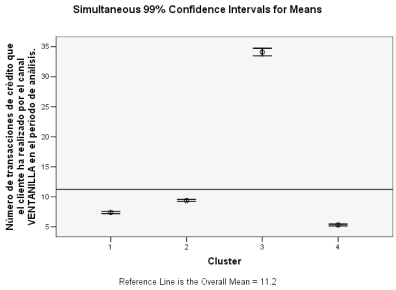

Figura 6.8: Crédito Ventanilla

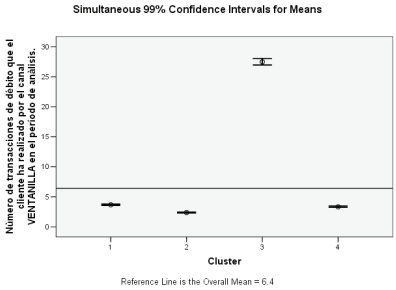

Figura 6.9: Débito en Ventanilla

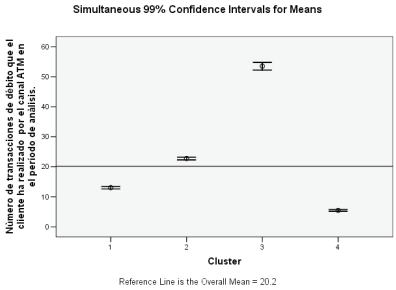

Figura 6.10: Débito ATM

6.3 Resultados Análisis de Supervivencia

6.3.1 Cálculo Supervivencia

Después de haber clasificado al grupo de clientes en cuatro conglomerados, cada uno enriquecido del carácter demográfico y transaccional, se procede a calcular la Supervivencia para cada cliente, para el cual se añadieron las siguientes variables:

Muerte pasiva	variable nominal
Antigüedad meses	variable continua

La variable muerte pasiva es dicotómica en la que 1 indica ocurrencia del evento de interés, en este el retiro como cliente del Banco. Un resumen del total de clientes que presenta el evento de interés se la ve en la siguiente tabla:

Total	muertos	%	
Conglomerado 1	36,058	2,850	7.90%
Conglomerado 2	39,450	3,337	8.46%
Conglomerado 3	51,313	4,633	9.03%
Conglomerado 4	20,233	5,423	26.80%

El proceso para el cálculo de Supervivencia (ejecutado en SPSS) se lo realiza de la siguiente manera:

- La variable temporal de análisis es la antigüedad del cliente
- El estado viene dado por la variable muerte pasiva, que es variable dicotómica en la el evento de interés es saber si dentro del período de análisis el cliente dejó de pertenecer a la institución.
- Las covariables son: ciclo de vida en meses, sexo, estado civil y la provincia de origen
- El método utilizado es el LR
- Algo adicional que se debe pedir al programa es la generación de las $\mathbf{x}\beta$ que son los coeficientes de la regresión, la función fuerza de mortandad o función de hazard y la función de supervivencia, cabe señalar a a partir de esta última uno puede generar la función $S_o = S^{\frac{1}{\exp(\mathbf{x}\beta}})$.

6.3.2 Resultados Supervivencia

■Conglomerado 1

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	2855	7.9%
	Censored	32931	91.3%
	Total	35786	99.2%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	272	.8%
	Total	272	.8%
Total		36058	100.0%

a. Dependent Variable: cl_antiguedad_meses

Figura 6.11: Resumen de la Regresión

En el gráfico (6.11) se observa un resumen de la regresión, donde consta el número de clientes, el número de eventos que han ocurrido, casos censurados y casos con tiempos negativos.

Omnibus Tests of Model Coefficients[d,e]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	54776.022	198.043	7	.000	150.882	7	.000	150.882	7	.000
2[b]	54695.759	279.363	12	.000	80.263	5	.000	231.145	12	.000
3[c]	54637.539	332.186	16	.000	58.220	4	.000	289.365	16	.000

a. Variable(s) Entered at Step Number 1: cl_codigo_provincia

b. Variable(s) Entered at Step Number 2: cl_ciclo_vida_meses

c. Variable(s) Entered at Step Number 3: cl_estado_civil

d. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 54926.904

e. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Figura 6.12: Pruebas de la Regresión

En el gráfico (6.12) se observa el test del método Forward Stepwise (likelihood ratio) que consiste en evaluar si la regresión realizada es la adecuada, se puede observar que el valor de la significancia es menor que 0.05% por tanto la regresión es válida.Tambien se debe tomar en cuenta que los valores de -2 log likelihood en todos los pasos son menores que el valor inicial 54926.904 condición necesaria para el ajuste de la regresión con respecto a los datos.

En el gráfico (6.13)se observa los estadísticos de Wald y significancia para cada variable;observando que son significativas;también se observa dentro de cada variable la significancia de cada una de las categorías que tiene dicha variable.

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	99.0% CI for Exp(B)	
								Lower	Upper
Step 1	cl_codigo_provincia			185.457	7	.000			
	cl_codigo_provincia(1)	-1.108	.103	116.804	1	.000	.330	.254	.430
	cl_codigo_provincia(2)	-1.042	.103	103.289	1	.000	.353	.271	.459
	cl_codigo_provincia(3)	-.898	.113	63.245	1	.000	.407	.304	.545
	cl_codigo_provincia(4)	-.786	.139	31.766	1	.000	.456	.318	.653
	cl_codigo_provincia(5)	-.514	.132	15.239	1	.000	.598	.426	.840
	cl_codigo_provincia(6)	-.468	.239	3.836	1	.050	.626	.338	1.159
	cl_codigo_provincia(7)	-.207	.180	1.321	1	.250	.813	.512	1.292
Step 2	cl_ciclo_vida_meses			80.344	5	.000			
	cl_ciclo_vida_meses(1)	.896	.579	2.396	1	.122	2.449	.552	10.876
	cl_ciclo_vida_meses(2)	.820	.174	22.228	1	.000	2.271	1.451	3.556
	cl_ciclo_vida_meses(3)	.456	.058	61.878	1	.000	1.578	1.359	1.832
	cl_ciclo_vida_meses(4)	.311	.051	37.191	1	.000	1.365	1.197	1.557
	cl_ciclo_vida_meses(5)	.245	.054	21.028	1	.000	1.278	1.114	1.467
	cl_codigo_provincia			184.342	7	.000			
	cl_codigo_provincia(1)	-1.116	.103	118.493	1	.000	.328	.251	.427
	cl_codigo_provincia(2)	-1.028	.103	100.570	1	.000	.358	.275	.466
	cl_codigo_provincia(3)	-.893	.113	62.507	1	.000	.409	.306	.548
	cl_codigo_provincia(4)	-.799	.139	32.780	1	.000	.450	.314	.644
	cl_codigo_provincia(5)	-.507	.132	14.852	1	.000	.602	.429	.845
	cl_codigo_provincia(6)	-.467	.239	3.812	1	.051	.627	.339	1.161
	cl_codigo_provincia(7)	-.239	.180	1.770	1	.183	.787	.496	1.251
Step 3	cl_ciclo_vida_meses			78.909	5	.000			
	cl_ciclo_vida_meses(1)	1.172	.580	4.081	1	.043	3.228	.724	14.381
	cl_ciclo_vida_meses(2)	.803	.175	21.064	1	.000	2.231	1.422	3.501
	cl_ciclo_vida_meses(3)	.459	.060	58.291	1	.000	1.583	1.356	1.848
	cl_ciclo_vida_meses(4)	.348	.053	43.378	1	.000	1.416	1.236	1.623
	cl_ciclo_vida_meses(5)	.264	.055	23.251	1	.000	1.303	1.131	1.500
	cl_estado_civil			54.834	4	.000			
	cl_estado_civil(1)	-.182	.082	4.894	1	.027	.834	.675	1.030
	cl_estado_civil(2)	-.094	.100	.897	1	.344	.910	.704	1.176
	cl_estado_civil(3)	-.525	.092	32.449	1	.000	.592	.467	.750
	cl_estado_civil(4)	.046	.344	.018	1	.892	1.048	.432	2.538
	cl_codigo_provincia			177.199	7	.000			
	cl_codigo_provincia(1)	-1.101	.103	115.198	1	.000	.333	.255	.433
	cl_codigo_provincia(2)	-1.045	.103	103.755	1	.000	.352	.270	.458
	cl_codigo_provincia(3)	-.881	.113	60.742	1	.000	.414	.310	.554
	cl_codigo_provincia(4)	-.820	.140	34.539	1	.000	.440	.307	.631
	cl_codigo_provincia(5)	-.532	.132	16.323	1	.000	.587	.419	.825
	cl_codigo_provincia(6)	-.465	.239	3.789	1	.052	.628	.339	1.162
	cl_codigo_provincia(7)	-.272	.180	2.297	1	.130	.762	.479	1.210

Figura 6.13: Variables de la Ecuación de Supervivencia

En el gráfico (6.14)de igual forma se observa las variables que no pertenecen a la ecuación o afectan a la supervivencia (son no significativas).

En el gráfico (6.15)se observa la función de supervivencia del cluster 1, es decir se observa la probabilidad que un individuo que tiene en una antigüedad dada siga en el banco.En este cluster la probabilidad de que un cliente permanezca en el banco más baja es de 0.85 correspondiente a la antigüedad de 150 meses(12.5 años).Se observa que la caída de la supervivencia empieza en la antigüedad de 50 meses(4.1 años).

En el gráfico (6.16)se observa la función de Hazard del cluster 1, es decir se observa el riesgo de que un individuo que tiene en una antigüedad dada cambie de estado en ese instante.

■Conglomerado 2

En el gráfico (6.17) se observa un resumen de la regresión, donde consta el número de clientes, el número de eventos que han ocurrido, casos censurados y casos con tiempos negativos

Variables not in the Equation[b,c]

		Score	df	Sig.
Step 1	cl_ciclo_vida_meses	81.516	5	.000
	cl_ciclo_vida_meses(1)	1.335	1	.248
	cl_ciclo_vida_meses(2)	11.685	1	.001
	cl_ciclo_vida_meses(3)	28.851	1	.000
	cl_ciclo_vida_meses(4)	6.179	1	.013
	cl_ciclo_vida_meses(5)	.037	1	.848
	cl_sexo	.049	1	.826
	cl_estado_civil	54.985	4	.000
	cl_estado_civil(1)	21.118	1	.000
	cl_estado_civil(2)	4.253	1	.039
	cl_estado_civil(3)	53.496	1	.000
	cl_estado_civil(4)	.827	1	.363
Step 2	cl_sexo	.050	1	.824
	cl_estado_civil	55.462	4	.000
	cl_estado_civil(1)	10.328	1	.001
	cl_estado_civil(2)	4.757	1	.029
	cl_estado_civil(3)	48.531	1	.000
	cl_estado_civil(4)	.765	1	.382
Step 3	cl_sexo	.048	1	.826

a. Residual Chi Square = 134.831 with 10 df Sig. = .000

b. Residual Chi Square = 55.499 with 5 df Sig. = .000

c. Residual Chi Square = .048 with 1 df Sig. = .826

Figura 6.14: Variables excluidas en la Ecuación de Supervivencia

En el gráfico (6.18)se observa el test del método Forward Stepwise (likelihood ratio) que consiste en evaluar si la regresión realizada es la adecuada, se puede observar que el valor de la significancia es menor que 0.05% por tanto la regresión es válida.Tambien se debe tomar en cuenta que los valores de -2 log likelihood en todos los pasos es menor al valor inicial 64751.126 condición necesaria para el ajuste de la regresión con respecto a los datos

En el gráfico (6.19) se observa los estadísticos de Wald y significancia para cada variable;observando que son significativas;también se observa dentro de cada variable la significancia de cada una de las categorías que tiene dicha variable.

En el gráfico (6.20) de igual forma se observa las variables que no pertenecen a la ecuación o afectan a la supervivencia (son no significativas)

En el gráfico (6.21)se observa la función de supervivencia del cluster 2, es decir se observa la probabilidad que un individuo que tiene en una antigüedad dada siga en el banco.En este cluster la probabilidad de que un cliente permanezca en el banco más baja es de 0.82 correspondiente a la antigüedad de 150 meses(12.5 años).Se observa que la caída de la supervivencia empieza en la antigüedad de 50 meses(4.1 años).

En el gráfico (6.22)se observa la función de Hazard del cluster 2, es decir se observa el riesgo de que un individuo que tiene en una antigüedad dada cambie de estado en ese instante.

■Conglomerado 3

En el gráfico (6.23) se observa un resumen de la regresión, donde consta el

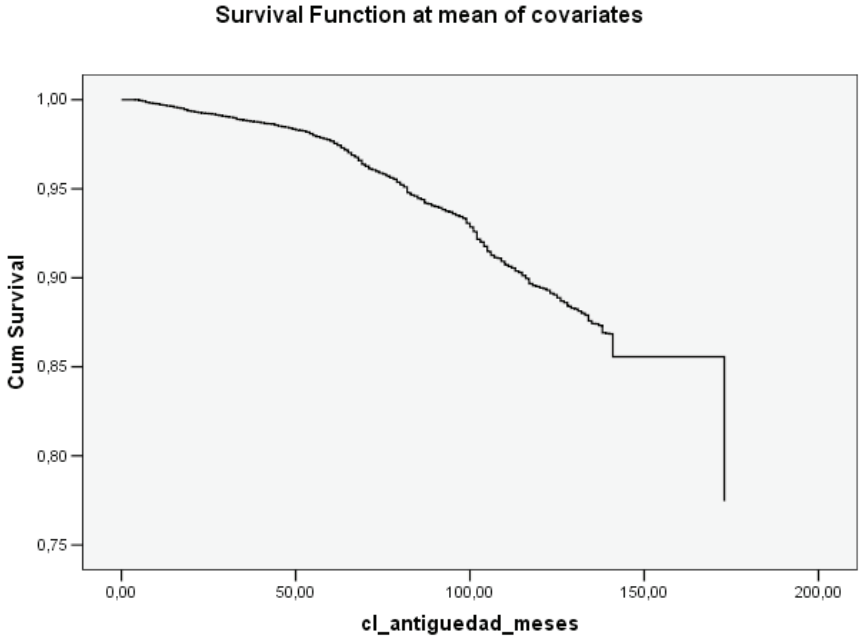

Figura 6.15: Función de supervivencia

número de clientes, el número de eventos que han ocurrido, casos censurados y casos con tiempos negativos

En el gráfico (6.24)se observa el test del método Forward Stepwise (likelihood ratio) que consiste en evaluar si la regresión realizada es la adecuada, se puede observar que el valor de la significancia es menor que 0.05% por tanto la regresión es válida.Tambien se debe tomar en cuenta que el valor -2 log likelihood en el paso 3 90684.140 es menor al valor inicial 91018.527 condición necesaria para el ajuste de la regresión con respecto a los datos

En el gráfico (6.25) se observa los estadísticos de Wald y significancia para cada variable; observando que son significativas;también se observa dentro de cada variable la significancia de cada una de las categorías que tiene dicha variable.

En el gráfico (6.26) se observa la función de supervivencia del cluster 3, es decir se observa la probabilidad que un individuo que tiene en una antigüedad dada siga en el banco.En este cluster la probabilidad de que un cliente permanezca en el banco más baja es de 0.6 correspondiente a la antigüedad de 150 meses(12.5 años). Se observa que la caída de la supervivencia empieza en la antigüedad de 24 meses(1año).

En el gráfico (6.27) se observa la función de Hazard del cluster 3, es decir se

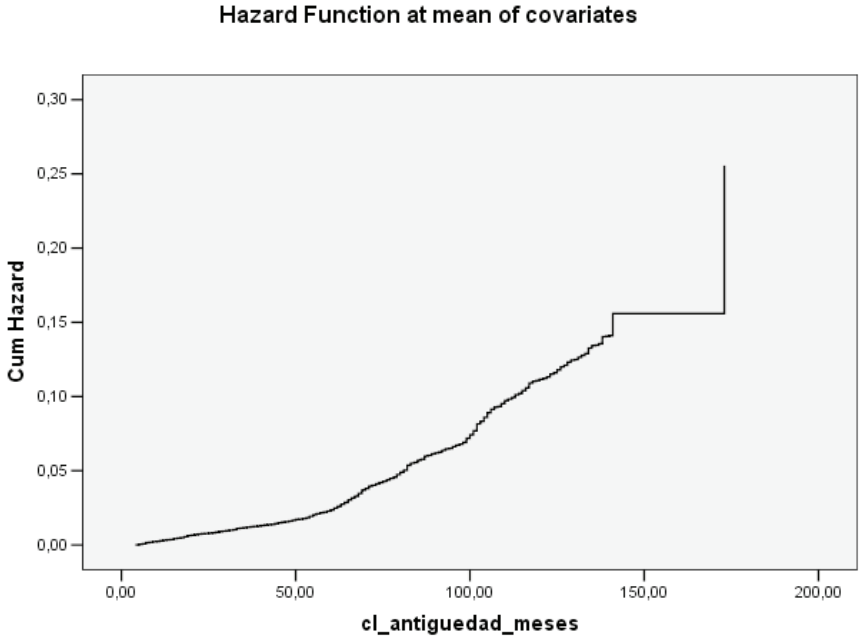

Figura 6.16: Función de Hazard

observa el riesgo de que un individuo que tiene en una antigüedad dada cambie de estado en ese instante.

■Conglomerado 4

En el gráfico (6.28) se observa un resumen de la regresión, donde consta el número de clientes, el número de eventos que han ocurrido, casos censurados y casos con tiempos negativos

En el gráfico (6.29) se observa el test del método Forward Stepwise (likelihood ratio) que consiste en evaluar si la regresión realizada es la adecuada, se puede observar que el valor de la significancia es menor que 0.05% por tanto la regresión es válida.Tambien se debe tomar en cuenta que el valor -2 log likelihood en los pasos 1 y 2 es menor al valor inicial 98966.372 condición necesaria para el ajuste de la regresión con respecto a los datos

En el gráfico (6.30) se observa los estadísticos de Wald y significancia para cada variable;observando que son significativas;también se observa dentro de cada variable la significancia de cada una de las categorías que tiene dicha variable.

En el gráfico (6.31)de igual forma se observa las variables que no pertenecen a la ecuación o afectan a la supervivencia (son no significativas)

En el gráfico (6.32)se observa la función de supervivencia del cluster 4, es decir

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	3351	8.5%
	Censored	35675	90.4%
	Total	39026	98.9%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	424	1.1%
	Total	424	1.1%
Total		39450	100.0%

a. Dependent Variable: cl_antiguedad_meses

Figura 6.17: Resumen de la Regresión

Omnibus Tests of Model Coefficients[d,e]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	64448.707	330.528	4	.000	302.419	4	.000	302.419	4	.000
2[b]	64284.903	532.634	11	.000	163.804	7	.000	466.223	11	.000
3[c]	64207.455	590.762	15	.000	77.448	4	.000	543.671	15	.000

a. Variable(s) Entered at Step Number 1: cl_ciclo_vida_meses

b. Variable(s) Entered at Step Number 2: cl_codigo_provincia

c. Variable(s) Entered at Step Number 3: cl_estado_civil

d. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 64751.126

e. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Figura 6.18: Pruebas de la Regresión

se observa la probabilidad que un individuo que tiene en una antigüedad dada siga en el banco.En este cluster la probabilidad de que un cliente permanezca en el banco más baja es de 0.59 correspondiente a la antigüedad de 150 meses(12.5 años).Se observa que la caída de la supervivencia empieza en la antigüedad de 75 meses(6 años).

En el gráfico (6.33)se observa la función de Hazard del cluster 4, es decir se observa el riesgo de que un individuo que tiene en una antigüedad dada cambie de estado en ese instante.

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	99.0% CI for Exp(B)	
								Lower	Upper
Step 1	cl_ciclo_vida_meses			318.343	4	.000			
	cl_ciclo_vida_meses(1)	1.093	.107	103.372	1	.000	2.983	2.261	3.934
	cl_ciclo_vida_meses(2)	.801	.052	234.430	1	.000	2.227	1.946	2.548
	cl_ciclo_vida_meses(3)	.322	.052	38.886	1	.000	1.380	1.208	1.576
	cl_ciclo_vida_meses(4)	.223	.056	15.848	1	.000	1.249	1.082	1.443
Step 2	cl_ciclo_vida_meses			303.381	4	.000			
	cl_ciclo_vida_meses(1)	1.061	.108	97.161	1	.000	2.888	2.189	3.810
	cl_ciclo_vida_meses(2)	.786	.052	225.548	1	.000	2.195	1.918	2.511
	cl_ciclo_vida_meses(3)	.324	.052	39.365	1	.000	1.383	1.211	1.580
	cl_ciclo_vida_meses(4)	.217	.056	15.068	1	.000	1.243	1.076	1.435
	cl_codigo_provincia			190.315	7	.000			
	cl_codigo_provincia(1)	-.894	.088	103.486	1	.000	.409	.326	.513
	cl_codigo_provincia(2)	-.784	.088	79.330	1	.000	.457	.364	.573
	cl_codigo_provincia(3)	-.696	.100	48.804	1	.000	.499	.386	.644
	cl_codigo_provincia(4)	-.565	.118	23.073	1	.000	.569	.420	.770
	cl_codigo_provincia(5)	-.316	.112	7.969	1	.005	.729	.546	.973
	cl_codigo_provincia(6)	.023	.163	.021	1	.885	1.024	.673	1.557
	cl_codigo_provincia(7)	-.222	.153	2.099	1	.147	.801	.539	1.189
Step 3	cl_ciclo_vida_meses			259.876	4	.000			
	cl_ciclo_vida_meses(1)	1.019	.108	89.148	1	.000	2.770	2.098	3.658
	cl_ciclo_vida_meses(2)	.749	.053	199.719	1	.000	2.114	1.844	2.423
	cl_ciclo_vida_meses(3)	.359	.052	47.237	1	.000	1.432	1.252	1.638
	cl_ciclo_vida_meses(4)	.228	.056	16.395	1	.000	1.256	1.086	1.452
	cl_estado_civil			66.832	4	.000			
	cl_estado_civil(1)	-.059	.159	.140	1	.708	.942	.626	1.419
	cl_estado_civil(2)	-.125	.178	.492	1	.483	.883	.559	1.395
	cl_estado_civil(3)	-.639	.172	13.756	1	.000	.528	.339	.823
	cl_estado_civil(4)	-.415	.279	2.208	1	.137	.660	.322	1.356
	cl_codigo_provincia			175.780	7	.000			
	cl_codigo_provincia(1)	-.859	.088	95.451	1	.000	.424	.338	.531
	cl_codigo_provincia(2)	-.795	.088	81.568	1	.000	.451	.360	.566
	cl_codigo_provincia(3)	-.684	.100	47.095	1	.000	.505	.390	.652
	cl_codigo_provincia(4)	-.575	.118	23.915	1	.000	.563	.416	.762
	cl_codigo_provincia(5)	-.332	.112	8.752	1	.003	.718	.538	.958
	cl_codigo_provincia(6)	.016	.163	.010	1	.921	1.016	.668	1.545
	cl_codigo_provincia(7)	-.222	.153	2.100	1	.147	.801	.539	1.189

Figura 6.19: Variables de la Ecuación de Supervivencia

Variables not in the Equation[a,b,c]

		Score	df	Sig.
Step 1	cl_sexo	.311	1	.577
	cl_estado_civil	79.468	4	.000
	cl_estado_civil(1)	53.229	1	.000
	cl_estado_civil(2)	.003	1	.955
	cl_estado_civil(3)	75.682	1	.000
	cl_estado_civil(4)	1.860	1	.173
	cl_codigo_provincia	198.557	7	.000
	cl_codigo_provincia(1)	44.009	1	.000
	cl_codigo_provincia(2)	2.142	1	.143
	cl_codigo_provincia(3)	1.002	1	.317
	cl_codigo_provincia(4)	4.911	1	.027
	cl_codigo_provincia(5)	34.317	1	.000
	cl_codigo_provincia(6)	31.492	1	.000
	cl_codigo_provincia(7)	16.674	1	.000
Step 2	cl_sexo	.285	1	.593
	cl_estado_civil	68.581	4	.000
	cl_estado_civil(1)	45.222	1	.000
	cl_estado_civil(2)	.020	1	.888
	cl_estado_civil(3)	65.260	1	.000
	cl_estado_civil(4)	1.479	1	.224
Step 3	cl_sexo	.304	1	.582

a. Residual Chi Square = 265.301 with 12 df Sig. = .000

b. Residual Chi Square = 68.904 with 5 df Sig. = .000

c. Residual Chi Square = .304 with 1 df Sig. = .582

Figura 6.20: Variables excluidas en la Ecuación de Supervivencia

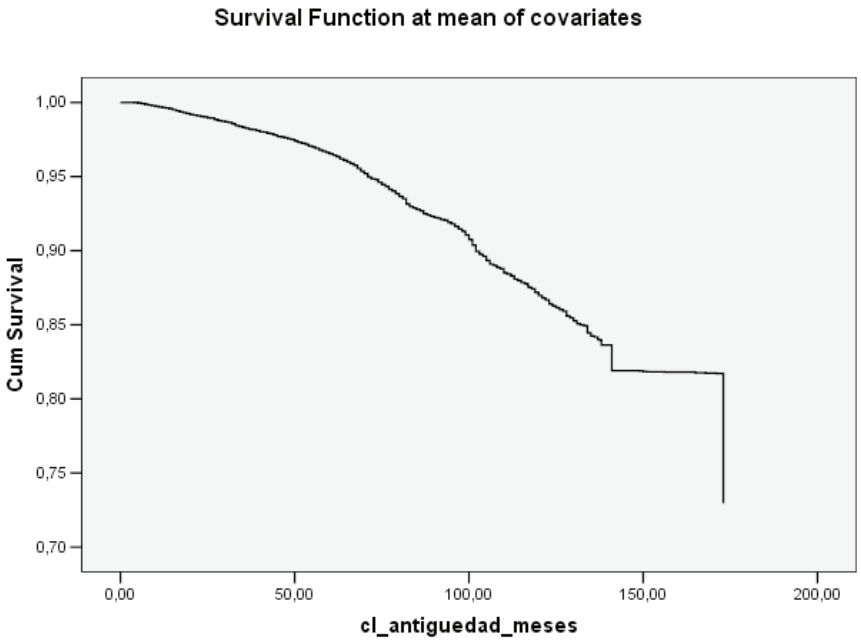

Figura 6.21: Función de supervivencia

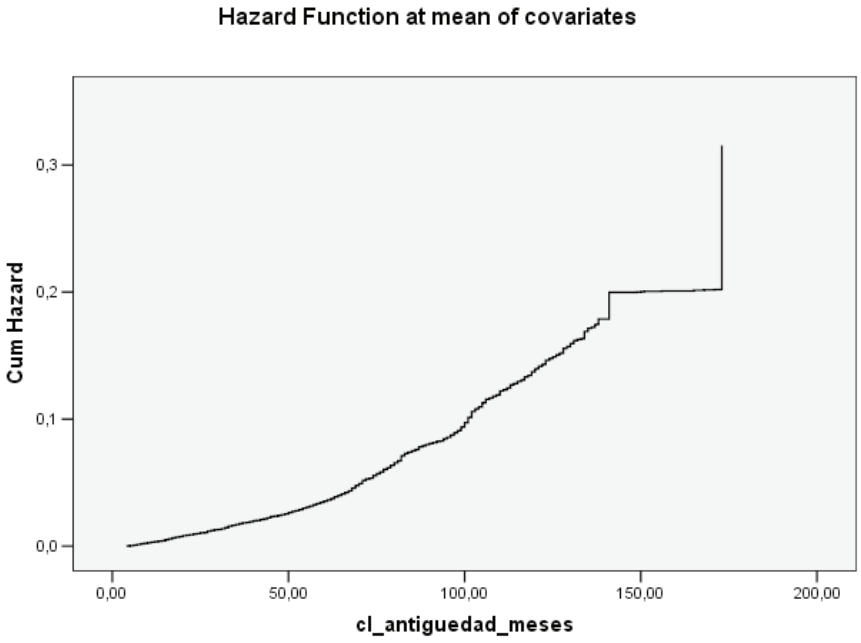

Figura 6.22: Función de Hazard

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	4668	9.1%
	Censored	46645	90.9%
	Total	51313	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		51313	100.0%

a. Dependent Variable: cl_antiguedad_meses

Figura 6.23: Resumen de la Regresión

Omnibus Tests of Model Coefficients[b]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.
3	90684.140	375.879	11	.000	334.387	11	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 91018.527

b. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Figura 6.24: Pruebas de la Regresión

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	99.0% CI for Exp(B)	
								Lower	Upper
Step 3	cl_sexo	-.097	.030	10.752	1	.001	.907	.840	.979
	cl_estado_civil			135.061	3	.000			
	cl_estado_civil(1)	.855	.121	49.909	1	.000	2.351	1.721	3.210
	cl_estado_civil(2)	1.015	.218	21.659	1	.000	2.759	1.573	4.837
	cl_estado_civil(3)	.373	.114	10.679	1	.001	1.452	1.082	1.948
	cl_codigo_provincia			157.383	7	.000			
	cl_codigo_provincia(1)	-.726	.088	67.550	1	.000	.484	.385	.607
	cl_codigo_provincia(2)	-.445	.089	25.096	1	.000	.641	.510	.806
	cl_codigo_provincia(3)	-.508	.093	29.810	1	.000	.602	.474	.765
	cl_codigo_provincia(4)	-.444	.124	12.880	1	.000	.642	.467	.882
	cl_codigo_provincia(5)	-.246	.120	4.202	1	.040	.782	.574	1.065
	cl_codigo_provincia(6)	-.125	.142	.777	1	.378	.882	.612	1.272
	cl_codigo_provincia(7)	-.047	.152	.095	1	.758	.954	.646	1.411

Figura 6.25: Variables de la Ecuación de Supervivencia

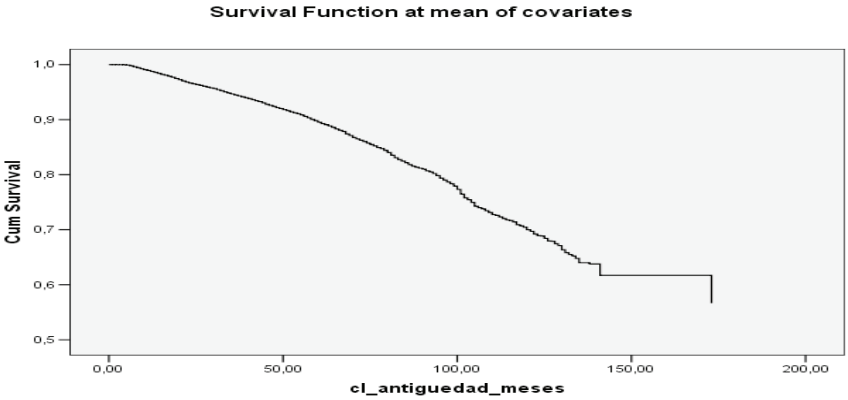

Figura 6.26: Función de supervivencia

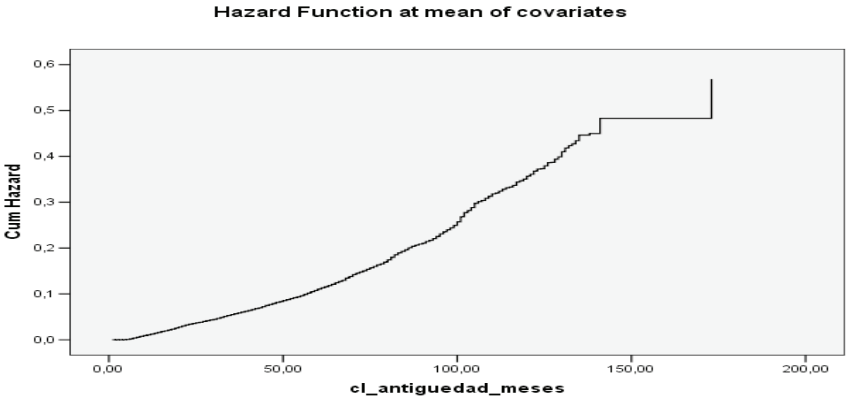

Figura 6.27: Función de Hazard

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	5426	26.8%
	Censored	14797	73.1%
	Total	20223	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	10	.0%
	Total	10	.0%
Total		20233	100.0%

a. Dependent Variable: cl_antiguedad_meses

Figura 6.28: Resumen de la Regresión

Omnibus Tests of Model Coefficients[d]

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	98140.513	1027.118	5	.000	825.859	5	.000	825.859	5	.000
2[b]	97931.385	1274.887	12	.000	209.127	7	.000	1034.986	12	.000

a. Variable(s) Entered at Step Number 1: cl_ciclo_vida_meses

b. Variable(s) Entered at Step Number 2: cl_codigo_provincia

c. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 98966.372

d. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Figura 6.29: Pruebas de la Regresión

Variables in the Equation

		B	SE	Wald	df	Sig.	Exp(B)	99.0% CI for Exp(B)	
								Lower	Upper
Step 1	cl_ciclo_vida_meses			932.557	5	.000			
	cl_ciclo_vida_meses(1)	-4.677	38.014	.015	1	.902	.009	.000	3E+040
	cl_ciclo_vida_meses(2)	1.668	.064	675.639	1	.000	5.299	4.492	6.251
	cl_ciclo_vida_meses(3)	.781	.047	271.272	1	.000	2.184	1.933	2.468
	cl_ciclo_vida_meses(4)	.347	.050	49.082	1	.000	1.415	1.246	1.608
	cl_ciclo_vida_meses(5)	.228	.054	17.961	1	.000	1.256	1.094	1.443
Step 2	cl_ciclo_vida_meses			939.114	5	.000			
	cl_ciclo_vida_meses(1)	-4.806	38.758	.015	1	.901	.008	.000	2E+041
	cl_ciclo_vida_meses(2)	1.675	.064	679.447	1	.000	5.336	4.523	6.297
	cl_ciclo_vida_meses(3)	.774	.047	265.610	1	.000	2.168	1.919	2.450
	cl_ciclo_vida_meses(4)	.336	.050	45.906	1	.000	1.400	1.232	1.591
	cl_ciclo_vida_meses(5)	.225	.054	17.392	1	.000	1.252	1.090	1.438
	cl_codigo_provincia			240.614	7	.000			
	cl_codigo_provincia(1)	-.663	.080	68.838	1	.000	.515	.420	.633
	cl_codigo_provincia(2)	-.914	.077	141.002	1	.000	.401	.329	.489
	cl_codigo_provincia(3)	-.686	.084	66.580	1	.000	.503	.405	.625
	cl_codigo_provincia(4)	-.605	.111	29.587	1	.000	.546	.410	.727
	cl_codigo_provincia(5)	-.709	.103	47.001	1	.000	.492	.377	.642
	cl_codigo_provincia(6)	.015	.159	.009	1	.923	1.016	.675	1.528
	cl_codigo_provincia(7)	-.182	.148	1.510	1	.219	.833	.569	1.221

Figura 6.30: Variables de la Ecuación de Supervivencia

Variables not in the Equation

		Score	df	Sig.
Step 3	cl_estado_civil	2.708	4	.608
	cl_estado_civil(1)	2.016	1	.156
	cl_estado_civil(2)	.459	1	.498
	cl_estado_civil(3)	1.583	1	.208
	cl_estado_civil(4)	.005	1	.943

a. Residual Chi Square = 2.708 with 4 df Sig. = .608

Figura 6.31: Función de supervivencia

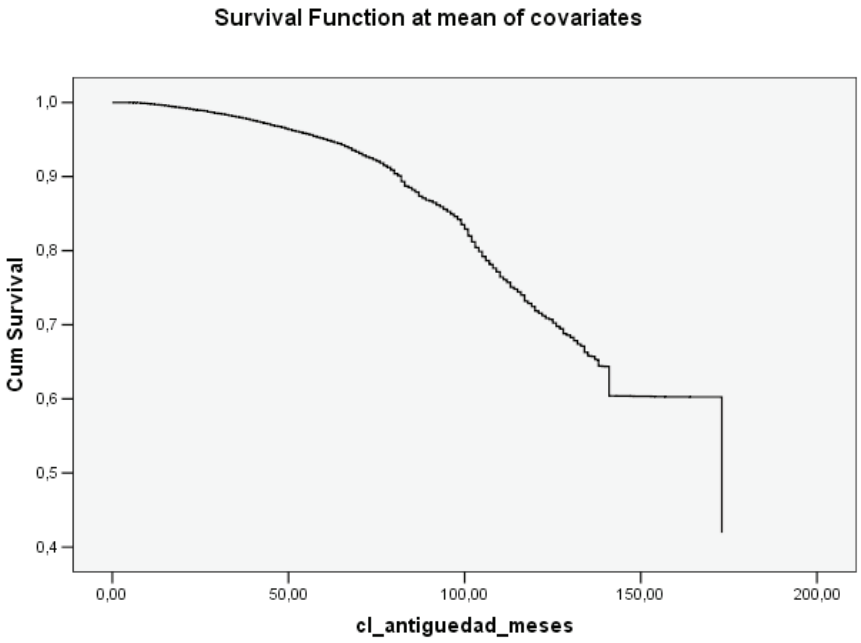

Figura 6.32: Función de supervivencia

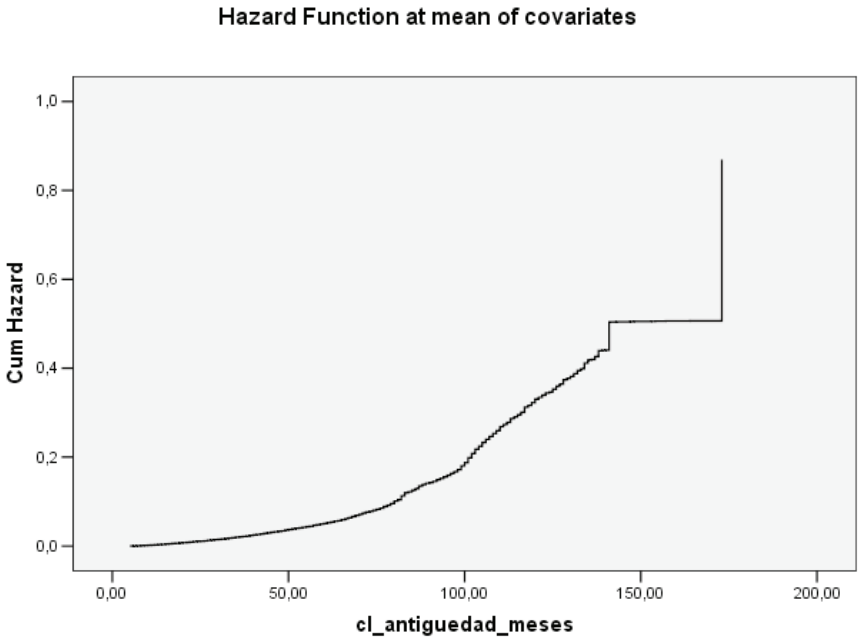

Figura 6.33: Función de Hazard

6.3.3 Análisis LT y LTV

LT

Después de haber obtenido la función de supervivencia, el siguiente paso es el cálculo del LT descrito en el Capítulo 5,el cálculo de la integral se la realizó en el programa MATLAB7, se deben considerar los siguientes aspectos para evaluar (5.5)

- Se debe colocar el número de intervalos para evaluar la integral
- Se debe tomar en cuenta que la integral se evalúa para cliente desde su edad hasta la edad ($2\pi a$) estimada por el cluster al que él pertenece

previo al cálculo del LT de cada cliente se debe ajusta la S_o usando (5.6), los resultados obtenidos fueron los siguientes[3]:

- **Conglomerado 1**

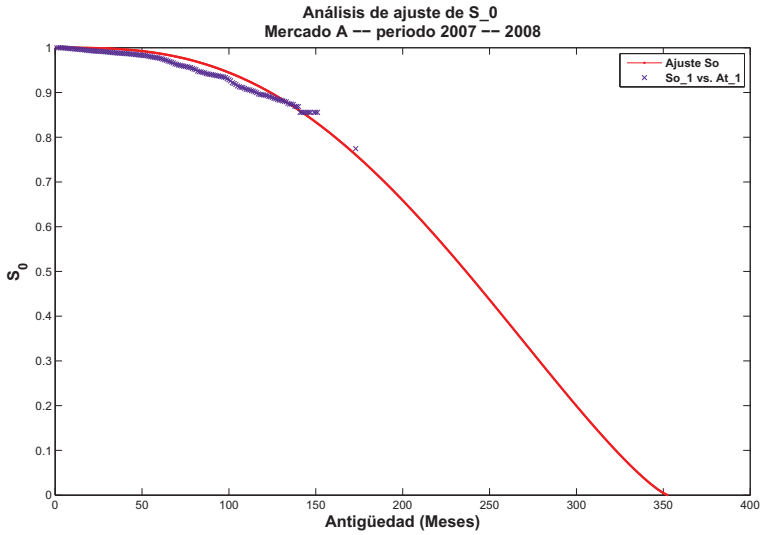

Figura 6.34: Aproximación de S_0, Conglomerado 1

$$
\begin{aligned}
S_0^1(x) &= \left(\frac{\sin(\frac{x}{a}) - \frac{x}{a} + 2\,\pi}{2\,\pi}\right)^b \\
a &= 56.13 \qquad (48.59, 63.67) \\
b &= 0.4171 \qquad (0.2561, 0.5782) \\
R^2 &= 0.9593
\end{aligned}
\tag{6.1}
$$

[3]La implementación de este cálculo en MATLAB se encuentra detallado en el Anexo A.2

- **Conglomerado 2**

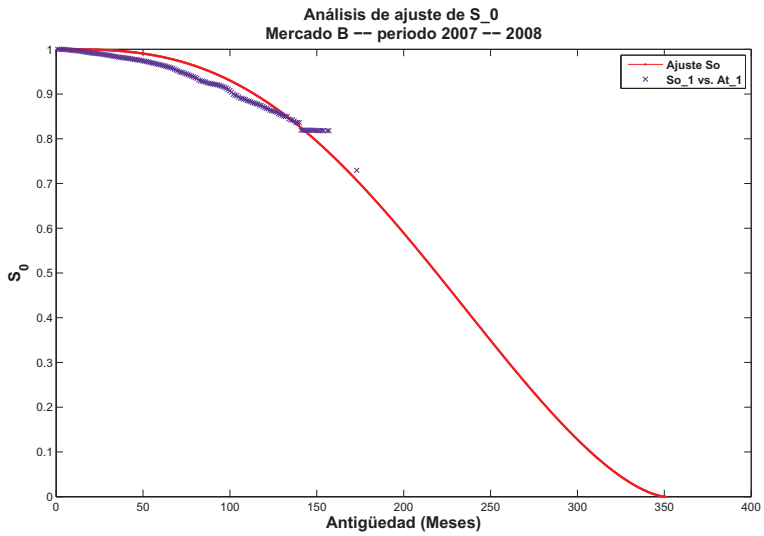

Figura 6.35: Aproximación de S_0, Conglomerado 2

$$
\begin{aligned}
S_0^2(x) &= \left(\frac{\sin(\frac{x}{a}) - \frac{x}{a} + 2\,\pi}{2\,\pi}\right)^b \\
a &= 55.84 \qquad (47.74, 63.94) \\
b &= 0.5206 \qquad (0.3036, 0.7377) \\
R^2 &= 0.9435
\end{aligned}
\tag{6.2}
$$

- **Conglomerado 3**

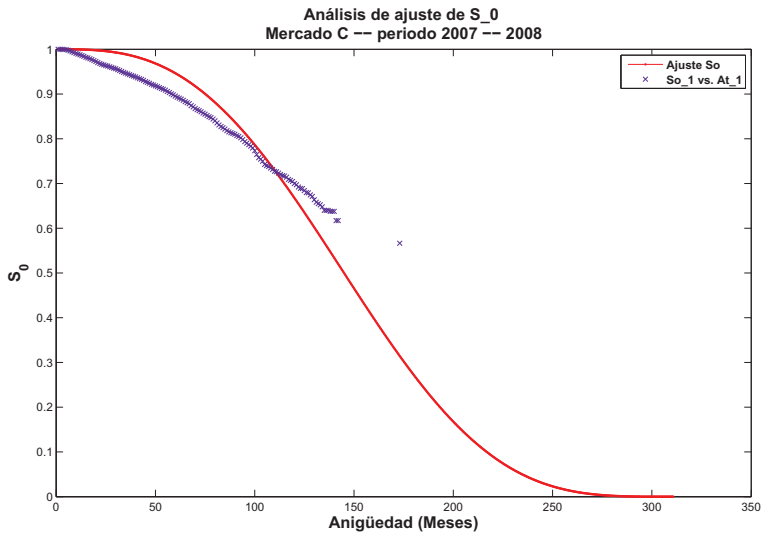

Figura 6.36: Aproximación de S_0, Conglomerado 3

$$
\begin{aligned}
S_0^3(x) &= \left(\frac{\sin(\frac{x}{a}) - \frac{x}{a} + 2\,\pi}{2\,\pi}\right)^b \\
a &= 49.51 \qquad (42.7, 56.32) \\
b &= 1.224 \qquad (0.7404, 1.708) \\
R^2 &= 0.7611
\end{aligned}
\tag{6.3}
$$

- **Conglomerado 4**

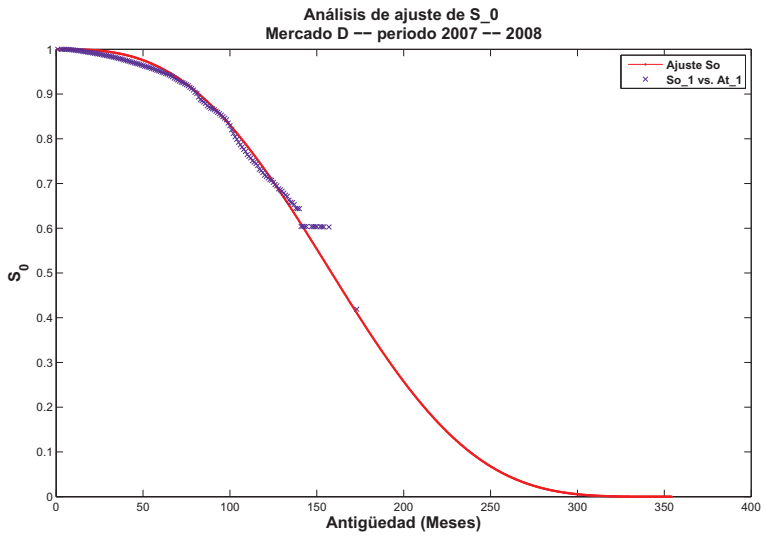

Figura 6.37: Aproximación de S_0, Conglomerado 4

$$
\begin{aligned}
S_0^3(x) &= \left(\frac{\sin(\frac{x}{a}) - \frac{x}{a} + 2\,\pi}{2\,\pi}\right)^b \\
a &= 56.38 \qquad (53.67, 59.09) \\
b &= 1.374 \qquad (1.185, 1.564) \\
R^2 &= 0.9974
\end{aligned}
\tag{6.4}
$$

Observación 6.3.1. *Nótese que para este conglomerado el ajuste con la función (5.6) es la manera obtenida que el resto de conglomerados.*

LTV

Después de haber obtenido el LT para cliente, el objetivo siguiente es el cálculo del LTV, que no es más que los flujos de rentabilidad generados en su LT traídos a valor presente.

Se deben tomar en cuenta las siguientes consideraciones:

- Se calculó la rentabilidad promedio para cada edad por cluster, para esto se ajustó una función de rentabilidad con los datos obtenidos de cada cliente

– **Conglomerado 1**

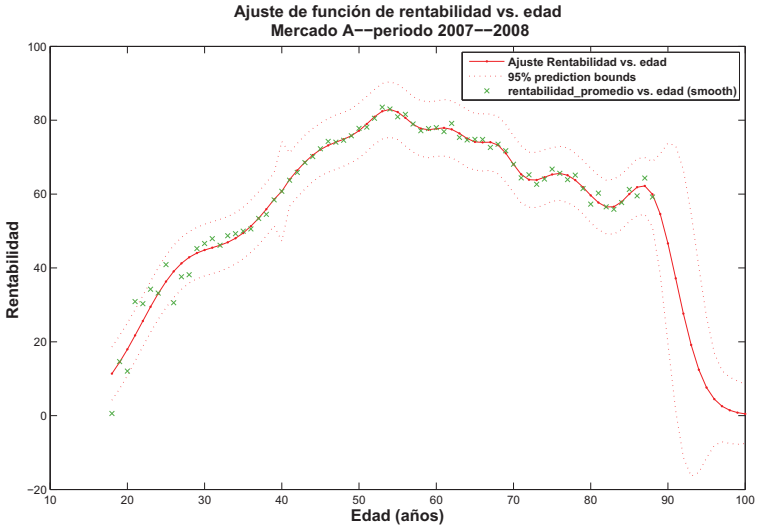

Figura 6.38: Función de rentabilidad vs. edad, Conglomerado 1

$$
\begin{aligned}
\overline{\mathfrak{R}}^{C^1}(x) = {} & a1\ \exp(-((x-b1)/c1)^2) + a2\ \exp(-((x-b2)/c2)^2) + \\
& a3\ \exp(-((x-b3)/c3)^2) + a4\ \exp(-((x-b4)/c4)^2) + \\
& a5\ \exp(-((x-b5)/c5)^2) + a6\ \exp(-((x-b6)/c6)^2) \\
& a7\ \exp(-((x-b7)/c7)^2) + a8\ \exp(-((x-b8)/c8)^2) \\
a1 = {} & 73.31; \qquad (59.57, 87.05) \\
b1 = {} & 47.32; \qquad (36.41, 58.22) \\
c1 = {} & 16.21; \qquad (-7.05, 39.46) \\
a2 = {} & 35.37; \qquad (-48.35, 119.1) \\
b2 = {} & 62.13; \qquad (57.15, 67.1) \\
c2 = {} & 5.457; \qquad (-14.34, 25.25) \\
a3 = {} & 16.79; \qquad (-87.71, 121.3) \\
b3 = {} & 54.36; \qquad (44.47, 64.25) \\
c3 = {} & 4.121; \qquad (-5.183, 13.43) \\
a4 = {} & -2.728e+006; \qquad (-7.349e+022, 7.349e+022) \\
b4 = {} & 39.67; \qquad (-2.083e+015, 2.083e+015) \\
c4 = {} & 0.08408; \qquad (-5.287e+014, 5.287e+014) \\
a5 = {} & 16.08; \qquad (-110.9, 143.1) \\
b5 = {} & 68.06; \qquad (62.88, 73.23) \\
c5 = {} & 3.311; \qquad (-4.275, 10.9) \\
a6 = {} & 26.15; \qquad (-18.65, 70.95) \\
b6 = {} & 26.44; \qquad (23.29, 29.59) \\
c6 = {} & 8.003; \qquad (2.393, 13.61) \\
a7 = {} & 62.42; \qquad (28.11, 96.73) \\
b7 = {} & 76.58; \qquad (71.41, 81.75) \\
c7 = {} & 10.43; \qquad (-4.97, 25.82) \\
a8 = {} & 40.99; \qquad (-27.06, 109) \\
b8 = {} & 88.05; \qquad (86.07, 90.03) \\
c8 = {} & 4.757; \qquad (-0.8613, 10.37) \\
R^2 = {} & 0.9787
\end{aligned}
\tag{6.5}
$$

– **Conglomerado 2**

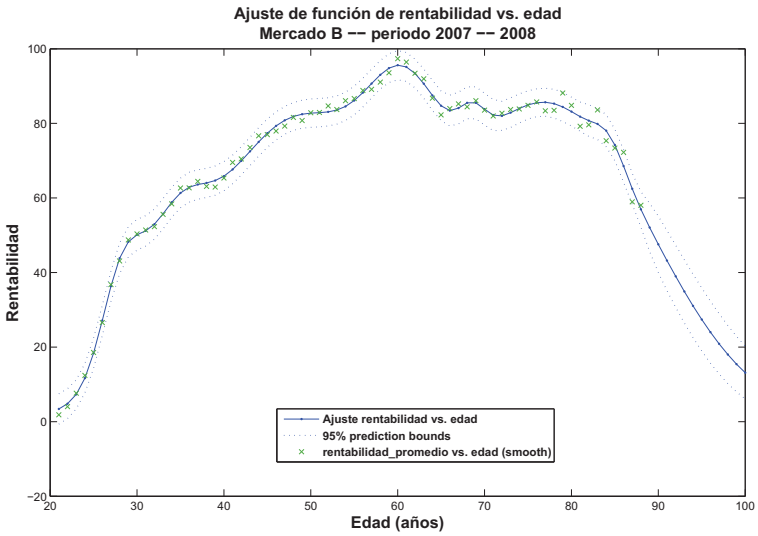

Figura 6.39: Función de rentabilidad vs. edad, Conglomerado 2

$$
\begin{aligned}
\overline{\mathfrak{R}}^{C^1}(x) = {} & a1\ \exp(-((x-b1)/c1)^2) + a2\ \exp(-((x-b2)/c2)^2) + \\
& a3\ \exp(-((x-b3)/c3)^2) + a4\ \exp(-((x-b4)/c4)^2) + \\
& a5\ \exp(-((x-b5)/c5)^2) + a6\ \exp(-((x-b6)/c6)^2) \\
& a7\ \exp(-((x-b7)/c7)^2) \\
a1 = {} & 28.4; \qquad (18.88, 37.91) \\
b1 = {} & 60.64; \qquad (60.1, 61.18) \\
c1 = {} & 5.758; \qquad (4.52, 6.995) \\
a2 = {} & 5.869; \qquad (1.62, 10.12) \\
b2 = {} & 84.41; \qquad (83.18, 85.63) \\
c2 = {} & 2.428; \qquad (0.1157, 4.741) \\
a3 = {} & 7.622; \qquad (3.494, 11.75) \\
b3 = {} & 68.54; \qquad (67.74, 69.34) \\
c3 = {} & 2.246; \qquad (0.8706, 3.622) \\
a4 = {} & 83.97; \qquad (81.23, 86.7) \\
b4 = {} & 77.57; \qquad (76.25, 78.9) \\
c4 = {} & 16.47; \qquad (14.06, 18.87) \\
a5 = {} & 24.42; \qquad (17.26, 31.58) \\
b5 = {} & 33.98; \qquad (32.04, 35.93) \\
c5 = {} & 5.075; \qquad (2.909, 7.242) \\
a6 = {} & 78.28; \qquad (74.26, 82.3) \\
b6 = {} & 47.88; \qquad (46.29, 49.46) \\
c6 = {} & 14.99; \qquad (12.71, 17.26) \\
a7 = {} & 24.34; \qquad (10.13, 38.55) \\
b7 = {} & 28.18; \qquad (27.3, 29.07) \\
c7 = {} & 3.32; \qquad (2.329, 4.311) \\
R^2 = {} & 0.996
\end{aligned}
\tag{6.6}
$$

– **Conglomerado 3**

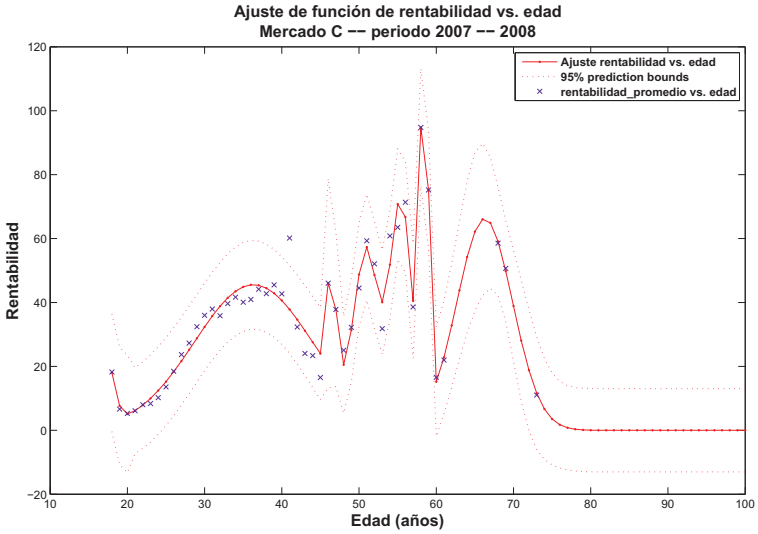

Figura 6.40: Función de rentabilidad vs. edad, Conglomerado 3

$$
\begin{aligned}
\overline{\mathfrak{R}}^{C^1}(x) = {} & a1\ \exp(-((x-b1)/c1)^2) + a2\ \exp(-((x-b2)/c2)^2) + \\
& a3\ \exp(-((x-b3)/c3)^2) + a4\ \exp(-((x-b4)/c4)^2) + \\
& a5\ \exp(-((x-b5)/c5)^2) + a6\ \exp(-((x-b6)/c6)^2) \\
& a7\ \exp(-((x-b7)/c7)^2) \\
a1 = 967.4; \quad & (-4.492e+010, 4.492e+010) \\
b1 = 58.49; \quad & (-1.557e+005, 1.558e+005) \\
c1 = 0.3082; \quad & (-2.719e+006, 2.719e+006) \\
a2 = 69.62; \quad & (58.32, 80.93) \\
b2 = 55.41; \quad & (55, 55.82) \\
c2 = 1.996; \quad & (1.027, 2.966) \\
a3 = 23.24; \quad & (-652.3, 698.8) \\
b3 = 17.05; \quad & (-32.33, 66.43) \\
c3 = 1.483; \quad & (-23.36, 26.33) \\
a4 = 49.54; \quad & (38.56, 60.52) \\
b4 = 50.99; \quad & (50.37, 51.62) \\
c4 = 2.081; \quad & (1.187, 2.974) \\
a5 = 66.2; \quad & (47.23, 85.18) \\
b5 = 66.28; \quad & (65.26, 67.3) \\
c5 = 5.095; \quad & (3.77, 6.42) \\
a6 = 2.482e+004; \quad & (-8.731e+019, 8.731e+019) \\
b6 = 46.5; \quad & (-2.433e+012, 2.433e+012) \\
c6 = 0.1887; \quad & (-4.728e+013, 4.728e+013) \\
a7 = 45.56; \quad & (40.74, 50.37) \\
b7 = 36.33; \quad & (35.3, 37.37) \\
c7 = 10.83; \quad & (9.098, 12.55) \\
R^2 = 0.9438 &
\end{aligned}
\tag{6.7}
$$

– **Conglomerado 4**

c

$$
\begin{aligned}
\overline{\Re}^{C^1}(x) = {} & a1\ \exp(-((x-b1)/c1)^2) + a2\ \exp(-((x-b2)/c2)^2) + \\
& a3\ \exp(-((x-b3)/c3)^2) + a4\ \exp(-((x-b4)/c4)^2) + \\
& a5\ \exp(-((x-b5)/c5)^2) + a6\ \exp(-((x-b6)/c6)^2) \\
a1 = 66.99; \quad & (36.64, 97.34) \\
b1 = 86.16; \quad & (84.76, 87.57) \\
c1 = 4.399; \quad & (1.375, 7.423) \\
a2 = -17.83; \quad & (-37.8, 2.144) \\
b2 = 67.23; \quad & (64.63, 69.84) \\
c2 = 3.418; \quad & (-1.129, 7.966) \\
a3 = -12.92; \quad & (-43.69, 17.86) \\
b3 = 54.76; \quad & (52.87, 56.66) \\
c3 = 0.9196; \quad & (-1.767, 3.606) \\
a4 = 20.81; \quad & (-4.58, 46.2) \\
b4 = 74.15; \quad & (72.37, 75.93) \\
c4 = 1.934; \quad & (-0.9734, 4.842) \\
a5 = 195.4; \quad & (187.3, 203.4) \\
b5 = 58.77; \quad & (53.74, 63.8) \\
c5 = 49.77; \quad & (42.22, 57.32) \\
R^2 = 0.8417 &
\end{aligned}
\tag{6.8}
$$

- Se tomó como tasa de interés 0.1433 dado por el banco

El cálculo del LT y LTV para cada individuo fue calculado utilizando la herramienta científica MATLAB7.

6.3.4 Score

En tabla (6.5) se detallan los valores de corte para el LTV por cada conglomerado.

como se observa en la tabla (6.5) existen grandes diferencias entre los conlomerados hallados incluso tomando como función de score a los cuartiles.

Conglomerado	Q1	Q2	Q3
1	$11.59	$93.05	$314.85
2	$3.46	$56.82	$273.30
3	$0.51	$6.90	$69.89
4	$73.16	$242.57	$702.48

Tabla 6.5: Cuartiles de LTV por conglomerado

- Llevar a cabo un modelo de segmentación en una institución bancaria e incluso en cualquier otra empresa que provea servicios o venda sus productos debe ser el pilar fundamental si pretende identificar perfiles de consumidores, para de esta manera poder tomar acciones acordes con la realidad ya sea de la empresa o la institución bancaria.

- Todo estudio que utilize más de dos variables para realizar cualquier tipo de análisis debe necesariamente realizar una depuración de datos atípicos, pero no utilizando técnicas univariantes sino más bien multivariantes por toda las implicaciones que conlleva y que se discutieron al inicio de este trabajo.

- El modelo de segmentación desarrollado aquí es un intento por presentar una alternativa diferente de segmentación a las utilizadas hasta la actualidad.

- Las funciones no lineales utilizadas, tanto para el ajuste de la función de supervivencia, como para el ajuste de rentabilidad vs. edad presentaron buenos resultados como se vió en el ajuste de la función de supervivencia del conglomerado 4 que se alcanzó un R^2 del 99% y en la que se nota claramente en el gráfico ser un ajuste casi perfecto.

- Por último, el desarrollo de modelos de segmentación de clientes son diversos en la actualidad y el llevar a cabo uno en específico sólo dependerá de la necesidad del negocio, las herramientas tecnológicas que se posean y el tiempo que se desee invertir en el mismo.

Apéndice A

Códigos fuente de implementación

A.1 Algoritmo de datos atípicos multivariantes

Para la ejecución del algoritmo de datos atípicos se realizó un pequeño ejecutable en MATLAB que a continuación se detalla:

```
function varargout = atipicos(varargin)
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @atipicos_OpeningFcn, ...
                   'gui_OutputFcn',  @atipicos_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end

global salida;
global data;
global carga;
global porc_atipicos

function atipicos_OpeningFcn(hObject, eventdata, handles, varargin)
handles.output = hObject;
guidata(hObject, handles);

function varargout = atipicos_OutputFcn(hObject, eventdata, handles)
```

```
 varargout{1} = handles.output;

function carga_Callback(hObject, eventdata, handles)

 global carga;
 [FileName,PathName] = uigetfile('*.dat','Cargar archivo BASE atipicos');
 carga = importdata(FileName);

function calcular_Callback(hObject, eventdata, handles)
 global salida;
 global carga;
 global porc_atipicos

 previo=carga.data;
 id=previo(:,1);
 data=previo(:,2:end-1);
 idx=main(data,1,porc_atipicos);

 salida=[id idx];

function save_Callback(hObject, eventdata, handles)
 global salida;
 uisave('salida','atipicos_resultados');

function porc_atipicos_Callback(hObject, eventdata, handles)

global porc_atipicos

porc_atipicos = str2double(get(hObject, 'String'));

if isnan(porc_atipicos)|porc_atipicos<0 |porc_atipicos>=10
    set(hObject, 'String', 0);
    errordlg(...
        'Ha ingresado una cantidad negativa, o tal vés su cantidad ha sobrepasado el r
        'Error al ingresar datos');
else
    set(handles.porc_atipicos, 'Enable','off','Value',porc_atipicos);
```

```
end


handles.metricdata.atipicos = porc_atipicos/100;
guidata(hObject,handles)


function porc_atipicos_CreateFcn(hObject, eventdata, handles)

if ispc && isequal(get(hObject,'BackgroundColor'), get(0,'defaultUicontrolBackgroundCo
    set(hObject,'BackgroundColor','white');
end
```

Y el archivo maestro para la ejecución de esta GUI es:

```
 function [idx] = main(x,mode,maximo_atipicos)

% Dimensión de matriz de datos
[n,p] = size(x);

% Dimensión de matriz de datos
if (n > 350000)|(p > 40),
  disp('Conjunto de datos muy grande para este código');
  return
end

if nargin < 2,
  mode = 0;
end
maxit = Inf;
if mode < 0,
  maxit = abs(mode);
  mode = -1;
elseif mode > 0,
  maxit = mode;
  mode = 1;
end
```

```
% Puntos de corte para las proyecciones
 ctf2 = 5.984 + 1.644*p;
 ctf1 = exp(3.752-1.3*log(p));
 ctf3 = 0.8;
 ctf4 = 2.0;

% mínimo número de observaciones a ser retiradas
cut=1-maximo_atipicos;
lmn1 = max(floor(cut*(n+p+1)),2*p);
lmn2 = min(2*p,max(p+2,floor(0.25*n)));

% removiendo atípicos sospechosos
xx0 = [x (1:n)'];
xx = xx0;
nn = n;
it = 1;

while (it <= maxit),
  xx1 = xx(:,1:p);
  V = dir_max_min_kur(xx1,mode);
  pr = xx1*V;
  md = median(pr);
  prn = abs(pr - ones(nn,1)*md);
  mad = median(prn);
  tt = prn./(ones(nn,1)*mad);

  if p > 2,
    ctfo = ctf1 - (0:(p-2))*0.75*(ctf1-ctf2)/(p-2);
  else
    ctfo = ctf1;
  end
  if mode >= 0,
    tctf = ones(nn,1)*[ ctfo ctf2 ctf3*ones(1,p-1) ctf4 ];
  else
    tctf = ones(nn,1)*[ ctfo ctf2 ];
  end
  taux = tt./tctf;
  t = max(taux')';
```

```
  in = find(t > 1);
  nn = length(in);
  if nn == 0,
    break
  end
  inn = find(t <= 1);
  nu = length(inn);

  if nu < lmn2,
    [v,ix] = sort(t);
    ixx = ix(1:lmn2);
    xx = xx(ixx,:);
    break
  end
  xx = xx(inn,:);
  [nn,pp] = size(xx);
  if nn <= lmn2,
    break
  end
  it = it + 1;
end

% Extracción de índices de atípicos
idx = ones(n,1);
[nn,pp] = size(xx);
idx(xx(:,pp)) = zeros(nn,1);

% Rechequeando las observaciones y etiquetándolas

% Puntos de corte para las distancias de Mahalanobis, basado en t2-Hotelling
% fijo a 0.99. Se puede cambiar este parámetro a continuación
ctflvl = 0.98;
ctf = (p*(n-1)/(n-p))*finv(ctflvl,p,n-p);

% Distancia de Mahalanobis usando estimadores de escala y centro
% basados en observaciones ''buenas''
sidx = sum(idx);
```

```
s1 = find(idx);
s2 = find(idx == 0);
if sidx > 0,
  xx1 = xx0(s1,:);
  xx1r = xx1(:,1:p);
end
xx2r = xx0(s2,1:p);

mm = mean(xx2r);
Ss = cov(xx2r);
if sidx > 0,
  aux1 = xx1r - ones(sidx,1)*mm;
  dd = Ss\aux1';
  dms = sum((aux1.*dd')');
end

% Fase de aseguramiento de que al menos lmn1 observaciones son
% consideradas como ''buenas''
ado = sidx + lmn1 - n;
if ado > 0,
  [dmss,idms] = sort(dms);
  idm1 = idms(1:ado);
  ido = xx0(s1,pp);
  s3 = ido(idm1);
  idx(s3) = zeros(ado,1);

  sidx = sum(idx);
  s1 = find(idx);
  s2 = find(idx == 0);
  if sidx > 0,
    xx1 = xx0(s1,:);
    xx1r = xx1(:,1:p);
  end
  xx2r = xx0(s2,1:p);

  mm = mean(xx2r);
  Ss = cov(xx2r);
  if sidx > 0,
```

```
    aux1 = xx1r - ones(sidx,1)*mm;
    dd = Ss\aux1';
    dms = sum((aux1.*dd')');
  end
end

% Chequeando observaciones restantes y reetiquetándolas si es apropiado
while sidx > 0,

  s1 = find(dms <= ctf);
  s2 = length(s1);
  if s2 == 0,
    break
  end
  s3 = xx1(s1,pp);
  idx(s3) = zeros(s2,1);
  sidx = sum(idx);

  s1 = find(idx);
  s2 = find(idx == 0);
  if sidx > 0,
    xx1 = xx0(s1,:);
    xx1r = xx1(:,1:p);
  end
  xx2r = xx0(s2,1:p);

  mm = mean(xx2r);
  Ss = cov(xx2r);
  if sidx > 0,
    aux1 = xx1r - ones(sidx,1)*mm;
    dd = Ss\aux1';
    dms = sum((aux1.*dd')'); %#ok<UDIM>
  end
end

% Valores de retorno
aux1 = x - ones(n,1)*mm;
dd = Ss\aux1';
```

```
dms = sum((aux1.*dd')');
dm = sqrt(dms);


 num_ati=sum(idx);

 porcentaje=(num_ati/n)*100;

 function Vv = dir_max_min_kur(x0,mode)

if nargin < 2,
  mode = 0;
end

% Inicializaciones

%% Parametros (tolerancias)

maxit = 100;

tol = 1.0e-5;
tol1 = 1.0e-6;
beta = 1.0e-4;
rho0 = 0.1;

[n,p0] = size(x0);

%% Initialization of vectors

Vv = [];
ti = 1;

%% Estandarizando los datos
mm = mean(x0);
S = cov(x0);
x = x0 - ones(n,1)*mm;

Rr = chol(S);
```

```
x = ((Rr')\(x'))';

% Direcciones de cálculo escogiendo el máximo o mínimo
cff = 1;
if mode < 0,
  cffmax = 2;
else
  cffmax = 3;
end

%% Principal loop para calcular las 2p direcciones
while (cff < cffmax),
  xx = x;
  p = p0;
  pin = p - 1;
  M = eye(p0);

  for i = 1:pin,
    if cff == 1,
      a = maxima_kurtosis(xx);
    else
      a = minima_kurtosis(xx);
    end
    la = length(a);
    za = zeros(la,1); za(1) = 1;
    w = a - za; nw = w'*a;
    if abs(nw) > eps,
      Q = eye(la) - w*w'/nw;
    else
      Q = eye(la);
    end

%% Cálculos de los valores proyectados
    Vv = [ Vv (M*a) ];
    Qp = Q(:,2:p);
    M = M*Qp;
    ti = ti + 1;
```

```
%% Reducción de dimensión
    Tt = xx*Q;
    xx = Tt(:,(2:p));

    p = p - 1;

  end

%% Calcular la última proyección

  Vv = [ Vv M ];
  ti = ti + 1;

%% Proceso de minimización
  cff = cff + 1;
end

% Deshacer la transformación de estandarización
Vv = Rr\Vv;
uaux = sum(Vv.*Vv);
Vv = Vv*diag(1../sqrt(uaux));

function [v,f] = maxima_kurtosis(x,d0,maxit)

maxitdefault = 30;


if nargin < 3,
  maxit = maxitdefault;
end
if maxit <= 0,
  maxit = maxitdefault;
end
if nargin < 2,
  d0 = [];
end

maxitini = 1;
```

```
tol = 1.0e-3;
tol1 = 1.0e-3;
tol2 = 1.0e-2;
beta = 1.0e-3;
rho0 = 0.1;


[n,p] = size(x);

%%  Estimación inicial de la dirección
if length(d0) == 0,
 eps = 0.01;
  uv = sum((x.*x)')';
  uw = 1../(eps + sqrt(uv));
  uu = x.*(uw*ones(1,p));

     Su = cov(uu);
  [V,D] = eig(Su);

  r = [];
  for i = 1:p,
    r = [ r coef_kurtosis(x,V(:,i)) ];
  end

  [v,ik] = max(r);
  a = V(:,ik(1));

  itini = 1;
  difa = 1;
  while (itini <= maxitini)&(difa > tol2),
    z = x*a;
    zaux = z.^2;
    xaux = x'.*(ones(p,1)*zaux');
    H = 12*xaux*x;
    [V,E] = eig(H);
    [vv,iv] = max(diag(E));
    aa = V(:,iv(1));
    difa = norm(a - aa);
```

```
    a = aa;
    itini = itini + 1;
  end
else
  a = d0/norm(d0);
end

%% Valores en la iteración 0 para la optimización del algoritmo
z = x*a;
sk = sum(z.^4);
lam = 2*sk;
f = sk;
g = (4*(z.^3)'*x)';
zaux = z.^2;
xaux = x'.*(ones(p,1)*zaux');
H = 12*xaux*x;

al = 0;
it = 0;
diff = 1;
rho = rho0;
clkr = 0;

c = 0;

while 1,

% Chequeo de las condiciones de terminación
  gl = g - 2*lam*a;

  A = 2*a';
  [Q,W] = qr(a);
  Z = Q(:,(2:p));

  crit = norm(gl) + abs(c);
  if (crit <= tol)|(it >= maxit),
    break
  end
```

```
%% Calcular la dirección de búsqueda
  Hl = H - 2*lam*eye(p,p);
  Hr = Z'*Hl*Z;
  [V,E] = eig(Hr);
  Es = min(-abs(E),-1.0e-4);
  Hs = V*Es*V';

  py = - c/(A*A');
  rhs = Z'*(g + H*A'*py);
  pz = - Hs\rhs;
  pp = Z*pz + py*A';

  dlam = (2*a)\(gl + H*pp);

%% Ajuste al parámetro de penalidad
  f0d = gl'*pp - 2*rho*c*a'*pp - dlam*c;
  crit1 = beta*norm(pp)^2;

  if f0d < crit1,
    rho1 = 2*(crit1 - f0d)/(eps + c^2);
    rho = max([2*rho1 1.5*rho rho0]);
    f = sk - lam*c - 0.5*rho*c^2;
    f0d = gl'*pp - 2*rho*c*a'*pp - dlam*c;
    clkr = 0;
  elseif (f0d > 1000*crit1)&(rho > rho0),
    rho1 = 2*(crit1 - gl'*pp + dlam*c)/(eps + c^2);
    if (clkr == 4)&(rho > 2*rho1),
      rho = 0.5*rho;
      f = sk - lam*c - 0.5*rho*c^2;
      f0d = gl'*pp - 2*rho*c*a'*pp - dlam*c;
      clkr = 0;
    else
      clkr = clkr + 1;
    end
  end
  if (abs(f0d/(norm(g-2*rho*a*c)+norm(c))) < tol1),
    break
```

```
  end

%% Línea de búsqueda
  al = 1;
  itbl = 0;
  while itbl < 20,
    aa = a + al*pp;
    lama = lam + al*dlam;
    zz = x*aa;
    cc = aa'*aa - 1;
    sk = sum(zz.^4);
    ff = sk - lama*cc - 0.5*rho*cc^2;

    if ff > f + 0.0001*al*f0d,
      break
    end
    al = al/2;
    itbl = itbl + 1;
  end
  if itbl >= 20,
    break
  end

%% Actualización de valores para la próxima iteración
  a = aa;
  lam = lama;
  z = zz;

  nmd2 = a'*a;
  c = nmd2 - 1;
  f = sk - lam*c - 0.5*rho*c^2;
  g = (4*(z.^3)'*x)';
  zaux = z.^2;
  xaux = x'.*(ones(p,1)*zaux');
  H = 12*xaux*x;

  it = it + 1;
```

```
end

% Valores de retorno
v = a/norm(a);
xa = x*v;
f = sum(xa.^4)/n;

function [v,f] = minima_kurtosis(x,d0,maxit)

% Inicialización
maxitdefault = 30; % valor máximo de iteraciones

if nargin < 3,
  maxit = maxitdefault;
end
if maxit <= 0,
  maxit = maxitdefault;
end
if nargin < 2,
  d0 = [];
end

%% Tolerancias
maxitini = 1;
tol = 1.0e-4;
tol1 = 1.0e-7;
tol2 = 1.0e-2;
beta = 1.0e-4;
rho0 = 0.1;

[n,p] = size(x);

%% Estimación de la dirección inicial
if length(d0) == 0,

  uv = sum((x.*x)')';
  uw = 1../(eps + sqrt(uv));
  uu = x.*(uw*ones(1,p));
```

```
    Su = cov(uu);
  [V,D] = eig(Su);

  r = [];
  for i = 1:p,
    r = [ r coef_kurtosis(x,V(:,i)) ];
  end
  [v,ik] = min(r);
  a = V(:,ik(1));

  itini = 1;
  difa = 1;
  while (itini <= maxitini)&(difa > tol2),
    z = x*a;
    zaux = z.^2;
    xaux = x'.*(ones(p,1)*zaux');
    H = 12*xaux*x;
    [V,E] = eig(H);
    [vv,iv] = min(diag(E));
    aa = V(:,iv(1));
    difa = norm(a - aa);
    a = aa;
    itini = itini + 1;
  end
else
  a = d0/norm(d0);
end

%% Valores en la iteración 0 para la optimización del algoritmo
z = x*a;
sk = sum(z.^4);
lam = 2*sk;
f = sk;
g = (4*(z.^3)'*x)';
zaux = z.^2;
xaux = x'.*(ones(p,1)*zaux');
H = 12*xaux*x;
```

```
al = 0;
it = 0;
diff = 1;
rho = rho0;
clkr = 0;

c = 0;

while 1,
%% Chequeo de las condiciones de terminación
  gl = g - 2*lam*a;
  A = 2*a';
  [Q,W] = qr(a);
  Z = Q(:,(2:p));
  crit = norm(gl) + abs(c);
  if (crit <= tol)|(it >= maxit),
    break
  end

%% Cálculo de la dirección de búsqueda
  Hl = H - 2*lam*eye(p,p);
  Hr = Z'*Hl*Z;
  [V,E] = eig(Hr);
  Es = max(abs(E),1.0e-4);
  Hs = V*Es*V';
  py = - c/(A*A');
  rhs = Z'*(g + H*A'*py);
  pz = - Hs\rhs;
  pp = Z*pz + py*A';
  dlam = (2*a)\(gl + H*pp);

%% Ajuste al parámetro de penalidad
  f0d = gl'*pp + 2*rho*c*a'*pp - dlam*c;
  crit1 = beta*norm(pp)^2;
  if f0d > -crit1,
    rho1 = 2*(crit1 + f0d)/(eps + c^2);
    rho = max([2*rho1 1.5*rho rho0]);
```

```
    f = sk - lam*c + 0.5*rho*c^2;
    f0d = gl'*pp + 2*rho*c*a'*pp - dlam*c;
    clkr = 0;
  elseif (f0d < -1000*crit1)&(rho > rho0),
    rho1 = 2*(crit1 - gl'*pp + dlam*c)/(eps + c^2);
    if (clkr == 4)&(rho > 2*rho1),
      rho = 0.5*rho;
      f = sk - lam*c + 0.5*rho*c^2;
      f0d = gl'*pp + 2*rho*c*a'*pp - dlam*c;
      clkr = 0;
    else
      clkr = clkr + 1;
    end
  end
  if (abs(f0d) < tol1),
    break
  end

%% Línea de búsqueda
  al = 1;
  itbl = 0;
  while itbl < 20,
    aa = a + al*pp;
    lama = lam + al*dlam;
    zz = x*aa;
    cc = aa'*aa - 1;
    sk = sum(zz.^4);
    ff = sk - lama*cc + 0.5*rho*cc^2;
    if ff < f + 0.0001*al*f0d,
      break
    end
    al = al/2;
    itbl = itbl + 1;
  end
  if itbl >= 20,
    break
  end
```

```
%% Actualización de valores para la próxima iteración
  a = aa;
  lam = lama;
  z = zz;
  nmd2 = a'*a;
  c = nmd2 - 1;
  f = sk - lam*c + 0.5*rho*c^2;
  g = (4*(z.^3)'*x)';
  zaux = z.^2;
  xaux = x'.*(ones(p,1)*zaux');
  H = 12*xaux*x;
  it = it + 1;

end

% Valores de retorno
v = a/norm(a);
xa = x*v;
f = sum(xa.^4)/n;

function mc = coef_kurtosis(x,d,km)

if nargin < 3,
  km = 4;
end

[n,p] = size(x);
[p1,p2] = size(d);
if p ~= p1,
  disp('La dimensión de los datos no es correcta');
  return
end

t = x*d;
tm = mean(t);
tt = abs(t - tm);
vr = sum(tt.^2)/(n-1);
kr = sum(tt.^km)/n;
```

```
mc = kr/vr^(km/2);
```

A.2 Implementación del LT

Para el cálculo del LT se programaron las siguientes funciones

A.2.1 Conglomerado 1

```
clc;
 paso=input(' Ingrese el numero de nodos en la malla n : ');

  sur  = data(:,2);
  xbe  = data(:,4);
 At_1 = data(:,5);

 sur_2 = sur(~isnan(sur));

 buenos = length(sur_2);
   nn_2 = length(sur);

   if buenos <nn_2
       for i=(buenos+1):nn_2
           sur(i)=1;
           xbe(i)=0;
       end
   end

 So_1 = sur.^(1./exp(xbe));

% xbe por clusters
xbe_1 = xbe;
clear xbe;
clear buenos;

% General model:
%        f(x) = ((sin(x/a)-x/a+2*pi)/(2*pi))^b
% Coefficients (with 95% confidence bounds):
```

```
       a =       56.13; %  (48.59, 63.67)
       b =      0.4171; %  (0.2561, 0.5782)

% Goodness of fit:
%   SSE: 7.135
%   R-square: 0.9593
%   Adjusted R-square: 0.9593
%   RMSE: 0.01407

    ind_cl1  = length(At_1);
       LT_1  = sparse(ind_cl1,1);
  val_xbe_1  = xbe_1(~isnan(xbe_1));
      sur_1  = sur(~isnan(sur));
      num_b  = length(val_xbe_1);
        fin  = 2*pi*a;
      clear xbe_1;
      clear ind_cl1;
      clear sur

for i=1:num_b
              h = (fin-At_1(i))/paso;
              x = At_1(i):h:fin;
             ff = ((sin(x/a)-(x/a)+2*pi)/(2*pi)).^b;
    f_integrando= ff.^(exp(val_xbe_1(i)));
             g = f_integrando;

    clear x;
    clear ff;

      %Cálculo de la integral por el método del trapecio compuesto
      for j=2:paso
          f_integrando(j) = 2*g(j);
      end

    clear j;
    clear g;

    LT_1(i)=((h/2)*(sum(f_integrando)))/sur_1(i);
```

```
    clear f_integrando;
end

    clear a
    clear b
    clear At_1
    clear i
    clear num_b
    clear val_xbe_1
    clear fin
    clear h
    clear So
    clear sur_1

%************************ Resultados *****************
 total_LT_1 = full(LT_1);
 clear LT_1
 clear nn
 clear nn_2
```

A.2.2 Conglomerado 2

```
clc;
 paso=input(' Ingrese el numero de nodos en la malla n : ');

  sur  = data(:,2);
  xbe  = data(:,4);
 At_1 = data(:,5);

 sur_2 = sur(~isnan(sur));

 buenos = length(sur_2);
   nn_2 = length(sur);

   if buenos <nn_2
       for i=(buenos+1):nn_2
           sur(i)=1;
```

```
           xbe(i)=0;
       end
   end

 So_1 = sur.^(1./exp(xbe));

% xbe por clusters
xbe_1 = xbe;
clear xbe;
clear buenos;

% General model:
%        f(x) = ((sin(x/a)-x/a+2*pi)/(2*pi))^b
% Coefficients (with 95% confidence bounds):
       a =       55.84; %  (47.74, 63.94)
       b =      0.5206; %  (0.3036, 0.7377)

% Goodness of fit:
%   SSE: 14.13
%   R-square: 0.9435
%   Adjusted R-square: 0.9435
%   RMSE: 0.01893

    ind_cl1  = length(At_1);
       LT_1  = sparse(ind_cl1,1);
  val_xbe_1  = xbe_1(~isnan(xbe_1));
      sur_1  = sur(~isnan(sur));
      num_b  = length(val_xbe_1);
        fin  = 2*pi*a;
      clear xbe_1;
      clear ind_cl1;
      clear sur

for i=1:num_b
              h = (fin-At_1(i))/paso;
              x = At_1(i):h:fin;
             ff = ((sin(x/a)-(x/a)+2*pi)/(2*pi)).^b;
    f_integrando= ff.^(exp(val_xbe_1(i)));
```

```
            g = f_integrando;

    clear x;
    clear ff;

      %Cálculo de la integral por el método del trapecio compuesto
      for j=2:paso
          f_integrando(j) = 2*g(j);
      end

    clear j;
    clear g;

    LT_1(i)=((h/2)*(sum(f_integrando)))/sur_1(i);

    clear f_integrando;
end

    clear a
    clear b
    clear At_1
    clear i
    clear num_b
    clear val_xbe_1
    clear fin
    clear h
    clear So
    clear sur_1

%************************ Resultados *****************
 total_LT_1 = full(LT_1);
 clear LT_1
 clear nn
 clear nn_2
```

A.2.3 Conglomerado 3

```
clc;
```

```
 paso=input(' Ingrese el numero de nodos en la malla n : ');

  sur  = data(:,2);
  xbe  = data(:,4);
 At_1 = data(:,5);

 sur_2 = sur(~isnan(sur));

 buenos = length(sur_2);
   nn_2 = length(sur);

   if buenos <nn_2
       for i=(buenos+1):nn_2
           sur(i)=1;
           xbe(i)=0;
       end
   end

 So_1 = sur.^(1./exp(xbe));

% xbe por clusters
xbe_1 = xbe;
clear xbe;
clear buenos;

% General model:
%        f(x) = ((sin(x/a)-x/a+2*pi)/(2*pi))^b
% Coefficients (with 95% confidence bounds):
       a =       49.51; %  (42.7, 56.32)
       b =       1.224; %  (0.7404, 1.708)

% Goodness of fit:
%   SSE: 89.48
%   R-square: 0.7611
%   Adjusted R-square: 0.7611
%   RMSE: 0.04177

    ind_cl1  = length(At_1);
```

```
       LT_1  = sparse(ind_cl1,1);
  val_xbe_1  = xbe_1(~isnan(xbe_1));
      sur_1  = sur(~isnan(sur));
      num_b  = length(val_xbe_1);
        fin  = 2*pi*a;
      clear xbe_1;
      clear ind_cl1;
      clear sur

for i=1:num_b
              h = (fin-At_1(i))/paso;
              x = At_1(i):h:fin;
             ff = ((sin(x/a)-(x/a)+2*pi)/(2*pi)).^b;
    f_integrando= ff.^(exp(val_xbe_1(i)));
             g = f_integrando;

    clear x;
    clear ff;

      %Cálculo de la integral por el método del trapecio compuesto
      for j=2:paso
          f_integrando(j) = 2*g(j);
      end

    clear j;
    clear g;

    LT_1(i)=((h/2)*(sum(f_integrando)))/sur_1(i);

    clear f_integrando;
end

    clear a
    clear b
    clear At_1
    clear i
    clear num_b
    clear val_xbe_1
```

```
    clear fin
    clear h
    clear So
    clear sur_1

%************************ Resultados ****************
 total_LT_1 = full(LT_1);
 clear LT_1
 clear nn
 clear nn_2
```

A.2.4 Conglomerado 4

```
clc;
 paso=input(' Ingrese el numero de nodos en la malla n : ');

  sur  = data(:,2);
  xbe  = data(:,4);
 At_1 = data(:,5);

 sur_2 = sur(~isnan(sur));

 buenos = length(sur_2);
   nn_2 = length(sur);

   if buenos <nn_2
       for i=(buenos+1):nn_2
           sur(i)=1;
           xbe(i)=0;
       end
   end

 So_1 = sur.^(1./exp(xbe));

% xbe por clusters
xbe_1 = xbe;
clear xbe;
clear buenos;
```

```
% General model:
%        f(x) = ((sin(x/a)-x/a+2*pi)/(2*pi))^b
% Coefficients (with 95% confidence bounds):
       a =       56.38; %  (53.67, 59.09)
       b =       1.374; %  (1.185, 1.564)

% Goodness of fit:
%   SSE: 1.866
%   R-square: 0.9974
%   Adjusted R-square: 0.9974
%   RMSE: 0.009605

    ind_cl1  = length(At_1);
       LT_1  = sparse(ind_cl1,1);
  val_xbe_1  = xbe_1(~isnan(xbe_1));
      sur_1  = sur(~isnan(sur));
      num_b  = length(val_xbe_1);
        fin  = 2*pi*a;
      clear xbe_1;
      clear ind_cl1;
      clear sur

for i=1:num_b
              h = (fin-At_1(i))/paso;
              x = At_1(i):h:fin;
             ff = ((sin(x/a)-(x/a)+2*pi)/(2*pi)).^b;
    f_integrando= ff.^(exp(val_xbe_1(i)));
             g = f_integrando;

    clear x;
    clear ff;

      %Cálculo de la integral por el método del trapecio compuesto
      for j=2:paso
          f_integrando(j) = 2*g(j);
      end
```

```
    clear j;
    clear g;

    LT_1(i)=((h/2)*(sum(f_integrando)))/sur_1(i);

    clear f_integrando;
end

    clear a
    clear b
    clear At_1
    clear i
    clear num_b
    clear val_xbe_1
    clear fin
    clear h
    clear So
    clear sur_1

%************************ Resultados *****************
 total_LT_1 = full(LT_1);
 clear LT_1
 clear nn
 clear nn_2
```

A.3 Implementación del LTV

Para el cálculo del LT se programaron las siguientes funciones

A.3.1 Conglomerado 1

```
clc;
% Life Time Value nicho
porc = input(' Ingrese la tasa j para el cálculo de LTV: ');

% Orden de variables en el archivo
% 1. id_cliente
```

```
% 2. edad_anios
% 3. lt_meses
% 4. suma_rentabilidad

  lt = data(:,3);
  lt = lt(~isnan(lt));
  lt = round(lt/12);

buenos = length(lt);

edad_cliente = data(:,2);
edad_cliente = edad_cliente(1:buenos);

% General model Gauss8:
%        f(x) =
%               a1*exp(-((x-b1)/c1)^2) + a2*exp(-((x-b2)/c2)^2) +
%               a3*exp(-((x-b3)/c3)^2) + a4*exp(-((x-b4)/c4)^2) +
%               a5*exp(-((x-b5)/c5)^2) + a6*exp(-((x-b6)/c6)^2) +
%               a7*exp(-((x-b7)/c7)^2) + a8*exp(-((x-b8)/c8)^2)
% Coefficients (with 95% confidence bounds):
       a1 =       73.31; %  (59.57, 87.05)
       b1 =       47.32; %  (36.41, 58.22)
       c1 =       16.21; %  (-7.05, 39.46)
       a2 =       35.37; %  (-48.35, 119.1)
       b2 =       62.13; %  (57.15, 67.1)
       c2 =       5.457; %  (-14.34, 25.25)
       a3 =       16.79; %  (-87.71, 121.3)
       b3 =       54.36; %  (44.47, 64.25)
       c3 =       4.121; %  (-5.183, 13.43)
       a4 = -2.728e+006; %  (-7.349e+022, 7.349e+022)
       b4 =       39.67; %  (-2.083e+015, 2.083e+015)
       c4 =     0.08408; %  (-5.287e+014, 5.287e+014)
       a5 =       16.08; %  (-110.9, 143.1)
       b5 =       68.06; %  (62.88, 73.23)
       c5 =       3.311; %  (-4.275, 10.9)
       a6 =       26.15; %  (-18.65, 70.95)
       b6 =       26.44; %  (23.29, 29.59)
       c6 =       8.003; %  (2.393, 13.61)
```

```
       a7 =       62.42; %  (28.11, 96.73)
       b7 =       76.58; %  (71.41, 81.75)
       c7 =       10.43; %  (-4.97, 25.82)
       a8 =       40.99; %  (-27.06, 109)
       b8 =       88.05; %  (86.07, 90.03)
       c8 =       4.757; %  (-0.8613, 10.37)

% Goodness of fit:
%   SSE: 477.3
%   R-square: 0.9787
%   Adjusted R-square: 0.9682
%   RMSE: 3.187

LTV = sparse(buenos,1);

% Para la parte del factor tomemos tabla como matriz de entrada (data)
% Orden de datos en la matriz tabla
% 1. Edad (tabla(:,1))
% 2. Rentabilidad Global ...
% 3. Rentabilidad promedio ...
% 4. Número de clientes ...

edad_vc = tabla(:,1);

 for i=1: buenos
          fin = edad_cliente(i)+lt(i);
            x = edad_cliente(i)+1:1:fin;
            a = ones(1,length(x));
            R = a1*exp(-((x-b1*a)./(c1*a)).^2) + ...
                a2*exp(-((x-b2*a)./(c2*a)).^2) + ...
                a3*exp(-((x-b3*a)./(c3*a)).^2) + ...
                a4*exp(-((x-b4*a)./(c4*a)).^2) + ...
                a5*exp(-((x-b5*a)./(c5*a)).^2) + ...
                a6*exp(-((x-b6*a)./(c6*a)).^2) + ...
                a7*exp(-((x-b7*a)./(c7*a)).^2)+ ...
                a8*exp(-((x-b8*a)./(c8*a)).^2) ;

     part_LTV = zeros(lt(i),1);
```

```
     clear x;
     clear a;

    for j=1:lt(i)
        fila = find(edad_vc == edad_cliente(i)+j);

        if isempty(fila)==1 | tabla(fila,4)==0

            part_LTV(j) = R(j)/((1+porc)^j);
        else
            part_LTV(j) = (tabla(fila,3))/((1+porc)^j);
        end
    end

      LTV(i) = sum(part_LTV);


        index = find(edad_vc == edad_cliente(i));

        if isempty(index)==1|tabla(index,4)==0
            factor= (data(i,4))/R(j);
        else
        factor = (data(i,4))/(tabla(index,3));

        end

        LTV(i) = factor*LTV(i);

      clear index;
      clear factor;
      clear part_LTV;
      clear fin;
      clear R;
      clear i;
      clear j;
end
```

A.3.2 Conglomerado 2

```
clc;
% Life Time Value nicho
porc = input(' Ingrese la tasa j para el cálculo de LTV: ');

% Orden de variables en el archivo
% 1. id_cliente
% 2. edad_anios
% 3. lt_meses
% 4. suma_rentabilidad

  lt = data(:,3);
  lt = lt(~isnan(lt));
  lt = round(lt/12);

buenos = length(lt);

edad_cliente = data(:,2);
edad_cliente = edad_cliente(1:buenos);

% General model Gauss7:
%        f(x) =
%               a1*exp(-((x-b1)/c1)^2) + a2*exp(-((x-b2)/c2)^2) +
%               a3*exp(-((x-b3)/c3)^2) + a4*exp(-((x-b4)/c4)^2) +
%               a5*exp(-((x-b5)/c5)^2) + a6*exp(-((x-b6)/c6)^2) +
%               a7*exp(-((x-b7)/c7)^2)
% Coefficients (with 95% confidence bounds):
       a1 =        28.4; %  (18.88, 37.91)
       b1 =       60.64; %  (60.1, 61.18)
       c1 =       5.758; %  (4.52, 6.995)
       a2 =       5.869; %  (1.62, 10.12)
       b2 =       84.41; %  (83.18, 85.63)
       c2 =       2.428; %  (0.1157, 4.741)
       a3 =       7.622; %  (3.494, 11.75)
       b3 =       68.54; %  (67.74, 69.34)
       c3 =       2.246; %  (0.8706, 3.622)
       a4 =       83.97; %  (81.23, 86.7)
```

```
        b4 =       77.57; %  (76.25, 78.9)
        c4 =       16.47; %  (14.06, 18.87)
        a5 =       24.42; %  (17.26, 31.58)
        b5 =       33.98; %  (32.04, 35.93)
        c5 =       5.075; %  (2.909, 7.242)
        a6 =       78.28; %  (74.26, 82.3)
        b6 =       47.88; %  (46.29, 49.46)
        c6 =       14.99; %  (12.71, 17.26)
        a7 =       24.34; %  (10.13, 38.55)
        b7 =       28.18; %  (27.3, 29.07)
        c7 =        3.32; %  (2.329, 4.311)

% Goodness of fit:
%   SSE: 140.4
%   R-square: 0.996
%   Adjusted R-square: 0.9943
%   RMSE: 1.728

LTV = sparse(buenos,1);

% Para la parte del factor tomemos tabla como matriz de entrada (data)
% Orden de datos en la matriz tabla
% 1. Edad (tabla(:,1))
% 2. Rentabilidad Global ...
% 3. Rentabilidad promedio ...
% 4. Número de clientes ...

edad_vc = tabla(:,1);

 for i=1: buenos
          fin = edad_cliente(i)+lt(i);
            x = edad_cliente(i)+1:1:fin;
            a = ones(1,length(x));
            R = a1*exp(-((x-b1*a)./(c1*a)).^2) + ...
                a2*exp(-((x-b2*a)./(c2*a)).^2) + ...
                a3*exp(-((x-b3*a)./(c3*a)).^2) + ...
                a4*exp(-((x-b4*a)./(c4*a)).^2) + ...
                a5*exp(-((x-b5*a)./(c5*a)).^2) + ...
```

```
            a6*exp(-((x-b6*a)./(c6*a)).^2) + ...
            a7*exp(-((x-b7*a)./(c7*a)).^2) ;

part_LTV = zeros(lt(i),1);

 clear x;
 clear a;

for j=1:lt(i)
    fila = find(edad_vc == edad_cliente(i)+j);

    if isempty(fila)==1 | tabla(fila,4)==0

        part_LTV(j) = R(j)/((1+porc)^j);
    else
        part_LTV(j) = (tabla(fila,3))/((1+porc)^j);
    end
end

  LTV(i) = sum(part_LTV);


    index = find(edad_vc == edad_cliente(i));

    if isempty(index)==1|tabla(index,4)==0
        factor= (data(i,4))/R(j);
    else
    factor = (data(i,4))/(tabla(index,3));

    end

    LTV(i) = factor*LTV(i);

  clear index;
  clear factor;
  clear part_LTV;
  clear fin;
  clear R;
```

```
       clear i;
       clear j;
 end

 LTV_nicho = full(LTV);
```

A.3.3 Conglomerado 3

```
clc;
% Life Time Value nicho
porc = input(' Ingrese la tasa j para el cálculo de LTV: ');

% Orden de variables en el archivo
% 1. id_cliente
% 2. edad_anios
% 3. lt_meses
% 4. suma_rentabilidad

  lt = data(:,3);
  lt = lt(~isnan(lt));
  lt = round(lt/12);

buenos = length(lt);

edad_cliente = data(:,2);
edad_cliente = edad_cliente(1:buenos);

% General model Gauss7:
%        f(x) =
%               a1*exp(-((x-b1)/c1)^2) + a2*exp(-((x-b2)/c2)^2) +
%               a3*exp(-((x-b3)/c3)^2) + a4*exp(-((x-b4)/c4)^2) +
%               a5*exp(-((x-b5)/c5)^2) + a6*exp(-((x-b6)/c6)^2) +
%               a7*exp(-((x-b7)/c7)^2)
% Coefficients (with 95% confidence bounds):
       a1 =       967.4; %  (-4.492e+010, 4.492e+010)
       b1 =       58.49; %  (-1.557e+005, 1.558e+005)
       c1 =      0.3082; %  (-2.719e+006, 2.719e+006)
       a2 =       69.62; %  (58.32, 80.93)
```

```
        b2 =       55.41; %  (55, 55.82)
        c2 =       1.996; %  (1.027, 2.966)
        a3 =       23.24; %  (-652.3, 698.8)
        b3 =       17.05; %  (-32.33, 66.43)
        c3 =       1.483; %  (-23.36, 26.33)
        a4 =       49.54; %  (38.56, 60.52)
        b4 =       50.99; %  (50.37, 51.62)
        c4 =       2.081; %  (1.187, 2.974)
        a5 =        66.2; %  (47.23, 85.18)
        b5 =       66.28; %  (65.26, 67.3)
        c5 =       5.095; %  (3.77, 6.42)
        a6 =  2.482e+004; %  (-8.731e+019, 8.731e+019)
        b6 =        46.5; %  (-2.433e+012, 2.433e+012)
        c6 =      0.1887; %  (-4.728e+013, 4.728e+013)
        a7 =       45.56; %  (40.74, 50.37)
        b7 =       36.33; %  (35.3, 37.37)
        c7 =       10.83; %  (9.098, 12.55)

% Goodness of fit:
%   SSE: 1043
%   R-square: 0.9438
%   Adjusted R-square: 0.9006
%   RMSE: 6.333


LTV = sparse(buenos,1);

% Para la parte del factor tomemos tabla como matriz de entrada (data)
% Orden de datos en la matriz tabla
% 1. Edad (tabla(:,1))
% 2. Rentabilidad Global ...
% 3. Rentabilidad promedio ...
% 4. Número de clientes ...

edad_vc = tabla(:,1);

 for i=1: buenos
          fin = edad_cliente(i)+lt(i);
```

```
       x = edad_cliente(i)+1:1:fin;
       a = ones(1,length(x));
       R = a1*exp(-((x-b1*a)./(c1*a)).^2) + ...
           a2*exp(-((x-b2*a)./(c2*a)).^2) + ...
           a3*exp(-((x-b3*a)./(c3*a)).^2) + ...
           a4*exp(-((x-b4*a)./(c4*a)).^2) + ...
           a5*exp(-((x-b5*a)./(c5*a)).^2) + ...
           a6*exp(-((x-b6*a)./(c6*a)).^2) + ...
           a7*exp(-((x-b7*a)./(c7*a)).^2) ;

part_LTV = zeros(lt(i),1);

 clear x;
 clear a;

for j=1:lt(i)
    fila = find(edad_vc == edad_cliente(i)+j);

    if isempty(fila)==1 | tabla(fila,4)==0

        part_LTV(j) = R(j)/((1+porc)^j);
    else
        part_LTV(j) = (tabla(fila,3))/((1+porc)^j);
    end
end

  LTV(i) = sum(part_LTV);


    index = find(edad_vc == edad_cliente(i));

    if isempty(index)==1|tabla(index,4)==0
        factor= (data(i,4))/R(j);
    else
    factor = (data(i,4))/(tabla(index,3));

    end
```

```
        LTV(i) = factor*LTV(i);

      clear index;
      clear factor;
      clear part_LTV;
      clear fin;
      clear R;
      clear i;
      clear j;
end

LTV_nicho = full(LTV);
```

A.3.4 Conglomerado 4

```
clc;
% Life Time Value nicho
porc = input(' Ingrese la tasa j para el cálculo de LTV: ');

% Orden de variables en el archivo
% 1. id_cliente
% 2. edad_anios
% 3. lt_meses
% 4. suma_rentabilidad

  lt = data(:,3);
  lt = lt(~isnan(lt));
  lt = round(lt/12);

buenos = length(lt);

edad_cliente = data(:,2);
edad_cliente = edad_cliente(1:buenos);

% General model Gauss5:
%        f(x) =
%              a1*exp(-((x-b1)/c1)^2) + a2*exp(-((x-b2)/c2)^2) +
%              a3*exp(-((x-b3)/c3)^2) + a4*exp(-((x-b4)/c4)^2) +
```

```
%               a5*exp(-((x-b5)/c5)^2)
% Coefficients (with 95% confidence bounds):
       a1 =       66.99; %  (36.64, 97.34)
       b1 =       86.16; %  (84.76, 87.57)
       c1 =       4.399; %  (1.375, 7.423)
       a2 =      -17.83; %  (-37.8, 2.144)
       b2 =       67.23; %  (64.63, 69.84)
       c2 =       3.418; %  (-1.129, 7.966)
       a3 =      -12.92; %  (-43.69, 17.86)
       b3 =       54.76; %  (52.87, 56.66)
       c3 =      0.9196; %  (-1.767, 3.606)
       a4 =       20.81; %  (-4.58, 46.2)
       b4 =       74.15; %  (72.37, 75.93)
       c4 =       1.934; %  (-0.9734, 4.842)
       a5 =       195.4; %  (187.3, 203.4)
       b5 =       58.77; %  (53.74, 63.8)
       c5 =       49.77; %  (42.22, 57.32)

% Goodness of fit:
%   SSE: 9131
%   R-square: 0.8417
%   Adjusted R-square: 0.8014
%   RMSE: 12.88

LTV = sparse(buenos,1);

% Para la parte del factor tomemos tabla como matriz de entrada (data)
% Orden de datos en la matriz tabla
% 1. Edad (tabla(:,1))
% 2. Rentabilidad Global ...
% 3. Rentabilidad promedio ...
% 4. Número de clientes ...

edad_vc = tabla(:,1);

 for i=1: buenos
          fin = edad_cliente(i)+lt(i);
            x = edad_cliente(i)+1:1:fin;
```

```
       a = ones(1,length(x));
       R = a1*exp(-((x-b1*a)./(c1*a)).^2) + ...
           a2*exp(-((x-b2*a)./(c2*a)).^2) + ...
           a3*exp(-((x-b3*a)./(c3*a)).^2) + ...
           a4*exp(-((x-b4*a)./(c4*a)).^2) + ...
           a5*exp(-((x-b5*a)./(c5*a)).^2)  ;

part_LTV = zeros(lt(i),1);

 clear x;
 clear a;

for j=1:lt(i)
    fila = find(edad_vc == edad_cliente(i)+j);

    if isempty(fila)==1 | tabla(fila,4)==0

        part_LTV(j) = R(j)/((1+porc)^j);
    else
        part_LTV(j) = (tabla(fila,3))/((1+porc)^j);
    end
end

  LTV(i) = sum(part_LTV);


    index = find(edad_vc == edad_cliente(i));

    if isempty(index)==1|tabla(index,4)==0
        factor= (data(i,4))/R(j);
    else
    factor = (data(i,4))/(tabla(index,3));

    end

    LTV(i) = factor*LTV(i);

  clear index;
```

```
      clear factor;
      clear part_LTV;
      clear fin;
      clear R;
      clear i;
      clear j;
end

LTV_nicho = full(LTV);
```

A.3.5 Tabla Rentabilidad promedio vs. edad

Todos ellos necesitan de una tabla promedio de rentabilidad por edad para poder realizar los ajuste de cueva Rentabilidad promedio vs. edad

```
 % Cálculo de la tabla de promedio de rentabilidad por edad
% Orden para este archivo
% 1. id_cliente
% 2. edad
% 3. lt_meses
% 4. cl_suma_rentabilidad
% 5. cl_antiguedad_anios

edad = data(:,2);
edad = edad(~isnan(edad));

min_edad = min(edad);
max_edad = max(edad);

long_tabla = (max_edad - min_edad)+1;

tabla = zeros(long_tabla,4);

 for i=1:long_tabla
       ref_edad = min_edad + (i-1);
     tabla(i,1) = ref_edad;

             index = find(edad == ref_edad);
```

```
    tot_clientes = length(index);

       renta_global = 0;

       for j=1:tot_clientes
           renta_global = renta_global + data(index(j),4);
       end

    clear index;
    clear j;

    tabla(i,2) = renta_global;

    if tot_clientes == 0
        tabla(i,3) = 0;
    else
        tabla(i,3) = renta_global/tot_clientes;
    end


    tabla(i,4) = tot_clientes;

    clear tot_clientes;
    clear renta_global;
    clear ref_edad;

end


edad = tabla(:,1);
rentabilidad_promedio = tabla(:,3);

tot_clientes_edad=tabla(:,4);

search=find(tot_clientes_edad==0);
tabla(search,:)=[];

clear i;
```

```
clear min_edad;
clear max_edad;
clear long_tabla;
clear search;
clear tot_clientes_edad;

% Orden de la salida
% 1. edad
% 2. rentabilidad global
% 3. rentabilidad promedio
% 4. cantidad de clientes en esa edad
```

BIBLIOGRAFÍA

[1] Abascal E., Grande I. *Métodos Multivariantes para la investigación comercial*, Editorial Ariel, S.A., Barcelona, 1989.

[2] Allison P., *Survival Analysis using SAS: A practical guide*, SAS Institute Inc., USA, 2004.

[3] Ayres F., *Matemáticas Financieras*, McGraw/Hill, México, 1979.

[4] Bowers, *Actuarial Mathematics, Society of Actuaries*, USA, 1986.

[5] Collica R., *CRM Segmentation and Clustering Using SAS Enterprise Miner*, SAS Institute Inc., USA, 2007.

[6] Cuadras C., *Nuevos Métodos de Análisis Multivariantes*, CMC Editions, Barcelona, 2007.

[7] Fernández R., *Segmentación de mercados*, Thomson Learning, México, 2002.

[8] Martínez W., Martínez A., *Exploratory Data Analysis with MATLAB*, Computer Science and Data Analysis Series, Chapman & HALL,CRC Press UK, UK, 2005.

[9] McDonald M., Dunbar I., *Market Segmentation, how to do it, How to profit from it*, Elsevier Butterworth-Heinemann, UK, 2004.

[10] Nakamura S., *Análisis Numérico y visualización gráfica con MATLAB*, Pearson, Prentice Hall, México, 1997.

[11] Nocedal J., Wright S., *Numerical Optimization*, Springer-Verlag New York, 1998.

[12] Peña, *Análisis de datos multivariantes*, McGraw-Hill/Interamericana de España, S.A., Madrid, 2003.

[13] Peña D., Prieto F., *Multivariate outlier detection and Robust Covariance Matrix Estimation*, Technometrics, Agust 2001, vol. 43, No. 3, p. 286 - 310.

[14] Smith P., *Analysis of failure and survival data*, Chapman & Hall/CRC, USA, 2002.

[15] The Mathworks, *Statistics Toolbox 7 User´s Guide*, The Mathworks Inc., USA, 2008.

[16] SPSS, *The SPSS TwoStep Cluster Component, A scalable component enabling more efficient customer segmentation*, SPSS Inc, USA, 2001.

[17] SPSS, *TwoStep Cluster Analysis*, SPSS Inc, USA, 2001.