UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

CONSTRUCCIÓN DE MODELOS ECONOMÉTRICOS PARA LA ESTIMACIÓN DE ESTADOS FINANCIEROS DE MICROEMPRESAS DEL SECTOR AGRÍCOLA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

YANXI HU

PROFESOR GUÍA:
RICHARD WEBER HAAS

MIEMBROS DE LA COMISIÓN:
JOSÉ MIGUEL CRUZ GONZÁLEZ
EDUARDO ARIOL CONTRERAS VILLABLANCA

SANTIAGO DE CHILE
2015

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniera Civil Industrial
POR: Yanxi Hu
FECHA: 09/10/2015
PROFESOR GUÍA: Richard Weber Haas

CONSTRUCCIÓN DE MODELOS ECONOMÉTRICOS PARA LA ESTIMACIÓN DE ESTADOS FINANCIEROS DE MICROEMPRESAS DEL SECTOR AGRÍCOLA

Para BancoEstado Microempresas (BEME), el monto de las colocaciones en el sector microempresa ha crecido más de 40% en los últimos cinco años. Debido a lo anterior, es importante mejorar el proceso de la otorgación de los créditos, respondiendo de forma más rápida las solicitudes.

La Tecnología de Evaluación de Riesgo (TER) es una herramienta que evalúa a los clientes de BEME que solicitan créditos. Actualmente, para la mayoría de los clientes, este proceso consiste en visitas a terreno y entrevistas por parte de los ejecutivos, con la finalidad de corroborar la información otorgada por éstos, lo cual hace de la TER un procedimiento lento. Con el fin de disminuir este tiempo, se desarrollaron modelos de estimaciones lineales de las variables del estado financiero. De este modo se espera disminuir considerablemente el tiempo que se toman los ejecutivos al otorgar crédito alguno a microempresarios del sector agrícola. De igual manera, al usar el menor valor entre la estimación y lo declarado por el cliente para las variables Venta y Costo fijo y el mayor valor para la variable Margen, se reduce el riesgo de no pago dado que bajaría el monto de otorgamiento, reduciendo el riesgo de crédito del BEME. Asimismo, el almacenamiento de información reduce el riesgo operacional en la verificación de ellas. Además, la información guardada en el sistema le permite al banco generar propuestas comerciales para los clientes. También ayuda a la retención de los clientes, evitando la fuga de ellos, porque se genera un mejor imagen de banco dado que el proceso es más eficiente. En el presente trabajo, se desarrolla la implementación de TER Express en el sector agrícola, actualmente inexistente.

Se estimaron seis modelos de regresión lineal generalizada, ya que este método es más robusto que la regresión lineal ordinaria, para las variables Venta, Costo Fijo y Margen. Esto se realizó para dos casos: por un lado, para los clientes que tienen más de un informe técnico en los últimos 36 meses y por otro, para clientes que sólo tienen uno. Los modelos se han seleccionado al comparar los sub-grupos, eligiendo aquello que tiene menor estadístico BIC.

Las variables, que explican de mejor forma los resultados de este trabajo, son: formalidad del cliente, el sub-segmento al cual pertenece, el tipo de vivienda, entre otro. Se arroja el coeficiente de determinación R^2de 86% para el modelo Venta con historia y 77% para Venta sin historia; en el caso de Costo Fijo con historia y sin historia, el estadístico es de 81% y 48%, respectivamente; y por último, para los modelos de Margen, el estadístico es de 40% y 13%, respectivamente.

Durante el trabajo, se ha intentado agregar la variación de precio de los productos como variable independiente con los datos publicado por la Oficina de Estudios y Policitas Agrarias (ODEPA), pero no se ha logrado encontrar suficientes datos para todos los productos y el tiempo que se requieren.

Finalmente, se recomienda al banco integrar variables exógenas al modelo, como el PIB, la tasa de desempleo sectorial y la variación del precio de los productos que venden los microempresario cuando están disponibles, de modo que éstos lleguen a ser más robustos.

DEDICATORIA

A mi familia, especialmente a mi mamá.
Y a mí misma.

AGRADECIMIENTOS

Al Centro de Finanza por confiarme la realización de este trabajo.

A mi guía Giorgio por su disposición, apoyo y paciencia durante el desarrollo de esta tesis. Sin él, esta memoria no hubiese sido terminada.

A mi comisión por el apoyo que sus miembros me brindaron y por sus comentarios, que enriquecieron la presentación final de esta tesis.

A mi familia por criarme, educarme y apoyarme en estos años, gracias a ellos soy como soy.

A mi hermana por crecer junto a mí, por soportarme y acompañarme día a día.

A Fran y Coni por ayudarme en la revisión de esta memoria

A mis amigos de la universidad por estos años de convivencia, aprendí mucho de cada uno de ellos y espero que sigan siendo así.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

1. INTRODUCCIÓN

1.1. ANTECEDENTES GENERALES

1.1.1. BANCOESTADO MICROEMPRESA

En el año 1995, el BancoEstado creó una filial, la cual tiene el nombre de BancoEstado Microempresas. Esta filial está inspirada en la misión institucional de generar igualdad de oportunidades en el acceso a los servicios financieros para todos los chilenos. El programa surge como respuesta para los sectores microempresariales, hasta entonces marginados del sistema financiero bancario[1].

Durante los últimos años, la cantidad de microempresarios que han pedido algún crédito en el banco ha aumentado sostenidamente. En la Ilustración 1, obtenida de la memoria anual de BancoEstado 2014, se muestra gráficamente el monto de colocaciones en miles de millones y en la Tabla 1, el número de clientes por canal.

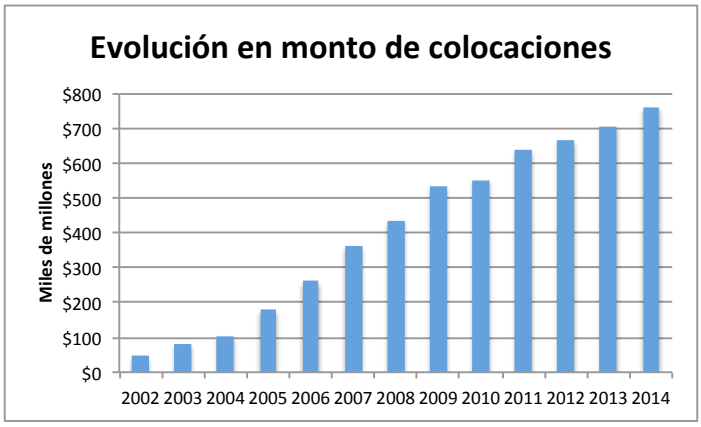


Ilustración 1: Evolución en monto de colocaciones de BEME
Fuente: Memoria anual de BEME 2014

	2010	2011	2012	2013	2014
Clientes Cuenta Rut	2,9 MM	4,2 MM	5,3 MM	6,4 MM	7,4 MM
Clientes por internet	1,0 MM	1,1 MM	1,4 MM	1,8 MM	2,3 MM
Clientes Microempresas	430.000	429.000	446.000	464.000	491.000
Puntos de atención CajaVecina	4.500	7.100	10.200	11.400	13.520
Sucursal BancoEstado	344	341	345	360	364
Oficina ServiEstado	84	86	96	100	107
Cajeros automáticos	1.812	1.891	2.366	2.388	2.405

Tabla 1: Evolución en clientes por canales de atención
Fuente: Memoria anual de BEME 2014

Según información de BEME publicada en el año 2014, el Banco atendió a más de 491.000 microempresarios, de los cuales el 70% tenía un crédito o había realizado una

transacción en los últimos 3 meses de ese año. En materia de colocaciones, su saldo asciende a $760.000 MM con un crecimiento anual del 6% [2].

Debido a la fuerte competencia de esta unidad de negocio en los últimos años y con el fin de tomar acción al respecto, BEME ha decido mejorar los servicios entregados a sus clientes. En particular, ha decidido mejorar el proceso de otorgamiento de créditos de manera que sea más rápida y eficiente.

1.1.1.1. TECNOLOGÍA DE EVALUACIÓN DE RIESGO

Actualmente, la herramienta utilizada para el otorgamiento de créditos para los clientes del sector agrícola es TER (Tecnología de Evaluación de Riesgo), a cargo del área de riesgo. El procedimiento para la asignación de los créditos es el siguiente:

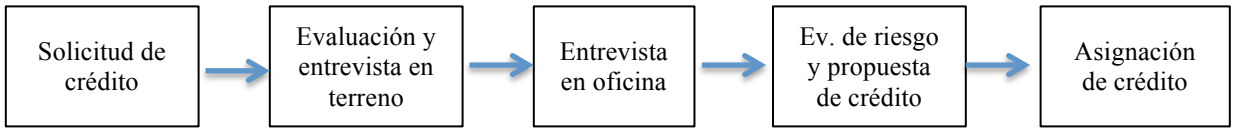


Ilustración 2: Sistema de asignación de créditos bancarios actual

A continuación, se detallan las etapas del proceso de otorgamiento de créditos. Todas estas son desarrolladas por los ejecutivos comerciales del banco. Además, el proceso es el mismo para todos los clientes, ya sean antiguos o nuevos.

1. **Solicitud de crédito**: Esta etapa se lleva a cabo en una sucursal bancaria y corresponde a la solicitud de un crédito por parte de los microempresarios. Esta etapa finaliza con la definición de la fecha para una entrevista en terreno.
2. **Evaluación y entrevista en terreno**: Consiste en una visita al lugar de trabajo del microempresario, en donde se toma nota acerca de los diferentes aspectos del negocio.
3. **Entrevista en oficina**: Se le pide al cliente que lleve algunos documentos que permitan complementar y acreditar la información obtenida en terreno, tales como dividendos, cuentas de luz, agua, teléfono, nivel de estudios, entre otros.
4. **Evaluación de riesgo y propuesta de crédito**: Luego de recopilar toda la información obtenida, el ejecutivo de cuentas evalúa al cliente y calcula el estado de resultados de éste, con el fin de estimar su capacidad de pago. Según la capacidad de pago, se determina el monto máximo de crédito y plazo a otorgar.
5. **Asignación de crédito**: El ejecutivo se comunica con el cliente y le ofrece un monto final de crédito, el cual se otorga una vez que el cliente firma el contrato con el banco.

El mayor inconveniente del proceso es la visita a terreno. Esta etapa puede llegar a demorar semanas si es que no se dan las condiciones para que el ejecutivo realice dicha visita. Según la contraparte de BEME, éste ha sido un problema frecuente durante los últimos dos años, dado el auge que han tenido las microempresas.

A modo de afrontar lo anterior, se formula TER Express. A través de esta herramienta, se realizan estimaciones de las variables de los estados de resultados y con esto se optimiza el proceso de TER.

Durante 2013 y 2014 se desarrolló e implementó TER Express para los segmentos del sector de comercio, de transportes, y de servicios profesionales y manufactura. En esta memoria se desarrollará la implementación de esta herramienta para el sector agrícola.

1.2. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN

Este proyecto nace de la necesidad de BEME de mejorar el proceso de asignación de créditos a sus clientes, dado que el requerimiento de tiempo del proceso actual presenta ineficiencias (especialmente en la etapa de visitas a terreno para la corroboración de la información entregada).

Por lo tanto, el presente proyecto radica en transformar el otorgamiento de micro-créditos de proceso TER a TER Express, donde este último consiste en confeccionar modelos de regresión múltiple para estimar las variables más importantes en el cálculo del estado de resultados de un cliente, las que son Venta, Costo Fijo y Margen.

Los datos utilizados provienen de los informes técnicos, las operaciones cursadas, información demográfica interna del BancoEstado y los datos externos del sistema bancario de los clientes del sector agrícola de BEME, en general.

El modelo de TER Express que se propone en esta tesis, puede ser ilustrado como:

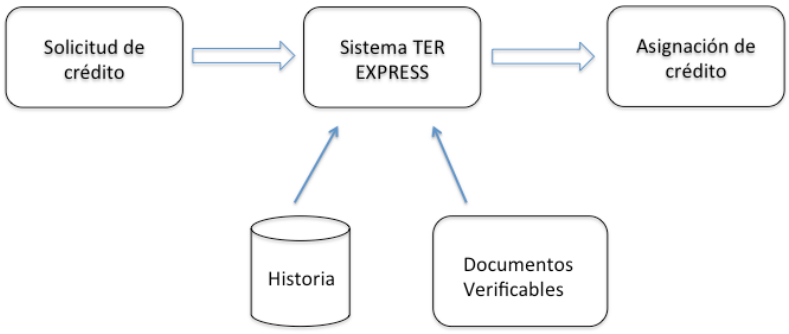


Ilustración 3: Sistema de asignación de créditos bancarios usando TER Express.

Se ajustarán dos modelos para cada una de estas tres variables (Venta, Costo Fijo y Margen): uno para el caso en que el cliente no presenta historia de informes técnicos con el banco (es decir, que no haya solicitado un crédito en los últimos tres años) y otro para el caso en que el cliente si presenta historia de informes técnicos.

A partir de las bases de datos de los clientes, se realizan regresiones lineales múltiples para estimar estas variables. En base a los modelos lineales, generados según variable y tipo de cliente, se construyen las variables que permiten estimar los estados de resultado operacionales (R.O.) de los clientes de BEME.

$$R.O. = Ventas - Costo\ Fijo - Costo\ Variable$$

Donde el Costo Variable se estima a partir del variable Margen que se obtiene de:

$$Margen = \frac{Ventas - Costo\ Variable}{Ventas}$$

Los modelos que se formularán permitirán disminuir el tiempo en que los ejecutivos comerciales obtienen el monto máximo de crédito a otorgar a los clientes, reduciendo el tiempo de la respuesta. El proceso de registro de datos para la evaluación de riesgo y propuesta de crédito actualmente dura una hora. Además, se deben tomar en cuenta los tiempos que implican el traslado a terreno y las reuniones con los clientes. Con TER Express, se espera disminuir el proceso de traspaso de datos a treinta minutos, liberando de esta forma treinta minutos, además de eliminar la visita a terreno que implica un mayor costo para el banco.

Se espera que los ejecutivos comerciales puedan atender un número mayor de clientes en comparación con los que atienden ahora, donde gastan casi un día en las visitas a terreno para corroborar la verosimilitud de los datos entregados. Con la herramienta TER Express, se elimina la etapa de visita a terreno (excepto si el cliente es totalmente nuevo: no tiene solicitudes de créditos anteriores). De igual manera, al usar el menor valor entre la estimación y lo declarado por los clientes para las variables Venta y Costo fijo y el mayor valor para la variable Margen, se reduce el riesgo de no pago, es decir, se logra disminuir el riesgo de crédito del BEME. También ayuda a la retención de los clientes evitando la fuga de ellos, porque se genera un mejor imagen del banco, dado que el proceso es más eficiente. Asimismo, el almacenamiento de la información reduce el riesgo operacional en la verificación de esta. Además, la información guardada en el sistema permite al banco generar propuestas comerciales para los clientes.

2. ASPECTOS FUNDAMENTALES DEL PROYECTO

2.1. OBJETIVOS

2.1.1. Objetivo General

Formular modelos econométricos para estimar las variables Ventas, Costos Fijos y Margen dentro de BEME con el fin de apoyar a los ejecutivos en el otorgamiento de microcréditos al sector agrícola.

2.1.2. Objetivos Específicos

- Conocer el procedimiento de otorgamiento de los créditos en el sector agrícola dentro del banco.
- Estudiar la herramienta TER Express, implementada para los casos del sector de comercio y el sector de servicios profesionales y manufactura.
- Validar el origen de los datos en el sistema de información de BEME y la construcción de la base de datos, a partir de las información histórica de los clientes.
- Generar modelos de regresión lineal generalizada para el sector agrícola.
- Verificar la predictibilidad de los modelos.

2.2. METODOLOGÍA

La estimación de las variables Venta, Costo Fijo y Margen, se realizará mediante una regresión lineal generalizada.

En primer lugar, se integra la base de datos analítica, sobre la cual se trabajan los distintos modelos con la información proveniente de los informes técnicos urbanos, operaciones cursadas de los clientes y la información demográfica interna del sistema financiero. Esto se puede apreciar detalladamente en la sección 4.1.

Luego, se realiza un análisis descriptivo de las variables de la base analítica armada usando el programa estadístico informático SPSS, donde se estudian las variables, de manera de tener una idea general de su comportamiento.

A continuación, se realiza una limpieza de la base de datos de acuerdo al resultado de análisis descriptivo anterior, donde se procede a limpiar aquellas variables que no cambian en la base o registros de variables que presentan inconsistencias, por ejemplo: duplicación de registros, clientes cuya edad es menor a 18 años, entre otros. El detalle de esto se puede ver en la sección 4.4.

Posteriormente, se categorizan las variables, ya sea continuas o discretas, en segmentos que presenten un comportamiento similar. Para esto se ocupa la herramienta del software SPSS, “árbol de clasificación”, con el algoritmo “CHAID exhaustivo”, que utiliza el criterio de “chi-cuadrado”. El detalle de estos se puede apreciar en la sección 5.1.

Después, se realiza una correlación de variables donde primero se eliminarán aquellas variables continuas que se correlacionan linealmente menos del 20% con la variable dependiente; luego, tomando las variables dependientes que quedan, se eliminarán aquellas variables continuas donde se correlaciona linealmente más del 50% entre ellas, lo cual es la estándar de la industria utilizado por el BEME , dejando la variable con mayor correlación con la variable dependiente; y posteriormente, se eliminarán aquellas variables categóricas que se correlacionan linealmente más del 50% entre ellas, dejando la variable con mayor correlación con la variable dependiente. El detalle de esto se puede apreciar en la sección 5.1.

Para ajustar los modelos, se detectan primero los datos *outliers*; posterior a eso, se aplican las transformaciones necesarias con el fin de arreglar problemas con los supuestos del modelo, como por ejemplo: la transformación de Box Cox, la cual permite normalizar las variables en el caso que no lo sean. Seguidamente se extraen los datos *outliers* de la muestra, se verifica si fueron eliminadas las categorías completas de las variable; después se ajustan los modelo con todas las variables resultantes de la eliminación, con regresión lineal generalizada, eliminando aquellas variables que tiene nivel de significancia mayor de 5% para después transformar las variables categóricas en variables *dummy*. A continuación, se seleccionan las variables continuas y las variables *dummy* a través del algoritmo *forward*, donde las variables más explicativas entran primero con regresión lineal multivariada. Por último, se generan y ordenan distintos modelos de regresión lineal generalizada, a partir de las variables seleccionadas en el modelo de regresión lineal multivariada anterior, seleccionando aquel modelo donde se presente un cambio menor al 0,05% en el estadístico BIC, al agregar más variables en el modelo, ya que se ha visto empíricamente desde los gráficos, que a partir de ese número, se establece la curva de estadístico BI. El detalle se puede apreciar en la sección 5.2.

Para la robustez de los modelos, es necesario validar los modelos ajustados, donde se contrasta la distribución empírica v/s la distribución teórica de la variable respuesta (sea Venta, Costo Fijo o Margen) dentro del segmento agrícola. El detalle de esto se puede apreciar en la sección 5.2.

2.3. ALCANCES

Los alcances de este proyecto consisten, primeramente, en levantar una base de datos a partir de las distintas bases de datos existentes, validando la correcta construcción de ella, mitigando el riesgo operacional que puede haber en dicho proceso.

Luego, acorde al objetivo general, se construirán seis modelos estadísticos para las microempresas que pertenecen al segmento agrícola. Los otros segmentos se dejan fuera de este estudio, ya que pueden presentar comportamientos distintos dados los segmentos y algunos de ellos ya se han estudiados en otros trabajos anteriores o se encuentran fuera del alcance del presente trabajo.

Los seis modelos corresponden a modelos de regresión lineal generalizada para Costo Fijo, Venta y Margen, en clientes con y sin historia. Una vez construidos los modelos, los ejecutivos podrán calcular estimaciones de los estados de resultados de los clientes del banco.

3. MARCO CONCEPTUAL

Para comprender y abordar de forma óptima el trabajo, se introducen conceptos y metodologías necesarias, proporcionando una base teórica que sustenta el trabajo desarrollado.

3.1. MICROEMPRESA y TER EXPRESS

A continuación, se muestran las definiciones preliminares para entender la relación de las Microempresas y cómo son conceptualizadas en BEME, a través de los datos de éstas en sus modelos de TER Express.

3.1.1. Tipos de Empresas

Se trabaja con todos los tipos de microempresas existentes en el país. Una clasificación adecuada para evaluar si una empresa será Formal, Semi-formal o Informal es la siguiente:

- Empresas Formales: Son aquellas que mantienen registros vigentes en el Sistema de Impuestos Internos (SII) y por lo tanto, pagan impuestos. Además, poseen todas las patentes, permisos y autorizaciones legales correspondientes. Aparecen de manera segura en los registros de la SBIF.
- Empresas Semi-formales: Son las que no tienen iniciación de actividades en el SII, es decir, no pagan impuestos; pero si tienen patentes municipales y aparecen registradas con su marca. Estas pueden aparecer o no en la SBIF dependiendo si han pedido crédito o no anteriormente.
- Empresas Informales: Son aquellas que no tienen registro alguno en el SII ni en la municipalidad y por lo tanto no poseen la documentación necesaria para la formalidad y solo poseen cuadernos de registros. Aparecen en los datos del SBIF sólo si han logrado conseguir crédito en un banco.

3.1.2. Cálculo de Resultado Operacional de la TER Express

En el modelo TER Express se realiza una estimación de las variables dependientes de Venta, Costo Fijo y Margen mediante la evaluación *express* que hacen los ejecutivos comerciales a los microempresarios con las variables que se requieren según los distintos modelos de regresión lineal generalizada múltiple. Calculando estas tres variables, finalmente se obtiene el Resultado Operacional (RO) de cada uno de los microempresarios.

Luego de obtener el RO mediante los modelos y lo declarado por el cliente, se escoge el de menor valor para las variables Venta y Margen. Por otro lado, se utiliza el valor máximo para la variable Costo Fijo; dado ese valor, se define la capacidad de pago.

Esta capacidad de pago es ajustada a ciertos ponderadores, entre ellos, por ejemplo: la tasa de interés. Más adelante se explica el proceso que da origen a la capacidad de pago ajustada sobre la cual se origina la oferta de crédito que se le ofrece al microempresario. A continuación, se puede apreciar el diagrama de funcionamiento de TER Express, en cuanto a la información y variables.

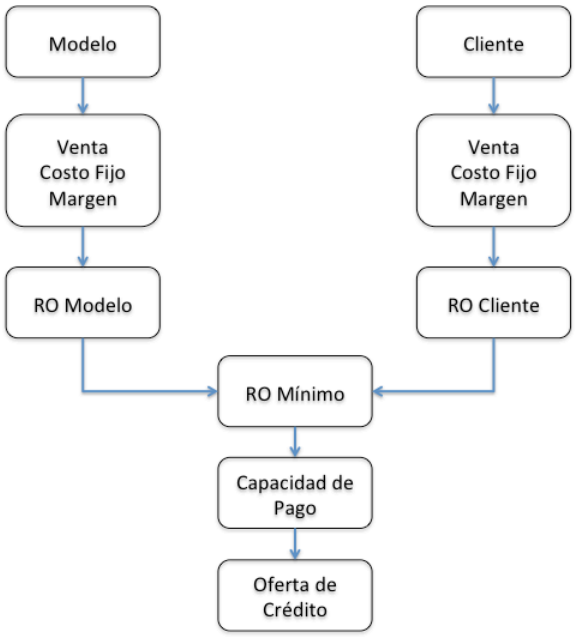


Ilustración 4: Funcionamiento de TER Express

Es importante destacar que si la persona es cliente por primera vez, el procedimiento de TER Express no es viable, ya que funciona sólo con personas que tienen antecedentes en el banco y no se aplicará para una persona que no ha pedido crédito nunca.

3.2. CLASIFICACIÓN DE VARIABLES

Las variables que se presentan en la base de datos analítica pueden clasificarse como:

- **Cualitativas nominales**: son variable no numéricas y no presentan un orden. Ejemplo: Género del cliente.
- **Cualitativas ordinales**: son variable no numéricas, sin embargo, si presentan un orden. Ejemplo: Formalidad del cliente, lo que puede ser Formal, Semi-Formal, o Informal.
- **Cuantitativa discreta**: con valores numéricos ordenados aislados (sin intermedios). Por ejemplo: Cantidad de hijos.
- **Cuantitativa continúa**: estos valores numéricos que poseen valores intermedios. Por ejemplo: Gastos en arriendos.
- **Variables *dummy* cualitativas**: son el tipo de variable *dummy* que en vez de tener valor 1 o 0, tiene un valor no numérico como Sí o No.

3.3. CLASIFICACIÓN DE DATOS

3.3.1. Árbol de Clasificación

Un árbol de decisión es un método de clasificación no paramétrico, cuya estructura es semejante a un diagrama de flujo, en donde cada nodo interno denota un test en cierto atributo, cada rama representa un resultado de ese test, y cada nodo terminal conlleva una etiqueta de clase, representando un segmento diferente.

La inducción en árboles de decisión es el aprendizaje a través de estas estructuras, a partir de datos de entrenamiento marcados con etiquetas de su clase correspondiente. La popularidad de estos métodos se debe a que la construcción de estos clasificadores no requiere conocimiento del dominio particular desde donde provienen los datos. Esto los hace particularmente atractivos para análisis exploratorio de datos. Más aún, estos métodos pueden manejar una alta dimensionalidad y su estructura de representación es intuitiva y fácil de asimilar [3].

Los árboles de clasificación permiten:

1. Caracterizar: Identificar qué propiedades caracterizan o discriminan a los casos que pertenecen a cada una de las categorías de la variable dependiente.
2. Segmentar: Agrupar los casos con características similares.
3. Conocer cuáles de las variables independientes influyen más en la clasificación.
4. Clasificar: Crear reglas que permitan clasificar nuevos casos en esos grupos ya predefinidos.

Algunos de los algoritmos más populares que implementan el concepto de árbol de clasificación son:

- ID3: Utiliza el criterio de "ganancia de información" para seleccionar atributos.
- C4.5: Utiliza el criterio de "razón de ganancia" para seleccionar atributos.
- CART: Utiliza el criterio del "índice Gini" para seleccionar atributos.
- CHAID: Utiliza el criterio "chi-cuadrado" para seleccionar atributo.
- CHAID Exhaustivo: Una modificación del CHAID que permite explora todas las divisiones posibles de cada predictor.

Para esta memoria, será de particular interés el algoritmo CHAID exhaustivo, ya que este examina todas las posibles divisiones, eligiendo la más robusta. A continuación, se explicará en más detalle el funcionamiento de esta herramienta.

3.3.1.1. CHAID (Chi-square Automated Interaction Detector)

La técnica CHAID divide el conjunto de datos en sub conjuntos que son mutuamente excluyentes y que discrimina de mejor manera el comportamiento de la variable dependiente. Los subconjuntos se construyen utilizando un grupo pequeño del total de predictores disponibles. Esto se hace para especificar mejor y ver relevancia de la estructura de variables. Los predictores seleccionados pueden ser usados más tarde en análisis posteriores, en la predicción de la variable dependiente, en lugar del total de predictores

El algoritmo bajo el cual se implementa CHAID, que es iterativo en esencia, procede de la manera que se defina a continuación.

1. Se identifica la mejor partición por cada predictor disponible.
2. Los predictores se comparan y se escoge al óptimo.

3. Los datos se subdividen de acuerdo a la partición del atributo seleccionado y cada una de las particiones vuelve a someterse al proceso descrito anteriormente, hasta que no se encuentren más particiones significativas a 95%.

3.3.1.2. CHAID Exhaustivo

El CHAID exhaustivo es una modificación de CHAID que analiza todas las posibles fusiones de las variables predictores nominales u ordinales hasta encontrar la relación con el mayor nivel de significancia con la variable dependiente. De esta forma, el CHAID exhaustivo puede encontrar la mejor segmentación de las variables predictores respecto a la variable dependiente.

En Anexo I se mostrará un ejemplo de esto.

3.4.COEFICIENTE DE CORRELACIÓN PEARSON

Este coeficiente indica el grado de asociación lineal entre dos variables, es decir, la tendencia de los puntos de la nube a situarse alineadamente cerca de una recta, con excepción de las rectas horizontales y verticales.

Dicho coeficiente oscila entre –1 y +1. Un valor de 1 indica una relación lineal o línea recta positiva perfecta. Una correlación próxima a cero indica que no hay relación lineal entre las dos variables.

Una correlación elevada y estadísticamente significativa no tiene que asociarse a causalidad. Cuando objetivamos que dos variables están correlacionadas, diversas razones pueden ser la causa de dicha correlación: a) puede que X influencie o cause Y, b) puede que Y influencie o cause X, c) X e Y pueden estar influenciadas por terceras variables que hace que se modifiquen ambas a la vez. El coeficiente de correlación mide el grado de asociación entre dos cantidades pero no mira el nivel de acuerdo o concordancia. Si los instrumentos de medida registran sistemáticamente cantidades diferentes uno del otro, la correlación puede ser 1 y su concordancia ser nula [4] .

La fórmula se expresa como:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

donde

σ_{xy} es la covarianza de las variables X e Y

σ_x es la desviación típica de la variable X

σ_y es la desviación típica de la variable y

3.5. ANÁLISIS CON TABLA DE CONTINGENCIA

Para analizar la relación de dependencia entre dos variables cualitativas nominales o discretas, es necesario estudiar su distribución conjunta, lo que se resume en un objeto que se llama tabla de contingencia.

La tabla de contingencia es una tabla de doble entrada, donde en cada casilla figura el número de casos o individuos que posee un nivel de uno de los factores o características analizadas y otro nivel del otro factor analizado. En la Tabla 2 se muestra un ejemplo de este, donde se analiza la relación entre la variable Género y Estado civil.

		Estado Civil				Total
		0	1	2	3	
Género	0	10.759	63	123	9	10.954
	1	2.350	154	135	40	2.679
	2	17.473	1.093	2.071	51	20.688
Total		30.582	1.310	2.329	100	34.321

Tabla 2: Ejemplo de tabla de contingencia entre la variable estado civil y género

La tabla de contingencia se define por el número de atributos o variables que se analizan conjuntamente y el número de modalidades o niveles de los mismos. El ejemplo propuesto es una tabla de contingencia 3 filas y 4 columnas, ya que tiene dos atributos (Género y Estado civil) y la variable Estado civil tiene 4 niveles, mientras que la variable Género tiene 3 niveles.

Para identificar relaciones de dependencia entre variables cualitativas se utiliza un contraste estadístico basado en el estadístico χ^2 (Chi-cuadrado), cuyo cálculo permite afirmar con un nivel de confianza estadístico determinando si los niveles de una variable cualitativa influyen o no en los niveles de la otra variable nominal analizada. Siguiendo con el ejemplo propuesto, el cálculo de Chi-cuadrado permitirá saber si el sexo de una persona es un factor determinante en que dicha persona tenga cierto estado civil.

La hipótesis nula a contrastar será la independencia entre los factores, siendo la hipótesis alternativa la dependencia entre los factores.

El valor de χ^2 calculado se compara con el valor tabulado de una χ^2 para un nivel de confianza determinado y (n-1) (k-1) grados de libertad. Si el valor calculado es mayor que el valor de tablas de una χ^2, significará que las diferencias entre las (n-1) (k-1) frecuencias observadas y las frecuencias teóricas o esperadas son muy elevadas y por tanto, se dice con un determinado nivel de confianza, que existe dependencia entre los factores o atributos analizados.

Cuando se utiliza el programa estadístico SPSS se dará el nivel de significación, es decir, la probabilidad de rechazar la hipótesis nula siendo cierta y por tanto, la probabilidad de equivocarse si rechazamos la hipótesis nula. Si esta probabilidad es muy pequeña (<0,05), se rechaza la hipótesis nula y en consecuencia, se dice que los atributos son dependientes. Por el contrario, si el nivel de significación fuera superior a 0,05, la probabilidad de equivocarse al concluir que los factores son dependientes sería muy alta y por tanto aceptaremos la hipótesis nula de independencia.

Cuando el valor de Chi-cuadrado indica que existe una relación de dependencia entre las variables, este no indica nada sobre la fuerza de asociación entre las variables estudiadas debido a que su valor está afectado por el número de casos incorporados en la muestra, ya que a mayor número de casos analizados (a mayor N), el valor de la χ^2tiende a aumentar; por lo que cuanto mayor sea la muestra, más fácil será que rechacemos la hipótesis nula de independencia, cuando a lo mejor podrían no ser dependientes.

Por eso, se realiza el test correlación, donde se calcula intervalo por intervalo el estadístico R de Pearson, el cual es análogo al coeficiente de correlación de Pearson.

3.6. TRANSFORMACIONES DE BOX-COX

A pesar de que existen transformaciones particulares que pueden corregir específicamente alguno de los supuestos, Box y Cox encontraron que frecuentemente una transformación simple puede rectificar simultáneamente los problemas de falta de normalidad y falta de homogeneidad de varianzas [5].

El método de Box-Cox es una transformación sobre la variable dependiente. Su diseño está pensado para valores estrictamente positivos y elige la transformación que mejor se ajuste a los datos. El método transforma la variable dependiente Y en $Y(\lambda)$, donde la familia de transformaciones indexada por λ es la indicada a continuación:

$$Y(\lambda)\begin{cases}\dfrac{Y^{\lambda}-1}{\lambda} & \lambda \neq 0\\ \log(Y) & \lambda = 0\end{cases}$$

El valor de λ es elegido por el método de máxima verosimilitud. Asumiendo la normalidad de los errores, su valor vendrá dado a través de la siguiente ecuación:

$$\mathrm{U}(\lambda)\begin{cases}\dfrac{\mathrm{y}^{\lambda}-1}{\lambda\tilde{\mathrm{y}}^{(\lambda-1)}} & \lambda \neq 0\\ \tilde{\mathrm{y}}\log(\mathrm{y}) & \lambda = 0\end{cases}$$

Siendo ỹ la media geométrica de la variable y. Para cada valor de λ se obtiene el conjunto de valores de $\{\mathrm{U_i}(\lambda)\}_{\mathrm{i=1}}^{\mathrm{n}}$. La función de verosimilitud es

$$L(\lambda) = -\frac{n}{2}\ln\left(\sum_{i=1}^{n}\left(U_i\,(\lambda) - \overline{U}\,(\lambda)\right)^2\right)$$

Se elige el parámetro λ que maximiza la función $\mathrm{L}(\lambda)$, pero normalmente $\mathrm{L}(\lambda)$ se maximiza sobre un conjunto de valores tales como {-2, -1, -1/2, 0, 1/2, 1, 2}, con el fin de asegurar que λ tenga una interpretación más sencilla.

Los valores más utilizados en las transformaciones son:

λ	Transformación
-1	$Z(\lambda) = \frac{1}{y}$
-1/2	$Z(\lambda) = \frac{1}{\sqrt{y}}$
0	$Z(\lambda) = \ln(y)$
1/2	$Z(\lambda) = \sqrt{y}$
1	$Z(\lambda) = y$

Tabla 3: Valores más utilizados de la transformación BOX-COX

3.7. REGRESIÓN LINEAL MULTIVARIADA

La técnica de regresión más popular es la de regresión lineal. Dentro de los problemas de regresión lineal, las técnicas a utilizar varían según la cantidad de variables predictoras. El caso básico trata aquellos problemas en los que existe una variable predictora y una variable respuesta y se denomina regresión simple. En los casos en que existen varias variables predictoras pero sólo una de respuesta, el problema se denomina regresión multivariada [6].

En este trabajo sólo analizaremos casos de regresión lineal multivariada, por lo que el siguiente desarrollo sólo abarcará este caso.

En el análisis univariado, el modelo lineal general surge de la necesidad de cuantificar las relaciones entre un conjunto de variables, en la que una de ellas se denomina variable respuesta o dependiente Y y las restantes son las variables explicativas o independientes X_I (continuas o categóricas) y un término de error ε.

En el ámbito de la regresión lineal multivariada, se asume que la variable respuesta o dependiente Y está linealmente relacionada con las variables explicativas o independientes X_I (continuas o categóricas) y el término de error ε, de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2, \ldots n$$

donde:

y_i= Variable respuesta o dependiente.

$x_{1i}, x_{2i}, \ldots, x_{pi}$=Variables explicativas o independientes.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$= Coeficientes de regresión (parámetros a estimar)

ε_i es una perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto, se asocian al azar o error de especificación del modelo, es decir, toma en cuenta las variables independientes que afectan a Y, pero que no están incluidas en el modelo. Este término se define como :

$$\varepsilon_i = y_i - \hat{y}_i$$

Donde y_i es el valor real de la variable dependiente e $\hat{y}_i$ es su valor estimado.

El problema de la regresión consiste en buscar unos valores determinados para los parámetros desconocidos β_p, de modo que la ecuación quede completamente especificada. Para ello, se necesita un conjunto de observaciones, con el fin de registrar el comportamiento simultáneo de la variable dependiente y las variables explicativas.

Para garantizar la validez de los modelos, es necesario cumplir algunos supuestos que son una serie de condiciones, las cuales son descritas a continuación [7].

1. **Linealidad**

   La variable dependiente es la suma de un conjunto de elementos, es decir, una combinación lineal de variables independientes y el residuo.

   El incumplimiento de este supuesto se denomina error de especificación. Este se puede dar por la no linealidad de la variable dependiente respecto a las variables independiente, por la omisión de variables independientes importantes o la inclusión de variables independientes irrelevantes, entre otros.

2. **Independencia de los errores**

   Los residuos son independiente entre sí, es decir, constituyen una variable aleatoria y no están correlacionados entre ellos.

   La falta de independencia se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo, esto es cuando se trabaja con series temporales.

   Como los modelos con los que se trabajarán en esta memoria corresponden a regresiones lineales que no dependen del tiempo (corte transversal), entonces se asume la no correlación de los errores.

3. **Homocedasticidad**

   Para cada valor de la variable independiente o combinación de valores de las variables independientes, la varianza de los residuos es constante.

   La homocedasticidad implica que la variación de los residuos es uniforme en todo el rango de valores de los pronósticos. Hay diversas causas que pueden afectar la homocedasticidad de los residuos. Entre ellas se encuentra: que no se haya incorporado en el modelo alguna variable de importancia, que existan muchos valores extremos o atípicos (outliers) o que las unidades de información se encuentren particionadas en grupos heterogéneos.

4. **Normalidad**

   Para cada valor de la variable independiente o combinación de valores de las variables independientes, los residuos se distribuyen normalmente con media cero. Y si a eso se le agrega el principio de homocedasticidad, se puede inferir que los errores están normalmente distribuidos con media cero y varianza constante.

   Dada esta situación y en base a la cantidad de observaciones que se tiene, puede aplicarse la ley de los grandes números y el supuesto de normalidad puede considerarse, al menos aproximadamente, para todos los casos. Se dice entonces que el supuesto de normalidad es innecesario para el modelo de regresión lineal excepto en los casos donde se asume explícitamente alguna distribución alternativa, cuyo caso no es el estudiado en este trabajo.

5. **No colinealidad**

   No existe relación lineal entre ninguna de las variables independientes, el incumplimiento de este supuesto da origen a multicolinealidad.

   Este problema puede volver inestable al modelo y afectar la varianza de los estimadores, ya que genera valores más altos en cuanto a la varianza.

   Para evitar problemas de este tipo, se realizará un análisis de las correlaciones entre todas las variables independientes del modelo y se excluirán las variables redundantes dentro de este.

3.7.1. Algoritmo Stepwise

Un concepto básico importante que se debe conocer, es el algoritmo stepwise con el cual se irán incluyendo variables a los modelos de Regresión Lineal para poder encontrar las variables que son significativas para el cálculo de Venta, Costo Fijo y Margen. Este algoritmo utiliza los siguientes procedimientos:

- Backward Stepwise Regression: es el procedimiento que parte del modelo de regresión con todas las variables dependientes y en cada etapa se elimina la variable menos influyente, según el contraste individual de la t (o de la F) hasta una cierta regla de parada.
- Fordward Stepwise Regression: es el algoritmo que funciona de forma inversa que el anterior, parte del modelo sin ninguna variable dependiente y en cada etapa se introduce la más significativa hasta una cierta regla de parada.
- Stepwise comienza como el algoritmo Fordward Stepwise, pero en cada etapa se plantea si todas las variables introducidas deben permanecer. Este termina cuando ninguna variable entra o sale del modelo.

3.8. REGRESIÓN LINEAL GENERALIZADA

Los modelos lineales generalizados (MLG) son una extensión del clásico Modelo Lineal a una familia más general e involucran un variedad de distribuciones tales como Poisson, Binomial, Normal, Gamma, entre otros.

Generalmente, el modelo lineal generalizado se puede escribir de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2, \ldots n$$

y la otra forma es:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \qquad i = 1,2, \ldots n$$

donde $\mu_i$es la media de una distribución normal con varianza constante σ^2. Se puede observar que el modelo está formado por los siguientes tres componentes [8]:

1. **Componente aleatorio**

   Las y_i (i=1, …, n) son variables aleatorias independientes con media μ_i. Estas proporcionan alguna distribución de la familia de distribución exponencial (Normal, Binomial, Poisson, Gamma y Binomial Negativa) con una constante como parámetro escala.

2. **Componente sistemático**

   El componente sistemático se refiere a un vector $\eta_i, \ldots, \eta_n$ de variables independientes mediante un modelo lineal. Sea x_{pi} el valor del predictor $j (j = 1, \ldots, p)$ para el sujeto i. Entonces:

   $$\eta_i = \sum_j x_{ij}\beta_j \quad i = 1, \ldots n$$

   Esta combinación lineal de variables independientes es llamada predictor lineal.

3. **Función de enlace**

   La función de enlace es una transformación $g(\mu)$ que permite relacionar el componente aleatorio y el componente sistemático, vinculando el predictor lineal η con el valor esperado de la variable respuesta, $E(Y/X) = \mu$, a través de la función:

   $$\eta = g(\mu), \mu = g^{-1}(\eta)$$

   donde g es una función diferenciable, monótona e invertible.

La regresión lineal generalizada funciona de la misma forma que la regresión lineal multivariada (propuesta en 3.7) cuando la variable dependiente se distribuye de forma normal. La diferencia está en que el resultado entregado por la regresión lineal generalizada es más robusto; y la forma de ingresar las variables al programa estadístico SPSS ocupado en dicha memoria es distinta, donde en el caso de la regresión lineal multivariada se deben ingresar los segmentos una variable categóricas como variables dummy, es decir, variables individuales; además la existencia de algoritmo stepwise, que genera un orden del ingreso de las variables, dependiendo el aporte en la variable dependientes.

3.9. BONDAD DE AJUSTE

3.9.1. Coeficiente de Determinación R^2

El coeficiente de determinación R^2 es un medida de bondad de ajuste del modelo que ajuste a un conjunto de datos [9].

Este coeficiente se basa en la siguiente descomposición:

$$SCT = SCE + SCR$$

donde SCT es la suma de cuadrados totales, SCE es la suma de cuadrados explicados y SCR es la suma de cuadrados residuales.

Basándose en esta ecuación, el coeficiente de determinación se define como:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

donde $\widehat{Y}_i$ es el valor ajustado de las variables reales dependiente Y_i y $\overline{Y}$ es el valor promedio de los Y.

Se nota que $0 \leq R^2 \leq 1$:

- Si $R^2 = 0$, la variabilidad de la Y es explicada exclusivamente por el error o bien las X´s en nada se ajustan a la Y.
- Si $R^2 = 1$, la variabilidad de la Y es explicada exclusivamente por las X´s (el error no interviene en la variabilidad de la Y) o bien se dice que el ajuste es máximo.

Este indicador tiene que ser considerado como uno más a tener en cuenta a la hora de valorar si un modelo es adecuado, pero no se le debe dar más importancia de la que tiene. Obtener un valor del R^2 cercano a 1 no indica que los resultados sean fiables [10].

Obtener un valor más o menos alto del coeficiente de determinación puede estar influenciado por el tipo de datos que se estén analizando. Normalmente con datos de series temporales, donde las variables pueden presentar tendencias similares en el tiempo, es fácil obtener R^2 altos, mientras que con datos de sección cruzada eso no suele ocurrir ya que normalmente las variables presentan mayor dispersión.

Además, a medida que se ingresan más variables al modelo, el R^2 tienden a aumentar, y en el peor de los casos se mantiene, por lo cual este coeficiente no sirve para la comparación de modelos.

3.10. COMPARACIÓN DE MODELOS

Dado que el coeficiente de determinación no es apto para seleccionar modelos, comúnmente se ocupan los criterios de información de Akaike (AIC) y Bayesiano de Schwarz (BIC). Estos criterios se calculan en función de la suma de cuadrados residual y de algún factor que penalice por la pérdida de grados de libertad. Un modelo más complejo, con más variables explicativas, reducirá la suma de cuadrados residual pero aumentará el factor de penalización, eliminando el sesgo introducido al agregar variables. Utilizando estos criterios, se escoge aquel modelo con un menor valor de AIC o BIC.

3.10.1. Criterio de Información Akaike

Uno de los criterios que se usa en la selección de modelos, para elegir el mejor entre un conjunto de modelos admisibles, es el criterio de información Akaike (AIC). Este se ha definido como:

$$AIC = -2 * LL(\widehat{\beta}) + 2 * k$$

donde :

LL es el máximo valor de la función de verosimilitud del modelo.

k es el número de parámetros libres de ser estimados.

n es el número de observaciones del modelo.

3.10.2. Criterio de Información Bayesiano

Otro de los criterios que se usa en la selección de modelos, para elegir el mejor entre un conjunto de modelos admisibles, es el criterio de información Bayesiano(BIC). Se define como:

$$BIC = -2 * LL(\widehat{\beta}) + k * Ln(n)$$

donde :

LL es el máximo valor de la función de verosimilitud del modelo.

k es el número de parámetros libres de ser estimados.

n es el número de observaciones del modelo.

Cuantitativamente, dado que el procedimiento de BIC difiere del AIC sólo en que la dimensión es multiplicada por ln(n), el BIC se inclina más que el AIC hacia modelos de

dimensión más baja cuando el número de observaciones es $n \geq 8$. Para un número elevado de observaciones, existen diferencias notables entre el AIC y el BIC [11].

Otra diferencia entre estos dos criterios, el AIC y el BIC, radica en que tienen objetivos distintos. El criterio BIC trata de seleccionar el modelo correcto, con máxima probabilidad a posteriori, y puede demostrarse que es un criterio consistente, de manera que la probabilidad de seleccionar el modelo correcto tiende a 1 cuando aumenta el tamaño muestral. El criterio AIC no pretende seleccionar el modelo correcto, ya que admite que el modelo verdadero puede no estar entre los estimados, y trata de obtener el modelo que proporcione mejores predicciones entre los existentes. Puede demostrarse que, suponiendo que el modelo verdadero puede aproximarse arbitrariamente bien con los estimados, al aumentar el tamaño de la muestra, el criterio AIC es eficiente, escogiendo el modelo que proporciona (en promedio) mejores predicciones. Sin embargo, en muestras pequeñas o medianas, el criterio AIC tiende a seleccionar modelos con más parámetros de los necesarios [12].

4. PREPARACIÓN DE LAS BASES DE DATOS

4.1. DISEÑO DE LA BASE ANALÍTICA

Tal como se señala en la sección de metodología, primero se construye la base de datos analítica, para eso se procede a integrar la información contenida en:

- Los informes técnicos urbanos, donde aparecen datos tales como formalidad del microempresario, permanencia en el rubro, número de clientes, forma de pago a los proveedores, número de trabajadores, entre otros.
- Operaciones cursadas, contienen datos como cantidades de operaciones cursadas en los últimos 12 meses, en los últimos 24 meses, venta, costos fijos y margen del último informe técnico, el plazo de la operación cursadas, el monto origen de la operación, entre otros.
- Información demográfica interna ambiental del sistema financiero, son datos como puntaje ambiente SMRT01, puntaje SICA, deuda en el BancoEstado, deuda en INDAP, deuda en otros banco, entre otros.

El proceso de integración de las distintas fuentes de información de BEME, para la base de datos analítica del sector agrícola, tomó 9 meses. Este proceso fue desarrollado en el contexto de esta memoria y se puede destacar lo siguiente:

1. Primero se verifica que la información de los informes técnicos urbanos, guardados en la base de datos, sea la misma que está en la plataforma web del banco. De esa forma, se visualiza la existencia de errores en el traspaso de información, evitando así equivocaciones en la creación de los datos.
2. Luego, se calculan las variables que se relacionan con los informes pasados, tales como operaciones cursadas en los últimos 12 meses, la venta del cliente que aparece en el informe pasado, en el último informes y entre otros El tiempo de ejecución requerido para el cálculo de estas variables puede demorarse hasta un día y, dado que esto se hace con el computador del funcionario de BEME, solamente se pueden ejecutar los cálculos los días viernes en la tarde.
3. Una vez calculadas estas variables, se procede a verificarlas con las bases históricas del cliente para ver si fueron correctamente computadas. En el caso de encontrarse discrepancia, se procede a calcular y ejecutar el proceso anterior de nuevo, hasta que se verifican correctamente.
4. Después de verificar el origen de los datos, se procede a unificar las variables de las distintas bases a la base final. Dado que solamente se puede ocupar el computador los viernes en la tarde y los fines de semana para este labor, y el proceso es demoroso, primero se agrega un set de variables para que se procese el fin de semana. Luego, se validan estas variables con las bases de origen para evitar errores operacionales. Después se agrega otro set de variables y se validan. Así sucesivamente, hasta terminar de agregar todos las variables a la base final.

A partir de la base de datos analítica armada, se realizan trabajos de modelamiento predictivo y análisis estadístico para la estimación de las variables: Venta, Costo Fijo, Margen. El Margen se calcula a partir de la siguiente fórmula:

$$MARGEN = \frac{\text{VENTA} - COSTO\ VARIABLE}{\text{VENTA}}$$

Se construye un Modelo por cada una de las variables dependientes a estimar. Estos son explotados en el marco del sistema de evaluación de créditos BEME denominado TER Express.

4.2. UNIVERSO DE ESTUDIO

Se consideran parte del universo de estudio a los clientes del sector Agrícola, los cuales presenten al menos un informe técnico asociado al mismo rubro en el pivote seleccionado.

Debido a la capacidad predictiva de las variables independientes, para estimar las variables dependientes (Venta, Costo, Fijo y Margen), BEME definió a priori una regla que segmenta a los clientes en dos categorías. Éstas se definen dependiendo de la cantidad de informes técnicos que presenten en el periodo de Observación, con el mismo rubro que el Pivote.

La regla es la siguiente:

1. Categoría SH: Cliente sin historia es aquel que presenta menos de dos informes técnicos en el mes pivote respecto a los 36 meses móviles anteriores.
2. Categoría CH: Cliente con historia es aquel que presenta más de un informe técnico en el periodo de observación del mismo rubro indicado.

4.3. HORIZONTE DE TIEMPO

Se contempla un total de cinco años de información, a partir de enero de 2010 a noviembre de 2014, los cuales se describen a continuación:

- Periodo de Observación: Considera la información disponible los 36 meses anteriores al Informe Técnico del pivote respectivo, además de la historia financiera del cliente disponible al momento de la evaluación, según normativa vigente (enero de 2010 a noviembre de 2014).
- Pivote: Se define este concepto con el objetivo de simular el instante de la evaluación del cliente en donde se extrae tanto la información de las variables dependientes o predictoras como de las variables independientes del Informe Técnico y de las otras Bases disponibles. En el desarrollo de la memoria se consideran 59 meses pivotes (enero de 2010 a noviembre de 2014).
- Se considera que los periodos de observación son igual a los meses de pivotes porque en el primer mes de pivote ya está considerada la información observada en los 36 meses anteriores, ya que esta es la convención del BEME para la base de datos general.

4.4. LIMPIEZA DE DATOS GENERALES

De acuerdo a la metodología descrita anteriormente, una vez que la base de datos analítica está armada, y antes de la construcción de los modelos, es necesario realizar una revisión de los datos, buscando qué hacer con los datos faltantes y sin olvidarse de la distribución de cada variable objetivo.

Se cuenta con información correspondiente a los clientes con datos históricos, los cuales presentan variables como promedio de Venta, promedio de Margen y promedio de Costo Fijo, entre otros, en los informes técnicos anteriores.

A la base general se le realizó la siguiente limpieza:

- Clientes con menos de 18 años de edad y mayores a 86 años

  Se realiza este filtro para borrar las edades que se pueden haber ingresado mal en el sistema, ya que un menor de edad no puede pedir crédito alguno. Por otro lado, no se otorgan microcréditos a personas que sobrepasen los 86 años de edad. Por lo tanto, se eliminan 178 registros.

- Clientes sin información en el sistema financiero, es decir, clientes sin registros en las bases de las cuales se informa.

  Se eliminan las observaciones de clientes que no tienen registro en las bases de la SBIF. Son 9 variables en este campo y se eliminan aquellas que tengan un valor nulo en alguno de los campos. Los registros eliminados son 3.906 casos.

- Eliminar los casos donde el cliente registra más de 1 sub segmento.

  Se eliminan 10 registros por duplicación de sub-segmentos

- Eliminar los casos duplicados

  Se eliminan 10 registros por duplicación de datos por error en la construcción de la base.

- Recodificación de valores faltantes

  Para las variables que presentan valores faltantes, si su naturaleza es porque no existe información asociada a los clientes, se asigna el valor equivalente a cero. Por ejemplo, para la variable el ingreso del hijo, la ausencia de dicha información es debido a que el cliente no tiene hijo o no hay ingreso asociado al hijo. Sin embargo, si la causa de los valores faltantes es debido a la pérdida de la información asociada, se dejará el campo vacío y se categorizará la variable, ya sea continúa o categórica, dejando como si fuera un segmento más de la variable.

  Posterior a este paso de eliminación, la base queda con 95,56% de la base original

5. CONSTRUCCIÓN DE MODELOS

5.1. FILTROS Y ELIMINACIÓN DE VARIABLES

Una vez limpiada la base de datos analítica general, el paso siguiente es eliminar las variables correlacionadas de los distintos modelos de acuerdo a la metodología, para eso primero se separa por un lado una base de datos para clientes con historia (CH), y por otra, una de clientes sin historia (SH).

Para los clientes sin historia se eliminan 33 variables, las cuales corresponden a variables de los informes anteriores que tiene el cliente y se debe tomar valor 0 por ser clientes sin historias, por lo tanto, se excluyen en la estimación de los modelos SH. Estas variables se encuentra en el Anexo II.

Después, se separan las bases de datos CH y SH respectivamente en 3 bases distintas, dependiendo de la variable predictora a estimar, es decir, las bases de Venta CH, Venta SH, Costo Fijo CH, Costo Fijo SH, Margen CH y Margen SH.

En los modelos de Costo Fijo se eliminan 6 variables que forman parte de la variable dependiente, ya que la variable Costo Fijo está conformada por las otras 6 variables.

Posteriormente, se eliminan los clientes que registran el 2,5% inferior y el 97,5% superior del segmento en el caso que la variable dependiente sea Margen, y el 5% superior cuando se trata de estimar Venta y Costo Fijo. El objetivo es excluir los casos *outlier* que se puede presentar en las bases. El objetivo es eliminar los sesgos que presenta la distribución para la muestra de modelamiento o entrenamiento

A continuación, a modo de ejemplo se presenta la distribución de la variable dependiente del modelo Venta CH antes y después de realizar este filtro, el resto se encuentran en el Anexo III.

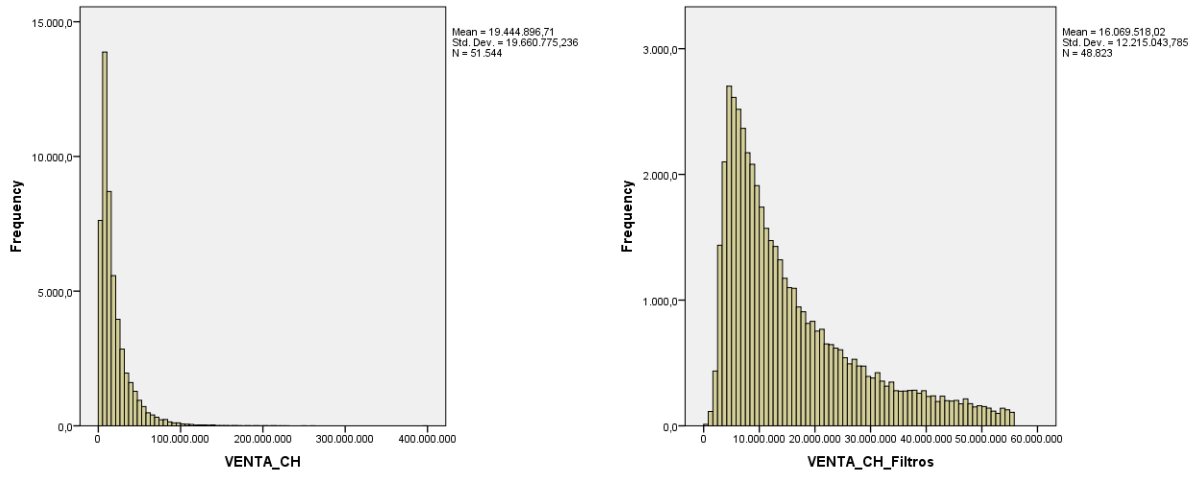


Ilustración 5: Distribución de las variables dependientes

Para la estimación de los modelos, se separa la base por muestra estatificada, donde el 70% de los datos corresponde a las bases de entrenamiento y el 30% restante a las bases de validación, prácticamente frecuente en la estimación y validación del modelo.

Sobre las bases de entrenamiento, estas son donde se construyen los distintos modelos.

Para los seis modelos, en primer lugar, se realizó un filtro de correlaciones entre las variables continuas y la variable respuesta; luego se eliminaron todas aquellas que tienen un estadístico Pearson entre -0,2 y 0,2; ya que se considera que al existir una correlación tan baja es muy poco probable que tengan alguna dependencia lineal entre ellas.

Además, se eliminaron todas aquellas variables continuas que correlacionan más de 50% entre ellas, dejando aquellas variables que tenga mayor correlación con la variable respuesta; ya que una correlación mayor a 50% significa que las variables son muy semejantes y no aporta mucho el agregar ambas en la análisis de regresión posterior.

Luego, se categorizaron las variables continuas con valores faltantes y las variables categóricas con respeto a la variable dependiente con el algoritmo CHAID exhaustivo de árbol de clasificación en el programa estadístico SPSS, donde se obtuvieron grupos por cada variable. Se eliminaron todas aquellas variables que presentan menos de dos segmentos, ya que este tipo de variable no sirve para discriminar en el modelo.

Posteriormente, se realizó la correlación entre las variables categóricas con la función análisis de tabla de contingencia, con lo cual se eliminaron variables que tuvieran un estadístico R de Pearson menor a -0,5 o mayor a 0,5, dejando aquellas variables que tienen mayor explicación con la variable dependiente, esto se sabe al realizar la regresión lineal multivariada y siendo consistente con lo utilizado anteriormente.

Finalmente, las cantidades de observaciones y variables con las que se trabajará en cada uno de los modelos se puede observar en la tabla 3 y la tabla 4. Se observa que el porcentaje de datos eliminados en las distintas bases es parejo y no supera el 10%, independiente de la cantidad de datos que tienen, también se puede apreciar que los modelos tiene distinta cantidad de variables, eso se debe a las distintas correlaciones que existen con las distintas variables dependientes.

Observaciones	Muestra	Validación	Porcentaje de datos eliminados
Venta Con Historia	34.321	14.502	9,4%
Venta Sin Historia	14.344	6.141	9,6%
Costo Fijo Con Historia	34.255	14.696	9,1%
Costo Fijo Sin Historia	14.325	6.147	9,7%
Margen Con Historia	34.262	14.701	9,1%
Margen Sin Historia	14.334	6.151	9,6%

Tabla 4: Cantidad de observaciones por modelo

	Cantidad de Variables
Venta Con Historia	63
Venta Sin Historia	54
Costo Fijo Con Historia	54
Costo Fijo Sin Historia	47
Margen Con Historia	34
Margen Sin Historia	34

Tabla 5: Cantidad de variables que se ingresan por modelo

5.2. MODELOS PARA LA ESTIMACIÓN DE LAS VARIABLES DEPENDIENTES

Dado que algunas de las variables a predecir presentas una distribución asintótica, es necesario aplicar la transformación box-cox mencionada anteriormente para normalizarlas y que estas cumplan con el criterio de normalidad. Las variables dependientes que requieren de la transformación son: Venta CH, Venta SH, Costo Fijo CH, Costo Fijo SH.

Para estimar los parámetros lambdas asociados a la transformación, se utiliza el paquete complementario de Excel, QI Macro 2015, el cual incluye dicha transformación.

El parámetro lambda resultante para todos los modelos que requieren transformación es de 0,5; lo cual equivale a una transformación con la raíz cuadrada.

A continuación, se realiza un análisis descriptivo de estas variables dependientes, antes y después de la transformación, además del análisis de las variables Margen CH y Margen SH, los cuales arrojan los siguientes resultados:

	Promedio	Mediana	Moda	Mínimo	Máximo
Venta CH	16.043.983	11.985.030	9.450.000	352.800	55.809.000

Tabla 6: Análisis descriptivo Venta con Historia

	Promedio	Mediana	Moda	Mínimo	Máximo
Venta CH Transformada	3.741	3.461	3074	594	7.470

Tabla 7: Análisis descriptivo Venta con Historia transformada

	Promedio	Mediana	Moda	Mínimo	Máximo
Venta SH	10.623.187	7.965.000	4.860.000	8100	38.961.000

Tabla 8: Análisis descriptivo Venta sin Historia

	Promedio	Mediana	Moda	Mínimo	Máximo
Venta SH Transformada	3.063	2.822	2.204	90	6.241

Tabla 9: Análisis descriptivo Venta con sin transformada

	Promedio	Mediana	Moda	Mínimo	Máximo
Cosot Fijo CH	1.522.720	841.800	345.000	8290	7.907604

Tabla 10: Análisis descriptivo Costo Fijo con Historia

	Promedio	Mediana	Moda	Mínimo	Máximo
Costo Fijo CH Transformada	1.091	917	587	91	2812

Tabla 11: Análisis descriptivo Costo Fijo con Historia transformada

	Promedio	Mediana	Moda	Mínimo	Máximo
Cosot Fijo SH	936.785	579.600	276.000	138	5.25900

Tabla 12: Análisis descriptivo Costo Fijo con Historia

	Promedio	Mediana	Moda	Mínimo	Máximo
Costo Fijo SH Transformada	869	761	525	12	2293

Tabla 13: Análisis descriptivo Costo Fijo con Historia transformada

	Promedio	Mediana	Moda	Mínimo	Máximo
Margen CH	0,505	0,503	0,56	0,338	0,6991

Tabla 14: Análisis descriptivo Margen con Historia

	Promedio	Mediana	Moda	Mínimo	Máximo
Margen SH	0,53	0,5277	0,56	0,354	0,7358

Tabla 15: Análisis descriptivo Margen sin Historia

Además se presenta gráficamente la distribución de las variables antes y después de realizar la transformación, así también de la distribución de las variables Margen CH y Margen SH.

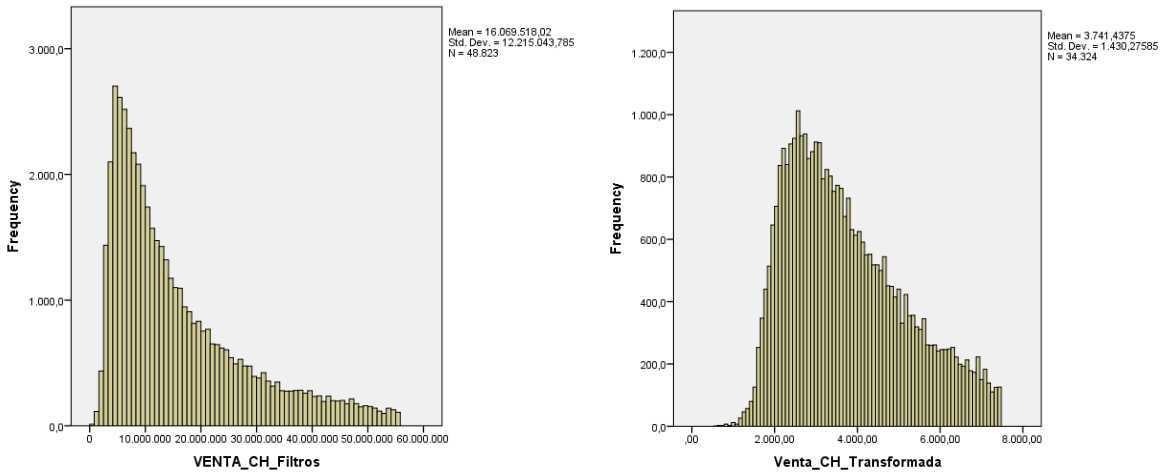


Ilustración 6:Distribución de Venta CH antes y después de la transformación

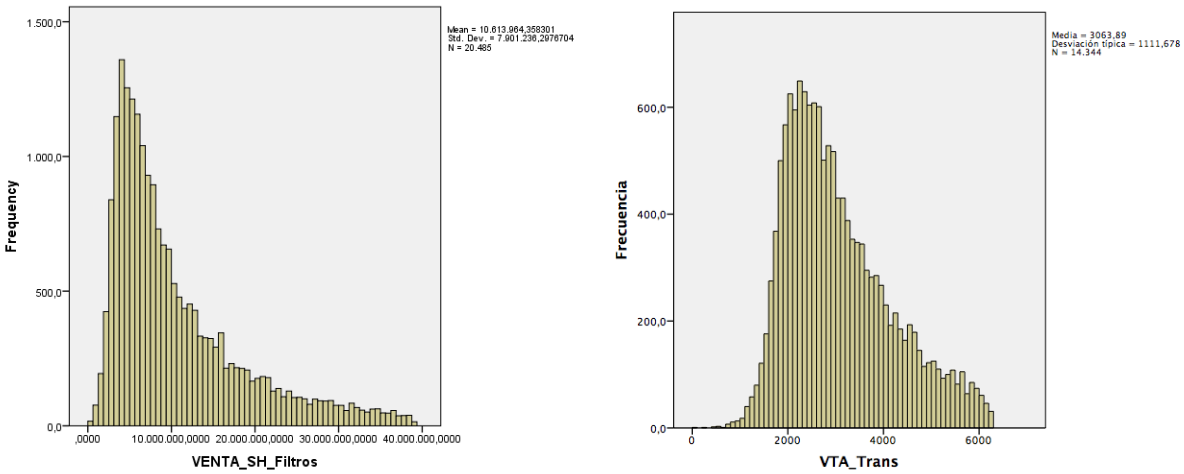


Ilustración 7:Distribución de Venta SH antes y después de la transformación

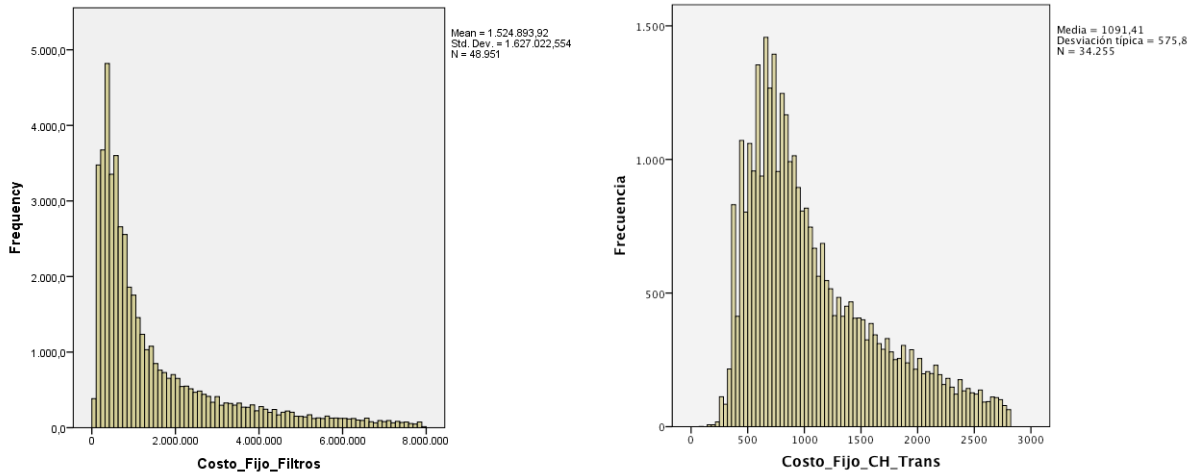


Ilustración 8:Distribución de Costo Fijo CH antes y después de la transformación

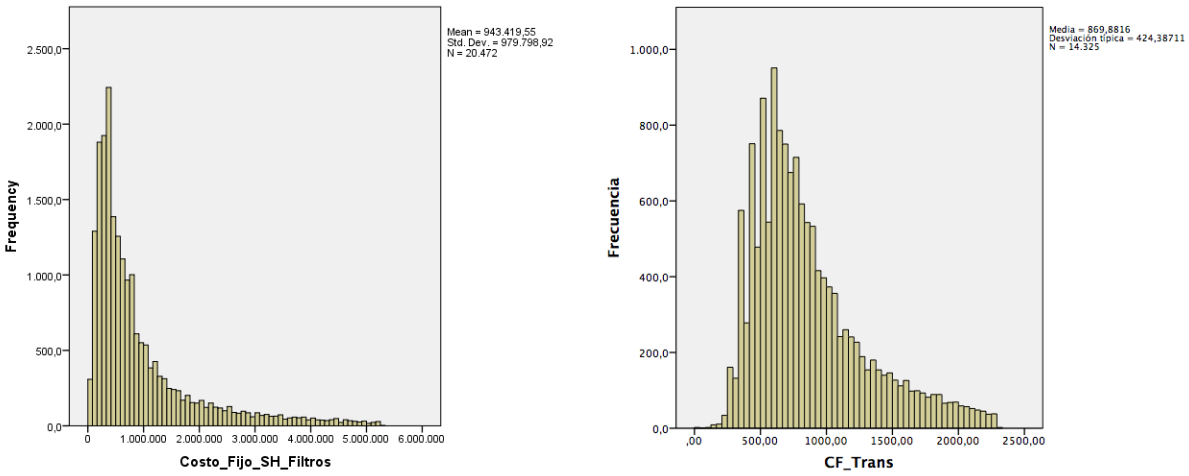


Ilustración 9:Distribución de Costo Fijo SH antes y después de la transformación

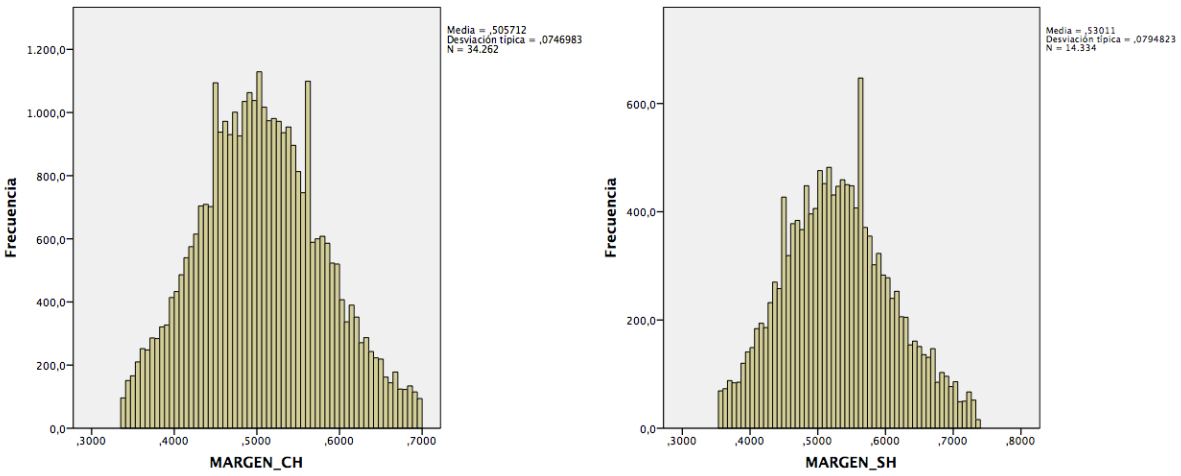


Ilustración 10:Distribución de Margen CH y Margen SH

Luego, se estiman las variables dependientes a través de la regresión lineal generalizada. Las variables resultantes del modelo Venta CH se pueden apreciar en Anexo IV, las de Venta SH en Anexo V, Costo Fijo CH en Anexo VI, Costo Fijo SH en Anexo VII, Margen CH en Anexo VIII y Margen SH en Anexo XI. Se debe tener en consideración que todas aquellas variables que termina en _N son variables categóricas que están agrupadas por sus segmentos.

A raíz de que los seis modelos resultantes presentan más de 30 variables en cada uno de ellos, y algunas de ellas pueden no presentar valor agregada en el modelo, se decidió buscar el subgrupo más explicativo.

Por dicha razón, se realiza una regresión lineal multivariada y se debe transformar los n segmentos de las variables categóricas en n-1 variables *dummy* independientes para poder ingresar al modelo en el programa SPSS.

El algoritmo ocupado para dicha regresión es fordward stepwise regression (3.8.1.), donde comienza con un modelo sin ninguna variable explicativa y en cada etapa se genera un nuevo modelo, para lo cual se agrega la variable más explicativa hasta una cierta regla de parada, en este caso, cuando las variables ya no son significativas.

El resultado de la regresión lineal multivariada entrega una lista de 85 posibles modelos para la variable Venta CH, unos 61 posibles modelos para Venta SH, 73 para Costo Fijo CH, 45 para Costo Fijo SH, 63 para Margen SH y 33 para Margen SH. Las variables que se han introducido en cada uno de los modelos se puede observar en Anexo X, Anexo XI, Anexo XII, Anexo XIII, Anexo XIV y Anexo XV, respectivamente.

A continuación, para comparar los diferentes posibles modelos de las 6 variables dependientes, se ocupa el criterio de información bayesiano (BIC). Finalmente, se eligió como se ven en las siguientes ilustraciones, aquel modelo que presenta una variabilidad menor a 0,05% con respecto al post ingresar una variable adicional.

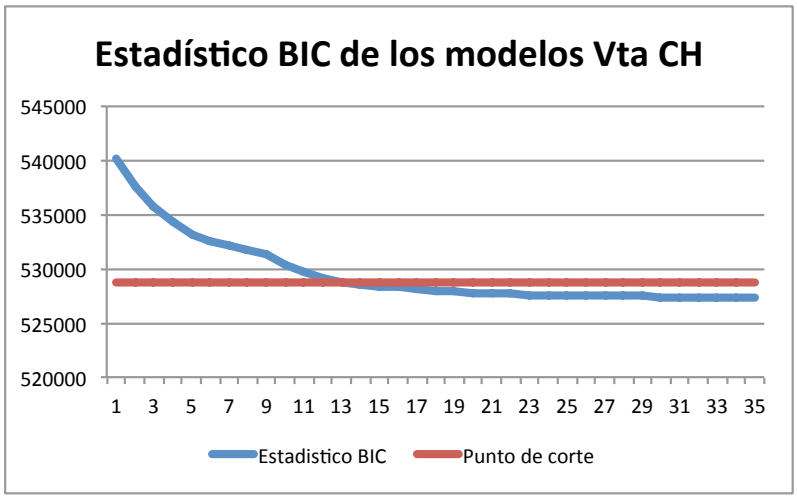


Ilustración 11:Estadístico BIC, Modelo Venta CH

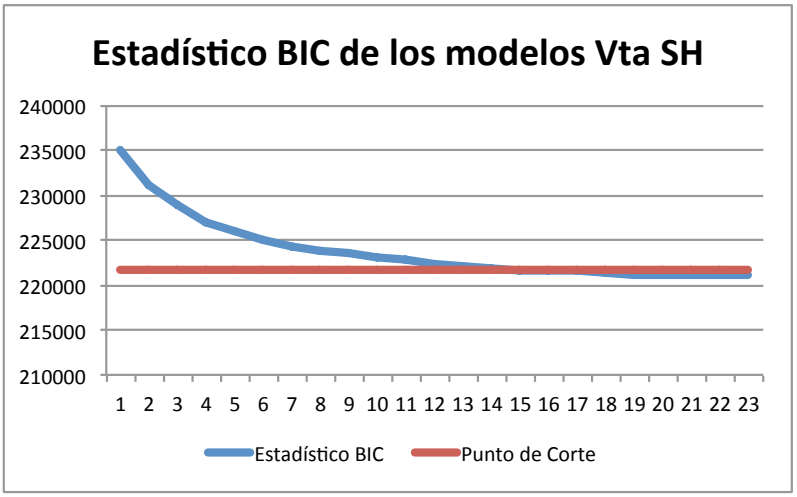


Ilustración 12:Estadístico BIC, Modelo Venta SH

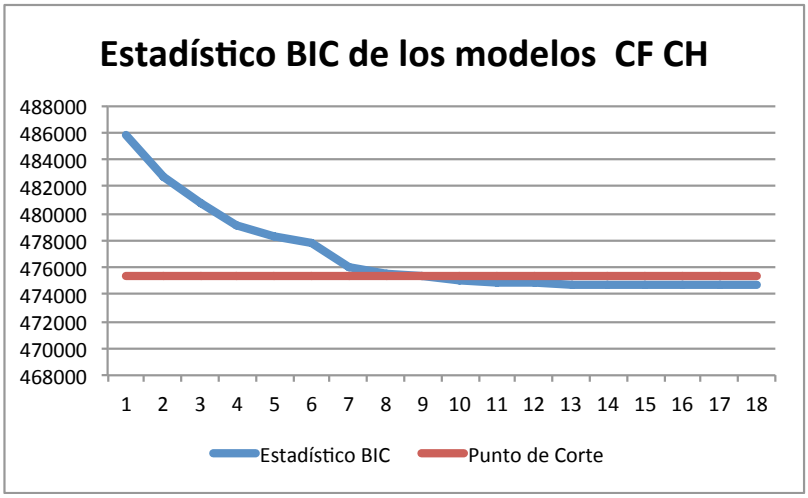


Ilustración 13:Estadístico BIC, Modelo Costo Fijo CH

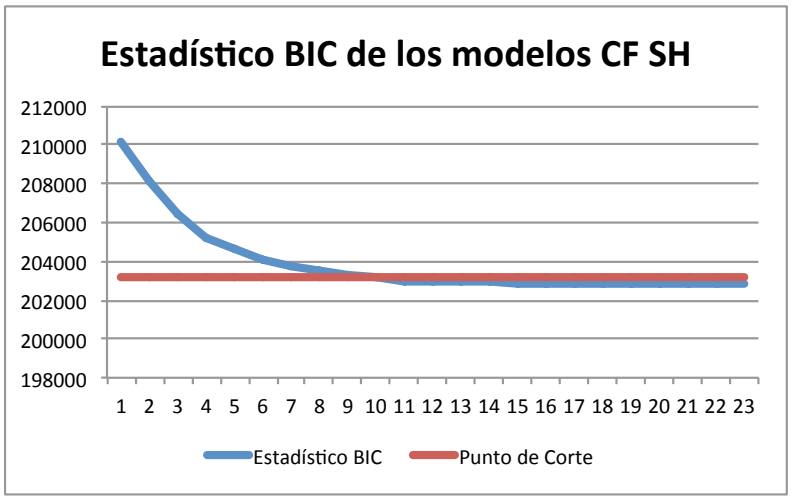


Ilustración 14:Estadístico BIC, Modelo Costo Fijo SH

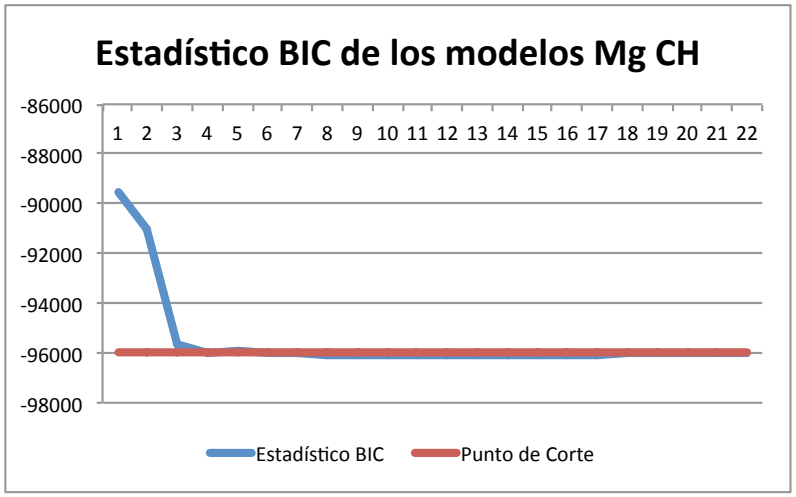


Ilustración 15:Estadístico BIC, Modelo Margen CH

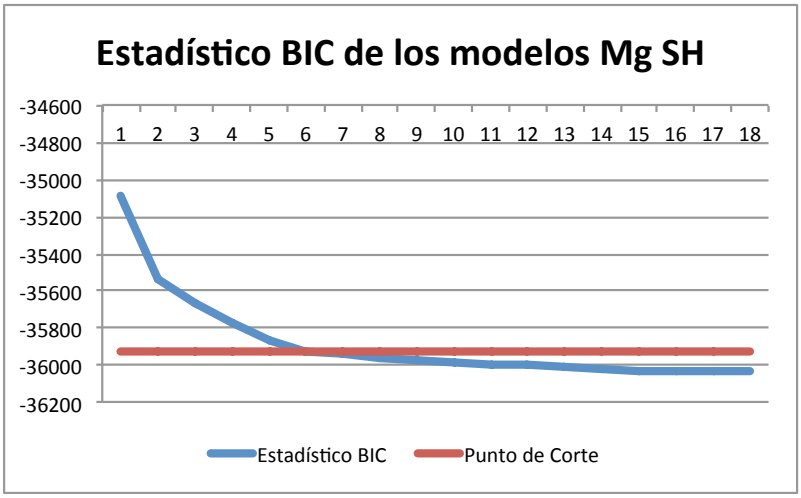


Ilustración 16:Estadístico BIC, Modelo Margen SH

El modelo elegido para la variable Venta CH se muestra en la Tabla Nº 16, donde se pueden observar los coeficientes de las variables del modelo. Dado que las variables continuas presentan comportamiento similar a la variable dependiente, dentro de la base, es decir, tienen

una distribución asimétrica a la izquierda, también se realiza una transformación con la raíz cuadrada para ingresar al modelo. La distribución de estas variables se pueden observar en el Anexo XVI.

Parámetros	Betas	Típ. Error	Chi-cuadrado	Sig.
Contante	392,99	17,08	529,60	0,00
Tr_vta_ant	0,69	0,00	43011,26	0,00
Tr_MONTO_CUOTA	0,12	0,00	1294,69	0,00
Tr_OTROS_ACTIVOS_CIRCULANTES	0,13	0,00	1975,63	0,00
Tr_PRODUCTOS_ALMACENADOS	0,10	0,00	829,32	0,00
Tr_MOVILIZACION	0,86	0,04	380,95	0,00
Tr_ARRIENDOS	0,06	0,01	105,88	0,00
Tr_DEUDA_INDAP	0,04	0,00	80,89	0,00
Tr_CAJA	0,15	0,01	522,36	0,00
Tr_TOTAL_ACTIVOS	0,04	0,00	1013,05	0,00
Tr_Tenencia_ARRENDATARIO	54,37	1,53	1255,45	0,00
Tr_SERVICIOS	0,409	0,0443	85,052	0
Formalidad= No es formal	-156,81	13,08	143,78	0,00
Formalidad= De 2 a 5 años	-29,27	13,54	4,68	0,03
Formalidad= más de 5 años	-94,00	13,24	50,37	0,00
Ptje_Amb_SMRT01 <= 616	157,23	12,05	170,28	0,00
616< Ptje_Amb_SMRT01<= 700	168,17	11,94	198,45	0,00
700< Ptje_Amb_SMRT01<= 759	126,76	10,25	153,01	0,00
759< Ptje_Amb_SMRT01<= 787	73,70	12,43	35,13	0,00
787< Ptje_Amb_SMRT01<= 805	66,57	11,24	35,10	0,00
805< Ptje_Amb_SMRT01<= 823	50,09	11,60	18,65	0,00
823< Ptje_Amb_SMRT01<= 868 o no tiene información	21,64	9,95	4,73	0,03

Tabla 16: Variables de Venta CH

El coeficiente de determinación para este modelo es de 86,7%, y un coeficiente de 85,6% para la validación del modelo, es decir, el modelo explica el 86% de la variabilidad.

En la Tabla Nº 17 se pueden observar los coeficientes de las variables del modelo elegido para Venta SH. Dado que algunas variables continuas presentan comportamiento similar a la variable dependiente dentro de la base, es decir, tiene una distribución asimétrica a la izquierda, también se realiza una transformación con la raíz cuadrada para ingresar al modelo con la excepción de la variable Gasto_ME (Gasto Familiar) que presenta una distribución normal. La distribución de las variables se pueden observar en el Anexo XVII.

Parámetros	Betas	Típ. Error	Chi-cuadrado	Sig.
(Intersección)	960,277	20,4728	2200,085	0
Tr_OPE_MONTO_ORIGEN_PES	0,346	0,0072	2293,077	0
Tr_TOTAL_ACTIVOS_CIRCULANTES	0,231	0,0062	1380,598	0
Tr_ARRIENDO_TERRENO	1,275	0,0326	1530,366	0
Formalidad = No es Formal	-200,143	12,5683	253,589	0
Formalidad = Menos de 1 Año	-72,255	21,1979	11,619	0,001
Formalidad = De 2 a 5 Años	-34,72	14,6653	5,605	0,018

Tr_INSUMOS	0,242	0,0077	999,908	0
Tr_MOVILIZACION	1,723	0,0721	571,339	0
Tr_OTROS_ACTIVOS_CIRCULANTES	0,147	0,006	606,364	0
Tr_SERVICIOS	1,207	0,0704	294,042	0
Tr_PRODUCTOS_ALMACENADOS	0,127	0,0072	308,155	0
Gastos _ME	0,002	8,74E-05	392,178	0
Tr_SUELDOS_FIJOS	1,312	0,0529	615,861	0
Tr_DEUDA_INDAP	0,077	0,0085	80,908	0
Tr_OTROS	0,999	0,0684	213,564	0
Tr_TOTAL_PASIVOS	0,053	0,0046	133,118	0

Tabla 17: Variables de Venta SH

El coeficiente de determinación para este modelo es de 76,8%, y un coeficiente de 76,6% para la validación del modelo, es decir, el modelo explica el 77% de la variabilidad.

En la Tabla Nº 18 se pueden observar los coeficientes de las variables del modelo elegido para Costo Fijo CH. Dado que algunas variables continuas presentan comportamiento similar a la variable dependiente, dentro de la base, es decir, tiene una distribución asimétrica a la izquierda, también se realiza una transformación con la raíz cuadrada para ingresar al modelo con la excepción de la variable Gasto_ME (Gasto Familiar) que presenta una distribución normal. La distribución de las variables pueden observar en Anexo XVIII.

Parámetros	Betas	Típ. Error	Chi-cuadrado	Sig.
Costante	92,712	7,737	143,59	0
Tr_cf_ant	0,759	0,003	62601,484	0
Formalidad= No es Formal	-40,022	3,4584	133,924	0
Formalidad=Menos de 1 Año	-6,782	3,289	4,252	0,039
Tr_ARRIENDOS	0,061	0,0025	582,554	0
Tr_TOTAL_ACTIVOS_CIRCULANTES	0,036	0,0012	882,07	0
Gastos_ME	0,0004	0,00002	407,155	0
Propiedad_Vivienda= PROPIETARIO	-12,778	6,2297	4,207	0,04
Tr_Tenencia_ARRENDATARIO	27,461	0,7249	1435,183	0
Tr_MONTO_TOTAL_ANT	0,015	0,0018	73,712	0

Tabla 18: Variables de Costo Fijo CH

El coeficiente de determinación para este modelo es de 81,3%, y un coeficiente de 81,2% para la validación del modelo, es decir, el modelo explica el 81% de la variabilidad.

En la Tabla Nº 19 se puede observar los coeficientes de las variables del modelo para Costo Fijo SH. Dado que algunas variables continuas presentan comportamiento similar a la variable dependiente, dentro de la base, es decir, tiene una distribución asimétrica a la izquierda, a estas también se les realiza una transformación con la raíz cuadrada para ingresar al modelo con la excepción de la variable Gasto_ME(Gasto Familiar) que presenta una distribución normal. La distribución de las variables pueden observar en Anexo XIX.

Parámetros	Betas	Típ. Error	Chi-cuadrado	Sig.
Constante	486,359	12,009	1640,205	0
Tr_OPE_MONTO_ORIGEN_PES	0,083	0,0039	459,816	0
Tr_ARRIENDOS	0,284	0,0077	1348,134	0
Gastos_ME	0,001	0,0000	764,004	0
Tr_TOTAL_ACTIVOS_CIRCULANTES	0,067	0,0029	518,431	0
Tr_VEHICULO	0,044	0,0026	287,066	0
Tr_INSUMOS	0,065	0,0042	242,561	0
Tr_OTRAS_MAQUINARIAS	0,041	0,0035	138,302	0
Formalidad= NO ES FORMAL	-138,254	7,7515	318,111	0
Formalidad = MENOS DE 1 Año	-20,325	8,2539	6,064	0,014
Propiedad_Vivienda=Propietario u otros	-132,582	6,9838	360,395	0

Tabla 19: Variables de Costo Fijo SH

El coeficiente de determinación para este modelo es de 47,7%, y un coeficiente de 49,5% para la validación del modelo, es decir, el modelo explica el 48% de la variabilidad.

En la Tabla Nº 20 se puede observar los coeficientes de las variables del modelo para Margen CH:

Parámetros	Betas	Típ. Error	Chi-cuadrado	Sig.
Constante	0,822	0,0061	17976,711	0
mg_ant	0,331	0,0028	14125,698	0
Tr_vta_ant	-0,027	0,0004	5494,281	0
Sub_segmento=CULTIVOS	-0,037	0,0017	480,065	0
Sub_segmento= AVICOLA o RUTICULTOR o HORTALIZAS o LECHERIA o VITIVINICULTORES	-0,028	0,0018	245,336	0
Sub_segmento=GANADERO	-0,019	0,0018	111,281	0
Formalidad = No es Formal	-0,01	0,0016	39,518	0

Tabla 20: Variables de Margen CH

El coeficiente de determinación para este modelo es de 37%, y un coeficiente de 40% para la validación del modelo, es decir, el modelo explica el 40% de la variabilidad.

En la Tabla Nº 21 se puede observar los coeficientes de las variables del modelo para Margen SH:

Parámetros	Betas	Típ. Error	Chi-cuadrado de Wald	Sig.
Constante	0,516	0,0029	30851,233	0
Sub_segmento= Cultivos o Fruticultor o Vitivinicultor	-0,047	0,0026	308,788	0
Sub_segmento= Hortaliza o Lechería	-0,039	0,0031	154,03	0
Sub_segmento= Ganadero o Avícola	-0,027	0,0028	92,485	0
INSUMO<=0	0,044	0,002	497,333	0
INSUMO>0 y INSUMO<350000	0,043	0,0018	549,792	0
INSUMO>350000 y INSUMO<1000000	0,021	0,0016	159,085	0
OTROS_ACTIVOS_CIRCULANTES <=0	0,014	0,0016	85,011	0
OTROS_ACTIVOS_CIRCULANTES >0 y OTROS_ACTIVOS_CIRCULANTES <=1200000	0,024	0,0017	201,72	0
PRODUCTOS_ALMACENADOS<=0	0,021	0,0016	170,191	0
PRODUCTOS_ALMACENADOS>0 y PRODUCTOS_ALMACENADOS<=1000000	0,023	0,0016	192,112	0

Tabla 21: Variables de Margen SH

El coeficiente de determinación para este modelo es de 15%, y un coeficiente de 13% para la validación del modelo, es decir, el modelo explica el 13% de la variabilidad.

Adicionalmente a estos indicadores, se calcula la sobre y subestimación del modelo con un nivel de significancia de un 10% para la base de datos de entrenamiento y para la de validación. En las siguientes tablas se pueden ver los resultados.

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	12.031	35,1%	4.515	31,1%
Entre 90% y 110%	13.453	39,2%	5.642	38,9%
Subestima 10%	8.837	25,7%	4.345	30,0%
Total	34.321	100%	14.503	100%

Tabla 22: Indicador de Estimación Venta CH

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	5.647	39,4%	2.614	42,6%
Entre 90% y 110%	3.822	26,7%	1.613	26,3%
Subestima 10%	4.875	33,9%	1.914	31,1%
Total	14.344	100%	14.503	100%

Tabla 23: Indicador de Estimación Venta SH

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	14.431	42,1%	6.128	41,7%
Entre 90% y 110%	11.699	34,2%	5.028	34,2%
Subestima 10%	8.125	23,7%	3.340	24,1%
Total	34.255	100%	14.696	100%

Tabla 24: Indicador de Estimación Costo Fijo CH

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	7.137	49,8%	2.784	45,3%
Entre 90% y 110%	1.881	13,1%	790	12,9%
Subestima 10%	5.307	37,1%	2.573	41,8%
Total	14.325	100%	6.147	100%

Tabla 25: Indicador de Estimación Costo Fijo SH

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	6.914	20%	3.160	21,5%
Entre 90% y 110%	22.470	65,6%	9465	64,4%
Subestima 10%	4.878	14,2%	2.076	14,1%
Total	34.262	100%	14.701	100%

Tabla 26: Indicador de Estimación Margen CH

	Entrenamiento		Validación	
Estimación v/s Real	N	%	N	%
Sobre 10%	3.760	26,2%	1.834	29,8%
Entre 90% y 110%	7.530	52,6%	3.122	50,8%
Subestima 10%	3.044	21,2%	1.195	19,4%
Total	14.334	100%	6.147	100%

Tabla 27: Indicador de Estimación Margen SH

6. ANÁLISIS DE LOS RESULTADOS

A continuación, se presentan algunos análisis obtenidos de los seis modelos presentados en el capítulo anterior.

Dado que las variables independientes en los modelos Venta CH, Venta SH, Costo Fijo CH, y Costo Fijo SH estaban transformadas, no se puede interpretar fácilmente cómo se explican las variables independientes a las variables dependiente, ya que éstas últimas están explicadas por la combinación de los X´s (de dos variables independientes).

Se observa que las constantes de los seis modelos explican bastante en los modelos, esto puede deberse a que existe sesgo en los datos, ya que las informaciones provienen de los clientes con créditos aprobados y no todos los clientes que solicitan crédito, por lo tanto, había selección de clientes antes.

La variable formalidad es una de las variables que más se repite en los modelos, mientras que el cliente se vuelve formal, este genera más nivel de ventas, pero a la vez es más costosos ser formal.

Se observa que la variable Margen tiene harto nivel de persistencia, donde la principal variable que lo explica es el valor anterior.

Los tres modelos asociados a clientes CH (clientes con al menos dos informe técnico en el horizonte de tiempo contemplado) presentan un mejor ajuste estadístico con respecto a los clientes SH (clientes con un solo informe técnico). Los mejores ajustes ordenados en forma decreciente son: Venta, Costo Fijo y Margen, para ambos segmento de clientes.

En el análisis de resultados solamente se ocupa el coeficiente de determinación para medir el ajuste de los modelos, no tiene ruido en el análisis, dado que el número de variables a comparar en modelos, no es un atributo controlable, sino resultado del proceso iterativo de selección de variables.

7. CONCLUSIONES

A continuación, se presentan las principales conclusiones tras la realización de este trabajo. Se muestran los aprendizajes que surgen directamente del trabajo, con respecto a la aplicación de la metodología, y tras ello se discute respecto los resultados obtenidos.

Sin duda, fue una experiencia muy enriquecedora conocer el proceso actual de otorgamiento de créditos a segmento agrícola, para así mejorarlo a un proceso más automático, lo cual corresponde al desarrollo de este proyecto.

Además, participar en el proceso de integración de la base de datos analítica, que es la base para el desarrollo de los modelos, ayuda a entender mejor cómo funciona el negocio en el cual se aplican los modelos, a encontrar sentidos los resultados, a ver las limitaciones que existen con los datos son muy valorado, ya que esto no se ha visto en los trabajos anteriores realizados en la universidad y es un proceso sumamente importante, ya que una mala base de daros genera malos resultados, por ende, mala interpretaciones.

En cuanto a los modelos, se observa un buen ajuste del modelo Venta y Costo Fijo con historia, pese a que se tuvieron que transformar las variables a estimar porque no cumplían los supuestos de un modelo de regresión múltiple.

En el porcentaje de pronósticos de errores, donde las valores estimados están entre el 95% y 105% del valor real de la variable de respuesta, más la subestimación en el caso de la estimación de Venta y Margen y sobreestimación en el caso de Costo Fijo, son un porcentaje mayor al 60% lo cual es muy bueno y ha de quedar satisfecho el trabajo, al cumplir el objetivo central.

Finalmente, se concluye que los modelos estimados son buenos, dado la limitaciones de los datos disponibles, ya que presenta un buen ajuste.

8. TRABAJOS FUTUROS

Dentro de los futuros trabajos, se recomienda analizar la posibilidad de incluir variables exógenas en el modelo. Estas pueden predecir cómo está la economía del país y por lo tanto, podrían servir para realizar pronósticos. Por ejemplo, el PIB podría ser un buen indicador y podría verse reflejado en el crecimiento de las ventas de microempresarios; el precio de los productos a vender puede ser una variable muy importante para estimar la capacidad de pago de los clientes, ya que a diferencia de otros sectores, el precio de los productos agrícolas varían mucho entre temporadas y el BEME puede estimar de forma propia los precios de ventas de estos productos agrícolas en la temporada donde el banco percibe el pago e incluir la variación de precio estimado como una variable exógena más.

9. BIBLIOGRAFÍA

[1]BancoEstado. www.bancoestado.cl. Revisado en 10 de 7 de 2015 desde https://www.bancoestado.cl/imagenes/corporativo/pub_pdf_microfinanzas.pdf

[2]BancoEstado Microempresa. (2013). Memoria anual. Santiago

[3]Biron, M. (Abril,2012). Desarrollo y Evaluación de Metodologías para la Aplicación de Regresiones Logísticas en Modelos de Comportamiento Bajo Supuesto de Independencia. Santiago. Memoria de Ingeniería Civil, Universidad de Chile

[4]Fernández, P. & Díaz, P. (1997). Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña. Cad Aten Primaria.

[5] Box, G. & Cox, D. Journal of the Royal Statistical Society. Series B (Methodological)Vol. 26, No. 2, pp. 211-252

[6] Talento, S. (2001). Bases para un sistema de predicción de caudales de aporte a rincón del Bonete y Salto Grande. Montevideo, Uruguay

[7] Miranda, C. (2013). Diseño de modelos econométricos para la estimación de estados financieros de microempresas que se desempeñan en servicio profesionales y manufactura. Santiago. Memoria de Ingeniería Civil, Universidad de Chile

[8] McCullagh, P. & Nelder, J. (1989). Generalized Linear Models

[9] Wooldridge,J. Introducción a la econometría, un enfoque moderno. 4ª edn

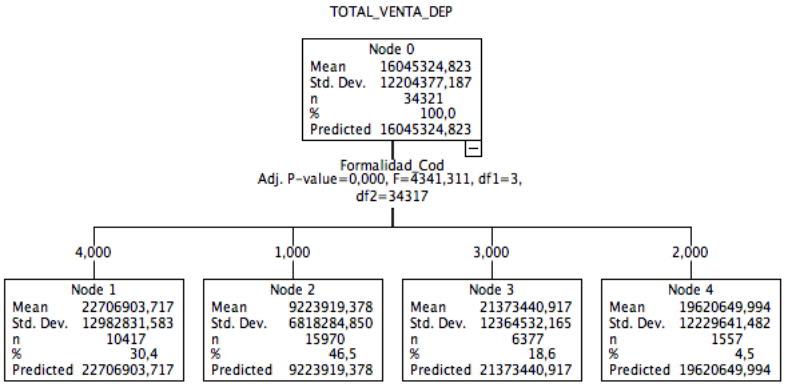
[10] Ramanathan, R. (2002). Introductory Econometrics with Applications. 5ª edn., South-Western.

[11] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 2, 461-464.

[12] Peña, D. (2002). Análisis de datos multivariantes. Madrid: McGraw-Hill.

10. ANEXOS

10.1. ANEXO I: Resultado algoritmo CHAID

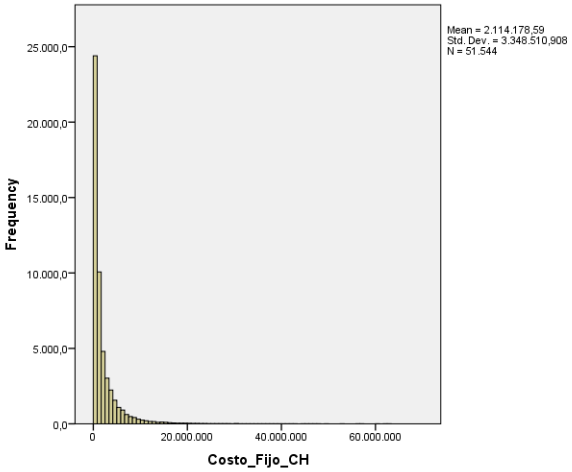


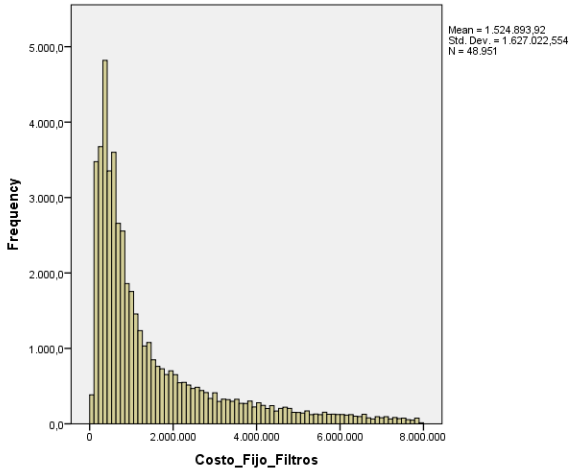
10.2. ANEXO II: Variables eliminadas para modelos Costos Fijos

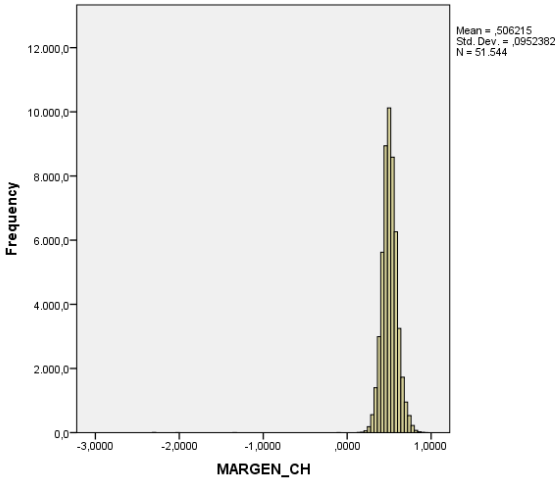
Variables
Otros_ingresos_old
Otros_ingresos_ant
Otros_ingresos_prom
DEUDAS_FAM_NEG_OLD
DEUDAS_FAM_NEG_ant
DEUDAS_FAM_NEG_PROM
cant_cuotas_OLD
MONTO_TOTAL_OLD
MONTO_CUOTA_OLD
MAX_ID
MIN_ID
vta_old
cf_old
mg_old
GF_old
vta_ant
cf_ant
mg_ant
GF_ant
vta_PROM
cf_PROM
mg_prom
GF_prom
cant_cuotas_ANT
MONTO_TOTAL_ANT
MONTO_CUOTA_ANT
CANTIDAD_IT_CLIENTE_RUBRO_APR
CANTIDAD_IT_CLIENTE_RUBRO_APRRE
CANTIDAD_IT_CLIENTE
Antiguedad_dias
Antiguedad_meses
Antiguedad_dias_old
Antiguedad_meses_old

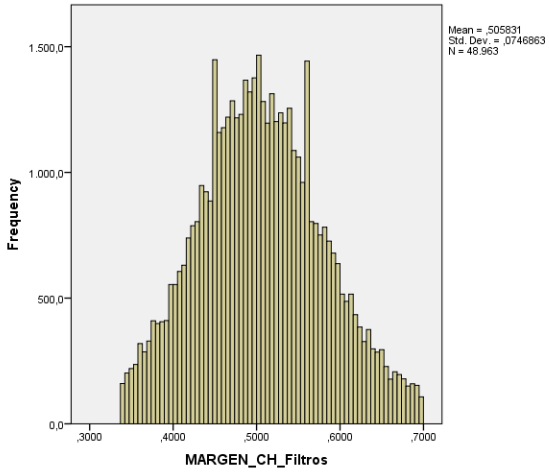
10.3. ANEXO III: Distribución de las variables dependientes

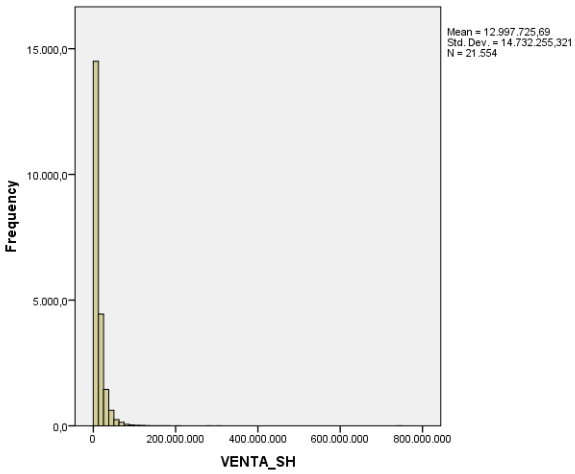
A continuación, se presenta la distribución de la variable dependiente de los modelos antes y después de realizar el filtro de outlier.

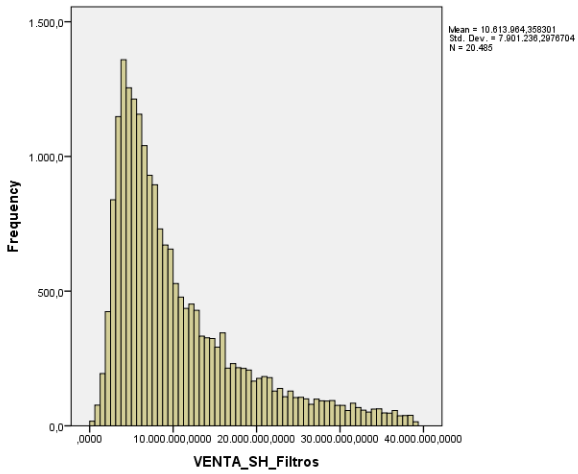


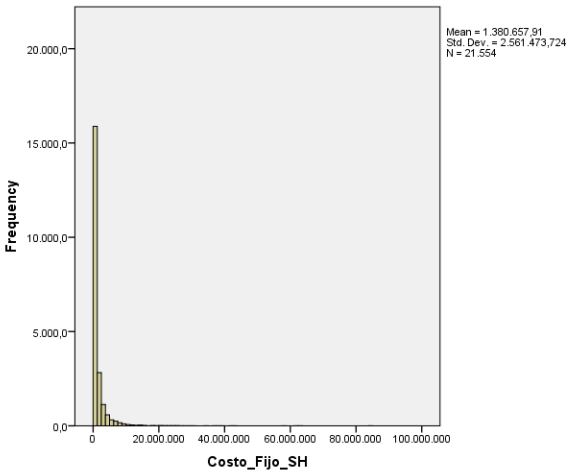


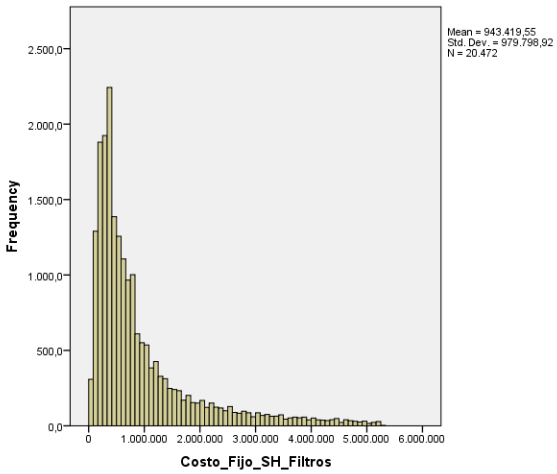


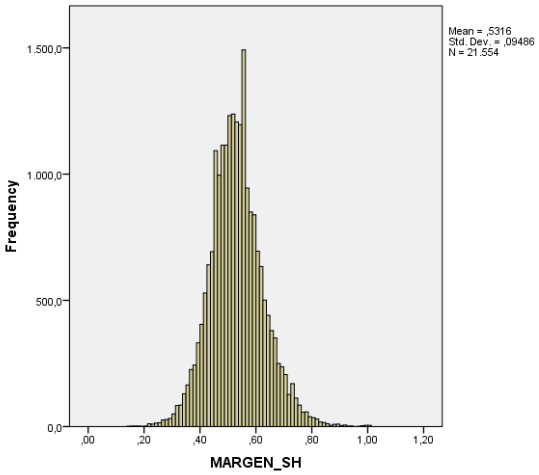


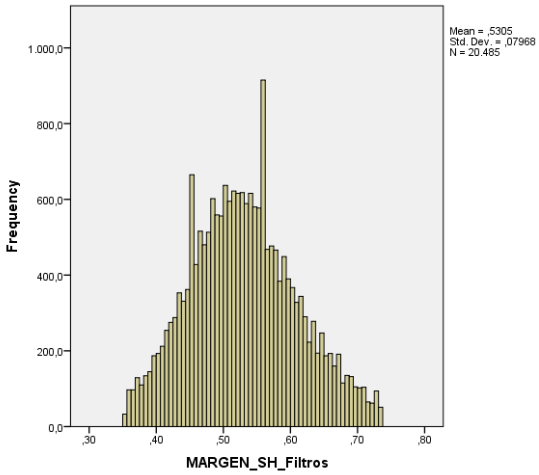





Mean = 1.380.657,91
Std. Dev. = 2.561.473,724
N = 21.554
Frequency
20.000,0
15.000,0
10.000,0
5.000,0
0,0
0
20.000.000
40.000.000
60.000.000
80.000.000
100.000.000
Costo_Fijo_SH


Mean = 943.419,55
Std. Dev. = 979.798,92
N = 20.472
Frequency
2.500,0
2.000,0
1.500,0
1.000,0
500,0
0,0
0
1.000.000
2.000.000
3.000.000
4.000.000
5.000.000
6.000.000
Costo_Fijo_SH_Filtros


Mean = ,5316
Std. Dev. = ,09486
N = 21.554
Frequency
1.500,0
1.000,0
500,0
0,0
,00
,20
,40
,60
,80
1,00
1,20
MARGEN_SH


Mean = ,5305
Std. Dev. = ,07968
N = 20.485
Frequency
1.000,0
800,0
600,0
400,0
200,0
0,0
,30
,40
,50
,60
,70
,80
MARGEN_SH_Filtros

10.4. ANEXO IV: Variables preliminares del modelo Venta CH

Variables	Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
(Intersección)	29.496,859	1	0,000
Sub_segmento_Cod_N	100,937	3	,000
sexo_N	34,371	2	,000
Region_N	57,172	7	,000
Comuna_N	33,017	9	,000
Ptje_Amb_SMRT01_N	845,533	7	,000
Mora_6M_N	5,927	1	,015
Ptje_SICA_N	186,336	8	,000
Formalidad_Cod_N	398,948	3	,000
Forma_de_Pago_Cod_N	11,940	3	,008
BAF_Cod_N	31,515	1	,000
Tipo_registro_Cod_N	32,916	3	,000
Gestion_Cod_N	4,807	1	,028
Propiedad_Vivienda_Cod_N	31,847	2	,000
Evolucion_Ahorro_Cod_N	21,774	2	0,000
prov_ult_3a_Cod_N	12,361	1	0,000
nuevo_antiguo_Cod_N	7,473	1	0,006
Modulo_Cod_N	118,235	9	0,000
Codofi_N	347,447	14	0,000
LCR_Cod_N	23,678	1	0,000
Cheq_Cod_N	11,545	1	0,001
max_dmora_6M_N	92,696	4	,000
vta_ant	14.520,431	1	,000
OTROS_ACTIVOS_CIRCULANTES	1.828,414	1	,000
MONTO_CUOTA	1.714,085	1	,000
TOTAL_ACTIVOS	218,967	1	,000
MOVILIZACION	355,170	1	,000
DEUDAS_FAM_NEG_PROM	50,208	1	,000
OTRAS_MAQUINARIAS	70,694	1	,000
GASTOS_FAMILIARES	32,113	1	,000
DEUDA_BECH	192,582	1	0,000
TOTAL_PASIVOS_LARGO_PLAZO	78,816	1	0,000
CAJA	249,220	1	0,000
PRODUCTOS_ALMACENADOS	1.139,415	1	0,000
CANTIDAD_IT_CLIENTE_TODAS	48,205	1	0,000
ARRIENDOS	345,975	1	0,000
SERVICIOS	205,234	1	0,000
MONTO_CUOTA_OLD	7,273	1	0,007
LC_disp	16,616	1	0,000
SUELDOS_FIJOS	57,969	1	0,000
OTROS_ACTIVOS_FIJOS	8,126	1	0,004
DEUDA_OTROS_BANCOS	64,927	1	0,000

RASTRA	10,658	1	0,001
DEUDA_INDAP	304,967	1	0,000
Tenencia_ARRENDATARIO	281,308	1	0,000
AHORRO	141,648	1	0,000
PROVEEDORES	107,543	1	0,000
OTROS	140,9	1	0

10.5. ANEXO V: Variables preliminares del modelo Venta SH

Variables	Chi-cuadrado	gl	Sig.
(Intersección)	14777,401	1	0
Sub_segmento_Cod_N	37,324	3	0
sexo_N	19,211	2	0
Region_N	76,191	5	0
Comuna_N	85,225	7	0
Ptje_Amb_SMRT01_N	241,934	6	0
Cod_giro_N	30,251	2	0
Ptje_SICA_N	72,169	5	0
Formalidad_Cod_N	156,618	3	0
Forma_de_Pago_Cod_N	32,14	2	0
BAF_Cod_N	13,361	1	0
Asistencia_Tecnica_Cod_N	8,064	2	0,018
Tipo_registro_Cod_N	34,467	2	0
Propiedad_Vivienda_Cod_N	9,985	2	0,007
Comportamiento_Beme_IT_Cod_N	19,321	3	0
prov_ult_3a_Cod_N	7,131	1	0,008
Modulo_Cod_N	100,679	3	0
Codofi_N	320,585	10	0
LCR_Cod_N	47,117	1	0
Cheq_Cod_N	11,669	1	0,001
max_dmora_6M_N	7,85	1	0,005
OPE_MONTO_ORIGEN_PES	1015,821	1	0
TOTAL_ACTIVOS_CIRCULANTES	362,797	1	0
Gastos_ME	250,91	1	0
INSUMOS	1175,079	1	0
TOTAL_PASIVOS	9,048	1	0,003
ARRIENDO_TERRENO	1215,122	1	0
OTROS_ACTIVOS_CIRCULANTES	704,409	1	0
MOVILIZACION	410,219	1	0
PATRIMONIO	91,673	1	0
OTRAS_MAQUINARIAS	67,346	1	0
SERVICIOS	181,704	1	0
CAJA	43,895	1	0
PRODUCTOS_ALMACENADOS	382,299	1	0
SUELDOS_FIJOS	261,035	1	0
DEUDA_BECH	22,779	1	0
DEUDA_INDAP	247,065	1	0
OTROS	163,919	1	0
LC_disp	84,342	1	0
AHORRO	4,981	1	0,026
DEUDA_OTROS_BANCOS	9,094	1	0,003

10.6. ANEXO VI: Variables preliminares del modelo Costo Fijo CH

Variables	Chi-cuadrado	gl	Sig.
(Intersección)	15510,719	1	0
Sub_segmento_Cod_N	68,081	3	0
sexo_N	60,139	2	0
Region_N	111,177	5	0
Comuna_N	111,757	11	0
Ptje_Amb_SMRT01_N	239,024	8	0
Mora_6M_N	5,141	1	0,023
Ptje_SICA_N	58,807	9	0
Permanencia_Cod_N	25,63	2	0
Formalidad_Cod_N	297,465	2	0
Nro_clientes_Cod_N	39,298	2	0
Asistencia_Tecnica_Cod_N	10,777	2	0,005
Tipo_registro_Cod_N	40,982	2	0
Propiedad_Vivienda_Cod_N	154,993	3	0
Manejo_Ahorro_Cod_N	4,109	1	0,043
Comportamiento_Beme_IT_Cod_N	12,747	3	0,005
nuevo_antiguo_Cod_N	14,204	1	0
Modulo_Cod_N	69,402	7	0
Codofi_N	660,214	14	0
LCR_Cod_N	12,375	1	0
max_dmora_6M_N	40,851	4	0
cf_ant	20853,824	1	0
TOTAL_ACTIVOS_CIRCULANTES	545,505	1	0
MONTO_TOTAL_ANT	23,56	1	0
ARRIENDOS	1805,886	1	0
INSUMOS	53,168	1	0
Gastos_ME	203,378	1	0
vta_PROM	19,472	1	0
VEHICULO	44,15	1	0
MONTO_CUOTA	153,866	1	0
OTRAS_MAQUINARIAS	96,829	1	0
Tenencia_ARRENDATARIO	542,396	1	0
TOTAL_PASIVOS_LARGO_PLAZO	7,645	1	0,006
CANTIDAD_IT_CLIENTE_RUBRO_APRRE	4,525	1	0,033
PRODUCTOS_ALMACENADOS	10,76	1	0,001
MONTO_CUOTA_OLD	29,653	1	0
OTROS_ACTIVOS_FIJOS	7,38	1	0,007

10.7. ANEXO VII: Variables preliminares del modelo Costo Fijo SH

Variables	Chi-cuadrado	gl	Sig.
(Intersección)	4979,208	1	0
sexo_N	42,024	2	0
Region_N	29,538	4	0
Comuna_N	30,853	7	0
Ptje_Amb_SMRT01_N	82,246	6	0
Mora_6M_N	4,352	1	0,037
Ptje_SICA_N	14,158	3	0,003
Formalidad_Cod_N	183,81	3	0
Nro_clientes_Cod_N	19,901	1	0
BAF_Cod_N	19,03	1	0
Asistencia_Tecnica_Cod_N	12,473	2	0,002
Tipo_registro_Cod_N	46,231	2	0
Propiedad_Vivienda_Cod_N	295,476	1	0
Manejo_Ahorro_Cod_N	17,364	1	0
Comportamiento_Beme_IT_Cod_N	18,741	3	0
ComportamientoSBFI_Cod_N	5,141	1	0,023
Modulo_Cod_N	120,882	5	0
Codofi_N	184,166	10	0
LCR_Cod_N	8,819	1	0,003
Cheq_Cod_N	8,304	1	0,004
max_dmora_6M_N	7,306	1	0,007
OPE_MONTO_ORIGEN_PES	182,353	1	0
Gastos_ME	392,531	1	0
TOTAL_ACTIVOS_CIRCULANTES	158,208	1	0
INSUMOS	181,211	1	0
ARRIENDOS	1344,133	1	0
VEHICULO	138,295	1	0
OTROS_ACTIVOS_CIRCULANTES	29,782	1	0
OTRAS_MAQUINARIAS	121,848	1	0
CAJA	7,631	1	0,006
PRODUCTOS_ALMACENADOS	51,933	1	0
OTROS_ACTIVOS_FIJOS	23,493	1	0
OTROS_INGRESOS_TOTAL	23,567	1	0
DEUDA_BECH	6,34	1	0,012

10.8. ANEXO VIII: Variables preliminares del modelo Margen CH

Variables	Chi-cuadrado	gl	Sig.
(Intersección)	29.496,859	1	0,000
Sub_segmento_Cod_N	100,937	3	,000
sexo_N	34,371	2	,000
Region_N	57,172	7	,000
Comuna_N	33,017	9	,000
Ptje_Amb_SMRT01_N	845,533	7	,000
Mora_6M_N	5,927	1	,015
Ptje_SICA_N	186,336	8	,000
Formalidad_Cod_N	398,948	3	,000
Forma_de_Pago_Cod_N	11,940	3	,008
BAF_Cod_N	31,515	1	,000
Tipo_registro_Cod_N	32,916	3	,000
Gestion_Cod_N	4,807	1	,028
Propiedad_Vivienda_Cod_N	31,847	2	,000
Evolucion_Ahorro_Cod_N	21,774	2	0,000
prov_ult_3a_Cod_N	12,361	1	0,000
nuevo_antiguo_Cod_N	7,473	1	0,006
Modulo_Cod_N	118,235	9	0,000
Codofi_N	347,447	14	0,000
LCR_Cod_N	23,678	1	0,000
Cheq_Cod_N	11,545	1	0,001
max_dmora_6M_N	92,696	4	,000
vta_ant	14.520,431	1	,000
OTROS_ACTIVOS_CIRCULANTES	1.828,414	1	,000
MONTO_CUOTA	1.714,085	1	,000
TOTAL_ACTIVOS	218,967	1	,000
MOVILIZACION	355,170	1	,000
DEUDAS_FAM_NEG_PROM	50,208	1	,000
OTRAS_MAQUINARIAS	70,694	1	,000
GASTOS_FAMILIARES	32,113	1	,000
DEUDA_BECH	192,582	1	0,000
TOTAL_PASIVOS_LARGO_PLAZO	78,816	1	0,000
CAJA	249,220	1	0,000
PRODUCTOS_ALMACENADOS	1.139,415	1	0,000
CANTIDAD_IT_CLIENTE_TODAS	48,205	1	0,000
ARRIENDOS	345,975	1	0,000
SERVICIOS	205,234	1	0,000
MONTO_CUOTA_OLD	7,273	1	0,007
LC_disp	16,616	1	0,000
SUELDOS_FIJOS	57,969	1	0,000
OTROS_ACTIVOS_FIJOS	8,126	1	0,004
DEUDA_OTROS_BANCOS	64,927	1	0,000
RASTRA	10,658	1	0,001
DEUDA_INDAP	304,967	1	0,000
Tenencia_ARRENDATARIO	281,308	1	0,000

AHORRO	141,648	1	0,000
PROVEEDORES	107,543	1	0,000
OTROS	140,9	1	0

10.9. ANEXO IX: Variables preliminares del modelo Margen SH

Variables	Chi-cuadrado	gl	Sig.
(Intersección)	26602,99	1	0
Sub_segmento_Cod_N	169,604	3	0
sexo_N	18,266	1	0
Region_N	62,213	4	0
Comuna_N	187,419	7	0
Cod_giro_N	15,165	1	0
Formalidad_Cod_N	32,276	2	0
Nro_clientes_Cod_N	12,063	1	0,001
Asistencia_Tecnica_Cod_N	22,316	1	0
ComportamientoBEME_Cod_N	5,02	1	0,025
nuevo_antiguo_Cod_N	3,191	1	0,074
Codofi_N	1487,221	10	0
LCR_Cod_N	6,685	1	0,01
Cheq_Cod_N	12,648	1	0
INSUMOS	232,336	1	0
MONTO_CUOTA	25,879	1	0
OTROS_ACTIVOS_CIRCULANTES	110,919	1	0
PRODUCTOS_ALMACENADOS	62,822	1	0
ARRIENDO_TERRENO	37,703	1	0
Gastos_ME	14,832	1	0

10.10. ANEXO X: Distintos modelos de Venta CH

Modelo	Variables introducidas
1	vta_ant
2	Formalidad_Cod_N_Dummy_1
3	MONTO_CUOTA
4	OTROS_ACTIVOS_CIRCULANTES
5	PRODUCTOS_ALMACENADOS
6	MOVILIZACION
7	ARRIENDOS
8	DEUDA_INDAP
9	CAJA
10	Ptje_Amb_SMRT01_N_Dummy_6
11	Ptje_Amb_SMRT01_N_Dummy_7
12	TOTAL_ACTIVOS
13	Codofi_N_Dummy_11
14	Tenencia_ARRENDATARIO
15	Codofi_N_Dummy_12
16	SERVICIOS
17	DEUDA_BECH
18	TOTAL_PASIVOS_LARGO_PLAZO
19	AHORRO
20	Codofi_N_Dummy_1
21	OTROS
22	PROVEEDORES
23	Ptje_Amb_SMRT01_N_Dummy_5
24	DEUDA_OTROS_BANCOS
25	Codofi_N_Dummy_14
26	Codofi_N_Dummy_9
27	Sub_segmento_Cod_N_Dummy_1
28	OTRAS_MAQUINARIAS
29	Ptje_SICA_N_Dummy_6
30	Codofi_N_Dummy_7
31	Ptje_Amb_SMRT01_N_Dummy_4
32	Ptje_SICA_N_Dummy_3
33	SUELDOS_FIJOS
34	Codofi_N_Dummy_8
35	CANTIDAD_IT_CLIENTE_TODAS
36	DEUDAS_FAM_NEG_PROM
37	Propiedad_Vivienda_Cod_N_Dummy_1
38	GASTOS_FAMILIARES
39	BAF_Cod_N_Dummy_1
40	Codofi_N_Dummy_6
41	LCR_Cod_N_Dummy_1
42	Modulo_Cod_N_Dummy_6
43	Codofi_N_Dummy_10
44	Codofi_N_Dummy_13
45	sexo_N_Dummy_2

46	Region_N_Dummy_2
47	Modulo_Cod_N_Dummy_4
48	Sub_segmento_Cod_N_Dummy_2
49	max_dmora_6M_N_Dummy_4
50	Ptje_SICA_N_Dummy_1
51	Ptje_SICA_N_Dummy_2
52	Tipo_registro_Cod_N_Dummy_2
53	Ptje_SICA_N_Dummy_4
54	max_dmora_6M_N_Dummy_2
55	Formalidad_Cod_N_Dummy_3
56	Codofi_N_Dummy_2
57	Modulo_Cod_N_Dummy_2
58	Propiedad_Vivienda_Cod_N_Dummy_2
59	Evolucion_Ahorro_Cod_N_Dummy_1
60	Modulo_Cod_N_Dummy_1
61	prov_ult_3a_Cod_N_Dummy_1
62	Region_N_Dummy_7
63	Ptje_Amb_SMRT01_N_Dummy_2
64	Ptje_Amb_SMRT01_N_Dummy_3
65	Ptje_Amb_SMRT01_N_Dummy_1
66	Cheq_Cod_N_Dummy_1
67	LC_disp
68	max_dmora_6M_N_Dummy_1
69	Modulo_Cod_N_Dummy_3
70	RASTRA
71	OTROS_ACTIVOS_FIJOS
72	nuevo_antiguo_Cod_N_Dummy_1
73	MONTO_CUOTA_OLD
74	Region_N_Dummy_5
75	Comuna_N_Dummy_5
76	Mora_6M_N_Dummy_1
77	Ptje_SICA_N_Dummy_5
78	Forma_de_Pago_Cod_N_Dummy_3
79	Comuna_N_Dummy_9
80	Region_N_Dummy_3
81	Region_N_Dummy_1
82	Evolucion_Ahorro_Cod_N_Dummy_2
83	Modulo_Cod_N_Dummy_7
84	Comuna_N_Dummy_2
85	Gestion_Cod_N_Dummy_1

10.11. ANEXO XI: Distintos modelos de Venta SH

Modelo	Variables introducidas
1	OPE_MONTO_ORIGEN_PES
2	TOTAL_ACTIVOS_CIRCULANTES
3	ARRIENDO_TERRENO
4	Formalidad_Cod_N_Dummy_1
5	INSUMOS
6	MOVILIZACION
7	OTROS_ACTIVOS_CIRCULANTES
8	SERVICIOS
9	PRODUCTOS_ALMACENADOS
10	Gastos_ME
11	SUELDOS_FIJOS
12	DEUDA_INDAP
13	Codofi_N_Dummy_6
14	OTROS
15	TOTAL_PASIVOS
16	LCR_Cod_N_Dummy_1
17	Codofi_N_Dummy_9
18	PATRIMONIO
19	Ptje_Amb_SMRT01_N_Dummy_1
20	Ptje_Amb_SMRT01_N_Dummy_2
21	OTRAS_MAQUINARIAS
22	LC_disp
23	Ptje_Amb_SMRT01_N_Dummy_3
24	Modulo_Cod_N_Dummy_1
25	Ptje_Amb_SMRT01_N_Dummy_4
26	Comuna_N_Dummy_2
27	CAJA
28	Cod_giro_N_Dummy_2
29	Forma_de_Pago_Cod_N_Dummy_1
30	Tipo_registro_Cod_N_Dummy_2
31	Ptje_SICA_N_Dummy_1
32	Cheq_Cod_N_Dummy_1
33	DEUDA_BECH
34	Codofi_N_Dummy_8
35	Modulo_Cod_N_Dummy_3
36	Ptje_SICA_N_Dummy_3
37	Codofi_N_Dummy_10
38	Codofi_N_Dummy_7
39	Codofi_N_Dummy_1
40	BAF_Cod_N_Dummy_1
41	Comportamiento_Beme_IT_Cod_N_Dummy_3
42	Sub_segmento_Cod_N_Dummy_3
43	DEUDA_OTROS_BANCOS

44	sexo_N_Dummy_1
45	Comuna_N_Dummy_1
46	Region_N_Dummy_1
47	Sub_segmento_Cod_N_Dummy_2
48	Sub_segmento_Cod_N_Dummy_1
49	Region_N_Dummy_2
50	Comuna_N_Dummy_4
51	Region_N_Dummy_4
52	Propiedad_Vivienda_Cod_N_Dummy_2
53	Ptje_SICA_N_Dummy_2
54	Codofi_N_Dummy_5
55	max_dmora_6M_N_Dummy_1
56	Cod_giro_N_Dummy_1
57	prov_ult_3a_Cod_N_Dummy_1
58	Asistencia_Tecnica_Cod_N_Dummy_2
59	AHORRO
60	Formalidad_Cod_N_Dummy_4
61	Formalidad_Cod_N_Dummy_2

10.12. ANEXO XII: Distintos modelos de Costo Fijo CH

Modelo	Variables introducidas
1	cf_ant
2	Formalidad_Cod_N_Dummy_1
3	ARRIENDOS
4	TOTAL_ACTIVOS_CIRCULANTES
5	Gastos_ME
6	Propiedad_Vivienda_Cod_N_Dummy_1
7	Codofi_N_Dummy_8
8	Tenencia_ARRENDATARIO
9	Modulo_Cod_N_Dummy_11
10	Codofi_N_Dummy_11
11	MONTO_TOTAL_ANT
12	Codofi_N_Dummy_3
13	MONTO_CUOTA
14	Codofi_N_Dummy_10
15	Ptje_Amb_SMRT01_N_Dummy_10
16	Codofi_N_Dummy_16
17	Codofi_N_Dummy_5
18	OTRAS_MAQUINARIAS
19	Codofi_N_Dummy_14
20	Codofi_N_Dummy_15
21	INSUMOS
22	VEHICULO
23	Permanencia_Cod_N_Dummy_1
24	Nro_clientes_Cod_N_Dummy_1
25	MONTO_CUOTA_OLD
26	Tipo_registro_Cod_N_Dummy_3
27	Sub_segmento_Cod_N_Dummy_3
28	Comuna_N_Dummy_8
29	Ptje_SICA_N_Dummy_5
30	Codofi_N_Dummy_6
31	Sub_segmento_Cod_N_Dummy_2
32	Codofi_N_Dummy_12
33	Modulo_Cod_N_Dummy_9
34	Comportamiento_Beme_IT_Cod_N_Dummy_1
35	Ptje_Amb_SMRT01_N_Dummy_4
36	Ptje_Amb_SMRT01_N_Dummy_5
37	Ptje_Amb_SMRT01_N_Dummy_6
38	Ptje_Amb_SMRT01_N_Dummy_3
39	Ptje_Amb_SMRT01_N_Dummy_8
40	Region_N_Dummy_7
41	Codofi_N_Dummy_7
42	Codofi_N_Dummy_13
43	Codofi_N_Dummy_4
44	Codofi_N_Dummy_9
45	Region_N_Dummy_5

46	sexo_N_Dummy_1
47	sexo_N_Dummy_2
48	Region_N_Dummy_4
49	Comuna_N_Dummy_7
50	Region_N_Dummy_6
51	Comuna_N_Dummy_5
52	LCR_Cod_N_Dummy_1
53	vta_PROM
54	Ptje_Amb_SMRT01_N_Dummy_7
55	Modulo_Cod_N_Dummy_10
56	Comuna_N_Dummy_12
57	Ptje_Amb_SMRT01_N_Dummy_9
58	nuevo_antiguo_Cod_N_Dummy_1
59	PRODUCTOS_ALMACENADOS
60	Asistencia_Tecnica_Cod_N_Dummy_2
61	Comuna_N_Dummy_4
62	Nro_clientes_Cod_N_Dummy_2
63	Modulo_Cod_N_Dummy_7
64	Mora_6M_N_Dummy_1
65	OTROS_ACTIVOS_FIJOS
66	TOTAL_PASIVOS_LARGO_PLAZO
67	Ptje_SICA_N_Dummy_3
68	Comportamiento_Beme_IT_Cod_N_Dummy_3
69	Propiedad_Vivienda_Cod_N_Dummy_3
70	Propiedad_Vivienda_Cod_N_Dummy_2
71	Tipo_registro_Cod_N_Dummy_2
72	Modulo_Cod_N_Dummy_4
73	Ptje_SICA_N_Dummy_4

10.13. ANEXO XIII: Distintos modelos de Costo Fijo SH

Modelo	Variables introducidas
1	Formalidad_Cod_N_Dummy_1
2	OPE_MONTO_ORIGEN_PES
3	ARRIENDOS
4	Gastos_ME
5	TOTAL_ACTIVOS_CIRCULANTES
6	Modulo_Cod_N_Dummy_4
7	Propiedad_Vivienda_Cod_N_Dummy_1
8	VEHICULO
9	INSUMOS
10	OTRAS_MAQUINARIAS
11	Modulo_Cod_N_Dummy_1
12	Codofi_N_Dummy_2
13	Codofi_N_Dummy_5
14	Ptje_Amb_SMRT01_N_Dummy_1
15	Tipo_registro_Cod_N_Dummy_2
16	PRODUCTOS_ALMACENADOS
17	Codofi_N_Dummy_9
18	OTROS_ACTIVOS_CIRCULANTES
19	OTROS_ACTIVOS_FIJOS
20	Ptje_Amb_SMRT01_N_Dummy_2
21	Modulo_Cod_N_Dummy_3
22	sexo_N_Dummy_2
23	OTROS_INGRESOS_TOTAL
24	Comportamiento_Beme_IT_Cod_N_Dummy_1
25	Codofi_N_Dummy_19
26	Nro_clientes_Cod_N_Dummy_1
27	BAF_Cod_N_Dummy_1
28	LCR_Cod_N_Dummy_1
29	Codofi_N_Dummy_8
30	Codofi_N_Dummy_7
31	Modulo_Cod_N_Dummy_5
32	Manejo_Ahorro_Cod_N_Dummy_1
33	Asistencia_Tecnica_Cod_N_Dummy_2
34	Region_N_Dummy_2
35	Region_N_Dummy_1
36	sexo_N_Dummy_1
37	Codofi_N_Dummy_1
38	Ptje_Amb_SMRT01_N_Dummy_6
39	Codofi_N_Dummy_4
40	CAJA
41	DEUDA_BECH
42	Cheq_Cod_N_Dummy_1
43	Modulo_Cod_N_Dummy_2
44	Ptje_SICA_N_Dummy_3
45	Comuna_N_Dummy_6

10.14. ANEXO XIV: Distintos modelos de Margen CH

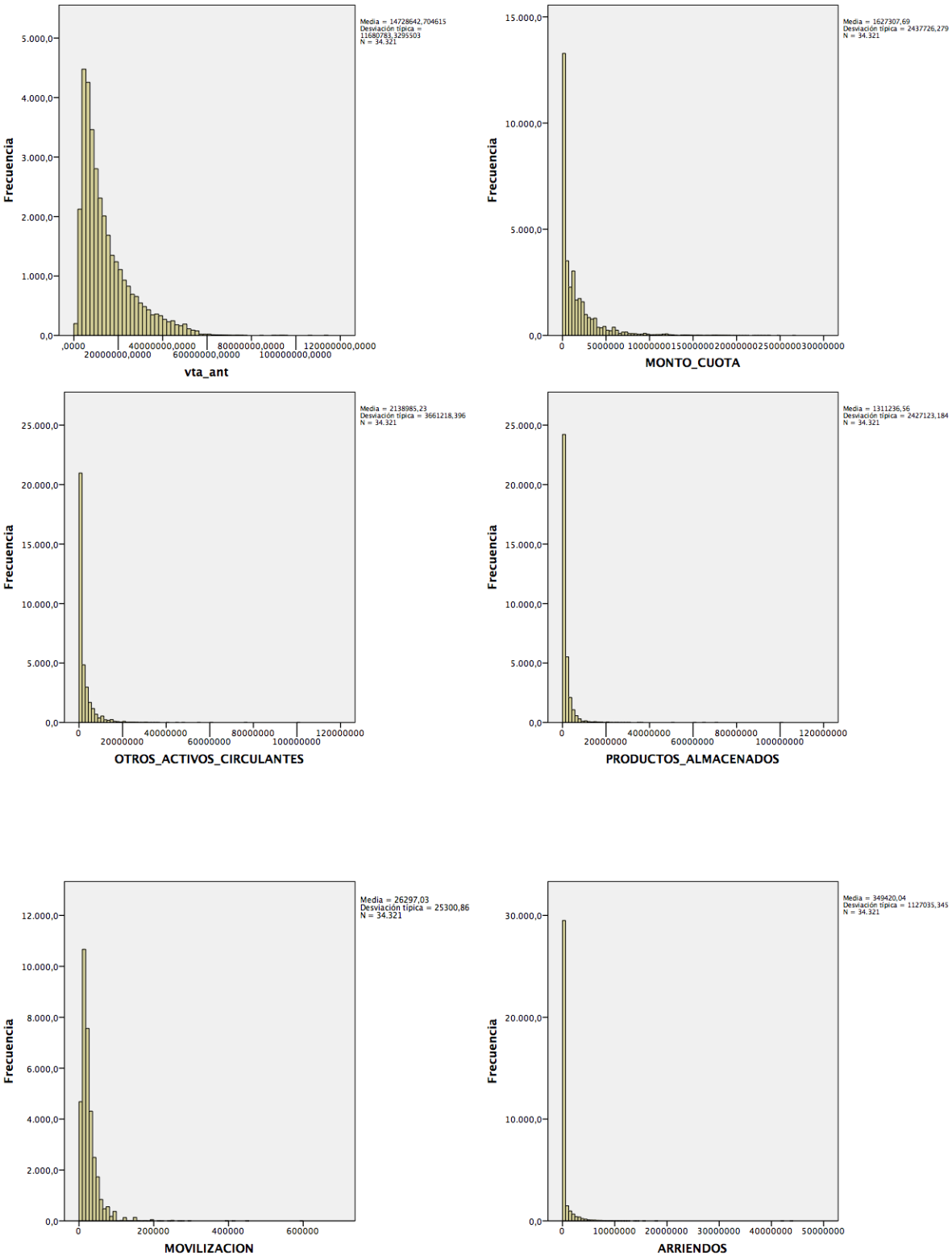
Modelo	Variables introducidas
1	mg_ant
2	vta_ant
3	Codofi_N_Dummy_11
4	Codofi_N_Dummy_2
5	Codofi_N_Dummy_3
6	Codofi_N_Dummy_1
7	Sub_segmento_Cod_N_Dummy_3
8	Modulo_Cod_N_Dummy_5
9	Codofi_N_Dummy_13
10	Codofi_N_Dummy_4
11	Codofi_N_Dummy_7
12	Codofi_N_Dummy_14
13	Codofi_N_Dummy_12
14	Codofi_N_Dummy_8
15	Codofi_N_Dummy_5
16	Formalidad_Cod_N_Dummy_2
17	Comuna_N_Dummy_2
18	Codofi_N_Dummy_9
19	Modulo_Cod_N_Dummy_2
20	Codofi_N_Dummy_10
21	Codofi_N_Dummy_6
22	Ptje_Amb_SMRT01_N_Dummy_4
23	Comuna_N_Dummy_14
24	LCR_Cod_N_Dummy_1
25	sexo_N_Dummy_2
26	Asistencia_Tecnica_Cod_N_Dummy_2
27	max_dmora_6M_N_Dummy_5
28	Ptje_SICA_N_Dummy_2
29	Cheq_Cod_N_Dummy_1
30	Comuna_N_Dummy_13
31	Region_N_Dummy_1
32	Sub_segmento_Cod_N_Dummy_2
33	max_dmora_6M_N_Dummy_1
34	Comuna_N_Dummy_7
35	Region_N_Dummy_4
36	Nro_clientes_Cod_N_Dummy_1
37	Comuna_N_Dummy_10
38	Comuna_N_Dummy_9
39	Comuna_N_Dummy_11
40	Comuna_N_Dummy_12
41	max_dmora_6M_N_Dummy_3
42	Cod_giro_N_Dummy_2
43	Propiedad_Vivienda_Cod_N_Dummy_1
44	Comuna_N_Dummy_4
45	Modulo_Cod_N_Dummy_7

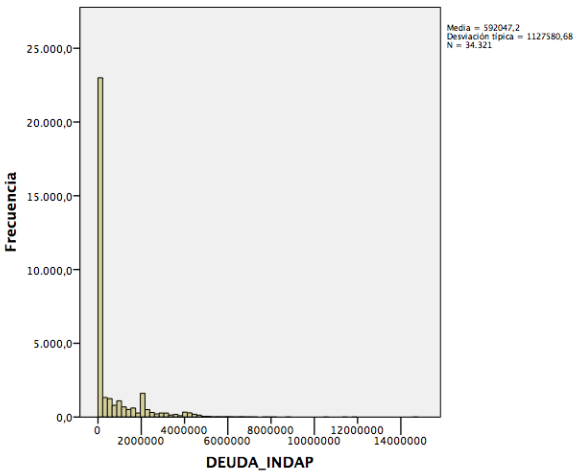
46	Asistencia_Tecnica_Cod_N_Dummy_1
47	Comportamiento_Beme_IT_Cod_N_Dummy_3
48	Modulo_Cod_N_Dummy_3
49	Evolucion_Ahorro_Cod_N_Dummy_1
50	Region_N_Dummy_2
51	Comuna_N_Dummy_3
52	Comuna_N_Dummy_8
53	Region_N_Dummy_3
54	Comuna_N_Dummy_1
55	Comuna_N_Dummy_5
56	Comuna_N_Dummy_6
57	prov_ult_3a_Cod_N_Dummy_1
58	Modulo_Cod_N_Dummy_6
59	Propiedad_Vivienda_Cod_N_Dummy_2
60	Cod_giro_N_Dummy_1
61	ComportamientoBEME_Cod_N_Dummy_2
62	max_dmora_6M_N_Dummy_2
63	BAF_Cod_N_Dummy_1

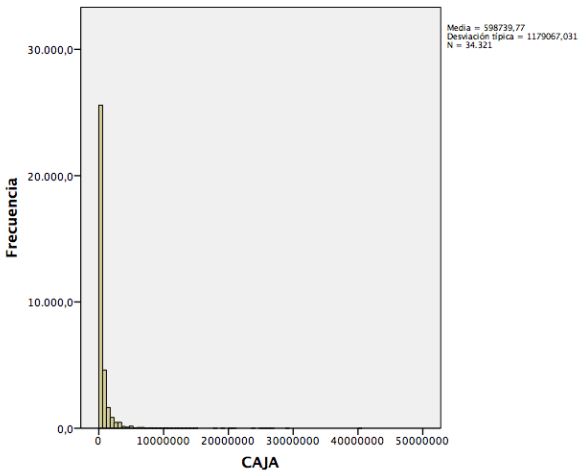
10.15. ANEXO XV: Distintos modelos de Margen SH

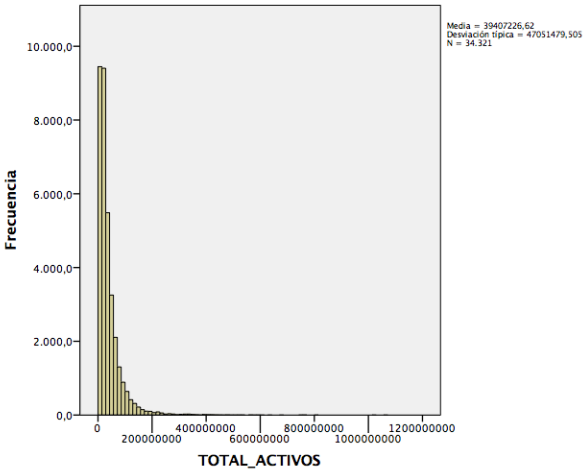
Modelo	Variables introducidas
1	Codofi_N_Dummy_1
2	Codofi_N_Dummy_2
3	INSUMOS
4	Codofi_N_Dummy_4
5	Codofi_N_Dummy_3
6	Codofi_N_Dummy_5
7	Codofi_N_Dummy_6
8	Sub_segmento_Cod_N_Dummy_3
9	Codofi_N_Dummy_7
10	OTROS_ACTIVOS_CIRCULANTES
11	Comuna_N_Dummy_6
12	PRODUCTOS_ALMACENADOS
13	Codofi_N_Dummy_10
14	Codofi_N_Dummy_9
15	Comuna_N_Dummy_5
16	Codofi_N_Dummy_8
17	Formalidad_Cod_N_Dummy_1
18	ARRIENDO_TERRENO
19	Cod_giro_N_Dummy_1
20	Comuna_N_Dummy_7
21	Comuna_N_Dummy_4
22	sexo_N_Dummy_1
23	MONTO_CUOTA
24	Gastos_ME
25	Comuna_N_Dummy_1
26	Region_N_Dummy_4
27	Region_N_Dummy_3
28	Cheq_Cod_N_Dummy_1
29	Asistencia_Tecnica_Cod_N_Dummy_1
30	Nro_clientes_Cod_N_Dummy_1
31	ComportamientoBEME_Cod_N_Dummy_1
32	LCR_Cod_N_Dummy_1
33	Sub_segmento_Cod_N_Dummy_2

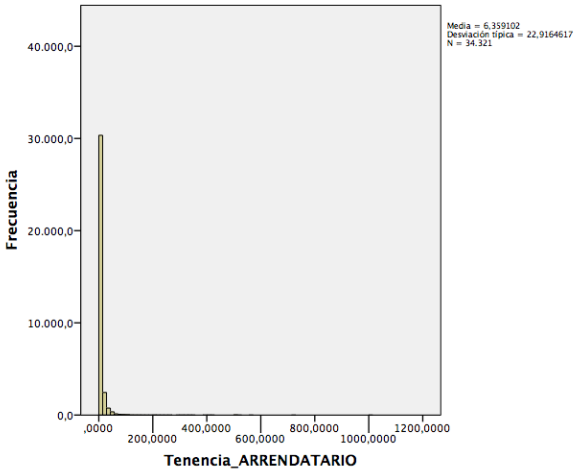
10.16. ANEXO XVI: Distribución de variables del modelo de Venta CH

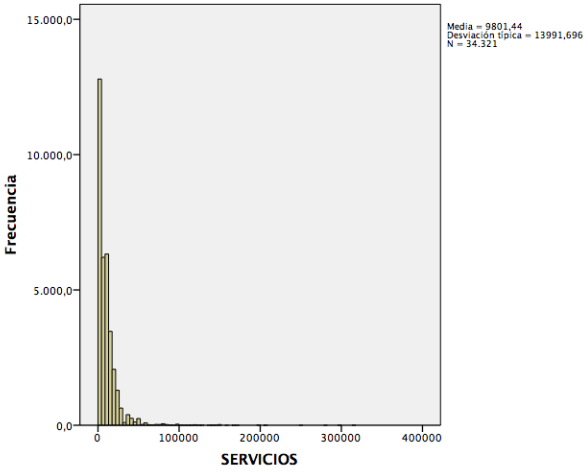



Media = 592047,2
Desviación típica = 1127580,68
N = 34.321
Frecuencia
25.000,0
20.000,0
15.000,0
10.000,0
5.000,0
0,0
0
2000000
4000000
6000000
8000000
10000000
12000000
14000000
DEUDA_INDAP

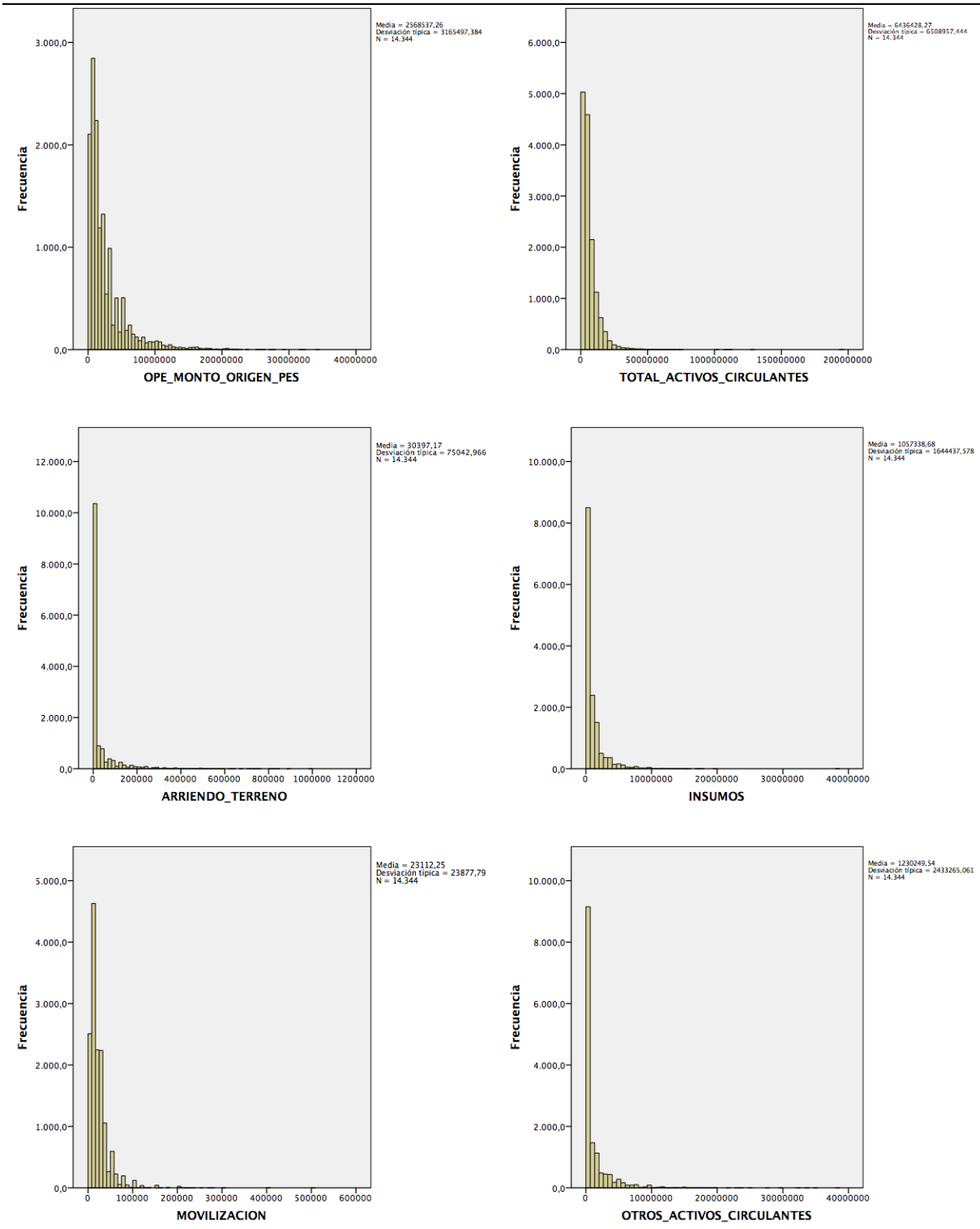

Media = 598739,77
Desviación típica = 1179067,031
N = 34.321
Frecuencia
30.000,0
20.000,0
10.000,0
0,0
0
10000000
20000000
30000000
40000000
50000000
CAJA

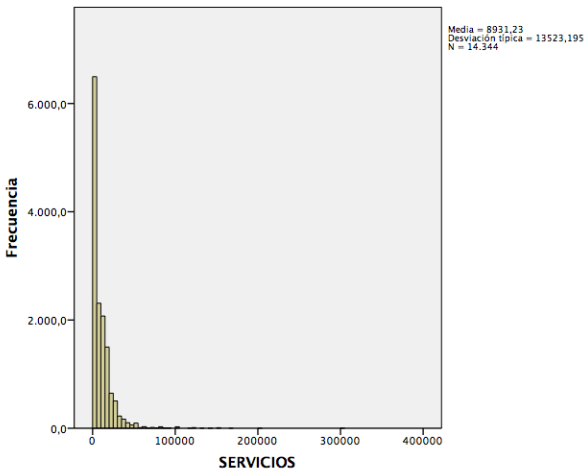

Media = 39407226,62
Desviación típica = 47051479,505
N = 34.321
Frecuencia
10.000,0
8.000,0
6.000,0
4.000,0
2.000,0
0,0
0
200000000
400000000
600000000
800000000
1000000000
1200000000
TOTAL_ACTIVOS

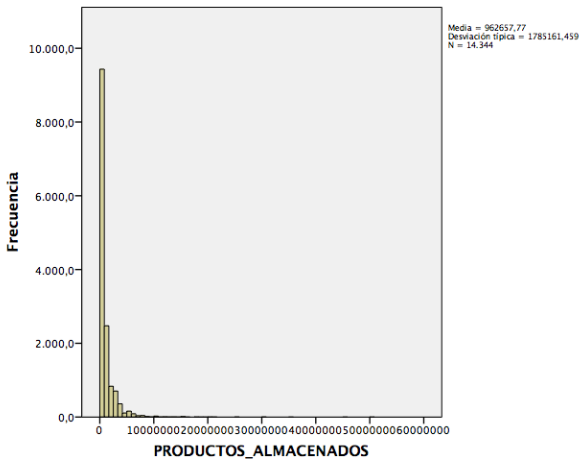

Media = 6,359102
Desviación típica = 22,9164617
N = 34.321
Frecuencia
40.000,0
30.000,0
20.000,0
10.000,0
0,0
,0000
200,0000
400,0000
600,0000
800,0000
1000,0000
1200,0000
Tenencia_ARRENDATARIO

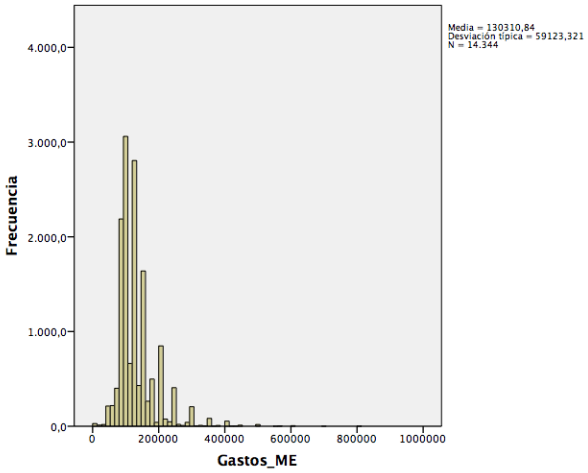

Media = 9801,44
Desviación típica = 13991,696
N = 34.321
Frecuencia
15.000,0
10.000,0
5.000,0
0,0
0
100000
200000
300000
400000
SERVICIOS

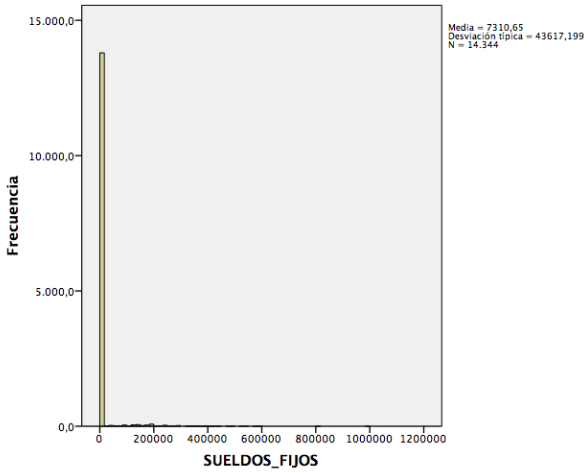
10.17. ANEXO XVII: Distribución de variables del modelo de Venta SH

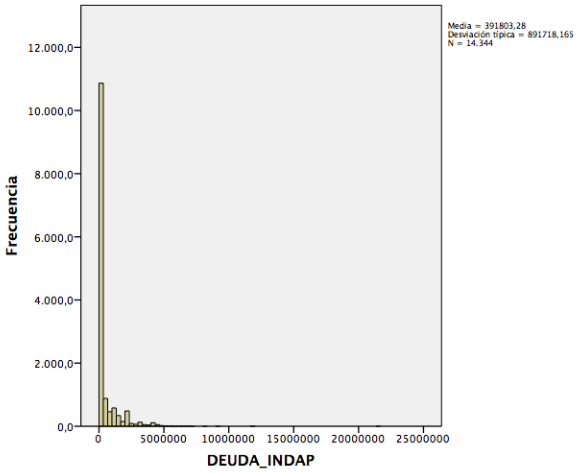


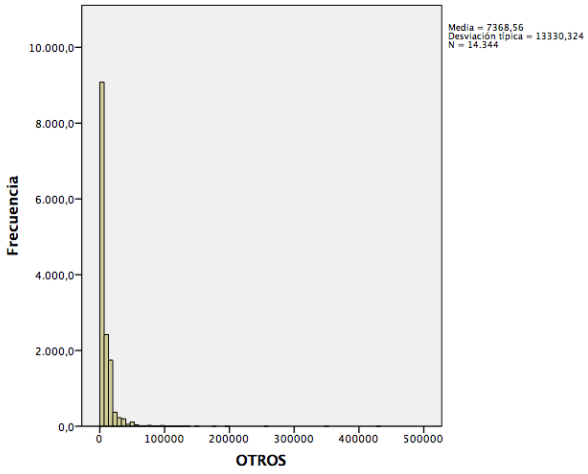

Media = 8931,23
Desviación típica = 13523,195
N = 14.344
Frecuencia
6.000,0
4.000,0
2.000,0
0,0
0
100000
200000
300000
400000
SERVICIOS

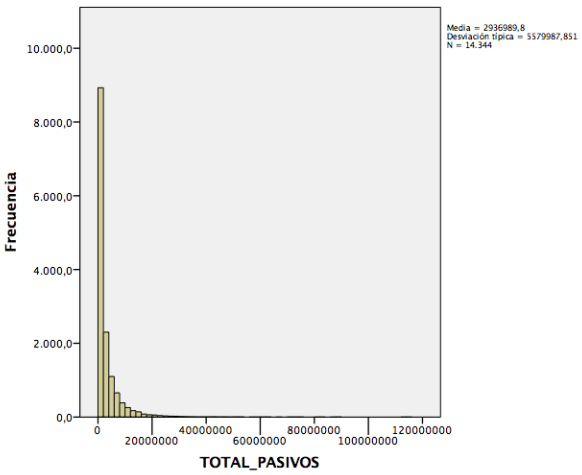

Media = 962657,77
Desviación típica = 1785161,459
N = 14.344
Frecuencia
10.000,0
8.000,0
6.000,0
4.000,0
2.000,0
0,0
0
10000000
20000000
30000000
40000000
50000000
60000000
PRODUCTOS_ALMACENADOS


Media = 130310,84
Desviación típica = 59123,321
N = 14.344
Frecuencia
4.000,0
3.000,0
2.000,0
1.000,0
0,0
0
200000
400000
600000
800000
1000000
Gastos_ME

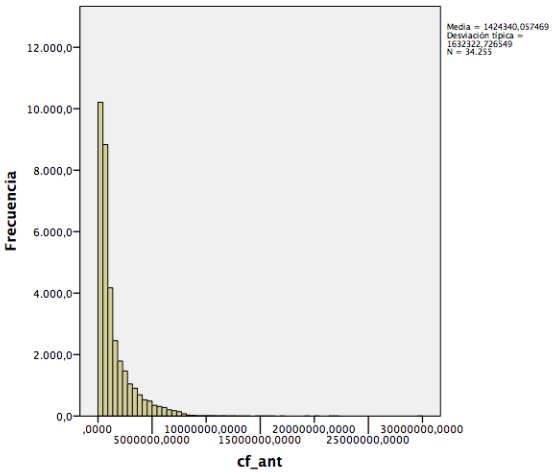

Media = 7310,65
Desviación típica = 43617,199
N = 14.344
Frecuencia
15.000,0
10.000,0
5.000,0
0,0
0
200000
400000
600000
800000
1000000
1200000
SUELDOS_FIJOS

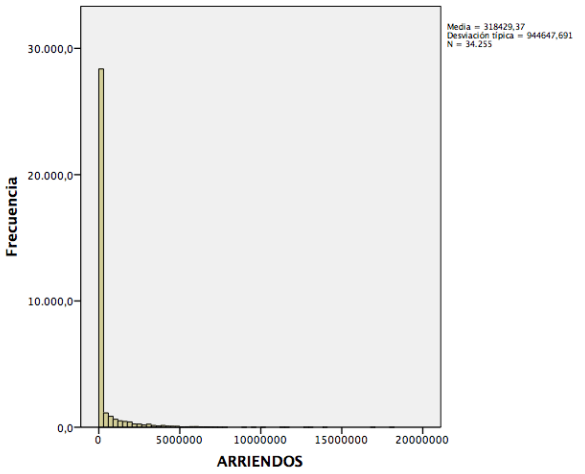

Media = 391803,28
Desviación típica = 891718,165
N = 14.344
Frecuencia
12.000,0
10.000,0
8.000,0
6.000,0
4.000,0
2.000,0
0,0
0
5000000
10000000
15000000
20000000
25000000
DEUDA_INDAP

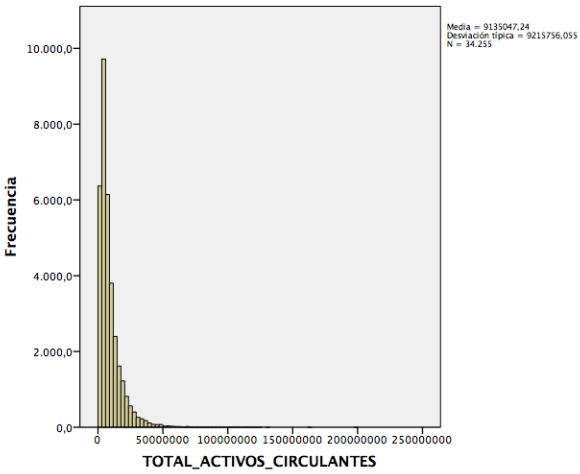

Media = 7368,56
Desviación típica = 13330,324
N = 14.344
Frecuencia
10.000,0
8.000,0
6.000,0
4.000,0
2.000,0
0,0
0
100000
200000
300000
400000
500000
OTROS

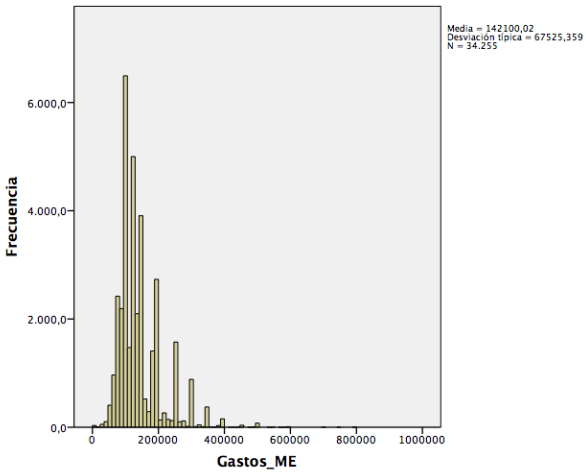

Media = 2936989,8
Desviación típica = 5579987,851
N = 14.344
Frecuencia
10.000,0
8.000,0
6.000,0
4.000,0
2.000,0
0,0
0
20000000
40000000
60000000
80000000
100000000
120000000
TOTAL_PASIVOS

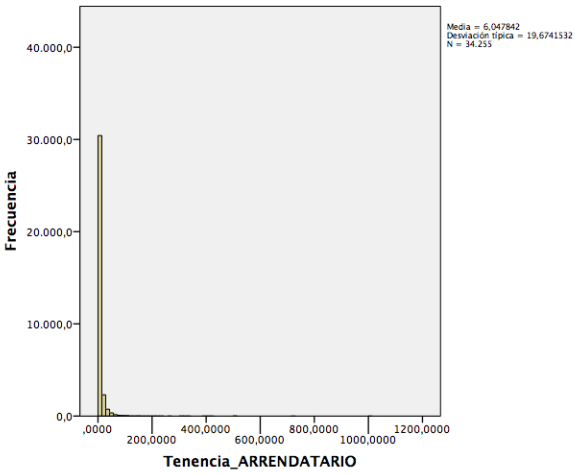
10.18. ANEXO XVIII: Distribución de variables del modelo de Costo Fijo CH

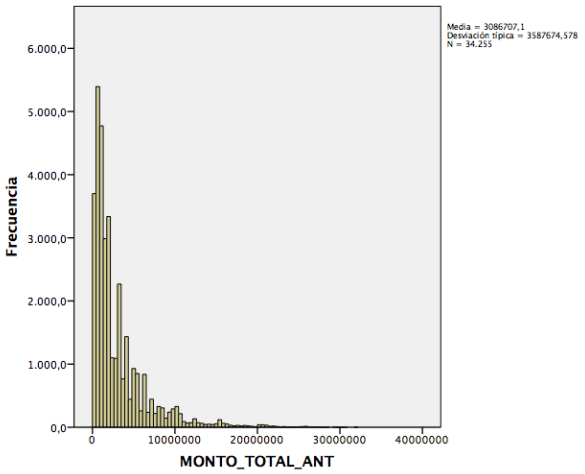




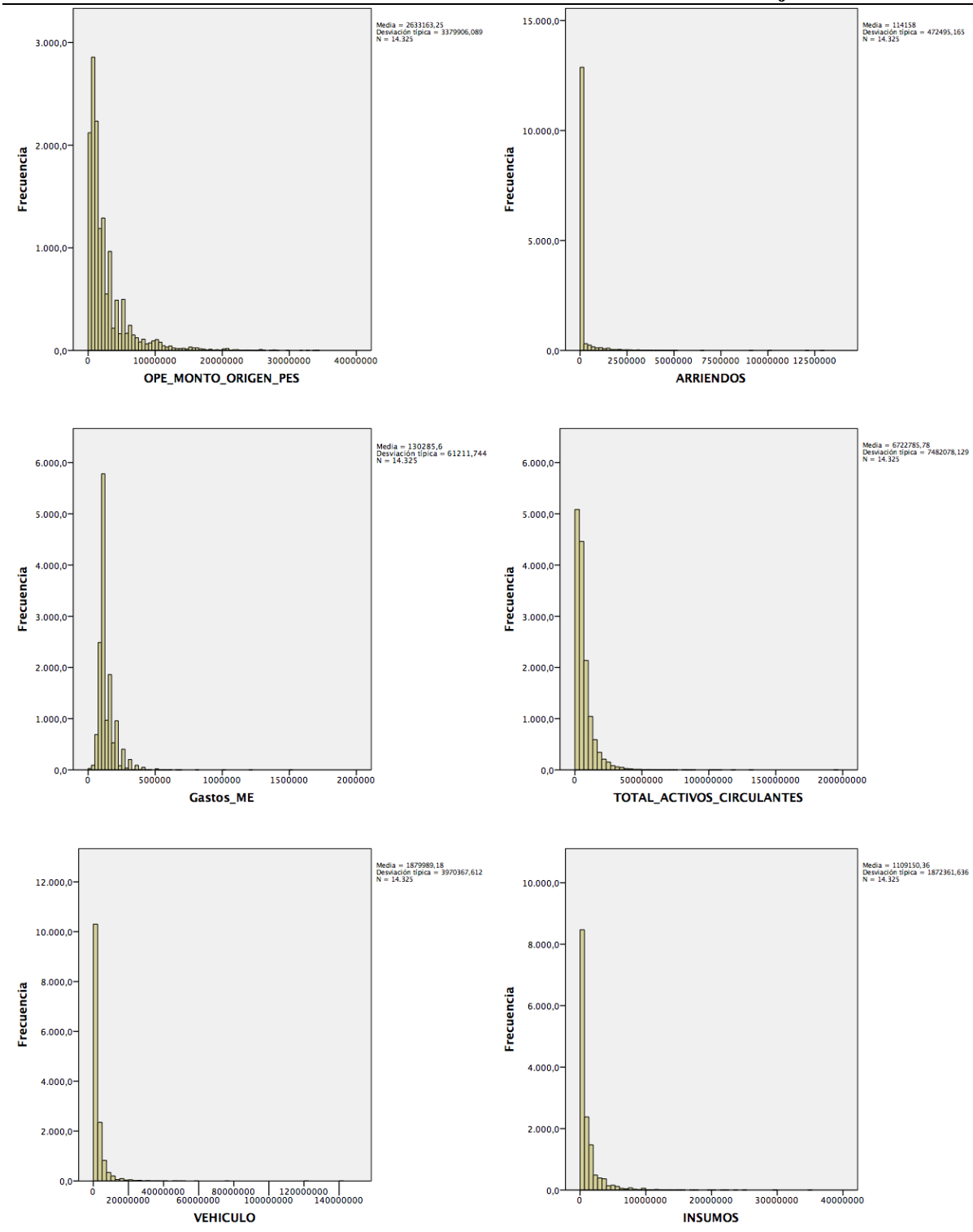


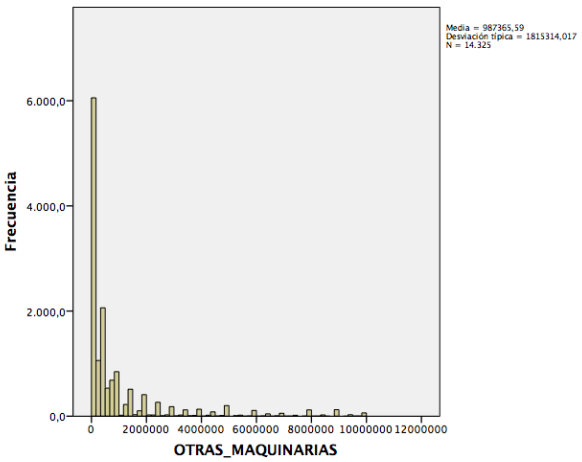






10.19. ANEXO XIX: Distribución de variables del modelo de Costo Fijo SH




Media = 987365,59
Desviación típica = 1815314,017
N = 14.325
Frecuencia
6.000,0
4.000,0
2.000,0
0,0
0
2000000
4000000
6000000
8000000
10000000
12000000
OTRAS_MAQUINARIAS