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

CREACIÓN DE UNA METODOLOGÍA PARA EL LEVANTAMIENTO DE UN PANEL DE CLIENTES, A PARTIR DE DATOS ALMACENADOS EN UN DATAWAREHOUSE

TESIS PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

CAROLINA VIVIANA NAVARRETE GUAJARDO

PROFESOR GUÍA:
MÁXIMO BOSCH PASSALACQUA

MIEMBROS DE LA COMISIÓN:
JULIA ROSS ZEBALLOS
MANUEL REYES JARA

SANTIAGO DE CHILE
MARZO 2010

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR: CAROLINA NAVARRETE G.
FECHA: 29/03/10
PROF. GUIA: SR. MAXIMO BOSCH P.

RESUMEN EJECUTIVO

Este trabajo tiene como objetivo la creación de un panel de clientes de una institución bancaria, que permitiera el cálculo de indicadores de su comportamiento a través del tiempo con el fin de observar el efecto de acciones comerciales de dicha entidad.
El foco investigativo fue diseñar una metodología adecuada para lograr una muestra representativa y de calidad de los 8.297.628 clientes del segmento personas, que pudiera ser mantenida y que asegurara entregar la información requerida en el tiempo.
Para su construcción se revisó la bibliografía de datos de panel, lo que permitió realizar un consolidado de los tópicos estadísticos que sustentan la presente tesis, destacando las ventajas del uso de datos de panel, sus diseños y la teoría base de toda inventigación que utiliza muestras.
Para asegurar que el panel fuera una solución efectiva al problema de negocio, se decidió elegir las variables en las que el panel pudiera entregar mediciones con un cierto nivel de confianza. A través del estudio de la información disponible y entrevistas con ejecutivos, se determinó que las variables de interés del panel serían: transaccionalidad de canales, tenencia de productos, volumen del negocio y uso de productos cruzados.
Las grandes diferencias del valor de dichas variables y la alta variabilidad en el número de individuos en los distintos cruces de la población objetivo hicieron concluir que el panel correspondería a una muestra aleatoria estratificada – no equilibrada .
Para la selección de las variables de estratificación se realizó un estudio de las distribuciones poblacionales en las variables clasificadoras de los clientes, dejando como candidatas aquellas con información completa (al menos 97%) y que presentaban una concentración mayor al 50% en al menos un nivel. Las variables seleccionadas fueron: segmento, renta, región y GRS (grupo de relación similar, creados en base a la tenencia de productos de los clientes). Luego de demostrar dichas variables tenían un efecto no nulo en las variables de interés del panel, a través de las pruebas de contraste de Games-Howell se definieron los niveles de corte, obteniendo un total de 312 estratos.
El tamaño de muestra, 49.770, fue calculado en base al objetivo de cumplir con la estimación de proporciones en varios niveles. La extracción de clientes no resultó constante para cada estrato. El desbalance de la muestra obtenida respecto a la población hizo necesaria la utilización de ponderadores para calibrar el panel, devolviéndole así la representatividad de la población total de clientes.
Los principales resultados de este trabajo son la estructuración de la información, el levantamiento del panel y la automatización de procesos: para extraer la muestra, calcular los ponderadores y obtener las métricas de comportamiento de clientes.
Para validar, se contrastó la mayor cantidad de información poblacional disponible con las estimaciones del panel, obteniendo errores inferiores al 3%. Dicho proceso permitió concluir que las decisiones de diseño tomadas favorecían el control del panel y aportaban flexibilidad en el manejo de información de los clientes de la institución, al permitir la estimación en 176 dimensiones con bajo error.
Finalmente, se concluyó que este trabajo posibilita realizar futuros estudios ad-hoc mediante la selección de clientes del panel para la aplicación de encuestas, así cómo análisis econométricos para estudios de comportamiento complejo. Sin embargo, como trabajo prioritario, se recomienda una exhaustiva revisión y perfeccionamiento de la calidad del *data warehouse* de la entidad.

Agradecimientos

A mis padres y hermana, por el cariño, apoyo, comprensión y ejemplo de constancia que me han entregado toda la vida.

A mis amigos, por todos los notables momentos que hemos compartido y que sin duda crearon los lazos que nos mantendrán unidos en el futuro.

A mis sobrinos, por toda su alegría y jovialidad que de manera inocente ayudan y renuevan las fuerzas para seguir adelante.

Índice general

Capítulo 1

Antecedentes del Problema

El gran aumento en la oferta de productos y la competitividad en la que se ven envueltos los negocios en la actualidad hacen que la relación compañía-cliente sea un tema fundamental para éstas. Cualquier error cometido en dicha relación puede ser fatal, en términos de pérdida de imagen o fuga de clientes, así cómo primordial para la captación de nuevos clientes y la fidelización de los actuales [1].

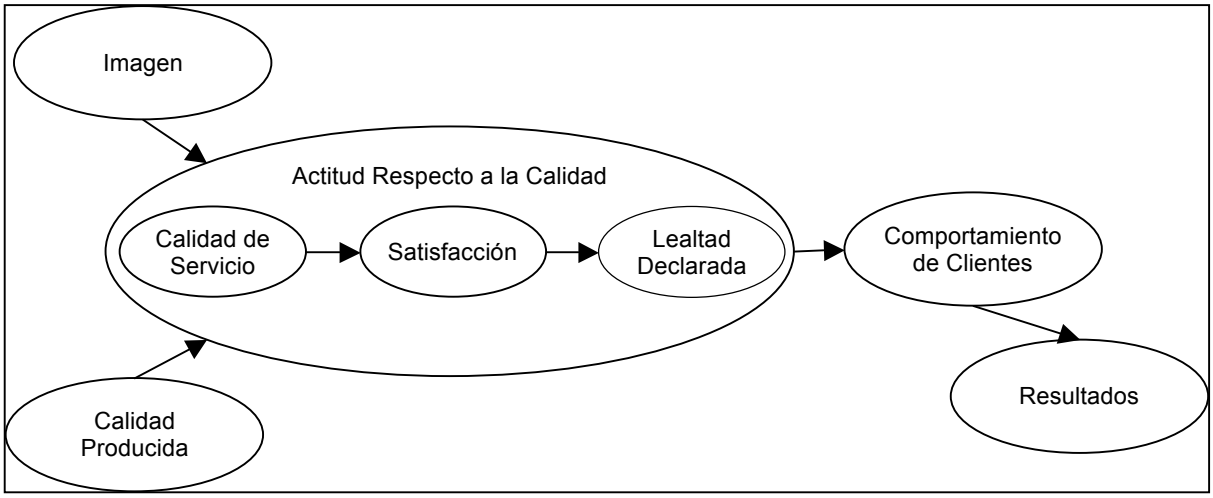
Es en este contexto que BancoEstado decidió emprender un proceso de Gestión de Calidad Transversa cuya misión es coordinar las distintas actividades de relación con los clientes intentando con ello la obtención de objetivos comunes, el establecimiento de programas y calendarios de actividades conjuntos y la creación de un proceso completo de medición que permita la retroalimentación de los resultados al sistema de gestión de la entidad.

Para esta tarea Banco Estado definió una organización que a través de un comité definió tres áreas principales de acción: Canales, Procesos y Métricas.

Como parte del desarrollo de esta última área, el banco ha estado llevando a cabo el proyecto “Métricas para la Calidad de Servicio BancoEstado” cuyo objetivo es el diseño de un conjunto de indicadores que permitan tener tanto una visión integral del desempeño de la calidad de servicio en la entidad como considerar el impacto de la calidad de servicio en la creación de valor para el banco y sus clientes.

La solución propuesta plantea un modelo respecto a cómo se generan los resultados del banco a partir de su operación que se describe en la siguiente figura:

Figura 1: Esquema del modelo conceptual de Calidad



Fuente: INGENOVA, 2008. Diseño de Indicadores para un Modelo de Calidad de Servicio Banca Retail [18].

Este modelo es de carácter conceptual y permite vincular la calidad de servicio producida e imagen, acciones operativas de la entidad, con el comportamiento de los clientes y los resultados de la institución.

La descripción de sus componentes es [18]:

- Calidad de Servicio Producida: Responde a las características operacionales con que el banco brinda efectivamente el servicio a sus clientes.
- Imagen: Corresponde a la percepción sobre el banco y sus servicios en la población en general y en sus clientes en particular.
- Actitud de los Clientes: Los clientes forman su actitud hacia el Banco en base a su experiencia y fuentes de información.
- Comportamiento de los Clientes: Corresponde a la interacción de los clientes a través de los diferentes procesos y canales. El comportamiento es, normalmente, registrado en las Bases del Banco.
- Resultados: El Banco obtiene un cierto nivel de desempeño en el cumplimiento de los distintos aspectos de su misión estratégica.

El eje articulador entre la operación y los resultados es el concepto llamado actitud pues este corresponde a la predisposición que tiene un sujeto (cliente) a actuar frente a un determinado objeto (servicio/oferta) que en relación con la calidad está conformada por tres componentes [18]:

- Calidad de Servicio Percibida: Es una percepción respecto al servicio recibido y su comparación con las alternativas relevantes a su disposición. Es un constructo, en el sentido que está compuesto por varias percepciones específicas.
- Satisfacción: Es el grado de cumplimiento de las expectativas que sienten los clientes respecto al servicio. La satisfacción es una evaluación global.
- Lealtad Declarada: La lealtad es la predisposición de los clientes a seguir operando con el Banco en sus distintos servicios. Es declarada pues es realizada por los mismos clientes a diferencia de la Lealtad Revelada que es la que el Banco puede observar (comportamiento).

1.1 Planteamiento del Problema y Justificación

La presente tesis será realizada en la empresa Ingenova, empresa consultora encargada de la asesoría en el desarrollo e implementación del sistema “Métricas para la Calidad de Servicio BancoEstado” expuesto anteriormente. Esta empresa requiere del diseño y la construcción de una herramienta que permita la alimentación continua de las métricas de comportamiento de clientes del modelo antes descrito.

Se entenderá por comportamiento todas las acciones que los clientes realizan en relación a los productos del banco y los canales que utilizan.

El cómputo de las métricas de comportamiento a partir de los registros administrativos del Banco podría desarrollarse de varias maneras, entre ellas

destacan: el análisis de la base de datos completa, la extracción de muestras aleatorias aisladas o sucesivas y la construcción de un panel de datos.

La primera alternativa queda descartada tanto por el alto costo a nivel de recursos tecnológicos (no disponibles en la entidad) como por restricciones de recursos humanos pues los requerimientos en el manejo a nivel usuario necesario se contraponen con la gran cantidad de datos que implicaría el análisis de las bases de datos completas.

Además se entiende que a pesar de que en la entidad actualmente existe un conjunto de mediciones del comportamiento de los consumidores estas no están orientadas a satisfacer los requerimientos de información de un sistema de gestión de calidad y por esto no cuentan con ni con la velocidad de cómputo ni la flexibilidad necesaria en este contexto.

La extracción de muestras aisladas es conocida como la obtención de cortes transversales, estos cortes ofrecen información valiosa para un momento determinado del tiempo pero como no tienen la dimensión tiempo limitan los análisis posibles. La realización de muestreos sucesivos mejora este punto, sin embargo, se sigue perdiendo poder de análisis pues es probable que las observaciones no provengan de los mismos individuos y por lo tanto no será posible la realización de estudios dinámicos [28].

La solución propuesta en la presente tesis corresponde a la tercera alternativa y es la creación de un panel de clientes el cual estará compuesto por una muestra representativa de los distintos tipos de clientes con el fin de poder calcular sobre éstos, de forma rápida y eficiente, los indicadores de comportamiento necesarios permitiendo además proyectar con validez estadística los resultados al total de clientes de la institución.

Un panel está definido como una serie de observaciones repetidas de muchas variables que caracterizan a las unidades muestrales a lo largo del tiempo [10].

Como se discutirá en el marco teórico, la dimensión temporal enriquece la estructura de los datos y es capaz de aportar información que no aparece en estudios con un único corte transversal.

De lo anterior se tiene que la justificación de la creación de un panel de clientes es la necesidad del cálculo continuo y de una gran cantidad de indicadores de comportamiento para los clientes del banco dónde la importancia de la solución propuesta radica en la flexibilidad que entrega para el manejo de información de los clientes.

Además, dado que la fuente de información que alimenta el panel son las bases de datos almacenadas en el datawarehouse del Banco es posible hacer el reconocimiento de los individuos y mantener la información de estos en el tiempo.

Con esto se evita el riego de valores perdidos y datos atípicos [28] haciendo ideal los análisis con perspectiva temporal necesarios para evaluar el comportamiento y fundamental en el contexto de desarrollo del proyecto en que se enmarca esta tesis.

1.2 Objetivos

1.2.1 Objetivo General

El objetivo del presente trabajo consiste en crear la metodología y realizar el levantamiento de un panel de clientes a partir de datos administrativos para permitir el cálculo eficiente y continuo de las métricas de comportamiento creadas en la primera etapa del proyecto “Métricas para la Calidad de Servicio BancoEstado”

1.2.2 Objetivos Específicos

Los objetivos específicos que envuelve la presenta memoria corresponden a:

(i) Examinar los patrones de cambio en el comportamiento de los clientes para definir los factores y/o variables que influyen en dichos cambios.
(ii) Levantar el panel en un programa computacional ad-hoc para el manejo a nivel usuario dentro de la entidad bancaria.
(iii) Calcular la primera camada de indicadores de comportamiento.
(iv) Realizar los análisis estadísticos correspondientes para establecer la validez de la solución propuesta.
(v) Definir posibles usos futuros de la solución creada, en función de aumentar la accesibilidad a la información del comportamiento de clientes y servir de apoyo a la gestión.

1.3 Metodología

Para obtener resultados exitosos, la metodología utilizada para resolver el problema de investigación propuesto en esta tesis debe considerar el uso ordenado en los datos disponibles a modo de generar una estructura que facilite la planificación y la dirección del proyecto.

Para el presente trabajo de tesis se definen los siguientes aspectos metodológicos a considerar:

(i) Estudio de literatura: La revisión bibliográfica es fundamental en la presente tesis y se concentra principalmente en:
  a) Estudio sobre la construcción de paneles en bases de datos
  b) Estudio de teoría subyacente a los datos de panel
  c) Teoría de muestreo y análisis estadísticos
(ii) Análisis del negocio: Consiste en la comprensión del problema a estudiar y la determinación de las posibles líneas de negocio a abordar.
(iii) Análisis de información disponible:
  a) Revisión de trabajos relacionados realizados anteriormente en la institución
  b) Caracterización de la base de datos disponible
  c) Limpieza, eliminación y/o transformación de datos
  d) Análisis de las principales componentes de la base de datos
(iv) Diseño de muestreo:
  a) Estudio de los clientes de la base: caracterización y definición de criterios para selección de la población objetivo.
  b) Selección de la técnica de muestreo a utilizar
  c) Determinación del tamaño de la muestra

(v) Levantamiento del Panel
   a) Definición del software a utilizar
   b) Confección de macros, procedimientos y consultas para procesamiento automático de los inputs computacionales requeridos.
   c) Creación de la programación necesaria para el cálculo de los indicadores de comportamiento

(vi) Validación
   a) Selección de muestras aleatorias y/o aplicación de métodos de re-muestreo
   b) Realización de pruebas estadísticas para validación de resultados obtenidos.
   c) Realización de calibraciones y ajustes necesarios (muestreo, pesos).

(vii) Mantención del Panel: Definición de los procedimientos de administración del panel en el tiempo.

1.4 Resultados Esperados

Con esta tesis se pretende entregar a la empresa consultora un proceso sistemático, metodológico y neutral que facilite la identificación, monitoreo y análisis del comportamiento de los clientes de BancoEstado a través del desarrollo de un Panel de Clientes.

El producto final consiste en la especificación comercial del panel de clientes para su posterior desarrollo por parte de la entidad bancaria y contempla los siguientes aspectos a cumplir:

- Primera muestra y criterios de selección
- Definiciones de fórmulas y análisis para calcular los indicadores
- Levantamiento del panel en una herramienta computacional de uso común
- Cálculo de la primera camada de indicadores
- Detalle de procedimientos para la administración y mantención del panel

1.5 Alcances

El fin del presente proyecto de tesis es realizar el estudio metodológico de la creación de un panel de clientes que tiene por finalidad el monitoreo del comportamiento de los clientes de la entidad bancaria.

Quedan fuera de los alcances del presente trabajo el desarrollo de las herramientas y procesos para el cálculo de los indicadores de los demás elementos del modelo de calidad que ampara este proyecto.

Además la investigación se restringe al estudio del segmento personas del banco y por lo tanto, quedan fuera de los análisis los datos provenientes de otros segmentos de la institución, por ejemplo: empresas.

Para la creación del panel se utilizarán técnicas estadísticas y de análisis de datos de panel. Además, los datos de los clientes serán recolectados sin entrevistas personales, esto se realizará uniendo los registros temporales de fuentes administrativas y bases de datos permitiendo, de esta manera, controlar y disminuir la falta de datos así como estudiar y determinar las políticas adecuadas de manejo y gestión del panel.

Cabe señalar que como es necesario asegurar la calidad de los resultados a lo largo del tiempo, además de seleccionar las técnicas adecuadas para la estructura de la información y la confección de la muestra será necesario definir el mantenimiento adecuado de la herramienta a construir a modo de minimizar las tendencias y desbalances que dañen las mediciones en el tiempo [22] [40].

El levantamiento computacional se realizará sobre una herramienta de uso computacional común como Excel, Access o SQL Server por ser estos los programas licenciados para los usuarios internos del panel de clientes quedando fuera del fuera del alcance la mejora de los tiempos de procesamiento de datos y la creación de la programación de un software o interfaz específica para el uso del panel.

Capítulo 2

Marco Teórico

A continuación se presentan los resultados de la investigación de literatura base del desarrollo de la presente tesis el cual es obtenido tanto de la revisión de los aspectos técnicos relacionados al problema a resolver como de la comprensión de estudios y publicaciones afines.

2.1 Datos de Panel [17] [26]

Los datos de las ciencias económicas pueden ser clasificados en tres formas: datos de corte transversal, series de tiempo y datos panel o longitudinales.

Los datos de corte transversal son aquellos que describen las actividades de los individuos, instituciones, hogares u otras unidades de muestreo en un momento único en el tiempo. Por el contrario, los datos de series de tiempo describen el movimiento de una única variable a lo largo del tiempo.

Los datos de panel son aquellos que combinan los dos tipos de datos anteriores, es decir, describen las actividades de un mismo grupo de unidades muestrales a lo largo del tiempo.

Por lo tanto, la característica más importante que distingue dichos datos es que las observaciones provienen de los mismos individuos en distintos instantes de tiempo [10].

2.2 Importancia de los Datos de Panel

El análisis de datos de panel es cada vez más utilizado entre los investigadores de las ciencias sociales y las ciencias del comportamiento [10]. Con frecuencia es utilizado en temas relacionados con innovación [30], gasto social [6], productividad [29], educación [33], pobreza [11], trabajo [12] [29], comunicaciones [2], etc. En general, el análisis de datos se utiliza para medir alguna característica particular de un grupo de personas a la cual se le hace seguimiento durante un período de tiempo determinado.

Entre los paneles más importantes destacan [10] [17]:

- NLS, Encuesta Nacional Longitudinal de Experiencia del Mercado Laboral. EEUU.
- PSID, Estudio de Panel de Dinámica de la Renta de Michigan, estudia aproximadamente 6000 familias (15000 individuos) desde 1968.

- SEP, Socioeconomic Panel (Holanda), seguimiento de 5000 familias en que se registran datos de variables socio-económicas como educación, ingreso, bienes durables, etc.

Formalmente se dice que un conjunto de datos es de panel cuando para una variable Y_{it} se tiene i=1,…, N observaciones de corte transversal (países, ciudades, empresas, artículos, individuos) y t= 1,…., T observaciones de series temporales.

Con las observaciones repetidas de muchas variables que caracterizan a las unidades muestrales a lo largo del tiempo, el análisis de panel permitirá de manera inteligente examinar dinámicas de cambio en el comportamiento de estas unidades.

El estudio continuo de la movilidad y del cambio de las características de los individuos implica la realización de un seguimiento de las unidades de estudio a lo largo del tiempo. Si bien es cierto que mediante los datos de corte transversal es posible obtener una "fotografía" de la población que permite su caracterización en un momento y luego al replicar dicha "fotografía" es factible determinar cambios brutos entre el antes y el después del tiempo que las separa, estos datos no permitirán profundizar en las causas y factores asociados a dichos cambios pues carecen de la dimensión temporal [11].

En su contraste los datos de panel se asemejan a tener una "película" a nivel de la unidad de observación y por ende hacen factible la revisión de los factores que provocan el cambio y generando un enfoque dinámico en el seguimiento de procesos [11].

Cabe señalar que cualquier estudio longitudinal no es una tarea sencilla pues se requiere asegurar tanto la cantidad como la calidad de datos a lo largo del tiempo. A su vez las dimensiones NxT[1] generarán una enorme cantidad de variables y número de casos.

De lo anterior, se desprende que un correcto estudio de datos de panel implicará una serie de consideraciones entre las que destacan: el esquema del diseño del panel, las políticas de mantenimiento, la gestión del panel y los métodos de análisis adecuados; los cuales serán diferentes a los realizados con datos de corte transversal y/o con series de tiempo.

2.3 Ventajas de los Datos de Panel

- Son fáciles de obtener de los registros administrativos de las empresas y permiten disminuir los costos (tiempo, dinero, tecnología) que implicaría trabajar con las bases de datos completas [28].
- Permiten realizar análisis de comportamiento más complejos y robustos pues posibilitan la consideración de efectos microeconómicos y dinámicos que no es posible obtener con data transversal y/o series de tiempo [17] [35].
- Mediante el uso de un panel se obtienen beneficios prácticos de la información analizada, dicha información permite tomar y retroalimentar

[1] N: Nº de elementos muestrales
T: Periodos de tiempo de registro de información

decisiones de gestión al permitir evaluar la evolución de las características frente a dichas decisiones [7] [21].

- El disponer de un mayor número de observaciones incrementa los grados de libertad y reduce la colinealidad entre las variables explicativas, lo que proporciona eficiencia estadística y se consiguen estimaciones con supuestos menos restrictivos [16] [26].
- Con N elementos muestrales y T períodos es posible estimar N modelos de series de tiempo y T modelos de corte transversal lo que generalmente crea dificultades de análisis, sin embargo las metodologías para datos de panel permiten agregar la información por ejemplo estimando: $y_{it}=x_{it}+u_{it}$ [31] [35].

2.4 Desventajas de los Datos de Panel

En términos generales, la principal desventaja asociadas a la técnica de uso de datos de panel es la llamada atrición o desgaste de la muestra [3] [17] [23] [34] [35].

Esto se refiere a la erosión gradual de la calidad de los datos a lo largo del tiempo debido a bajas en la muestra que traen como resultado la obtención de un menor número de observaciones disminuyendo la eficiencia de los estimadores y si, además, dicho desgaste ocurre sólo para un grupo seleccionado de individuos la muestra se desbalanceará respecto a la población objetivo y por lo tanto además de la pérdida de representatividad los resultados estarán sesgados.

2.5 Diseños de Paneles de Datos

En general existen cinco marcos conceptuales de diseño de panel [23]:

- Panel Fijo: Intenta recolectar los datos desde la muestra original en múltiples ocasiones y no son aceptadas modificaciones en dicha muestra.
- Panel Fijo más “nacimientos”: Es como un panel fijo que acepta la incorporación de nuevos miembros de la población “nacimientos” en la muestra. Diseño preferible al anterior pues intenta representar a la población transversal en cada ciclo y por lo tanto será posible realizar estimaciones transversales en paralelo con las estimaciones longitudinales.
- Panel Repetido: de cortes transversales en distintos intervalos de tiempo, que puede tener o no elementos en común. Generalmente son diseñados para la representación equivalente de la población de acuerdo a alguna característica específica. Por ejemplo, jóvenes de 17 años en el año 2000 y jóvenes de 17 años del año 2008.
- Panel Rotativo: Es aquel en que una proporción predeterminada de la muestra es reemplazada en cada ocasión. La muestra seleccionada en cada ciclo pretende representar la misma población o universo de estudio, permitiendo la combinación de los paneles para las estimaciones.
- Panel Dividido o Pool de Datos: Es una combinación de cortes transversales independiente en distintos intervalos de tiempo, dónde por lo tanto las observaciones no provienen de las mismas unidades muestrales a lo largo del tiempo.

De las definiciones anteriores es posible distinguir dos tipos de paneles: estáticos y dinámicos. Los paneles estáticos no cambian sus integrantes durante la vida del panel a diferencia de los dinámicos en los cuales se realiza rotación de sus miembros.

Una segunda distinción importante corresponde a los paneles balanceados y desbalanceados, mientras que en los primeros se mantiene fijo el número de observaciones periódica para cada uno de los individuos, en el segundo se aceptan valores perdidos.

2.6 Panel Rotativo

Dentro de los paneles dinámicos el panel rotativo es aquel en que se pretende representar la misma población o universo de estudio realizando la el reemplazo de una proporción de sus miembros en función de suavizar el desbalance del panel.

Los desequilibrios entre muestra de un panel y población pueden darse principalmente por tres condiciones generales [23]:

- "Muerte" de la unidad muestral, esto quiere decir que el individuo ya no se encuentra en la población de interés y por lo tanto no es elegible para continuar en la muestra.
- "Nacimientos" esto es que individuos nuevos son incorporados a la población de interés.
- Cambios naturales en las características de las unidades muestrales a lo largo del tiempo.

De esto se entiende, por ejemplo, que los paneles fijos no permitirán realizar los análisis de las características específicas de los nuevos integrantes de la población o de los que dejan de serlo presentando por tanto limitaciones en la operación del panel.

De lo anterior se tiene que utilizando una serie de unidades de muestreo demográficamente idénticas es posible realizar el reemplazo adecuado para formar un panel dinámico-rotativo que logre asegurar la representatividad de la muestra a través del tiempo.

Para cumplir dicho objetivo en la presente tesis será necesario definir un proceso de mantención del panel a construir pues en ningún caso los datos podrán ser tratados como una muestra común realizada para un estudio de corte transversal [32] [34].

2.7 Teoría de Muestreo [1] [24]

El muestreo es un proceso sistemático mediante el cual se selecciona un grupo de elementos (muestra) de un grupo completo (población).

La necesidad de inferir el comportamiento de ciertas características (ventas en una empresa, fallas en la fabricación de productos, etc.) que describen a la población hace necesario el uso de muestras cuando el análisis de la población completa no es posible debido a restricciones de costo, tiempo, operación y tecnología disponible.

Para que la muestra sea útil debe reflejar el comportamiento de toda la población objetivo y por lo tanto el muestreo debe ser realizado en función de cumplir con esta misión, cuando esto ocurre se dice que la muestra es representativa en caso contrario, la muestra es sesgada.

El proceso de muestreo se divide básicamente en los siguientes pasos [24]:

(i) Definir la Población Objetivo

La población objetivo es el conjunto conformado por todos los elementos que poseen la información buscada en la investigación.

(ii) Determinar el Marco Muestral

La construcción del marco muestral consiste en generar los listados de los elementos de la población objetivo.

El marco muestral debe cumplir tres requisitos:

- Individualizar a cada una de las unidades que podrían ser seleccionadas.
- Ser comprehensivo: es decir contener la máxima cobertura posible de unidades de la población objetivo.
- Permitir la ubicación, sin ambigüedad, de la unidad seleccionada.

(iii) Seleccionar la Técnica de Muestreo

La división más amplia de las técnicas de muestreo son: probabilístico y no probabilístico.

El muestreo no probabilístico se basa en el juicio del entrevistador o de responsable de la investigación pues es éste el que decide de manera arbitraria cuales son los elementos a incluir en la muestra. Este tipo de muestreo puede arrojar buenas aproximaciones de las características de la población, sin embargo, al no poder determinar la probabilidad con que un elemento es seleccionado no permite calcular la precisión, acotar el error cometido o proyectar los resultados de forma estadística la población.

Por el contrario en el muestreo probabilístico las unidades son seleccionadas al azar. Para llevar a cabo de manera correcta este tipo de muestreo, la confección del marco muestral es muy importante pues es éste el que permite realizar los cálculos de las probabilidades de selección de los elementos, conocer el error muestral y por lo tanto hace posible la realización de las proyecciones de las estimaciones a la población objetivo completa.

Debido a que las necesidades del trabajo expuesto en la presente tesis involucran la necesidad de realizar proyecciones a nivel poblacional, sólo se detallan a continuación los principales métodos de muestreo probabilístico.

a) Muestreo Aleatorio Simple (M.A.S)

En el M.A.S cada elemento de la población, y cada muestra posible, tienen la misma probabilidad de ser seleccionadas. Existen dos distinciones en la extracción de la muestra: con reposición y sin reposición, dónde la diferencia es el reingreso (o no) de los elementos al marco muestral una vez que son extraídos.

b) Muestreo Aleatorio Sistemático

La muestra es seleccionada a través de un paso de sucesión escogiendo un punto de inicio al azar. El paso de sucesión se determina dividiendo el tamaño de la población y el tamaño de la muestra (aproximando al entero más cercano).

En el muestreo sistemático los elementos pueden encontrarse ordenados (o no) de manera coincidente con las características de interés (por ejemplo, orden decreciente del monto de las deuda).

Si los elementos no se ordenan de manera relacionada con las características de interés, se obtendrán resultados similares al M.A.S.

c) Muestreo Estratificado

Es un proceso de dos etapas en el cual primero se divide la población en subgrupos o estratos mutuamente excluyentes y colectivamente exhaustivos lo que significa que cada elemento de la población debe ser asignado solamente a un estrato y no se debe omitir ningún elemento.

La segunda etapa corresponde a la selección de una muestra aleatoria independiente para cada estrato, generalmente a través de muestreo aleatorio simple o muestreo sistemático.

Se tienen que los principales criterios para la elección de las variables de estratificación son: homogeneidad, heterogeneidad, costo y relación.

Lo anterior es debido a que la estratificación debe intentar formar estratos cuyos elementos sean lo más homogéneos posibles y también debe lograr que los elementos de distintos estratos sean lo más heterogéneos entre sí manejando el trade-off en costos (complejidad, dificultad de aplicación, etc.) que implique construir una mayor cantidad de estratos.

De lo anterior se tiene que las variables de estratificación deben estar relacionadas a las características variables de interés en la investigación y este tipo de muestreo tendrá sentido cuando ellas presenten una alta variabilidad en la población muestreada pues con la estratificación se logra disminuir dicha variabilidad.

Existen dos tipos de muestreos estratificados dependiendo de la afijación utilizada:

- Afijación proporcional: reparte la muestra proporcionalmente a la población de cada estrato.
- Afijación no proporcional: reparte la muestra de manera no proporcional puede ser fija o según algún criterio de varianza en los estratos[2].

La afijación no proporcional crea sesgos en la muestra ya que las proporciones reales de la población no son mantenidas, por lo tanto para la realización de inferencias correctas se hace indispensable el cálculo y utilización de ponderadores que representan el peso real de cada elemento de la muestra en la población y permiten restituir la representatividad de la muestra [9].

[2] Mayor varianza mayor cantidad de elementos seleccionados en la muestra.

Estos ponderadores corresponden al inverso de la probabilidad de selección quedando determinados por:

$$W_i = \frac{\%\, poblacional\ estrato\ i}{\%\, muestral\ estrato\ i}$$

d) Muestreo por Etapas

En general el muestreo por etapas cumple con los pasos del muestreo estratificado: división de la población en grupos y luego se seleccionan los elementos en base a una técnica de muestreo probabilístico.

La diferencia entre ambas técnicas es que en el muestreo por etapas se elije sólo una muestra de subpoblaciones (grupos) y no todos los estratos. Además, el criterio para la formación de grupos o conglomerados es el opuesto al muestreo aleatorio estratificado pues los elementos dentro de un grupo deberán ser tan heterogéneos como la población.

A continuación se presenta un cuadro resumen de las fortalezas y debilidades de las técnicas de muestreo probabilísticas:

Tabla 1: Resumen de fortalezas y debilidades de las técnicas de muestreo probabilístico

Técnica	**Fortalezas**	**Debilidades**
Muestreo aleatorio simple (MAS)	De fácil comprensión, resultados proyectables	Marco de muestreo difícil de construir , costoso, menos precisión, no hay seguridad de representatividad
Muestreo sistemático	Puede aumentar la representatividad, más fácil de poner en práctica que el MAS, marco de muestreo no necesario	Puede disminuir la representatividad
Muestreo Estratificado	Incluye todas las subpoblaciones importantes y logra precisión de las estimaciones	Difícil de seleccionar las variables de estratificación pertinentes, no es fácil de estratificar en muchas variables, costoso
Muestreo por Etapas	Fácil de poner en práctica, eficaz en costos	Imprecisión, difícil de computar e interpretar los resultados

Fuente: Malhotra, Naresh K et al. “Investigación de mercados: un enfoque aplicado” pp -338.

(iv) Determinación del tamaño de Muestra

El tamaño de muestra se refiere al número de elementos que se utilizarán para realizar el estudio, para definir esto se deben tener en consideración razones cualitativas y cuantitativas como [19]:

- Número y naturaleza de las variables a estimar
- Limitaciones de recursos y tiempo
- El nivel de confianza para los resultados estimados
- Naturaleza de los análisis a realizar con la información recopilada

Para el cálculo estadístico del tamaño de muestra se distinguen principalmente dos casos de estimación: medias o proporciones.

a) Cálculo de tamaño de muestra para estimar medias [24]

Dado un nivel de confianza Z[3], ε es el error fijo y realizando la estimación por intervalos de confianza se tiene:

$$P(|\bar{x} - \mu| < \varepsilon) = (1 - \alpha)$$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Por el teorema del límite central se sabe que $\frac{\sigma}{\sqrt{n}}$ sigue una normal de media 0 y varianza 1 dónde:

$$P(\bar{x} - \frac{Z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{Z\sigma}{\sqrt{n}}) = (1 - \alpha)$$

De donde se obtiene que el error máximo de estimación estará dado por:

$$\varepsilon = \frac{Z\sigma}{\sqrt{n}}$$

Elevando al cuadrado ambos lados de esta ecuación y despejando n se obtiene que el tamaño de muestra es:

$$n = \left(\frac{Z\sigma}{\varepsilon}\right)^2$$

Para el caso de una población finita de tamaño N y muestreo sin reemplazo debe realizarse la corrección de finitud en la estimación del error máximo:

$$\varepsilon = \frac{Z\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

De dónde finalmente despejando n se obtiene:

$$n = \frac{Z^2\sigma^2 N}{\varepsilon^2(N-1) + Z^2\sigma^2}$$

b) Cálculo de tamaño de muestra para estimar proporciones [24]

La estimación de proporciones es aquella en que el interés principal es conocer la proporción de personas que cumple o no cumple cierta condición en la población.

En dicho caso la variable se distribuirá como una binomial cuya varianza está dada por el término $p(1-p)$ y reemplazando en la fórmula del error máximo de estimación antes descrita se tiene:

$$\varepsilon = \frac{Z\sqrt{p(1-p)}}{\sqrt{n}}$$

Por lo tanto, el tamaño de muestra queda definido por:

[3] Z=valor inverso de la distribución normal calculado para (1- α).

$$n = \left(\frac{Z\sqrt{pq}}{\varepsilon} \right)^2$$

Y realizando la corrección por finitud el tamaño muestral queda dado por:

$$n = \frac{Z^2 pq\ N}{\varepsilon^2 (N-1) + Z^2 pq}$$

Donde:

N = Total de la población
Z = Refleja el nivel de confianza de la estimación[4]
ε = error de estimación o precisión
p = proporción esperada
$q = 1 - p$

Cuando no se conocen los valores de p y q, se utiliza p=q=0.5 ya que dichos valores permiten obtener el mayor tamaño de muestra pues maximizan la varianza.

Además del interés por realizar estimaciones de proporciones para una variable en 2 niveles, puede ser parte de la investigación la necesidad de calcular proporciones simultáneas para variables con una mayor cantidad de niveles.

Del estudio de literatura realizado se tiene que en Medina [27] se realiza una completa discusión de los factores que llevan a determinar el tamaño óptimo de muestra destacando entre estas "el tipo de variables e indicadores a estimar y los dominios de estudio que se requiere realizar".

En función de esto Medina plantea que es necesario realizar una corrección al tamaño de muestra cuando lo que se desea es estimar la distribución de proporciones de una variable de más de dos niveles y que a pesar de que es poco considerada por los investigadores en la práctica debe ser considerada en a modo de incorporar toda la calidad estadística disponible en la investigación a realizar.

Medina expone que la mejor manera de realizar la corrección del tamaño muestral antes descrita fue presentado por Tortora [27 pp 7-12] quien propone que el cálculo del tamaño de muestra n[5] para estimar proporciones cuando es necesario realizar inferencias para k niveles o categorías mutuamente excluyentes y exhaustivas es posible a partir del procedimiento de cálculo de intervalos de confianza simultáneos propuesto por Goodman [14].

Sea:

n = el tamaño de la muestra
k = la cantidad de categorías a estimar proporción
n_i = la frecuencia de observación de la i-ésima categoría
π_i =la proporción de la población ubicada en la i-ésima categoría
α = nivel de confianza

[4] En el caso en que se escoja un 95% de confianza, Z vale 1.96.
[5] para cierto nivel de confianza y error requerido.

El autor plantea que una aproximación de los intervalos de confianza inferior y superior es respectivamente:

$$\pi_i^{\,i} = \pi_i - \left[B\,\pi_i\,(1-\pi_i)/\,n\right]^{\frac{1}{2}}$$

$$\pi_i^{\,s} = \pi_i + \left[B\,\pi_i\,(1-\pi_i)/\,n\right]^{\frac{1}{2}}$$

Donde para k>2 B es el percentil superior (α/k *100) de una distribución chi-cuadrada con 1 grado de libertad.

A lo cual Tortora [27] para la determinar el tamaño de muestra adecuado agrega que si la precisión requerida para cada una de las categorías a estimar b_i las ecuaciones quedan de la siguiente manera:

$$\pi_i - b_i = \pi_i - \left[B\,\pi_i\,(1-\pi_i)/\,n\right]^{\frac{1}{2}}$$

$$\pi_i + b_i = \pi_i - \left[B\,\pi_i\,(1-\pi_i)/\,n\right]^{\frac{1}{2}}$$

Despejando b_i en las ecuaciones anteriores se tiene:

$$b_i = \left[B\,\pi_i\,(1-\pi_i)/\,n\right]^{\frac{1}{2}}$$

De despejando n se tendrá que el tamaño de muestra necesario para estimar cada celda con precisión b_i es:

$$n = \max\left\{B\,\pi_i\,(1-\pi_i)/\,b_i^{\,2}\right\}$$

Y para el caso de población finita el tamaño de muestra queda dado por:

$$n = \max\left\{\frac{B\,\pi_i\,(1-\pi_i)}{b_i^{\,2}(N-1)+B\,\pi_i\,(1-\pi_i)}\right\}$$

2.8 Pruebas Estadísticas

Las justificaciones de las decisiones tomadas en el desarrollo de la presente tesis envuelven sin lugar a dudas la realización de diferentes test de análisis de estadísticos que las avalen, entre los principales se encuentran:

2.8.1 Coeficiente de Variación (CV) [13]

El coeficiente de variación corresponde al número de veces que la desviación estándar contiene a la media, esto es:

CV= desviación estándar/media aritmética

Por lo tanto a mayor valor de dicho coeficiente mayor será la dispersión de los datos y por lo tanto menor la representatividad de la media (CV>1 => media no representativa)

Es independiente de las unidades de medida y no debe utilizarse si la media tiene valor cercano a cero.

Para mediciones experimentales la exigencia al CV es aún mayor para asegurar la precisión de las estimaciones realizadas, por ejemplo para un intervalo de confianza de 95% la interpretación es la siguiente [20]:

Tabla 2: Interpretación CV

C.V. (%)	Precisión obtenida
Hasta 5	Muy buena
De 5 a 10	Buena
De 10 a 20	Aceptable
Más de 20	No confiable

Fuente: Elaboración Propia

2.8.2 Análisis de Igualdad de Medias [25]

Uno de los test más utilizados para la comparación de medias es test ANOVA de un factor el cual genera un análisis de varianza para una variable dependiente cuantitativa respecto a una variable independiente categórica de 2 o más niveles los cuales constituyen los grupos a los cuales se desea comparar la media.

Esta prueba exige el cumplimiento de determinados supuestos respecto a los parámetros de las distribuciones de las poblaciones a comparar, en efecto para su validez se requiere:

- Normalidad: Cada grupo formado por los niveles de la variable independiente deber ser un muestreo aleatorio que proveniente de una población normal.
- Homocedasticidad: Los grupos provienen de poblaciones con varianzas iguales.

Si no existe normalidad el test ANOVA sigue siendo robusto para muestras grandes y sólo es necesario que los datos sean simétricos. Si hay heterocedasticidad pueden obtenerse resultados erróneos por lo tanto es muy importante la verificación de este supuesto.

El test ANOVA contrasta la hipótesis nula de que las medias de los grupos poblacionales formados por los niveles de la variable independiente (factor) son iguales:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_i = ... = \mu_m$$

Donde μ_j es la media de los distintos m grupos.

El análisis se realiza a través del estadístico F-Fisher el cual corresponde a la división de la estimación de la varianza poblacional basada en dispersión existente entre las medias de cada grupo y la estimación de dicho valor en base a la variabilidad existente dentro de cada grupo:

$$F = \frac{n\hat{\sigma}^2_{\bar{Y}}}{\overline{S}^2_j}$$

El valor de F reflejará la similitud existente entre las medias poblacionales pues si ellas son iguales también lo serán las medias muestrales y por lo tanto ambas

estimaciones de la varianza poblacional serán parecidas y harán que el valor de F sea cercano a 1.

Si las medias son distintas, la estimación de la varianza poblacional basada en la variabilidad existente entre grupos será mayor y por lo tanto F tomará valores mayores que 1.

Si se cumplen los supuestos de normalidad y heterogeneidad, antes descritos, el estadístico creado se distribuye según el modelo de probabilidad F de Fisher el cual permitirá concluir a través de su respectivo nivel de significancia.

En caso que dicho valor sea inferior a 0,05 se rechaza la hipótesis nula y eso significa que los grupos no poseen medias iguales y por ende existe una relación entre la variable dependiente y el factor.

Para estudiar el efecto de dos o más variables independientes (categóricas) sobre una variable dependiente (cuantitativa) se utiliza el test anova factorial el cual proporciona un análisis de regresión del modelo evaluado y testea de manera análoga una hipótesis nula por cada variable independiente y por cada posible interacción entre ellas.

2.8.3 Test de Igualdad de Varianzas [25]

El test de Levene permite contrastar la hipótesis nula de igualdad de varianzas el cual es requisito para asegurar la validez del test ANOVA descrito anteriormente.

El test de Levene es rechazado si el nivel de significación es < 0,05.

2.8.4 Pruebas Robustas para la Igualdad de medias [25]

Los métodos de Welch y Brown-Forsythe son variaciones del estadístico F para contrastar la igualdad de medias en casos en que no es posible verificar el supuesto de homocedasticidad.

Al igual que en el test ANOVA la a hipótesis nula es la igualdad de medias en los distintos niveles de la variable independiente y por lo tanto si su significación es menor a 0,05 ésta debe ser rechazada.

2.8.5 Test no parámetricos de Igualdad de Medias [25]

El test no paramétrico de igualdad de medias es equivalente al análisis ANOVA descrito anteriormente pero permite trabajar cuando no se cumplen los requisitos de normalidad y homogeneidad antes descritos.

Esta prueba corresponde al test Kruskal y Wallis la cual considera M muestras aleatorias e independientes de tamaños n_1, n_2,…,n_M extraídas de la misma población o de M poblaciones idénticas.

El método consiste en ordenar los datos de mayor a menor y asignar rangos de 1 a n a dicho conjunto (al existir datos repetidos se asigna el rango promedio a cada uno de estos) para crear el estadístico H.

$$H = \frac{12}{n \cdot (n+1)} \sum_{j}^{M} \frac{R_j^2}{n_j} - 3 \cdot (n+1)$$

Dónde:

- n el conjunto de todas las observaciones de las M muestras, $n = \sum_{j}^{M} n_j$

- R_j la suma de los rangos asignados a las n_j observaciones,

- R_{ij} el rango asignado a las observaciones i de la muestra j

H se distribuye según una chi-cuadrado con M-1 grados de libertad y al contrastar la hipótesis nula de que los M promedios poblacionales son iguales con si el nivel de significancia es menor a 0.05 será posible concluir que dichos promedios no son iguales y por lo tanto existe una relación entre la variable dependiente y la variable categórica independiente.

2.8.6 Pruebas Kolmogorov-Smirnov [25]

Este test es ampliamente utilizado para probar si una variable cuantitativa se ajusta o no a una determinada función teórica de distribución (normal, exponencial, etc.) funciona en base a la comparación de las funciones de distribución acumuladas contrastando los datos teóricos con los empíricos formando esta última mediante la ordenación ascendente de sus valores y calculando posteriormente la función de distribución de la forma:

F(Xi)=i/n donde n es el total de observaciones e i corresponde al rango de ordenación obtenido para cada observación.

Para la cual posteriormente se calcula el estadístico K-S a partir de la mayor diferencia entre ambas distribuciones y la cantidad de observaciones de la forma:

$$Z_{k-s} = \max|D|\sqrt{n}$$

El cual permite finalmente concluir ya que se distribuye N(0,1).

Este test puede ser extendido para comparar dos muestras independientes permitiendo contrastar si 2 muestras proceden de la misma población.

En este caso el valor de la función de distribución empírica es calculado para ambas muestras o grupos y posteriormente se contrastan sus diferencias de igual forma que el caso anterior rechazando cuando la significancia asintótica es menor a 0.05.

Capítulo 3

Análisis de la Información Disponible

Todo trabajo con datos, debe contar con un proceso de análisis en profundidad de la información disponible con el objetivo de lograr el conocimiento en detalle de dicha información.

Los objetivos del presente capítulo son comprender el vínculo del problema a resolver con la construcción del panel y el negocio; a su vez se espera lograr una adecuada familiarización de los datos disponibles mediante su comprensión, limpieza y transformación.

3.1 Análisis del negocio

La herramienta a construir en la presente tesis debe realizarse a partir de los datos almacenados en el datawarehouse de BancoEstado y su principal foco corresponde a disponibilizar la información necesaria para la posterior realización de estudios del comportamiento de los clientes de la institución a lo largo del tiempo.

El panel de clientes se basa en la extracción y mantención de una muestra representativa y controlada de los clientes del Banco, a modo de poder proyectar con validez estadística los resultados a la población total.

Para lograr la comprensión de los outputs del panel requeridos por la entidad fue necesario realizar entrevistas en profundidad a los ejecutivos de la banca comercial, además de una encuesta (ver anexo A) y varias presentaciones del proyecto a modo de recibir los feedbacks para definirlos.

De los principales resultados obtenidos en la comprensión del negocio destaca la necesidad lograr que el panel de clientes sea un aporte en:

- Facilitar el acceso y uso de la información contenida en los registros administrativos del Banco.
- Generar una fuente de información para estudios ad-hoc y para estudios periódicos.
- Permitir el cálculo rápido, eficiente y continuo de variables de comportamiento de clientes relacionadas a la actividad transaccional de los clientes, sus saldos en inventario, sus deudas y sus productos.
- Disminuir los costos en tiempo, dinero y personal especializado que implica trabajar con las bases de datos completas.

Además de se detectó un gran interés por generar nuevas dimensiones de observación del comportamiento, además de las variables demográficas clásicas, a modo de ser capaces de responder cuestionamientos como:

¿Qué características tienen y cómo se comportan los clientes qué?

- Tienen un uso sobre el promedio de Internet este trimestre
- Realizaron más compras que el promedio con tarjeta de crédito el mes pasado
- Acuden a la caja vecina y poseen cuenta rut
- Aumentaron la tenencia de productos este último año
- Aumentaron (disminuyeron) sus saldos en cuenta corriente el mes anterior

De esto se tiene que el panel deberá facilitar el estudio del comportamiento desde el punto de vista del consumidor y ser una herramienta global de aporte a la gestión aportando en las siguientes líneas de negocio:

- Transaccionalidad de Canales
- Tenencia de Productos
- Volumen del Negocio
- Uso de Productos Cruzados

3.2 Caracterización, limpieza y transformación de datos disponibles

El manejo de datos de la entidad no se realiza de forma centralizada y depende del área de negocio la información que se maneje.

El panel deberá ser desarrollado en base a la información disponible en el Datamart del área de Marketing por ser esta área de negocio la que se encuentra a cargo del proyecto.

La base de datos disponible para el desarrollo de esta tesis es una base de datos relacional de la cual se solicitó la siguiente información:

Tabla 3: Resumen de principales tablas disponibles

TABLA	DESCRIPCIÓN
CANALES	Tabla que almacena los canales dónde transaccionan los clientes.
CLIENTES1	Tabla que almacena información general de clientes del banco, tanto personas como organizaciones.
CLIENTES2	Almacena información de personas.
SEGMENTOS1	Tabla que registra los segmentos estratégicos de la institución.
SEGMENTOS2	Guarda la información de subsegmentos estratégicos con su segmento estratégico asociado.
CONTRATO1	Detalle de los contratos de productos menos ahorro.
CONTRATO2	Detalle de los contratos de los productos ahorro.
SBIF	Información de deudas proveniente de la Superintendencia (SBIF).
PRODUCTOS	Tabla de productos.
TRANSACCIONES	Transacciones realizadas por los clientes a través de los distintos canales (Mesón, Cajas, Internet, etc.), identificando la Unidad Organizacional de la transacción y del cliente.
OFICINA ASIGNADA	Almacena el código de la oficina asignada a cada cliente.

OFICINAS	Registra los datos de las oficinas.
SALDO	Almacena el saldo promedio mensual de cada producto y cliente.

Fuente: Elaboración propia en base a documento "Diccionario de Datos Datawarehouse BancoEstado 2004"

A continuación se detallan las características más relevantes de las tablas expuestas anteriormente:

3.2.1 Canales

La tabla canales almacenan los códigos de los 58 canales a través de los cuales el total de clientes pueden realizar transacciones y otros considerados canales internos. Entre los no internos se encuentran: cajas, internet, cajeros, buzoneras, redbanc, caja vecina, call center, etc.

De los antecedentes preliminares es posible notar que los canales utilizados en el panel deben enmarcarse dentro de los mismos estándares utilizados para el resto de componentes del modelo de calidad siendo ellos: máquinas, sucursales e internet.

Para lograr dicho objetivo fue realizada una clasificación de canales en dichas categorías y para ello se consideraron las transacciones realizadas en cada canal y las opiniones de los ejecutivos de la banca comercial respecto a la relevancia de las transacciones y de los canales a clasificar.

El resultado de la clasificación realizada se presenta a continuación, para detalle del análisis realizado ver anexo B.

Tabla 4: Clasificación Canales

Máquinas	ATM RedBanc Buzonera Dispensadores La Polar Presto Coopeuch
Internet	Internet
Presenciales	Sucursales Casa Matriz Caja Vecina ServiEstado

Fuente: Elaboración propia

Cabe señalar que canales como call center y wap no fueron considerados por ser de uso menor y encontrarse fuera de los alcances de la tesis.

3.2.2 Clientes

Estas tablas almacenan la información de distintos atributos del total clientes de la entidad. Además del identificador único de cada cliente (id) las variables almacenadas pueden ser clasificadas como:

(i) Variables de vínculo:
   a) Antigüedad del cliente
   b) Convenio Pago Abono de Remuneraciones (PAR): Marca tenencia de convenio mediante el cual el empleador paga las remuneraciones del cliente

y le entrega algún producto para poder extraer los sueldos por Cajeros Automáticos (ATM) o Redbanc.

(ii) Variables demográficas: Género, Nacionalidad, Edad, Renta, Región, Comuna, Estado Civil, Nivel Educacional y Profesión.

Cabe recordar que el panel se restringe al estudio del comportamiento de las personas naturales (sin giro comercial) quedando, por tanto, exentos de los análisis los segmentos Empresas, Micro y Pequeñas Empresas.

Una vez solicitada la información de los clientes la primera tarea realizada fue la integración de todos los atributos disponibles uniendo a través del id (rut del cliente) los datos de los clientes.

Posteriormente se verificó el poblamiento de dichos atributos obteniendo los siguientes resultados:

Tabla 5: Completitud de Información

ATRIBUTO	COMPLETITUD
Región	97%
Comuna	97%
Segmento	100%
Estado Civil	97%
Nivel Educacional	63%
Profesión	30%
Renta	80%
Antigüedad	97%
Género	98%
Edad	98%

Fuente: Elaboración propia

La posibilidad de completar los datos con información confiable sólo fue posible para las variables región y comuna.

Para cumplir con este objetivo se realizó el contraste entre el dato demográfico de la región de los clientes y el dato correspondiente a la región de la oficina asignada por el Banco a cada uno de ellos, obteniendo una diferencia del 0,1% entre ambos datos del cliente.

Además utilizando tres muestras aleatorias pilotos (25000 clientes c/u) se llevó a cabo una segunda revisión la cual consistió en la comparación de la región demográfica y la región dónde el cliente ha transaccionado más durante los 2 últimos años de datos disponibles obteniendo los siguientes resultados[6]:

Tabla 6: Resultado Análisis Región Transaccional

MUESTRA	% DE COINCIDENCIAS
M 1	59,6
M2	74,7
M3	63,9

Fuente: Elaboración propia

La necesidad de utilizar muestras pilotos para realizar la comparación antes expuesta es por imposibilidad de manejar los datos transaccionales de la población completa.

[6]Detalles Anexo C,

De los resultados de las comparaciones anteriores se decidió utilizar el dato región de la oficina asignada como dato sustituto en los casos de valores perdidos por ser el valor más próximo al real.

Cabe señalar que la diferencia de un 30% entre la región transaccional del cliente y el dato demográfico obtenido en las tres muestran pilotos hacen tener en cuenta posibles des-actualizaciones en el valor demográfico que registra la entidad.

3.2.3 Segmentos

Las tablas de segmentos complementan la información de los clientes de la entidad y son obtenidas como resultado del proceso de segmentación llevado a cabo por la institución en base a ciertos atributos[7] del cliente.

Los registros almacenados son:

a) Segmento: Clasificación en diferentes segmentos, de acuerdo a su naturaleza: Personas, Empresas, Micro y Pequeñas Empresas.
b) Subsegmento: Corresponde a los 6 grupos de clasificación del segmento personas y desde ahora en adelante serán designados con los nombres (S1–S6).

3.2.4 Contratos

Las tablas de contratos almacenan los datos de los contratos establecidos por los clientes entre cuya información destaca: la fecha de apertura y término del contrato, el tipo de producto contratado, un código único asociado al producto contratado y el cliente al que pertenece.

3.2.5 Productos

Esta tabla contiene la lista los códigos de los productos manejados por la entidad a la fecha de realización de la tesis.

Cabe señalar que la entidad utiliza el concepto “producto” de forma ampliada pues califican con códigos de productos algunos servicios adicionales ofrecidos por la institución como la venta y compra de dólares o algunos asociados a la tenencia de algún contrato regular, como por ejemplo: productos complementarios a créditos hipotecarios o la tarjeta de débito asociada a cuentas vistas y corrientes.

Además de los códigos de identificación de los productos la entidad cuenta con dos agrupaciones para ellos, las cuales fueron estudiadas pues no son consistentes ya que por ejemplo la clasificación 1 asocia al producto cuenta corriente con 5 códigos (esto es 5 tipos de contratos) y en la segunda clasificación sólo a 1.

Y por lo tanto al realizar el cómputo de clientes con contratos por producto los resultados son diferentes.

Con dichos datos calculados (contratos por producto) y con know how de ejecutivos de la entidad fue definida la clasificación producto_segmentacion2 la cual clasifica los productos gestionables en 17 niveles y serán los utilizados como productos en el desarrollo de esta tesis (P1-P17)[8].

[7] Los atributos y los segmentos no serán definidos por confidencialidad de la información.
[8] La clasificación de productos no es ilustrada por ser datos confidenciales de la entidad.

3.2.6 Oficinas

La tabla oficinas guarda la información de las distintas sucursales de la entidad, los datos utilizados para la construcción del panel de clientes son: código de la oficina, región y comuna a la que pertenece.

Además como fue explicado anteriormente para completar los datos región y comuna de los clientes sin esta información se trabajó con la tabla oficina asignada la cual registra la asociación de los clientes a cada una de las oficinas registradas en la tabla oficinas[9].

3.2.7 Transacciones

La tabla de transacciones almacena todas las actividades realizadas por los clientes en un periodo de 24 meses.

De los atributos de esta tabla son de utilidad para el proyecto: el código de la transacción, el identificador del cliente, el código del canal utilizado, el monto de la transacción, el código del producto asociado y la fecha de realización de la transacción.

Debido a la alta cantidad de registros contenidos en esta tabla su manejo es externo al Datamart de Marketing, disponible para la presente tesis, y la información que se manejará de ella corresponderá de manera exclusiva a los datos de pruebas y los datos finales del panel.

3.2.8 Saldos

Esta tabla contiene información del saldo promedio diario mensual de cada cliente con cada producto que posee a lo largo del tiempo.

Sus principales atributos son: identificador del cliente, identificador del contrato, fecha, saldo promedio.

3.2.9 Deuda SBIF

Esta tabla almacena la información entregada mensualmente por la SBIF a los bancos en la cual se incluyen las deudas consolidadas de las personas. La historia almacenada corresponde a los últimos 12 meses entre sus registros se encuentra:

a) Monto deuda vigente: Deuda que posee una persona, sea esta de consumo, hipotecaria, comercial.
b) Monto deuda vencida: Es la cantidad, en pesos, de endeudamiento vencido (sobre 120 días de atraso de pago) en el sistema, de una persona.
c) Monto deuda consumo: Es la cantidad, en miles de pesos, de deuda de la persona en cualquier institución financiera. Incluye las deudas de consumo de líneas de crédito, crédito de consumo y tarjeta de crédito.
d) Monto deuda hipotecaria vivienda: Es la cantidad, en miles de pesos, de deuda de la persona en cualquier institución financiera por concepto de créditos hipotecarios.
e) Monto deuda comercial: Corresponde a los créditos solicitados por las empresas.

[9] Para detalles ver Anexo D

Capítulo 4

Comprensión de variables de interés

El estudio de las variables de interés presentado a continuación tiene por finalidad entender las características de las variables que son necesarias para la realización de estudios futuros del comportamiento de clientes y por lo tanto el objetivo del capítulo es lograr tener el conocimiento dicha información para la toma de decisiones de muestreo para el panel a construir[10].

4.1 Variable Transacciones

En el contexto de una institución financiera la data transaccional se refiere tanto el monto como a la cantidad de transacciones que los clientes realizan a través de los distintos canales disponibles, tal como fue explicado en el capítulo anterior estos registros son almacenados por un período de 24 meses para todos los clientes.

Dada la magnitud de dichas tablas estas no se encuentran disponibles para la población completa y por lo tanto para comprender el comportamiento de las variables monto y cantidad se solicitaron tres muestras aleatorias pilotos de 25000 clientes con las cuales se realizó un análisis de las transacciones mensuales realizadas dentro de 1 año de datos (102007-102008).

Cabe señalar en este punto que la data transaccional solicitada correspondió a 2 años de registros y la cantidad de clientes contenidos no superó 18500 clientes en los tres pilotos levantados.

La solicitud de tres muestras con un número elevado de clientes tuvo como objetivo validar que para todas se obtuvieran características similares y por lo tanto una vez verificado aquello se procedió a utilizar una única muestra obteniendo los siguientes estadísticos descriptivos:

Tabla 7: Estadísticos Descriptivos Variable Monto de Transacciones Mensuales

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Monto	73074	1	277037362	465912,026	2363421	5,58576E+12	5,07

Fuente: Elaboración propia

De dichos resultados es posible notar el elevado orden de magnitud en la varianza (10^{12}) obtenida y que sumado al hecho de un coeficiente de variación demasiado elevado permiten notar la dispersión excesiva de los datos.

[10] Para detalles ver Capítulo 2 – Sección 2.7

A modo intentar disminuir la dispersión de la variable monto se procedió a identificar los posibles valores *outliers* utilizando el método del gráfico de cajas.

Dicho método plantea que el rango intercuartílico ($dQ= Q_3-Q_1$) que corresponde a la distancia entre el tercer y primer cuartil[11]) es resistente a la presencia de valores *outliers* y por lo tanto sirven para definir los valores atípicos moderados y severos de la siguiente manera[12]:

Valor atípico moderado:

$$< Q_1-1,5(Q_3-Q_1)$$

$$> Q_1+1,5(Q_3-Q_1)$$

Valor atípico extremo:

$$< Q_1-3(Q_3-Q_1)$$

$$>Q_1+3(Q_3-Q_1)$$

Utilizando el software SPSS fueron obtenidos los percentiles de la variable y con ello se obtuvo el valor de los cuartiles requerido para el cálculo de valores outliers:

Tabla 8: Resultado de Percentiles y Cálculo de Distancia Intercuartílica

Percentiles									
5	**10**	**25**	**50**	**75**	**90**	**95**	**dQ**	**1,5*dQ**	**3*dQ**
8000	15000	40035,8	108000	313882	786285	1474624	273847	410770	821540

Fuente: Elaboración propia

De dónde se obtienen las siguientes cotas:

Tabla 9: Cotas de Valores Atípicos

Inf Moderado	-370734
Sup Moderado	724652
Inf Severo	-781504
Sup Severo	1135422

Fuente: Elaboración propia

Los valores de las cotas inferiores son descartados debido a la naturaleza positiva de la variable.

Como valor de cota superior se decide utilizar el valor atípico moderado pues a pesar de que este deja fuera un 10% de los datos se desea visualizar el efecto que tiene eliminar la mayor cantidad de posibles valores fuera de rango.

Tabla 10: Estadísticos Descriptivos Variable Monto Sin Valores Atípicos

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Monto	65117	1	724525	154675,29	163219,104	26640476052	1,06

Fuente: Elaboración propia

De los resultados obtenidos se tiene que la eliminación de los valores extremos logra disminuir el coeficiente de variación, sin embargo es aún insuficiente para asegurar la representatividad de la media (CV>1)[13].

[11] Q_1 valor que deja debajo el 25% de las observaciones y Q_3 el 75% de las observaciones
[12] Fuente : <http://es.wikipedia.org/wiki/valor_atípico>
[13] Ver Capítulo 2 - Sección 2.8.1

A su vez la varianza también tiene una disminución importante lo que muestra su alta sensibilidad a la presencia de valores *outliers* y a lo poco recomendable que es su utilización para distribuciones asimétricas (en este caso asimétrica positiva por la existencia de una mayor concentración de valores a la izquierda de la media).

El elevado coeficiente de variación, la alta diferencia en la varianza, la existencia de una distribución asimétrica, la sensibilidad a la presencia de valores extremos y la disminución de un 67% en el valor de la media obtenido para la población estudiada hacen concluir que es sospechable que el 10% eliminado no corresponda a valores fuera de rango sino que posiblemente corresponda a una segunda población dentro de los clientes de la entidad.

Para la variable cantidad de transacciones mensuales se realizó el mismo análisis obteniendo:

Tabla 11: Estadísticos Descriptivos Variable Cantidad de Transacciones (Txns.)

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Txns Cantidad	81734	1	588	7,00	12,79	163,64	1,83

Fuente: Elaboración propia

Tabla 12: Curtiles Variable Cantidad de Txns.

Cuartiles			Distancia Intercuartil		
25	**50**	**75**	**dQ**	**1.5dQ**	**3dQ**
1	3	7	6	9	18

Fuente: Elaboración propia

Tabla 13: Cotas de Valores Atípicos

Inf Moderado	-8
Sup Moderado	16
Inf Severo	-17
Sup Severo	25

Fuente: Elaboración propia

De los resultados de la tabla 11 nuevamente es posible observar la alta heterogeneidad de la variable cantidad de transacciones, además de las cotas obtenidas para los valores atípicos se tiene que los límites inferiores no deben ser considerados y analizado los casos de ambos límites superiores se tiene:

a) Eliminando los valores atípicos extremos se obtiene:

Tabla 14: Estadísticos Descriptivos Variable Cantidad de Tnxs

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Txns Cantidad	77498	1	25	4,75	4,85	23,56	1,02

Fuente: Elaboración propia

b) Eliminando todos valores atípicos se obtiene:

Tabla 15: Estadísticos Descriptivos Variable Cantidad de Tnxs

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Txns Cantidad	74058	1	16	4,03	3,54	12,55	0,88

Fuente: Elaboración propia

De la tabla 15 se tiene que la eliminación de todos los valores extremos logra una disminución notable en la dispersión de los datos pero nuevamente la eliminación corresponde aproximadamente al 10% de los datos, a pesar de obtener un CV de variación pequeño el cambio en la media es aproximadamente de 47%[14] y por lo tanto es sospechable que dicho valor no sea representativo de la población.

Además en este caso es fácil notar que dada la asimetría positiva de la distribución de la cantidad de transacciones mensuales de los clientes (pues los valores de encuentran más concentrados a la izquierda de la media) la mediana (Q_2=3, valor por el cual se encuentra debajo el 50% de la población) es más representativa y menos sensible a los valores extremos pues aún eliminando el 10% correspondiente a valores atípicos el resultado sigue siendo 2 como se ilustra en la siguiente tabla:

Tabla 16: Tabla de percentiles sin valores extremos considerados

	Percentiles						
	5	10	25	50	75	90	95
Cantidad	1	1	1	2	4	9	12

Fuente: Elaboración propia

De los análisis realizados y lo expuesto anteriormente se concluye que los estudios de comportamiento de clientes basados en la actividad transaccional mensuales (monto y cantidad) deberán ser efectuados a partir de la estimación de proporciones en vez de la estimación puntual de la media por encontrarse dicho valor poco fiable y además por la sospecha de la existencia de una segunda población debido a la coincidente y elevada proporción de valores atípicos en ambos casos.

De lo anterior se obtuvo la necesidad de crear una clasificación de dichas variables con el fin de permitir el objetivo de clasificación de la población en las categorías definidas por dicha clasificación y por lo tanto hacer viable la estimación de proporciones en los futuros estudios de comportamiento a realizar con el panel.

4.1.1 Clasificación de Transacciones según Tipo

Para obtener el conocimiento en detalle de la información que agrupa la variable transacciones fue necesario crear una estructura sencilla que permitiera dar un esbozo de las distintas acciones realizadas por los clientes.

Esta estructura fue replicada para cada uno de los productos del banco y como ejemplo a continuación se presenta la matriz construida para el producto ahorro:

Tabla 17: Acciones Realizadas por los clientes con producto Ahorro

PRODUCTO	ACCION CLIENTE	CANALES INVOLUCRADOS				
		SUCURSAL	INTERNET	CAJERO	BUZONERA	C.CENTER
AHORRO						
	Actualiza libreta	1		1	1	
	Abre cuenta	1				
	Cambia Libreta	1				
	Solicita Cartola		1		1	
	Depósito en Doc.	1		1	1	

[14] Cambio de 7 a 4 transacciones mensuales.

	Depósito en Efec.	1	1	1	1	1
	Dep. por Traspaso			1		
	Giro	1	1	1		1
	Pago Cargo Ahorr.		1			
	Transferencia entre ctas.		1			

Fuente: Elaboración propia

Del análisis de las acciones de los clientes en el banco fue posible distinguir que en los registros existen distintos tipos de transacciones. Por ejemplo, es interesante notar que algunas de ellas son producto de operaciones internas del banco (por ej. intereses, reajustes, etc.) o que otras no implican movimientos de dinero (por ej. consultas, aperturas de cuenta, etc.)[15].

A partir de esta estructura, junto al apoyo de los ejecutivos de marketing y de dataminig, se llevo a cabo un estudio de las 7667 transacciones totales de la base con el objetivo de realizar una clasificación de ellas para su manejo en el panel.

El resultado de dicho proceso fue almacenado en una tabla de clasificación de transacciones que asocia el código de cada transacción con su correspondiente clasificación siguiendo los siguientes criterios:

Tabla 18: Clasificación de Transacciones 1

VOLUNTARIA:	Aquella en que el cliente es consciente que está realizando una acción sobre un producto contratado. Por ejemplo: giros, depósitos, consultas, transferencias, etc.
INVOLUNTARIA:	Aquella efectuada sobre un producto y no depende de la voluntad del cliente. Por ejemplo: reajustes, chequeos de información, intereses, etc.
FINANCIERA:	Es aquella que está asociada a algún tipo de movimiento de dinero. Por ejemplo: Transferencia de Cuenta Corriente a Línea de Crédito.
NO FINANCIERA:	Es aquella que no está asociada a movimiento de dinero. Por ejemplo: Consulta de Saldo.

Fuente: Elaboración propia

Esta clasificación realizada para la variable transacciones da origen a la necesidad de estimar la distribución de la población en 4 categorías (transacciones financieras– no voluntarias, financieras voluntarias, no financieras – voluntarias, no financieras – no voluntarias) y por lo tanto corresponde a la primera cota de la cantidad de clientes requeridos para la realización de estudios confiables en cada nivel de inferencia necesario[16].

4.1.2 Clasificación de Transacciones por Canal

Del análisis del negocio se tiene que otro de los requerimientos a realizar con el panel de clientes corresponde a la determinación de la transaccionalidad de canales.

Para cumplir con dicho objetivo se tiene por lo tanto el panel deberá ser capaz de responder a la estimación de las proporciones en dicha categoría y de lo

[15] Ver ejemplos de otros productos en Anexo E.
[16] Ver Capítulo 2 - Sección 2.8: Determinación del tamaño de muestra

explicado en el capítulo anterior se tiene que dicha clasificación comprende 3 niveles siendo ellos: máquinas, internet y sucursales o canales presenciales.

4.1.3 Clasificación de Transacciones según Monto y Cantidad

Otra propuesta de clasificación para la realización de inferencia en los estudios posteriores del panel es la que se obtiene al realizar la transformación de la variable continua monto de transacciones mensuales a una variable discreta que sea capaz de agrupar en intervalos los valores posibles de dicha variable.

Más que el nivel de detalle en la determinación de los montos de corte de los intervalos el interés, que posteriormente pueden ser modificados según el nivel de inferencia y contexto general de los estudios a generar, este análisis tiene por objetivo determinar la cantidad de categorías o grupos de clasificación para los cuales el panel deberá ser capaz de responder desde la perspectiva menos buena (distintos clientes, distintos productos, alta variabilidad, etc.) pues, como se expondrá en el siguiente capítulo, dicha cantidad de categorías será el que determine el tamaño de muestra requerido[17] para cada nivel de inferencia necesario.

A modo de incluir la máxima información disponible, evitar sesgos incorporados por estacionalidades de algún mes en particular e incluir la máxima heterogeneidad de los datos para definir la cantidad de niveles se consideraron todas las transacciones financieras realizadas mensualmente por los clientes de la entidad para 1 año de información.

Para definir la cantidad óptima de categorías se utilizó el criterio de discretización por intervalos de igual frecuencia analizando para distintas cantidades de grupos la distancia de los cortes de los intervalos resultantes en cada caso.

En el estudio se comenzó analizando 5 intervalos pues como se expuso anteriormente ya se tiene la cota de clasificación según tipo de transacción correspondiente a 4 niveles, posteriormente se agregaron niveles obteniendo en resumen los siguientes resultados:

Tabla 19: Análisis de Discretización Variable Monto de Txns.

Intervalos a 5 Niveles		**Intervalos a 6 Niveles**		**Intervalos a 7 Niveles**		**Intervalos a 8 Niveles**	
Variable Monto		Variable Monto		Variable Monto		Variable Monto	
%	Cortes	%	Cortes	%	Cortes	%	Cortes
20	30000	16,67	20400	14,29	20000	12,5	18000
40	61720	33,33	50000	28,57	41000	25,0	38000
60	121934	50,00	90000	42,86	70000	37,5	59432
80	260000	66,67	160000	57,14	110000	50,0	90000
		83,33	300000	71,43	194990	62,5	137000
				85,71	328159	75,0	210701
						87,5	356076

Fuente: Elaboración propia

De lo cual es posible concluir que la mejor clasificación corresponde a 6 niveles definidos de la siguiente manera:

[17] Detalles en Capítulo 2 - Sección 2.7. IV

- I1: Monto de transacciones inferior a 20000
- I2: Monto de transacciones entre 20000 y 50000
- I3: Monto de transacciones entre 50000 y 100000
- I4: Monto de transacciones entre 100000 y 200000
- I5: Monto de transacciones entre 200000 y 300000
- I6: Monto de transacciones mayor a 300000

Incluir una cantidad mayor de niveles no aporta mejor conocimiento del comportamiento de los consumidores pues los montos de cortes para 7 u 8 niveles no incluyen grandes diferencias en las categorías propuestas, dicho de otra forma generan intervalos de ancho muy pequeño.

Para la variable cantidad de transacciones mensuales se realizó un estudio de frecuencia de sus valores considerando los mismos datos descritos anteriormente obteniendo:

Tabla 20: Frecuencia Variable Cantidad de Txns.

Percentiles	Cantidad
5	1
10	1
15	1
20	1
25	1
30	2
35	2
40	2
45	2
50	3
55	3
60	3
65	4
70	4
75	7
80	7
85	9
90	10
95	16

Fuente: Elaboración propia

A partir de los cual se decide definir 5 categorías[18] las cuales corresponden a: 1, 2, 3, 4 a 7 y más de 7 transacciones mensuales.

4.2 Variable Tenencia

La tenencia por producto corresponde a la marca de si un cliente posee o no la existencia de un contrato vigente para dicho producto a una fecha determinada y por lo tanto se basa en la información de las tablas de contratos descritas anteriormente.

[18] Que corresponderán a 6 en caso de incluir la realización de cero transacciones mensuales.

Para darle valor a dicha variable se realizó la programación de consultas que permitieron lograr la generación de una nueva tabla que almacena dicha marca para todos los productos[19] y clientes de la entidad.

Dada la gran magnitud de información a evaluar fue necesario crear un procedimiento de consulta de bases de datos para el cual fue utilizando el software SQL Server el cual corresponde al motor de base de datos seleccionado para todo el desarrollo de la presente tesis[20].

A modo general las consultas programadas chequean la existencia de contratos vigentes al día de interés en ambas tablas de contratos para posteriormente agrupar y transponer los resultados con el fin de obtener la tenencia por producto y por cliente como variable binaria tal como se ilustra en el siguiente ejemplo[21]:

Tabla 21: Ejemplo Tabla Tenencia por Producto

ID_CLIENTE	TENENCIA_AHORRO	TENENCIA_CTA_CTE
153380457	1	1
124721098	0	0

Fuente: Elaboración propia

Además de la tenencia por producto se estudió la tenencia total por cliente (cantidad de productos que posee cada cliente) la cual es obtenida sumando las columnas de la tabla antes descrita.

Para comprender el comportamiento de dicha variable se utilizó el mismo programa para calcular los estadísticos descriptivos de la población completa obteniendo:

Tabla 22: Estadísticos Descriptivos Variable Tenencia Total

	MÍNIMO	MÁXIMO	MEDIA	DESV. TÍP.	VARIANZA	CV
T_total	1	15	1,82	1,35	2	0,7

Fuente: Elaboración propia

Estos valores muestran la factibilidad de la estimación de medias en esta variable debido tanto a la baja variabilidad observada en la varianza y en el coeficiente de variación que además por ser inferior a 1 expresa su buena representatividad.

Si bien es cierto que la variable tenencia total, recién descrita, puede ser utilizada por la entidad para medir ciertos índices de gestión también hace mucho sentido evaluar la tenencia de ciertas combinaciones específicas de productos (tenencia cruzada) cuyos resultados pueden ser utilizados por ejemplo: para focalizar campañas de marketing directo, campañas de fidelización, ente otros.

En relación a esto y a los objetivos globales del proyecto de gestión de calidad total que enmarca la realización de esta tesis fue calculada la variable GRS (Grupos de Relación Similar) la cual es descrita a continuación.

[19] Utilizando la clasificación de productos descrita en el capítulo anterior
[20] Seleccionado debido a su disponibilidad en la institución y familiaridad con el lenguaje sql de la alumna investigadora.
[21] Para detalle de la programación ver anexo F.

4.2.1 Variable GRS

La variable GRS es la variable clasificadora de la tenencia de productos que se tener en consideración para el presente trabajo de tesis pues es la que se utiliza a lo largo de todo el desarrollo del proyecto "Métricas para la calidad de Servicio BancoEstado".

Esta clasificación nace debido a la alta heterogeneidad que caracteriza a la población de clientes a estudiar la cuál fue posible observar, por ejemplo, en el estudio de las variable monto y cantidad presentado anteriormente y que tiene su fundamento en Misión de la entidad:

"Existimos para que cualquier chileno en cualquier lugar, pueda emprender y desarrollarse",

"Ser un banco universal, de todos y para todos"[22]

La idea principal de un GRS es conseguir la agrupación de los clientes según sus distintas experiencias y formas que tengan para relacionarse con el banco justificada en el hecho de que la calidad de servicio es una percepción del cliente sobre el servicio recibido el cual a su vez obtiene dicho concepto contrastado el servicio recibido con sus expectativas y dichas expectativas dependen de sus experiencias pasadas, de sus alternativas y de otras experiencias de consumo [18].

Como las experiencias de consumo de los clientes están fuertemente determinadas por el tipo de productos que estos poseen cada GRS fue definido en función de la variable tenencia descrita anteriormente y a continuación se presentan sus definiciones:

(i) GRS Relación Básica: Considera clientes que se relacionan exclusivamente a través de las Cuentas de Ahorro.
(ii) GRS Relación en Desarrollo: Son aquellos clientes en que la relación con el Banco se basa en el débito que ellos poseen, es decir, clientes sin tenencia de créditos de ningún tipo y que operan al menos a través de uno de los siguientes productos: Cuenta Vista, Cuenta RUT, Chequera Electrónica y Línea de Crédito Cta. Vista.
(iii) GRS Alta Interacción: Considera todos aquellos clientes que tienen una relación integral con el Banco, son sujetos de débito y crédito que interactúan al menos a través de alguno de los siguientes productos: Cuenta Corriente, Línea de Crédito Cta. Cte. y Tarjeta de Crédito.
(iv) GRS Relación Enriquecida: Son clientes que han sido clasificados en alguno de los GRS anteriores pero su relación se ve fortalecida por la tenencia de al menos un producto de crédito (consumo, universitario, hipotecario).

Debido a este requerimiento fue necesario realizar la programación de la clasificación de todos los clientes de la entidad en las 6 distintas categorías pues el modelo de calidad completo así lo requiere[23].

[22] BANCO ESTADO, 2007: Plan Estratégico 2007-2010: Un Banco para todos los chilenos [18].
[23] Desde ahora llamados indistintamente G1-G6 por confidencialidad de la información.

Dicha programación se basa en la tabla construida para la variable tenencia por productos pues de acuerdo a dichos resultados se clasifica a los clientes en sus respectivos GRS. Detalles de programación en anexo G.

4.3 Variable Saldo

Tal como fue explicado en el capítulo anterior dentro de la información disponible para el panel a construir se encuentran los datos del saldo promedio diario mensual, esto es el promedio mensual del saldo diario, de cada cliente con todos los productos que este tiene a lo largo del tiempo.

Para entender estos datos fue necesario realizar el estudio de dicha variable cuyos resultados permite obtener conclusiones similares a las expuestas anteriormente para las variables monto y cantidad de transacciones en relación a la dispersión de los datos y la representatividad de la media obtenida en la población[24].

Además cabe señalar que en el caso de productos de créditos el saldo puede ser negativo y corresponderá al monto de deuda del cliente con la entidad.

Debido a ello también se realizó la discretización de la variable con el objetivo de conocer la cantidad de niveles para los cuales se deberá estimar la proporción de clientes en los futuros estudios y por lo tanto también incluir esta variable en el diseño de la muestra.

Los niveles definidos son 5 si sólo se consideran productos de acreedores del cliente (aquellos en que el saldo a favor) y 6 si se incluyen todos los productos.

4.4 Variable Deuda con el Sistema

Cómo fue expuesto en el capítulo anterior esta la información corresponde a la deuda vigente de cada cliente y proviene de la SuperIntendencia de Bancos e Instituciones Financieras y es parte de la información disponible del panel de clientes.

Con apoyo de los ejecutivos de la entidad se definió que las métricas relevantes con respecto a esta variable corresponden a:

$$Market\ \ Share = \frac{Deuda\_Banco}{Total\_Deuda}$$

$$Market\ Share\ Clientes = \frac{Numero\_Clientes\_con\_Deuda\_BE}{Número\_total\_clientes}$$

De ellos se tiene que la clasificación necesaria para esta variable corresponde a 2 grupos los cuales son: cliente con deuda en BE o en otra entidad.

[24] Ver detalles anexo H.

Capítulo 5

Diseño de muestreo

A continuación se presentan las principales actividades desarrolladas para la toma de decisiones que guiaron la extracción de la muestra necesaria para el levantamiento del panel de clientes.

5.1 Selección del Tipo de Panel a Construir

El tipo de panel a construir debe tener en consideración el principal foco de éste: entregar información de calidad de forma rápida y eficiente para permitir el estudio del comportamiento de los clientes a lo largo del tiempo.

Luego de las definiciones presentadas en el marco teórico es posible desprender que dada la necesidad de asegurar la calidad de los resultados a lo largo del tiempo el panel debe ser dinámico-rotativo.

La representatividad de una única muestra a lo largo del tiempo, sólo podrá ser asegurada en los casos en que los elementos muestrales se mantengan estáticos a lo largo del tiempo, en este caso no es así pues los clientes cambian: se mudan, envejecen, cierran contratos, abren contratos, cambian de empleo, se casan, etc.

Según lo discutido en el marco teórico un panel dinámico-rotativo aportará la flexibilidad necesaria y permitirá controlar los desbalances ocasionados por los cambios en la muestra y población a lo largo del tiempo sin afectar la dirección de los resultados.

Además que sea rotativo aporta riqueza en la utilización del panel pues al permitir el ingreso de nuevos clientes y, a su vez, mantener los antiguos haciendo posible realizar varios tipos de análisis entre estos: análisis retrospectivos, seguimiento de efectos de acciones de gestión, análisis de camadas o la construcción de perfiles basados en comportamientos pasados.

5.2 Definición de Población objetivo y Marco Muestral

Para la definición de la población objetivo se determinó que el primer requisito para que un cliente de la base de datos pudiera ser considerado como miembro potencial del panel era tener al menos la tenencia de un producto vigente al último día de datos disponibles para el levantamiento del panel (31-10-2008)

Además debido a la necesidad de congruencia entre los objetivos del panel y los objetivos generales del proyecto de calidad fue necesario refinar dicho requisito

exigiendo además la pertenencia de estos clientes a alguno de los GRS definidos por el proyecto y anteriormente explicados.

Dicho de otro modo, el universo considerado es formado por los clientes que poseen un contrato vigente de al menos uno de los productos que definen los GRS.

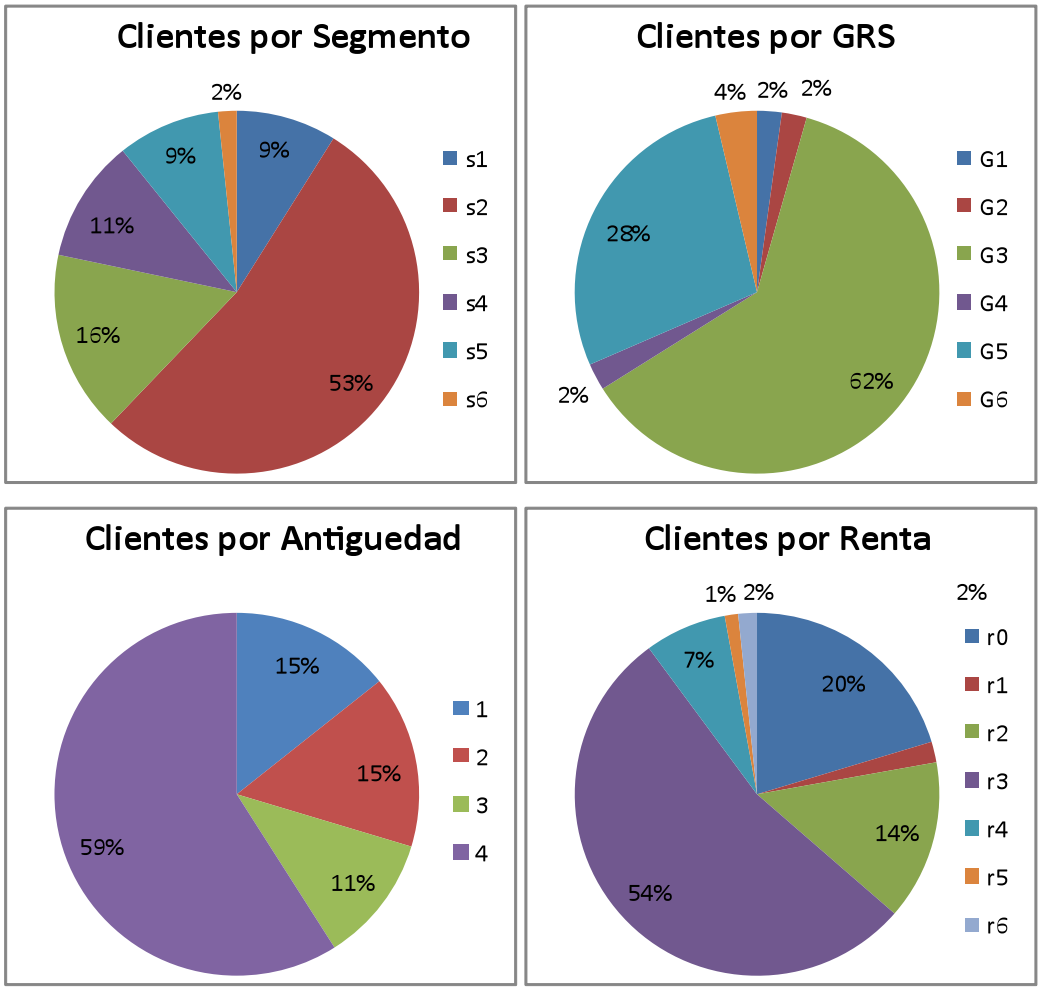
Cabe hacer notar que la cifra de clientes no clasificados en los GRS corresponde al 1,3% del total de clientes personas con contratos vigentes a la fecha en estudio de los cuales un 1,0% son sólo clientes de crédito (principalmente hipotecario) y un 0,3% de productos de inversión (depósitos a plazo, compra de acciones y fondos mutuos).

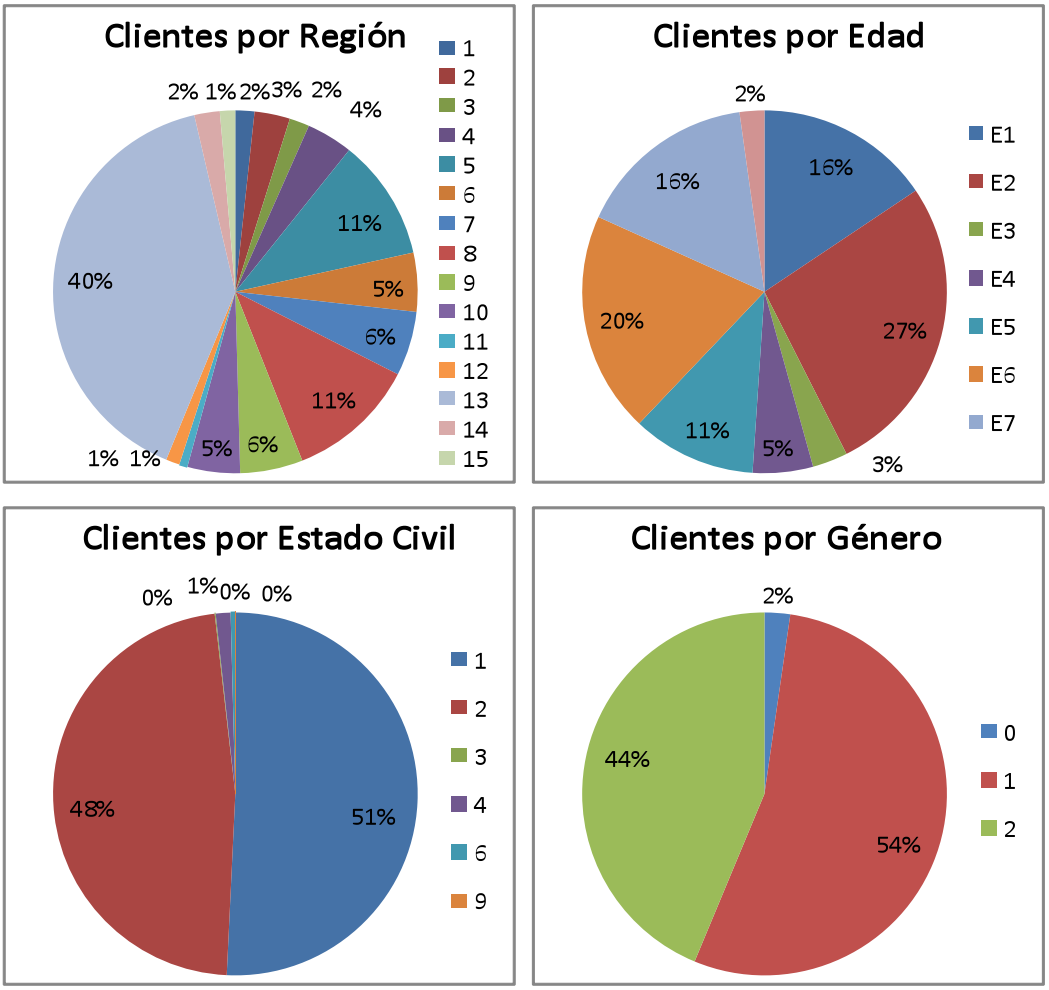
Debido a que es un porcentaje menor y a que la relación de estos clientes con la institución es considerada de vínculo menor para el proyecto en que se enmarca la tesis se decide que su exclusión de la población objetivo es correcta.

5.3 Selección de la Técnica de Muestreo

Para comprender las decisiones tomadas en este punto a continuación se describirán las distribuciones de la población en las distintas variables demográficas descritas anteriormente considerando sólo aquellas variables cuyo poblamiento de datos superior al 80%.

Cabe señalar que el nombre de los niveles de las variables ha sido reemplazado por letras y números sin un significado específico como medida de resguardo de la confidencialidad de los datos de la entidad.





De la investigación bibliográfica se determinó que los estudios de panel utilizan las mismas técnicas de muestreo que aseguran una buena calidad de datos en los estudios de corte transversal y estos son muestreos probabilísticos: aleatorios simples o estratificados [35].

Que sea probabilístico es necesario para poder proyectar con validez estadística los resultados de los estudios a toda la población pues con esto se asegura la existencia de una probabilidad conocida y distinta de cero de selección para todos los individuos.

La gran cantidad de variables y niveles para cada una de ellas genera un alto número de cruces en los cuales es posible encontrar clientes. Esto, sumado a la misión de inclusión que tiene la entidad[25] implica que existe muchos clientes distintos y con ello una alta probabilidad de comportamientos transaccionales diferentes entre ellos.

Según lo discutido en el marco teórico esta composición heterogénea de la población hace que un muestreo estratificado sea el indicado pues permite disminuir la variabilidad de los estimadores y por tanto mejorar las estimaciones a realizar con el panel.

[25] BANCO ESTADO, 2007: Plan Estratégico 2007-2010: Un Banco para todos los chilenos. [20]

Además de los gráficos de torta presentados se observa una alta variación de concentración en los distintos niveles de las variables descritas y por lo tanto un muestreo aleatorio simple hará obtener datos insuficientes para aquellos cruces de niveles menos poblados.

Debido a lo anterior se concluye que un muestreo aleatorio estratificado es la mejor técnica de extracción para los elementos del panel a construir, además, por lo referente a la concentración de individuos en ciertos cruces, se tendrá que una afijación proporcional no será suficiente para dar la flexibilidad buscada en los ángulos de observación del comportamiento de clientes y por lo tanto será necesario utilizar afijación no proporcional con el fin asegurar una cantidad suficiente de individuos en cada estrato definido y con esto en cada nivel de inferencia deseado.

5.4 Selección de Variables de Estratificación

Según lo discutido anteriormente lo que se espera de la estratificación a realizar es que dicho proceso ayude a cumplir con el objetivo de lograr la representatividad de todos los clientes en el panel a construir.

De lo expuesto en el marco teórico se sabe que es un requisito que los estratos sean discriminados por variables relacionadas a las características de interés pues de este modo se logrará formar grupos más homogéneos que la población completa y por ende reducir la variabilidad de dicha población permitiendo aumentar la eficiencia estadística de los estimadores.

El primer requisito impuesto a los atributos de los clientes para ser considerados como variables candidatas para estratificar fue la completitud de su información en la base de datos soporte del panel; esto es justificado en el hecho de que dicha información será necesaria para poder agrupar de forma excluyente y exhaustiva a todos los individuos de la población y por lo tanto con información faltante esto no es posible.

Se exigió un llenado mínimo del 97% de datos quedando como candidatas las siguientes variables:

Tabla 23: Resumen de la Completitud y Distribución de Atributos en la Población

Variable	% de Datos	Distribución (en % para niveles de la variable)													
Región	97%	40,1	11,5	10,8	5,7	5,6	5,2	4,7	4,1	3,2	2,3	1,8	1,7	1,4	1,2	0,7
Segmento	100%	53,2	16,1	10,9	9,1	8,9	1,7									
Estado Civil	97%	49,4	46,1	1,3	0,4	0,1	0,0	0,0								
Renta*	100%	53,5	20,3	14,2	7,3	1,8	1,7	1,2								
Antigüedad	97%	59,0	15,3	14,3	11,3											
Sexo	98%	54,0	44,0	2,0	0,0											
Edad	98%	31,9	23,3	19,0	13,1	6,4	3,7	2,6								
GRS	100%	61,7	27,8	3,7	2,4	2,2	2,2									

* dada la magnitud de clientes sin dato renta (20%) fue creada dicha categoría de renta sin dato

Fuente: Elaboración propia

A continuación se revisaron las distribuciones de la población en dichas variables y se realizó un filtro de concentración que se refiere a mantener como

variables candidatas a aquellas que presentan muchos clientes en ciertos niveles y muy pocos clientes en otros.

En este paso fueron descartadas las variables género, estado civil y antigüedad pues al no presentar los problemas de asimetría en sus distribuciones a pesar de que no se utilicen como variables de control una extracción aleatoria asegurará una cantidad suficiente de individuos en sus niveles (hombre/mujer, soltero/casado, cliente nuevo/antiguo).

El paso siguiente fue definir como primeras seleccionadas las variables segmento y grs pues además de que ambas variables presenten concentraciones muy altas (sobre el 50%) en un único nivel y muy bajas (inferior al 2%) para otros, de los antecedentes del proyecto en que se enmarca la presente tesis, se tiene que ambas variables son necesarias para las inferencias y estudio a realizar con el panel a construir y por lo tanto son un requisito de control fundamental para la extracción de la muestra.

Posteriormente fueron analizadas la variable renta y edad ya que ambas presentaban distribuciones asimétricas en la población. De dicho análisis se tiene que edad, a pesar de presentar los problemas de concentración, se encuentra bien diferenciada a través de la variable segmento.

A su vez, la variable renta posee los problemas de concentración mencionados y es considerada un aporte en las dimensiones de observación de comportamiento pues permite refinar los segmentos de clientes ya que en varios de ellos los rangos de renta que los definen son muy amplios[26].

Finalmente se decidió seleccionar la variable región pues, además pasar el filtro concentración, su control es considerado un apoyo frente a asuntos de gestión territorial de la institución.

Las cuatro variables de estratificación propuestas fueron presentadas y analizadas junto a ejecutivos de la entidad (clientes del proyecto en la institución) los cuales además de aprobarlas, contribuyeron en su selección apoyando con sus juicios y posibilidades de futuros usos de la herramienta a construir.

Cómo fue expuesto en el marco teórico las variables de estratificación deben estar relacionadas a las características de interés y por lo tanto debido a ello fue necesario estudiar dicho requerimiento para validar la selección realizada.

Para poder cumplir con esta tarea se trabajó con una de las muestras aleatorias pilotos de 25000 clientes para la cual se estudiaron las variables: monto de transacciones, cantidad de transacciones, saldo promedio y tenencia por ser estas las características de interés más relevantes de comportamiento y base del panel a construir.

Cabe señalar que los clientes considerados para determinar los valores de las variables monto y cantidad de transacciones son aquellos clientes que realizan transacciones financieras (monto>0) para el período Octubre 2007-Octubre 2008 y la

[26] Cabe recordar que las definiciones de los segmentos no pueden ser presentadas por ser información confidencial de la entidad.

variable tenencia es calculada, como fue expuesto en el capítulo anterior, al día 31-10-2008.

Un primer intento de verificación fue con la realización de un test anova factorial debido a que este además de contrastar la existencia de diferencias significativas entre las medias definidas por los niveles de cada factor y la combinación de estos permite obtener un análisis de regresión para las variables testeadas.

A pesar de que los resultados de las combinaciones de las 4 variables de estratificación seleccionadas son los mejores en términos de varianza explicada para las tres variables de interés (valor del R^2 ajustado en la regresión) el test no es considerado válido por no cumplir los requerimientos de homocedasticidad y normalidad necesarios (ver detalles en anexo I).

Finalmente, dado a que el interés no era definir un modelo de regresión para las variables de interés sino que determinar la existencia de una relación entre las variables seleccionadas para estratificar y las variables de estudio se decidió realizar test estadísticos individuales de contrastes de medias.

Estas pruebas permiten concluir respecto la existencia de diferencias significativas entre los promedios de las poblaciones formadas por los niveles de cada una de las variables de estratificación y por lo tanto permiten afirmar la dependencia o independencia entre ellas y las variables de interés.

A continuación se presenta el resumen de los test efectuados[27]:

Tabla 24: Estadísticos de Levene y Nivel de Significación

Variable	Segmento		GRS		Renta		Región	
	Est.	Sig.	Est.	Sig.	Est.	Sig.	Est.	Sig.
monto	420,4	0	631,4	0	503,7	0	2,6	0
cantidad	863,3	0	1584,5	0	954,9	0	5,3	0
tenencia	564,5	0	484,9	0	637,2	0	5,8	0
saldo	383,3	0	281,9	0	327,6	0	2,7	0

Fuente: Elaboración Propia

Como ha sido descrito en el marco teórico para la utilización del estadístico F-Fisher entregado por el test de igualdad de varianzas deben cumplirse los requisitos de normalidad y homocedasticidad.

Sin embargo, de la tabla 24 es posible observar que el nivel de significancia de las prueba de homogeneidad de varianzas es menor a 0,05 en todos los casos y por lo tanto permiten concluir que no se cumple dicha condición.

De lo discutido en el marco teórico se tiene que frente a estos casos es posible utilizar estadísticos robustos y/o algún test no paramétrico de contraste de medias.

Debido a ello fueron realizadas las pruebas robustas de Welch y Brown-Forsythe además del test no paramétrico de Kruskal-Wallis, descritas en el marco teórico, cuyos resultados se resumen a continuación (ver detalles en anexo J):

[27] Para detalles de las pruebas realizadas ver anexo J

Tabla 25: Significancia para Pruebas Robustas

Variable de agrupación	Segmento			GRS		
Variable de interés	Welch	Brown-Forsythe	Kruskal-Wallis	Welch	Brown-Forsythe	Kruskal-Wallis
Monto	,0000	1,7E-124	,0000	7E-155	4E-173	,0000
Cantidad	,0000	8E-230	,0000	,0000	,0000	,0000
Tenencia	,0000	,0000	,0000	,0000	,0000	,0000
Saldo	,0000	,000	,000	,000	,000	,000
Variable de agrupación	Renta			Región		
Variable de interés	Welch	Brown-Forsythe	Kruskal-Wallis	Welch	Brown-Forsythe	Kruskal-Wallis
Monto	,0000	3E-91	,0000	8E-05	0,001167	7,4E-26
Cantidad	,0000	9,2E-287	,0000	6E-18	1,95E-14	1,4E-29
Tenencia	,0000	,0000	,0000	8E-30	2,17E-28	6,9E-34
	,000	,000	,000	,000	,000	,000

Fuente: Elaboración Propia

De los resultados expuestos en la tabla 25 se tiene que los valores significativos para todos los estadísticos son inferiores a 0,05 rechazar la hipótesis nula de igualdad de medias permitiendo concluir que el efecto de cada variable de estratificación seleccionada no es igual en todos los niveles y por lo tanto si están relacionadas con las características de interés.

Debido al cumplimiento de este objetivo y al gran número de estratos formados como resultado del cruce de estas cuatro variables (1980 casillas) se concluye que la incorporación de una 5ta variable conlleva a un aumento en la complejidad de la estratificación deseada y por lo tanto se seleccionan como variables de estratificación las cuatro variables verificadas.

5.5 Definición de niveles de estratificación

Las variables de estratificación seleccionadas conforman un total de 132 cruces por región (figura 2) lo que conlleva a que si por ejemplo se consideraran 100 elementos por casilla la muestra obtenida correspondería a 198 mil clientes.

Dicha cantidad de clientes hace no viable la mantención de los datos transaccionales a lo largo del tiempo, como referencia se tiene que para un rango de 25000 sólo los datos de transacciones, saldos y SBIF clientes implican aproximadamente 3 millones de registros en 1 año y por lo tanto debido a las restricciones computacionales (capacidad y velocidad de respuesta) y a la necesidad de mantener en el tiempo el respaldo de todos los datos de la base descrita en el capítulo 3 se decide profundizar el estudio a modo de definir los cortes a considerar para la estratificación de la población.

Figura 2: Cruce de variables por región

		SEGMENTO/RENTA																				
		S1							S2			S3							S4	S5			S6
REGION	GRS	R0	R1	R2	R3	R4	R5	R6	R0	R2	R3	R0	R1	R2	R3	R4	R5	R6	R0	R4	R5	R6	R1
1	G1																						
	G2																						
	G3																						
	G4																						
	G5																						
	G6																						

Fuente: Elaboración propia.

Para las variables segmento y grs es imposible disminuir sus nivele pues ambas son pilares de las dimensiones de observación del proyecto en que está inmersa la presente tesis y por lo tanto son requisitos de control para el panel a construir.

Para el análisis de las diferencias significativas entre los niveles de las variables de estratificación restantes (renta y región) fue realizado utilizando el software SPSS el cual junto a la ejecución de los test de igualdad de medias descritos en el punto anterior permite la realización de los llamados “análisis post-hoc” los cuales realizan la comparación de igualdad de medias para todos los pares de niveles de la variable factor.

Las pruebas que se realizan en los análisis post-hoc deben ser seleccionadas según el cumplimiento o no del supuesto de igualdad de varianza y, como fue expuesto en el punto anterior, dicho supuesto no se cumple para ninguna de las variables de interés en los factores a estudiar (renta y región)[28] y por lo tanto se utilizó la prueba de Games-Howell[29].

Para la determinación de los niveles de la variable renta debe tenerse en consideración que ésta es agregada como variable de estratificación para refinar los niveles en la variable segmento y el estudio fue realizado para los grupos formados por dicha variable en cada uno de los segmentos en que dicha se encuentra presente en más de un nivel (S1, S2, S3 y S5).

La tabla presentada a continuación resume los niveles con diferencias significativas para cada uno de los segmentos y niveles estudiados[30]:

[28] Ver Anexo J.
[29] Prueba que mejor controla la tasa de error bajo el incumplimiento de la homocedasticidad [25]
[30] Ver Anexo K.

Tabla 26: Diferencias significativas por segmento según prueba de Games-Howell

Segmento	(I) Renta	(J) Renta	Sig.	Segmento	(I) Renta	(J) Renta	Sig.
1	0	2	0,000	3	0	2	0,000
	0	3	0,000		0	3	0,000
	0	4	0,000		0	4	0,017
	0	6	0,001		2	3	0,000
	3	2	0,000		3	4	0,034
	3	4	0,000	5	4	6	0,011
	3	6	0,009		4	5	0,044
2	0	2	0,000		5	4	0,044
	0	3	0,000		6	4	0,011
	2	3	0,000	*La diferencia de medias es significativa al nivel .05.			
	3	0	0,000				

Fuente: Elaboración propia

De lo cual se decide agrupar la variable renta en los siguientes niveles:

Tabla 27: Agrupación tramos de Renta por Segmento

Segmento	Agrupación Renta
S1	R0 R3 R1-R2-R4-R5-R6
S2	R0 R2 R3
S3	R0 R3 R1-R2-R4-R5-R6
S4	R0
S5	R4 R5-R6
S6	R1

Fuente: Elaboración propia

Cabe señalar que el resultado de corte en los niveles de la variable renta fue validado por los ejecutivos de la entidad quienes corroboraron los grupos propuestos en función de su know how respecto a las diferencias de comportamiento de clientes según sus niveles de ingreso.

El segundo estudio correspondió a los niveles de la variable región, para ello se realizó una agrupación en zonas considerando su ubicación geográfica, la cantidad de clientes en cada una de ellas y que no se incluyeran en un mismo grupo regiones cuyos promedios de las variables de interés no presentaran diferencias significativas.

Para cumplir con el último requisito se realizó el estudio de diferencias significativas expuesto en el anexo L.

Dicho análisis permite notar que las mayores diferencias significativas son obtenidas por las regiones VII, V y XIII; dado esto y el hecho de que dichas regiones son aquellas con mayor proporción de población se decide mantenerlas como zonas independientes pues el permitirá extraer una mayor cantidad de elementos muestrales para ellas y por lo tanto disminuir los efectos de una mayor variabilidad debido a la existencia de una mayor proporción de individuos en dichos niveles.

El resto de diferencias significativas es obtenido entre regiones de norte y del sur lo que no presenta inconvenientes pues no existe la intención de agruparlas.

De los resultados obtenidos se decide disminuir los 15 niveles de la variable región a 6 zonas las cuales, sin considerar la distribución para la región metropolitana, logran mejorar la simetría de la cantidad de clientes por zona y cumplen el requisito de las diferencias significativas planteado anteriormente.

Tabla 28: Agrupación de Regiones en Zonas

Zona	% de clientes	Regiones
1	12,14	XV-I-II-III-IV
2	10,77	V
3	10,93	VI-VII
4	11,51	VIII
5	14,51	IX-X-XI-XII-XIV
6	40,13	XIII

Fuente: Elaboración propia.

Reduciendo de esta forma los 1980 cruces a 468 que si posibilitan una extracción del orden de 100 clientes por casilla.

5.6 Tamaño de la Muestra

Cómo fue discutido en el marco teórico la definición del tamaño de muestra debe estar guiado tanto por la naturaleza de las variables a estudiar, el nivel de confianza, el error aceptado y el tipo de inferencia o dominio de los estudios a realizar con el panel.

Del estudio de las variables de interés presentado en el capítulo 4 se tiene que el tamaño de muestra a extraer deberá basarse en los tamaños de muestra para la estimación de proporciones en varios niveles o categorías.

La revisión bibliográfica realizada para la presente tesis permite concluir la mejor estimación del tamaño de muestra a realizar en dichos casos corresponde a:

$$n = \max\left\{ \frac{B\pi_i\left(1-\pi_i\right)}{b_i^{\,2}\left(N-1\right)+B\pi_i\left(1-\pi_i\right)} \right\}$$

Donde:

n = tamaño de la muestra

B es el percentil superior (α/k *100) de una distribución chi-cuadrada con 1 grado de libertad.

k = cantidad categorías a estimar proporción

n_i = frecuencia de observación de la i-ésima categoría

π_i = proporción de la población ubicada en la i-ésima categoría

α = nivel de confianza

b_i = precisión aceptada

Derivando la fórmula antes descrita e igualando a cero se obtiene que el valor máximo tamaño de muestra obtenido es para $\pi_i = 1/2$, en efecto:

$$\frac{\partial n}{\partial \pi_i} = \frac{(1-2\pi_i)(c + B\pi_i(1-\pi_i) - (1-2\pi_i)B\pi_i(1-\pi_i)}{(c + B\pi_i(1-\pi_i))^2} = \frac{(1-2\pi_i)\cdot c}{(c + B\pi_i(1-\pi_i))^2} = 0$$

De donde $(1-2\pi_i)\cdot c = 0 \Rightarrow \pi_i = 1/2$, ya que $c = b_i^2\,(N-1) \neq 0$

Además de lo discutido en el capítulo anterior la cantidad máxima de categorías a estimar corresponde a 6 para lo cual considerando una precisión del 3% y una confianza del 95% aplicando la fórmula antes descrita se tiene que el tamaño de muestra requerido para cada una de las categorías de inferencia corresponde a 1933[31] clientes.

De no realizar la corrección del tamaño de muestra por la cantidad de niveles necesarios a estimar proporción (o si estos fueran sólo 2) el tamaño de muestra necesario correspondería a 1067 individuos[32] para el mismo nivel de confianza y error.

Cabe señalar que los niveles de inferencia para los cuales el panel deberá cumplir este requisito corresponden a[33]:

1. Segmento
2. GRS
3. Zona
4. Renta
5. Nivel Educacional
6. Género
7. Edad
8. Antigüedad
9. Segmento-GRS
10. Segmento-Renta
11. Segmento-Zona
12. GRS-Zona
13. GRS-Renta
14. Segmento-GRS-Renta (GRS=5 y Segmento=2)
15. Productos

[31] Valor sobreestimado pues es aquel que no considera la corrección por finitud. Ver tabla 30 y 31 tamaños ajustados.
[32] Ver Capítulo 2-Sección 2.7 (VI.b)
[33] Determinados del análisis del negocio explicado en el Capítulo 3

Para tomar la decisión de la cantidad de clientes por estrato se realizó la clasificación de todos clientes en los 468 grupos definidos determinando que muchos de ellos poseían muy pocos clientes (0,4% en total), analizando dichos casos se tiene que aparecen por ejemplo niños con productos de crédito o cuentas corrientes, personas con cuenta corriente sin datos de renta y por lo tanto se excluyen de la población objetivo pues es altamente probable que correspondan a errores en los registros.

Eliminando los clientes antes descritos la estratificación queda en una grilla de 312 estratos la cual corresponden a 8297628 clientes y representa el 99,6% de la población total de clientes de la entidad.

Luego de esto se realizó la cuantificación poblacional de los clientes por cada categoría de inferencia requerido yse programó en Excel una planilla para el cómputo automático del tamaño muestral necesario para cada una de ellas en función de los tamaños poblacionales y los parámetros descritos en la tabla 29.

Cabe señalar que el tamaño de cada casilla de la grilla de estratificación fue definido a prueba y error a modo de cumplir con todos los requisitos antes descritos y a continuación se presenta como ejemplo el listado de cálculos finales obtenidos para el nivel de inferencia Segmento - GRS:

Tabla 29: Parámetros cálculo del tamaño de muestra

Parámetro	Var	Valor
Chi-Cuadrada	B	7,0
Precisión	b	2,0%
Nivel de Confianza	alpha	5,0%
Peor Caso	Pi	50,0%
Categorías	k	6,0

SegmentoXGRS	Total	N Teórico	N Logrado
S1-G3	560928	1927	2700
S1-G5	143506	1908	2700
S2-G1	87628	1892	1880
S2-G3	2605809	1932	2710
S2-G4	128511	1905	1910
S2-G5	1392151	1931	300
S2-G6	176678	1913	1910
S3-G3	755577	1929	2950
S3-G4	18669	1752	1800
S3-G5	511398	1926	3280
S3-G6	29077	1813	1750
S5-G1	49691	1861	1800
S5-G2	91914	1894	1920
S5-G3	275419	1920	2010
S5-G4	39432	1843	1800
S5-G5	221788	1917	2050
S5-G6	84400	1890	1880
S6*-G1	12670	1002	1050
S6*-G2	23794	1040	1050
S6*-G3	65350	1070	1050
S6*-G4	3872	849	900
S6*-G5	24219	1041	1050
S6*-G6	8740	967	1050

Como será ilustrado capítulo 7 el chequeo del tamaño teórico y el tamaño obtenido para la muestra definida fue realizado para cada nivel de inferencia y además se decidió utilizar la zona 6 para compensar la existencia de una menor cantidad de clientes en la población en las otras zonas extrayendo una cantidad mayor de clientes en ella a modo de cumplir con los requerimientos.

Para cada zona la extracción de clientes en cada estrato debe cumplir con lo ilustrado las siguientes tablas:

Tabla 30: Cantidad de Clientes a Extraer por Casilla por Zona (1 – 5)

Zona 1	Segmento 1			Segmento 2			Segmento 3			Segmento 4	Segmento 5		Segmento 6
	R3	R1-R2-R4-R5-R6	R0	R3	R2	R0	R3	R1-R2-R4-R5-R6	R0	R0	R4	R5-R6	R1
GRS 1		150		150	160			150			150	150	150
GRS 2					150			150			160	160	150
GRS 3	150	150	150	160	150	140	170	170	150	150	170	160	150
GRS 4				150	160		150	150			150	150	150
GRS 5	150	150	150	200	180	120	200	170	170	160	170	160	150
GRS 6				150	160		120	160			160	150	150

Fuete: Elaboración Propia

Tabla 31: Cantidad de Clientes a Extraer por Casilla por Zona 6

Zona 6	Segmento 1			Segmento 2			Segmento 3			Segmento 4	Segmento 5		Segmento 6
	R3	R1-R2-R4-R5-R6	R0	R3	R2	R0	R3	R1-R2-R4-R5-R6	R0	R0	R4	R5-R6	R1
GRS 1		150		160	170			150			150	150	300
GRS 2					150			150			160	160	300
GRS 3	150	150	150	160	160	140	170	170	160	150	200	160	300
GRS 4				180	180		150	150			150	150	150
GRS 5	150	150	150	200	180	120	200	200	180	200	200	200	300
GRS 6				180	180		150	200			170	160	300

Fuete: Elaboración Propia

Capítulo 6

Levantamiento del Panel

El levantamiento del panel de clientes consiste a la realización ordenada de una secuencia de pasos que permiten generar la estructura de información base del mismo y por ende posibilita tanto la extracción de la muestra definida en el capítulo anterior como la realización de las programaciones para generar la información necesaria para el cálculo de los indicadores de comportamiento requeridos.

6.1 Pasos del levantamiento del Panel

La primera tarea realizada correspondió a la creación del marco muestral, dicho proceso fue llevado a cabo recopilando la información demográfica y de segmentación de los clientes de la entidad.

Además debido a las variables de estratificación seleccionadas se tiene que el marco muestral debe incluir la variable de clasificación GRS descrita anteriormente y por lo tanto la información contenida en las tablas de contratos, la creación de la variable tenencia y GRS también deben ser consideras en este punto.

Una vez obtenido el marco muestral fueron realizados los procedimientos de limpieza, eliminación de datos y reemplazo de información faltante, discutidos en el capítulo 3, para generar con ello la clasificación de los elementos de la población en los 312 estratos definidos necesaria tanto para la extracción aleatoria de los clientes en cada uno de ellos como para tener el registro de los tamaños poblacionales de los estratos.

La cuantificación de los estratos fue realizada a través de la programación de un procedimiento sql que realiza el conteo de los clientes que cumplen las restricciones de clasificación para cada estrato y simultáneamente va guardando dichos valores en una tabla cuyas columnas corresponden a los segmentos refinados por la variable renta que se va llenado por filas hasta cubrir todos los niveles de los grs y zonas de la población (ver detalles en el Anexo M).

		Segmentos + Renta												
grs	zona	S1_R3	S1_Arenta	S1_R0	S2_R3	S2_Arenta	S2_R0	S3_R3	S3_Arenta	S3_R0	S4	S5_R4	S5_Arenta	S6
G1	1													
G2	1													
G3	1													
G4	1													
G5	1													
G6	1													

Figura 3: Estructura de la tabla resultado de procedimiento cuenta población

Una vez cuantificados los estratos y definidos los tamaños de extracción descritos en el capítulo anterior la siguiente tarea correspondió a la extracción de la muestra.

Cabe señalar que además de la extracción de los elementos muestrales, tal como se expuso en el marco teórico, una técnica de muestreo estratificada no proporcional requiere del cálculo de ponderadores para la correcta interpretación de resultados.

La extracción aleatoria fue realizada en cada estrato utilizando la opción de ordenamiento newid[34] disponible en SqlServer y la para cada elemento extraído se realizó el cálculo automático de los ponderadores o pesos definiendo para cada estrato i[35]:

$$peso_i = \frac{\%\, poblacional\ estrato\ i}{\%\, muestral\ estrato\ i} = \frac{\frac{cantidad\_clientes\_poblacion\_en\_el\_estrato\ i}{total\_poblacion}}{\frac{cantidad\_clientes\_a\_extraer\_en\_el\_estrato\ i}{total\_muestra}}$$

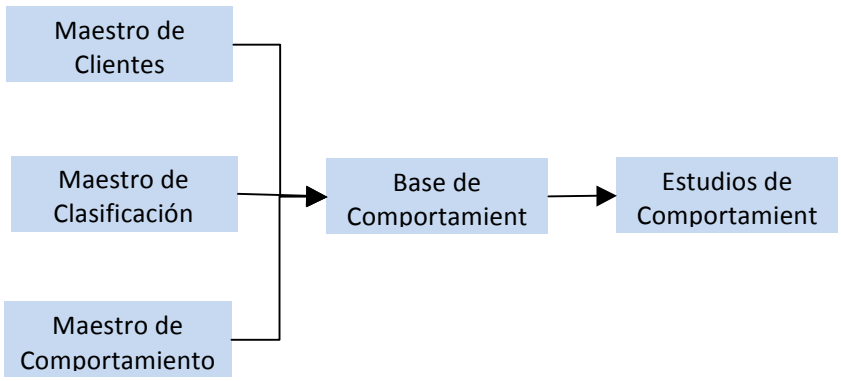
Cabe hacer notar que la consideración del peso de cada cliente es fundamental para la realización de inferencias poblacionales pues debido a la naturaleza del muestreo utilizado la probabilidad de selección de los individuos no es la misma y serán, por lo tanto, dichos pesos los que logren recuperar la representatividad de la población y finalmente permitan la proyección adecuada de los resultados.

Una vez definida la muestra del panel el paso final correspondió a la solicitud de los registros transaccionales, de saldo y deudas para los individuos seleccionados.

6.2 Estructura de la información contenida en el Panel

La estructura de información del panel de clientes puede ser resumida en el siguiente diagrama:

Figura 4: Estructura de información del panel de clientes.



Fuente: Elaboración Propia

El maestro de clientes corresponde a cada una de las tablas que registran información acerca de las características de los clientes y que posibilitan la definición

[34] Fuente: <www.manualespdf.es/manual-sql-server-200/>
[35] Ver anexo N.

del marco muestral. Entre ellas se tienen las tablas que incluyen las características demográficas, de segmentación, de contratos, grs, tenencia y de oficina asignada.

El maestro de clasificación está compuesto tanto por las tablas clasificatorias de la entidad como las creadas en el proyecto y permiten agrupar las acciones de los clientes: canales, productos y tipo de transacciones.

El maestro de comportamiento corresponde a las tablas que registran las acciones de los clientes a lo largo del tiempo entre ellas: transacciones, saldo, sbif.

Además dado a que el principal objetivo de la herramienta a construir es permitir la generación continua y dinámica de los indicadores de comportamiento necesarios para el modelo de calidad en que se enmarca la presente tesis la tarea final del levantamiento del panel fue la creación de un procedimiento almacenado posibilita el cómputo mensual del siguiente listado de variables a nivel de cada cliente.

1. Tenencia por producto
2. GRS
3. Cantidad de Transacciones
4. Monto de Transacciones
5. Saldo Inventario
6. Cantidad y Monto de Transacciones voluntarias en Máquinas
7. Cantidad y Monto de Transacciones voluntarias en Sucursales
8. Cantidad y Monto de Transacciones voluntarias en Internet
9. Monto Deuda Banco Estado y Total

Dicho procedimiento almacenado[36] se base en el input de los tres maestros descritos anteriormente y en aproximadamente 25 minutos da origen a la base de comportamiento.

Además del cálculo de las variables listadas anteriormente la base registra en una misma tabla las características demográficas de los clientes (segmento, renta, zona, renta, edad, etc.), sus correspondientes pesos y el mes en que el proceso fue ejecutado.

Con lo que se obtiene una única base que reúne toda la información mensual necesaria para la realización de estudios regulares o específicos del comportamiento de clientes evitando de esta forma recurrir a los maestros antes descritos cada vez que se necesite realizar alguna investigación y por lo tanto aportando la flexibilidad y eficiencia requerida para el cómputo de los indicadores necesarios para el modelo de calidad descrito en el primer capítulo.

[36] Ver detalles Anexo M.

Capítulo 7

Validación

Una muestra de calidad será aquella que asegura una cantidad suficiente de individuos para la obtención de resultados representativos de la población en estudio, debido a esto una vez levantado el panel fue necesario verificar tanto el número de individuos obtenido en cada nivel de inferencia como la realización del contraste de las distribuciones población muestra utilizando la mayor cantidad de información disponible para el universo completo.

7.1 Validación del Tamaño de Muestra y Ponderadores

Como se ha mencionado anteriormente el tamaño de muestra fue diseñado en función de cumplir con la cantidad de clientes requeridos para garantizar la validez de los estudios en cada una de las categorías de inferencia o dimensiones de observación del comportamiento diseñadas para el panel construido.

Además debido a la naturaleza no proporcional de la muestra se sabe que los pesos cumplen un rol fundamental para devolver la representatividad de la población completa.

Debido a ambos conceptos en el proceso de validación se realizaron 2 tareas principales:

(i) La comparación del tamaño de muestra poblacional (N teórico) y el inferido con el panel (N empírico).

(ii) El análisis de las distribuciones poblacionales y las obtenidas por el panel.

Permitiendo con esta última a modo verificar la calibración de los ponderadores utilizados en su cómputo.

A continuación se presentan los resultados obtenidos para las variables de estratificación:

Tabla 32: Validación Variable Segmento

	[illegible]		[illegible]		[illegible]		[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	717762	8,65	1928	6800	4305	8,65	0
[illegible]	4458184	53,49	1909	12810	28591	53,49	0
[illegible]	1265581	15,1	1901	12560	8011	15,1	0
[illegible]	909932	10,96	1929	1900	5457	10,96	0
[illegible]	762544	9,19	1929	11460	4574	9,18	0
[illegible]	139645	1,67	1907	6150	892	1,67	0

Fuente: Elaboración Propia

Tabla 33: Validación Variables GRS, Zona y Renta

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	2,04	[illegible]	[illegible]	[illegible]	2,04	0
[illegible]	[illegible]	2,08	[illegible]	[illegible]	[illegible]	2,08	0
[illegible]	[illegible]	62,09	[illegible]	[illegible]	[illegible]	62,09	0
[illegible]	[illegible]	2,3	[illegible]	[illegible]	[illegible]	2,3	0
[illegible]	[illegible]	27,95	[illegible]	[illegible]	[illegible]	27,95	0
[illegible]	[illegible]	8,6	[illegible]	[illegible]	[illegible]	8,6	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	12,15	[illegible]	[illegible]	[illegible]	12,15	0
[illegible]	[illegible]	10,75	[illegible]	[illegible]	[illegible]	10,75	0
[illegible]	[illegible]	10,94	[illegible]	[illegible]	[illegible]	10,94	0
[illegible]	[illegible]	11,53	[illegible]	[illegible]	[illegible]	11,53	0
[illegible]	[illegible]	14,52	[illegible]	[illegible]	[illegible]	14,52	0
[illegible]	[illegible]	40,11	[illegible]	[illegible]	[illegible]	40,11	0
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	20,45	[illegible]	[illegible]	[illegible]	20,45	0
[illegible]	[illegible]	1,83	[illegible]	[illegible]	[illegible]	1,82	0,01
[illegible]	[illegible]	14,18	[illegible]	[illegible]	[illegible]	14,15	0,03
[illegible]	[illegible]	53,42	[illegible]	[illegible]	[illegible]	53,42	0
[illegible]	[illegible]	7,28	[illegible]	[illegible]	[illegible]	7,32	-0,04
[illegible]	[illegible]	1,66	[illegible]	[illegible]	[illegible]	1,64	0,02
[illegible]	[illegible]	1,18	[illegible]	[illegible]	[illegible]	1,2	-0,02

Fuente: Elaboración Propia

De la tabla 32 y 33 se tiene que los requisitos del tamaño de muestra son cumplidos para las variables analizadas, además los diferenciales nulos obtenidos para las distribuciones de las variables grs, segmento, zona son los esperados pues ellas son variables de estratificación y por lo tanto son utilizadas para calibrar los ponderados del panel.

También se tiene que dicho valor es distinto de cero para la variable renta pues como se señaló en el capítulo 5 para la estratificación realizada se utilizaron criterios de agrupación en dicha variable.

A su vez también se verificaron los conceptos de tamaño de muestra y distribución para el resto de variables demográficas no consideras como variables de control en la extracción de la muestra y a continuación se exponen los resultados:

Tabla 34: Validación Variables Género y Antiguedad

Género	Total	%	N Teórico	N Empírico	Total	%	Diferencial
V1	4E+06	54	1933	21524	26729	53,7	0,33
V2	4E+06	43,7	1932	26957	21817	43,8	-0,16
NULL	189727	2,29	N/A	N/A	1224	2,46	-0,17
Antigüedad	**Total**	**%**	**N Teórico**	**N Empírico**	**Total**	**%**	**Diferencial**
A1	1E+06	13,9	1930	7473	6625	13,3	0,63
A2	1E+06	14,9	1930	7052	7309	14,7	0,16
A3	913224	11	1929	6013	5734	11,5	-0,51
A4	5E+06	57,3	1933	27759	28583	57,4	-0,12
NULL	240410	2,9	N/A	N/A	1518	3,05	-0,15

Fuente: Elaboración Propia

Tabla 35: Validación Variables Edad, Nivel Educacional y Región

Edad	Total	%	N Teórico	N Empírico	Total	%	Diferencial
E1	922315	11,1	1929	1866	5504	11,1	0,06
E2	1E+06	15,6	1931	11211	7702	15,5	0,11
E3	2E+06	19,7	1931	8164	9790	19,7	0,03
E4	2E+06	27	1932	13254	13467	27,1	-0,03
E5	1E+06	16,1	1931	8223	8013	16,1	0
E6	425638	5,13	1925	3874	2612	5,25	-0,12
E7	258808	3,12	1919	1937	1492	3	0,12
NULL	183651	2,21	N/A	N/A	1192	2,4	-0,19
N_Educac.	**Total**	**%**	**N Teórico**	**N Empírico**	**Total**	**%**	**Diferencial**
N1	1E+06	17,8	1931	3863	8839	17,8	0,01
N2	3E+06	30,3	1932	14338	15400	30,9	-0,62
N3	194997	2,35	1914	1875	1148	2,31	0,04
N4	335207	4,04	1922	7064	2039	4,1	-0,06
N5	494870	5,96	1926	6275	2860	5,75	0,21
N6	10339	0,12	N/A	N/A	50	0,1	0,02
N7	95415	1,15	N/A	N/A	571	1,15	0
N8	122389	1,47	N/A	N/A	737	1,48	-0,01
NULL	3E+06	36,8	1932	14907	18127	36,4	0,39
Región	**Total**	**%**	**N Teórico**	**N Empírico**	**Total**	**%**	**Diferencial**
I	140388	1,69	1907	1087	755	1,52	0,17
II	263692	3,18	1919	2188	1603	3,22	-0,04
III	147951	1,78	1909	1193	931	1,87	-0,09
IV	342147	4,12	1923	2757	2053	4,13	-0,01
V	892404	10,8	1929	8090	5353	10,8	0
VI	432842	5,22	1925	4044	2693	5,41	-0,19
VII	474641	5,72	1926	4046	2750	5,53	0,19
VIII	956601	11,5	1930	8090	5738	11,5	0
IX	464643	5,6	1925	2740	2785	5,6	0
X	389028	4,69	1924	2752	2415	4,85	-0,16
XI	61176	0,74	1874	496	316	0,64	0,1
XII	101929	1,23	1897	833	638	1,28	-0,05
XIII	3E+06	40,1	1932	9250	19962	40,1	0
XIV	188280	2,27	1914	1269	1074	2,16	0,11
XV	113866	1,37	1901	865	704	1,41	-0,04

Fuente: Elaboración Propia

De las tablas 34 y 35 se tiene que para el resto de las variables demográficas las distribuciones de la muestra son acertadas en cuanto a su similitud con la población completa así también se ve el cumplimiento de las restricciones del tamaño de muestral.

Cabe señalar que la variable región no ha sido planteada como requisito de dimensión de observación del comportamiento, en su reemplazo se usa la variable región, de ser necesaria en el futuro deberán incluirse la cantidad de individuos faltantes para asegurar la calidad estadística de las inferencias en este nivel o también es posible utilizar la misma muestra relajando el error máximo aceptado para aquellos casos en que se tiene un menor tamaño muestral.

Los análisis de validación también fueron realizados para los niveles de inferencia con cruces presentados en la sección 5.9 pues de hecho son estas las

restricciones que imponen mayores restricciones a la muestra extraída al exigir una cantidad de clientes determinada para niveles más específicos tal como se muestra en la tabla 29 de dicha sección.

A continuación se presentan como ejemplo las validaciones para los cruces Segmento-Renta y Grs-Zona[37]:

Tabla 36: Validación Segmento-Renta

SegmentoXRenta	Total	%	N Teórico	N Empírico	Total	%	Diferencial
S1-RO	78862	0,95	1887	1800	473	0,95	0
S1-R3	506083	6,1	1926	1800	3036	6,1	0
S1-R_Agrupado	132817	1,6	1906	2700	796	1,6	0
S2-R0	528779	6,37	1926	1960	3172	6,37	0
S2-R3	3E+06	35,4	1932	4930	17629	35,42	0
S2-R2	965169	11,6	1930	5420	5789	11,63	0
S3-R0	179199	2,16	1913	1940	1074	2,16	0
S3-R3	987342	11,9	1930	3870	5922	11,9	0
S3-R_Agrupado	169040	2,04	1912	5770	1014	2,04	0
S4-R0	909832	11	1929	1900	5457	10,96	0
S5-R4	540051	6,51	1927	5830	3239	6,51	0
S5-R5 y R6	222593	2,68	1917	5630	1335	2,68	0
S6-R1	138645	1,67	1907	6150	832	1,67	0

Fuente: Elaboración Propia

Tabla 37: Validación GRS-Zona

GRSXZona	Total	%	N Teórico	N Empírico	Total	%	Diferencial
G1 y G2-Z1	48448	0,58	1859	1830	291	0,58	0
G1 y G2-Z2	37418	0,45	1838	1830	224	0,45	0
G1 y G2-Z3	30811	0,37	1819	1830	185	0,37	0
G1 y G2-Z4	35352	0,43	1833	1830	212	0,43	0
G1 y G2-Z5	51829	0,62	1864	1830	311	0,62	0
G1 y G2-Z6	138414	1,67	1907	2150	830	1,67	0
G3 y G4-Z1	652672	7,87	1928	3080	3915	7,87	0
G3 y G4-Z2	589947	7,11	1927	3080	3539	7,11	0
G3 y G4-Z3	646552	7,79	1928	3080	3878	7,79	0
G3 y G4-Z4	664435	8,01	1928	3080	3985	8,01	0
G3 y G4-Z5	829566	10	1929	3080	4976	10	0
G3 y G4-Z6	2E+06	23,6	1932	3330	11721	23,55	0
G5 y G6-Z1	306924	3,7	1921	3180	1841	3,7	0
G5 y G6-Z2	265039	3,19	1919	3180	1590	3,19	0
G5 y G6-Z3	230120	2,77	1917	3180	1380	2,77	0
G5 y G6-Z4	256814	3,1	1919	3180	1541	3,1	0
G5 y G6-Z5	323661	3,9	1922	3180	1941	3,9	0
G5 y G6-Z6	1E+06	14,9	1930	3770	7411	14,89	0
G5-Z1	277280	3,34	1815	2130	1663	3,34	0
G5-Z2	237798	2,87	1805	2130	1426	2,87	0
G5-Z3	202049	2,44	1809	2130	1212	2,44	0
G5-Z4	230961	2,78	1799	2130	1385	2,78	0
G5-Z5	289543	3,49	1830	2130	1737	3,49	0
G5-Z6	1E+06	13	1909	2430	6487	13,03	0

Fuente: Elaboración Propia

[37] Para resto de cruces ver anexo O.

7.2 Validación de Variable Tenencia

Un tercer aspecto abordado en la validación al contraste de la data poblacional y del panel para la variable tenencia cuyos resultados se muestran en la siguiente tabla:

Tabla 38: Validación Inferencia Nivel Producto

Productos	% Población	N Teórico	N Empírico	% Panel	Diferencial
P1	3,85	1933	11759	3,94	-0,09
P2	4,25	1933	5419	4,27	-0,02
P3	55,28	1933	45527	55,2	0,08
P4	1,65	1933	7368	1,65	0
P5	0,49	1933	3503	0,5	-0,01
P6	1,08	1933	6146	1,07	0,01
P7	12,12	1933	13595	11,93	0,19
P8	3,76	1933	8933	3,82	-0,06
P9	1,08	1933	2207	1,11	-0,03
P10	2,48	1933	8616	2,47	0,01
P11	0,6	1933	3125	0,6	0
P12*	0,27	1080	1174	0,27	0
P13	4,65	1933	15905	4,75	-0,1
P14	6,68	1933	18111	6,69	-0,01
P15	0,42	1933	2160	0,42	0
P16	0,92	1933	4808	0,92	0
P17*	0,4	1420	1484	0,39	0,01

Fuente: Elaboración Propia

De la tabla 38 de verifica que la variable tenencia por producto posee la cantidad adecuada de individuos y que los diferenciales del error obtenido están bien bajo a lo esperado, atribuible al control realizado por la variable GRS.

A su vez de los resultados expuestos en la tabla 32 y 33 se tiene que los tamaños de muestra obtenidos para los niveles de inferencia superiores (segmento, grs, zona) por lo tanto se decide indagar los resultados obtenidos en la estimación de la variable tenencia para cada uno de ellos.

A continuación se presenta las distribuciones calculadas con la data poblacional completa en contraste con las estimadas a partir de los datos del panel de clientes verificando previamente el tamaño de muestra requerido (ejemplo tabla 39) para la estimación de proporciones para una variable a 17 niveles para cada dimensión de observación requerida.

A continuación se ilustran los resultados obtenidos para la dimensión segmento, para resto de niveles ver anexo O.

Tabla 39: N necesario para estimación Variable Tenencia por Producto a Nivel de Segmento

Variable 17 categorias	N Teórico*	N Empírico
Segmento 1	5485	6300
Segmento 2	5520	12310
Segmento 3	5504	11580
Segmento 4	5494	1900
Segmento 5	5487	11460
Segmento 6	5315	6150

*Error al 2%, 95% de confianza

Fuente: Elaboración Propia

Tabla 40: Estimación de Proporciones Variable Tenencia a Nivel de Segmento

S1	% Población	% Panel	Error	S2	% Población	% Panel	Error	S3	% Población	% Panel	Error
P1	3,61	3,67	-0,06	P1	3,16	3,39	-0,23	P1	2,35	2,39	-0,04
P2	5,09	4,92	0,17	P2	5,33	5,4	-0,07	P2	2,59	2,5	0,09
P3	73,55	73,3	0,25	P3	56,55	56,33	0,22	P3	57,93	57,5	0,43
P4	0	0	0	P4	1,18	1,2	-0,02	P4	1,5	1,51	-0,01
P5	0	0	0	P5	0,2	0,21	-0,01	P5	1,05	1,06	-0,01
P6	1,03	1,04	-0,01	P6	0,3	0,28	0,02	P6	0,22	0,21	0,01
P7	8,36	8,05	0,31	P7	12,75	12,47	0,28	P7	19,08	19,33	-0,25
P8	2,46	2,76	-0,3	P8	3,39	3,4	-0,01	P8	5,51	5,76	-0,25
P9	3,3	3,47	-0,17	P9	0,96	1,06	-0,1	P9	0,27	0,24	0,03
P10	0	0	0	P10	2,77	2,74	0,03	P10	0,37	0,36	0,01
P11	0,15	0,17	-0,02	P11	0,16	0,15	0,01	P11	0,2	0,2	0
P12	0,02	0,02	0	P12	0,21	0,22	-0,01	P12	0,29	0,31	-0,02
P13	0,29	0,27	0,02	P13	4,58	4,67	-0,09	P13	3,01	3,34	-0,33
P14	1,68	1,91	-0,23	P14	7,11	7,16	-0,05	P14	4,74	4,39	0,35
P15	0,14	0,13	0,01	P15	0,16	0,16	0	P15	0,11	0,11	0
P16	0,28	0,28	0	P16	0,56	0,56	0	P16	0,66	0,67	-0,01
P17	0,05	0,03	0,02	P17	0,62	0,61	0,01	P17	0,1	0,11	-0,01
S4	**% Población**	**% Panel**	**Error**	**S5**	**% Población**	**% Panel**	**Error**	**S6**	**% Población**	**% Panel**	**Error**
P1	0,04	0,01	0,03	P1	8,14	7,9	0,24	**P1**	10,02	9,95	0,07
P2	0,11	0,12	-0,01	P2	3,49	3,48	0,01	**P2**	2,29	2,17	0,12
P3	97,07	97,14	-0,07	P3	29,52	30,02	-0,5	**P3**	28,13	28,37	-0,24
P4	0	0	0	P4	4,22	4,2	0,02	**P4**	4,75	4,86	-0,11
P5	0	0	0	P5	1,21	1,25	-0,04	**P5**	1,44	1,45	-0,01
P6		0	0	P6	4,01	3,97	0,04	**P6**	7,33	7,45	-0,12
P7	2,65	2,63	0,02	P7	9,57	9,33	0,24	**P7**	7,34	7,31	0,03
P8	0,09	0,1	-0,01	P8	5,46	5,5	-0,04	**P8**	3,75	3,56	0,19
P9	0,04	0	0,04	P9	1,53	1,39	0,14	**P9**	2,04	2,22	-0,18
P10	0	0	0	P10	5,03	5,05	-0,02	**P10**	4,54	4,34	0,2
P11	0	0	0	P11	2,32	2,29	0,03	**P11**	4,22	4,33	-0,11
P12	0	0	0	P12	0,65	0,6	0,05	**P12**	0,39	0,39	0
P13	0	0	0	P13	9,45	9,52	-0,07	**P13**	7,85	7,76	0,09
P14	0	0	0	P14	11,04	11,1	-0,06	**P14**	9,87	9,81	0,06
P15	0	0	0	P15	1,52	1,54	-0,02	**P15**	2,58	2,55	0,03
P16	0	0	0	P16	2,63	2,65	-0,02	**P16**	3,41	3,44	-0,03
P17	0	0	0	P17	0,22	0,22	0	**P17**	0,06	0,05	0,01

Fuente: Elaboración Propia

Los resultados del análisis anterior hablan también de la robustez de los tamaños de muestra definidos para cada estrato pues permiten tanto disminuir el nivel de error máximo establecido como aumentar la cantidad de categorías a estimar las proporciones de la población.

En efecto la mayor sensibilidad del tamaño de muestra es para el error máximo aceptado y no para el cambio marginal en el nº de categorías a estimar en proporción y permitiendo flexibilizar el supuesto de estimación de proporciones para 6 niveles expuesto en el capítulo 5 de acuerdo a lo ilustrado en la siguiente tabla[38]:

Tabla 41: Tamaños de muestra en función del error y categorias

Nº de categorías	e=1%	e=2%	e=3%	e=4%	e=5%
5	16587	4147	1843	1037	663
6	17401	4350	1933	1088	696
7	18092	4523	2010	1131	724
8	18692	4673	2077	1168	748
9	19223	4806	2136	1201	769
10	19699	4925	2189	1231	788
11	20130	5032	2237	1258	805
12	20524	5131	2280	1283	821
13	20888	5222	2321	1305	836
14	21225	5306	2358	1327	849
15	21538	5385	2393	1346	862
16	21833	5458	2426	1365	873
17	22109	5527	2457	1382	884
18	22370	5592	2486	1398	895
19	22617	5654	2513	1414	905
20	22851	5713	2539	1428	914

Fuente: Elaboración propia

7.3 Validación de Variables Cantidad de Transacciones y Saldo Promedio

Además de las verificaciones realizadas para la variable tenencia descritas en el punto anterior se decidió estudiar el comportamiento de las variables cantidad de transacciones mensuales y saldo promedio diario mensual a través de los datos obtenidos en las muestras aleatorias pilotos por considerarse estas las más representativas a nivel de la población completa[39] obteniendo:

(i) Análisis de la distribución de la cantidad transacciones mensuales

a) Cantidad de transacciones por segmento

Tabla 42: Validación Variable Cantidad de Transacciones a Nivel Segmento

PANEL

SEGMENTO	200801	200802	200803	200804	200805	200806	200807	200808	200809
S1	5,27	5,40	4,96	5,01	4,28	4,22	4,41	4,41	4,30
S2	43,22	43,64	42,55	44,32	42,99	42,47	42,33	43,09	43,73
S3	11,81	11,69	12,15	12,97	12,77	12,91	12,91	13,55	13,36
S4	1,45	1,72	1,47	1,82	1,76	1,57	1,85	1,57	1,48
S5	30,19	29,52	30,57	28,11	30,04	30,25	30,23	29,24	29,03
S6	8,06	8,03	8,29	7,77	8,15	8,57	8,27	8,14	8,09

Fuente: Elaboración propia

[38] Cálculos realizados sin corrección de finitud

[39] No para niveles de inferencia específicos pues la muestra aleatoria no posee la cantidad de datos adecuada para asegurar la calidad estadística de los resultados en ellos.

MUESTRA ALEATORIA									
SEGMENTO	200801	200802	200803	200804	200805	200806	200807	200808	200809
S1	6,12	6,62	5,93	5,94	4,85	4,79	4,67	4,74	4,63
S2	43,83	44,94	44,62	45,95	43,43	43,56	45,06	44,57	44,68
S3	10,69	11,44	11,57	13,07	13,61	12,64	12,77	12,59	13,40
S4	1,67	1,91	1,65	1,91	1,81	1,53	1,61	1,70	1,61
S5	29,49	27,67	28,29	26,05	29,01	29,28	28,06	28,71	28,59
S6	8,20	7,42	7,94	7,08	7,29	8,20	7,83	7,69	7,09

Fuente: Elaboración propia

b) Cantidad de transacciones por GRS

Tabla 43: Validación Variable Cantidad de Transacciones a Nivel GRS

PANEL									
GRS	200801	200802	200803	200804	200805	200806	200807	200808	200809
GRS1	12,64	11,99	12,31	11,43	11,90	12,35	12,33	11,05	11,82
GRS2	21,53	21,26	22,48	20,50	22,26	23,44	22,63	20,61	22,30
GRS3	13,12	13,51	12,55	14,21	11,98	11,64	11,93	12,04	11,05
GRS4	1,57	1,59	1,42	1,50	1,29	1,25	1,24	1,06	1,27
GRS5	39,97	40,48	40,36	41,76	42,27	41,10	41,29	44,60	42,96
GRS6	11,17	11,17	10,89	10,60	10,31	10,22	10,59	10,64	10,59
MUESTRA ALEATORIA									
GRS	200801	200802	200803	200804	200805	200806	200807	200808	200809
GRS1	12,53	11,44	12,02	11,41	12,22	13,14	12,24	11,37	12,49
GRS2	22,05	21,19	21,96	19,29	21,50	22,99	22,02	20,39	22,12
GRS3	13,55	13,92	13,06	14,00	11,92	11,11	11,56	11,35	11,26
GRS4	1,52	1,39	1,34	1,44	1,16	1,12	1,18	0,94	1,26
GRS5	39,98	41,31	40,58	42,56	42,44	41,10	42,32	45,23	42,80
GRS6	10,38	10,74	11,04	11,31	10,76	10,53	10,68	10,72	10,07

Fuente: Elaboración propia

c) Cantidad de transacciones por Canal (no agregados)

Tabla 44: Validación Variable Cantidad de Transacciones a Nivel Canal

PANEL									
Canal	200801	200802	200803	200804	200805	200806	200807	200808	200809
1	22,35	23,32	22,97	23,39	20,58	18,96	20,62	21,26	20,33
2	0,93	1,09	1,18	1,17	0,86	0,72	0,79	0,78	0,88
3	7,30	6,83	6,51	6,65	5,50	5,09	5,44	5,00	4,62
4	0,73	0,79	0,77	0,74	0,66	0,75	1,03	0,96	1,20
5	9,60	11,44	11,06	10,91	10,34	9,19	10,00	11,52	10,10
6	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00
7	2,51	2,49	2,25	2,20	2,05	1,76	1,97	1,57	1,73
8	17,74	13,93	17,90	16,22	27,78	31,85	26,28	25,22	26,08
9	0,01	0,01	0,02	0,01	0,01	0,01	0,00	0,01	0,01
10	17,15	17,94	17,29	17,78	15,62	14,55	15,70	17,39	16,96
11	2,09	2,30	1,99	2,21	1,76	1,74	1,99	2,03	1,89
12	1,80	1,95	2,30	1,65	1,56	1,53	1,65	0,00	1,80
13	17,77	17,91	15,75	17,06	13,26	13,85	14,52	14,27	14,40

MUESTRA ALEATORIA									
Canal	200801	200802	200803	200804	200805	200806	200807	200808	200809
1	22,82	23,44	23,48	23,42	20,52	19,43	20,35	21,13	20,43
2	1,32	1,25	1,05	1,35	0,89	0,76	0,73	0,82	0,75
3	6,78	7,39	6,36	6,95	5,37	4,98	4,99	4,89	4,36
4	0,57	0,61	0,61	0,90	0,88	0,76	1,70	2,14	2,62
5	8,86	10,96	10,46	10,75	10,09	8,77	10,44	11,20	10,02
6	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00
7	2,28	2,31	2,22	2,09	2,02	1,60	1,89	1,53	1,71
8	17,67	13,93	17,64	15,06	27,00	31,39	26,04	25,07	25,07
9	0,01	0,02	0,01	0,01	0,02	0,02	0,01	0,00	0,01
10	17,83	18,50	17,50	18,22	16,19	15,11	15,91	17,58	17,34
11	1,85	1,99	1,85	1,97	1,65	1,64	1,73	1,72	1,72
12	1,62	1,72	2,11	1,43	1,34	1,41	1,42	0,00	1,57
13	18,39	17,88	16,71	17,84	14,02	14,13	14,78	13,91	14,36

Fuente: Elaboración propia

De las tablas 42, 43 y 44 tiene que tanto con la muestra piloto de 25000 clientes como el panel (considerando los pesos individuales del aporte realizado por cada cliente) obtienen valores similares.

Además para validar la información transaccional contenida en el panel también fue estudiada la distribución de cantidad y monto de transacciones a lo largo del tiempo realizando el mismo análisis de distribuciones ilustrado anteriormente y además realizó la prueba Kolmogorov-Smirnov para dos muestras independientes[40] la cual permitió concluir que tanto los datos de monto y cantidad de la muestra aleatoria como el panel de clientes provienen de la misma población (ver detalles en anexo P).

(ii) Análisis de la distribución de saldos

Los saldos fueron estudiados a nivel de su distribución por segmento para los cuatro productos acreedores más relevantes utilizando un año de de datos y utilizando la variable media de saldo promedio definida por:

$$\frac{suma\ saldo\ prom\ prod\ j\ para\ segmento\ k\ en\ periodo\ en\ estudio}{n^{o}\ meses * clientes\ que\ poseen\ producto\ j\ en\ segmento\ k}$$

Obteniendo las siguientes distribuciones:

Tabla 45: Distribución Saldos por Producto Nivel Segmento

Segmentos		S1	S2	S3	S4	S5	S6
P7	Aleatoria	14,4	11,3	6,6	1	27,7	39
	Panel	13,5	10,1	6,6	1,7	25,7	42,4
P3	Aleatoria	37,76	12,95	4,17	4,09	17,21	23,92
	Panel	35,86	11,11	4,39	4,47	18,19	25,98
P6	Aleatoria	45,71	10,23	3,76	0	14,11	26,19
	Panel	47,66	8,61	3,29	0	12,84	27,6
P8	Aleatoria	38,51	8,42	4,16	0,71	17,52	30,68
	Panel	36,33	7,42	3,96	0,97	18,27	33,05

Fuente: Elaboración propia

[40] Presentado en el Capítulo 2 - Sección 2.8.6.

Además se realizó un análisis del comportamiento de saldos promedios mensual por producto en el tiempo realizando para ello el contraste de K-S descrito anteriormente utilizando la variable media de saldo para cada uno de los 22 meses de datos disponibles (200612-200809).

$$media\ \ de\,saldo = \frac{suma\,saldo\ prom\ prod\ \ j\ \ en\ tiempo\ t}{clientes\ \ que\ poseen\ \ producto\ \ j\,en\ tiempo\,t}$$

A continuación se presentan los resultados obtenidos:

Tabla 46: Resultados Análisis K-S para Distribución de Saldos en el Tiempo

		P7	P3	P6	P8
Diferencias más extremas	Absoluta	0,136	0,318	0,318	0,316
	Positiva	0,136	0,318	0,318	0,316
	Negativa	-0,136	0	-0,45	-2,11
Z de Kolmogorov-Smirnov		0,452	1,055	1,055	0,973
Sig. asintót. (bilateral)		0,987	0,215	0,215	0,3

Fuente: Elaboración propia

De dónde se obtiene que para todos los productos estudiados la probabilidad asintótica es mayor que 0,05 y por lo tanto se concluye que ambas muestras provienen de la misma población.

Capítulo 8

Mantención del Panel de Clientes

El proceso de mantención del panel de clientes corresponde a una serie de procedimientos necesarios tanto para lograr mantener la estructura de información del panel de clientes actualizada como para asegurar la representatividad de la muestra a lo largo del tiempo.

En general será importante la determinación de los patrones que generen brechas de representatividad de la población real [8] [23] y por lo tanto se tendrá que los principales gatilladores de mantención o ajustes serán los cambios en las variables individuales de la población objetivo y de los clientes del panel en el tiempo.

Estos cambios pueden ser efecto de acciones de la misma institución (como por ejemplo: la realización de campañas masivas para contratos en algún producto) o por variaciones naturales en los individuos del panel y la población (cambios en tenencia por producto, edad, renta, etc.).

De lo anterior se tiene que además de realizar una actualización periódica a las tablas maestros descritas anteriormente se deberán ingresar clientes nuevos a la muestra a modo de reemplazo de aquellos clientes que ya no lo son permitiendo con ello estudiar tanto las características de los nuevos clientes como de aquellos que ya no lo son.

A su vez, junto a los procesos de actualización de datos y renovación de clientes el mantenimiento del panel implicará necesariamente el ajuste de los pesos de cada unidad muestral debido a que estos son pieza fundamental en el proceso de inferencia estadística y dado a que estos son definidos en función de las características del individuo en la muestra y en la población en el instante de su cómputo deberá mantenerse el registro de los valores antiguos como los nuevos para permitir la realizar de estudios en periodos de cambios estructurales solapados.

El detalle y la estructura de la operación de mantención del observatorio de clientes se expone a continuación:

(i) Proceso de Actualización:

Cada tres meses se deberán actualizar las tablas relativas a las características demográficas y a las acciones de los clientes que conforman el panel.

Esto implica que al menos cada 3 meses será necesario conseguir para los clientes del panel los datos correspondientes a:

- Información de Transacciones

- Información de Saldos Promedio Diario Mensual
- Información de la SBIF
- Información de Contratos
- Información de Demográfica y de Segmentación

Una vez conseguida dichos registros deberán ejecutarse los procedimientos almacenados de tenencia (para ser consistentes con toda la estructura de datos que compone el panel de clientes la fecha de dicho procedimiento[41] deberá ser la más actual en la que sea posible conseguir el resto de datos) y el proceso de asignación de clientes a los GRS[42] para definir con estos los nuevos valores de dichas variables para los clientes del panel.

Además de modificarse alguna definición de la variable GRS o agregar un nuevo producto los procesos automáticos deberán ser corregidos de manera ad-hoc.

A su vez en conjunto con la información demográfica y de segmentación conseguida deberá realizarse la cuantificación de los estratos en la muestra[43] para corregir las posibles diferencias existentes entre la estructura de la muestra inicial y la actual a través del re-cómputo de pesos.

Además después cada actualización, para cada uno de los meses incorporado, deberá ejecutarse el procedimiento Creación Base Comportamiento descrito en el capítulo 6.

(ii) Proceso de Renovación:

Cada 12 meses deberá obtenerse la información poblacional del maestro de clientes y junto con ello se deberá realizar el reemplazo de la variable región para aquellos clientes sin dato tal como fue ha sido expuesto en el tercer capítulo[44].

Con esa información reunida deberán ejecutarse los procesos de tenencia, asignación de clientes a los GRS y cuantificación de los estratos para la población completa.

Con dichos datos renovados se deberá comprender la nueva distribución de la población pues de manera natural existirán cambios como por ejemplo: clientes con nuevos contratos que cambiaron de grs, clientes que se mudan y cambian región, clientes que de manera natural han dejado de ser parte de la población de la entidad, etc.

Además será posible comprender los cambios en la población objetivo y junto al proceso de actualización descrito en el punto anterior podrán definirse las brechas entre la población objetivo y del panel permitiendo incorporar/eliminar clientes en los estratos dónde sea necesario a modo de velar por el mantenimiento de los tamaños de muestra descritos en el capítulo 5.

El proceso de renovación, dependiendo de los resultados obtenidos, permitirá incorporar clientes que no pertenecían a la población objetivo o que han modificado su tenencia de productos desde la última renovación.

[41] Detalles Anexo F
[42] Detalles Anexo G
[43] Detalles Anexo M
[44] Detalles Anexo D

Esto es, si la diferencia en las cuantificaciones poblacionales para un estrato es significativa entre un periodo y otro se deberá extraer una muestra aleatoria de dichos clientes y, a su vez, eliminar del observatorio la misma cantidad de clientes que han ingresado.

Por ejemplo, si entraron 2000 clientes al S3, Zona 1, GRS3 y R0 se eliminarán (si no ha disminuido de forma natural su tamaño) 30 clientes de dicho estrato y se ingresará una muestra aleatoria de clientes nuevos en igual magnitud (o la necesaria para compensar una posible baja natural en dicho estrato).

Además con las nuevas cuantificaciones de la población y la muestra deberá realizarse la reasignación de pesos a los elementos muestrales productos de los cambios en las cantidades de clientes en cada estrato.

(iii) Eliminación de clientes Inactivos

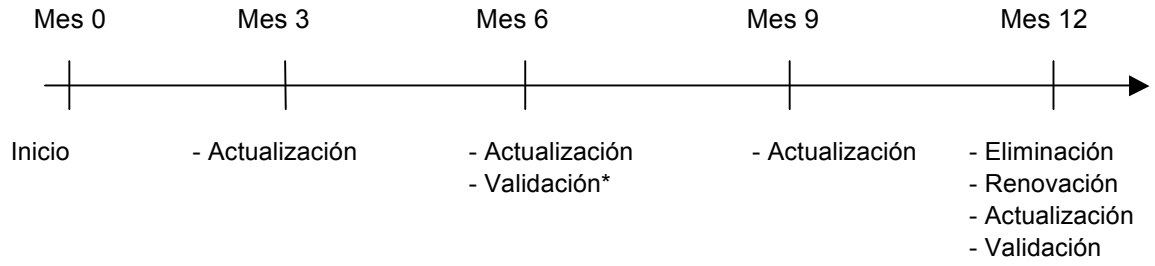
Anualmente, se eliminarán del observatorio aquellos clientes que ya no tienen productos vigentes con el banco a la fecha de eliminación y clientes que presenten periodos de inactividad prolongados en sus cuentas[45].

En su reemplazo, ingresarán al observatorio de manera aleatoria y equivalente la misma cantidad de clientes que han sido eliminados en cada estrato.

(iv) Validación del maestro de clasificación

Anualmente, o antes* si se tiene el conocimiento de acciones de gestión de la entidad que impliquen cambios en este contexto, deberá efectuarse tanto la mantención como la revisión de las tablas de: canales, productos y clasificación de transacciones a modo de incluir posibles nuevos valores en dichos registros y modificar los procedimientos creados que se relacionan con ellas.

Estos cuatro procesos de mantención deberán ser constantes a lo largo del tiempo generando la siguiente estructura, que se replica de igual manera cada año:



[45] Ver Anexo Q

Capítulo 9

Conclusiones y Comentarios Finales

9.1 Conclusiones del trabajo realizado

Como ha sido expuesto en el presente informe el objetivo principal de la realización de esta tesis corresponde al levantamiento de una herramienta que sea de fácil de utilizar y que constituya una fuente de información continua y eficiente para realizar estudios de tendencia del comportamiento de clientes.

A lo largo del informe presentado es posible notar que todas las decisiones tomadas para el cumplimiento de dichos objetivos han sido realizadas intentando considerar con máxima rigurosidad los métodos estadísticos disponibles con el fin de validar sus resultados y permitir una correcta proyección al total de la población de clientes.

Del análisis del negocio es posible concluir que BancoEstado no cuenta con una definición formal de cliente, a pesar de lo importante que es, y por lo tanto fue necesario crear una definición propia a modo de mantener la consistencia estadística antes descrita.

De los estudios de las variables de interés es posible determinar la relevancia que tienen los análisis exploratorios en cualquier estudio que involucre el trabajo con datos pues, por ejemplo, de no haber considerado los altos índices de variabilidad definidos en esta etapa o la necesidad de estimación de más de dos proporciones se podrían haber obtenido resultados completamente inválidos.

De las variables de estratificación y el tamaño de muestra definido se concluye que aportan una alta robustez a la información contenida en el panel de clientes pues además de permitir la obtención de una cantidad suficiente de clientes en 312 estratos poblacionales posibilitan tanto la disminución del error como el aumento de categorías a estimar.

Del levantamiento del panel se tiene que en todo momento se intentó realizar la automatización de la mayor cantidad de procedimientos involucrados permitiendo de con ello contar con procesos repetibles en el tiempo para la futura mantención y réplica de la herramienta construida.

Finalmente se tiene que el panel levantado cumple los todos objetivos planteados en el proyecto de investigación pues logra entregar la información del

comportamiento de clientes para 176 categorías de comportamiento con errores de proyección inferiores al 1%.

9.2 Discusiones Finales y Trabajos Futuros

Uno de los primeros aspectos a considerar en la realización de trabajos futuros corresponde a profundizar la sospecha de existencia de dos poblaciones expuesta en el capítulo 3 y junto con ello se propone realizar los análisis metodológicos propuestos para definir tanto la necesidad de la extracción de dos muestras independientes para cada una de ellas con la posibilidad de disminuir la variabilidad de cada población a modo poder asegurar la calidad estadística necesaria para la estimación puntual de las variables monto, cantidad, saldo y tenencia.

Otra línea de investigación es el estudio exhaustivo y corrección de la calidad de la data del Data Warehouse de la entidad esto pues además de los aspectos discutidos para la variable región y a la falta de una definición formal de cliente en la realización de la tesis se puedo comprobar que para una muestra aleatoria de 25000 clientes, cuyos registros señalan la tenencia de al menos 1 producto vigente a la fecha de muestreo, la data transaccional obtenida para 2 años de transacciones sólo corresponde en promedio al 70% de los clientes de la muestra.

Si bien es cierto que dicho valor aumenta un 15% con las variables de control estipuladas para la extracción de la muestra de todas formas es sospechable la actualización de los datos pues si un clientes posee al menos 1 producto es seguro que se generarán transacciones (del tipo involuntarias por ejemplo) en una ventana temporal de 2 años.

Del levantamiento de la base de comportamiento es posible realizar estudios de inactividad de los clientes a lo largo del tiempo y con ellos definir la cantidad de clientes que efectivamente aportan información transaccional permitiendo con ellos estimar una tasa de no respuesta y ajustar con ellos los requerimientos en el tamaño de muestra exigido para cada estrato.

Del proceso de mantención planteado es posible concluir que una investigación formal de los patrones de la población y del panel en el tiempo es sin duda una actividad complementaria a la presente tesis realizada, esto permitirá definir tanto el período óptimo para la realización de los pasos de mantención estipulados como por la definición y programación de *triggers* de inactividad, u otros, que anuncien la necesidad de eliminar y reemplazar clientes del panel de manera automática.

Finalmente se tiene que el aumento de potencia en los recursos informáticos que se destinen a la herramienta creada facilitará la incorporación de nuevas dimensiones de observación lo que sumado a las mejoras propuestas y a el mantenimiento de los registros a lo largo del tiempo permitirán el desarrollo paulatino de una base de datos cada vez más rica en cuanto a la cantidad de información recolectada que posibilita la realización de estudios econométricos de panel complejos que sin duda aportarán al nivel de conocimiento de los clientes.

Referencias

1. AAKER, D., KUMAR, V. y DAY, G. 2003. Marketing research. 8a Ed. Wiley.
2. ARELLANO, SOLEDAD y BENAVENTE, JOSÉ MIGUEL. 2006. Evidencia preliminar de sustitución entre telefonía fija y móvil en Chile
3. BENDENZÚ, LUIS et al. [s.a.] Análisis de la atrición de la muestra en la encuesta panel CASEN. <http://www.osuah.cl/encuestapanelcasen>
4. BENDENZÚ, LUIS et al. [s.a.] La encuesta panel CASEN: metodología y calidad de los datos. <http://www.osuah.cl/encuestapanelcasen>
5. BUCK, N., ERMISCH, J. F. y JENKINS, S. P. 1995. Choosing a longitudinal survey design: the issues.
6. CASTRO, JUAN F. 2006. Política y gasto social en el Perú: cuánto se ha avanzado y qué más se puede hacer para reducir la vulnerabilidad de los hogares [documento de discusión]. Lima, Centro de Investigación de la Universidad del Pacífico.
7. DEATON et al. 1986. Collecting panel data in developing countries: does it make sense? [working paper] <http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED282819>.
8. ELLIOT, D., LYNN, P. y SMITH, P. 2009. Sample design for longitudinal surveys. En: Methodology of longitudinal surveys. John Wiley
9. FERNÁNDEZ, TABARÉ. 2003-2004. Programa de doctorado de ciencia social: estadística I [fichas]. México (ficha 23, pp4-5).
10. FREES, EDWARD W. y KIM, JEE-SON. 2007. Longitudinal and panel data. Madison, University of Wisconsin
11. FUNDACIÓN PARA LA SUPERACIÓN DE LA POBREZA, MIDEPLAN, OBSERVATORIOS SOCIAL UNIVERSIDAD ALBERTO HURTADO. 2007. La encuesta panel CASEN 1996, 2001, 2006: Primera fase de análisis. <http://www.osuah.cl/encuestapanelcasen>
12. GARCÍA, INMACULADA y MONTUENGA, VÍCTOR MANUEL. [s.a.] Determinantes de la siniestralidad laboral en España. Proyecto "Determinantes de la siniestralidad laboral" de la Fundación de Estudios de Economía Aragonesa.
13. GARRIDO, M. A. [s.a.] Estadística descriptiva. Universidad de Sevilla. pp 22.
14. GOODMAN, LEO. 1965. On Simultaneous Confidence Intervals for Multinomial Proportions. Technometrics 7 (2). pp 247 – 254.
15. GREENE, WILLIAM. 2003. Problemas en los datos. En: Econometric Analysis. 5ª Ed. Prentice Hall. pp 363 – 420.
16. GREENE, WILLIAM. 2003. Modelos para datos de panel. En: Econometric Analysis. 5ª Ed. Prentice Hall. pp 531 - 556.
17. HSIAO, CHENG. 2003. Analysis of panel data. 2ª Ed. Cambridge, University Press.
18. INGENOVA, 2008. Proyecto Métricas para la Calidad de Servicio, Informe 2. pp 6 - 18
19. INSTITUTO NACIONAL DE ESTADÍSTICAS. [s.a.] Código de buenas prácticas de las estadísticas chilenas [en línea]. Santiago, Chile. http://www.ine.cl/canales/corporativo/buenas_practicas/pdf/buenaspracticas_pag.pdf
20. INSITITUTO NACIONAL DE ESTADÍSTICAS E INFORMÁTICA. 1997. Anexo metodológico. En: Percepción de los jefes de hogar sobre el consumo de

drogas en su barrio o localidad [en línea]. Lima, Perú. <http://www1.inei.gob.pe/biblioineipub/bancopub/Est/Lib0054/anexmeto.htm>
21. JENKINS, STEPHEN P. [s.a.] The value of longitudinal data. University of Essex, Institute for Social and Economic Research [en línea] <http://melbourneinstitute.com/hilda/conf/conf2003/pdffiles/SJenkins.pdf>.
22. LEPKOWSKI, JAMES M. 1990. Sampling the difficult-to-sample. Ann Harbor, University of Michigan, Institute for Social Research.
23. LYNN, PETER. 2005. Metodología de las encuestas longitudinales. En: XIX Seminario de Estadística: Bilbao, 28 – 30 de Noviembre de 2005.
24. MALHOTRA, NARESH K. et al. 2004. Investigación de mercados: un enfoque aplicado. 4ª Ed. Ciudad de México, Pearson Education. pp312 – 331.
25. MANUAL SPSS [s.a.]. Análisis de Varianza Anova de un Factor, Capítulo 14. Análisis Anova Factorial, Capítulo 15. Test No Paramétricos, Capítulo 19 pp12-15.
26. MAYORGA, MAURICIO y MUÑOZ, EVELYN. [s.a.] La técnica de datos de panel, una guía para su uso e interpretación. San Juan, Departamento de Investigaciones Económicas del Banco Central de Costa Rica.
27. MEDINA, FERNANDO. [s.a.] Tamaño óptimo de muestra de encuestas de propósitos múltiples. CEPAL.
28. PÉREZ, C., DÍAZ DE SERRALDE, S. y PICOS S., F. 2008. Construcción de paneles de datos con información de fuentes administrativas a partir de muestreo. Aplicación al panel de datos del impuesto sobre la renta. En: XV Encuentro de Economía Pública: Salamanca, 7 y 8 de Febrero de 2008. <www.usal.es/~XVEEP/PAPERS/V3S2/XVEEP-89%20PEREZ%20ET%20AL.pdf>
29. SÁNCHEZ PÉREZ, ROSARIO. [s.a.] Productividad y desempleo: un estudio a través de salarios de eficiencia. Valencia, Departamento de Análisis Económico de la Universidad de Valencia.
30. SANTAMARÍA, LUIS y SURROCA, JORDI. 2004. Idoneidad del socio tecnológico. Un análisis con datos de panel. Madrid, Serie de Economía de Empresa de la Universidad Carlos III.
31. SOSA ESCUDERO, WALTER. [s.a.] Modelos lineales para datos en paneles [Documentos de clases de econometría avanzada] <http://cablemodem.fibertel.com.ar/wsosa/econometriaunlp/PanelGraduate.pdf >. Buenos Aires, Universidad de San Andrés.
32. SUDMAN, S. y WANSINK, B. 2002. Consumer panels. 2ª Ed. American Marketing Association.
33. VALLEJO RUIZ, MÓNICA. 2005. Estudio longitudinal de la producción española de tesis doctorales en educación matemática. Tesis Doctoral. Granada, Universidad de Granada, Departamento de Métodos de Investigación y Diagnóstico en Educación.
34. VANDECASTEELE, L. y DEBELS, A. 2007. Attrition in panel data: the effectiveness of weighting. European Sociological Review 23 (1). Oxford Journals
35. WOOLDRIDGE, J.M. 2002. Sample, selection, attrition and stratified sampling. En: Econometric analysis of cross section and panel data. Cambridge, The MIT Press. pp 551 – 602.

Capítulo 10

ANEXO A: Encuesta "Observatorio de Clientes BE"

Ligado al proyecto "Métricas para la Calidad de Servicio BancoEstado" se requiere construir un observatorio de clientes cuyo objetivo es monitorear de forma continua el comportamiento de los clientes para comprender y analizar los vínculos existentes con la calidad y los resultados.

Por comportamiento se entiende todas las acciones que los clientes realizan en relación a los productos del banco y los canales que utilizan.

El Panel de Clientes estará compuesto por una muestra representativa de los distintos tipos de clientes del Banco, de modo de poder calcular sobre ellos los indicadores de comportamiento de forma rápida y eficiente, y proyectar con validez estadística los resultados al total de clientes del banco.

Los objetivos del panel son:

(i) Permitir estudiar el comportamiento desde el punto de vista del consumidor, para enriquecer medidas como: tenencia, frecuencia de uso de productos, monto de transacciones y uso de canales, considerando variables demográficas de los clientes BancoEstado (segmento, género, lugar de residencia, renta,etc).

(ii) Crear una clasificación a través de variables de comportamiento como: antigüedad, saldo promedio en productos vigentes, canal de atención, productos contratados y frecuencia de uso de productos, para establecer índices de relación cruzada entre las variables de comportamiento señaladas.

Logrando reunir información de buena calidad y suficiente para responder cuestionamientos como:

¿Qué características tienen y cómo se comportan los clientes qué?

- Tienen un uso más intenso de Internet este trimestre
- Realizaron más compra con tarjeta de crédito el mes pasado
- Aumentaron la tenencia de productos este último año
- Acuden a la caja vecina y poseen cuenta rut
- Aumentaron (disminuyeron) sus saldos en cuenta corriente el mes anterior
- Aumentaron sus créditos en la competencia

En virtud de esto le solicitamos contestar:

Si ud. pudiera observar todas las operaciones de un cliente del banco y los canales que utiliza durante un periodo de tiempo, un año por ejemplo.

1. ***¿Qué enfoques utilizaría para observar el comportamiento?***

   Como ejemplos considere:

   - Clientes con renta inferior a $200.000 v/s clientes con renta superior
   - Clientes de cuenta corriente
   - Clientes de cuenta rut
   - Clientes que utiliza internet
   - Clientes de la región metropolitana

- Clientes de tarjeta de crédito

ANEXO B: Estudio de Canales

A continuación se presenta un ejemplo del análisis realizado para definir la clasificación de los distintos medios a través de los cuales el cliente se relaciona con la entidad.
Dicho análisis se basa en el entendimiento de las acciones realizadas por los clientes a través de cada canal (transacciones) y de acuerdo a ello se define su categoría.
En el ejemplo buzonera: es definido como máquinas pues las transacciones asociadas a dicho canal son realizadas de manera remota, de forma desatendida y no requieren del personal de la institución para ser efectuadas.

Tabla 47: Ejemplo Estudio Transacciones por Canal

N°	Canales	Transacciones	Financieras	Consultas	Considerar Tx?	Considerar Canal?	Clasificación Canal
1	Batch Terminal	LIBERAC.RETENC.POR DEP.E	1	0	NO	NO	Canal Interno
2	Buzonera	Actualizar libreta	0	1	SI	SI	Máquinas
		Cargo por pago de servicios	1	0	NO		
		Cargo por traspaso	1	0	NO		
		Cargo por traspaso Internet tard	1	0	NO		
		CONSULTA POR CAJERO AUT	0	1	SI		
		Depos.con.Doc.Misma plaza po	1	0	SI		
		Deposito efectivo por cajero aut	1	0	SI		
		Deposito en documento por caje	1	0	SI		
		Deposito en efectivo por cajero	1	0	SI		
		Giro sin/lbrt Internet mañana	1	0	NO		
		Pago cuota Internet dia	1	0	NO		
		Pago cuota Internet tarde	1	0	NO		
		Pago hipotecario Internet Dia	1	0	NO		
		Pago hipotecario Internet tarde	1	0	NO		
		Pago M/N (Internet mañana)	1	0	NO		
		Pago M/N (Internet tarde)	1	0	NO		
		REVERSA LIBERAR RETENCIO	0	1	SI		
3	Caja vecina	Abono por caja	1	0	SI	NO	Sucursal
		Cargo por caja	1	0	SI		
		Cargo por pago de servicios	1	0	NO		
		Deposito efectivo sin/librt.	1	0	NO		
		Giro sin/librt p./corr.	1	0	NO		
		Pago cuota crédito	1	0	NO		
		Pago hipotecario	1	0	NO		

Fuente: Elaboración Propia

ANEXO C: Comparación Región Demográfica v/s Región Transaccional del cliente

```
SELECT SUM(T1.CNT) CANT_TRANS, T1.CST_ID, T2.RGN_ID
FROM TRANSACCIONES AS T1,TABLA_OU AS T2
WHERE T1.OU_IP_ID=T2.OU_IP_ID
GROUP BY T1.CST_ID, T2.RGN_ID

EXEC DBO.INVIERTE_TRAS '#PRUEBA_REGION', 'RGN_ID', 'CST_ID', 'CANT_TRANS', 'SUM',
'REG_TRANS_MUESTRA1'

SELECT * FROM REG_TRANS_MUESTRA1

SELECT CST_ID, CASE
WHEN (SUM_1>SUM_2 AND SUM_1>SUM_3 AND SUM_1>SUM_4 AND SUM_1>SUM_5 AND SUM_1>SUM_6 AND
SUM_1>SUM_7 AND SUM_1>SUM_8 AND SUM_1>SUM_9 AND SUM_1>SUM_10 AND SUM_1>SUM_11 AND SUM_1>SUM_12
AND SUM_1>SUM_13 AND SUM_1>SUM_14 AND SUM_1>SUM_15) THEN 1
WHEN (SUM_2>SUM_1 AND SUM_2>SUM_3 AND SUM_2>SUM_4 AND SUM_2>SUM_5 AND SUM_2>SUM_6 AND
SUM_2>SUM_7 AND SUM_2>SUM_8 AND SUM_2>SUM_9 AND SUM_2>SUM_10 AND SUM_2>SUM_11 AND SUM_2>SUM_12
AND SUM_2>SUM_13 AND SUM_2>SUM_14 AND SUM_2>SUM_15) THEN 2
```

```
WHEN (SUM_3>SUM_1 AND SUM_3>SUM_2 AND SUM_3>SUM_4 AND SUM_3>SUM_5 AND SUM_3>SUM_6 AND
SUM_3>SUM_7 AND SUM_3>SUM_8 AND SUM_3>SUM_9 AND SUM_3>SUM_10 AND SUM_3>SUM_11 AND SUM_3>SUM_12
AND SUM_3>SUM_13 AND SUM_3>SUM_14 AND SUM_3>SUM_15) THEN 3
WHEN (SUM_4>SUM_1 AND SUM_4>SUM_2 AND SUM_4>SUM_3 AND SUM_4>SUM_5 AND SUM_4>SUM_6 AND
SUM_4>SUM_7 AND SUM_4>SUM_8 AND SUM_4>SUM_9 AND SUM_4>SUM_10 AND SUM_4>SUM_11 AND SUM_4>SUM_12
AND SUM_4>SUM_13 AND SUM_4>SUM_14 AND SUM_4>SUM_15) THEN 4
WHEN (SUM_5>SUM_1 AND SUM_5>SUM_2 AND SUM_5>SUM_3 AND SUM_5>SUM_4 AND SUM_5>SUM_6 AND
SUM_5>SUM_7 AND SUM_5>SUM_8 AND SUM_5>SUM_9 AND SUM_5>SUM_10 AND SUM_5>SUM_11 AND SUM_5>SUM_12
AND SUM_5>SUM_13 AND SUM_5>SUM_14 AND SUM_5>SUM_15) THEN 5
WHEN (SUM_6>SUM_1 AND SUM_6>SUM_2 AND SUM_6>SUM_3 AND SUM_6>SUM_4 AND SUM_6>SUM_5 AND
SUM_6>SUM_7 AND SUM_6>SUM_8 AND SUM_6>SUM_9 AND SUM_6>SUM_10 AND SUM_6>SUM_11 AND SUM_6>SUM_12
AND SUM_6>SUM_13 AND SUM_6>SUM_14 AND SUM_6>SUM_15) THEN 6
WHEN (SUM_7>SUM_1 AND SUM_7>SUM_2 AND SUM_7>SUM_3 AND SUM_7>SUM_4 AND SUM_7>SUM_5 AND
SUM_7>SUM_6 AND SUM_7>SUM_8 AND SUM_7>SUM_9 AND SUM_7>SUM_10 AND SUM_7>SUM_11 AND SUM_7>SUM_12
AND SUM_7>SUM_13 AND SUM_7>SUM_14 AND SUM_7>SUM_15) THEN 7
WHEN (SUM_8>SUM_1 AND SUM_8>SUM_2 AND SUM_8>SUM_3 AND SUM_8>SUM_4 AND SUM_8>SUM_5 AND
SUM_8>SUM_6 AND SUM_8>SUM_7 AND SUM_8>SUM_9 AND SUM_8>SUM_10 AND SUM_8>SUM_11 AND SUM_8>SUM_12
AND SUM_8>SUM_13 AND SUM_8>SUM_14 AND SUM_8>SUM_15) THEN 8
WHEN (SUM_9>SUM_1 AND SUM_9>SUM_2 AND SUM_9>SUM_3 AND SUM_9>SUM_4 AND SUM_9>SUM_5 AND
SUM_9>SUM_6 AND SUM_9>SUM_7 AND SUM_9>SUM_8 AND SUM_9>SUM_10 AND SUM_9>SUM_11 AND SUM_9>SUM_12
AND SUM_9>SUM_13 AND SUM_9>SUM_14 AND SUM_9>SUM_15) THEN 9
WHEN (SUM_10>SUM_1 AND SUM_10>SUM_2 AND SUM_10>SUM_3 AND SUM_10>SUM_4 AND SUM_10>SUM_5 AND
SUM_10>SUM_6 AND SUM_10>SUM_7 AND SUM_10>SUM_8 AND SUM_10>SUM_9 AND SUM_10>SUM_11 AND
SUM_10>SUM_12 AND SUM_10>SUM_13 AND SUM_10>SUM_14 AND SUM_10>SUM_15) THEN 10
WHEN (SUM_11>SUM_1 AND SUM_11>SUM_2 AND SUM_11>SUM_3 AND SUM_11>SUM_4 AND SUM_11>SUM_5 AND
SUM_11>SUM_6 AND SUM_11>SUM_7 AND SUM_11>SUM_8 AND SUM_11>SUM_9 AND SUM_11>SUM_10 AND
SUM_11>SUM_12 AND SUM_11>SUM_13 AND SUM_11>SUM_14 AND SUM_11>SUM_15) THEN 11
WHEN (SUM_12>SUM_1 AND SUM_12>SUM_2 AND SUM_12>SUM_3 AND SUM_12>SUM_4 AND SUM_12>SUM_5 AND
SUM_12>SUM_6 AND SUM_12>SUM_7 AND SUM_12>SUM_8 AND SUM_12>SUM_9 AND SUM_12>SUM_10 AND
SUM_12>SUM_11 AND SUM_12>SUM_13 AND SUM_12>SUM_14 AND SUM_12>SUM_15) THEN 12
WHEN (SUM_13>SUM_1 AND SUM_13>SUM_2 AND SUM_13>SUM_3 AND SUM_13>SUM_4 AND SUM_13>SUM_5 AND
SUM_13>SUM_6 AND SUM_13>SUM_7 AND SUM_13>SUM_8 AND SUM_13>SUM_9 AND SUM_13>SUM_10 AND
SUM_13>SUM_11 AND SUM_13>SUM_12 AND SUM_13>SUM_14 AND SUM_13>SUM_15) THEN 13
WHEN (SUM_14>SUM_1 AND SUM_14>SUM_2 AND SUM_14>SUM_3 AND SUM_14>SUM_4 AND SUM_14>SUM_5 AND
SUM_14>SUM_6 AND SUM_14>SUM_7 AND SUM_14>SUM_8 AND SUM_14>SUM_9 AND SUM_14>SUM_10 AND
SUM_14>SUM_11 AND SUM_14>SUM_12 AND SUM_14>SUM_13 AND SUM_14>SUM_15) THEN 14
WHEN (SUM_15>SUM_1 AND SUM_15>SUM_2 AND SUM_15>SUM_3 AND SUM_15>SUM_4 AND SUM_15>SUM_5 AND
SUM_15>SUM_6 AND SUM_15>SUM_7 AND SUM_15>SUM_8 AND SUM_15>SUM_9 AND SUM_15>SUM_10 AND
SUM_15>SUM_11 AND SUM_15>SUM_12 AND SUM_15>SUM_13 AND SUM_15>SUM_14) THEN 15
END
"REGION"
INTO PRUEBA_MUESTRA1
FROM REG_TRANS_MUESTRA1

/*COMPARA REGION DEMO CON REGION DONDE MÁS TRANSACCIONA EL CLIENTE*/
SELECT SUM (CASE WHEN T1.REGION=T2.REGION THEN 1 ELSE 0 END) COMPARA, COUNT (DISTINCT
T2.CST_ID) POBLACION
FROM MARCO_MUESTRAL AS T1, PRUEBA_MUESTRA1 AS T2
WHERE T1.RUT_DEMO=T2.CST_ID
```

ANEXO D: Reemplazo Dato Región y Comuna

Como ha sido explicado anteriormente tanto para el levantamiento del panel como para su posterior mantención se debe tener en consideración el reemplazo de los valores nulos para los datos demográficos de región y comuna de la siguiente manera:

```
/*EXTRAE REGISTRO DE COMUNA ASIGNADA PARA CLIENTES SIN DATO*/
SELECT T1.RUT_DEMO RUT, T3.OFI_RGN_COD,T3.OFI_CMN_COD
INTO REGION_DE_REEMPLAZO
FROM TABLA_MARCO_MUESTRAL2 AS T1, OFICINA_CLI AS T2, TABLA_OFICINAS AS T3
WHERE T1.RUT_DEMO=T2.CLI_RUT AND T2.CLI_OFI_COD=T3.OFI_COD

/*REEMPLAZA REGIÒN*/
UPDATE TABLA_MARCO_MUESTRAL
SET REGION= T2.OFI_RGN_COD
FROM TABLA_MARCO_MUESTRAL T, REGION_DE_REEMPLAZO T2
WHERE T.RUT_DEMO=T2.RUT AND REGION=NULL

/*REEMPLAZA COMUNA*/
UPDATE TABLA_MARCO_MUESTRAL
```

```
SET COM_COD= T2.OFI_CMN_COD
FROM TABLA_MARCO_MUESTRAL T, REGION_DE_REEMPLAZO T2
WHERE T.RUT_DEMO=T2.RUT AND COM_COD=NULL
```

ANEXO E: Estudio de Transacciones Producto/Canal

Tabla 48: Transacciones por Producto y Canal

PRODUCTO	ACCION CLIENTE	CANAL				
		SUCURSAL	INTERNET	CAJERO	BUZONERA	CALL CENTER
CTA CTE						
	ACLARACION PROTESTO	1				
	ANULA ORDEN DE NO PAG(	1				
	ANUNCIO ORDEN DE NO P/	1				1
	APERTURA CUENTA	1	1			
	CONSULTA SALDO		1	1		
	DEPOSITO EFECTIVO	1		1	1	
	DEPOSITO DOCUMENTO	1		1	1	
	GIRO			1		
	PAGO CON CARGO CTA CTE		1			
	SOLICITUD CARTOLA	1	1			1
	TRANSFERENCIA ENTRE CUENTAS		1			
	TRANSFERENCIA OTROS BANCOS		1			
TARJETA CREDITO						
	AVANCE EFECTIVO	1		1		
	CONSULTA ESTADO		1			1
	CONSULTA CARTOLA		1	1		
	PAGO CUOTA	1	1	1	1	
	COMPRAS					
	TRANSFERENCIA A OTRA CTA		1			

Fuente: Elaboración Propia

ANEXO F: Procedimiento para definir la Tenencia por Producto

```
SELECT T1.PRIM_CST_ID AS RUT_PRODS, T1.PD_ID, T2.PROD_PANEL, COUNT(*) AS N_PRODS
INTO CARO_TENENCIA_TEMP
FROM TABLA_AR AS T1, TABLA_PRODUCTOS_AGRUPADOS AS T2
WHERE T1.EFF_DT<='20081031' AND T1.END_DT >'20081031' AND T1.PD_ID=T2.PD_ID
GROUP BY T1.PRIM_CST_ID, T1.PD_ID, T2.PROD_PANEL

INSERT INTO CARO_TENENCIA_TEMP
SELECT T1.PRIM_CST_ID AS RUT_PRODS, T1.PD_ID, T2.PROD_PANEL, COUNT(*) AS N_PRODS
FROM TABLA_AR_AHO AS T1, TABLA_PRODUCTOS_AGRUPADOS AS T2
WHERE T1.EFF_DT<='20081031' AND T1.END_DT >'20081031' AND T1.PD_ID=T2.PD_ID
GROUP BY T1.PRIM_CST_ID, T1.PD_ID, T2.PROD_PANEL

/* AGRUPA POR PROD_PANEL */
SELECT RUT_PRODS, PROD_PANEL, SUM(N_PRODS) AS SUM_N_PRODS
INTO CARO_TENENCIA_TEMP2
FROM CARO_TENENCIA_TEMP
GROUP BY RUT_PRODS, PROD_PANEL

/* OBTIENE TENENCIA DE PRODUCTOS (TRANSPUESTO) */
EXEC PR_CROSSTAB_TABLE 'CARO_TENENCIA_TEMP2', 'PROD_PANEL', 'RUT_PRODS', 'PROD_PANEL',
'COUNT', 'TABLA_TENENCIA_VIGENTES'
```

Como se ve para el cálculo de la tenencia por producto fue necesario utilizar el procedimiento almacenado PR_crosstab el cual transpone una matriz de acuerdo a los campos de agrupación dados.

Dicho procedimiento se presenta a continuación pues no es parte de las variables de inicio de SQL server.[46]

```
--SE CREA EL PROCEDIMIENTO ALMACENADO
CREATE PROCEDURE PR_CROSSTAB
@TABLA VARCHAR(255), @PIVOT VARCHAR(255), @AGRUPACION VARCHAR(255), @CAMPO VARCHAR(255),
@CALCULO VARCHAR(20)
AS
--SE DECLARAN LAS VARIABLES QUE PERMITIRÁN CREAR EL SQL CON LOS "CASES"
DECLARE @STRG AS VARCHAR(8000)
DECLARE @SQL AS VARCHAR(8000)
CREATE TABLE #PIVOT (
PIVOT VARCHAR (8000) )
-- SE LIMPIAN LAS VARIABLES
SET @STRG=''
SET @SQL=''
-- SE REALIZA UN "SELECT DISTINCT" DEL CAMPO SE UTILIZA COMO PIVOTE, A CADA REGISTRO SE LE
CONCATENA SU CORRESPONDIENTE "CASE" Y SE ALMACENA EN UNA TABLA TEMPORAL LLAMADA #PIVOT
SET @STRG=@STRG + 'INSERT INTO #PIVOT SELECT DISTINCT ''' + @CALCULO + '(CASE WHEN ' + @PIVOT
+ '=''''''+ RTRIM(CAST(' + @PIVOT + ' AS VARCHAR(500)))
+ '''''' THEN ' + @CAMPO + ' ELSE NULL END) AS ''''' + @CALCULO + '_'' +
RTRIM(CAST(' + @PIVOT + ' AS VARCHAR(500))) + ''''''', '' AS PIVOT
FROM ' + @TABLA + ' WHERE ' + @PIVOT + ' IS NOT NULL'

EXECUTE (@STRG)

/* CONSULTA FINAL: SELECCIONA LAS COLUMNAS SEGÚN LA TABLA #PIVOT Y REALIZA LA AGRUPACIÓN
CORRESPONDIENTE.
SET @SQL ='SELECT '
SELECT @SQL= @SQL + RTRIM(CONVERT(VARCHAR(500), PIVOT))
FROM #PIVOT ORDER BY PIVOT
IF @AGRUPACION<>'*'
BEGIN
SET @SQL=@SQL + @AGRUPACION + ' FROM ' + @TABLA + ' GROUP BY ' +
@AGRUPACION
END
ELSE
BEGIN
SET @SQL=@SQL + '''TODOS'' AS T FROM ' + @TABLA
END
```

ANEXO G: Procedimiento de Clasificación de Clientes en GRS

A continuación, de modo ilustrativo, se presenta la programación realizada para la asignación de clientes a cada GRS pues por motivos de confidencialidad de la información no se expondrán los nombres de los productos clasificadores y los GRS resultantes son codifcados desde G1 a G6.

```
/* ASIGNA CLIENTES A CADA GRS A PARTIR DE LA MATRIZ DE TENENCIA DE CLIENTES*/
SELECT *,
CASE
WHEN (([P6]=1 OR [P13]=1 OR [P19]=1 OR [P20]=1 OR [P21]=1)  AND ([P4]=0 AND [P5]=0 AND [P9]=0
AND [P10]=0 AND [P11]=0 AND  [P12]=0)) THEN 'G1'
WHEN (([P6]=1 OR [P13]=1 OR [P19]=1 OR [P20]=1 OR [P21]=1) AND ([P4]=1 OR [P5]=1 OR [P9]=1 OR
[P10]=1 OR [P11]=1 OR [P12]=1)) THEN 'G2'
WHEN (([P2]=1 OR [P7]=1 OR [P8]=1 OR [P14]=1) AND ([P4]=0 AND [P5]=0 AND [P9]=0 AND [P10]=0
AND [P11]=0 AND  [P12]=0) AND ([P6]=0 AND [P13]=0 AND [P19]=0 AND [P20]=0 AND [P21]=0)) THEN
'G3'
WHEN (([P2]=1 OR [P7]=1 OR [P8]=1 OR [P14]=1) AND ([P4]=1 OR [P5]=1 OR [P9]=1 OR
[P10]=1 OR [P11]=1 OR [P12]=1) AND ([P6]=0 AND [P13]=0 AND [P19]=0 AND [P20]=0 AND [P21]=0))
THEN 'G4'
WHEN (([P1]=1 OR [P3]=1) AND ([P4]=0 AND [P5]=0 AND [P9]=0 AND [P10]=0 AND [P11]=0 AND
[P12]=0) AND ([P2]=0 AND [P7]=0 AND [P8]=0 AND [P14]=0) AND ([P6]=0 AND [P13]=0 AND [P19]=0
AND [P20]=0 AND [P21]=0)) THEN 'G5'
```

[46] Fuente: < www.lawebdelprogramador.com>

```
WHEN (([P1]=1 OR [P3]=1) AND ([P4]=1 OR [P5]=1 OR [P9]=1 OR [P10]=1 OR [P11]=1 OR [P12]=1) AND
([P2]=0 AND [P7]=0 AND [P8]=0 AND [P14]=0) AND ([P6]=0 AND [P13]=0 AND [P19]=0 AND [P20]=0 AND
[P21]=0)) THEN 'G6'
ELSE 'NO GRS ASOCIADO'
END
"GRS"
INTO CLI_GRS_VIGENTES
FROM TABLA_TENENCIA_VIGENTES
```

ANEXO H: Estudio Variable Saldo Promedio Mensual

Para el estudio de la variable saldo promedio mensual se utilizó una muestra aleatoria de 25 mil clientes para los cuales se solicitó su información del saldo promedio mensual diario para su análisis con el sofware SPSS.

Los resultados de los estadísticos descriptivos de esta variable, considerando todos los productos, son:

Tabla 49: Estadísticos Descriptivos Saldo

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Saldo	273935	-91007	312973072	814777	3524448	12421734404560	4,33

Fuente Elaboración Propia

De dichos resultados es posible notar una alta magnitud en la dispersión de los datos y un coeficiente de variación que permite concluir la mala representatividad de la media obtenida valor de la población completa (CV>1).

Además la existencia de un máximo tan elevado hace necesario estudiar la existencia de outliers y para ello se utilizó el método de cajas descrito en el capítulo 4 obteniendo en primer lugar los cuartiles y posteriormente las respectivas distancias para definir las cotas de los valores atípicos moderados y severos:

Tabla 50: Cuartiles Variable Saldo

Cuartiles		
25	50	75
1052	14295	200000

Fuente Elaboración Propia

Tabla 51: Distancia Intercuartil Variable Saldo

Distancia Intercuartil		
dQ	1.5dQ	3dQ
198948	298422	596844

Fuente Elaboración Propia

Y por lo tanto las cotas quedan definidas por:

Tabla 52: Cotas Valores Atípicos Variable Saldo

Inf Moderado	-297370
Sup Moderado	498422
Inf Severo	-595792
Sup Severo	796844

Fuente Elaboración Propia

Los valores de la cota inferior hacen posible mantener todos los datos menores que cero de la muestra pues el valor del mínimo en ella es superior a dichas cotas.

Para los valores superiores se decidió estudiar ambos casos obteniendo:

a) Caso eliminación valores atípicos severos

Tabla 53: Estadísticos Descriptivos Variable Saldo Sin Datos Atípicos Severos

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Saldo	233990	-91007,5	796728,86	71394,1	144745,5	20951255549	2,03

Fuente Elaboración Propia

b) Caso eliminación valores atípicos moderados

Tabla 54: Estadísticos Descriptivos Variable Saldo Sin Datos Atípicos Moderados

	N	Mínimo	Máximo	Media	Desv. típ.	Varianza	CV
Saldo	225506	-91007,5	498319	50460,95	96811,7	9372505182	1,92

Fuente Elaboración Propia

El cambio abrupto del valor de la media al eliminar dichos valores hace imposible sostener que la estimación de valores puntuales para ella será factible con el panel de clientes.

Además en ambos casos luego de realizar una alta eliminación de datos (15% y 17% respectivamente) la dispersión de los datos sigue siendo alta y los valores del coefiente de variación confirman la declaración del párrafo anterior.

Realizando el estudio de frecuencias con los datos mensuales disponibles para 1 año se definieron 6 intervalos de agrupación:

NS1: Saldo ≤ 0
NS2: 0 < Saldo ≤ 10000
NS3: 10000 < Saldo ≤ 35000
NS4: 35000 < Saldo ≤ 100000
NS5: 100000 < Saldo ≤ 250000
NS6: Más de 250000

ANEXO I: Test Anova Factorial para Variables de Estratificación

Como se explicó en el marco teórico en la prueba anova factorial existe una hipótesis nula por cada factor y por cada posible combinación de factores que plantea la inexistencia de diferencias en las medias de los grupos definidos por los niveles y los cruces de los factores.

Debido a la gran cantidad de combinaciones de niveles de las variables candidatas para estratificar el procesamiento conjunto del anova factorial es computacionalmente imposible por lo que el estudio fue dividido en etapas, a través

de la realización de múltiples anovas factoriales incluyendo una variable candidata en cada paso y verificando los valores de los R^2 obtenidos en cada etapa por ser este el valor considerado como el mejor indicador disponible de la calidad de la relación entre la variable dependiente y los factores pues es este el valor que representa el porcentaje de varianza explicada por el modelo generado por incluir todos los efectos de las variables estudiadas y por ende realiza una cuantificación concreta del aporte otorgado por el factor incluido en cada paso.

A pesar no poder afirmar la validez de las bondades de ajuste obtenidas en estos resultados se decide exponerlos con fines ilustrativos para comprensión de las distintas actividades llevadas a cabo para el desarrollo de la presente tesis.

A continuación se presenta los cuadros resumen de los resultados obtenidos en cada paso:

i) Paso 1: Análisis Segmento + GRS

Matriz de Varianza Explicada (%)

Paso	Variables	Monto	Cantidad	Tenencia	Saldo
1	Segmento + GRS	23,2	37,6	76,2	13,6

ii) Paso 2: Análisis Segmento + GRS + Candidatas

Matriz de Varianza Explicada (%)

Paso	Variables	Monto	Cantidad	Tenencia	Saldo
2.a	Segmento+GRS+Género	24,5	38,5	76,4	16,1
2.b	Segmento+GRS+Antigüedad	26,7	41,3	76,5	63,8
2.c	Segmento+GRS+E_Civil	24,5	38,9	76,4	66,2
2.d	Segmento+GRS+Renta	25,9	40,6	76,6	66,6
2.e	Segmento+GRS+Edad	25,6	40,1	76,5	67,3
2.f	Segmento+GRS+Región	37,3	42,7	76,8	85,6

iii) Paso 3: Análisis Segmento + GRS + Región + Candidatas

Matriz de Varianza Explicada (%)

Paso	Variable	Monto	Cantidad	Tenencia	Saldo
3.a	Segmento+GRS+Región+Antigüedad	80,4	48,9	77,4	78,9
3.b	Segmento+GRS+Región+Renta	85,8	53,0	79,0	89,1
3.c	Segmento+GRS+Región+Edad	82,3	47,9	77,6	84,8
3.d	Segmento+GRS+Región+Género	79,5	47,1	77,2	85,9
	Segmento+GRS+Región+E Civil	79,1	47,6	77,3	83,4

De dónde se tiene que los aportes de una 4ta variable de estratificación son significativos para la variable monto y la mejor combinación es la obtenida en el paso 3b cuyos resultados se muestran a continuación:

Tabla 55: Resultados Anova Factoria Variable Monto de Transacciones

Pruebas de los efectos inter-sujetos

Variable dependiente: monto

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	3,258E+020[a]	801	4,1E+017	92,167	,000
Intersección	7,510E+018	1	7,5E+018	1701,940	,000
segmento	1,197E+018	5	2,4E+017	54,261	,000
grs	1,167E+019	5	2,3E+018	528,987	,000
region	4,652E+018	14	3,3E+017	75,302	,000
renta	5,522E+017	6	9,2E+016	20,857	,000
segmento * grs	5,751E+018	19	3,0E+017	68,591	,000
segmento * region	2,494E+018	60	4,2E+016	9,420	,000
grs * region	1,771E+019	70	2,5E+017	57,334	,000
segmento * grs * region	6,395E+018	108	5,9E+016	13,418	,000
segmento * renta	2,394E+017	10	2,4E+016	5,424	,000
grs * renta	4,475E+018	24	1,9E+017	42,257	,000
segmento * grs * renta	2,166E+018	23	9,4E+016	21,343	,000
region * renta	1,468E+019	74	2,0E+017	44,953	,000
segmento * region * renta	1,110E+018	56	2,0E+016	4,490	,000
grs * region * renta	3,248E+019	157	2,1E+017	46,888	,000
segmento * grs * region * renta	2,323E+018	48	4,8E+016	10,968	,000
Error	5,378E+019	12186	4,4E+015		
Total	3,909E+020	12988			
Total corregida	3,796E+020	12987			

a. R cuadrado = ,858 (R cuadrado corregida = ,849)

Fuente Resultados Análisis SPSS

Tabla 56: Resultados Anova Factoria Variable Cantidad de Transacciones

Pruebas de los efectos inter-sujetos

Variable dependiente: cantidad

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	9414777439[a]	801	11753780	17,163	,000
Intersección	518312568	1	5,2E+008	756,830	,000
segmento	49201098,3	5	9840219,7	14,368	,000
grs	762654053	5	1,5E+008	222,723	,000
region	72969347,8	14	5212096,3	7,611	,000
renta	51379422,7	6	8563237,1	12,504	,000
segmento * grs	237080072	19	12477899	18,220	,000
segmento * region	97330707,8	60	1622178,5	2,369	,000
grs * region	270447438	70	3863534,8	5,641	,000
segmento * grs * region	180597012	108	1672194,6	2,442	,000
segmento * renta	61237324,9	10	6123732,5	8,942	,000
grs * renta	196909125	24	8204546,9	11,980	,000
segmento * grs * renta	177243760	23	7706250,4	11,253	,000
region * renta	319463179	74	4317070,0	6,304	,000
segmento * region * renta	19449965,2	56	347320,806	,507	,999
grs * region * renta	549967683	157	3502978,9	5,115	,000
segmento * grs * region * renta	14705926,1	48	306373,461	,447	1,000
Error	8345543158	12186	684846,804		
Total	1,884E+010	12988			
Total corregida	1,776E+010	12987			

a. R cuadrado = ,530 (R cuadrado corregida = ,499)

Fuente: Resultados Análisis SPSS

Tabla 57: Resultados Anova Factoria Variable Cantidad

Pruebas de los efectos inter-sujetos

Variable dependiente: tenencia

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	26628,703[a]	801	33,244	57,063	,000
Intersección	5257,278	1	5257,278	9023,929	,000
segmento	12,961	5	2,592	4,449	,000
grs	1512,506	5	302,501	519,233	,000
region	15,166	14	1,083	1,859	,026
renta	11,764	6	1,961	3,365	,003
segmento * grs	22,672	19	1,193	2,048	,005
segmento * region	29,816	60	,497	,853	,784
grs * region	105,079	70	1,501	2,577	,000
segmento * grs * region	89,891	108	,832	1,429	,002
segmento * renta	12,247	10	1,225	2,102	,021
grs * renta	35,251	24	1,469	2,521	,000
segmento * grs * renta	22,737	23	,989	1,697	,020
region * renta	70,663	74	,955	1,639	,000
segmento * region * renta	36,600	56	,654	1,122	,248
grs * region * renta	157,114	157	1,001	1,718	,000
segmento * grs * region * renta	64,807	48	1,350	2,317	,000
Error	7099,478	12186	,583		
Total	101347,000	12988			
Total corregida	33728,181	12987			

a. R cuadrado = ,790 (R cuadrado corregida = ,776)

Fuente: Resultados Análisis SPSS

Tabla 58: Resultados Anova Factoria Variable Saldo

Pruebas de los efectos inter-sujetos

Variable dependiente: saldo

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	5,119E+022[a]	801	6,4E+019	124,464	,000
Intersección	1,967E+020	1	2,0E+020	383,114	,000
segmento	8,367E+019	5	1,7E+019	32,590	,000
grs	3,960E+020	5	7,9E+019	154,243	,000
region	1,776E+020	14	1,3E+019	24,712	,000
renta	1,239E+019	6	2,1E+018	4,021	,000
segmento * grs	2,608E+020	19	1,4E+019	26,728	,000
segmento * region	1,544E+020	60	2,6E+018	5,012	,000
grs * region	5,724E+020	70	8,2E+018	15,926	,000
segmento * grs * region	2,417E+020	108	2,2E+018	4,359	,000
segmento * renta	9,666E+018	10	9,7E+017	1,882	,043
grs * renta	5,514E+019	24	2,3E+018	4,474	,000
segmento * grs * renta	2,790E+019	23	1,2E+018	2,362	,000
region * renta	2,117E+020	74	2,9E+018	5,571	,000
segmento * region * renta	1,345E+019	56	2,4E+017	,468	1,000
grs * region * renta	4,996E+020	157	3,2E+018	6,198	,000
segmento * grs * region * renta	1,595E+019	48	3,3E+017	,647	,972
Error	6,257E+021	12186	5,1E+017		
Total	5,790E+022	12988			
Total corregida	5,745E+022	12987			

a. R cuadrado = ,891 (R cuadrado corregida = ,884)

Fuente: Resultados Análisis SPSS

Los resultados obtenidos permiten confirmar de inmediato las variables de estratificación seleccionadas, tanto por su significancia individual como por el aporte de sus interacciones, sin embargo no son confiables ya que los datos no cumplen los requisitos de homocedasticidad necesario para concluir con el estadístico F-Fisher.

Pues por ejemplo para la variable monto el test de Levene corresponde a:

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente: monto

F	gl1	gl2	Significación
18,598	801	12186	,000

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño:
Intersección+segmento+grs+region+renta+segmento * grs+segmento * region+grs * region+segmento * grs * region+segmento * renta+grs * renta+segmento * grs * renta+region * renta+segmento * region * renta+grs * region * renta+segmento * grs * region * renta

Fuente: Resultados Análisis SPSS

Como fue señalado anteriormente, a pesar de no poder concluir con este estudio realizado se expone a modo de compresión del lector.

ANEXO J: Contrastes de medias para variables de estratificación

Cabe señalar que las variables de interés saldo, monto y cantidad de transacciones fueron calculadas en promedio para el año de datos de cada cliente.

(i) Variable Segmento

Tabla 59: Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
saldo	382,264	5	12982	,000
monto	420,405	5	12982	,000
cantidad	863,307	5	12982	,000
tenencia	565,455	5	12982	,000

Fuente: Resultados Análisis SPSS

Para todas las variables de interés la prueba de Levene entrega un p_valor < 0,05 por lo tanto para todas ellas se rechaza la igualdad de varianza en las poblaciones testeadas.

Debido al incumplimiento de la homocedasticidad la prueba F-Fisher de del test anova convencional no sirve aceptar o rechazar la hipótesis nula de igualdad de medias y por lo tanto, para concluir, fue necesario realizar las pruebas robustas y/o la prueba no parámetrica de Kruskal-Wallis descritas en el marco teórico.

Tabla 60: Pruebas robustas de igualdad de las medias

		Estadístico(a)	gl1	gl2	Sig.
saldo	Welch	65,340	5	1608,315	,000
	Brown-Forsythe	14,900	5	206,770	,000
monto	Welch	113,932	5	1609,013	,000
	Brown-Forsythe	58,016	5	368,006	,000
cantidad	Welch	229,996	5	1608,622	,000
	Brown-Forsythe	91,493	5	323,499	,000
tenencia	Welch	1631,826	5	1628,014	,000
	Brown-Forsythe	394,467	5	957,345	,000

Tabla 61: Prueba no paramétrica de Kruskal-Wallis Estadísticos de contraste(a,b)

	saldo	monto	cantidad	tenencia
Chi-cuadrado	2449,593	3094,456	2289,239	2197,325
gl	5	5	5	5
Sig. asintót.	,000	,000	,000	,000

a Prueba de Kruskal-Wallis y b Variable de agrupación: segmento

Los estadísticos robustos presentados en la tabla 60 y el estadístico no paramétrico de Kruskal-Wallis (tabla 61) permiten rechazar la igualdad de medias y concluir que si existe alguna relación entre las variables de interés y segmento.

A continuación a modo ilustrativo se presentan los resultados de la prueba Games-Howell para la variable monto la cual permite contrastar las diferencias efectivas entre las medias de los distintos niveles de la variable segmento cuando no se cumple el requisito de homocestacidad (tabla 59).

Comparaciones múltiples

Variable dependiente: monto

Games-Howell

(I) segmento	(J) segmento	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1,00	2,00	25009603,55*	7222965,6	,007	4393554,7	5E+007
	3,00	34010686,85*	7190419,0	,000	13486940	5E+007
	4,00	38473643,35*	7171636,9	,000	18003159	6E+007
	5,00	-71484160,7*	11059277	,000	-1,0E+008	-4E+007
	6,00	-298738124*	46140462	,000	-4,3E+008	-2E+008
2,00	1,00	-25009603,5*	7222965,6	,007	-45625652	-4393555
	3,00	9001083,302*	1005652,1	,000	6134641,7	1E+007
	4,00	13464039,80*	861154,10	,000	11009350	2E+007
	5,00	-96493764,2*	8462672,4	,000	-1,2E+008	-7E+007
	6,00	-323747728*	45587843	,000	-4,5E+008	-2E+008
3,00	1,00	-34010686,8*	7190419,0	,000	-54534434	-1E+007
	2,00	-9001083,30*	1005652,1	,000	-11867525	-6134642
	4,00	4462956,496*	521995,77	,000	2973859,2	5952054
	5,00	-105494848*	8434910,7	,000	-1,3E+008	-8E+007
	6,00	-332748811*	45582698	,000	-4,6E+008	-2E+008
4,00	1,00	-38473643,3*	7171636,9	,000	-58944128	-2E+007
	2,00	-13464039,8*	861154,10	,000	-15918730	-1E+007
	3,00	-4462956,50*	521995,77	,000	-5952054	-2973859
	5,00	-109957804*	8418905,4	,000	-1,3E+008	-9E+007
	6,00	-337211768*	45579739	,000	-4,7E+008	-2E+008
5,00	1,00	71484160,66*	11059277	,000	39945341	1E+008
	2,00	96493764,21*	8462672,4	,000	72346328	1E+008
	3,00	105494847,5*	8434910,7	,000	81426214	1E+008
	4,00	109957804,0*	8418905,4	,000	85934600	1E+008
	6,00	-227253963*	46350705	,000	-3,6E+008	-9E+007
6,00	1,00	298738124,2*	46140462	,000	1,7E+008	4E+008
	2,00	323747727,7*	45587843	,000	1,9E+008	5E+008
	3,00	332748811,0*	45582698	,000	2,0E+008	5E+008
	4,00	337211767,5*	45579739	,000	2,1E+008	5E+008
	5,00	227253963,5*	46350705	,000	93949068	4E+008

*. La diferencia de medias es significativa al nivel .05.

Fuente: Resultados Análisis SPSS

De los resultados ilustrados en la tabla anterior se verifica la existencia de diferencias significativas en la media del monto de transacciones de los clientes para todos los grupos formados por los niveles de la variable segmento.

(ii) Variable GRS

Tabla 62: Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
saldo	281,896	5	12982	,000
monto	631,442	5	12982	,000
cantidad	1584,522	5	12982	,000
tenencia	484,946	5	12982	,000

Fuente: Resultados Análisis SPSS

Como no se cumple la condición de homocedasticidad en los grupos formados por la variable grs, igual que para la variable segmento, fue necesario determinar la conclusión de igualdad de medias entre los distintos niveles de la variable grs observando los resultados de las pruebas robustas y/o el test no paramétrico presentados a continuación:

Tabla 63: Pruebas robustas de igualdad de las medias

		Estadístico(a)	gl1	gl2	Sig.
saldo	Welch	74,148	5	1565,692	,000
	Brown-Forsythe	34,516	5	871,827	,000
monto	Welch	79,723	5	1597,800	,000
	Brown-Forsythe	94,283	5	1060,061	,000
cantidad	Welch	250,966	5	1593,061	,000
	Brown-Forsythe	237,343	5	896,460	,000
tenencia	Welch	5137,827	5	1629,337	,000
	Brown-Forsythe	4137,391	5	2149,094	,000

Fuente: Resultados Análisis SPSS- a Distribuidos en F asintóticamente.

Tabla 64: Prueba no paramétrica de Kruskal-Wallis

	saldo	monto	cantidad	tenencia
Chi-cuadrado	4344,642	4352,983	5621,540	8108,952
gl	5	5	5	5
Sig. asintót.	,000	,000	,000	,000

Fuente: Resultados Análisis SPSS- a Prueba de Kruskal-Wallis
b Variable de agrupación: grs

De las tablas 63 y 64 se observa que tanto los valores significativos del estadístico de Welch, Brown Rorsythe y Kruskal-Wallis permiten concluir que efectivamente existen diferencias de medias entre los niveles de la variable grs para todas las variables testeadas, si no existieran dichas diferencias en los promedio se podría afirmar que el efecto de cada nivel de grs sería el mismo para todas las poblaciones

definidas en la variable monto y como esto no es así se concluye que ambas variables comparadas están relacionadas.

(iii) Variable Renta

A pesar de que es sospechable que la variable renta se encuentre relacionada con las variables de interés se realizaron los test antes descritos para corroborar dicho supuesto.

Tabla 65: Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
monto	503,721	6	12981	,000
saldo	327,578	6	12981	,000
cantidad	954,892	6	12981	,000
tenencia	673,109	6	12981	,000

Fuente: Resultados Análisis SPSS

Tabla 66: Pruebas robustas de igualdad de las medias

		Estadístico(a)	gl1	gl2	Sig.
monto	Welch	87,498	6	1111,115	,000
	Brown-Forsythe	42,080	6	570,721	,000
saldo	Welch	82,698	6	1111,392	,000
	Brown-Forsythe	15,939	6	264,447	,000
cantidad	Welch	175,370	6	1111,305	,000
	Brown-Forsythe	73,616	6	650,738	,000
tenencia	Welch	1129,552	6	1117,968	,000
	Brown-Forsythe	309,394	6	1368,544	,000

a Distribuidos en F asintóticamente.

Tabla 67: Estadísticos de contraste(a,b)

	saldo	monto	cantidad	tenencia
Chi-cuadrado	3337,359	4147,385	3375,383	2741,381
gl	6	6	6	6
Sig. asintót.	,000	,000	,000	,000

a Prueba de Kruskal-Wallis

b Variable de agrupación: renta

Debido a que al no cumplimiento de igualdad de varianza (tabla 9) n los grupos formados por la variable renta, el nivel de significancia (p_valor<0,05) de los estadísticos Welch, Brown Rorsythe y Kruskal-Wallis presentados en las tablas 66 y 67 respectivamente permiten rechazar la hipótesis de igualdad de medias y concluir que si diferencias significativas en los promedios de las variables de interés entre los niveles de la variable renta.

(iv) Variable Región

Tal como en los casos anteriores, la variable región fue estudiada a modo de definir la existencia de una relación entre ésta y las características de interés.

Tabla 68: Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
monto	2,644	14	12973	,001
saldo	2,727	14	12973	,000
cantidad	5,229	14	12973	,000
tenencia	5,821	14	12973	,000

Fuente: Resultados Análisis SPSS

Tabla 69: Pruebas robustas de igualdad de las medias

		Estadístico(a)	gl1	gl2	Sig.
monto	Welch	3,093	14	5620,118	,000
	Brown-Forsythe	2,551	14	11734,632	,001
cantidad	Welch	8,292	14	5487,202	,000
	Brown-Forsythe	6,946	14	9200,597	,000
tenencia	Welch	12,651	14	5496,127	,000
	Brown-Forsythe	12,038	14	12934,521	,000

Fuente: Resultados Análisis SPSS - a Distribuidos en F asintóticamente.

Tabla 70: Estadísticos de contraste(a,b)

	saldo	monto	cantidad	tenencia
Chi-cuadrado	65,470	61,592	71,738	94,471
gl	14	14	14	14
Sig. asintót.	,000	,000	,000	,000

a Prueba de Kruskal-Wallis

b Variable de agrupación: region

Nuevamente los resultados obtenidos permiten concluir que la variable región está relacionada con las variables tenencia, monto, saldo y cantidad de transacciones.

ANEXO K: Análisis Post Hoc Variable Renta

(i) Segmento 1

a) Prueba de Homocedasticidad Variable Monto

Tabla 71: Prueba de homogeneidad de varianzas

Estadístico de Levene	gl1	gl2	Sig.
82,550	6	3719	,000

Fuente: Resultados Análisis SPSS

Los resultados expuestos en la tabla 71 muestran un nivel de significancia<0,05 para la prueba de Levene y por lo tanto se rechaza la hipótesis de igualdad de varianzas.

b) Contraste de Medias para Variable Monto de Transacciones

Tabla 72: Pruebas robustas de igualdad de las medias

	Estadístico(a)	gl1	gl2	Sig.
Welch	42,530	6	165,606	,000
Brown-Forsythe	3,453	6	38,279	,008

Fuente: Resultados Análisis SPSS - a Distribuidos en F asintóticamente.

Ambas pruebas robustas rechazan (p_valor <0,05) la igualdad de montos promedios entre los niveles de la variable renta.

Para conocer entre que niveles existe realmente tal diferencia y dado el incumplimiento de la igualdad de varianzas (tabla 1) es necesario realizar la prueba de Games-Howell (tabla 3), la cual indica la existencia de diferencias significativas entre todos aquellos pares de niveles con valor de significación < 0,05.

Tabla 73: Comparaciones múltiples – Prueba de Games-Howell

(I) renta	(J) renta	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite superior	Límite inferior
0	1	-43285757,262	17411935,441	,183	-96401903,75	9830389,23
	2	-6837814,285(*)	876900,110	,000	-9433293,26	-4242335,31
	3	-2133514,963(*)	230615,398	,000	-2813938,92	-1453091,00
	4	-10449938,549(*)	1128008,091	,000	-13804459,74	-7095417,36
	5	-70743465,690	45852353,823	,717	-221347190,45	79860259,07
	6	-13313028,281(*)	3000855,265	,001	-22545796,35	-4080260,21
1	0	43285757,262	17411935,441	,183	-9830389,23	96401903,75
	2	36447942,977	17433997,916	,371	-16726335,93	89622221,89
	3	41152242,299	17413457,813	,232	-11967914,22	94272398,82
	4	32835818,713	17448430,640	,500	-20376514,53	86048151,95
	5	-27457708,427	49047055,780	,997	-184568808,61	129653391,76
	6	29972728,982	17668629,314	,621	-23823879,22	83769337,19
2	0	6837814,285(*)	876900,110	,000	4242335,31	9433293,26
	1	-36447942,977	17433997,916	,371	-89622221,89	16726335,93
	3	4704299,322(*)	906626,160	,000	2022133,60	7386465,04
	4	-3612124,264	1428702,134	,151	-7840864,43	616615,91
	5	-63905651,405	45860736,351	,799	-214524961,97	86713659,16
	6	-6475213,995	3126326,895	,383	-16032397,17	3081969,18
3	0	2133514,963(*)	230615,398	,000	1453091,00	2813938,92
	1	-41152242,299	17413457,813	,232	-94272398,82	11967914,22
	2	-4704299,322(*)	906626,160	,000	-7386465,04	-2022133,60
	4	-8316423,586(*)	1151268,624	,000	-11737808,22	-4895038,96
	5	-68609950,727	45852931,949	,744	-219214750,16	81994848,70
	6	-11179513,318(*)	3009675,981	,009	-20434806,61	-1924220,02

4	0	10449938,549(*)	1128008,091	,000	7095417,36	13804459,74
	1	-32835818,713	17448430,640	,500	-86048151,95	20376514,53
	2	3612124,264	1428702,134	,151	-616615,91	7840864,43
	3	8316423,586(*)	1151268,624	,000	4895038,96	11737808,22
	5	-60293527,141	45866224,906	,837	-210923047,15	90335992,87
	6	-2863089,731	3205833,480	,972	-12630512,81	6904333,34
5	0	70743465,690	45852353,823	,717	-79860259,07	221347190,45
	1	27457708,427	49047055,780	,997	-129653391,76	184568808,61
	2	63905651,405	45860736,351	,799	-86713659,16	214524961,97
	3	68609950,727	45852931,949	,744	-81994848,70	219214750,16
	4	60293527,141	45866224,906	,837	-90335992,87	210923047,15
	6	57430437,409	45950444,146	,866	-93356258,93	208217133,75
6	0	13313028,281(*)	3000855,265	,001	4080260,21	22545796,35
	1	-29972728,982	17668629,314	,621	-83769337,19	23823879,22
	2	6475213,995	3126326,895	,383	-3081969,18	16032397,17
	3	11179513,318(*)	3009675,981	,009	1924220,02	20434806,61
	4	2863089,731	3205833,480	,972	-6904333,34	12630512,81
	5	-57430437,409	45950444,146	,866	-208217133,75	93356258,93

* La diferencia de medias es significativa al nivel .05.

(ii) Segmento 2

a) Prueba de Homocedasticidad Variable Monto

Tabla 74: Prueba de homogeneidad de varianzas

Estadístico de Levene	gl1	gl2	Sig.
112,065	2	23180	,000

Fuente: Resultados Análisis SPSS

Dado el nivel de significancia < 0,05 se rechaza la hipótesis de igualdad de varianzas.

b) Contraste de Medias para Variable Monto de Transacciones

Tabla 75: Pruebas robustas de igualdad de las medias

	Estadístico(a)	gl1	gl2	Sig.
Welch	484,656	2	10494,331	,000
Brown-Forsythe	262,493	2	12797,124	,000

Fuente: Resultados Análisis SPSS - a Distribuidos en F asintóticamente.

Se rechaza la igualdad de promedios en los montos de transacciones para las poblaciones formadas por los niveles de la variable renta pues p_valor < 0,05 en ambos estadísticos.

Tabla 76: Comparaciones múltiples – Prueba de Games-Howell

(I) renta	(J) renta	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite superior	Límite inferior
0	2	-3573271,976(*)	124864,234	,000	-3865981,03	-3280562,92

	3	-1223968,492(*)	86246,771	,000	-1426129,72	-1021807,27
2	0	3573271,976(*)	124864,234	,000	3280562,92	3865981,03
	3	2349303,484(*)	146826,301	,000	2005145,76	2693461,20
3	0	1223968,492(*)	86246,771	,000	1021807,27	1426129,72
	2	-2349303,484(*)	146826,301	,000	-2693461,20	-2005145,76

* La diferencia de medias es significativa al nivel .05.

Fuente: Resultados Análisis SPSS

La prueba de Games-Howell señala que para el segmento 2 existen diferencias significativas en los promedios de los montos de transacciones para todos los niveles de renta que lo definen.

(iii) Segmento 3

a) Prueba de Homocedasticidad Variable Monto

Tabla 77: Prueba de homogeneidad de varianzas

Estadístico de Levene	gl1	gl2	Sig.
79,392	6	6223	,000

b) Contraste de Medias para Variable Monto de Transacciones

Tabla 78: Pruebas robustas de igualdad de las medias

	Estadístico(a)	gl1	gl2	Sig.
Welch	45,398	6	40,908	,000
Brown-Forsythe	8,390	6	80,904	,000

a Distribuidos en F asintóticamente.

Fuente: Resultados Análisis SPSS

Los resultados de la tabla 78 muestran que la igualdad de medias de la variable monto de transacciones en las poblaciones definidas por las variable renta es rechazada por ambas pruebas.

A continuación se presenta la prueba de Games-Howell, la cual frente a la heterocedasticidad de la varianza (tabla 77) permite determinar entre cuales niveles de la variable testeada existen realmente diferencias de medias.

Tabla 79: Comparaciones múltiples
Variable dependiente: monto
Games-Howell

(I) renta	(J) renta	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite superior	Límite inferior
0	1	-7575142,027	5945493,625	,843	-32359250,78	17208966,73
	2	-2211923,488(*)	197326,550	,000	-2794997,67	-1628849,30
	3	-488597,987(*)	37580,415	,000	-599805,65	-377390,33
	4	-7048484,073(*)	2099304,350	,017	-13330221,84	-766746,30
	5	-9458776,617	3511792,348	,227	-23378486,31	4460933,08
	6	-9307851,675	3499562,525	,180	-21257204,19	2641500,84
1	0	7575142,027	5945493,625	,843	-17208966,73	32359250,78
	2	5363218,539	5948647,172	,959	-19415454,20	30141891,28
	3	7086544,040	5945492,217	,875	-17697567,15	31870655,23

	4	526657,954	6305120,472	1,000	-23893810,30	24947126,21
	5	-1883634,589	6905081,528	1,000	-26963548,85	23196279,67
	6	-1732709,648	6898869,710	1,000	-26556390,81	23090971,51
2	0	2211923,488(*)	197326,550	,000	1628849,30	2794997,67
	1	-5363218,539	5948647,172	,959	-30141891,28	19415454,20
	3	1723325,501(*)	197284,128	,000	1140378,23	2306272,77
	4	-4836560,585	2108219,033	,254	-11143404,32	1470283,15
	5	-7246853,128	3517128,692	,456	-21162321,17	6668614,91
	6	-7095928,187	3504917,489	,442	-19052783,15	4860926,78
3	0	488597,987(*)	37580,415	,000	377390,33	599805,65
	1	-7086544,040	5945492,217	,875	-31870655,23	17697567,15
	2	-1723325,501(*)	197284,128	,000	-2306272,77	-1140378,23
	4	-6559886,086(*)	2099300,362	,034	-12841612,63	-278159,54
	5	-8970178,630	3511789,964	,267	-22889890,26	4949533,00
	6	-8819253,688	3499560,133	,224	-20768602,87	3130095,49
4	0	7048484,073(*)	2099304,350	,017	766746,30	13330221,84
	1	-526657,954	6305120,472	1,000	-24947126,21	23893810,30
	2	4836560,585	2108219,033	,254	-1470283,15	11143404,32
	3	6559886,086(*)	2099300,362	,034	278159,54	12841612,63
	5	-2410292,544	4091251,058	,996	-16573016,18	11752431,09
	6	-2259367,602	4080758,213	,998	-15295447,70	10776712,49
5	0	9458776,617	3511792,348	,227	-4460933,08	23378486,31
	1	1883634,589	6905081,528	1,000	-23196279,67	26963548,85
	2	7246853,128	3517128,692	,456	-6668614,91	21162321,17
	3	8970178,630	3511789,964	,267	-4949533,00	22889890,26
	4	2410292,544	4091251,058	,996	-11752431,09	16573016,18
	6	150924,942	4957639,996	1,000	-16168425,93	16470275,81
6	0	9307851,675	3499562,525	,180	-2641500,84	21257204,19
	1	1732709,648	6898869,710	1,000	-23090971,51	26556390,81
	2	7095928,187	3504917,489	,442	-4860926,78	19052783,15
	3	8819253,688	3499560,133	,224	-3130095,49	20768602,87
	4	2259367,602	4080758,213	,998	-10776712,49	15295447,70
	5	-150924,942	4957639,996	1,000	-16470275,81	16168425,93

* La diferencia de medias es significativa al nivel .05.

Fuente: Resultados Análisis SPSS

(iv) Segmento 5

c) Prueba de Homocedasticidad Variable Monto

Tabla 80: Prueba de homogeneidad de varianzas

Estadístico de Levene	gl1	gl2	Sig.
57,640	2	4463	,000

Fuente: Resultados Análisis SPSS

Nivel de significancia < 0,05 se rechaza la hipótesis de igualdad de varianzas.

d) Contraste de Medias para Variable Monto de Transacciones

Tabla 81: Pruebas robustas de igualdad de las medias

	Estadístico(a)	gl1	gl2	Sig.
Welch	18,527	2	1037,135	,000
Brown-Forsythe	15,169	2	677,105	,000

Fuente: Resultados Análisis SPSS - a Distribuidos en F asintóticamente.

Tabla 82: Prueba de Games-Howell para comparaciones múltiples

(I) renta	(J) renta	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite superior	Límite inferior
4	5	-33636852,966(*)	13950373,036	,044	-66609421,22	-664284,71
	6	-3366609,270(*)	1167190,234	,011	-6112235,51	-620983,03
5	4	33636852,966(*)	13950373,036	,044	664284,71	66609421,22
	6	30270243,696	13986528,809	,080	-2784905,53	63325392,92
6	4	3366609,270(*)	1167190,234	,011	620983,03	6112235,51
	5	-30270243,696	13986528,809	,080	-63325392,92	2784905,53

* La diferencia de medias es significativa al nivel .05.

Fuente: Resultados Análisis SPSS

La prueba de Games-Howell señala que para el segmento 5 existen diferencias significativas en los promedios de los montos de transacciones para todas la combinaciones del tramo 4 (nivel 4 y nivel 5 – nivel 4 y nivel 6).

ANEXO L: Análisis Post Hoc Variable Región

Como la condición de homocedasticidad de las poblaciones definidas por los niveles de la variable región no se cumple (ver prueba de Levene en tabla 68 del anexo J), fue necesario utilizar el test de Games-Howell para contrastar la existencia de diferencias significativas de los promedios entre pares de regiones para las variables monto, cantidad y tenencia.

Debido a la magnitud de la tabla de resultados sólo serán expuestas aquellas combinaciones en que se encontraron diferencias, es decir aquellas que presentaron un nivel de significación menor a 0,05 en la prueba y que son los pares considerados como restricciones para la agrupación de regiones.

Tabla 83: Diferencias significativas para comparaciones múltiples de la variable región según prueba de Games-Howell

Variable dependiente	(I) region	(J) region	Diferencia de medias (I-J)	Sig.
monto	2	7	197090,032	0,0236
monto	2	9	234655,6772	0,0015
monto	3	6	243559,9099	0,0260
monto	3	7	270982,5098	0,0048
monto	3	9	308548,1551	0,0004
monto	4	6	243480,2317	0,0003
monto	4	7	270902,8316	0,0000
monto	4	8	174997,9143	0,0154
monto	4	9	308468,4769	0,0000
monto	5	6	251015,5627	0,0000
monto	5	7	278438,1626	0,0000
monto	5	8	182533,2454	0,0000
monto	5	9	316003,8079	0,0000

monto	5	13	103353,5928	0,0338
monto	6	13	-147661,9699	0,0015
monto	7	10	-175125,1824	0,0193
monto	7	13	-175084,5698	0,0000
monto	8	9	133470,5625	0,0233
monto	9	13	-212650,2151	0,0000
cantidad	2	5	-3,780691178	0,0437
cantidad	3	6	4,379251218	0,0217
cantidad	3	7	4,180861363	0,0363
cantidad	4	6	5,713898333	0,0000
cantidad	4	7	5,515508478	0,0000
cantidad	4	8	4,189362524	0,0022
cantidad	4	9	4,973639308	0,0003
cantidad	5	6	6,436131511	0,0000
cantidad	5	7	6,237741656	0,0000
cantidad	5	8	4,911595701	0,0000
cantidad	5	9	5,695872486	0,0000
cantidad	5	10	3,342389845	0,0359
cantidad	5	13	3,793327654	0,0000
cantidad	6	13	-2,642803857	0,0027
cantidad	6	14	-5,310066741	0,0028
cantidad	7	13	-2,444414002	0,0076
cantidad	7	14	-5,111676886	0,0050
tenencia	1	9	0,225018255	0,0267
tenencia	2	9	0,223944138	0,0017
tenencia	3	9	0,313183862	0,0001
tenencia	4	9	0,276151655	0,0000
tenencia	5	8	0,113988964	0,0283
tenencia	5	9	0,252443081	0,0000
tenencia	5	13	-0,10979398	0,0050
tenencia	6	9	0,177357742	0,0017
tenencia	6	13	-0,184879319	0,0000
tenencia	7	9	0,138073046	0,0314
tenencia	7	13	-0,224164015	0,0000
tenencia	8	9	0,138454117	0,0034
tenencia	8	13	-0,223782944	0,0000
tenencia	9	13	-0,362237061	0,0000
tenencia	10	13	-0,224353504	0,0000
tenencia	12	13	-0,309945341	0,0003
tenencia	13	14	0,224923813	0,0005
tenencia	13	15	0,241739187	0,0053

* La diferencia de medias es significativa al nivel .05.

Fuente: Resultados Análisis SPSS

ANEXO M: Procedimientos Almacenados para Extracción de Muestra

a) Procedimiento Cuenta Población

A modo de ejemplo se presenta el código que crea la tabla y llena la primera fila de la figura 3 descrita en el capítulo 6:

```
/*PRIMERO COPIAN Y CORREN ESTAS 4 LÍNEAS ES PARA BORRAR EL PROCEDIMIENTO Y LA TABLA SI
EXISTIERAN EN LA BASE POR EJ AL MANTENER EL PANEL PARA CONSIDERAR NUEVOS CLIENTES*/

IF OBJECT_ID ('TABLA_CUENTA_POBLACION') IS NOT NULL DROP TABLE TABLA_CUENTA_POBLACION
IF OBJECT_ID ('CUENTA_POBLACION') IS NOT NULL DROP PROCEDURE CUENTA_POBLACION

/* AHORA SE CORRE EL PROCEDIMIENTO COMPLETO*/
CREATE PROCEDURE CUENTA_POBLACION AS
SELECT T1.GRS, T3.ZONA, SUM (CASE WHEN T1.SEGMENTO=1 AND  T1.RENTA=3 THEN 1 ELSE 0 END) S1_R3
,SUM (CASE WHEN T1.SEGMENTO=1 AND  (T1.RENTA=2 OR T1.RENTA=4 OR T1.RENTA=6 OR T1.RENTA=5 OR
T1.RENTA=1) THEN 1 ELSE 0 END) S1_ARENTA ,SUM (CASE WHEN T1.SEGMENTO=1 AND  T1.RENTA=0 THEN 1
ELSE 0 END) S1_R0,SUM (CASE WHEN T1.SEGMENTO=2 AND  T1.RENTA=3 THEN 1 ELSE 0 END) S2_R3,SUM
(CASE WHEN T1.SEGMENTO=2 AND  T1.RENTA=2 THEN 1 ELSE 0 END) S2_ARENTA,SUM (CASE WHEN
T1.SEGMENTO=2 AND  T1.RENTA=0 THEN 1 ELSE 0 END) S2_A0,SUM (CASE WHEN T1.SEGMENTO=3 AND
(RENTA=3) THEN 1 ELSE 0 END) S3_R3,SUM (CASE WHEN T1.SEGMENTO=3 AND  (RENTA=2 OR T1.RENTA=4 OR
T1.RENTA=6 OR T1.RENTA=5OR T1.RENTA=1) THEN 1 ELSE 0 END) S3_ARENTA ,SUM (CASE WHEN
T1.SEGMENTO=3 AND  (RENTA=0) THEN 1 ELSE 0 END) S3_R0,SUM (CASE WHEN T1.SEGMENTO=4 THEN 1 ELSE
0 END) S4_,SUM (CASE WHEN T1.SEGMENTO=5 AND  T1.RENTA=4 THEN 1 ELSE 0 END) S5_R4,SUM (CASE
WHEN T1.SEGMENTO=5 AND  (T1.RENTA=6OR T1.RENTA=5) THEN 1 ELSE 0 END)S5_ARENTA,SUM (CASE WHEN
T1.SEGMENTO=6  THEN 1 ELSE 0 END) S6 INTO TABLA_CUENTA_POBLACION
FROM TABLA_MARCO_MUESTRAL_3 AS T1, TABLA_REGION_ZONA AS T3
WHERE T1.REGION=T3.REGION AND
(T1.REGION=1 OR T1.REGION=2 OR T1.REGION=3 OR T1.REGION=4 OR T1.REGION=15) AND T1.GRS=1
GROUP BY GRS,T3.ZONA ORDER BY GRS,ZONA

EXEC CUENTA_POBLACION

SELECT * FROM TABLA_CUENTA_POBLACION
ORDER BY ZONA, GRS
```

En el caso de realizar la cuantificación de clientes de la muestra en los estratos la consulta deberá ser realizada utilizando el rutero de los clientes del panel agregando dicha tabla en la cláusula *from*.

b) Procedimiento Extracción Muestral

A continuación se presenta la programación realizada para la extracción aleatoria y el cálculo de ponderadores para los tres primeros estratos del segmento S1:

```
IF OBJECT_ID ('RUTERO_PANEL') IS NOT NULL DROP TABLE RUTERO_PANEL
/*S1*/
SELECT TOP 150 T1.RUT_DEMO, T3.ADU_MENOR200*49770.0/(150*8297628) PESO
INTO RUTERO_PANEL
FROM TABLA_MARCO_MUESTRAL AS T1, TABLA_CUENTA_POBLACION AS T3
WHERE  T1.GRS=T3.GRS  AND    T1.SEGMENTO='S1'  AND  T1.RENTA='001-199'  AND  (T1.REGION=1  OR
T1.REGION=2 OR T1.REGION=3 OR T1.REGION=4 OR T1.REGION=15) AND T3.ZONA=1 AND T1.GRS='G3' GROUP
BY T1.RUT_DEMO, T3.ADU_MENOR200
ORDER BY NEWID ()

INSERT INTO RUTERO_PANEL
SELECT TOP 150 T1.RUT_DEMO, T3.ADU_MENOR200*49770.0/(150*8297628) PESO
FROM TABLA_MARCO_MUESTRAL AS T1, TABLA_CUENTA_POBLACION AS T3
WHERE  T1.GRS=T3.GRS  AND    T1.SEGMENTO='S1'  AND    T1.RENTA='001-199'  AND  (T1.REGION=1  OR
T1.REGION=2 OR T1.REGION=3 OR T1.REGION=4 OR T1.REGION=15) AND T3.ZONA=1 AND T1.GRS='G5'
GROUP BY T1.RUT_DEMO, T3.ADU_MENOR200 ORDER BY NEWID ()

INSERT INTO RUTERO_PANEL
SELECT TOP 150 T1.RUT_DEMO, T3.ADU_MAYOR200*49770.0/(150*8297628)
FROM TABLA_MARCO_MUESTRAL AS T1, TABLA_CUENTA_POBLACION AS T3
```

```
WHERE T1.GRS=T3.GRS AND T1.SEGMENTO='S1' AND (T1.RENTA='200-399' OR T1.RENTA='400-599' OR
T1.RENTA='600-799' OR T1.RENTA='800-999' OR T1.RENTA='1MM-MAS') AND (T1.REGION=1 OR
T1.REGION=2 OR T1.REGION=3 OR T1.REGION=4 OR T1.REGION=15) AND T3.ZONA=1 AND T1.GRS='G1'
GROUP BY T1.RUT_DEMO, T3.ADU_MAYOR200
ORDER BY NEWID ()
```

ANEXO N: Procedimiento para Cálculo Base de Comportamiento

```
CREATE PROCEDURE BASE_COMPORTAMIENTO
@AAAAMM NVARCHAR(6)
AS
BEGIN
/*SE OBTIENE INFORMACION DEMOGRÁFICA DE CADA CLIENTE*/
SELECT T1.RUT_DEMO RUT1,@AAAAMM TIEMPO1,T1.PESO, T2.SEGMENTO,T2.RENTA,T2.EDAD, T2.GRS,
T2.SEXO,T2.REGION,T2.ANTIGUEDAD,T3.ZONA INTO #T_1
FROM RUTERO_PANEL AS T1, TABLA_MARCO_MUESTRAL AS T2, TABLA_REGION_ZONA AS T3
WHERE T1.RUT_DEMO = T2.RUT_DEMO AND T2.REGION=T3.REGION
/*SE CALCULA EL MONTO Y LA CANTIDAD DE TRANSACCIONES FINANCIERAS VOLUNTARIAS PARA CADA CLIENTE
EN EL PERIODO EN ESTUDIO*/
SELECT DISTINCT CST_ID RUT2,@AAAAMM TIEMPO2,SUM(AMNT_LOC) MONTO_TXN_MENSUAL,SUM(CNT)
CANT_TXN_MENSUAL INTO #T_2
FROM RUTERO_PANEL AS T1, TABLA_TXN AS T2, TABLA_CLASIFICACION_TRANSACCIONES AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND VOLUNTARIA = 1 AND AMNT_LOC
>0 AND TIEMPO = @AAAAMM
GROUP BY CST_ID
INSERT INTO #T_2
SELECT RUT_DEMO,@AAAAMM,0,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT2 FROM #T_2)
/*SE CALCULA EL SALDO PROMEDIO DIARIO PROMEDIO PARA EL PERIODO EN ESTUDIO PARA PRODUCTOS
DISTINTOS A CRÉDITOS */
SELECT DISTINCT CST_ID RUT3,@AAAAMM TIEMPO3,SUM(IMS_SPM_DCM) SALDO INTO #T_3
FROM RUTERO_PANEL AS T1, TABLA_SMY_PANEL AS T2, TABLA_PRODUCTOS_AGRUPADOS3 AS T3
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.PD_ID = T3.PD_ID AND (T3.PROD_SBIF = 'MEDIO_PAGO' OR
PROD_MATRIZ IN ('DAP','CTA_CTE','CTA_VISTA','VALE_VISTA','AHORRO_PLAZO')) AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #T_3
SELECT RUT_DEMO,@AAAAMM,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT3 FROM #T_3)
/*SE ADJUNTA LA VARIABLE TENENCIA POR PRODUCTOS */
SELECT T2.*,@AAAAMM TIEMPO4 INTO #T_4
FROM TABLA_TENENCIA_VIGENTES2 AS T2, RUTERO_PANEL AS T1
WHERE T2.RUT_PRODS=T1.RUT_DEMO
/*CALCULO DEL TOTAL DE TRANSACCIONES VOLUNTARIAS EN LOS CANALES MAQUINAS, SUCURSALES E
INTERNET*/
SELECT DISTINCT CST_ID RUT5,@AAAAMM TIEMPO5,SUM(CNT) CANT_TXN INTO #T_5_I
FROM RUTERO_PANEL AS T1, TABLA_TXN AS T2,CANALES_OBSERVATORIO2 AS T3 ,
TABLA_CLASIFICACION_TRANSACCIONES AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
(CLASI_CANAL=1 OR CLASI_CANAL=2 OR CLASI_CANAL=3) AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #T_5_I
SELECT RUT_DEMO,@AAAAMM,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT5 FROM #T_5_I)
/*MAQUINAS*/
SELECT DISTINCT CST_ID,SUM(CNT) CANT_TXN_MAQ
INTO #T_5_1
FROM RUTERO_PANEL AS T1, TABLA_TXN AS T2,CANALES_OBSERVATORIO2 AS T3 ,
TABLA_CLASIFICACION_TRANSACCIONES AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=1 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #T_5_1
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #T_5_1)
/*SUCURSAL*/
SELECT DISTINCT CST_ID,SUM(CNT) CANT_TXN_SUC
INTO #T_5_2
FROM RUTERO_PANEL AS T1, TABLA_TXN AS T2,CANALES_OBSERVATORIO2 AS T3 ,
TABLA_CLASIFICACION_TRANSACCIONES AS T4
```

```
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=2 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #T_5_2
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #T_5_2)
/*INTERNET*/
SELECT DISTINCT CST_ID,SUM(CNT) CANT_TXN_INT
INTO #T_5_3
FROM   RUTERO_PANEL   AS   T1,   TABLA_TXN   AS   T2,CANALES_OBSERVATORIO2      AS   T3   ,
TABLA_CLASIFICACION_TRANSACCIONES  AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=3 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #T_5_3
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #T_5_3)
/*SE JUNTA EL TOTAL Y LA CANTIDAD DE TRANSACCIONES POR CADA CANAL*/
SELECT T1.RUT5,@AAAAMM TIEMPO5,CANT_TXN_MAQ,CANT_TXN_SUC,CANT_TXN_INT,CANT_TXN INTO #T_5
FROM #T_5_I AS T1, #T_5_1 AS T2, #T_5_2 AS T3, #T_5_3 AS T4
WHERE T1.RUT5 = T2.CST_ID AND T2.CST_ID = T3.CST_ID AND T3.CST_ID = T4.CST_ID
/*MONTO*/
/*SE CALCULA EL MONTO DE TRANSACCIONES VOUNTARIAS POR CANAL*/
/*TOTAL*/
SELECT DISTINCT CST_ID RUT6,SUM(AMNT_LOC) MNT_TXN INTO #M_6_I
FROM   RUTERO_PANEL   AS   T1,   TABLA_TXN   AS   T2,CANALES_OBSERVATORIO2      AS   T3   ,
TABLA_CLASIFICACION_TRANSACCIONES  AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
(CLASI_CANAL=1 OR CLASI_CANAL=2 OR CLASI_CANAL=3) AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #M_6_I
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT6 FROM #M_6_I)
/*MAQUINAS*/
SELECT DISTINCT CST_ID,SUM(AMNT_LOC) MNT_TXN_MAQ
INTO #M_6_1
FROM   RUTERO_PANEL   AS   T1,   TABLA_TXN   AS   T2,CANALES_OBSERVATORIO2      AS   T3   ,
TABLA_CLASIFICACION_TRANSACCIONES  AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=1 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #M_6_1
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #M_6_1)
/*SUCURSAL*/
SELECT DISTINCT CST_ID,SUM(AMNT_LOC) MNT_TXN_SUC
INTO #M_6_2
FROM   RUTERO_PANEL   AS   T1,   TABLA_TXN   AS   T2,CANALES_OBSERVATORIO2      AS   T3   ,
TABLA_CLASIFICACION_TRANSACCIONES  AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=2 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #M_6_2
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #M_6_2)
/*INTERNET*/
SELECT DISTINCT CST_ID,SUM(AMNT_LOC) MNT_TXN_INT
INTO #M_6_3
FROM   RUTERO_PANEL   AS   T1,   TABLA_TXN   AS   T2,CANALES_OBSERVATORIO2      AS   T3   ,
TABLA_CLASIFICACION_TRANSACCIONES  AS T4
WHERE T1.RUT_DEMO = T2.CST_ID AND T2.CNL_ID = T3.CNL_ID AND T2.TXN_TP_ID = T4.TXN_TP_ID AND
CLASI_CANAL=3 AND T4.VOLUNTARIA = 1 AND TIEMPO=@AAAAMM
GROUP BY CST_ID
INSERT INTO #M_6_3
SELECT RUT_DEMO,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT CST_ID FROM #M_6_3)
/*SE JUNTA EL TOTAL Y LA CANTIDAD DE TRANSACCIONES POR CADA CANAL*/
SELECT T1.RUT6,@AAAAMM TIEMPO6,MNT_TXN_MAQ,MNT_TXN_SUC,MNT_TXN_INT,MNT_TXN INTO #T_6
FROM #M_6_I AS T1, #M_6_1 AS T2, #M_6_2 AS T3, #M_6_3 AS T4
WHERE T1.RUT6 = T2.CST_ID AND T2.CST_ID = T3.CST_ID AND T3.CST_ID = T4.CST_ID
/*SE OBTIENE LA DEUDA PROMEDIO DEL CLIENTE CON LA ENTIDAD EN EL PERIODO EN ESTUDIO*/
SELECT DISTINCT SRU_RUT_NRT RUT7, @AAAAMM TIEMPO7,SUM(SRU_DEU_DCT_VGT) DEUDA_BE
```

```
INTO #T_7
FROM RUTERO_PANEL AS T1, SBIF AS T2
WHERE T1.RUT_DEMO = T2.SRU_RUT_NRT AND T2.TIEMPO=@AAAAMM AND T2.SRU_MDD_RST=12
GROUP BY SRU_RUT_NRT
INSERT INTO #T_7
SELECT RUT_DEMO,@AAAAMM,0
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT7 FROM #T_7)

/*SE OBTIENE LA DEUDA PROMEDIO TOTAL DEL CLIENTE CON TODAS LAS ENTIDADES FINANCIERAS EN EL
PERIODO EN ESTUDIO*/
SELECT DISTINCT SRU_RUT_NRT RUT8,@AAAAMM TIEMPO8,SUM(SRU_DEU_DCT_VGT) DEUDA_TOT
INTO #T_8
FROM RUTERO_PANEL AS T1, SBIF AS T2
WHERE T1.RUT_DEMO = T2.SRU_RUT_NRT AND T2.TIEMPO=@AAAAMM
GROUP BY SRU_RUT_NRT
INSERT INTO #T_8
SELECT RUT_DEMO,@AAAAMM,1
FROM RUTERO_PANEL
WHERE RUT_DEMO NOT IN (SELECT RUT8 FROM #T_8)
/*SE JUNTAN TODAS LAS VARIABLES DE COMPORTAMIENTO EN UNA SOLA TABLA*/
INSERT INTO BASE_COMPORTAMIENTO_MENSUAL
SELECT     T1.*,     T2.MONTO_TXN_MENSUAL     ,     T2.CANT_TXN_MENSUAL,     SALDO,     T4.*,
CANT_TXN_MAQ,CANT_TXN_SUC,CANT_TXN_INT,CANT_TXN,MNT_TXN_MAQ,MNT_TXN_SUC,MNT_TXN_INT,MNT_TXN,DE
UDA_BE,DEUDA_TOT
FROM #T_1 AS T1, #T_2 AS T2, #T_3 AS T3, #T_4 AS T4, #T_5 AS T5, #T_6 AS T6, #T_7 AS T7, #T_8
AS T8
WHERE T1.RUT1= T2.RUT2 AND T2.RUT2 = T3.RUT3 AND T3.RUT3 = T4.RUT_PRODS AND  T4.RUT_PRODS =
T5.RUT5  AND  T5.RUT5  =  T6.RUT6  AND  T6.RUT6  =  T7.RUT7  AND  T7.RUT7  =  T8.RUT8  AND
T1.TIEMPO1=T2.TIEMPO2 AND T2.TIEMPO2=T3.TIEMPO3
AND   T3.TIEMPO3=T4.TIEMPO4   AND   T4.TIEMPO4=T5.TIEMPO5   AND   T5.TIEMPO5=T6.TIEMPO6   AND
T6.TIEMPO6=T7.TIEMPO7 AND T7.TIEMPO7=T8.TIEMPO8
END
```

ANEXO O: Validación de Resultados

Tamaño de Muestra y Distribución

Tabla 84: Validación Nivel de Inferencia Segmento - Renta

GRSXRenta	Total	%	N Teórico	N Empírico	Total	%	Diferencial
G1-menos400	96610	1,16	1896	2886	579,4	1,16	0
G1-mas400	72699	0,88	1883	3644	436,2	0,88	0
G2-menos400	119701	1,44	1903	1576	715,7	1,44	0
G2-mas400	53262	0,64	1866	3194	626	1,26	-0,62
G3-menos200	3E+06	35,1	1932	2880	17459	35,08	0
G3-mas 200	823663	9,93	1929	4762	4940	9,93	0
G3-sin dato renta	1E+06	17	1931	3550	8472	17,02	0
G4-menos200	83083	1	1889	1830	498,3	1	0
G4-mas de 200	107401	1,29	1899	4580	644,2	1,29	0
G4-menos 400	146441	1,76	1908	4740	878	1,76	0
G4-mas400	44043	0,53	1852	2160	264,6	0,53	0
G5-menos200	1E+06	15,8	1931	3300	7871	15,82	0
G5-mas200	722638	8,71	1928	6130	4334	8,71	0
G5-sin dato renta	284286	3,43	1920	3650	1705	3,43	0
G6-menos400	203556	2,45	1915	3544	1218	2,45	0
G6-mas400	95339	1,15	1895	3046	574,4	1,15	0

Fuente: Elaboración Propia

Tabla 85: Validación Nivel de Inferencia Segmento - Género

Segmento-Género	Total	%	N Teórico	N Empírico	Total	%	Diferencial
S1-V1	427102	5,15	1925	3134	2524	5,07	0,08
S1-V2	281719	3,4	1920	2899	1704	3,42	-0,02
S2-V1	3E+06	31,6	1932	6328	15735	31,62	-0,05
S2-V2	2E+06	21	1931	5798	10444	20,99	0,01
S3-V1	659071	7,94	1928	5109	3917	7,87	0,07
S3-V2	608748	7,34	1927	6006	3677	7,39	-0,05
S5-V1	304661	3,67	1921	4300	1822	3,66	0,01
S5-V2	453527	5,47	1925	7112	2721	5,47	0
S6-V1	43428	0,52	1851	1792	268,6	0,54	-0,02
S6-V2	94489	1,14	1895	4330	558,1	1,12	0,02

Fuente: Elaboración Propia

Tabla 86: Validación Nivel de Inferencia Segmento – GRS - Renta

Segmento-GRS-Renta	Total	%	N Teórico	N Empírico	Total	%	Diferencial
S2-G5-R0+R3	1E+06	12,4	1933	1920	6160	12,38	0
S2-G5-Renta_otros	365218	4,4	1084	1080	2191	4,4	0

Fuente: Elaboración Propia

Tabla 87: Validación Nivel de Inferencia Segmento - Zona

SegmentoXZona	Total	%	N Teórico	N Empírico	Total	%	Diferencial
S1*-Z1	74873	0,9	1072	1050	449	0,9	0
S1*-Z2	93802	1,13	1075	1050	563	1,13	0
S1*-Z3	76188	0,92	1072	1050	457	0,92	0
S1*-Z4	81199	0,98	1073	1050	487	0,98	0
S1*-Z5	94306	1,14	1075	1050	566	1,14	0
S1*-Z6	297394	3,58	1084	1050	1784	3,58	0
S2-Z1	526856	6,35	1926	2030	3160	6,35	0
S2-Z2	475117	5,73	1926	2030	2850	5,73	0
S2-Z3	514868	6,21	1926	2030	3088	6,21	0
S2-Z4	532215	6,41	1926	2030	3192	6,41	0
S2-Z5	650444	7,84	1928	2030	3901	7,84	0
S2-Z6	2E+06	20,9	1931	2160	10399	20,9	0
S3-Z1	174752	2,11	1912	1910	1048	2,11	0
S3-Z2	141858	1,71	1907	1910	851	1,71	0
S3-Z3	145268	1,75	1908	1910	871	1,75	0
S3-Z4	155546	1,87	1910	1910	933	1,87	0
S3-Z5	205018	2,47	1915	1910	1230	2,47	0
S3-Z6	513139	6,18	1926	2030	3078	6,18	0
S5-Z1	103492	1,25	1898	1890	621	1,25	0
S5-Z2	75230	0,91	1885	1890	451	0,91	0
S5-Z3	60961	0,73	1874	1890	366	0,73	0
S5-Z4	76196	0,92	1886	1890	457	0,92	0
S5-Z5	93181	1,12	1894	1890	559	1,12	0
S5-Z6	353584	4,26	1923	2010	2121	4,26	0
S6-*Z1	16590	0,2	1021	900	100	0,2	0
S6-*Z2	13090	0,16	1004	900	79	0,16	0
S6-*Z3	8951	0,11	970	900	54	0,11	0
S6-*Z4	9882	0,12	980	900	59	0,12	0
S6-*Z5	14561	0,18	1012	900	87	0,18	0
S6-*Z6	75571	0,91	1072	1650	453	0,91	0

*Tamaño calculado al 3,5% y 95% de confianza

Fuente: Elaboración Propia

A continuación se presentan los resultados de la validación distribuciones de tenencia de productos para la variable GRS:

Tabla 88: Parámetros del Cálculo de N

Parámetro	Var	Valor
Chi-Cuadrada	B	8,8
Precisión	b	2,0%
Nivel de Confianza	alpha	5,0%
Peor Caso	Pi	50,0%
Categorías	k	17,0

Fuente: Elaboración Propia

Tabla 89: Nivel de Inferencia GRS

Variable 17 categorias	N Teórico*	N Empírico
GRS1	5353	6530
GRS2	5356	4770
GRS3	5521	12320
GRS4	5371	6410
GRS5	5514	13080
GRS6	5427	6590

*Error al 2%, 95% de confianza

Fuente: Elaboración Propia

Tabla 90: Estimaciones de Tenencia por GRS – G1 a G3

G1	% Población	% Panel	Error	G2	% Población	% Panel	Error	G3	% Población	% Panel	Error
P1	12,91	12,97	-0,06	**P1**	9,62	9,66	-0,04	**P1**	1,46	1,57	-0,11
P2	3,11	3,04	0,07	**P2**	1,55	1,49	0,06	**P2**	0	0	0
P3	25,36	25,62	-0,26	**P3**	11,05	11,22	-0,17	**P3**	93,44	93,28	0,16
P4	0	0	0	**P4**	9,55	9,58	-0,03	**P4**	0	0	0
P5	0	0	0	**P5**	1,51	1,5	0,01	**P5**	0	0	0
P6	10,76	10,72	0,04	**P6**	7,84	7,77	0,07	**P6**	0	0	0
P7	9,13	9,36	-0,23	**P7**	3,87	3,65	0,22	**P7**	0	0	0
P8	2,88	2,89	-0,01	**P8**	4,5	4,46	0,04	**P8**	0	0	0
P9	2,64	2,48	0,16	**P9**	1,11	1,09	0,02	**P9**	1,17	1,38	-0,21
P10	0	0	0	**P10**	5,43	5,44	-0,01	**P10**	0	0	0
P11	3,27	3,13	0,14	**P11**	5,67	5,7	-0,03	**P11**	0	0	0
P12	0,35	0,27	0,08	**P12**	1,65	1,63	0,02	**P12**	0	0	0
P13	1,84	1,74	0,1	**P13**	12,61	12,67	-0,06	**P13**	0,82	0,89	-0,07
P14	7,78	7,52	0,26	**P14**	12,8	12,81	-0,01	**P14**	3,11	2,87	0,24
P15	3,48	3,53	-0,05	**P15**	3,39	3,43	-0,04	**P15**	0	0	0
P16	7,49	7,6	-0,11	**P16**	7,48	7,53	-0,05	**P16**	0	0	0
P17	9,02	9,12	-0,1	**P17**	0,39	0,37	0,02	**P17**	0	0	0

Fuente: Elaboración Propia

Tabla 91: Estimaciones de Tenencia por GRS – G4 a G6

G4	% Población	% Panel	Error	G5	% Población	% Panel	Error	G6	% Población	% Panel	Error
P1	1,85	2,33	-0,48	P1	4,39	4,43	-0,04	P1	4,47	4,58	-0,11
P2	0	0	0	P2	11,44	11,6	-0,16	P2	3,49	3,43	0,06
P3	30,53	30,7	-0,17	P3	37,64	37,16	0,48	P3	16,42	16,8	-0,38
P4	3,62	3,49	0,13	P4	0	0	0	P4	6,82	6,79	0,03
P5	2,85	3,07	-0,22	P5	0	0	0	P5	2,23	2,2	0,03
P6	0	0	0	P6	0	0	0	P6	0	0	0
P7	0	0	0	P7	32,5	32,17	0,33	P7	10,61	10,57	0,04
P8	0	0	0	P8	7,52	7,72	-0,2	P8	8,74	8,77	-0,03
P9	0,5	0,5	0	P9	0,98	0,87	0,11	P9	0,63	0,61	0,02
P10	18,64	18,23	0,41	P10	0	0	0	P10	11,21	11,04	0,17
P11	0	0	0	P11	0	0	0	P11	0	0	0
P12	0	0	0	P12	0,05	0,05	0	P12	1,06	1,04	0,02
P13	21,41	21,36	0,05	P13	1,26	1,4	-0,14	P13	17,49	17,44	0,05
P14	20,6	20,31	0,29	P14	4,23	4,6	-0,37	P14	16,83	16,73	0,1
P15	0	0	0	P15	0	0	0	P15	0	0	0
P16	0	0	0	P16	0	0	0	P16	0	0	0
P17	0	0	0	P17	0	0	0	P17	0	0	0

Fuente: Elaboración Propia

A continuación se presentan los resultados de la validación distribuciones de tenencia de productos para la variable zona:

Tabla 92: Cálculo de N para Estimación de Tenencia por Producto a Nivel Zona

Variable 17 categorias	N Teórico*	N Empírico
Zona1	5497	8090
Zona2	5493	8090
Zona3	5494	8090
Zona4	5496	8090
Zona5	5502	8090
Zona6	5518	9250

*Error al 2%, 95% de confianza

Fuente: Elaboración Propia

Tabla 93: Estimaciones de Tenencia por Zona – Z1 a Z3

Z1	% Población	% Panel	Error	Z2	% Población	% Panel	Error	Z3	% Población	% Panel	Error
P1	4,34	4,86	-0,52	P1	4,41	4,17	0,24	P1	3,28	3,47	-0,19
P2	4,93	5,21	-0,28	P2	4,59	4,66	-0,07	P2	4,84	4,74	0,1
P3	55,84	55,09	0,75	P3	55,69	55,97	-0,28	P3	59,33	59,31	0,02
P4	1,89	1,9	-0,01	P4	1,62	1,6	0,02	P4	1,63	1,73	-0,1
P5	0,45	0,47	-0,02	P5	0,44	0,46	-0,02	P5	0,52	0,52	0
P6	1,25	1,26	-0,01	P6	1,24	1,22	0,02	P6	0,99	0,99	0
P7	9,94	9,75	0,19	P7	9,87	9,49	0,38	P7	7,98	8,14	-0,16
P8	4,15	4,04	0,11	P8	4,36	4,58	-0,22	P8	3,2	3,24	-0,04
P9	1,03	1,06	-0,03	P9	1,23	1,23	0	P9	1,02	1,2	-0,18
P10	2,16	2,13	0,03	P10	2,46	2,43	0,03	P10	2,94	2,89	0,05
P11	0,76	0,76	0	P11	0,72	0,71	0,01	P11	0,52	0,54	-0,02
P12	0,29	0,29	0	P12	0,3	0,26	0,04	P12	0,22	0,21	0,01
P13	4,5	4,7	-0,2	P13	4,49	4,55	-0,06	P13	5,15	5,09	0,06
P14	6,49	6,53	-0,04	P14	6,81	6,86	-0,05	P14	6,98	6,52	0,46
P15	0,4	0,4	0	P15	0,46	0,47	-0,01	P15	0,33	0,34	-0,01
P16	1,26	1,22	0,04	P16	0,96	0,96	0	P16	0,8	0,81	-0,01
P17	0,32	0,32	0	P17	0,35	0,36	-0,01	P17	0,28	0,29	-0,01

Fuente: Elaboración Propia

Tabla 94: Estimaciones de Tenencia por Zona – Z4 a Z6

Z4	% Población	% Panel	Error	Z5	% Población	% Panel	Error	Z6	% Población	% Panel	Error
P1	4,08	3,99	0,09	P1	4,15	4,1	0,05	P1	3,55	3,66	-0,11
P2	4,49	4,8	-0,31	P2	4,92	5,01	-0,09	P2	3,52	3,38	0,14
P3	58,45	58,13	0,32	P3	58,55	58,91	-0,36	P3	51,98	51,93	0,05
P4	1,58	1,51	0,07	P4	1,97	1,96	0,01	P4	1,5	1,51	-0,01
P5	0,57	0,62	-0,05	P5	0,42	0,42	0	P5	0,51	0,51	0
P6	1,13	1,1	0,03	P6	1,25	1,27	-0,02	P6	0,95	0,91	0,04
P7	8,31	7,98	0,33	P7	7,99	8,06	-0,07	P7	16,86	16,55	0,31
P8	4,46	4,69	-0,23	P8	4,31	4,1	0,21	P8	3,26	3,38	-0,12
P9	1,17	1,2	-0,03	P9	1,09	1,03	0,06	P9	1,04	1,1	-0,06
P10	2,04	2,05	-0,01	P10	1,87	1,9	-0,03	P10	2,8	2,77	0,03
P11	0,63	0,59	0,04	P11	0,69	0,69	0	P11	0,51	0,5	0,01
P12	0,26	0,31	-0,05	P12	0,28	0,27	0,01	P12	0,27	0,26	0,01
P13	4,07	4,13	-0,06	P13	4,33	4,35	-0,02	P13	4,87	5,03	-0,16
P14	7,17	7,27	-0,1	P14	6,39	6,15	0,24	P14	6,6	6,76	-0,16
P15	0,45	0,47	-0,02	P15	0,56	0,53	0,03	P15	0,39	0,39	0
P16	0,87	0,86	0,01	P16	0,9	0,93	-0,03	P16	0,87	0,87	0
P17	0,29	0,3	-0,01	P17	0,34	0,32	0,02	P17	0,51	0,49	0,02

Fuente: Elaboración Propia

ANEXO P: Análisis de la distribución de transacciones promedios mensuales durante periodo en estudio

(i) Distribución promedio mensual de cantidad de transacciones

MUESTRA ALEATORIA 25000 CLIENTES PERSONAS

	200701	200702	200703	200704	200705	200706	200707	200708	200709	200710	200711	200712	200801	200802	200803	200804	200805	200806	200807	200808	200809
%	3,62	3,47	3,80	3,78	3,75	3,98	4,06	4,50	3,99	4,66	4,61	5,01	4,45	4,20	4,59	4,35	5,19	5,32	5,24	5,12	5,29

PANEL DE CLIENTES

	200701	200702	200703	200704	200705	200706	200707	200708	200709	200710	200711	200712	200801	200802	200803	200804	200805	200806	200807	200808	200809
%	3,73	3,62	3,92	3,81	3,81	4,08	4,21	4,75	4,24	4,91	4,72	5,28	4,64	4,29	4,83	4,36	5,38	5,62	5,31	5,12	5,43

El valor del porcentaje es obtenido sobre el total de transacciones de los meses de estudio para cada muestra.

(ii) Análisis estadístico cantidad de transacciones promedio mensual 200612 a 200809

A continuación se muestran los resultados de la Prueba de Kolmogorov-Smirnov para dos grupos formados por el panel y la muestra aleatoria respectivamente:

Tabla 95: Prueba de Kolmogorov-Smirnov para dos muestras

		CANTIDAD
Diferencias más extremas	Absoluta	,194
	Positiva	,194
	Negativa	-,022
Z de Kolmogorov-Smirnov		,649
Sig. asintót. (bilateral)		,793

Fuente: Resultados Software SPP- a Variable de agrupación: GRUPO

Probabilidad asintótica = 0,97>0,05 por lo tanto se acepta la hipótesis nula y se concluye que ambas muestras provienen de la misma población.

i) Distribución promedio mensual del monto de transacciones

MUESTRA ALEATORIA 25000 CLIENTES PERSONAS

FECHA	200701	200702	200703	200704	200705	200706	200707	200708	200709	200710	200711	200712	200801	200802	200803	200804	200805	200806	200807	200808	200809
%	3,52	3,34	3,90	4,38	3,70	3,60	4,41	4,21	3,60	4,09	4,72	5,01	4,09	4,79	5,28	5,17	4,64	5,07	5,44	4,32	5,32

PANEL DE CLIENTES

FECHA	200701	200702	200703	200704	200705	200706	200707	200708	200709	200710	200711	200712	200801	200802	200803	200804	200805	200806	200807	200808	200809
%	4,06	3,67	4,39	4,05	4,14	4,04	4,76	4,74	4,10	4,62	4,64	4,93	4,28	4,62	5,11	4,86	4,90	5,04	5,26	4,34	5,47

El valor del porcentaje ilustrado es obtenido sobre el monto total de transacciones de los meses de estudio para cada muestra.

ii) Análisis estadístico monto de transacciones promedio mensual 200612 a 200809

Tabla 96: Prueba de Kolmogorov-Smirnov para dos muestras
Estadísticos de contraste(a)

		MONTO
Diferencias más extremas	Absoluta	,259
	Positiva	,259
	Negativa	-,170
Z de Kolmogorov-Smirnov		,868
Sig. asintót. (bilateral)		,438

Fuente: Resultados Software SPP- a Variable de agrupación: GRUPO

Probabilidad asintótica = 0,44>0,05 por lo tanto se acepta la hipótesis nula y se concluye que ambas muestras provienen de la misma población.

ANEXO Q: Definiciones de Fuga de Clientes[47]

Un fenómeno de comportamiento estudiado por la Gerencia de Marketing corresponde a la variable fuga de clientes, si bien es cierto, con los datos del panel no se pretende construir modelos predictivos éste deberá tener en consideración este fenómeno para la mantención de la vigencia de sus datos.

De los estudios de fuga desarrollados por la entidad para los productos cuenta corriente, chequera electrónica y tarjeta de crédito se tienen los siguientes criterios:

(i) Fuga Tipo I (Cierre de productos): Considera el cierre de contrato del cliente para los productos Cuenta Corriente, Chequera Electrónica y Tarjeta de Crédito.

(i) Fuga Tipo II (Inactividad de productos): Corresponde a la inactividad del cliente con respecto al producto en estudio durante un periodo de tiempo determinado.

Los umbrales de inactividad a partir de los cuales se considera un cliente fugado son:

- Cuenta Corriente: 4 meses
- Chequera Electrónica: 7 meses
- Tarjeta de Crédito: 5 meses

(ii) Fuga Tipo III (Clientes Eliminados): Clientes que no realizaron transacciones en los meses analizados con los productos en estudio.

[47] "BANCO ESTADO, Informes definiciones de Fuga de Clientes para modelos predictivos" (Octubre del 2007) y "Estudio de Abandonadores Cuenta Corriente" (Agosto 2005 y Septiembre 2006)