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

MEDICIÓN DE INCONSISTENCIAS DE PRECIOS EN SUPERMERCADOS: DIAGNÓSTICO A NIVEL AGREGADO & PROBABILIDAD DE OCURRENCIA A PARTIR DE INFORMACIÓN OBSERVABLE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

JAVIER IGNACIO FERRAZ SALAZAR

PROFESOR GUÍA:
MARCELO OLIVARES ACUÑA

MIEMBROS DE LA COMISIÓN:
MARCEL GOIC FIGUEROA
ALEJANDRA PUENTE CHANDIA

Este trabajo ha sido financiado por el Centro de Estudios del Retail

SANTIAGO DE CHILE
2017

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: INGENIERA CIVIL INDUSTRIAL
POR: JAVIER IGNACIO FERRAZ SALAZAR
FECHA: 05 MARZO 2017
PROFESOR GUÍA: MARCELO OLIVARES

MEDICIÓN DE INCONSISTENCIAS DE PRECIOS EN SUPERMERCADOS: DIAGNÓSTICO A NIVEL AGREGADO & PROBABILIDAD DE OCURRENCIA A PARTIR DE INFORMACIÓN OBSERVABLE

A partir de reclamos documentados por el Servicio Nacional del Consumidor entre 2011-2015 en contra de las principales cadenas de supermercados, se sostiene que un 15% del total de reclamos se atribuye a categorías de mal cobro de precios. Muchos de estos casos, atribuidos a un cobro de precio superior a lo exhibido. Algunos de éstos han sido de conocimiento público, involucrando la imagen de actores de la industria supermercadista y finalmente, siendo necesaria la intervención del SERNAC. Lo anterior motiva el cuestionamiento sobre los métodos de colocación de precios que utilizan las cadenas de supermercado actualmente. Más aún, por parte de los consumidores, se pone en duda sobre la existencia de prácticas sistemáticas que vulneren los derechos de los consumidores, por supuestos cobros inconsistentes con relación a los precios exhibidos. El contexto anterior motiva la elaboración de un estudio que permita caracterizar discrepancias en el cobro de precios en compras de supermercados y eventualmente sugerir métodos para futuras mediciones. El trabajo de investigación considera la obtención y uso de dos fuentes de datos: (1) compras realizadas por un conjunto de clientes quienes reportan precios declarados en góndola y precios cobrados mediante registro fotográfico, y (2) registros de compras realizadas para dos meses en una cierta cadena y un conjunto de fotografías que registra el precio declarado de productos en sala.

Resultados a partir de una base de datos de 2.128 muestras extraídas en compras de supermercado, aseguran la existencia de: (1) una tasa promedio de discrepancia de precios en torno al 14%, (2) contra intuitivamente, un 70% de los casos el cliente habría sido favorecido (cobro menor a lo exhibido) y (3) una desviación porcentual promedio del precio en torno al 16%. El uso de modelos de clasificación, sugiere diferencias significativas en probabilidades de inconsistencia en precios según: (1) la cadena de supermercado, (2) la localización geográfica de las salas y (3) la aplicación de promociones en los productos. Se evidencia un desempeño inferior, es decir, mayores tasas, para los supermercados Unimarc y Express de Líder y tasas menores para supermercados Jumbo y Santa Isabel. Con respecto a la localización, se obtiene proporciones mayores para compras en supermercados de zona sur y periferia de Santiago. A su vez, la mayor demanda operacional a partir de productos en promociones, parece tener un efecto positivo en la probabilidad de inconsistencias.

El análisis de 9.931 datos transaccionales de compras en salas de supermercados reafirma el orden de magnitud de la tasa de discrepancia agregada, la cual se estima en un 16%. Además, una caracterización de los productos a partir de la frecuencia de recambio de precios abre el cuestionamiento sobre posibles correlaciones con la propensión de inconsistencia en precios, lo cual queda como una línea futura de investigación. Finalmente se espera que el trabajo de investigación sea de conocimiento público, y particularmente sea de interés para el Servicio Nacional de Consumidor y la industria supermercadista con el objeto de establecer protocolos de seguimiento, métricas en el asunto y, por último, propuestas de mejoras.

DEDICATORIA

Me gustaría dedicar este trabajo a todas aquellas personas que contribuyeron en este largo camino, de prácticamente 19 años como estudiante.

Un deseo de gratitud y dedicación especial a mi madre Cecilia, quien ha sido el pilar fundamental de mi persona y mi padre Guillermo, quien me ha brindado su gran apoyo durante todos estos años. Finalmente, todo lo que soy y seré, se los debo a ellos.

También, quiero dedicar este trabajo a mis amigos más cercanos, aquellos inolvidables del colegio: Javier Labbe, Enrique Gomez, Daniel Gomez, y también, a mis grandes amigos de la universidad: Felipe Lorca, Tomás Lagos y Mario Novoa.

Finalmente, dedico además este trabajo al cuerpo académico de Ingeniería Civil Industrial de la Universidad de Chile, a mis profesores guías Marcelo O. y Marcel G., y miembros del Taller de Título, quienes han colaborado de forma imprescindible en este proceso de investigación.

AGRADECIMIENTOS

Nace de mí, el deseo de agradecer a todas aquellas personas que colaboraron de una u otra forma en este proceso que culmina, finalmente, luego de seis años de carrera universitaria.

Comenzando por las personas más cercanas, doy mis más sinceros agradecimientos a mi querida madre, quien más que nadie, creyó en mí y decidió brindar un apoyo incondicional en mi formación, desde siempre. A mi padre, por tratarse de una de las personas que más consejos me ha entregado, y ha sido un pilar fundamental en mi formación, como persona y profesional.

Quiero manifestar mis más sinceros agradecimientos a todas aquellas personas que han desempeñado un rol de profesor en mi camino como estudiante, desde mis inicios en el jardín de niños hasta la Universidad. Particularmente, quiero reconocer a los profesores: Juan Carlos Sáez, David Alvo, Carlos Vignolo, Alejandra Puente, Marcelo Olivares y Marcel Goic, quienes han contribuido de manera significativa en mi crecimiento profesional.

Al departamento de ingeniería civil industrial, quiero manifestar mis agradecimientos por haber sido prácticamente un segundo hogar durante estos últimos años. Ha sido un honor y un placer ser parte de una comunidad académica de reconocimiento mundial, que busca contribuir no solo en la dimensión académica, sino que también en la calidad como ciudadanos y personas. Confío plenamente, que seguirá marcando historia en nuestro país.

Me gustaría reconocer a todos mis compañeros de carrera, particularmente a mis amigos cercanos: Mario Novoa, Felipe Lorca & Tomás Lagos, quienes hicieron de este proceso universitario, una etapa absolutamente memorable en mi vida personal.

Finalmente, agradecer a todas las personas que aportaron, de alguna u otra forma en la realización de mi proyecto para ser Ingeniero Civil Industrial: familiares cercanos, amistades, compañeros de carrera, miembros del CERET, profesores del colegio y de la infancia, y muchos otros más. Sin ustedes, este camino habría sido ciertamente más difícil o quizás, imposible de recorrer.

TABLA DE CONTENIDO

1. Introducción 8
2. Descripción del proyecto 10
   2.1. Definición de objetivos 14
      2.1.1. Objetivo General 14
      2.1.2. Objetivos Específicos 14
   2.2. Alcances 14
3. Desarrollo Metodológico 15
   3.1. Diseño Muestral 15
   3.1.1. Base de datos construida 15
   3.1.2. Base de datos transaccionales 17
   3.2. Proceso KDD 18
      3.2.1. Pre-procesamiento de datos 18
      3.2.2. Transformación de datos 18
      3.2.3. Análisis Descriptivo 19
      3.2.3.1. Descriptivos Generales 19
      3.2.3.2. Análisis desagregado 20
      3.2.3.3. Fecha de colocación de etiquetas de precio 24
      3.2.3.4. Valencia 25
      3.2.4. Minería de Datos 26
      3.2.4.1. Modelo Logit (inicial) 26
      3.2.4.2. Árbol de Decisión 28
      3.2.4.3. Modelo Logit con interacción 29
      3.2.4.4. Evaluación de Modelos 30
      3.2.4.5. Análisis mediante datos transaccionales 32
   3.3. Evaluación de Resultados 34
4. Conclusiones 36
5. Bibliografía 38
6. Anexos 39

Anexo A: Índice de ventas de supermercados, desestacionalizada y tendencia (enero 2011-noviembre 2016) 39

Anexo B: Distribución geográfica a partir de nivel socioeconómico en zona metropolitana de Chile. 39

Anexo C: Medición de calidad de servicio en el retail supermercados desagregado según dimensiones 40

Anexo D: Resultados de encuesta sobre percepción de supermercados 40

Anexo E: Ficha técnica referente a recolección de datos 42
Anexo F: Detalles tasas de discrepancias desagregadas según la zona geográfica y oferta de productos..... 42
Anexo G: Gráfico P-P respecto al ajuste de función de densidad de probabilidad referente a la desviación en precio para productos con inconsistencia..... 43
Anexo H: Subconjunto de panel de datos construidos partir de levantamiento en sala 44
Anexo I: Resultados Árbol de Decisión..... 45
Anexo J: Modelo de regresión logística con interacción de variables..... 46

ÍNDICE DE TABLAS

Tabla 1: Ficha técnica de datos transaccionales dispuestos 18
Tabla 2: Estadísticos descriptivos a partir de base de datos construida 20
Tabla 3: Test ANOVA para diferencias en discrepancias a partir de lapsus en la actualización de precios 25
Tabla 4: Test de medias para desviación porcentual de precios con inconsistencia 26
Tabla 5: Resultados a partir de regresión logística 28
Tabla 6: Estadísticos Descriptivos mediante uso de datos transaccionales 32

ÍNDICE DE IMÁGENES

Ilustración 1: Distribución Promedio de Categorías de Reclamos en Supermercados (2011-2015) Fuente Datos: SERNAC, 2011-2015 8
Ilustración 2: Evolución de proporción de reclamos relacionados con mal cobro en supermercados chilenos 2011-2015 9
Ilustración 3: Distribuciones porcentuales de transacciones a lo largo de la semana. Fuente: Propia 11
Ilustración 4: Mapa conceptual en la actualización de precios en supermercados. 12
Ilustración 5: Medida de fluctuaciones de precios, agregado según categorías de productos por nivel de perecibilidad. Fuente: Propia 13
Ilustración 6: Mapa ilustrativo del proceso de muestreo de datos 16
Ilustración 7: Caracterización de muestras obtenidas en base de datos consolidada 16
Ilustración 8: Tasas de discrepancia para cadenas de supermercado estudiadas 21
Ilustración 9: Tasa de discrepancia desagregado a partir de la zona geográfica, formato y oferta de productos 22
Ilustración 10: Discrepancias a lo largo de la semana 23
Ilustración 11: Tasa de discrepancia según categorización de productos por nivel de perecibilidad 23
Ilustración 12: Histograma y comportamiento de ratios de discrepancias a partir de lapsus entre fecha de actualización de precio y fecha de compra 24
Ilustración 13: Histograma sobre distribución de desviaciones porcentuales de precios en casos con discrepancia acompañado de función de densidad de probabilidad de mejor ajuste 25
Ilustración 14: Curvas ROC obtenida a partir de modelos de clasificación 31
Ilustración 15: Evolución de precios de un producto ejemplo con altas fluctuaciones. Fuente: Datos transaccionales 33
Ilustración 16: Evolución de precios de un producto ejemplo con bajas fluctuaciones. Fuente: Datos transaccionales 33
Ilustración 17: Mapa de dispersión discrepancias vs fluctuaciones de precios. Fuente: Datos transaccionales 34

1. INTRODUCCIÓN

Se estima que las ventas de comercio para el año 2014, ronda entorno al 29% del PIB nacional, siendo la industria supermercadista responsable de un 19,7% (del 29%) aproximadamente. El índice de Ventas de Supermercados (ISUP) corresponde a un indicador utilizado por el INE, con el objetivo de capturar los niveles de ventas en los establecimientos nacionales de supermercados. El crecimiento de este indicador desde el año 2011-presente es un reflejo del incremento en ventas en supermercados. La gráfica en Anexo A ilustra la evolución en el índice de ventas desde 2011 hasta ahora.

Lo anterior da cuenta de la importancia del comercio generado en establecimientos de supermercados, tanto en términos de dinero transado como también en cantidad de clientes involucrados.

Uno de las principales preocupaciones de las empresas de supermercado corresponde a la calidad de servicio percibida por sus clientes, buscando siempre proyectar confianza, responsabilidad, y en general, una buena experiencia de compra. Lo anterior dada la existente correlación positiva con el número de ventas, ganancias y rentabilidad en el largo plazo.

La existencia e incremento aparente de reclamos gestionados por el Servicio Nacional del Consumidor entorno a problemas con el cobro de precios de productos en supermercados nacionales (ver Ilustración 1 y 2) ha puesto en cuestionamiento la transparencia de cara al consumidor. Lo anterior, pudiéndose explicar ya sea por un crecimiento real en la gestión de precios de productos o un posible aumento en la conciencia del consumidor en su experiencia de compra.

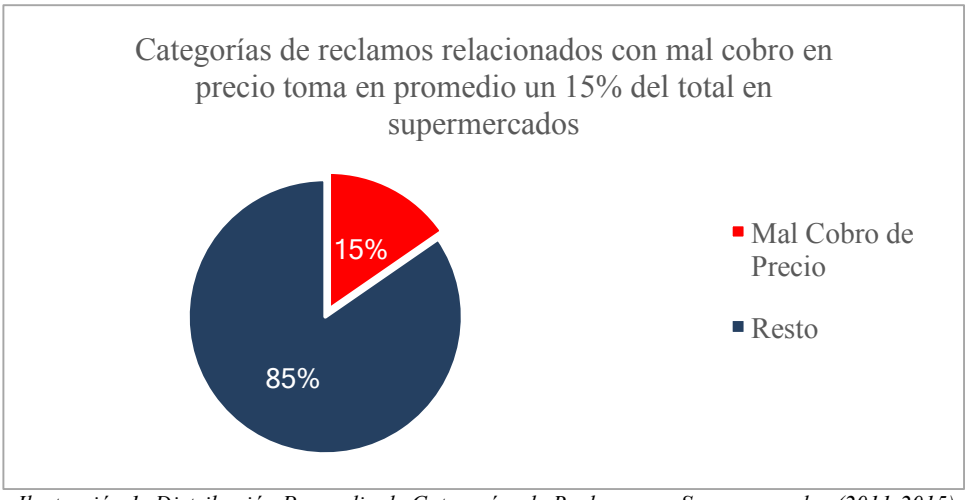


Ilustración 1: Distribución Promedio de Categorías de Reclamos en Supermercados (2011-2015)
Fuente Datos: SERNAC, 2011-2015

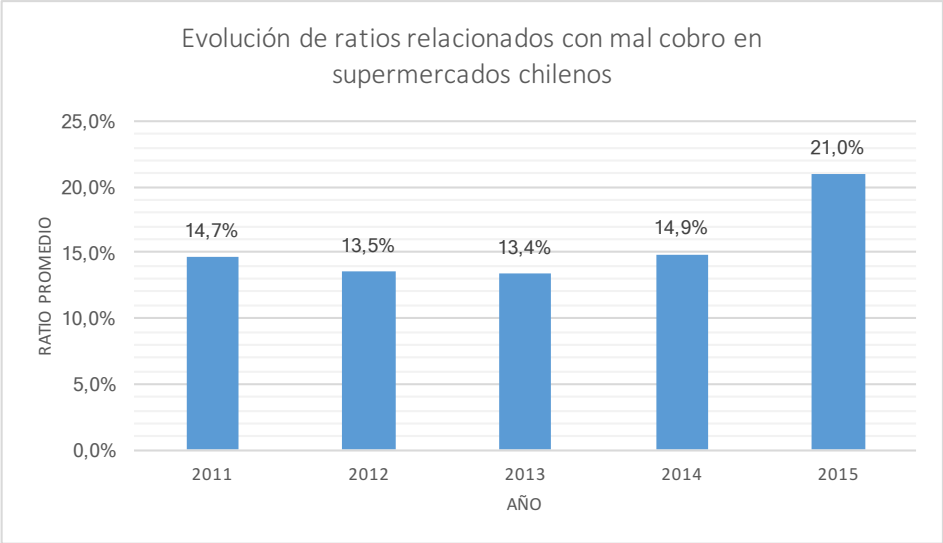


Ilustración 2: Evolución de proporción de reclamos relacionados con mal cobro en supermercados chilenos 2011-2015

Particularmente, de la categoría de mal cobro en precios, se identifica una subcategoría referida al cobro inconsistente entre el precio declarado en fleje de góndola y el precio finalmente cobrado. Estos casos se traducen a molestias y eventualmente, pérdidas de confianza por parte de los clientes. En adición, Valarie A. Zeithaml (1988), concluye sobre el rol que cumple el precio percibido en la probabilidad de compra de los consumidores, evidenciando una dependencia real entre ambas variables.

El Centro de Estudios del Retail (CERET) corresponde a una institución que desarrolla investigaciones relacionadas con el retail (físico como digital) en áreas como: gestión de productos, cadena de demanda y suministros, tiendas, clientes, sistemas transaccionales y minería de datos, entre otros dominios. Particularmente, ha desarrollado en el último tiempo, investigaciones relacionadas con percepción de los consumidores (calidad de servicio) a partir de compras en la industria supermercadista.

Actualmente, no se dispone de ninguna investigación formal relacionada con el estudio de casos con inconsistencias de precios. Bajo esta premisa, el presente trabajo de investigación, es acogido por el CERET con la finalidad de construir un diagnóstico sobre la situación actual entorno al mal cobro de precios (inconsistencias) en compras en los distintos supermercados chilenos.

El presente estudio hace uso de dos bases, una primera, construida mediante un exhaustivo levantamiento de datos en compras de diversos supermercados localizados en la Región Metropolitana. La base de datos construida se consolida mediante 2.128 datos, repartidos en seis cadenas de supermercados en tres zonas geográficas agrupadas y durante 18 semanas. Una segunda base de datos, es facilitada por una cadena de gran participación en el mercado supermercadista y consiste en datos transaccionales para ciertos meses del año 2016.

2. DESCRIPCIÓN DEL PROYECTO

En el marco de la investigación, se justifica el planteamiento de un conjunto de hipótesis con el motivo de caracterizar con mayor precisión el comportamiento de mal cobro de precios. Una primera aproximación apunta a entender efectos de primer orden: validar la existencia y significancia del fenómeno de discrepancias en términos de magnitud. Luego, una interrogante posterior concierne a la identificación de aquella entidad perjudicada. Según la serie de reclamos acogidos por el SERNAC, apuntaría a un eventual perjuicio de los clientes. Formalmente, la hipótesis se plantea:

H1. *Dada una inconsistencia en el precio de un producto, existe una mayor probabilidad en que tenga un impacto monetario negativo hacia los clientes en vez de la cadena de supermercado*.

Por otra parte, se estudia la relación entre un conjunto de factores y el fenómeno de mal cobro de precios en salas de supermercado. Por un lado, con el motivo de identificar potenciales correlaciones, como también para evaluar impactos de distinta índole. Su justificación se detalla a continuación:

1. *Localización geográfica*: Miembros de la industria supermercadista revelan la existencia de diferencias en la disposición de los productos, particularmente con relación al orden (limpieza, etiquetas de precios colocadas correctamente) en sala según las distintas zonas de la Región Metropolitana. El motivo se atribuye a: (1) diferencias del recurso humano entre salas y (2) diferencias en el comportamiento de los clientes debido a características adyacentes, particularmente a partir del estrato socioeconómico. Es por ello que se decide incorporar una componente asociada a la ubicación geográfica de los supermercados a estudiar. Para más detalles sobre la clasificación geográfica ligado a la distribución socioeconómica de la población, ver Anexo B.

2. *Cadenas de Supermercado:* En el último tiempo, la función logística ha tomado cada vez mayor importancia en las empresas de retail, dado el incremento en exigencias y competencia en el mercado. Lo anterior con el objetivo de generar precios más competitivos y con una mejor calidad en servicio. El mercado supermercadista en Chile no ha sido la excepción, y las cadenas de supermercado han percibido las herramientas logísticas como una ventaja competitiva fundamental. Sin embargo, existe heterogeneidad entre cadenas. Una explicación válida viene dada según las diferencias en propuestas de valor, lo que finalmente recae tanto en distinciones en la logística (en general) como también en el nivel de servicio entregado. Tomando como ejemplo Jumbo, quien declara su posicionamiento fuertemente basado en la calidad de su surtido y el nivel de servicio con el cliente. Es de esperar que, diferencias tanto

en niveles de logística como en la calidad del servicio hacia los clientes, tenga un impacto en la minuciosidad en la actualización de precios, tanto en caja (sistema) como también en los flejes de góndola.
El estudio de *"Medición de calidad de servicio percibido en la industria del retail supermercados"* (CERET, 2016) revela heterogeneidad en la percepción por parte de clientes hacia la calidad de las distintas cadenas de supermercado, particularmente a partir de la dimensión "tangibilidad", que alude a las instalaciones físicas, equipos, materiales asociados y producto físico, etc. El estudio califica Jumbo & Unimarc con mejor desempeño en contraste con Santa Isabel & Tottus Express con peor desempeño percibido (ver Anexo C para más detalles).

Adicionalmente, los resultados de una encuesta a diversos clientes que denotan un comportamiento exhaustivo en la revisión de sus compras, sugieren diferencias en el nivel de orden percibido y en tasas de discrepancia esperadas según cadena (ver Anexo D). Siguiendo esta línea se podría argumentar que, diferencias en los niveles de inconsistencias entre cadenas de supermercados, pueda ser explicado según diferencias en desempeño operacional (e.g. mayor orden en sala, limpieza, stock de productos) y con propuestas de valor enfocadas mayormente en el cliente.

3. <u>Período a lo largo de la semana:</u> Es de conocimiento común la existencia de estacionalidad en las ventas en supermercados, conforme al comportamiento particular de la demanda. Datos transaccionales en compras para una cierta cadena de supermercados para dos meses de actividad apoyan la siguiente distribución de ventas a lo largo de la semana (Ilustración 3), destacando el fin de semana como aquel período con mayor concentración de las ventas (sábado y domingo; 21.1% y 16.2% respectivamente):

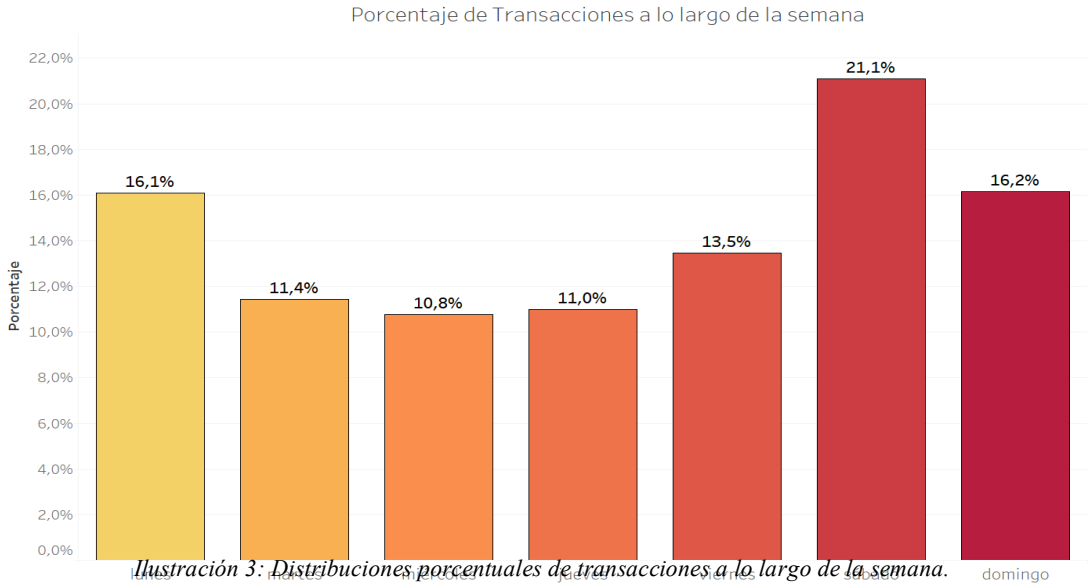


Ilustración 3: Distribuciones porcentuales de transacciones a lo largo de la semana.
Fuente: Propia

De lo anterior, surge el cuestionamiento sobre potenciales diferencias en la probabilidad de encontrar errores en el cobro de productos, según el día en que la compra es efectuada en el establecimiento, dada la diferencia en demanda operacional asociada al día.

4. *Oferta en productos:* Diferencias en los mecanismos para actualizar precios de productos en oferta relativo a productos en precio normal, en las bases de datos en supermercados, motivan la evaluación de esta variable dentro de la probabilidad inconsistencia en el cobro. Entrevistas con actores gerenciales de la industria supermercadista revelan procedimientos al momento del cobro de productos en salas resumidos (Ilustración 4): (1) se consulta a una matriz central de datos con precios en tiempo real, (2) consulta a una segunda matriz para verificar promociones asociadas al producto (en caso de existir) sobre-escribiendo el precio original, (3) se recibe el precio final el cual es cobrado al cliente. Las bases de datos con productos en oferta en general presentan altas volatilidades y variaciones, debido a la heterogeneidad existente en la gama de ofertas y promociones ofrecidas (e.g. programas de fidelidad, casas comerciales, descuentos por día, promociones especiales, etc). En este contexto, actores de la industria supermercadista insinúan complicaciones en la consistencia de cobros en caja, dada la mayor demanda de recursos para la actualización de etiquetas de precios en tienda.

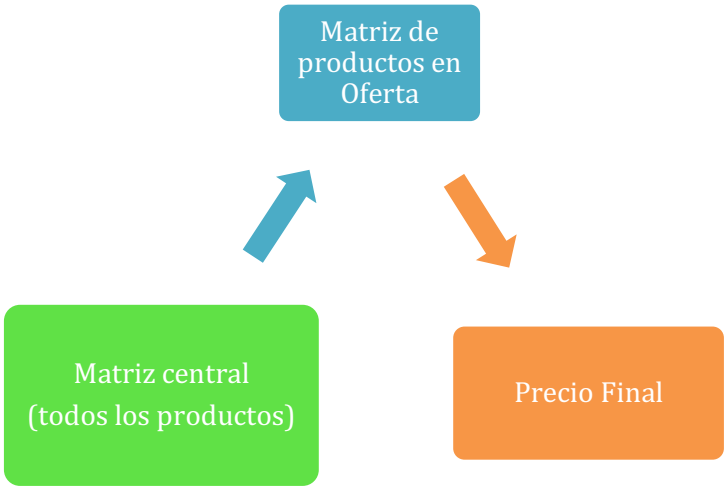


Ilustración 4: Mapa conceptual en la actualización de precios en supermercados.

En adición, Lichtenstein, D., Ridgway, N., & Netemeyer, R. (1993), sugieren, al igual que evidencia empírica y teórica, la relación positiva entre la precisión en recuerdo de precios por parte de clientes (*price recall accuracy*) y la percepción de precios en su rol positivo, particularmente, la propensión de uso en cupones y ofertas en productos. Es decir, clientes poseen una mayor probabilidad de recordar precios

(particularmente al momento de pagar) cuando los productos poseen oferta/promoción. El mal cobro de productos al momento de pagar tendría en tal caso una mayor probabilidad de ser recordado y manifestado en el instante de la compra. Si bien, lo anterior no posee un impacto en la probabilidad de inconsistencia de precios, si sugiere consecuencias negativas en la percepción hacia el supermercado.

5. *Categorías de productos:* El surtido de productos dispuestos en establecimientos de supermercado se caracteriza por su heterogeneidad, y su categorización es posible según diversos criterios, ya sea: nivel de rotación, perecibilidad, funciones de consumo individual (COICOP), entre otros. Para este caso particular, se testeará utilizando como criterio, la agrupación de productos a partir del nivel de perecibilidad de éstos. El motivo se fundamenta a partir de juicios comunicados por miembros de la administración en la industria, que apuntaría a una potencial conexión entre niveles de rotación de los productos, promociones generadas, periodicidad en las fluctuaciones de precios y finalmente, desempeños en la probabilidad de discrepancias de precios. En concreto, a partir de los datos transaccionales, se infiere que, a nivel agregado el promedio de fluctuaciones de precios en productos se ordena en forma decreciente: (1) perecibles, (2) no perecibles, (3) non food, (4) varios y (5) textiles. Los valores se presentan en la siguiente gráfica (Ilustración 5):

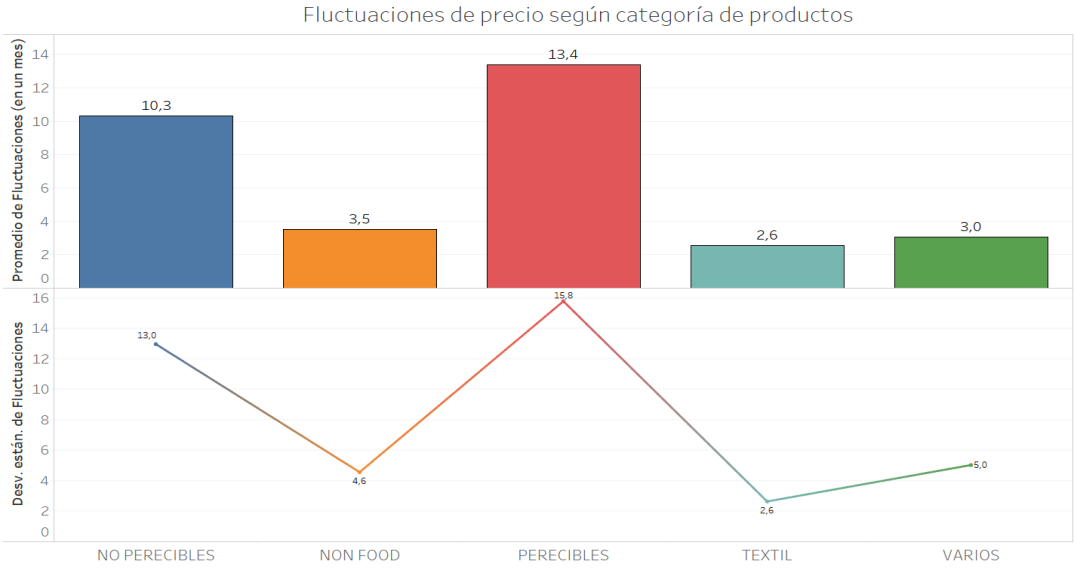


Ilustración 5: Medida de fluctuaciones de precios, agregado según categorías de productos por nivel de perecibilidad. Fuente: Propia

El análisis diferenciado (desagregado por factores) tiene un valor significativo, particularmente para las cadenas de supermercado, dado el potencial de *accionabilidad* que pudiesen tener los resultados. Eventualmente, lograr identificar salas con peores desempeños en términos de cobro de productos, permitiría no solo generar alertas, sino mejorar

desempeños a nivel cadena, implementando mejoras operacionales en salas críticas. Finalmente, el conjunto de hipótesis se sintetiza de la forma:

H2. Los siguientes factores tienen un impacto, ya sea positivo o negativo, en la probabilidad de inconsistencia de precio en productos:

a. *Tamaño (operacional) de la sala del supermercado o también denominado "formato".*
b. *Aplicación de ofertas en el producto*
c. *Localización geográfica*
d. *La cadena del supermercado*
e. *Nivel de rotación del producto*
f. *Tiempo transcurrido entre la realización de la compra y la fecha de colocación de la etiqueta de precio.*

2.1. Definición de objetivos

2.1.1. Objetivo General

Obtener una aproximación de primer orden sobre la existencia de discrepancias de precios en cadenas de supermercados nacionales y caracterizar factores relacionados con el fenómeno.

2.1.2. Objetivos Específicos

1. Caracterizar el fenómeno a nivel agregado, mediante estadísticos descriptivos.
2. Identificar variables correlacionadas con la probabilidad de discrepancia de precios en supermercados.
3. Determinar y cuantificar implicancias monetarias, tanto para los clientes como para las cadenas de supermercado.

2.2. Alcances

El análisis del estudio es válido solo para los supermercados de la región metropolitana, dada la estratificación para establecimientos pertenecientes sólo esta región. Por tanto, el estudio no considera como variable explicativa, la ubicación regional del supermercado, en la probabilidad de discrepancia de precio. Asimismo, el estudio es concluyente solo a partir de las cadenas de supermercados escogidas (Híper Líder, Líder Express, Tottus, Tottus Express, Jumbo, Santa Isabel y Unimarc).

Además, el proyecto de investigación no considera el diseño de mecanismos operativos que permitan minimizar la probabilidad de discrepancias de precios (fleje-boleta), tan solo se diagnostica el fenómeno a nivel agregado y desagregado según las variables explicativas.

Luego, el estudio no se considera un análisis de dependencia sobre la variable objetivo a partir de variables no observables (e.g. factores que explican deficiencias operativas, como

la cantidad de trabajadores a cargo de la actualización de precios, diferencias en protocolos de colocación de etiquetas, entre otras).

3. DESARROLLO METODOLÓGICO

3.1. Diseño Muestral

3.1.1. Base de datos construida

La inexistencia de datos de primera fuente, hace necesario el diseño e implementación de un proceso de recolección de datos que permita capturar el comportamiento de discrepancias de precios de productos en supermercados.

En efecto, la obtención de la base de datos se lleva a cabo en la Región Metropolitana, con una duración total de 18 semanas, de las cuales 15 de éstas se dedica a la recolección y procesamiento de datos. El proceso se resume en las siguientes etapas:

Etapa 1: Se selecciona las cadenas de supermercados que participan en la investigación. Se escogen las cuatro principales, en términos de su participación en la industria supermercadista; SMU, Tottus, Walmart y Cencosud. De cada cadena se toma en consideración la participación de supermercados en sus dos formatos: híper y estándar. Adicionalmente, se identifican tres zonas geográficas que logran segmentar la región metropolitana, utilizando como criterio la diferenciación del consumidor según sus características ligadas a la zona (nivel socioeconómico, preferencias, educación, etc). Las zonas establecidas son zona sur & periferia, centro poniente y finalmente zona oriente.

Etapa 2: Para la obtención de muestras, cada encuestador al momento de hacer sus compras en supermercado (perteneciente a una lista de establecimientos factibles según la segmentación por zona y cadena), extrae imágenes de la etiqueta de precio en fleje de góndola de cada uno de los productos a comprar mediante su dispositivo móvil (no se exige alguna categoría de productos especial ni condiciones específicas). Adicionalmente, se debe tomar fotografía del comprobante comercial (boleta) y hacer envío del total de fotografías al encargado de la investigación. Por cada fotografía, se compensa monetariamente al encuestador. Luego, con las imágenes de cada etiqueta de precio y el comprobante comercial, es posible contrastar ambos precios (expuesto vs cobrado). La secuencia de eventos involucrado en el proceso de muestreo se explicita en el siguiente diagrama (Ilustración 6):

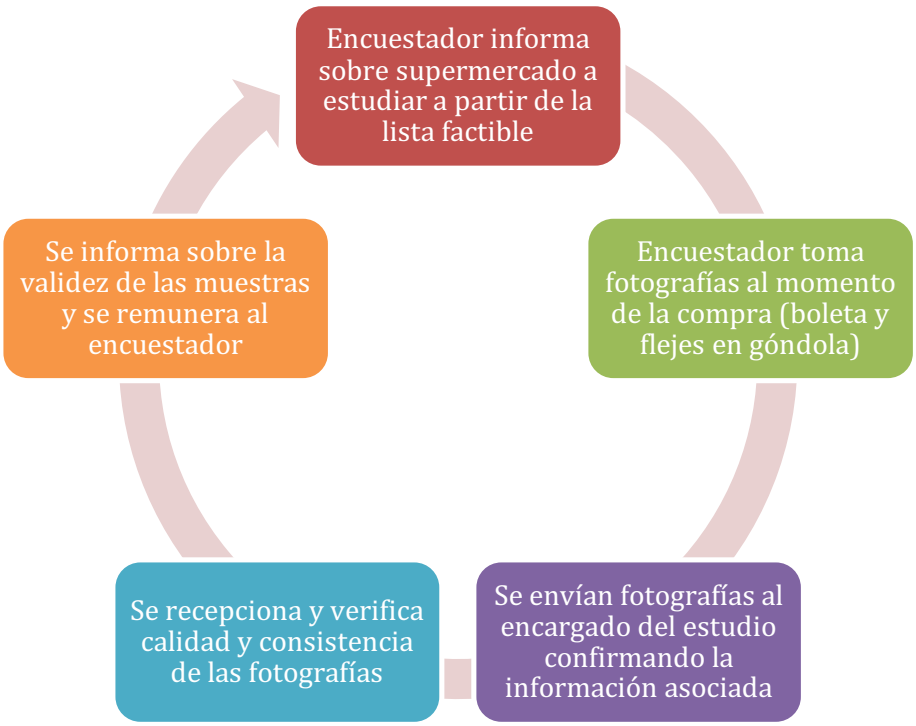


Ilustración 6: Mapa ilustrativo del proceso de muestreo de datos

En efecto, la base de datos final se construye mediante la recolección y procesamiento de un total de 2.128 muestras a partir de 28 salas repartidas en el Gran Santiago. Para detalles sobre la distribución de las muestras recolectadas a partir de los diferentes supermercados, ver Ilustración 7.

Supermercado	Formato	Salas Muestreadas	Items Muestreados	Discrepancias
Hiper Lider	Hiper	7	421	47
Express de Lider	Tradicional	5	216	36
Jumbo	Hiper	7	325	35
Santa Isabel	Tradicional	3	123	12
Tottus	Hiper	2	449	53
Tottus express	Tradicional	1	327	59
Unimarc	Tradicional	3	267	71
Total		**28**	**2128**	**313**

Ilustración 7: Caracterización de muestras obtenidas en base de datos consolidada

Mediante las imágenes capturadas en cada evento de compra de un encuestador (boleta y etiquetas de precio) se realiza el traspaso de la siguiente información a la base de datos:

- Id. Supermercado: Identificador del supermercado en donde se realiza la compra. A través de este código es posible relacionar con características observables subyacentes a cada sala, como la ubicación, tamaño, cadena de *retail*, entre otros.

- Id. Boleta: Identificador de la boleta, el cual corresponde a un código único ubicado en el comprobante comercial.
- Fecha de compra: Tiempo en que se realiza la compra (información en boleta).
- Código SKU: Identificador asociado al producto y al supermercado en particular.
- Nombre SKU: Breve descripción del producto, presente en boleta.
- Fecha Fleje: Corresponde a la fecha en que la etiqueta de precio fue impresa y colocada en el fleje de góndola.
- Oferta: En caso que el producto haya presentado alguna promoción se considera esta variable dicotómica con valor 1 (respaldado con la etiqueta de oferta en góndola).
- Precio Góndola (PG_i): Precio declarado en fleje de góndola.
- Precio Boleta (PB_i): Precio efectivamente cobrado hacia el cliente.

Para ver más detalles sobre el levantamiento de datos, ver ficha técnica en Anexo E

3.1.2. Base de datos transaccionales

En paralelo a los datos recolectados, se dispone de datos transaccionales en tienda, lo cual permite tener una mirada más completa sobre el comportamiento de discrepancias en compras realizadas. Por cada compra, se tiene información de temporalidad (día y hora), promociones asociadas, montos cobrados, cantidad de ítems adquiridos, métodos de pagos, etc. Particularmente, en el marco de la presente investigación se hará uso del registro de compras efectuadas en adición de un monitoreo físico en sala, mediante fotografías de un conjunto de ítems, lo que permite respaldar y luego comparar la consistencia en precios cobrados en las transacciones registradas. Datos transaccionales permiten la observación no sólo de una compra puntual, sino del espectro completo de compras a partir de SKUs (seguimiento en el tiempo), lo que facilita una mayor precisión en las medidas de desempeño entorno a inconsistencias de precios en compras.

Más adelante, se explicitan algunas consideraciones importantes a considerar entorno a los datos que posteriormente se utilizan para su análisis respectivo:

(1) El monitoreo físico en sala se realiza durante dos meses completos (octubre-noviembre, 2016) de un listado de productos específicos. Estos consisten en un total de entre 70-80 ítems por sala.
(2) Los ítems fueron escogidos a partir del nivel transaccional de éstos, siendo seleccionados aquellos en el top 100 productos con mayor participación en transacciones de cada sala.
(3) La recolección completa de datos considera un total de 8 salas a lo largo de Santiago. Sin embargo, el presente análisis solo considera el procesamiento de datos de una sala durante el mes de octubre, debidos a restricciones en recursos para el procesamiento de registros obtenidos físicamente.

Futuras observaciones e inferencias serán en función de las anteriores consideraciones. La siguiente tabla explicita la ficha técnica sobre el conjunto de datos utilizados para el análisis posterior.

Caracterización de datos transaccionales	
Tamaño muestral (n)	9.931
Cadenas supermercados	1
Salas	1
Período muestreo	Octubre 2016

Tabla 1: Ficha técnica de datos transaccionales dispuestos

3.2. Proceso KDD

En este proceso, se considera el tratamiento de ambas bases de datos utilizadas: la base recolectada y datos transaccionales dispuestos.

3.2.1. Pre-procesamiento de datos

Se verifica la calidad de los datos, según: la calidad intrínseca, contextual y representativa. Específicamente, para aquel conjunto de datos recolectados, se considera como factor clave, el nivel de objetividad de los datos, el valor aportado por las variables capturadas (completitud y relevancia) y el nivel de interpretabilidad de la información.

Una complicación identificada para datos recolectados, tiene relación con que algunas de las cadenas de supermercado no establecen, como información visible, la fecha de colocación de la etiqueta de precio. Por tanto, la variable "Fecha fleje", manifiesta valores inexistentes, para un conjunto significativo de observaciones. Dada la cantidad de valores faltantes (67% de la muestra final aproximado), se aislará esta variable para un análisis fuera de los modelos de clasificación.

3.2.2. Transformación de datos

Para la construcción de los modelos y análisis varios a realizar, es necesario una modificación en los datos recolectados. En específico se tiene:

i. *Creación de variables*:
   a. Discrepancia: Identificada como la variable objetivo, se constituye como una variable dicotómica de la forma:

$$Discrepancia_i = \begin{cases} 1, si\ PG_i \neq PB_i \\ 0, si\ PG_i = PB_i \end{cases}$$

   b. Desviación porcentual: Para los casos de inconsistencia, informa el porcentaje en que el precio cobrado se desvía del cobrado (i.e. $DP = 100 \cdot \frac{PG_i - PB_i}{PG_i}$). Es una variable normalizada, dada la necesidad de hacer comparable las desviaciones, sin importar el precio del producto con inconsistencia.
   c. Favor Cliente/Supermercado: Captura la valencia para cada observación (i.e. la dirección sobre quien se ve favorecido; ya sea cliente o supermercado).

ii. *Dicotomización de variables de naturaleza categóricas:* Se transforman todas las variables categóricas a variables dicotómicas (i.e. valores 0,1), tales como: categoría producto, formato, supermercado, día semana, favorecimiento.

3.2.3. Análisis Descriptivo

En esta sección se examinan los datos construidos, mediante estadísticos descriptivos y gráficas, que finalmente, se cristaliza en inferencias preliminares a partir bajo distintos niveles de agregación. Lo anterior con la motivación de justificar el análisis y respaldar el conjunto de hipótesis escogidas.

3.2.3.1. Descriptivos Generales

Un análisis preliminar muestra desagregación del total de inconsistencias de precios encontradas a lo largo de los seis establecimientos de supermercados. De los resultados exhibidos en Tabla 1 se concluye que, a nivel agregado existe una tasa de discrepancia del orden del 14% (i.e. 4 de cada 30 productos comprados). Tomando en consideración la desviación porcentual del precio cobrado versus el precio declarado ($\bar{P}_D$) se tiene que, en promedio, ésta se encuentra entorno al 9%. El signo positivo denota un favorecimiento hacia el cliente, cobrándose un 9% menos del precio exhibido.

Otro hallazgo interesante, se relaciona con la proporción de casos encontrados, en donde el cliente se ve perjudicado (i.e. precio cobrado mayor al exhibido). Según los datos recolectados, éste corresponde solo a un 29.5% del total de inconsistencias encontradas y por lo tanto un 70.5% de las veces, el cliente sería favorecido (se cobra menos de lo declarado).

Estadísticos Descriptivos			
Métrica	*Descripción breve*	*Cálculo*	*Valor*
$\bar{X}_{TD}$	Tasa de discrepancia promedio	$\frac{1}{n}\sum_{i}^{n} 1_{\{x_i\}}$	**14,3%**
$\bar{P}_D$	Promedio en variación porcentual en discrepancia de precio para casos con inconsistencia.	$\frac{1}{m}\sum_{i}^{m} 1_{\{x_i=1\}} \cdot \left(\frac{PG_i - PB_i}{PG_i}\right)$	+8,9%
$\lvert\bar{P}_D\rvert$	Promedio en variación porcentual en discrepancia de precio para casos con inconsistencia (en valor abs.).	$\frac{1}{m}\sum_{i}^{m} 1_{\{x_i=1\}} \cdot \left\lvert\frac{PG_i - PB_i}{PG_i}\right\rvert$	16,4%
σ_{TD}	Desv. estándar de porcentaje de desviación absoluta en precio para casos con discrepancia.	$\sqrt{\frac{1}{m}\sum_{i}^{n} 1_{\{x_i=1\}}(\lvert\Delta\%P_i - \bar{X}_{TD}\rvert)^2}$	17,1%
$\%n_{FS}$	Proporción de discrepancias a favor de supermercado.	$\frac{\sum_i^n 1_{\{PG_i - PB_i<0\}}}{\sum_i^n 1_{\{x_i=1\}}}$	29,5%
$\%n_{FC}$	Proporción de discrepancias a favor de cliente.	$\frac{\sum_i^n 1_{\{PG_i - PB_i>0\}}}{\sum_i^n 1_{\{x_i=1\}}}$	**70,5%**

Tabla 2: Estadísticos descriptivos a partir de base de datos construida

3.2.3.2. Análisis desagregado

Al desagregar los datos según la cadena de supermercado (ver Ilustración 8) destaca la cadena Unimarc como aquella en donde se encuentra la mayor tasa de discrepancias, en torno a un 27%. Cadenas asociadas a Tottus y Walmart siguen un peldaño por debajo, con tasas del orden entre 11-18%. Contrariamente, Cencosud sobresale con tasas de discrepancias inferiores en sus dos formatos, Santa Isabel y Jumbo, rondando en torno al 10%.

Por otra parte, es posible visualizar la proporción de casos en donde el cliente se ve perjudicado para las distintas cadenas (representado a partir del área por sobre la línea negra en cada columna en Ilustración 8). Se observa que, para todas las cadenas excepto en supermercado Santa Isabel, la proporción de casos encontrados, apuntaría a favor del cliente (cobro menor a lo exhibido).

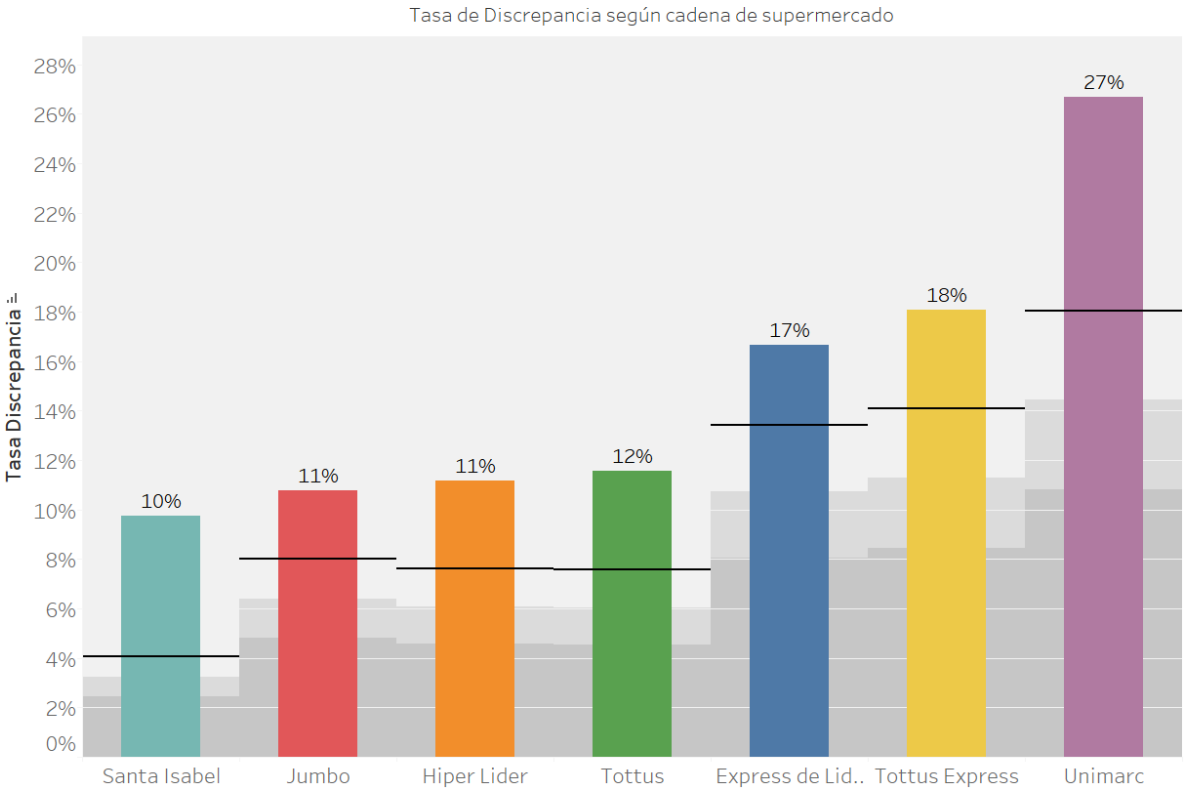


Ilustración 8: Tasas de discrepancia para cadenas de supermercado estudiadas

Tomando en consideración la zona geográfica, el formato del supermercado y la aplicación de ofertas, la Ilustración 9 señala:

(1) Las tasas de discrepancia desagregadas según zona, se ordenan de forma creciente: zona Oriente (10.6%), Centro/Poniente (11.3%) y Sur/Periferia (20.4%).
(2) Columnas correspondientes a aquellas observaciones con oferta, muestran para las tres zonas, mayor probabilidad de inconsistencia en precios.
(3) El formato (híper, estándar) no muestra una dirección clara.
(4) En zonas oriente y sur/periferia, compras con oferta parecen tener un impacto positivo en la tasa de discrepancia, versus aquellos productos con precio normal.

Para más detalles sobre los indicadores de tasas de discrepancias obtenidos, ver Anexo F.

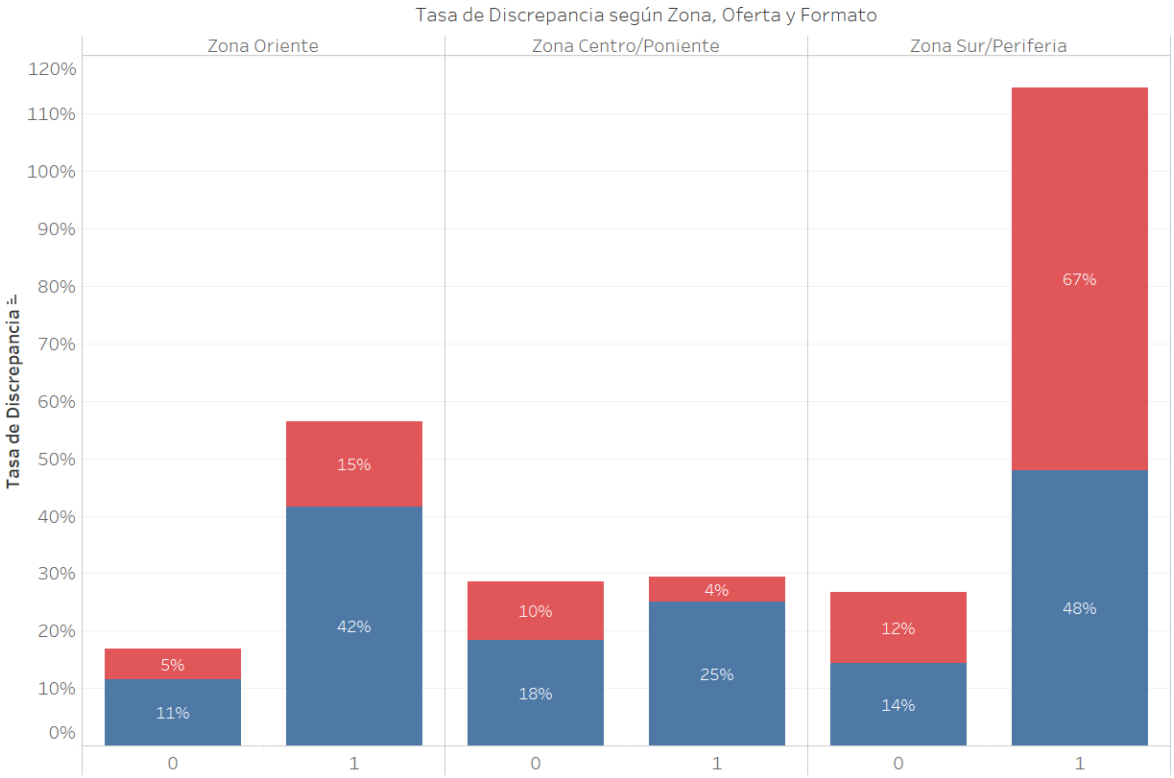


Ilustración 9: Tasa de discrepancia desagregado a partir de la zona geográfica, formato y oferta de productos

En términos operacionales, los supermercados distinguen diferentes demandas operacionales tanto a nivel semanal como mensual. Particularmente, es de conocimiento común que, la mayor carga y demanda de compras en salas de supermercado se concentre durante el fin de semana. Los resultados visibles en Ilustración 10, sin embargo, no muestran diferencias sustanciales a lo largo de la semana, y en particular se visualizan menores tasas promedio de inconsistencias durante los fines de semana. Para lo cual el día martes sería quien reporta mayores irregularidades ($\bar{X}_{TD}$=19%) y el domingo quien presenta menos ($\bar{X}_{TD}$=10%). Una potencial explicación nace en respuesta a posibles planificaciones (reflejado en el orden en salas, colocación de etiquetas de precios, limpieza, etc) anticipando el alto flujo esperado de ventas durante el fin de semana.

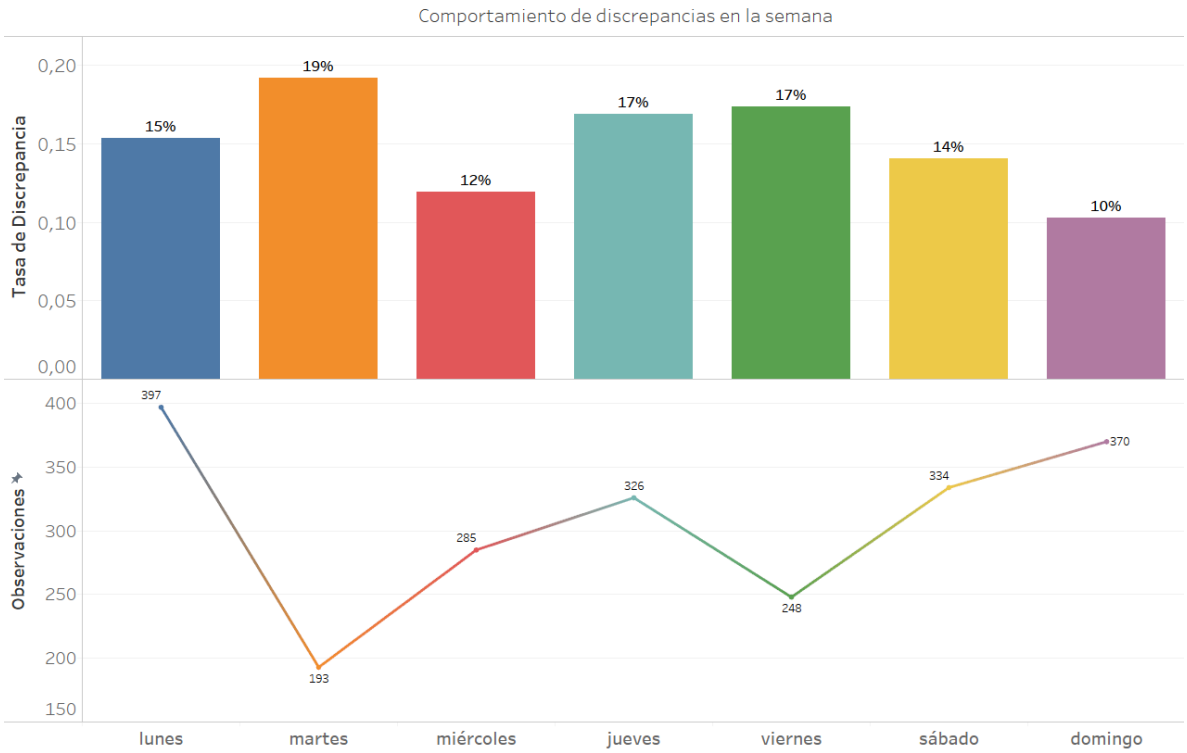


Ilustración 10: Discrepancias a lo largo de la semana

Finalmente, una desagregación a partir de las categorías de productos según su nivel de perecibilidad sugiere mayores niveles de inconsistencia de precios en aquellos productos de naturaleza perecible (Ilustración 11). En adición, no se visualizan diferencias sustanciales entre categorías, por lo que, preliminarmente, no habría una correlación entre el nivel de perecibilidad de los productos y la probabilidad de mal cobro de productos en caja.

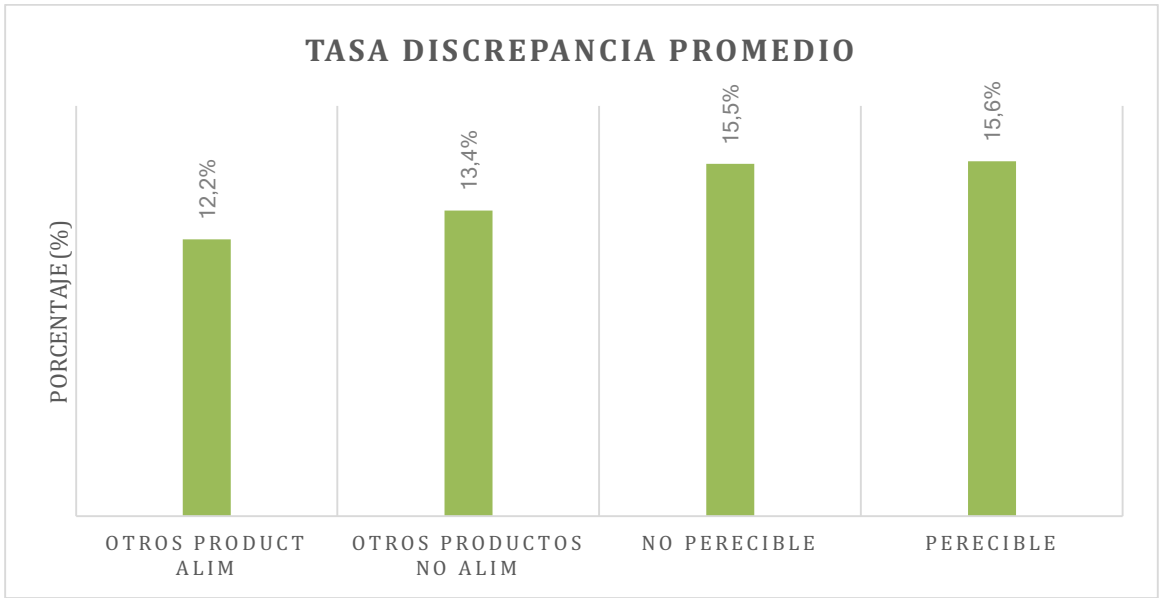


Ilustración 11: Tasa de discrepancia según categorización de productos por nivel de perecibilidad

3.2.3.3. Fecha de colocación de etiquetas de precio

Los datos utilizados para este análisis corresponden a un subconjunto de la matriz completa, dado que sólo algunas cadenas de supermercado declaran la fecha en que la etiqueta de precio fue colocada en su correspondiente fleje (total de 635 datos).

En este marco, el siguiente histograma (Ilustración 12) ilustra el comportamiento de tasas de discrepancias encontradas, a medida que el lapsus de tiempo (Δt) aumenta. Los datos disponibles se distribuyen en el histograma en un rango entre [0, 120] días, con una mayor concentración entre los 0-15 días, y un menor porcentaje después de los 70 días. La curva graficada, representa el porcentaje de productos con inconsistencias, de lo cual se observa un crecimiento positivo a medida que el lapsus de tiempo es mayor. Lo anterior responde a una lógica natural, pues, la permanencia de un precio para un producto dado durante un tiempo significativo debiese inducir a una mayor probabilidad de discrepancia entre el precio cobrado y el declarado.

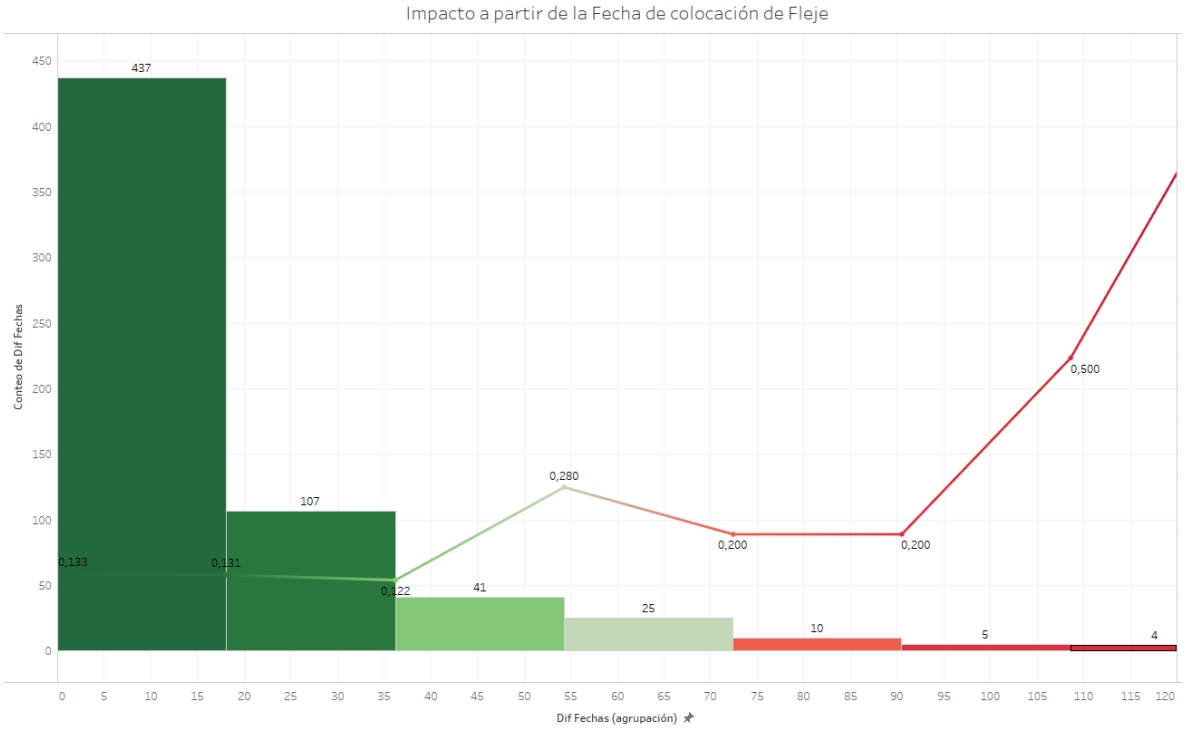


Ilustración 12: Histograma y comportamiento de ratios de discrepancias a partir de lapsus entre fecha de actualización de precio y fecha de compra

Luego, las inferencias a partir del análisis exploratorio anterior se complementan mediante un test ANOVA, que busca responder en este caso, si efectivamente existe una diferencia **significativa** en la probabilidad de discrepancia a medida que el lapsus de tiempo entre compra-actualización de precio incrementa. Los resultados de la Tabla 4 evidencian una diferencia significativa (sig ≈ 1,9%), por lo que se ratifica un impacto considerable a partir del lapsus de tiempo existente.

Test Anova

Dif_Fechas	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	3.236	1	3235,984	5,562	**0,019**
Dentro de grupos	368.281	633	581,803		
Total	371.517	634			

Tabla 3: Test ANOVA para diferencias en discrepancias a partir de lapsus en la actualización de precios

3.2.3.4. Valencia

Asimismo, en un contexto donde dos entidades se ven involucradas en una irregularidad de precio, puntualmente al momento de generar una compra, existe un interés por responder sobre cuál de ellas es la perjudicada y sobre si tal perjuicio logra ser significativo o no. Los descriptivos apoyados mediante la Ilustración 13, ilustran las distribuciones de los "porcentajes de diferencia relativos en precio" (histograma acompañado de la distribución de densidad de probabilidad mejor ajustada). Recordar que, para cada discrepancia encontrada, es posible obtener una desviación porcentual del precio, el cual puede tomar valores negativos o positivos (e.g. el rango entre 0-20% indicarían un precio cobrado 0-20% menos de lo declarado). Por construcción de esta variable, valores negativos indicarían un perjuicio del cliente y valores positivos, un perjuicio de la entidad supermercadista.

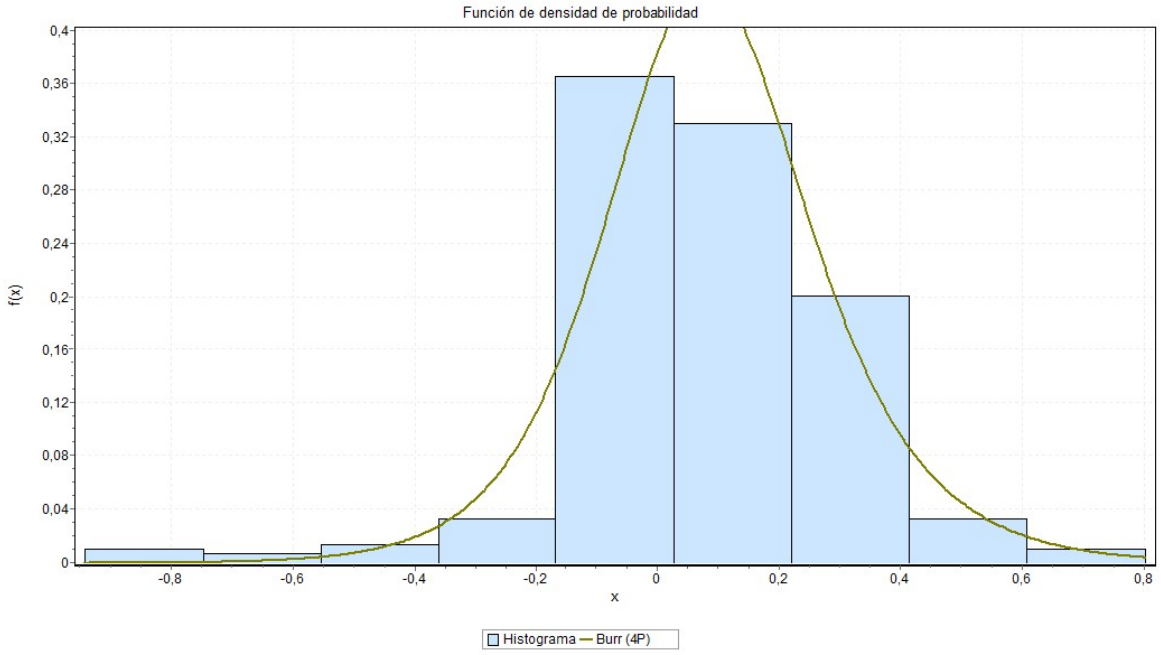


Ilustración 13: Histograma sobre distribución de desviaciones porcentuales de precios en casos con discrepancia acompañado de función de densidad de probabilidad de mejor ajuste

Los resultados muestran una carencia de simetría en la distribución de los datos y más bien indican una inclinación hacia el área positiva (para más detalle sobre el ajuste de función de densidad, ver gráfico P-P presente en Anexo G). Lo anterior indicaría un mayor perjuicio del supermercado, en relación a la contraparte (clientes), lo cual es consistente con la proporción

de discrepancias a favor de una entidad u otra, encontrado en los estadísticos descriptivos generales. En adición, el análisis se complementa mediante un test de medias que busca responder sobre la significancia en relación a las desviaciones porcentuales en precio, para aquellos casos inconsistentes. La siguiente tabla 5, resume los resultados arrojados, en donde se concluye que, en términos del porcentaje de desviación del precio cobrado relativo al exhibido, se habría favorecido al cliente en términos agregados.

Test de Medias	
Media %Desv Favor Super	12,68%
Media %Desv Favor Cliente	17,91%
N° Discr Favor Super	93
N° Discr Favor Cliente	222
Varianza Desv Favor Super	3,7%
Varianza Desv Favor Cliente	2,5%
T(estadistico)	2,31
gl(v)	147
t(alfa=5%,gl)	1,97
Rechazo(T>t)	Si

Tabla 4: Test de medias para desviación porcentual de precios con inconsistencia

3.2.4. Minería de Datos

En este apartado se opta por utilizar modelos supervisados, ya que la variable objetivo es identificable, tratándose de la variable dicotómica *discrepancia*. Algunos de los requerimientos para los modelos supervisados a utilizar considera: la interpretabilidad de resultados, capacidad explicativa, restricciones de la variable objetivo.
De la misma forma, los modelos buscan servir como herramienta para identificar factores relacionados con la irregularidad en el cobro de precios en supermercados y eventualmente, lograr encontrar patrones causales (coherente con el conjunto de hipótesis enunciado).

Es necesario hacer la distinción que, los modelos supervisados (particularmente de clasificación) utilizados en esta investigación poseen un fin descriptivo y no predictivo. Por lo tanto, indicadores de calidad de los modelos, no se basan en su capacidad predictiva (e.g. precisión, especificidad).

Finalmente, se presenta un análisis a partir de los datos transaccionales dispuestos, que permiten enriquecer la investigación. Luego se concluye a partir de los modelos utilizados, sobre el conjunto de variables que logran una correlación significativa con el fenómeno investigado.

3.2.4.1. Modelo Logit (inicial)

Modelo logit o también llamado, modelo de regresión logística corresponde a un método supervisado de regresión, en donde la variable dependiente es de naturaleza discreta (ya sea dicotómica o multinominal). Tiene como objetivo establecer el efecto que tienen ciertas variables explicativas sobre la probabilidad de ocurrencia de una cierta elección. Para su construcción se requiere de datos asociados a las variables de influencia (independientes) y a la variable objetivo (dependiente), para lo cual finalmente se obtiene una serie de parámetros que corresponden a los coeficientes de cada variable explicativa (asociado a la magnitud del efecto que poseen estas variables sobre la dependiente). Matemáticamente es posible expresar la estimación del modelo logit mediante:

$$\ln\left(\frac{\Pr(y=1|\boldsymbol{X})}{1-\Pr(y=1|\boldsymbol{X})}\right)=\beta\boldsymbol{X} \qquad \begin{aligned}\Pr(y=1|\boldsymbol{X}) &= \frac{e^{\beta\boldsymbol{X}}}{1+e^{\beta\boldsymbol{X}}}\\ &\equiv G(\beta\boldsymbol{X})\end{aligned}$$

Donde,
βX representa una probabilidad
G(.)toma valores entre [0,1]
y: variable dependiente discreta

En este marco, una primera aproximación mediante un modelo logit, es utilizado para identificar factores correlacionados con la variable discrepancia. Resulta apropiado su uso dado el alto nivel de interpretabilidad de este modelo de clasificación (coeficientes y significancia). En este marco, se procede a utilizar seis variables, cinco de naturaleza categóricas y una continua de la forma:

$$Discrepancia_i = \beta_0 + \beta_{Oferta}\cdot Oferta_i + \beta_{Precio}\cdot Precio\ Unitario_i + \sum_j \beta_j \cdot Supermercado_{ij} + \sum_k \beta_k \cdot Zona_{ik} + \sum_l \beta_l \cdot Día_{il} + \sum_m \beta_m \cdot Categoría\ Producto_{im}$$

En esta aproximación se decide incorporar variables como el precio individual del producto (exhibido), dada una potencial dependencia con productos con ticket promedio más alto. Las demás variables utilizadas, ya fueron argumentadas en secciones anteriores. En Anexo H es posible visualizar un subconjunto de los datos tipo panel, utilizados para construir el modelo.

Los resultados numéricos se pueden apreciar en la siguiente Tabla 5, los cuales indican la existencia de variables significativas: **zona**, **supermercado** y **oferta** (coeficientes con p-valor menores a 5% de significancia).

Discrepancia	Variable	Beta	Error estándar	gl	Sig.	Exp(B)
1	Intersección	0,352	0,224	1	0,12	
	[Zona=1]	-0,848	0,213	1	0,00	0,428
	[Zona=2]	-0,556	0,178	1	0,00	0,574
	[Zona=3]	0b	.	0	.	.
	[Supermercado=Express de Lider]	-0,51	0,249	1	0,04	0,601
	[Supermercado=Hiper Lider]	-0,978	0,232	1	0,00	0,376
	[Supermercado=Jumbo]	-0,987	0,236	1	0,00	0,373
	[Supermercado=Santa Isabel]	-1,158	0,344	1	0,00	0,314
	[Supermercado=Tottus]	-0,944	0,237	1	0,00	0,389
	[Supermercado=Tottus Express]	-0,854	0,214	1	0,00	0,426
	[Supermercado=Unimarc]	0b	.	0	.	.
	[Categ_Prod_Agr=No perecible]	-0,13	0,154	1	0,40	0,878
	[Categ_Prod_Agr=Otros product alim]	-0,303	0,209	1	0,15	0,739
	[Categ_Prod_Agr=Otros productos no alim]	-0,242	0,197	1	0,22	0,785
	[Categ_Prod_Agr=Perecible]	0b	.	0	.	.
	[Oferta=0]	-1,111	0,145	1	0,00	0,329
	[Oferta=1]	0b	.	0	.	.

a La categoría de referencia es: 0.

b Este parámetro está establecido en cero porque es redundante.

Tabla 5: Resultados a partir de regresión logística

Con relación al signo de los coeficientes obtenidos, se encuentra una consistencia con algunas de las inferencias sugeridas mediante el análisis descriptivo. En particular, aquella zona mayormente correlacionada con discrepancias de precios corresponde a la zona periferia/sur, en contraste de la zona de Santiago oriente con la menor tasa de discrepancia a nivel agregado. Finalmente, si bien se encuentra una lógica parcialmente consistente a los descriptivos encontrados y a la intuición formalizada en el conjunto de hipótesis.

3.2.4.2. Árbol de Decisión

Árboles de decisión corresponden a modelos supervisados de clasificación, siendo uno de los más utilizados para inferencia inductiva. El método consiste en clasificar instancias, ordenándolos a partir de un nodo raíz hasta cierto(s) nodo(s) hoja (*leaf nodes*) lo que finalmente entrega alguna clasificación de la instancia. Cada nodo especifica un test de algún atributo de la instancia y cada rama que desciende de aquel nodo, representa algún posible valor de aquel atributo. El algoritmo para construir los nodos y hojas se construye según distintos criterios. Algunos de los algoritmos en árboles de decisión más utilizados son ID3, ASSISTANT, CART, entre otros. Dentro de las ventajas de los modelos de árbol en relación a modelos de regresión logística se encuentra: la versatilidad en la interacción entre variables y la alta interpretabilidad sin tener la necesidad de conocimientos altamente técnicos. Otra

ventaja tiene relación con el hecho que relaciones no lineares entre parámetros, no afectan el desempeño (de estimación) de los árboles.

Para efectos prácticos de la investigación, se probará el algoritmo CART para estimar la probabilidad de discrepancia según variables explicativas. Estas son: cadena de supermercado, localización geográfica (zona), oferta, período de la compra (día semanal), categoría y precio producto. Los resultados arrojados se aprecian en Anexo I, de lo cual es posible obtener las siguientes inferencias:

- Compras en supermercados localizados en zona sur/periferia y oriente de Santiago, particularmente de productos en oferta, muestran un desempeño significativamente inferior y deficiente en términos del cobro consistente de precios. En efecto, se sugiere que un 46,5% de las ocasiones de compra (nodo 6), habría existido una diferencia entre el precio cobrado y el declarado en góndola.
- Una otra combinación de factores relacionados fuertemente con el mal cobro de precios encontrados, se halla en función de productos presentes en supermercados Express de Líder y Unimarc, particularmente ubicados en zona centro/poniente (nodo 13).
- En contraste, productos sin promociones u ofertas en supermercados Jumbo son aquellos que mostraron una menor proporción de productos con inconsistencia de precios (nodo 7 y ramificaciones), con tasas en torno al 5-6%, lo cual habla de una potencial interacción entre la cadena Jumbo y el manejo de productos con y sin oferta.

3.2.4.3. Modelo Logit con interacción

Finalmente, utilizando los aprendizajes obtenidos a partir de ambos modelos anteriores; modelo logit y árbol de decisión, se procede a generar un nuevo modelo de regresión logística, esta vez incorporando interacciones entre variables. Aquellas interacciones se hacen cargo del último análisis a partir del modelo de árbol de decisión, en donde se identifican cadenas en zonas con niveles críticos en términos del mal cobro de precios. Particularmente, se agrupan las siguientes variables esperando que, en conjunto, tengan un impacto significativo en la probabilidad de inconsistencia:

1. Zona Oriente – Productos con/sin oferta
2. Zona Sur/Periferia – Productos con/sin oferta
3. Cadena Express de Líder – Zona Sur/Periferia

Los resultados obtenidos a partir de un modelo de regresión logística con interacción de variables (ver detalle en Anexo J) revelan la dirección en que las siguientes variables correlacionaron con la probabilidad de inconsistencia de precio en compras de supermercado:

(1) Productos en oferta adquiridos en zona oriente y sur/periferia se mueven en la dirección positiva.
(2) Express de Líder en zona sur/periferia se mueve en dirección negativa
(3) Tasas de discrepancia encontradas se ordenan en forma creciente según zona: oriente, centro/poniente, sur/periferia, respectivamente

(4) Con relación a los demás factores considerados como la cadena, se encuentra una similitud con relación a lo concluido a partir del modelo logit sin interacción.

3.2.4.4. Evaluación de Modelos

Un gráfico de Característica Operativa del Receptor (curva ROC) es una técnica utilizada para visualizar, organizar y evaluar clasificadores basados en su desempeño. Este tipo de gráficas se ha utilizado bastamente en teorías de detección compensando tasas de ocurrencia versus tasas de falsa alarma en clasificadores (Egan, 1975; Swets et al., 2000). En otras palabras, curvas ROC capturan un *tradeoff* entre beneficios (verdaderos positivos) y costos (falsos negativos), paras las distintas combinaciones de valores límites existentes (i.e. *threshold*). Luego, para comparar el desempeño de distintos clasificadores, se utiliza como método la maximización del área bajo la curva, conocido como AUC (Bradley, 1997; Hanley and McNeil, 1982). Se entiende que, una curva ROC asociada a un clasificador, con mayor área bajo la curva tendrá un mejor desempeño promedio (para cualquier valor crítico). En este contexto, se estima apropiado el uso de curvas ROC para la comparación de los tres modelos de clasificación construidos anteriormente.

El desempeño de clasificadores, sin embargo, no es concluyente sin una evaluación en base a medidas de dispersión (Fawcett, 2006). En otras palabras, una curva ROC no es lo suficientemente robusta para concluir sobre el desempeño de un clasificador. Una solución más robusta considera la simulación de curvas (para un mismo clasificador) para luego obtener una curva "promedio" que logre capturar el comportamiento mejor estimado. En este caso particular, se aplica un cálculo de promedio a partir de simulación de valores críticos. Los resultados se visualizan en la Ilustración 14, para cada uno de los clasificadores. De lo anterior se tiene que:

(1) Los clasificadores se ordenan de peor a mejor según el área bajo la curva (AUC) de la forma: modelo logit sin interacción, modelo de árbol de decisión (CRT) y modelo logit con interacción. El último, por ende, se trataría del clasificador con mejor desempeño promedio.

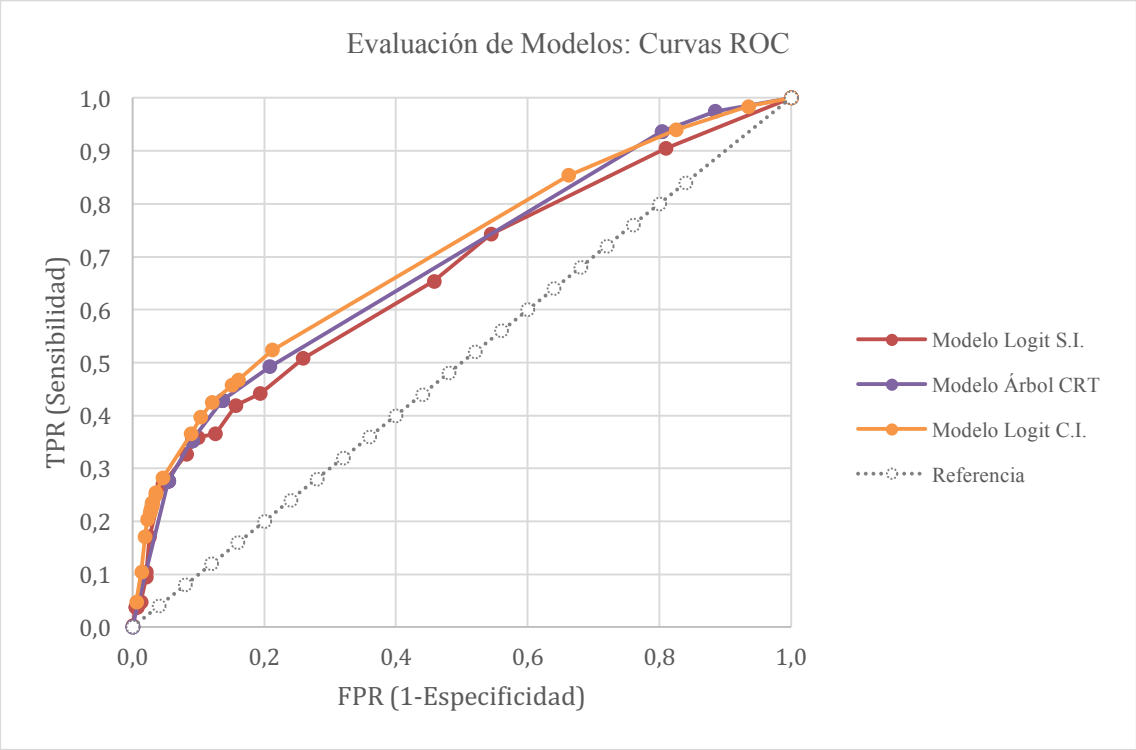


Ilustración 14: Curvas ROC obtenida a partir de modelos de clasificación

3.2.4.5. Análisis mediante datos transaccionales

En el contexto del uso de datos transaccionales para una empresa de supermercados y el análisis de inconsistencias en precios, resultados preliminares mediante el uso de estadísticos descriptivos, son exhibidos en el siguiente recuadro (Tabla 6):

Estadísticos Descriptivos: Cruce con datos transaccionales			
Métrica	*Descripción breve*	*Cálculo*	*Valor*
$\bar{X}_{TD}$	Tasa de discrepancia promedio	$\frac{1}{n}\sum_{i}^{n} 1_{\{x_i\}}$	**16,1%**
$\%n_{FS}$	Proporción de discrepancias a favor de supermercado.	$\frac{\sum_i^n 1_{\{PG_i - PB_i < 0\}}}{\sum_i^n 1_{\{x_i=1\}}}$	**72%**
$\%n_{FC}$	Proporción de discrepancias a favor de cliente.	$\frac{\sum_i^n 1_{\{PG_i - PB_i > 0\}}}{\sum_i^n 1_{\{x_i=1\}}}$	**28%**

Tabla 6: Estadísticos Descriptivos mediante uso de datos transaccionales

Del recuadro anterior, destaca: (1) el orden de magnitud de la tasa de discrepancia agregada en datos transaccionales se ajusta a las mediciones encontradas mediante la base de datos inicial. (2) Se observa un contraste de resultados en relación a la valencia, en donde en este caso, la mayor proporción de mal cobros de precios apunta al perjuicio del cliente (72% de las ocasiones). Lo último, sin embargo, no logra ser concluyente, pues corresponde a información de tan solo una sala de una cadena de supermercado particular, pero efectivamente sugiere sobre la heterogeneidad existente en relación a la valencia de discrepancias.

Una otra observación apunta a una potencial relación entre el número de fluctuaciones de precios de un cierto producto y la propensión en presentar una inconsistencia de precios. En otras palabras, fluctuaciones de precios corresponde a una medida de cuantas etiquetas de precios se han asociado a un cierto producto en un cierto tiempo. El siguiente gráfico (Ilustración 15) muestra un caso ejemplo de un producto en particular con alto nivel de recambio de precios cobrados:

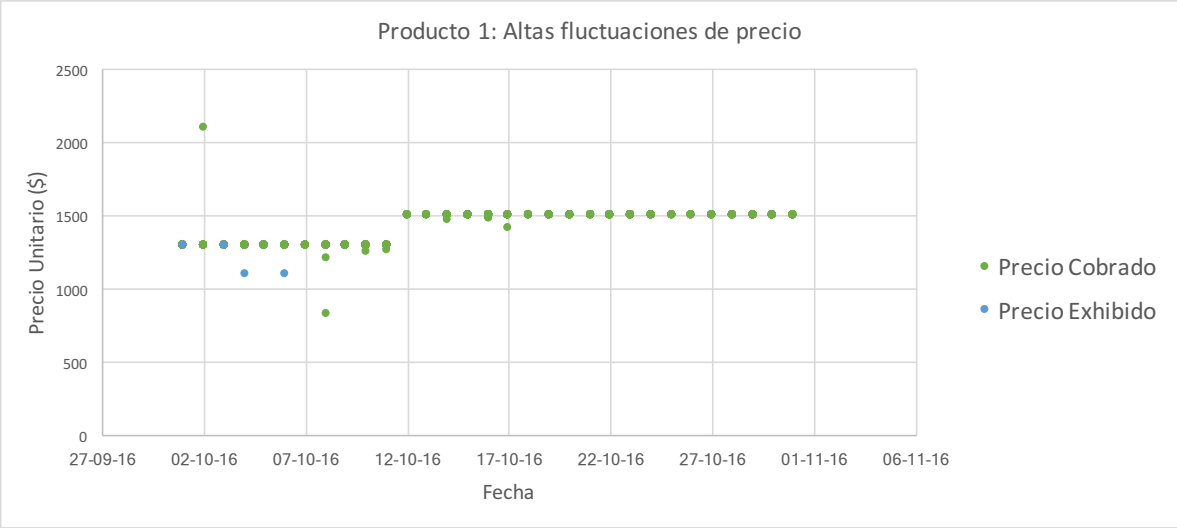


Ilustración 15: Evolución de precios de un producto ejemplo con altas fluctuaciones. Fuente: Datos transaccionales

Tal producto exhibe 12 precios diferentes en el período de un mes, y se registra un total de 159 ocasiones con mal cobro del producto. En contraste, la siguiente gráfica (Ilustración 16) denota el comportamiento de precios cobrados y exhibidos en el tiempo. En este caso, se registra tan solo dos magnitudes de precios (cobrados) y en la fecha estudiada, de un total de 415 transacciones se registraron solo dos cobros inconsistentes respecto a lo exhibido.

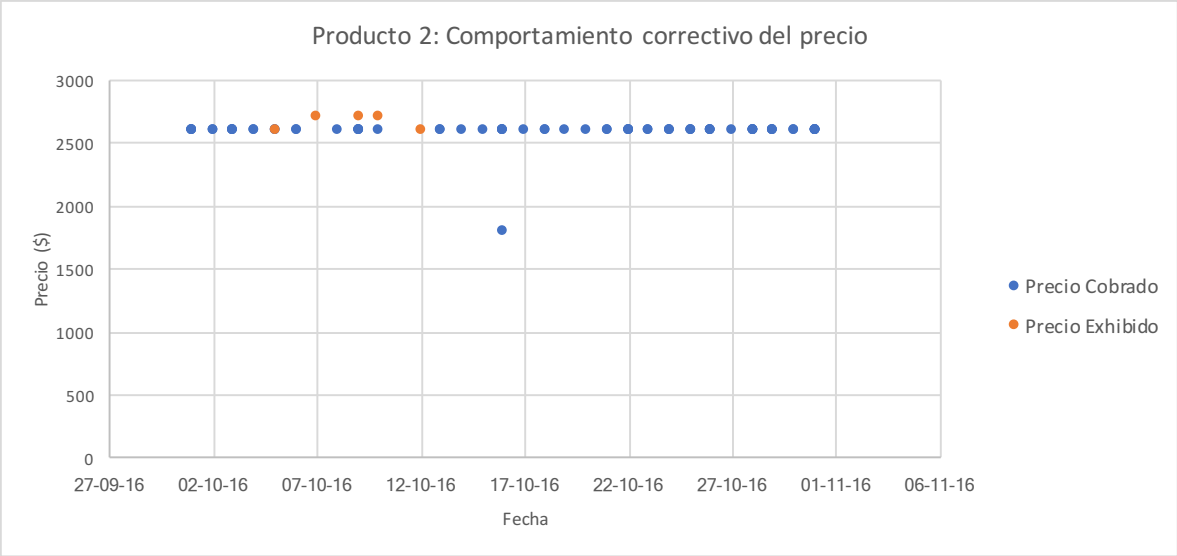


Ilustración 16: Evolución de precios de un producto ejemplo con bajas fluctuaciones. Fuente: Datos transaccionales

En este marco, es de interés un análisis sobre potenciales correlaciones entre tasas de discrepancias y número de fluctuaciones en precios. La gráfica siguiente (Ilustración 17) exhibe un mapa de dispersión acorde a un seguimiento de 59 productos, en donde se exhibe el número de fluctuaciones versus la tasa de discrepancia encontrada para cada uno de ellos (obtenida como el cociente entre nº de casos con cobro distinto al exhibido en góndola y el nº de transacciones registradas en tal período).

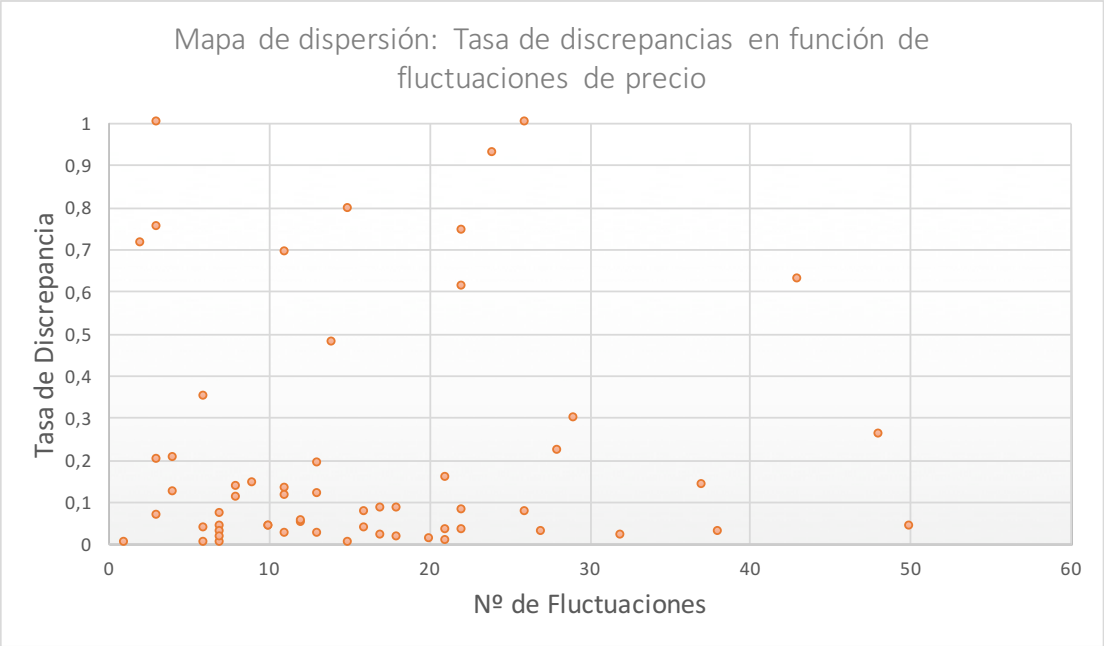


Ilustración 17: Mapa de dispersión discrepancias vs fluctuaciones de precios. Fuente: Datos transaccionales

La mayor cantidad de datos se concentra entre 0-20 fluctuaciones y tasa de discrepancia en torno al 0-20%. Luego, la gráfica señala una leve relación positiva entre ambas variables, aunque no concluyente. El número de observaciones en tal mapa, limita el análisis, sin embargo, deja abierto el cuestionamiento para potenciales trabajos futuros.

3.3. Evaluación de Resultados

Los múltiples modelos estadísticos utilizados en las secciones anteriores son útiles, particularmente permiten responder la validez del conjunto de hipótesis planteado. Si bien el análisis agregado entrega descriptivos e inferencias de interés, la significancia de los efectos encontrados es posibles de respaldar mediante las técnicas de análisis a nivel individual. De los resultados obtenidos se observa:

(1) Una concordancia entre inferencias a partir de descriptivos y modelos de clasificación, identificando variables significativas que impactan en la probabilidad de inconsistencia y finalmente permiten responder sobre la validez del conjunto de hipótesis enunciado:
   a. Existen diferencias significativas a partir de cadenas de supermercado.
   b. Compras de productos con o sin oferta en supermercados ubicados en zonas geográficas particulares demuestran correlación significativa.
   c. Inexistencia de diferencias significativas a partir del día semanal de compra y la categoría de productos (según perecibilidad).

(2) Coherencia a partir de los tres modelos de clasificación obtenidos. Se encuentran factores común denominadores.

(3) Un análisis mediante curvas ROC sugiere un mejor ajuste de los datos, mediante un modelo logit con interacción de variables (zona geográfica y oferta de productos).

En adición, los resultados mediante el uso de datos transaccionales respaldan, (4) el orden de magnitud de la proporción de discrepancias promedio entorno a un 16%, (5) existencia de

productos con distintas frecuencias en la actualización de precios (fluctuaciones). Se identifican productos con alto nivel de recambio de precios, y se estima, tenga un impacto en la probabilidad de discrepancia. Lo anterior queda como línea futura de trabajo, dado restricciones de recursos dispuestos para la investigación, destacando la exhaustividad en el procesamiento de datos obtenidos de forma presencial en sala.

4. CONCLUSIONES

La presente investigación tiene como objetivo analizar el fenómeno de inconsistencias en el cobro de productos en supermercados nacionales, en un contexto en que hasta ahora, no ha sido diagnosticado y caracterizado en estudios formales. La inexistencia de investigaciones pasadas, motiva la elaboración de una metodología para el desarrollo de un análisis del mal cobro de productos. Particularmente se construye una base con 2.128 datos que considera la participación de seis cadenas de supermercados, en salas divididas en tres zonas geográficas de la Región Metropolitana.

Adicionalmente, la investigación es respaldada mediante el acceso y uso de una base de datos transaccionales en compras para una cadena de supermercado de gran participación en la industria supermercadista nacional. En este contexto, destaca el uso de herramientas estadísticas, tales como test de hipótesis, regresiones y particularmente modelos de clasificación que permiten caracterizar y concluir sobre un conjunto de hipótesis planteado en el presente estudio. Dentro de los resultados primordiales, se tiene:

(1) Datos recolectados y transaccionales aseguran una tasa de discrepancia promedio en supermercados entorno a un 14%-16%.
(2) No existe evidencia que fundamente un beneficio por parte de los supermercados (ganancia) generalizado en casos con discrepancias. Particularmente el conjunto de datos recolectados sugiere contrariamente, un favorecimiento agregado por parte de los clientes (frecuencia y monto).
(3) A partir del análisis descriptivo y el uso de modelos de clasificación en datos recolectados se tiene:
    a. Existe una heterogeneidad en discrepancias a lo largo de las cadenas de supermercado:
        i. Peor desempeño: Unimarc y Ex. Lider
        ii. Mejor desempeño: Jumbo y Sta. Isabel
    b. Supermercados ubicados en zona oriente (1) demuestra menores probabilidad de discrepancia.
    c. Se encuentra una correlación positiva entre productos con oferta y probabilidad de discrepancia, y un efecto cruzado entre zonas y la aplicación de ofertas.
    d. Se evidencia una propensión a un aumento en discrepancias a medida que la fecha de actualización física del precio se aleja de la fecha de compra.
(4) El análisis de datos transaccionales permite clasificar productos según el nivel de recambio de precios mensual, y se admite una posible relación positiva con la probabilidad de discrepancia.

La evidencia encontrada y analizada sugiere una correlación del fenómeno a partir de deficiencias operacionales en la actualización física de precios. De esta forma, la implementación de mejoras operacionales en la actualización de precios (e.g. flejes electrónicos, mejoras en protocolos de revisión) podría inducir a una disminución en pérdidas a causa de este tipo de irregularidad operacional, no sólo del tipo monetarias, sino que también por perjuicios en la percepción de la marca.

Para lo anterior, se sugiere:

1. Identificación estratégica de items críticos (mayores tasas de discrepancias) a partir de una medición inicial. El estudio apoya una correlación con aquellos items que frecuentan en promociones, y con alto nivel de fluctuación (mensual) de precios.
2. Uso de herramientas tecnológicas para la medición y monitoreo físico de precio de productos en sala (e.g.pistola scanner) en tiempo real.
3. Uso de datos transaccionales para la obtención de métricas de desempeño, desagregando por categoría de productos, temporalidad y salas de estudio.

Finalmente, este estudio deja espacio para futuros trabajos, principalmente mediante el uso de datos transaccionales y motiva la inclusión de tasas de discrepancias como nueva métrica de desempeño operacional en supermercados, dado el posible impacto en la percepción de calidad de servicio por parte del cliente.

5. BIBLIOGRAFÍA

[1] Diario Pyme. 2016. Cliente denuncia como operaría Sistema de precios en supermercado [en línea] <http://www.diariopyme.com/cliente-denuncia-como-operaria-sistema-de-precios-en-supermercado-me-senti-estafado/prontus_diariopyme/2016-03-07/181025.html> [consulta: 04/04/2016]

[2] Lichtenstein, D., Ridgway, N., & Netemeyer, R. (1993). Price Perceptions and Consumer Shopping Behavior: A Field Study. Journal of Marketing Research, 30(2), 234-245

[3] Valarie A. Zeithaml (1988). Consumer Perceptions of Price, Quality and Value: A Means-End Model and Synthesis of Evidence, Journal of Marketing 52(3), 2-22.

[4] CERET. 2015. Medición de calidad de servicio percibido en la industria del retail supermercados [en línea] <http://www.ceret.cl/estudios-realizados/1-estudio-calidad-de-servicio-en-supermercado-2015/> [consulta 06/08/2016].

[5] Egan, J.P. (1975). Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York.

[6] Swets, J.A., Dawes, R.M., Monahan, J. (2000). Better decisions through science. Scientific American 283, 82–87.

[7] Daniel McFadden (2000). Statistical Tools for Economists (1), 155-158.

[8] William H. Greene (2002). Econometric Analysis (5), 7-18, 719-728.

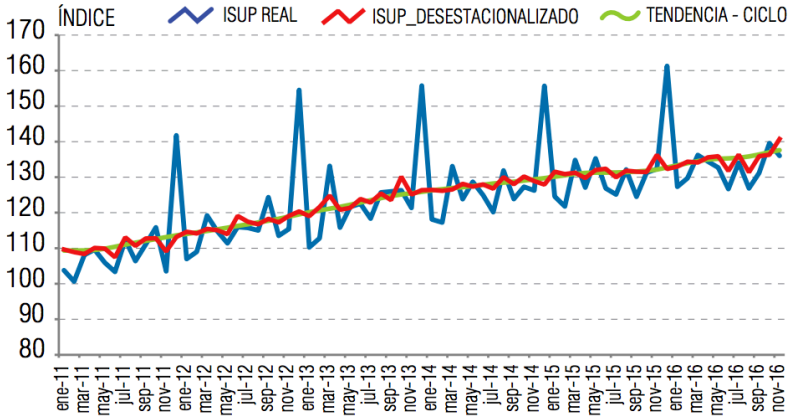
[9] Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

[10] Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29 – 36.

[11] Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

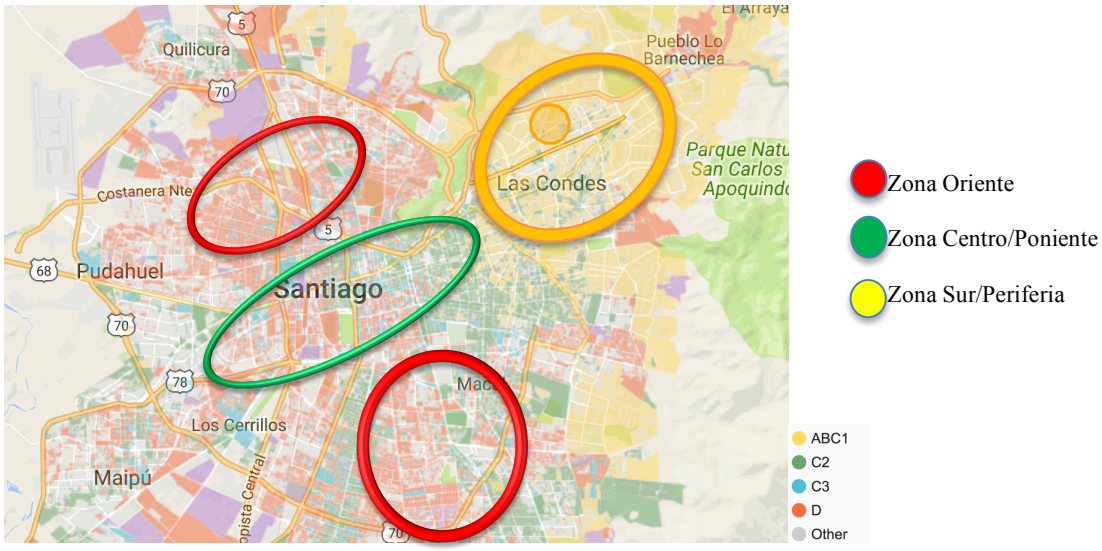
6. ANEXOS

Anexo A: Índice de ventas de supermercados, desestacionalizada y tendencia (enero 2011-noviembre 2016)



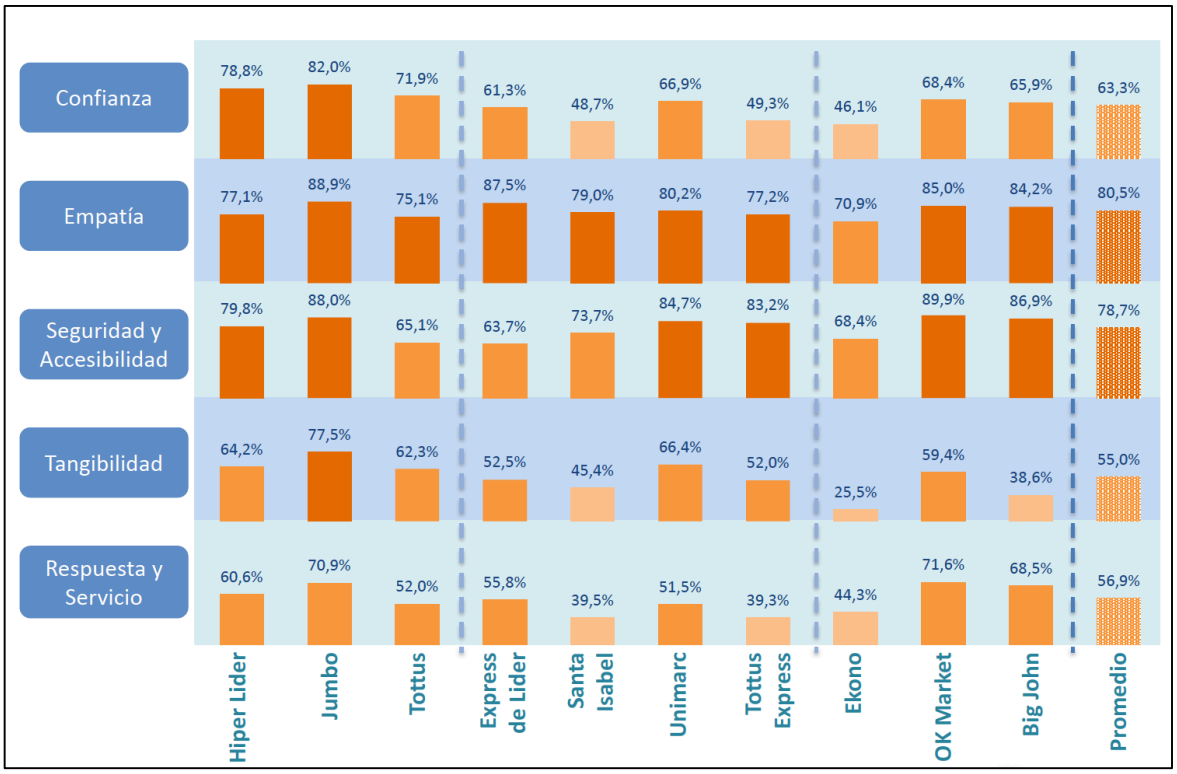
(*) Serie empalmada.
(**) La serie desestacionalizada y de tendencia-ciclo está construida utilizando la metodología X-12 ARIMA.

Anexo B: Distribución geográfica a partir de nivel socioeconómico en zona metropolitana de Chile.

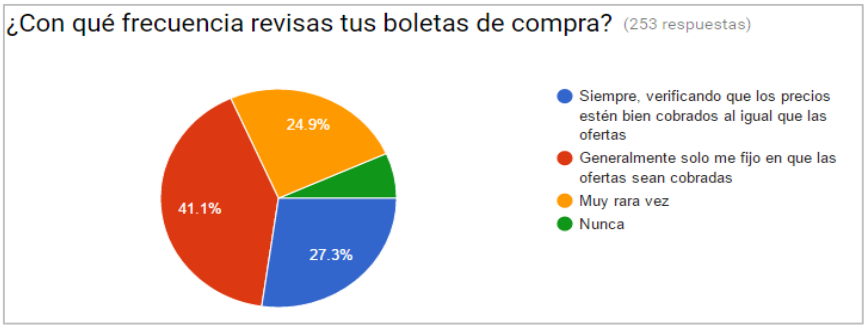


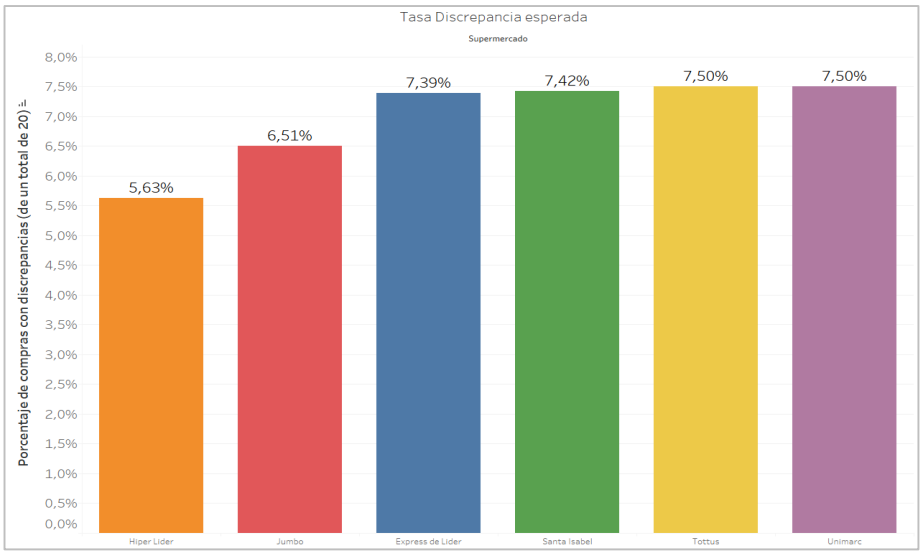
Fuente: http:www.mientorno.cl.
Información actualizada según censo 2012

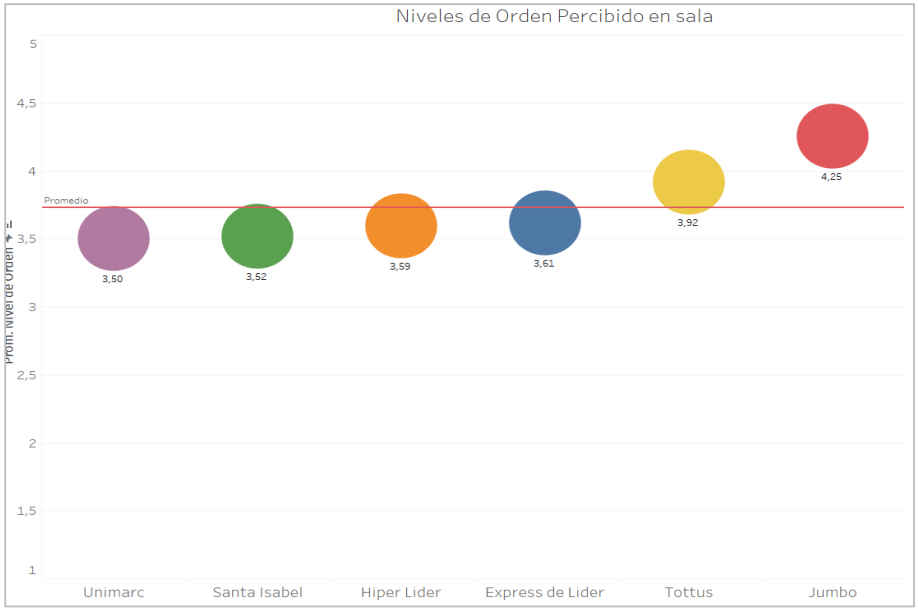
Anexo C: Medición de calidad de servicio en el retail supermercados desagregado según dimensiones



Anexo D: Resultados de encuesta sobre percepción de supermercados


¿Con qué frecuencia revisas tus boletas de compra? (253 respuestas)
24.9%
41.1%
27.3%
Siempre, verificando que los precios estén bien cobrados al igual que las ofertas
Generalmente solo me fijo en que las ofertas sean cobradas
Muy rara vez
Nunca


Tasa Discrepancia esperada
Supermercado
Porcentaje de compras con discrepancias (de un total de 20)
8,0%
7,5%
7,0%
6,5%
6,0%
5,5%
5,0%
4,5%
4,0%
3,5%
3,0%
2,5%
2,0%
1,5%
1,0%
0,5%
0,0%
5,63%
6,51%
7,39%
7,42%
7,50%
7,50%
Hiper Lider
Jumbo
Express de Lider
Santa Isabel
Tottus
Unimarc


Niveles de Orden Percibido en sala
Prom. Nivel de Orden
5
4,5
4
3,5
3
2,5
2
1,5
1
Promedio
3,50
3,52
3,59
3,61
3,92
4,25
Unimarc
Santa Isabel
Hiper Lider
Express de Lider
Tottus
Jumbo

Anexo E: Ficha técnica referente a recolección de datos

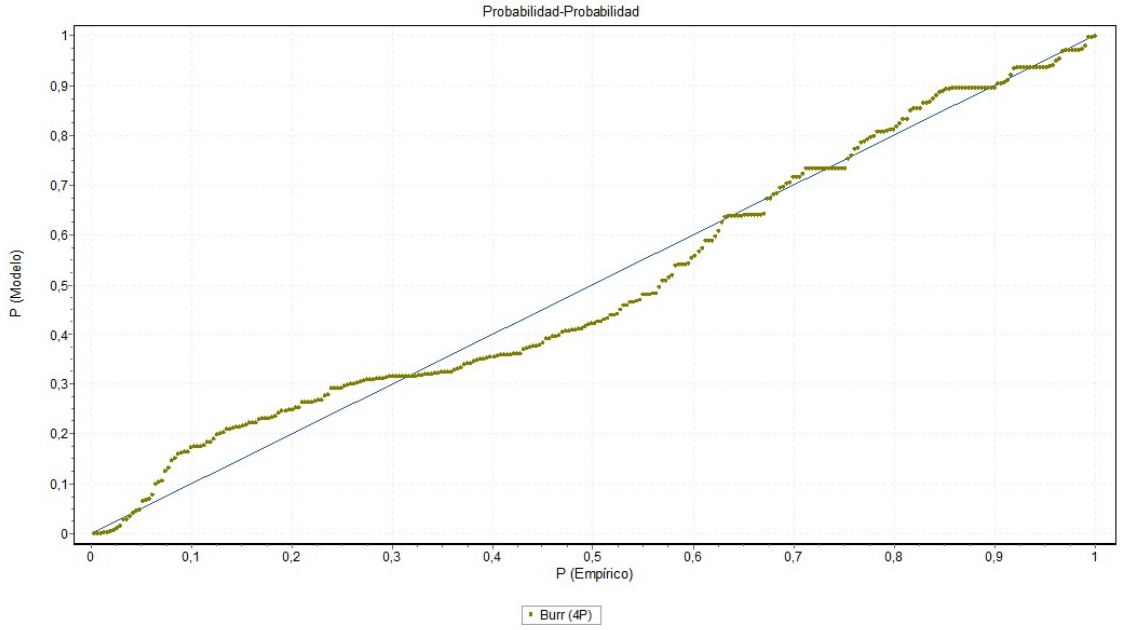
Levantamiento de Datos

Métrica	Valor
Tamaño muestral (n)	2153
Zonas involucradas	3
Cadenas estudiadas	7
Período de muestreo	Mayo-Octubre 2016

Anexo F: Detalles tasas de discrepancias desagregadas según la zona geográfica y oferta de productos

Zona	***Formato***	***Oferta***	***Tasa Discrepancia***	
Zona Oriente	Tradicional	1	10,6%	41,7%
	Tradicional	0		11,5%
	Hiper	1		14,8%
	Hiper	0		5,3%
Zona Centro/Poniente	Tradicional	1	11,3%	25,0%
	Tradicional	0		18,3%
	Hiper	1		4,4%
	Hiper	0		10,2%
Zona Sur/Periferia	Tradicional	1	20,4%	47,9%
	Tradicional	0		14,3%
	Hiper	1		66,7%
	Hiper	0		12,4%

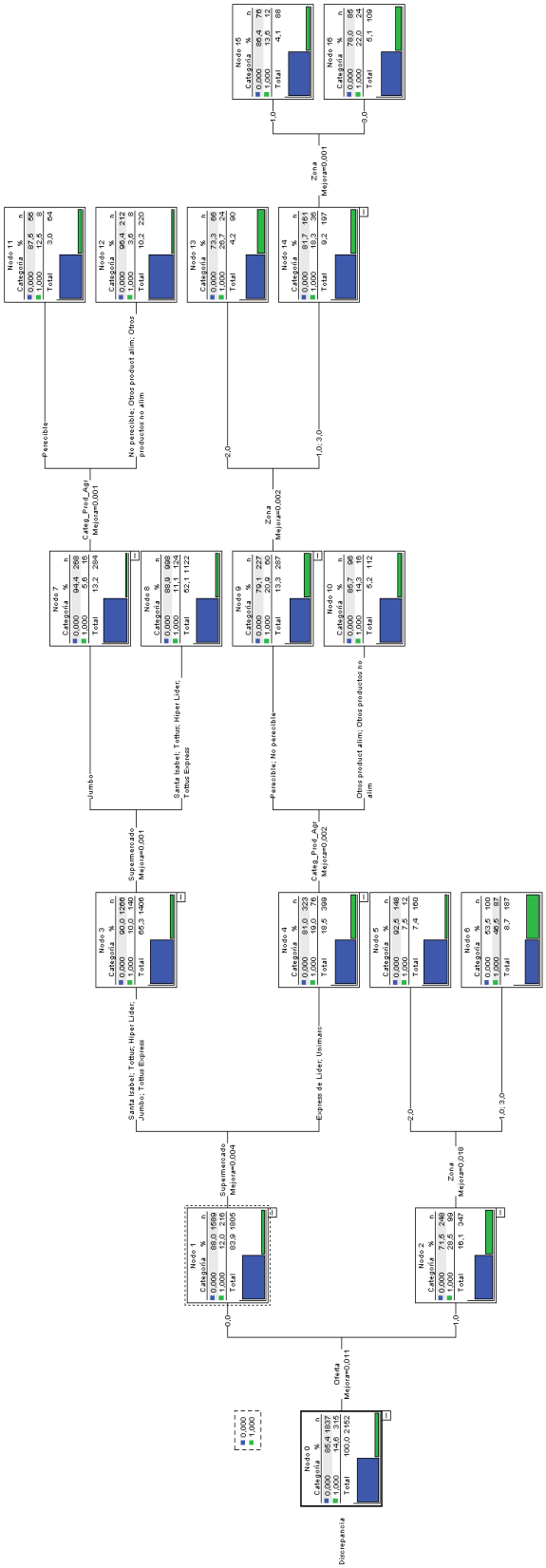
Anexo G: Gráfico P-P respecto al ajuste de función de densidad de probabilidad referente a la desviación en precio para productos con inconsistencia



Anexo H: Subconjunto de panel de datos construidos partir de levantamiento en sala

Zona	Cadena	Supermercado	Día	Categ_Prod_Agr	Oferta	Prec_Gon	*Discrepancia
2	Cencosud	Santa Isabel	domingo	Perecible	0	279	0
2	Cencosud	Santa Isabel	domingo	Perecible	0	149	0
2	Cencosud	Santa Isabel	domingo	No perecible	0	899	0
2	Cencosud	Santa Isabel	domingo	No perecible	0	819	0
2	Cencosud	Santa Isabel	domingo	Otros product alim	0	1499	0
2	Cencosud	Santa Isabel	domingo	Perecible	0	1789	0
2	Cencosud	Santa Isabel	domingo	No perecible	1	1090	0
2	Cencosud	Santa Isabel	domingo	No perecible	0	1889	0
2	Cencosud	Santa Isabel	domingo	No perecible	0	699	0
2	Cencosud	Santa Isabel	domingo	Perecible	0	399	0
3	Cencosud	Santa Isabel	lunes	Perecible	0	2899	1
3	Cencosud	Santa Isabel	lunes	Otros product alim	0	649	0
1	Cencosud	Jumbo	martes	No perecible	0	599	0
1	Tottus	Tottus	sábado	Otros productos no alim	0	1590	1

Anexo I: Resultados Árbol de Decisión



Anexo J: Modelo de regresión logística con interacción de variables

Discrepancia[a]	Variable	B	Error estándar	gl	Sig.	Exp(B)
	Intersección	-1,33	0,79	1	0,093	
	[Zona1*Oferta=0]	2,03	0,50	1	0,000	7,628
	[Zona1*Oferta=1]	0[b]		0		
	[Zona3*Oferta=0]	2,73	0,39	1	0,000	15,311
	[Zona3*Oferta=1]	0[b]		0		
	[ExpressLider*Zona3=0]	-1,97	0,59	1	0,001	0,140
	[ExpressLider*Zona3=1]	0[b]		0		
	[Zona=1]	0,86	0,26	1	0,001	2,373
	[Zona=2]	0,17	0,20	1	0,393	1,187
	[Zona=3]	0[b]		0		
	[Supermercado=Express de Lider]	0,22	0,28	1	0,424	1,247
0	[Supermercado=Hiper Lider]	0,92	0,24	1	0,000	2,511
	[Supermercado=Jumbo]	1,14	0,25	1	0,000	3,119
	[Supermercado=Santa Isabel]	1,21	0,35	1	0,001	3,363
	[Supermercado=Tottus]	0,85	0,24	1	0,000	2,342
	[Supermercado=Tottus Express]	1,03	0,23	1	0,000	2,801
	[Supermercado=Unimarc]	0[b]		0		
	[Categ_Prod_Agr=No perecible]	0,12	0,16	1	0,441	1,130
	[Categ_Prod_Agr=Otros product alim]	0,29	0,21	1	0,174	1,338
	[Categ_Prod_Agr=Otros productos no alim]	0,13	0,20	1	0,504	1,144
	[Categ_Prod_Agr=Perecible]	0[b]		0		
	[Oferta=0]	-0,53	0,32	1	0,097	0,587
	[Oferta=1]	0[b]		0		

a. La categoría de referencia es: 1.

b. Este parámetro está establecido en cero porque es redundante.