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

ESTIMACIÓN DE MODELOS UTILIZANDO CONJUNTOS DE CONSIDERACIONES LATENTES CONSTRUIDOS A PARTIR DE DATOS HISTÓRICOS

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL

BENJAMÍN IGNACIO GUERRERO JORQUERA

PROFESOR GUÍA:

ÁNGELO GUEVARA CUE

MIEMBROS DE LA COMISIÓN:

MARCELA MUNIZAGA MUÑOZ

RICARDO HURTUBIA GONZÁLEZ

SANTIAGO DE CHILE

2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TITULO DE: INGENIERIA CIVIL
POR: BENJAMIN GUERRERO JORQUERA
FECHA: 2020
PROF. GUÍA: ÁNGELO GUEVARA CUE

ESTIMACIÓN DE MODELOS DE ELECCIÓN UTILIZANDO CONJUNTOS DE CONSIDERACIONES LATENTES CONSTRUIDOS A PARTIR DE DATOS HISTÓRICOS

Los modelos de elección son cruciales para el análisis de sistemas de transporte pues ayudan a predecir la demanda de proyectos usualmente muy costosos, para así determinar su viabilidad social y/o privada. Un supuesto fundamental en la modelación de elecciones es que el investigador sabe exactamente el conjunto de consideración (alternativas sopesadas) de cada persona en su elección. Sin embargo, esto es difícil de cumplir en la práctica, ya que el investigador no sabe cómo piensa cada persona. y debe hacer supuestos que, de ser erróneos, se traducen en la imposibilidad de recuperar consistentemente los estimadores del modelo, comprometiendo su validez en interpretación y pronóstico.

El objetivo de este trabajo es avanzar hacia el desarrollo de una teoría que explique resultados obtenidos en un trabajo anterior que sugieren que, al usar un conjunto de consideración basado en alternativas experimentadas (elecciones anteriores de la persona), los parámetros de un modelo de elección pueden ser estimados consistentemente, aunque no se conozca el conjunto real de consideración.

De esta manera, este trabajo comienza por estudiar, usando simulaciones de Monte Carlo, las condiciones bajo las cuales el método basado en alternativas experimentadas puede estimar el conjunto de consideración real. Luego, se analiza una base de datos con un historial de elecciones para comprobar si el resultado se manifiesta en un escenario práctico. Finalmente, el trabajo propone un planteamiento preliminar que podría sentar las bases para el futuro desarrollo de una demostración formal que explique los resultados obtenidos.

A partir de los resultados de las simulaciones, usando un método basado en los días de medición para generar el conjunto de experiencias históricas y limitando la cantidad de alternativas consideradas por el usuario, se encuentra que la cantidad de alternativas totales no influye en el cumplimiento de la hipótesis.

Al analizar los datos reales de compras de supermercado, se encontró que, si bien el modelo basado en las experiencias históricas tiene una mejor bondad de ajuste en estimación y predicción que un modelo en el que se asume que el usuario considera todas las alternativas, tiene la limitación que imputa probabilidad nula a todas las alternativas nuevas. Para subsanar este problema se propuso y probó una metodología preliminar que no mostró resultados satisfactorios, y se discutieron posibles mejoras para investigaciones futuras en este tema.

Con los resultados obtenidos se puede concluir que el uso de conjuntos de consideración basados en alternativas experimentadas permite la creación de modelos de elección que predicen de igual o mejor manera los conjuntos de consideración reales de los usuarios en comparación con técnicas usadas en la actualidad. Sin embargo, aún se requiere avanzar en el desarrollo de demostraciones formales que permitan sustentar este resultado en la práctica, en casos realistas.

Agradecimientos

En primer lugar, quiero agradecer a mis papás, quienes no solo me inspiraron para estudiar esta profesión, sino que también me dieron su apoyo incondicional desde que nací. También quiero agradecer a mi hermano Javier, que me ha acompañado y alegrado el día por muchos años, siendo compañero de juegos y de conversaciones. También quiero agradecer a la Sra. Carmen, quien ayudó en el cuidado de la casa por 18 años y a quien quiero como si fuera mi abuela.

Además, quiero agradecer a todos mis compañeros de colegio y universidad, especialmente a los compañeros de la rama de Ingeniería de Transporte, los cuales siempre han estado dispuestos a ayudarme cuando los necesitaba, y han hecho estos años de estudio más soportables.

También quisiera agradecer a los profesores tanto del colegio como de la Universidad, quienes siempre se preocuparon de que sus alumnos aprendieran y crecieran para ser mejores personas. En particular, quisiera agradecer a mi profesor guía, que me sugirió realizar esta línea de trabajo, y que me ha apoyado en cada paso de este proyecto.

Por último, quiero agradecer a nuestra gata Nanita, la cual me ha acompañado a menudo durante la confección de este informe, y que siempre me hace feliz con lo cariñosa que es.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

Capítulo 1: Introducción

1.1 Motivación

En una ciudad, el sistema de transporte es vital debido a que permite la accesibilidad de las personas a los servicios existentes en la ciudad, y también las interacciones entre personas que no viven en lugares cercanos. Es por esto que, para muchas personas, la calidad del sistema de transporte genera un gran impacto tanto en el ámbito económico como en el emocional.

Los movimientos de vehículos y peatones que se observan en el sistema de transporte son el resultado de un gran número de elecciones de parte de las personas que deciden la ruta que van a usar para ir de un lugar a otro. Por esto, es fundamental entender cómo los usuarios realizan estas elecciones, y generar modelos para predecirlas, con el fin de implementar políticas que mejoren el sistema de transporte.

El método más usado para el modelamiento de elecciones discretas, formalizado por McFadden (1974), considera que el proceso de generación de datos se basa en el Modelo de Utilidad Aleatoria (RUM, por sus siglas en inglés). Dentro de este método, el modelo más utilizado es el modelo Logit.

Un supuesto fundamental del modelo de elección discreta es que el analista sabe todas las alternativas consideradas por el usuario. Este supuesto es especialmente cuestionable cuando el conjunto de alternativas es muy grande, como ocurre, por ejemplo, en problemas de elección de ruta o de elección de productos de compras. Esto se debe a que, según se explica en Hauser (2014), existe un costo al considerar una alternativa extra, causado por la limitación cognitiva del individuo (es decir, que a una persona le es difícil comparar un gran número de alternativas).

Para superar esta limitación, Manski (1977) propuso un enfoque de conjunto latente que, si bien resuelve el problema desde un punto de vista teórico al dar cuenta explícita de las limitaciones del analista, resulta en una formulación con muy poca utilidad práctica. Ben-Akiva, M., & Boccara, B. (1995) proponen una mejora práctica del modelo de Manski que permite reducir sustancialmente su complejidad, pero al costo de incorporar una serie de nuevos supuestos no verificables sobre la heurística que el usuario usa para construir su conjunto de consideración.

Utilizando simulaciones de Monte Carlo, Villalobos (2017) probó diferentes combinaciones de procesos de generación de datos y métodos de estimación, para luego calcular la consistencia de cada uno de estos métodos, mediante el uso del test-t, o mediante cobertura empírica con un 75% de confianza (es decir, el número de simulaciones en las cuales se recuperan los parámetros originales).[1] Los resultados obtenidos sugieren la siguiente hipótesis: El uso de conjuntos de consideración basados en las experiencias anteriores puede simular con buena certeza los conjuntos de consideración usados por los usuarios.

De esta forma, en este trabajo se busca validar este resultado, y además ver si éste puede ser aplicado en datos reales.

[1] El procedimiento de simulación usado por Villalobos se explica con mayor detalle en el Capítulo 3.

1.2 Objetivos

El objetivo principal del trabajo es demostrar la validez del uso de conjuntos de consideración basados en las experiencias anteriores de los usuarios para la generación de modelos de elección discretos.

Para que esto se cumpla, se proponen los siguientes objetivos específicos:

- Analizar la prevalencia del resultado obtenido por Villalobos mediante el uso de simulaciones de Monte Carlo, para formular una hipótesis sobre los casos en el que este es válido.

- Ilustrar una aplicación de este método en un caso con datos reales.

El primer objetivo específico planteado es, esencialmente, ver si el resultado obtenido por Villalobos es replicable, y luego ver con detalle si este se mantiene bajo diferentes condiciones (cambio del número de alternativas disponibles, cambio del conjunto de consideración, generación de datos, etc). Esto va a colaborar con la formulación de una hipótesis que señala las posibles condiciones bajo las cuales se cumple que el modelo de experiencias anteriores permite simular los conjuntos de consideración.

El segundo objetivo consiste en comprobar la hipótesis mediante la revisión de datos obtenidos de mediciones de preferencia, con la meta de estudiar un caso práctico.

1.3 Metodología

En primer lugar, antes de empezar el trabajo, hay que asumir un marco de modelación tal que el usuario construya un conjunto de consideración, desconocido para el investigador. Luego, se debe estudiar cómo se realiza el método de simulación de Monte Carlo en algún lenguaje computacional, para su uso en el trabajo. Para esto, se pueden usar códigos ya existentes.

Para cumplir el primer objetivo específico, se usa la experimentación de Monte Carlo para demostrar la robustez de varios métodos que el investigador puede considerar para la generación de conjuntos de consideración de rutas. El grado de robustez de los métodos va ser determinado en función a su habilidad de recuperar los parámetros β_k de la población, sin importar el verdadero proceso de generación de datos que podría estar detrás del conjunto de consideración observado (es decir, que los parámetros estimados sean consistentes). Específicamente, se intenta recrear la conjetura obtenida por Villalobos: que, al considerar un conjunto de consideración derivado de las elecciones pasadas del usuario, se pueden encontrar estimadores consistentes de los parámetros originales. Luego, se modifican diferentes parámetros de la simulación para ver si estos influyen en el cumplimiento de la conjetura.

Para cumplir el segundo objetivo, se debe observar un caso en el que la hipótesis se cumple en la realidad. Esto se va a hacer mediante el uso de una base de datos extraída de un "supermercado" online, que consiste de más o menos 3 millones de diferentes órdenes hechas por más de 200.000 usuarios diferentes. Con estos datos, se puede determinar si un método basado en las órdenes históricas de los usuarios predice de mejor manera la última elección que otros métodos. El análisis realizado puede ser extrapolado a la elección de rutas (ya que se cumple el mismo sistema de elección de alternativas).

1.4 Estructura de esta memoria

Esta memoria tiene 5 capítulos, comenzando por el capítulo actual que contiene la introducción del trabajo de título. El segundo capítulo describe la revisión bibliográfica de textos asociados al tema de interés, integrada por dos temas diferentes: modelación de elecciones y modelación del conjunto de consideración. En el tercer capítulo se presenta el estudio de la robustez del método basado en alternativas experimentadas para modelar el conjunto de consideración de un individuo, usando simulaciones de Monte Carlo. En el cuarto capítulo se analiza una base de datos observados para verificar si el método basado en alternativas experimentadas funciona en una situación práctica. Finalmente, en el quinto capítulo se presentan las conclusiones del trabajo, y además se muestran recomendaciones metodológicas y extensiones que se pueden realizar a partir de este trabajo.

Capítulo 2: Revisión Bibliográfica

2.1 Introducción

La revisión bibliográfica para la realización de este trabajo se basó fundamentalmente en dos áreas de la literatura: teoría de elecciones discretas, y modelación de conjuntos de consideración.

2.2 Modelación de elecciones discretas

El marco teórico en el que se sustenta este trabajo es el de elecciones discretas, donde lo que se hace es modelar el proceso de elección de alternativas para entender y predecir el comportamiento de las personas. El método más usado para el modelamiento de elecciones, formalizado por McFadden (1974), considera que el proceso de generación de datos es gobernado por el Modelo de Utilidad Aleatoria (RUM, por sus siglas en inglés) que se puede ver en la ecuación (2.1), en la que N agentes (personas) enfrentan la elección de una alternativa i de entre J alternativas en el conjunto de consideración C_n:

$$U_{in} = \sum_k \beta_k x_{kin} + \varepsilon_{in} = V_{in} + \varepsilon_{in} \quad n = 1, \dots, N \;\; ; \;\; i \in C_n \qquad (2.1)$$

$$y_{in} = 1\left[U_{in} = \max_{j \in C_n} \{U_{jn}\}\right]$$

Se asume que el agente n percibe una utilidad U_{in} si escoge la alternativa i. Esta utilidad se divide en una parte discreta V_{in}, que depende de forma lineal de los atributos x_{kin} con parámetros β_k, y en una parte aleatoria ε_{in}. Ahora, la utilidad no se puede medir. Lo que sí se puede medir son los atributos x_{kin}, y la variable de elección y_{in}, la cual es 1 si la ruta i le da la mayor utilidad al agente y 0 en caso contrario.

Usualmente, se asume que la utilidad aleatoria ε_{in} distribuye Gumbel o Valor Extremo Tipo 1. Esto da vida al modelo Logit, que se caracteriza por la ecuación de probabilidad (2.2):

$$P_{in} = \frac{\exp(\mu * U_{in})}{\sum_{j \in C_n} \exp(\mu * U_{jn})} \qquad i \in C_n \qquad (2.2)$$

Aquí, μ representa el parámetro de escala, que generalmente se fija en 1, y P_{in} representa la probabilidad de que la persona n escoja la alternativa i.

Como se vio en el Capítulo 1, otro supuesto importante que se hace en elecciones discretas es que el modelador conoce el conjunto de consideración de las personas, lo cual es sumamente cuestionable en un contexto donde las alternativas son muchas y en donde no se sabe con certeza el pensamiento de cada individuo al escoger una alternativa.

2.3 Modelación de conjunto de consideración

En esta sección se van a describir los trabajos que estudian la modelación de los conjuntos de consideración, tanto en teoría como en campos específicos, como por ejemplo la elección de ruta.

En primer lugar, existe una línea de trabajo originalmente desarrollada por Manski (1977), en donde se realiza un modelo en el cual se asumen que los conjuntos de consideración son conocidos, para poder hacerse cargo de los posibles errores que causa no conocer el verdadero conjunto de consideración de cada tomador de decisión. En la ecuación (2.3) se muestra la fórmula usada, que es, en esencia, una fórmula de probabilidades condicionales. Sin embargo, es prácticamente imposible conocer la probabilidad de cada posible conjunto de consideración para evaluar dicha expresión.

$$P_n(i) = \sum_{C \in G_n} P_n(i|C) * P_n(C) \qquad (2.3)$$

En donde $P_n(i)$ es la probabilidad de que la persona n escoja la alternativa i. C es un conjunto de consideración y G_n es un conjunto que contiene todos los conjuntos de consideración posibles. Luego, $P_n(i|C)$ es la probabilidad de que la persona n escoja la alternativa i si su conjunto de consideración es C, y $P_n(C)$ es la probabilidad de que el conjunto de consideración de la persona n sea C.

Tiempo después, Ben-Akiva y Boccara (1995) aplicaron una versión simplificada del método de Manski, en la cual se usan datos de preferencias declaradas. A partir de estos, se puede generar una expresión para las probabilidades de cada posible conjunto de consideración, por lo que resulta posible usar la ecuación (2.3). Sin embargo, se encontró que este método solo obtenía una leve mejora con respecto a un modelo logit en el que se consideran todas las alternativas (Conjunto de consideración Completo), a cambio de un gran gasto en términos computacionales.

Una alternativa a la fórmula de Manski fue propuesta por Martinez et al. (2009), la cual consiste un modelo de elección conocido como Logit Multinomial Restringido (CMNL, por sus siglas en inglés), cuyas funciones de utilidad y probabilidad se define de la siguiente manera:

$$U_{ni} = V_{ni} + \frac{1}{\mu}\ln(\phi_{ni}) + \varepsilon_{ni} \qquad (2.4)$$

$$P_{ni} = \frac{\phi_{ni} * \exp(\mu * V_{ni})}{\sum_{j \in C} \phi_{nj} * \exp(\mu * V_{nj})} \qquad (2.5)$$

En la cual el parámetro ϕ_{ni} representa la probabilidad de la persona n considere la alternativa i. Este término puede ser definido mediante un logit binario:

$$\phi_{ni}(X_{nik}; u_k, \omega_k) = \frac{1}{1 + \exp\big(\omega_k * (X_{nik} - u_k)\big)} \qquad (2.6)$$

En primer lugar, X_{nik} representa la k-ésima variable que afecta ϕ_{ni} (es decir que afecta que la persona n considere la alternativa i). Luego, u_k es la restricción superior de la variable X_{nik} y ω_k es el parámetro de escala del logit binario. De forma similar para una restricción inferior l_k:

$$\phi_{ni}(X_{nik}; l_k, \omega_k) = \frac{1}{1 + \exp\big(-\omega_k * (X_{nik} - l_k)\big)} \qquad (2.7)$$

La ventaja de este método es que es más fácil de estimar que la fórmula de Manski, ya que no requiere enumerar los conjuntos de consideración. Sin embargo, una investigación realizada por Bierlaire et al. (2010) reveló que, debido a que se realiza una serie de supuestos sobre la forma de la función de utilidad, este modelo no permite modelar la generación conjuntos de consideración de una forma consistente a la fórmula de Manski.

Por otro lado, en el área de marketing, se estudian los conjuntos de consideración con respecto a la compra de productos. Entre estos trabajos destaca el de Hauser (2014), el cual menciona que es esencial que un producto se encuentre en el conjunto de consideración de una persona, porque de lo contrario esa persona no va a considerar el producto en su compra. Además, se mencionan distintos tipos de heurísticas que un comprador puede usar para determinar su conjunto de consideración, y métodos usados para determinar qué heurística usa cada persona.

Finalmente, en el área de elección de rutas de transporte, se aprecia que en el artículo de Bekhor et. Al (2006) se estudian algoritmos basados en el método conocido como "Labeling Approach", en el cual se calcula un consto generalizado para luego determinar el conjunto de consideración a partir de las rutas que tienen menor costo. En efecto, la mayoría de métodos mostrados en esta área se basa en el modelo de k-ésimas rutas mínimas, el cual asume que el usuario va a considerar las rutas que más le convienen en términos de costo y tiempo.

Finalmente, se debe destacar la memoria hecha por Villalobos (2017), la cual tuvo como objetivo principal caracterizar el conjunto de consideración en elección de ruta con el propósito de comprobar que los métodos usados por otros expertos sean realmente útiles.

Las simulaciones hechas por Villalobos estudiaron el funcionamiento de cinco modelos de determinación de conjuntos de consideración, para cuatro heurísticas que pueden ser usadas por los usuarios. Los métodos y las heurísticas, así como el procedimiento hecho, se describen con más detalle en el Capítulo 3. A continuación, se muestra la Tabla 2.1, que muestra los métodos usados y sus abreviaciones, y la Figura 2.1, la cual resume los resultados de la simulación:

Tabla 2.1 Métodos generadores del conjunto de consideración supuesto.

Abreviación	Método
CV	Conjunto Verdadero
CE	Conjunto Experimentado
KRM_V	Conjunto K-ésima ruta mínima (parámetros verdaderos)
KRM_E	Conjunto K-ésima ruta mínima (parámetros erróneos)
CC	Conjunto Completo
CLB	Conjunto Logit Binario

Fuente: Villalobos (2017).

Métodos\Escenarios	I. Completo	II. Aleatorio	III. Eliminación por Aspectos	IV. CLB
i. CC				
ii. KRM_V				
iii. KRM_E				
iv. CLB				
v. CE				

Figura 2.1 Resultados de las simulaciones de Monte Carlo realizadas por Nicolás Villalobos (2017), con 4 heurísticas de elección (columnas) y 5 métodos de elección asumidos (filas).
Fuente: "Accounting for the Consideration-Set in Discrete Choice Model Estimation." Villalobos, N., & Guevara, A. (2019).

En la Figura 2.1, un cuadro marcado con color azul significa que el método de generación supuesto logra obtener parámetros consistentes con el conjunto de consideración real del usuario (generado por la heurística).

Este resultado sugiere, a diferencia de la literatura anterior, que los métodos basados en k-ésimas rutas mínimas no funcionan bien al modelar el conjunto de consideración, dependiendo de la heurística, o de si los parámetros usados en la modelación no corresponden exactamente a los parámetros reales (lo que es el caso más probable, ya que el modelador usualmente no sabe cuáles son estos). En cambio, se sugiere que el conjunto de alternativas experimentadas (CE) permite hacer un mejor modelo.

En una extensión de su trabajo posterior a la memoria, Villalobos investigó que es lo que pasaba si el costo monetario y el tiempo de cada ruta variaban en el tiempo. Él encontró que, si bien el método basado en alternativas experimentadas fallaba en este caso (debido a que los datos usados para generar el conjunto de consideración de alternativas experimentadas eran diferentes a los datos usados para generar el conjunto de consideración real), métodos modificados a partir del método CE generan mejores resultados.

Así, en resumen, esta rama de la literatura se presentan las investigaciones que expertos en diferentes sectores han hecho acerca de la determinación de los conjuntos de consideración, así como los métodos que se usan para este fin.

Capítulo 3: Estudio del Método de Experiencias Anteriores mediante simulaciones de Monte Carlo

3.1 Simulación inicial

Para comprobar la efectividad de los diferentes métodos de determinación del conjunto de consideración, se realizan simulaciones de Monte Carlo programadas en el software R (R Development Core Team. 2008). Estas simulaciones son idénticas a las realizadas en la memoria de Nicolás Villalobos (2017), por lo que la intención es comprobar si los resultados obtenidos en ese trabajo son reproducibles.

La simulación consiste en primer lugar de un proceso de generación de atributos. Con estos datos se crean cuatro escenarios de elección, donde en cada uno se genera el conjunto de consideración verdadero mediante distintos métodos o heurísticas. Luego, para cada uno de los escenarios se estiman modelos de elección utilizando distintos conjuntos de consideración supuestos por el modelador. Finalmente, se realiza un análisis de los sesgos causados por utilizar los distintos métodos de generación del conjunto de consideración supuesto para cada escenario presentado, y así determinar qué métodos se acercan más al conjunto de consideración real.

3.1.1 Generación de datos

Para cada iteración de la simulación de Monte Carlo, se genera una muestra de 2000 individuos, donde se generan parámetros de costo monetario y de tiempo para cada una de las 10 alternativas de manera tal que el costo monetario aumenta cuando el tiempo disminuye, y viceversa. Esto se hace de la siguiente forma:

- Se generan valores máximos y mínimos para el costo y el tiempo. En este caso, se tiene que $t_{minimo} = c_{minimo} = 10$ y que $t_{maximo} = c_{maximo} = 30$.
- Para cada alternativa, se calcula un valor promedio para el tiempo y el costo, usando los siguientes valores:

$$t_{prom} = t_{minimo} + (1 - id_{alt}) * \frac{t_{minimo} - t_{maximo}}{10}$$
$$c_{prom} = c_{maximo} + (1 - id_{alt}) * \frac{c_{maximo} - c_{minimo}}{10}$$

  En donde id_{alt} es el número de la alternativa.
- Luego, se calculan los siguientes valores de desviación estándar:

$$t_{desv} = 0.3 * t_{minimo}$$
$$c_{desv} = 0.3 * c_{minimo}$$

- Finalmente, con los promedios y las desviaciones estándar, se pueden generar 2000 valores de costo y tiempo usando los valores absolutos de números aleatorios generados a partir de una distribución normal.

Una vez generados los datos para cada una de las alternativas, la utilidad determinista V_{in} se encuentra con la siguiente ecuación:

$$V_{in} = \beta_c * costo_{in} + \beta_t * tiempo_{in} \qquad (3.1)$$

En la que $\beta_c = -0.2$ y $\beta_t = -0.25$.

3.1.2 Heurísticas (Métodos de generación del conjunto de consideración verdadero)
Las heurísticas utilizadas para la construcción del conjunto de consideración verdadero intentan simular la forma de pensar del individuo al armar un conjunto de consideración para realizar la elección. Se escogieron las siguientes cuatro heurísticas:

1. Conjunto Completo

Esta heurística hace que todas las personas consideren todas las alternativas disponibles, en este caso 10. Este escenario es poco realista (es decir, es muy poco probable que las personas elijan esta heurística), pero se usa con el fin de cubrir los extremos.

2. Aleatorio

Esta heurística genera el conjunto de consideración de manera aleatoria para cada persona. Esencialmente, se generan todos los conjuntos de consideración posibles para cada persona y se escoge uno de estos al azar. Tal como el escenario anterior, este escenario es poco realista.

3. Eliminación por aspectos (EBA)

En esta heurística se crea el conjunto de consideración verdadero mediante un algoritmo que se asemeja a eliminación por aspecto, eliminando las alternativas que no cumplan las siguientes restricciones:

$$c_{in} < c_{prom_n} + 0.8 * c_{desv_n} \qquad (3.2)$$

$$t_{in} < t_{prom_n} + 0.8 * t_{desv_n} \qquad (3.3)$$

Este escenario es más realista que los dos anteriores, ya que es una de las posibles heurísticas que se encuentran en la literatura al referirse a la construcción del conjunto de consideración por parte de las personas.

4. Conjunto de consideración logit binario (CLB)

La última regla consiste en utilizar un modelo logit binario, la especificación de la utilidad para este modelo se presenta en la ecuación 3.4 y la probabilidad de considerar una alternativa se calcula siguiendo la expresión de la ecuación 3.5. Además, los parámetros para calcular la utilidad son dados y se presentan en la Tabla 3.2.

$$V_{in} = \beta_1 + \beta_2 * costo_{in} + \beta_3 * tiempo_{in} \qquad (3.4)$$

$$P_n(i) = \frac{e^{\mu * V_{in}}}{e^{\mu * V_{in}} + 1} \qquad (3.5)$$

Tabla 3.1 Parámetros del conjunto de consideración CLB.

Parámetro	Valor
β_1	7.5
β_2	-0.2
β_3	-0.25

Fuente: Elaboración propia

Luego, para cada alternativa se calcula la probabilidad de ser considerada y mediante la simulación de números aleatorios uniformes entre 0 y 1, se realiza la decisión de considerar o no cada alternativa (si la probabilidad calculada es menor al número aleatorio simulado), generando de esta manera el conjunto de consideración verdadero. Este proceso resuelve el problema presente en la fórmula de Manski (1977), ya que permite encontrar las probabilidades de generar cada conjunto de consideración posible. Además, el escenario puede llegar a representar bien la realidad ya que un modelo de consideración puede actuar en representación de alguna heurística utilizada por las personas.

En todos los casos, por temas prácticos, se filtran todas las observaciones con menos de 3 alternativas consideradas, ya que algunos de los métodos utilizados para generar el conjunto supuesto requieren un mínimo de alternativas para funcionar. De esta manera, con los conjuntos de consideración generados en cada uno de los casos, se utiliza un modelo logit multinomial para generar las elecciones.

3.1.3 Métodos Generadores del Conjunto de Consideración supuesto

A continuación, se van a describir los seis métodos que el investigador asume para la generación del conjunto de consideración:

1. Conjunto Verdadero (CV)

   Básicamente, aquí se usa el conjunto verdadero que se usó para la generación de datos (sección anterior). Este escenario es muy poco realista y solo se usa para comprobar que la simulación de Monte Carlo está bien programada.

2. Conjunto Experimentado (CE)

   Este método consiste en simular suficientes elecciones de cada individuo hasta observar al menos 2 alternativas seleccionadas además de la original para cada uno. De esta manera el conjunto de consideración supuesto se compone por las alternativas que se observan como experimentadas, de manera similar a lo que sería una base de datos de serie de tiempo.

   Este método muestra ventajas en cuanto la factibilidad de su aplicación práctica. Las alternativas observadas pueden ser obtenidas mediante datos pasivos como GPS en el caso de elecciones de ruta o con el historial de compras de una tarjeta bancaria en el caso de cualquier producto.

3. Conjunto de k-ésimas rutas mínimas con parámetros verdaderos (KRM_V)

   Este método consiste en calcular la componente sistemática de la utilidad, con la misma forma funcional mostrada en la ecuación 3.1 utilizada por el modelo logit multinomial con el que se generan las elecciones, utilizando parámetros a priori. En este caso particular, los parámetros a priori son los verdaderos utilizados en el modelo de elección original. Luego, con la utilidad sistemática calculada se realiza un ranking y se eligen las 3 mayores para formar el conjunto de consideración supuesto, además si la elección original no se encuentra en este conjunto, entonces se agrega.

4. Conjunto de k-ésimas rutas mínimas con parámetros erróneos (KRM_E)

Es similar al anterior, pero se utilizan parámetros a priori diferentes al modelo de elección original, los que se ven en la siguiente tabla:

Tabla 3.2 Parámetros usados en el método KRM_E.

Parámetro	Valor
β_c	-0.2
β_t	-0.26

Fuente: Elaboración propia

Este método es de los más prácticos por su simpleza y además presenta un caso realista ya que los parámetros a priori nunca son los reales, o el modelador no estaría tratando de estimarlos.

5. Conjunto Completo (CC)

Este modelo es el más simple de modelar, ya que el conjunto de consideración supuesto incluye a todas las alternativas.

6. Conjunto de Logit Binario (CLB)

Lo que se hace en este caso, es estimar un modelo logit binario, donde a partir del conjunto de consideración verdadero, se utiliza cada alternativa como una observación. Luego, cada elección consiste en si una alternativa es considerada o no. Este método es equivalente a la heurística 4 mostrada en la sección anterior.

3.1.4 Resultados Obtenidos

A continuación, se presentan resultados que muestran el sesgo que causa cada uno de los 6 métodos generadores del conjunto de consideración supuesto (ver Tabla 3.3) bajo los 4 escenarios de generación del conjunto de consideración verdadero. Las abreviaciones de cada método generador se pueden ver en la Tabla 2.1.

En las Tablas 3.3, 3.4, 3.5 y 3.6, se presentan resultados numéricos, mostrando cuatro métricas para caracterizar el sesgo de cada método generador del conjunto de consideración supuesto. La primera métrica corresponde al sesgo promedio, el cual indica la diferencia entre el valor promedio de la razón de los parámetros estimados y el valor de la razón verdadera. La segunda corresponde al error de la raíz cuadrada de la media (RMSE), a modo de ejemplo, para un estimador insesgado, el RMSE es la raíz cuadrada de la varianza. La tercera corresponde a un test-t, donde se compara el valor de la razón de los parámetros estimados con el valor de la razón verdadera, el valor crítico para este test es 1.984 para 99 (iteraciones - 1) grados de libertad con 95% de confianza, donde un estadístico menor a este valor, representa que se acepta la hipótesis nula de que el valor estimado es estadísticamente igual al valor real. Finalmente, se presenta la cobertura empírica con un 75% de confianza, lo que muestra cuantas veces de las 100 iteraciones el valor de la razón de parámetros estimada es estadísticamente igual a la razón verdadera, a modo de ejemplo, para el método que recupera los parámetros utilizando el conjunto verdadero la cobertura empírica al 75% debe ser de alrededor de 75.

Tabla 3.3 Análisis de Monte Carlo de los conjuntos de consideración supuestos, siendo la heurística original “Aleatorio”.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.001	0.017	0.086	85
CE	-0.002	0.030	0.056	61
KRM_V	-0.004	0.020	0.193	81
KRM_E	-0.062	0.065	3.096	2
CC	-0.001	0.020	0.061	79
CLB	-0.002	0.021	0.108	80

Fuente: Elaboración propia

Tabla 3.4 Análisis de Monte Carlo de los conjuntos de consideración supuestos, siendo la heurística original “Conjunto Completo”.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.002	0.017	0.143	85
CE	-0.004	0.055	0.081	50
KRM_V	-0.005	0.036	0.138	59
KRM_E	-0.098	0.105	2.567	8
CC	-0.002	0.017	0.143	86
CLB	-0.002	0.017	0.143	86

Fuente: Elaboración propia

Tabla 3.5 Análisis de Monte Carlo de los conjuntos de consideración supuestos, siendo la heurística original “Eliminación por aspectos”.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.001	0.023	0.038	81
CE	-0.012	0.043	0.298	61
KRM_V	-0.525	0.532	5.906	0
KRM_E	-0.651	0.658	6.671	0
CC	0.134	0.134	13.383	1
CLB	0.111	0.112	9.112	11

Fuente: Elaboración propia

Tabla 3.6 Análisis de Monte Carlo de los conjuntos de consideración supuestos, siendo la heurística original “Conjunto Logit Binario”.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	0.003	0.015	0.181	91
CE	0.006	0.027	0.227	64
KRM_V	-0.174	0.182	3.175	0
KRM_E	-0.318	0.325	4.828	0
CC	0.047	0.049	3.973	35
CLB	0.011	0.019	0.669	78

Fuente: Elaboración propia

Estos resultados se pueden resumir en la siguiente figura:

Métodos\Escenarios	I. Completo	II. Aleatorio	III. Eliminación por Aspectos	IV. CLB
i. CC				
ii. KRM_V				
iii. KRM_E				
iv. CLB				
v. CE				

Figura 3.1 Resultados de las simulaciones de Monte Carlo, con 4 heurísticas de elección (columnas) y 5 métodos de elección asumidos (filas).

Fuente: Elaboración propia

En la figura anterior, si los parámetros son consistentes (lo que se evalúa verificando que el valor de la cobertura empírica sea mayor a 5), el cuadrado se marca con azul, si no, en rojo. Existen dos casos en los que la cobertura empírica es mayor a 5, pero menor a 15, los cuales son marcados en color púrpura. Nótese que, tal como en la memoria de Villalobos, el método CE (experiencias anteriores) entrega parámetros consistentes al conjunto de consideración real sin importar el conjunto de elección simulado. Así, el resultado encontrado por Villalobos puede ser reproducido y no fue resultado del azar.

3.2 Estudio y corrección del tamaño del conjunto de consideración real

Ahora, es posible que el resultado sea de esa manera debido al método usado para el cálculo. Por ese motivo, se solicitó que se modificara el código para descartar que el resultado mostrado en el capítulo anterior fuera resultado de un error en el código. Así, lo que se hizo fue analizar el código inicial, realizar cambios específicos y ejecutar el programa, para después observar si aún se mantiene la consistencia del método CE con estos cambios.

En el Anexo A, se describe un estudio preliminar de varias condiciones usando la heurística "Conjunto Logit Binario" (véase la sección 3.1.2 para ver detalles de esta heurística). Tras modificar el código y rehacer la simulación, se encontró el siguiente resultado, que muestra que el método CE permite modelar el conjunto de consideración real:

Tabla 3.7 Resultados de simulación nueva usando la heurística "Conjunto Logit Binario".

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.002	0.034	0.069	80
CE	0.071	0.097	1.051	37
CC	0.034	0.046	1.114	60

Fuente: Elaboración Propia

Luego, al estudiar si es que el número total de alternativas afecta el resultado, se obtuvo lo siguiente:

Tabla 3.8 Resultados de análisis de número de alternativas usando la heurística “Conjunto Logit Binario”.

	Cobertura Empírica	
Número de alternativas	CE	CV
10	37	81
12	35	78
15	39	82
18	37	84
21	32	76
30	30	84

Fuente: Elaboración Propia

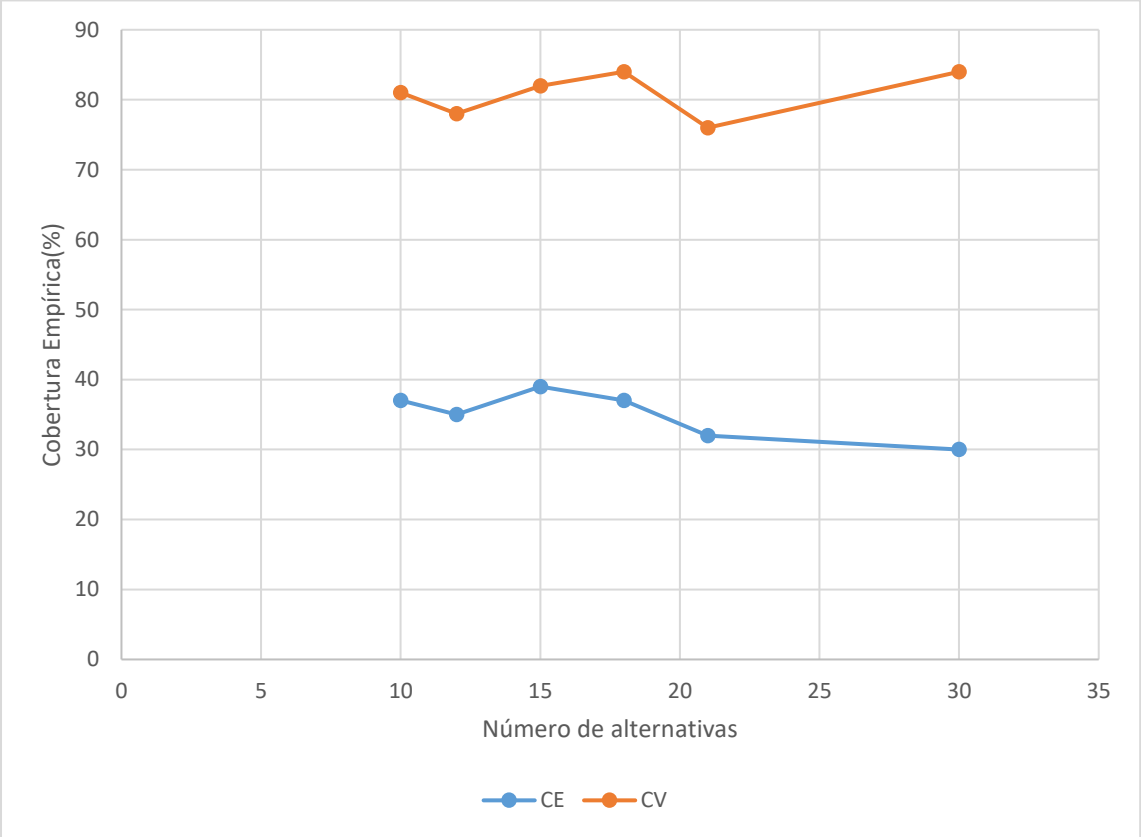

Figura 3.2 Comparación cobertura empírica entre métodos CE y CV.

Fuente: Elaboración Propia

Como el valor de la cobertura empírica se mantiene sobre 5, se puede decir que el número de alternativas totales no afecta el funcionamiento del método CE.

Para comprobar el resultado obtenido con la heurística CLB de forma más exhaustiva, se decidió repetir la simulación, esta vez asumiendo que la heurística usada por los usuarios es la "Eliminación por Aspectos" (véase la sección 3.1.2 para ver detalles).

Los cambios hechos al código original fueron los siguientes:

1) Se decidió cambiar el procedimiento por el cual se genera el conjunto de alternativas experimentadas. Originalmente, el código asignaba las alternativas de mayor utilidad al conjunto experimentado. Este método se cambió por uno que asigna las alternativas al azar en un número definido de días, de la siguiente manera:

- En primer lugar, se genera una matriz de utilidades (V_exp) usando los parámetros base señalados en el capítulo 3.1. V_exp [i,j] retorna la utilidad que le da la alternativa j a la persona i. Por lo tanto, esta matriz tiene un número de filas igual al número de personas totales, y un número de columnas igual al número de alternativas disponibles.

- Luego, se genera una matriz de ceros llamada $Avail\_exp$, con las mismas dimensiones que V_exp. Esta matriz representa las alternativas experimentadas de las personas, de la siguiente manera:

  $$Avail\_\exp[i,j] = \begin{cases} 1 & si\ la\ alternativa\ j\ fue\ experimentada\ por\ la\ persona\ i \\ 0 & si\ no \end{cases}$$

  Después, se le agrega a $Avail\_exp$ las alternativas reales que fueron escogidas por las personas (después de generar los datos, se define el conjunto de consideración real de cada persona usando EBA, y después se elige una de las alternativas consideradas).

- A partir de la matriz de utilidades se calcula la matriz de probabilidades P_exp (usando la fórmula de Logit Multinomial), para luego calcular la matriz de probabilidades acumuladas P_acum, en la cual $P\_acum_{kj} = \sum_{i=1}^{k} P\_exp_{ij}$. Cabe destacar que, si una persona eligió una alternativa con anterioridad, esa alternativa debería estar en su conjunto de consideración real. Por lo tanto, las alternativas no consideradas en el conjunto de consideración real de una persona van a tener una probabilidad de 0 para esta persona.

- Para cada usuario, se genera un número al azar entre 0 y 1. Si el número generado es menor que la probabilidad acumulada de la alternativa 1, se agrega la alternativa 1 a $Avail\_exp$. Si el número generado es menor que la probabilidad acumulada de la alternativa n, y mayor que la probabilidad acumulada de la alternativa n-1, se agrega la alternativa n a $Avail\_exp$.

- Esto se repite una cantidad definida de veces (Días de medición).

  De esta manera, este método de generación es equivalente a que los usuarios elijan una ruta cada día. El código modificado se puede ver en el Anexo D.

2) De forma similar, en el código original la elección real de cada persona se define usando la alternativa con mayor utilidad. Por esto, se decidió elegir la ruta escogida al azar, usando el método de probabilidades acumuladas.

Tras estos cambios, se obtuvo el siguiente resultado:

Tabla 3.9 Resultados de código con los cambios, siendo la heurística original "Eliminación por aspectos" y usando una generación por 30 días.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	0.002	0.027	0.076	64
CE	0.000	0.028	0.011	58
CC	0.135	0.135	12.671	0

Fuente: Elaboración Propia

En este caso, la hipótesis se cumple tanto para el test t como para la cobertura empírica.

Para comprobar que la hipótesis se cumple en este caso, se realizan simulaciones con las otras heurísticas (sección 3.1.2). El resumen de los resultados se puede ver en la figura 3.3 (en azul están los resultados con una cobertura empírica mayor a 5 y con un valor de test t menor a 1.984):

Métodos\Escenarios	I. Completo	II. Aleatorio	III. Eliminación por Aspectos	IV. CLB
i. CC				
ii. CE				

Figura 3.3 Resultados de las simulaciones con el código cambiado para todas las heurísticas.
Fuente: Elaboración Propia

Ahora, para estudiar la condición del tamaño del conjunto total de alternativas, se hacen tres simulaciones más: la primera usa un período de generación del CE de 10 días (en lugar de 30) y con 30 alternativas, la segunda también usa 10 días, pero usando 100 alternativas en lugar de 30, y la tercera limita el tamaño del CE a tres alternativas. Los resultados de estas simulaciones se pueden ver a continuación:

Tabla 3.10 Resultados de código con los cambios, siendo la heurística original "Eliminación por aspectos" y usando una generación por 10 días con 30 alternativas.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.003	0.019	0.170	67
CE	-0.032	0.060	0.641	42
CC	0.121	0.121	15.863	0

Fuente: Elaboración Propia

Tabla 3.11 Resultados de código con los cambios, siendo la heurística original "Eliminación por aspectos" y usando una generación por 10 días con 100 alternativas.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	0.001	0.021	0.039	65
CE	-0.032	0.133	0.249	27
CC	0.117	0.117	13.656	33

Fuente: Elaboración Propia

Tabla 3.12 Resultados de código con los cambios, siendo la heurística original "Eliminación por aspectos" y usando una generación por 10 días con 30 alternativas. El CE está limitado a 3 alternativas.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.004	0.025	0.141	63
CE	-0.016	0.046	0.365	44
CC	0.133	0.134	12.873	0

Fuente: Elaboración Propia

Se encontró un resultado anómalo en la tabla 3.17. En este caso, la cobertura empírica del método CC es mejor que la del método CE, a pesar de que el valor del test t rechaza la hipótesis nula de que los parámetros estimados son consistentes.

Para estudiar el valor anómalo, se realiza un análisis en el que se modifica el número de alternativas, y se ve el valor de la cobertura empírica para cada método, asumiendo un período de medición de 10 días para el método CE:

Tabla 3.13 Resultados de análisis de número de alternativas vs cobertura empírica para la heurística "Eliminación por aspectos".

	Cobertura Empírica		
Número de alternativas	**CV**	**CE**	**CC**
5	66	58	0
10	69	58	0
15	62	54	0
20	71	48	0
30	74	44	0
40	67	40	0
50	78	36	0
60	68	38	0
70	71	32	1
80	72	39	16
90	71	32	23
100	72	34	33
150	67	23	33
200	73	29	33

Fuente: Elaboración Propia

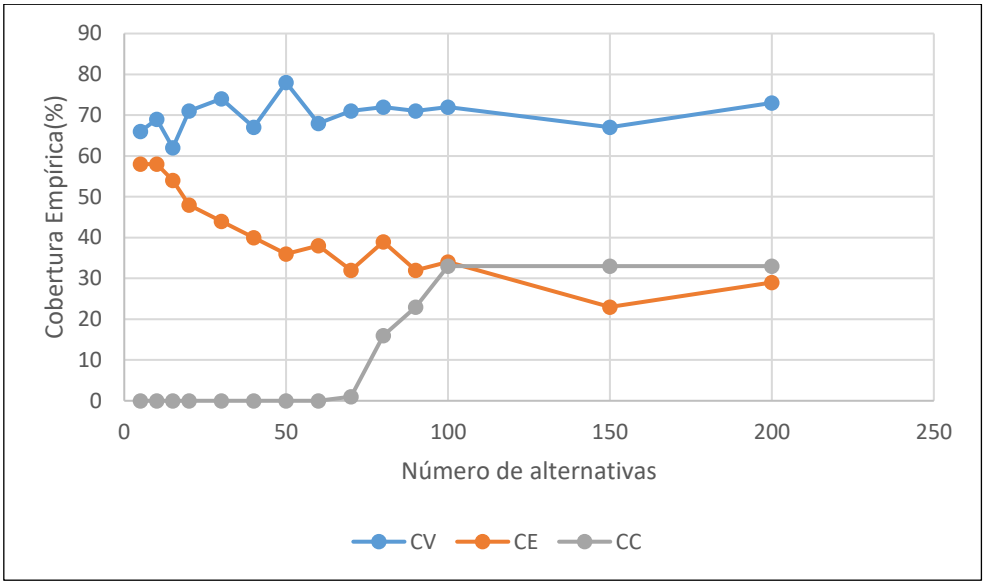

Figura 3.4 Comparación entre los valores de cobertura empírica de los tres métodos.
Fuente: Elaboración Propia

Se puede observar que la cobertura empírica del método CC empieza a aumentar a partir de las 70 alternativas, para luego mantenerse constante por sobre las 100 alternativas. Comparemos este resultado con el sesgo promedio y con el p-value del test t (es decir, la probabilidad de que en una medición se cumpla la hipótesis nula según el test t):

Tabla 3.14 Resultados de análisis de número de alternativas vs sesgo promedio para la heurística "Eliminación por aspectos".

	Sesgo Promedio		
Número de alternativas	CV	CE	CC
5	-0.001778	-0.006315	0.123545
10	0.000588	-0.007795	0.133990
15	0.002357	-0.009490	0.130149
20	0.002589	-0.010461	0.126821
30	-0.000655	-0.025189	0.121963
40	0.003965	-0.014996	0.122037
50	0.001699	-0.022688	0.120160
60	-0.004206	-0.046760	0.116697
70	-0.000667	-0.033623	0.117519
80	-0.001238	-0.040524	0.116845
90	0.000554	-0.030774	0.117387
100	-0.001989	-0.049143	0.116140
150	-0.000891	-0.038599	0.115282
200	-0.003212	-0.076916	0.114080

Fuente: Elaboración Propia

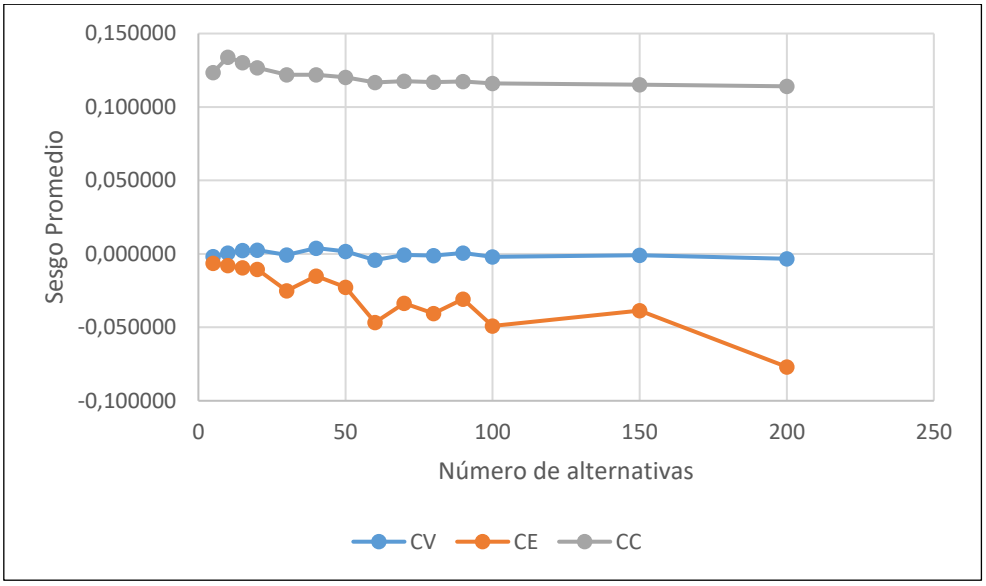

Figura 3.5 Comparación entre los valores de sesgo promedio de los tres métodos.
Fuente: Elaboración Propia

Tabla 3.15 Resultados de análisis de número de alternativas vs p-value del test t para la heurística "Eliminación por aspectos".

	p value del test t		
Número de alternativas	**CV**	**CE**	**CC**
5	95%	69%	0%
10	98%	82%	0%
15	92%	82%	0%
20	91%	82%	0%
30	98%	65%	0%
40	86%	82%	0%
50	93%	74%	0%
60	82%	53%	0%
70	97%	71%	0%
80	95%	69%	0%
90	98%	77%	0%
100	92%	69%	0%
150	96%	80%	0%
200	87%	74%	0%

Fuente: Elaboración Propia

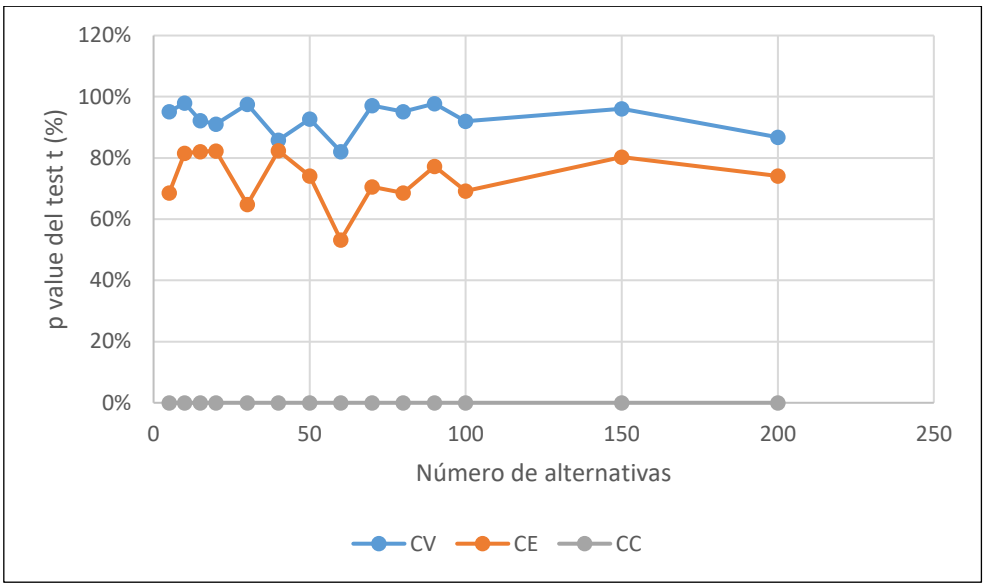

Figura 3.6 Comparación entre los valores del p-value del test t de los tres métodos.
Fuente: Elaboración Propia

La figura 3.5 muestra que el sesgo promedio de CC es mayor a 0.1, mientras que para los otros métodos el valor absoluto del sesgo se mantiene menor a 0.1. Mientras tanto, la figura 3.6 muestra que el método CC tiene un p-value del test t de 0% sin importar el número de alternativas totales.

Como se puede ver, estos valores confirman que el método CC no entrega parámetros consistentes con los reales. Con esto en mente, ¿por qué la cobertura empírica da un valor contradictorio a esto?

Una posible causa de este comportamiento es la definición del conjunto de consideración real. La fórmula usada en la heurística EBA para generar el conjunto de consideración real causa que el tamaño de este aumente de manera lineal al tamaño del conjunto de alternativas totales. Por ejemplo, con 100 alternativas totales, se puede ver que, en promedio, se consideran más o menos 50 alternativas.

Se recuerda que el concepto de conjuntos de consideración surgió porque a una persona le es difícil considerar un gran número de alternativas debido a limitaciones cognitivas. Por lo tanto, no es realista asumir que el conjunto de consideración va a crecer con el número de alternativas totales.

Debido a esto, se decidió cambiar el método de eliminación por aspectos. Ahora, para calcular el conjunto de consideración real, se realizó la siguiente iteración:

1) Una vez generados los datos de costo y tiempo, se decide eliminar las alternativas que no cumplan las siguientes restricciones:

$$c_{in} < c_{prom_n} + M * c_{desv_n} \qquad (3.7)$$

$$t_{in} < t_{prom_n} + M * t_{desv_n} \qquad (3.8)$$

Inicialmente, M va a tener un valor de 2.

2) Una vez eliminadas las alternativas que no cumplen con las restricciones, se usa el resto para crear la matriz $Avail$, que contiene el conjunto de consideración de cada usuario.

3) Si, en promedio, las personas tienen más de 10.5 alternativas consideradas, se vuelve al paso 1), y se le resta 0.02 a M (así la restricción descarta más alternativas). En caso contrario, se deja el conjunto de consideración tal como está.

Esto asegura que, sin importar la cantidad de alternativas, los usuarios van a considerar como máximo 10 (en promedio). El código que realiza este cambio se muestra en el Anexo E.

Con esto en mente, se hacen las simulaciones de nuevo con este método modificado, y se entregan los siguientes resultados:

Tabla 3.16 Resultados de análisis de número de alternativas vs cobertura empírica para la heurística "Eliminación por aspectos", con un conjunto de consideración promedio de 10 alternativas.

	Cobertura Empírica		
Número de alternativas	CV	CE	CC
10	81	56	82
15	74	51	0
20	73	50	0
50	59	41	0
75	63	44	0
100	61	44	0
150	61	46	0
200	61	40	0

Fuente: Elaboración Propia

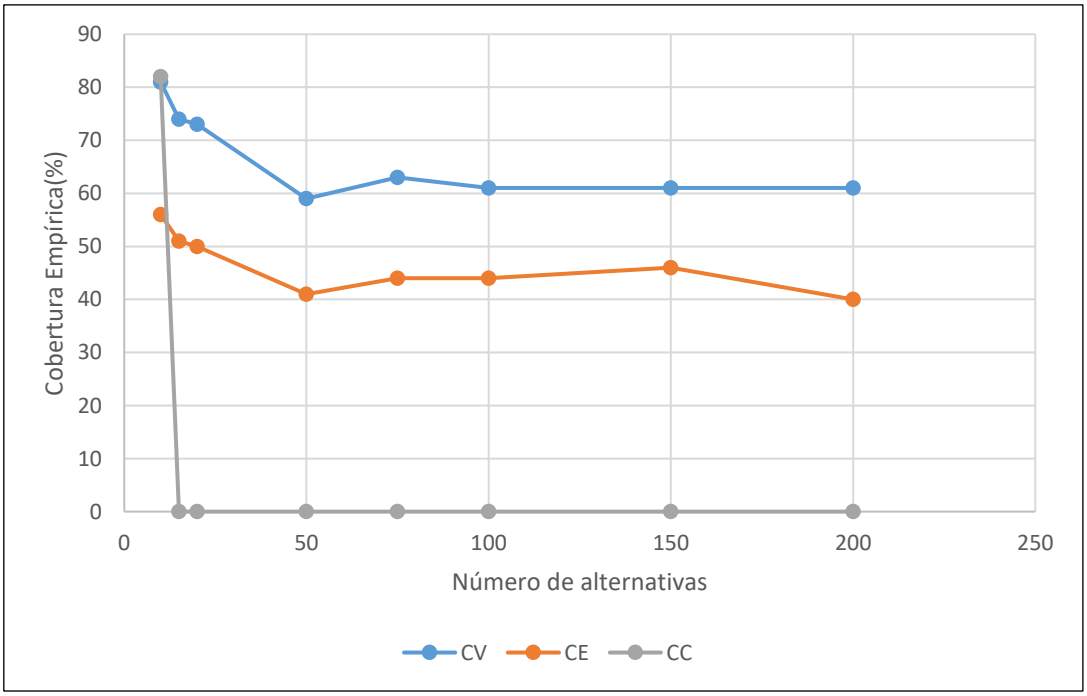

Figura 3.7 Comparación entre los valores de cobertura empírica de los tres métodos, con un conjunto de consideración promedio de 10 alternativas.

Fuente: Elaboración Propia

Como es de esperarse, como se consideran 10 alternativas, si hay 10 alternativas totales entonces el método CC va a entregar un resultado satisfactorio. Sin embargo, con 15 o más alternativas totales, CC obtiene una cobertura empírica de 0. Esto incluye la opción con 100 alternativas, lo que significa que la anomalía que estaba presente en las pruebas anteriores fue corregida. Lo importante que hay que destacar es que la cobertura empírica del método CE no baja de 40, por lo que sigue siendo válido.

3.3 Análisis de los resultados

A continuación, se presentarán los resultados del análisis de las condiciones para que sea posible usar los conjuntos de consideración de rutas basados en experiencias anteriores para simular con buena certeza los conjuntos de consideración reales usados por los usuarios.

Como se ve en la figura 3.7, la cobertura empírica del método basado en CE disminuye mientras aumenta la cantidad de alternativas totales. Sin embargo, aún con una diferencia considerable, en la que la cantidad alternativas es 20 veces mayor que la cantidad de días de medición, la cobertura empírica es lo suficientemente alta como para no rechazar la hipótesis. Esto se debe a lo descrito por Hauser (2014), que una persona no va a considerar un número elevado de alternativas, ya que se requiere de un esfuerzo mental muy alto para comparar un gran número de alternativas. Por lo tanto, el número total de alternativas no afecta el funcionamiento del método basado en CE para modelar el conjunto de consideración real.

Según lo visto en el Anexo 1 (en el cual se cambió el método generador de datos visto en la simulación inicial por uno más sencillo), también se puede decir que el método generador de datos no afecta la hipótesis.

Ahora, hay que señalar que existen varias extensiones para la simulación, para comprobar si existe alguna restricción que se debe cumplir para que el método basado en CE permita la modelación del conjunto de consideración de los usuarios. Por ejemplo, durante todo este proceso, se asume que las alternativas son independientes una de otra, ya que se asume que las elecciones son hechas usando Logit Multinomial. Para ver si la dependencia entre las alternativas impacta la hipótesis, se pueden hacer pruebas en la que los datos estén correlacionados de cierta manera. Esto se puede hacer asumiendo que el modelo de elección sea Probit o Logit Anidado.

Por otro lado, en este capítulo se asume que los valores de costo monetario y tiempo de viaje de cada ruta son fijos en el tiempo, lo que no siempre se cumple. Esta extensión está siendo trabajada en este momento por Villalobos, con resultados positivos hasta el momento (él ha encontrado variaciones del modelo CE que permiten encontrar el conjunto de consideración real aun cuando los valores de costo y tiempo de viaje son variables).

Capítulo 4: Estimación con datos reales

4.1 Introducción

En esta sección se busca conocer si es posible que el uso de un modelo basado en elecciones históricas permita una mejor estimación del conjunto de consideración real con datos reales, comprobando lo visto en las simulaciones de Monte Carlo. Debido a la limitada cantidad de bases de datos disponibles para elecciones de rutas, se decidió usar una base de datos de un rubro completamente diferente: compras de supermercado.

4.2 Descripción Base de Datos

La base de datos que se va a usar para este trabajo proviene de Instacart, una empresa estadounidense que ofrece compras en línea mediante un intermediario (el usuario selecciona una tienda asociada para después pedir un producto, luego una persona que trabaja para Instacart va a la tienda, compra el producto, y lo entrega a domicilio). Esta estructura le permite tener un registro de los productos que son comprados por cada usuario.

En mayo de 2017, Instacart publicó una base de datos que contiene más o menos 3 millones de diferentes órdenes hechas por más de 200 mil usuarios diferentes. Luego, se realizó un concurso público en el sitio web Kaggle para generar un código que pudiera predecir qué productos van a ser reordenados por un usuario.

Es necesario señalar que, debido a la naturaleza de la empresa, la base de datos no contiene los precios de los productos (debido a que diferentes tiendas pueden tener diferentes precios para el mismo producto).

Esta base se divide en los siguientes archivos:

1) **aisles.csv**

Este archivo contiene el número de identificación y el nombre de los 134 “pasillos” de la tienda.

2) **departments.csv**

Este archivo contiene el número de identificación y el nombre de las 21 “secciones”.

3) **order_products__prior.csv**

Este archivo especifica que productos fueron comprados en cada orden de compra, así como el orden en el que estos productos fueron añadidos. También se señala si un producto ha sido “recomprado” (es decir, si el producto se encuentra en una orden anterior del usuario). En particular, este archivo contiene órdenes que ya fueron hechas por el usuario.

4) **order_products__train.csv**

Similar al archivo anterior, con la diferencia de que estas órdenes son simuladas y se usan para entrenar (train) el modelo de predicción de compra.

5) **orders.csv**

Este archivo contiene las órdenes que realizó cada usuario, así como el tipo de orden realizada (prior, train, test). Además, contiene el día de la semana y la hora en la que se realizó la orden, así como los días que pasaron entre cada orden.

6) **products.csv**

Este archivo contiene el número de identificación y el nombre de los 49688 productos que se ofrecen, así como el pasillo y sección al que pertenece.

A continuación, se muestran tablas que contienen las primeras 10 filas de datos presentes en cada archivo:

Tabla 4.1 Primeras 10 filas de datos presentes en aisles.csv.

aisle_id	aisle
1	prepared soups salads
2	specialty cheeses
3	energy granola bars
4	instant foods
5	marinades meat preparation
6	other
7	packaged meat
8	bakery desserts
9	pasta sauce
10	kitchen supplies

Fuente: Base de Datos Instacart

Tabla 4.2 Primeras 10 filas de datos presentes en departments.csv.

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pasta
10	bulk

Fuente: Base de Datos Instacart

Tabla 4.3 Primeras 10 filas de datos presentes en order_products__prior.csv.

order_id	product_id	add_to_cart_order	reordered
2	33120	1	1
2	28985	2	1
2	9327	3	0
2	45918	4	1
2	30035	5	0
2	17794	6	1
2	40141	7	1
2	1819	8	1
2	43668	9	0
3	33754	1	1

Fuente: Base de Datos Instacart

Tabla 4.4 Primeras 10 filas de datos presentes en order_products__train.csv.

order_id	product_id	add_to_cart_order	reordered
1	49302	1	1
1	11109	2	1
1	10246	3	0
1	49683	4	0
1	43633	5	1
1	13176	6	0
1	47209	7	0
1	22035	8	1
36	39612	1	0
36	19660	2	1

Fuente: Base de Datos Instacart

Tabla 4.5 Primeras 10 filas de datos presentes en orders.csv.

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21
2254736	1	prior	4	4	7	29
431534	1	prior	5	4	15	28
3367565	1	prior	6	2	7	19
550135	1	prior	7	1	9	20
3108588	1	prior	8	1	14	14
2295261	1	prior	9	1	16	0
2550362	1	prior	10	4	8	30

Fuente: Base de Datos Instacart

Tabla 4.6 Primeras 10 filas de datos presentes en products.csv.

product_id	product_name	aisle_id	department_id
1	Chocolate Sandwich Cookies	61	19
2	All-Seasons Salt	104	13
3	Robust Golden Unsweetened Oolong Tea	94	7
4	Smart Ones Classic Favorites Mini Rigatoni With Vodka Cream Sauce	38	1
5	Green Chile Anytime Sauce	5	13
6	Dry Nose Oil	11	11
7	Pure Coconut Water With Orange	98	7
8	Cut Russet Potatoes Steam N' Mash	116	1
9	Light Strawberry Blueberry Yogurt	120	16
10	Sparkling Orange Juice & Prickly Pear Beverage	115	7

Fuente: Base de Datos Instacart

Sudalai Raj Kumar, un usuario de Kaggle, realizó un análisis preliminar de la base de datos, y se encontró lo siguiente:

- El producto más comprado por este servicio son plátanos:

Tabla 4.7 Frecuencia de productos comprados

	product_name	frequency_count
0	Banana	472565
1	Bag of Organic Bananas	379450
2	Organic Strawberries	264683
3	Organic Baby Spinach	241921
4	Organic Hass Avocado	213584
5	Organic Avocado	176815
6	Large Lemon	152657
7	Strawberries	142951
8	Limes	140627
9	Organic Whole Milk	137905
10	Organic Raspberries	137057
11	Organic Yellow Onion	113426
12	Organic Garlic	109778
13	Organic Zucchini	104823
14	Organic Blueberries	100060
15	Cucumber Kirby	97315
16	Organic Fuji Apple	89632
17	Organic Lemon	87746
18	Apple Honeycrisp Organic	85020
19	Organic Grape Tomatoes	84255

Fuente: "Simple Exploration Notebook – Instacart" de Sudalai Raj Kumar (2017)

- El pasillo con más productos comprados es el de fruta fresca:

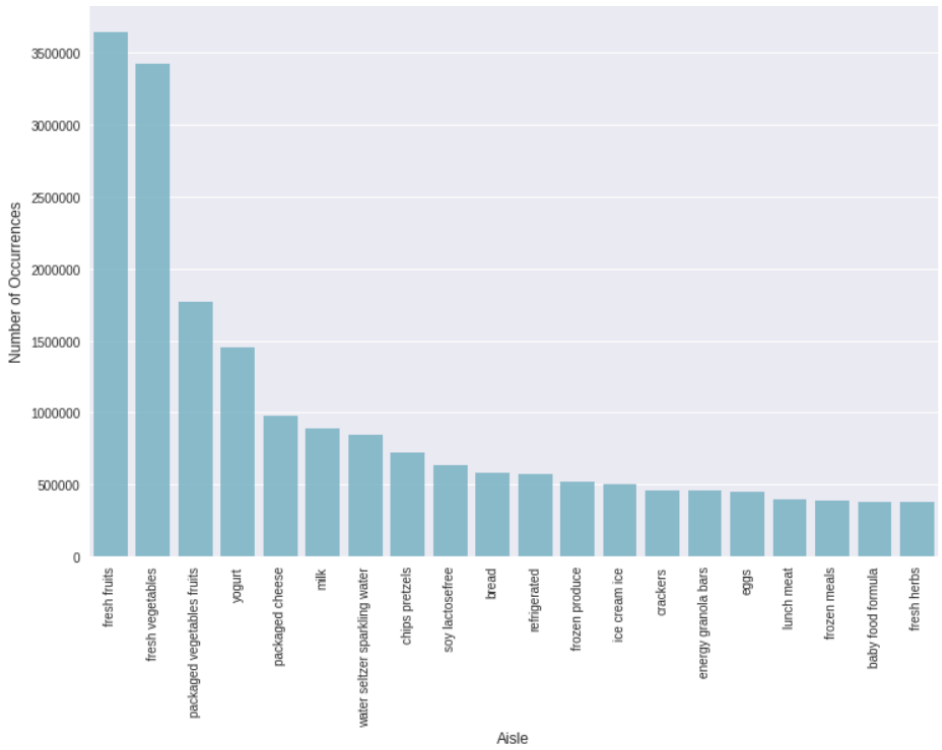

Figura 4.1 Gráfico de productos comprados por pasillo.
Fuente: "Simple Exploration Notebook – Instacart" de Sudalai Raj Kumar (2017)

- La sección con la mayor tasa de productos reordenados es la sección de huevos y lácteos:

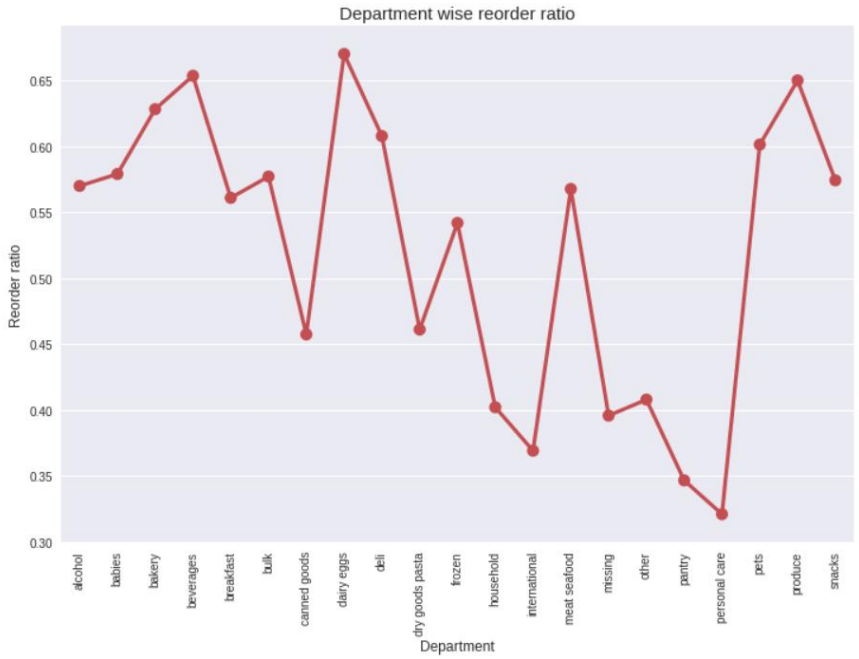

Figura 4.2 Gráfico de tasa de productos reordenados por sección.
Fuente: "Simple Exploration Notebook – Instacart" de Sudalai Raj Kumar (2017)

- Finalmente, los productos que se añaden primero al carro tienen una mayor tasa de reorden:

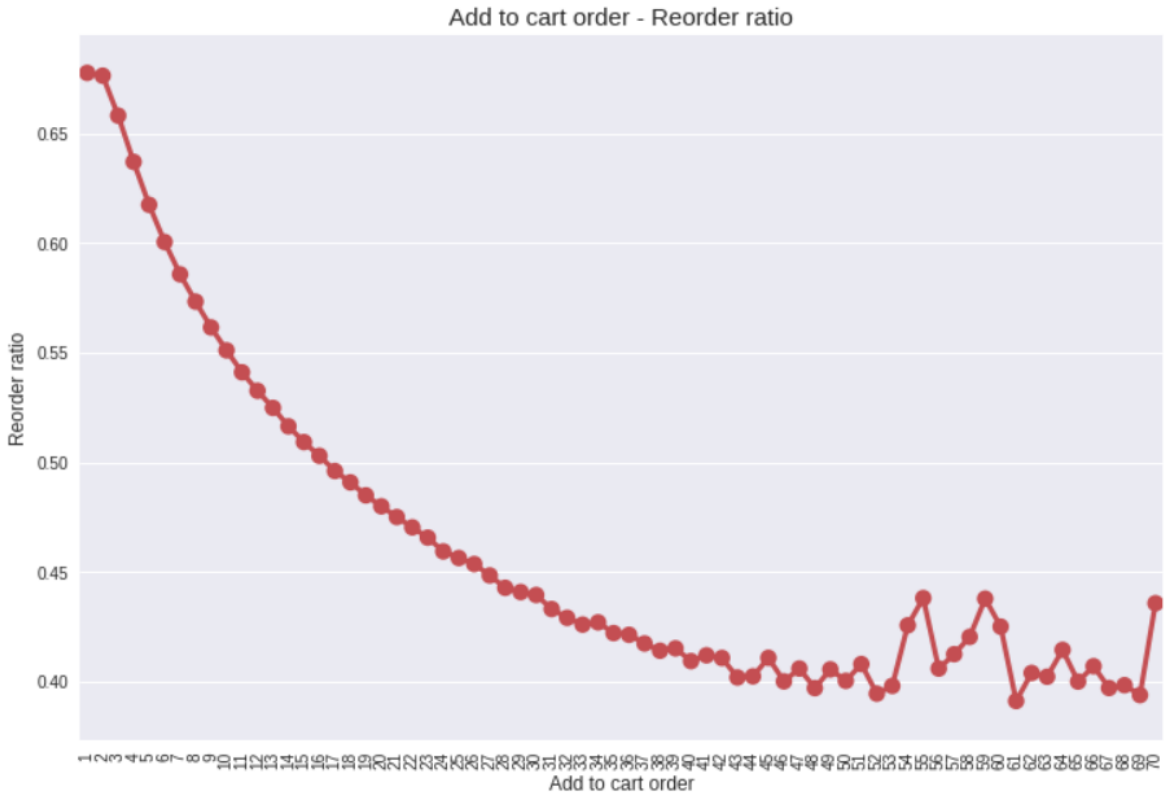

Figura 4.3 Gráfico de tasa de productos reordenados por orden de compra
Fuente: "Simple Exploration Notebook – Instacart" de Sudalai Raj Kumar (2017)

Este último gráfico tiene sentido lógico: Cuando una persona compra en el supermercado, esta tiende a comprar primero los productos que esta suele comprar, para luego ver los productos que no compra a menudo.

4.3 Esquema de Fórmulas de Probabilidades

Para efectos de este trabajo, debido a que no hay datos de precio, se va a asumir que la utilidad de un producto es de la siguiente forma:

$$U_{ni}(k) = \beta_{ni}(k) + \varepsilon_{ni}(k)$$

Se asume que el usuario n percibe una utilidad $U_{in}(k)$ si compra el producto k dentro de la orden i. $\beta_{ni}(k)$ es la constante específica de alternativa, y $\varepsilon_{ni}(k)$ es un número aleatorio que distribuye Gumbel con media 0. Para efectos de este estudio, $\varepsilon_{ni}(k) = 0$. Esto se hace para evitar la obtención de resultados erróneos debido a la aleatoriedad de $\varepsilon_{ni}(k)$.

Tal como se ve en la tabla 4.1.3, en cada orden se eligen varios productos y se muestra el orden de elección. Con esto en mente, se tiene la siguiente función de probabilidades para el primer producto electo:

$$P_{ni1}(k_1) = \frac{\exp\left(U_{ni}(k_1)\right)}{\sum_{k\prime \in C_p} \exp\left(U_{ni}(k')\right)}$$

La ecuación asume que C_p es el conjunto de todos los productos considerados por el usuario y que $k1$ es el primer producto elegido por el usuario. Asumiendo que no se escoge el mismo producto dos veces, se tiene que la probabilidad de escoger el producto k2 después del producto k1 es:

$$P_{ni2}(k_2|k_1) = \frac{\exp\left(U_{ni}(k_2)\right)}{\sum_{k' \in C_p - \{k_1\}} \exp\left(U_{ni}(k')\right)}$$

Usando probabilidades condicionales, se tiene que la probabilidad de escoger el producto $k2$ en segundo lugar es:

$$P_{ni2}(k_2) = \sum_{k_1 \in C_p} P_{ni2}(k_2|k_1) * P_{ni1}(k_1)$$

Así, para el producto k_f, el último producto electo en una orden, las probabilidades van a ser las siguientes:

$$P_{nif}\left(k_f \middle| k_1, k_2, \dots, k_{f-1}\right) = \frac{\exp\left(U_{ni}\left(k_f\right)\right)}{\sum_{k' \in C_p - \{k_1, k_2, \dots, k_{f-1}\}} \exp\left(U_{ni}(k')\right)}$$

$$P_{nif}\left(k_f\right) = \sum_{(k_1, k_2, \dots, k_{f-1}) \in C_p} P_{nif}\left(k_f \middle| k_1, k_2, \dots, k_{f-1}\right) * P_{ni}\left(k_1, k_2, \dots, k_{f-1}\right)$$

$$P_{nif}\left(k_f\right) = \sum_{(k_1, k_2, \dots, k_{f-1}) \in C_p} P_{nif}\left(k_f \middle| k_1, k_2, \dots, k_{f-1}\right) * P_{ni1}(k_1) * P_{ni2}(k_2) * \dots * P_{ni(f-1)}\left(k_{f-1}\right)$$

Con las probabilidades calculadas, se puede ver la función log- verosimilitud (likelihood) de la orden i:

$$L_{ni} = \sum_{j=1}^{f} \ln\left(P_{nij}\left(k_j\right)\right) * y_{nij}$$

$$y_{nij} = 1 \; si \; la \; persona \; n \; escogió, en \; la \; orden \; i, el \; producto \; k_j \; en \; el \; lugar \; j, 0 \; si \; no$$

Ahora, otra alternativa es ver las secciones que se escogieron, ya que se evalúa una menor cantidad de alternativas (21 en comparación con casi 50000). Este caso es más simple ya que se puede escoger la misma sección más de una vez:

$$P_{nij}\left(k_j\right) = \frac{\exp\left(U_{ni}\left(k_j\right)\right)}{\sum_{k' \in C_p} \exp\left(U_{ni}(k')\right)} \qquad \forall j \in [1:f]$$

$$L_{ni} = \sum_{j=1}^{f} \ln\left(P_{nij}\left(k_j\right)\right) * y_{nij}$$

Lo que se busca hacer es recuperar los parámetros $\beta_{ni}(k)$ de cada alternativa. Esto se puede hacer usando el método de estimador de máxima verosimilitud:

$$\hat{\beta}_{ni}(k) = \beta_{ni}(k) \; que \; maximiza \; L_{ni}^{*}$$

Como el logaritmo natural es una función monótona, los estimadores que maximizan la verosimilitud también maximizan la función log-verosimilitud. De esta manera, el estimador cumple lo siguiente:

$$\frac{\partial L_{ni}}{\partial \hat{\beta}_{ni}(k_j)} = 0 \qquad \forall j \in [1:f]$$

$$H\left(\hat{\beta}_{ni}(k)\right) = \begin{bmatrix} \frac{\partial^2 L_{ni}}{\partial \hat{\beta}_{ni}(k_1)^2} & \cdots & \frac{\partial^2 L_{ni}}{\partial \hat{\beta}_{ni}(k_1)\partial \hat{\beta}_{ni}(k_f)} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L_{ni}}{\partial \hat{\beta}_{ni}(k_f)\partial \hat{\beta}_{ni}(k_1)} & \cdots & \frac{\partial^2 L_{ni}}{\partial \hat{\beta}_{ni}(k_f)^2} \end{bmatrix} \quad es\ semidefinido\ negativo$$

Que el hessiano $H\left(\hat{\beta}_{ni}(k)\right)$ sea semidefinido negativo quiere decir que todos los elementos de este son menores o iguales a 0.

Para simplificar la ejecución del algoritmo, se va a usar el segundo método.

4.4 Comparación de métodos en base de datos

Teniendo en cuanta lo visto en el Capítulo 3, se va a realizar una prueba en la que van a comparar los siguientes métodos generadores del conjunto de consideración supuesto:

1. Conjunto Completo (CC)

   Este método asume que todas las personas consideran todas las secciones disponibles.

2. Conjunto Experimentado (CE)

   Este método asume que los usuarios solo consideran las secciones que contienen productos que fueron escogidos en el pasado.

Los modelos se van a comparar usando la bondad de ajuste y la verosimilitud promedio fuera de muestra.

La bondad de ajuste de un modelo describe que tan bien se ajusta este modelo a los valores observados. En este caso, la bondad de ajuste que se calcula va a ser el rho cuadrado ajustado ($\bar{\rho}^2$), el cual se determina de la siguiente manera:

$$\bar{\rho}^2 = 1 - \frac{L(\hat{\beta}) - f}{L(0)}$$

En esta ecuación, $L(\hat{\beta})$ es la log-verosimilitud del modelo con los parámetros estimados, f es el número de grados de libertad del modelo, y $L(0)$ es la log-verosimilitud de un modelo en el que todos los parámetros β son iguales a 0 (es decir, en el que todas las alternativas son equiprobables).

$\bar{\rho}^2$ es un valor que va entre 0 y 1. Por lo general, este valor no se usa para ver qué tan bueno es el ajuste de un solo modelo, sino para compararlo con el ajuste de otros modelos. Esencialmente, el modelo que tenga mayor $\bar{\rho}^2$ se va a ajustar mejor a los valores observados.

La verosimilitud promedio fuera de muestra del modelo es un parámetro que se calcula de la siguiente forma:

Sean $\widehat{\beta_i}$ el parámetro estimado de la alternativa i, usando las $n-1$ órdenes y todos los usuarios. Como solo hay constantes de alternativa en este caso, la probabilidad de escoger cada alternativa para cada usuario k va a ser:

$$P_k(i) = \frac{Y_{ki} * e^{\widehat{\beta_i}}}{\sum_{j \in C}\left(Y_{kj} * e^{\widehat{\beta_j}}\right)}$$

En la ecuación, C es el conjunto completo de alternativas, e Y_{ki} es el conjunto de consideración de la persona (de tal manera que Y_{ki} es 1 si la persona considera la alternativa i). Esto es diferente dependiendo del modelo. El modelo CE solo considera las alternativas que el usuario ha escogido en el pasado, mientras que en el modelo CC se consideran todas las alternativas ($Y_k = C$).

Con estas probabilidades, se puede calcular el siguiente valor:

$$\overline{P_k} = \frac{\sum_{i \in C}\left(Choice\_us_{ki} * P_k(i)\right)}{\sum_{i \in C} Choice\_us_{ki}}$$

En esta ecuación, $Choice\_us_k$ representa la orden n del usuario k, de tal manera que $Choice\_us_{ki}$ va ser 1 si la persona k elige la alternativa i en su última orden, y 0 si no. Finalmente, se puede obtener la verosimilitud promedio fuera de muestra del modelo promediando estos valores:

$$\bar{P} = \frac{1}{K} * \sum_{k=1}^{K} \overline{P_k}$$

Siendo K la cantidad total de usuarios medidos.

Así, la verosimilitud promedio fuera de muestra del modelo $\bar{P}$ representa que tan bien se ajusta el modelo a las elecciones n de los usuarios.

Con esto en mente, ahora se debe recopilar la información de la base de datos para obtener las últimas órdenes de los usuarios (para realizar la modelación) y el historial de elecciones de los usuarios (para diferenciar los dos modelos).

En primer lugar, se escogen K personas de la base de datos al azar. Luego, se extrae la información del archivo *orders.csv*, la cual contiene las órdenes que hace cada usuario. Después, se extraen los datos de los archivos *order_products__prior.csv* y *order_products__train.csv*, los cuales contienen los productos presentes en cada orden. Así, se seleccionan los productos que fueron elegidos por cada uno de los K usuarios.

Finalmente, se usa el archivo *products.csv* para identificar a qué secciones pertenece cada producto. Con esto se tienen las secciones que fueron escogidas por cada usuario, en cada una de sus órdenes. Con los datos obtenidos, se definen las siguientes variables:

$Choice\_ps$: Es una matriz de K filas y 21 columnas (se recuerda que hay 21 secciones en total) que representa las órdenes $(n-1)$ de K personas. Si el usuario 8 eligió la sección 5 en la penúltima orden, entonces $Choice\_n_{8,5}$ va a ser 1. En caso contrario, será 0.

$Choice\_us$: Es una matriz de K filas y 21 columnas que representa las órdenes n de K personas.

Y_u: Es una matriz de K filas y 21 columnas que representa los conjuntos de consideración de cada usuario. Por ejemplo, si el usuario 6 eligió la sección 7 en una o más órdenes, entonces $Y\_u_{6,7}$ será 1. En caso contrario, será 0.

Una vez que se tienen estos elementos, se puede generar una función que toma los parámetros β y retorna el valor de la log-verosimilitud para el caso CE:

$$L_{CE}(\beta_j) = \sum_{k=1}^{K} \sum_{j=1}^{21} \ln\big(P_{kj}(\beta_j)\big) * Choice\_ps_{kj}$$

$$Siendo \quad P_{ki}(\beta_i) = \frac{Y\_u_{ki} * \exp(\beta_i)}{\sum_{j=1}^{21}\big(Y\_u_{kj} * \exp\big(\beta_j\big)\big)} \qquad i \in [1,21]$$

En esta función:

β es un vector que representa las constantes específicas de alternativa. Debido a lo visto en la sección 4.2, la utilidad sistemática de cada alternativa (V) va a ser igual a la constante específica.

Luego, P es un vector que representa las probabilidades de escoger cada alternativa. Como se asume que hay repetición, la fórmula de probabilidades es la misma para todas las alternativas (Logit). A cada elemento se le multiplica su valor correspondiente de Y_u para descartar las secciones que no son consideradas por el usuario.

Si un usuario no considera una alternativa ($Y\_u_{ki} = 0$), entonces $P_{ki} = 0$. Debido a que $\ln(0) = -\infty$ (lo que causa problemas en el algoritmo), esta probabilidad se cambia por 1, ya que $\ln(1) = 0$, y así esa alternativa no es contada en el cálculo de la log-verosimilitud.

Finalmente, se calcula la log-verosimilitud aplicando la función vista en la sección anterior. En este caso, los logaritmos de las probabilidades se multiplican por el vector $Choice\_ps$.

El procedimiento es equivalente para el caso CC, con la diferencia de que en este caso no se multiplica Y_u en la fórmula de la matriz de probabilidades:

$$L_{CC}(\beta_j) = \sum_{k=1}^{K} \sum_{j=1}^{21} \ln\big(P_{kj}(\beta_j)\big) * Choice\_ps_{kj}$$

$$Siendo \quad P_{ki}(\beta_i) = \frac{\exp(\beta_i)}{\sum_{j=1}^{21}\exp\big(\beta_j\big)} \qquad i \in [1,21]$$

En este caso, se necesita $L(0)$ para el cálculo de la bondad de ajuste. En este caso, se calcula de la siguiente forma:

$$L(0) = \sum_{k=1}^{K} \sum_{j=1}^{n_{secc}} \ln\big(P(0)\big) = K * n_{secc} * \ln\big(P(0)\big)$$

$$Siendo \quad P(0) = \frac{\exp(0)}{\sum_{j=1}^{n_{secc}}\exp(0)} = \frac{\exp(0)}{n_{secc} * \exp(0)} = \frac{1}{n_{secc}}$$

Luego, se usa la función MaxLik para encontrar los estimadores de máxima verosimilitud y la log-verosimilitud máxima.

Sea $K = 10000$, es decir, que se ven los datos de diez mil personas al azar en la base de datos. En este caso, se obtiene que el valor de $L(0)$ es de -639349.4538. Realizando los procedimientos para el cálculo de la bondad de ajuste (con 20 grados de libertad, debido a que solo se fija un parámetro en 0) y para el cálculo de la verosimilitud promedio fuera de muestra, se obtienen los siguientes resultados:

Tabla 4.8 Comparación de resultados entre el método basado en alternativas experimentadas (CE) y el método basado en el conjunto completo de alternativas (CC).

Método	Log-verosimilitud	Bondad de Ajuste	Verosimilitud Promedio Fuera de Muestra
CC	-127869.9	0.7999688	0.09247863
CE	-113845.9	0.8219035	0.1324719

Fuente: Elaboración Propia

Los resultados mostrados señalan que el método CE obtiene una mayor bondad de ajuste y una mayor verosimilitud promedio fuera de muestra que el método CC. Por lo tanto, se puede considerar que el método CE es mejor que el método CC al modelar el conjunto de consideración real de las personas.

4.5 Prueba con métodos modificados para considerar elecciones de alternativas no consideradas

Hasta ahora, se ha asumido que todas las alternativas que el usuario escoge se encuentran dentro de su conjunto de consideración histórico. Sin embargo, un análisis de la base de datos revela que más o menos un 25% de personas (242 personas de 1000 escogidas al azar) escogen en su última orden una o más alternativas que no fueron elegidas en ninguna de las ordenes anteriores, es decir, que no se encuentran en el conjunto de consideración histórico. Este fenómeno puede ocurrir por varios motivos: una oferta, un cambio en los hábitos del usuario, o simple curiosidad, entre otros.

Esto puede causar que el método CE no represente adecuadamente el conjunto de consideración real de la población, ya que les da probabilidad nula a las alternativas no históricas. Por esto, se van a probar métodos alternativos para ver si el uso de estos obtiene un mejor resultado que el método CE.

En primer lugar, se va a asumir que, si la persona elige en su última orden al menos una alternativa que no está en sus elecciones históricas, entonces su conjunto de consideración va a ser el conjunto completo (CC). Por lo tanto, se propone el uso de la fórmula de Manski (ecuación 2.1) para la confección de los métodos:

$$P(i) = P(CE) * P(i|CE) + P(CC) * P(i|CC)$$

$$P(CC) = 1 - P(CE)$$

En esta ecuación, $P(i)$ es la probabilidad de que una persona elija la alternativa i, $P(CE)$ es la probabilidad de que la persona considere solo alternativas históricas, y $P(i|CE)$ es la posibilidad de escoger la alternativa i si es que el conjunto de consideración del usuario es en base a elecciones

históricas. $P(CC)$ y $P(i|CC)$ son análogos a lo anterior, asumiendo un conjunto de consideración que incluye a todas las alternativas.

A partir de la sección anterior, se tiene lo siguiente:

$$P(i|CE) = \frac{Y\_u_i * \exp(\beta_i)}{\sum_{j=1}^{n_{secc}} \left(Y\_u_j * \exp\left(\beta_j\right)\right)}$$

$$P(i|CC) = \frac{\exp(\beta_i)}{\sum_{j=1}^{n_{secc}} \left(\exp\left(\beta_j\right)\right)}$$

Por lo tanto, lo único que se debe averiguar es el valor de $P(CC)$. Para esto, se propone el uso de las órdenes históricas (preferencias observadas) para encontrar un valor que pueda representar la probabilidad buscada. Así, se deducen tres métodos posibles:

Mod1: Observar la penúltima orden de cada usuario para ver si elige una alternativa que no haya sido elegida en las órdenes anteriores. Si se elige una alternativa nueva en la penúltima orden, entonces se va a asumir que el usuario tiende a escoger elecciones nuevas, por lo que $P(CC) = 1$. De no ser así, $P(CC) = 0$.

Mod2: En lugar de observar la penúltima orden, se observan todas las órdenes anteriores de cada usuario, y se puede observar el número de órdenes en las que se eligió una alternativa que no había sido electa en las órdenes anteriores. Esto permite definir la probabilidad buscada de la siguiente manera:

$$P(CC) = \frac{\#ordenes\ con\ alternativas\ nuevas - 1}{\#\ ordenes - 1}$$

Así, las personas que tienden a escoger secciones nuevas tienen una mayor probabilidad de escoger una alternativa nueva en su última orden. La unidad que se resta en ambos términos de la fracción representa la primera orden de cada individuo, que, por razones obvias, siempre va a tener alternativas nuevas. Además, cabe destacar que el número de órdenes solo se cuenta hasta la penúltima orden.

Mod3: Se usa una combinación de los dos modelos anteriores, es decir, si se si se elige una alternativa nueva en la penúltima orden:

$$P(CC) = \frac{\#ordenes\ con\ alternativas\ nuevas - 1}{\#\ ordenes - 1}$$

De lo contrario, $P(CC) = 0$. Así, se eliminan posibles anomalías que pueden estar presentes en los modelos anteriores (por ejemplo, si una persona elige alternativas nuevas en sus primeras tres órdenes, pero no elige alternativas nuevas en el resto, entonces se asume que no va a elegir alternativas nuevas en su última orden).

Para comprobar la efectividad de los modelos, se encuentran los estimadores de máxima verosimilitud para cada uno de estos usando un universo de 1000 personas al azar, y después se calculan las bondades de ajuste y las verosimilitudes promedio fuera de muestra, de forma similar a la sección 4.4. Al hacer esto, se obtienen los siguientes resultados:

Tabla 4.9 Comparación de resultados entre el método basado en alternativas experimentadas (CE), el método basado en el conjunto completo de alternativas (CC), y los métodos modificados descritos anteriormente.

Método	Bondad de Ajuste	Verosimilitud Promedio Fuera de Muestra
CE	0.825	0.122
CC	0.804	0.092
Mod1	0.814	0.114
Mod2	0.814	0.11
Mod3	0.818	0.116

Fuente: Elaboración Propia

La tabla muestra que, a pesar de las modificaciones, el método CE sigue siendo el que mejor caracteriza el conjunto de consideración real de la población.

Capítulo 5: Conclusiones

5.1 Introducción

En este trabajo se presentan un par de aportes a la caracterización de conjuntos de consideración de los usuarios, mediante el uso de conjuntos con elecciones históricas. En primer lugar, se realiza una simulación de Monte Carlo para comprobar simulaciones anteriores que sugieren que el método basado en el conjunto de alternativas experimentadas representa de mejor manera el conjunto de consideración real de las personas. Luego, mediante el uso de datos de preferencias observadas obtenidos a partir de una base de datos de elecciones de supermercado, se puede ver si el conjunto de consideración histórico representa el conjunto de consideración real en una situación práctica.

La simulación de Monte Carlo logra demostrar que el método que genera el conjunto de consideración con alternativas experimentadas es robusto en condiciones bajo las cuales otros métodos, como el que está basado en el conjunto completo de alternativas, fallan. Esto se puede ver en las pruebas realizadas usando la heurística de eliminación por aspectos. Además, se comprobó que la cantidad total de alternativas no es una condición para que el método con alternativas históricas funcione de manera adecuada. Por otro lado, en el análisis de las preferencias observadas se logra establecer un modelo de utilidad para los datos disponibles, luego se calculan los parámetros que maximizan la log-verosimilitud, y luego se comparan los resultados obtenidos por el método que considera el conjunto completo con el método que considera las alternativas experimentadas. Este análisis muestra conclusiones similares a las simulaciones de Monte Carlo, aun considerando que a veces los individuos eligen alternativas que no están presentes en sus elecciones históricas.

5.2 Conclusiones Generales

Como primera conclusión, se puede decir que el análisis de las simulaciones de Monte Carlo, así como el análisis de la base de datos, indican que el uso de un método basado en alternativas experimentadas es factible para la caracterización de los conjuntos de consideración de los usuarios. Además, se puede decir que la razón por la que incrementar el número de alternativas disponibles no influye en el resultado es que un individuo no va a considerar un gran número de alternativas, por limitaciones cognitivas. Esto permite reducir la cantidad de alternativas que se van a modelar, por lo que también reduce el costo computacional del modelo.

Como segunda conclusión, se puede decir que, si bien la estimación con datos observados da un resultado práctico, se ve que el modelo con alternativas experimentadas no contempla que el usuario elija una alternativa que se encuentre fuera de sus elecciones históricas, cosa que sucede en un considerable número de casos. La razón por las que los modelos propuestos para resolver este problema resultaron ser peores que el modelo con alternativas experimentadas se puede deber a que se asumió que, si se consideraban alternativas nuevas, entonces el usuario consideraba todas las alternativas, lo que no siempre es cierto. Es posible que se pueda generar un modelo que dé un mejor ajuste, pero esto escapa a las limitaciones de este trabajo.

5.3 Recomendaciones Metodológicas

La principal recomendación metodológica que se puede hacer es el uso de bases de datos de elecciones de ruta con preferencias relevadas. Si bien la base de datos usada en este trabajo pudo ser usada debido a que usaba elecciones reales de la población, estas elecciones, dada la naturaleza de la compañía que publicó la base de datos, no tienen valores de costo para cada producto. Por eso, lo ideal sería obtener, de alguna manera, una base de datos que muestre el costo y el tiempo de cada ruta (como los datos de la tarjeta Bip!, por ejemplo), y comprobar si el modelo de alternativas históricas funciona de mejor manera que otros métodos usados para la caracterización del conjunto de consideración de la población.

5.4 Extensiones

En las simulaciones de Monte Carlo, es posible extender la investigación para que se vea que es lo que sucede si se asume que existe dependencia entre las alternativas (esto puede ser posible debido a que las rutas pueden compartir tramos). Como se mencionó en la sección 3.5, esto puede realizarse si se asume que el modelo de utilidad de las alternativas es Probit, o Logit Anidado.

En el análisis de la base de datos se realizaron varias simplificaciones para resolver este trabajo, por lo que existen varias extensiones posibles en este sector. En primer lugar, de ser posible, se pueden realizar pruebas con todos los productos en lugar de restringir las elecciones a las 21 secciones. Esto no se hizo en este trabajo debido a falta de capacidad computacional. Es importante recalcar que, de hacerse el trabajo de esta manera, los productos son electos sin probabilidad de ser reelegidos en la misma orden, lo que cambia la fórmula de probabilidades (lo que se ve en la sección 4.2).

Como se mencionó anteriormente, también sería conveniente obtener datos de preferencias declaradas en elección de ruta.

Finalmente, se sugiere a cualquiera que quiera seguir esta investigación que, en lo posible, use un computador server con alguna distribución de Linux instalada, ya que la versión de R disponible para Linux posee mayores recursos (incluyendo opciones para aprovechar más de un núcleo de la CPU) que la versión disponible para Windows.

Bibliografía

Bekhor, S., Ben-Akiva, M. E., y Ramming, M. S. (2006). Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research, 144(1),* 235-247.

Ben-Akiva, M., y Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing, 12(1),* 9-24.

Bierlaire, M., Hurtubia, R., y Flötteröd, G. (2010). Analysis of Implicit Choice Set Generation Using a Constrained Multinomial Logit Model. *Transportation Research Record: Journal of the Transportation Research Board, 2175,* 92-97.

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research, 67(8),* 1688-1699.

Instacart (2017). "The Instacart Online Grocery Shopping Dataset 2017", Descargado de https://www.instacart.com/datasets/grocery-shopping-2017 el 8 de Agosto de 2019.

Kumar, S., R. (2017, 2 de Junio). Simple Exploration Notebook – Instacart [Entrada de blog]. Descargado de https://www.kaggle.com/sudalairajkumar/simple-exploration-notebook-instacart/notebook

Manski, C. F. (1977). The structure of random utility models. *Theory and decision, 8(3),* 229-254.

Martinez, F., Aguila, F., y Hurtubia, R. (2009). The Constrained Multinomial Logit Model: A Semi-Compensatory Choice Model. *Transportation Research Part B, 43(3),* 365-377.

McFadden, D. L. (1984). Econometric analysis of qualitative response models. *Handbook of Econometrics, 2,* 1395-1457.

R Development Core Team. (2008). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de https://www.r-project.org/ (ISBN 3-900051-07-0)

Villalobos Zaid, G., N,. (2018). Caracterización del conjunto de consideración en elección de ruta. Disponible en http://repositorio.uchile.cl/handle/2250/159298.

Villalobos, G., N., y Guevara, A. (2019). *Accounting for the Consideration-Set in Discrete Choice Model Estimation* [Presentación de PowerPoint]. University of Central Florida.

Villalobos, G., N., y Guevara, A. (2019). *Caracterización del conjunto de consideración en elección de ruta* [Presentación de PowerPoint]. 19° Congreso de Ingeniería de Transporte, Santiago 2019.

Anexo A: Estudio de las condiciones usando la heurística "Conjunto Logit Binario" con simulaciones de Monte Carlo

En la sección 3.1, se detalló la simulación inicial que sugiere que el método basado en experiencias anteriores permite la modelación del conjunto de consideración usado por los usurarios en elecciones de ruta. Para comprobar que este resultado se mantiene bajo diferentes condiciones, se decidió hacer modificaciones al código, primero usando la heurística de "Conjunto Logit Binario", y luego usando la heurística de "Eliminación por Aspectos", la cual es más realista.

Se decidió ubicar el análisis de la heurística CLB en este anexo debido a que el procedimiento de análisis realizado en este es similar al análisis realizado con la otra heurística. Así, se evita la redundancia entre los dos análisis.

Los dos cambios de código que se probaron en este análisis fueron los siguientes:

1) Modificación método de generación de datos

En relación con lo anterior, se cambia el proceso de generación de datos para comprobar si el resultado cambia usando un proceso diferente. Para simplificar el código, se asume el siguiente costo promedio para todas las alternativas:

$$c_{prom} = \frac{(c_{maximo} + c_{minimo})}{2} = 20 \qquad (A.1)$$

El valor de la desviación estándar y la distribución usada son idénticas al código original.

2) Corregir método de construcción de conjunto de alternativas experimentadas.

Tras analizar el método que se usó en el código original para generar el conjunto de alternativas experimentadas para cada usuario, se encontró que se realizaba el siguiente proceso:

En primer lugar, se genera una matriz de utilidades determinísticas (V_exp) usando los parámetros base señalados en el capítulo 3.1. V_exp [i,j] retorna la utilidad que le da la alternativa j a la persona i. Por lo tanto, esta matriz tiene un número de filas igual al número de personas totales, y un número de columnas igual al número de alternativas disponibles.

Luego, se genera una matriz de ceros llamada $Avail\_exp$, con las mismas dimensiones que V_exp. Esta matriz representa las alternativas experimentadas de las personas, de la siguiente manera:

$$Avail\_\text{exp}\,[i,j] = \begin{cases} 1 & si\ la\ alternativa\ j\ fue\ experimentada\ por\ la\ persona\ i \\ 0 & si\ no \end{cases}$$

Después, se le agrega a $Avail\_exp$ las alternativas que fueron escogidas por las personas mediante MNL.

Una vez hecho esto, se realiza un ciclo "while" para agregar más alternativas experimentadas a cada persona. En primer lugar, se genera una matriz $epsilon\_exp$ de errores con distribución Gumbel (con las mismas dimensiones que V_exp) para calcular las utilidades totales de cada alternativa ($U\_exp = V\_exp + epsilon\_exp$), luego se le da una utilidad muy negativa a aquellas alternativas que no estén en el conjunto de consideración de la persona (que se simuló anteriormente mediante heurística). Finalmente, se le agrega la alternativa con mayor utilidad a $Avail\_exp$.

El código R que realiza el proceso descrito se muestra en el Anexo B.

El problema con este método es que es posible que el algoritmo agregue la misma alternativa una y otra vez al conjunto, si su utilidad determinística es lo suficientemente alta. Esto sucedió con el uso de los parámetros sin tiempo, lo que llevo a que el código quedara atascado en este proceso, y no obtuviera resultados.

Para resolver este problema, se decidió no seleccionar las alternativas experimentadas usando la mayor utilidad, sino simulando las probabilidades de elección, asumiendo un modelo Logit.

Así, se modificó el procedimiento de la siguiente manera:

Primero, se realiza el mismo procedimiento anterior hasta agregar la alternativa escogida por cada persona a $Avail\_exp$. Luego, en lugar de hacer un ciclo, solo se calculan las matrices de errores y utilidad total una sola vez.

A partir de la matriz de utilidades totales se calcula la matriz de probabilidades P_exp (usando la fórmula de Logit Multinomial) y luego, para cada usuario, se genera un vector de probabilidad acumulada ($probs\_acum_j = \sum_{i=1}^{j} probs_j$). Luego, se hace el siguiente ciclo para cada usuario:

1) Se genera un número uniforme entre 0 y max_p (la mayor probabilidad acumulada, la cual es 1 en un principio).

2a) Si el número generado es menor que la probabilidad acumulada de la alternativa 1, se agrega la alternativa 1 a $Avail\_exp$. Luego, se cambia la probabilidad de escoger la alternativa 1 a 0.

2b) Si el número generado es menor que la probabilidad acumulada de la alternativa n, y mayor que la probabilidad acumulada de la alternativa n-1, se agrega la alternativa n a $Avail\_exp$. Luego, se cambia la probabilidad de escoger la alternativa n a 0.

3) Luego, se recalcula el vector de probabilidad acumulada.

4) Si se escogieron n_alt_obs alternativas, se termina el ciclo y se pasa al siguiente usuario. De caso contrario, se actualiza max_p y se vuelve a 1.

El código de R que ejecuta este procedimiento se puede ver en el Anexo C.

El resto del código es similar al original, y se entregan los mismos resultados (sesgo promedio, RMSE, test-t y cobertura empírica). Por simplicidad, solo se van a mostrar los resultados para los siguientes métodos generadores del conjunto supuesto: Conjunto Verdadero (CV), Conjunto experimentado (CE) y Conjunto Completo (CC).

Tabla A.1 Resultados de código nuevo con parámetros de costo y tiempo.

Método	Sesgo Promedio	RMSE	test_t	Cobertura Empírica
CV	-0.002313875	0.03358014	0.06907023	80
CE	0.070505534	0.09730016	1.05146796	37
CC	0.034433693	0.04628017	1.11356354	60

Fuente: Elaboración Propia

Esto señala que la simulación de Monte Carlo funciona aún con diferentes métodos de generación, y con el método corregido de generación del conjunto de elecciones históricas. Se nota que los valores de test-t y de cobertura empírica son peores que con el código original, pero aún son valores relativamente aceptables, debido a que el test-t es menor que 1.984 y la cobertura empírica es mayor que 5.

Ahora, hay que ver si existe una condición bajo la cual este resultado no se cumple. Por lo tanto, se van a modificar varias condiciones iniciales del proceso de generación de datos, para luego correr la simulación de nuevo para ver si el valor de la cobertura empírica cambia o no.

1) Número de alternativas totales

En el experimento original, el número total de alternativas es 10. Por lo tanto, se va a incrementar el número de alternativas sin variar el número de alternativas experimentadas (siguen siendo 3). Se encontró lo siguiente:

Tabla A.2 Resultados de análisis de número de alternativas.

	Cobertura Empírica	
Número de alternativas	CE	CV
10	37	81
12	35	78
15	39	82
18	37	84
21	32	76
30	30	84

Fuente: Elaboración Propia

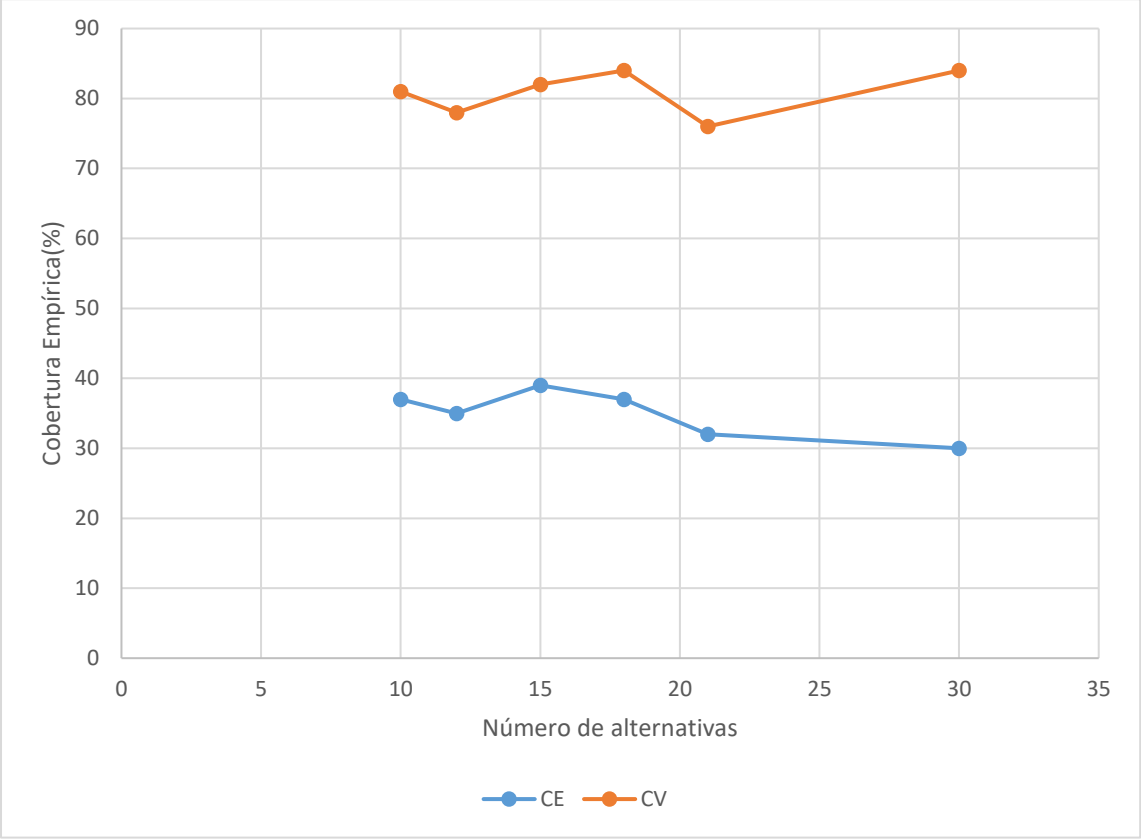

Figura A.1 Comparación cobertura empírica entre métodos CE y CV.

Fuente: Elaboración Propia

Se nota que la cobertura empírica del método CE tiende a descender (levemente) mientras aumenta el número de alternativas. Sin embargo, otro de los parámetros calculados, el test t, se mantiene debajo del límite superior:

Tabla A.3 Resultados de análisis de número de alternativas para el test t, en CE.

Número de alternativas	Test t
10	1.0895
12	1.0268
15	0.9338
18	1.0131
21	0.9279
30	1.0124

Fuente: Elaboración Propia

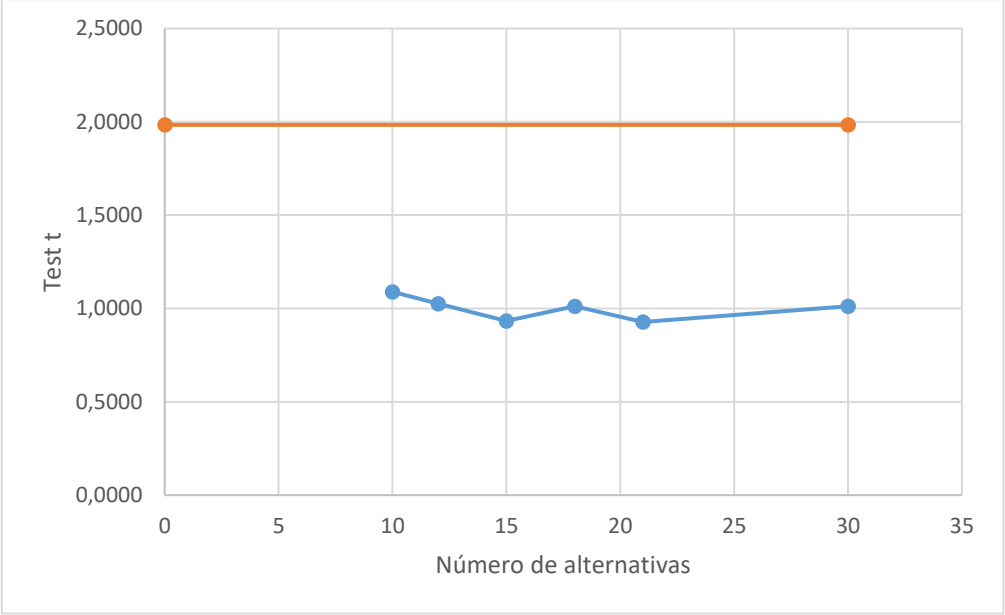

Figura A.2 Comparación test-t entre Método CE y valor límite.

Fuente: Elaboración Propia

En la figura 3.3, la línea anaranjada (1.984) representa el valor bajo el cual se dice que se recuperan los parámetros reales de forma exitosa.

2) Tamaño del conjunto de consideración

Para este análisis, se modificó el código de manera que, luego de generar los conjuntos de consideración, sólo se consideren las personas que consideraron un cierto número de alternativas.

Tabla A.4 Resultados de análisis de tamaño de conjunto de consideración en CE, para un universo de 10 alternativas (izquierda) y 15 alternativas (derecha).

Conjunto de consideración	Cobertura empírica
2	56
3	50
4	60
5	61
6	66
7	76
8	70
9	90

Conjunto de consideración	Cobertura empírica
3	91
4	80
5	56
6	62
7	58
8	61
9	71
10	69
11	69
12	78
13	86

Fuente: Elaboración Propia

Se debe notar que la simulación falla cuando el número de alternativas del conjunto de consideración es muy cercano al número total de alternativas. Esto puede deberse a que, en una de las iteraciones, no hay personas que consideren ese número de alternativas. Sin embargo, debido a que el método CE da una cobertura empírica sobre 50 si se asume que todas las personas consideran todas las alternativas (sección 3.1.4), se puede asumir que el tamaño del conjunto de consideración verdadero no afecta a la hipótesis.

3) Número de alternativas experimentadas

Aquí, simplemente se incrementa el número de alternativas experimentadas. Se recuerda que, para que el método funcione, se seleccionan a las personas que consideren un número igual o mayor de alternativas al número de alternativas experimentales.

Tabla A.5 Resultados de análisis de número de alternativas experimentadas en CE, para un universo de 10 alternativas.

Número de alternativas experimentales	Cobertura empírica
2	27
3	37
4	50
5	58
6	63
7	66
8	77
9	82

Fuente: Elaboración Propia

Claramente, se nota que la cobertura empírica aumenta con el número de alternativas experimentales.

Con esto, se puede decir que ninguna de estas tres condiciones afecta a la hipótesis. La primera condición (número total de alternativas) se estudia con mayor detalle en la sección 3.2.

Anexo B: Código de simulación original

```
V_exp <- mnl.v(betas)
colnames(V_exp) <- paste("V",seq(1,n_alt),sep = "")

Avail_exp <- matrix(rep(0,n_alt),ns_validas,n_alt,2)
colnames(Avail_exp) <- paste("Av",seq(1,n_alt),sep = "")

Avail_exp <- ifelse(Choice == 1, 1, Avail_exp)

while(min(rowSums(Avail_exp))<n_alt_obs){
  errores_exp <- replicate(ns_validas*n_alt, gumbel.gen(mu,beta))
  epsilon_exp <- matrix(data=errores_exp, nrow=ns_validas, ncol=n_alt)

  U_exp <- V_exp + epsilon_exp
  U_exp <- ifelse(Avail==1, U_exp, -100000)

  Avail_exp <- ifelse(rowSums(Avail_exp)<n_alt_obs & U_exp == apply(U_exp, 1, max), 1, Avail_exp)
}
```

Anexo C: Código con correcciones descritas en el Anexo A

```
Avail_exp <- matrix(rep(0,n_alt),np_validas,n_alt)
colnames(Avail_exp) <- paste("Av",seq(1,n_alt),sep = "")

Avail_exp <- ifelse(Choice == 1, 1, Avail_exp)

V_exp <- mnl.v(beta_asc, beta_cyt[1],beta_cyt[2])
colnames(V_exp) <- paste("V",seq(1,n_alt),sep = "")


errores_exp <- abs(rgumbel(n=np_validas*n_alt, loc=mu, scale=beta))
epsilon_exp <- matrix(data=errores, nrow=np_validas, ncol=n_alt)

U_exp <- V_exp + epsilon_exp

P_exp <- exp(U_exp)/rowSums(exp(U_exp))

while(min(rowSums(Avail_exp))<n_alt_obs){
  for (i in 1:np_validas){
    probs <- P_exp[i,]
    probs_acum <- rep(0,n_alt)
    for (j in 1:n_alt){
      if (j == 1){
        probs_acum[j] <- probs[j]
      }else{
        probs_acum[j] <- probs[j]+probs[j-1]
      }
    }
    alts_con<-sum(Avail_exp[i,])
    max_p <-max(probs_acum)

    while (alts_con<n_alt_obs){
      prob_u <- runif(1,0,max_p)  #se genera un numero entre 0 y max_p
      for (j in 1:n_alt){
        if (prob_u<probs_acum[j] & alts_con<n_alt_obs){  #se ve si prob_u es menor que la prob acumulada
          if (j == 1 ){
            Avail_exp[i,j]<-1
            probs[j]<-0
          }else if (prob_u>=probs_acum[j-1]){
              Avail_exp[i,j]<-1
              probs[j]<-0
          }
        }
        alts_con<-sum(Avail_exp[i,])
      }
      for (j in 1:n_alt){
        if (j == 1){
          probs_acum[j] <- probs[j]
        }else{
          probs_acum[j] <- probs[j]+probs[j-1]
        }
      }
      alts_con<-sum(Avail_exp[i,])
      max_p <- max(probs_acum)
      }
    }
}
```

Anexo D: Código con la corrección vista en la sección 3.2 (se determina conjunto de alternativas históricas con Días de medición)

```
Avail_exp <- matrix(rep(0,n_alt),np_validas,n_alt)
colnames(Avail_exp) <- paste("Av",seq(1,n_alt),sep = "")        # Nombra las columnas de Avail_exp

Avail_exp <- ifelse(Choice == 1, 1, Avail_exp)                  # Se agrega, si es que no estaba, la eleccion

V_exp <- mnl.v(beta_asc, beta_cyt[1],beta_cyt[2])               # Calcula la utilidad sistematica de las alternativas
colnames(V_exp) <- paste("V",seq(1,n_alt),sep = "")             # Nombra las columnas de V_p

P_exp <- (Avail*exp(V_exp))/rowSums(Avail*exp(V_exp))           # Calucla probabilidad Logit y elimina las alternativas que el usuario no considera

P_acum_exp <- matrix(0,np_validas,n_alt)                        # Genera matriz de probabilidades acumuladas
for (i in 1:np_validas){
  for (j in 1:n_alt){
    if (j==1){
      P_acum_exp[i,j] = P_exp[i,j]
    }else{
      P_acum_exp[i,j] = P_exp[i,j]+P_acum_exp[i,j-1]
    }
  }
}
while_counter <- 0                                              # Contador
while(while_counter<Dias){                                      # Asegura que se registren x Dias
  while_counter <- while_counter+1
  print(while_counter)
  prob <- runif(np_validas)                                     # Se generan números uniformes entre 0 y 1
  for (i in 1:np_validas){
    for (j in 1:n_alt){
      if (prob[i]<P_acum_exp[i,j]){
        if (j == 1){
          Avail_exp[i,j]<-1
        }else{
          if (prob[i]>=P_acum_exp[i,j-1]){
            Avail_exp[i,j]<-1
          }
        }
      }
    }
  }
}
```

Anexo E: Código que limita el tamaño del conjunto de consideración real a 10 sin importar el número de alternativas totales.

```
max_size<-10.5                                                  # Máximo tamaño de conjunto de consideración buscado
real_size<-n_alt
mult<-2.0                                                       #Valor inciial del multiplicador de la desviación estándar
cost_lim <- rowMeans(cost_info,1) + mult*apply(cost_info,1,sd)
time_lim <- rowMeans(time_info,1) + mult*apply(time_info,1,sd)
Avail <- matrix(rep(1,n_alt),np,n_alt,2)                        # Genera la estructura de la matriz de disponibilidad
Avail <- ifelse( cost_info <= cost_lim, Avail, 0)               # Elimina alternativas de costo superior al limite
Avail <- ifelse( time_info <= time_lim, Avail, 0)               # Elimina alternativas de tiempo superior al limite
colnames(Avail) <- paste("Av",seq(1,n_alt),sep = "")            # Nombra las columnas de Avail
real_size<-mean(rowSums(Avail))                                 # Actualiza tamaño de conjunto de consideración

while (real_size>max_size){           #Si el conjunto de consideración real es muy grande, se itera hasta que sea más o menos 10.
  mult<-(mult-0.02)
  cost_lim <- rowMeans(cost_info,1) + mult*apply(cost_info,1,sd)
  time_lim <- rowMeans(time_info,1) + mult*apply(time_info,1,sd)
  Avail <- matrix(rep(1,n_alt),np,n_alt,2)                      # Genera la estructura de la matriz de disponibilidad
  Avail <- ifelse( cost_info <= cost_lim, Avail, 0)             # Elimina alternativas de costo superior al limite
  Avail <- ifelse( time_info <= time_lim, Avail, 0)             # Elimina alternativas de tiempo superior al limite
  colnames(Avail) <- paste("Av",seq(1,n_alt),sep = "")          # Nombra las columnas de Avail
  real_size<-mean(rowSums(Avail))                               # Actualiza tamaño de conjunto de consideración
}

print(real_size)
alts_cons <- rowSums(Avail)
min_alt   <- min(alts_cons)
```