Ana Coimbra • Fernando Massa
Leonardo Moreno • Elena Vernazza

Ramón Álvarez-Vaz • Silvia Rodríguez-Collazo
coordinadores

La investigación en estadística desde la mirada de cuatro jóvenes investigadores

CSIC
UNIVERSIDAD DE LA REPÚBLICA URUGUAY
bibliotecaplural

La investigación en estadística desde la mirada de cuatro jóvenes investigadores

coordinadores de la publicación

Ramón Álvarez-Vaz

Silvia Rodríguez-Collazo

investigadores responsables

Ana Coimbra

Fernando Massa

Leonardo Moreno

Elena Vernazza

La publicación de este libro fue realizada con el apoyo de la Comisión Sectorial de Investigación Científica (CSIC) de la Universidad de la República.
Los libros publicados en la presente colección han sido evaluados por académicos de reconocida trayectoria, en las temáticas respectivas.
La Subcomisión de Apoyo a Publicaciones de la csic, integrada por Alejandra López, Luis Bértola, Carlos Demasi, Fernando Miranda y Andrés Mazzini ha sido la encargada de recomendar los evaluadores para la convocatoria 2015.

Ediciones Universitarias, Unidad de Comunicación de la Universidad de la República (UCUR)
18 de Julio 1824 (Facultad de Derecho, subsuelo Eduardo Acevedo)
Montevideo, CP 11200, Uruguay
Tels.: (+598) 2408 5714 - (+598) 2408 2906
Telefax: (+598) 2409 7720
Correo electrónico: infoed@edic.edu.uy
www.universidad.edu.uy/bibliotecas

isbn: 978-9974-0-1550-0

Tabla de contenidos

Presentación de la colección Biblioteca Plural

La Universidad de la República (Udelar) es una institución compleja, que ha tenido un gran crecimiento y cambios profundos en las últimas décadas. En su seno no hay asuntos aislados ni independientes: su rico entramado obliga a verla como un todo en equilibrio.

La necesidad de cambios que se reclaman y nos reclamamos permanentemente no puede negar ni puede prescindir de los muchos aspectos positivos que por su historia, su accionar y sus resultados, la Udelar tiene a nivel nacional, regional e internacional. Esos logros son de orden institucional, ético, compromiso social, académico y es, justamente, a partir de ellos y de la inteligencia y voluntad de los universitarios que se debe impulsar la transformación.

La Udelar es hoy una institución de gran tamaño (presupuesto anual de más de cuatrocientos millones de dólares, cien mil estudiantes, cerca de diez mil puestos docentes, cerca de cinco mil egresados por año) y en extremo heterogénea. No es posible adjudicar debilidades y fortalezas a sus servicios académicos por igual.

En las últimas décadas se han dado cambios muy importantes: nuevas facultades y carreras, multiplicación de los posgrados y formaciones terciarias, un desarrollo impetuoso fuera del área metropolitana, un desarrollo importante de la investigación y de los vínculos de la extensión con la enseñanza, proyectos muy variados y exitosos con diversos organismos públicos, participación activa en las formas existentes de coordinación con el resto del sistema educativo. Es natural que en una institución tan grande y compleja se generen visiones contrapuestas y sea vista por muchos como una estructura que es renuente a los cambios y que, por tanto, cambia muy poco.

Por ello es necesario:

a. Generar condiciones para incrementar la confianza en la seriedad y las virtudes de la institución, en particular mediante el firme apoyo a la creación de conocimiento avanzado y la enseñanza de calidad y la plena autonomía de los poderes políticos.

b. Tomar en cuenta las necesidades sociales y productivas al concebir las formaciones terciarias y superiores y buscar para ellas soluciones superadoras que reconozcan que la Udelar no es ni debe ser la única institución a cargo de ellas.

c. Buscar nuevas formas de participación democrática, del irrestricto ejercicio de la crítica y la autocrítica y del libre funcionamiento gremial.

El anterior rector, Rodrigo Arocena, en la presentación de esta colección, incluyó las siguientes palabras que comparto enteramente y que complementan adecuadamente esta presentación de la colección Biblioteca Plural de la Comisión Sectorial de Investigación Científica (CSIC), en la que se publican trabajos de muy diversa índole y finalidades:

"La Universidad de la República promueve la investigación en el conjunto de las tecnologías, las ciencias, las humanidades y las artes. Contribuye, así, a la creación de cultura; esta se manifiesta en la vocación por conocer, hacer y expresarse de maneras nuevas y variadas, cultivando a la vez la originalidad, la tenacidad y el respeto por la diversidad; ello caracteriza a la investigación - a la mejor investigación - que es, pues, una de la grandes manifestaciones de la creatividad humana.

Investigación de creciente calidad en todos los campos, ligada a la expansión de la cultura, la mejora de la enseñanza y el uso socialmente útil del conocimiento: todo ello exige pluralismo. Bien escogido está el título de la colección a la que este libro hace su aporte."

Roberto Markarian
Rector de la Universidad de la República
Mayo, 2015

Introducción

En 1998 comienza a funcionar la Licenciatura de Estadística en el marco de la Facultad de Ciencias Económicas y Administración de la Universidad de la República a partir del entusiasmo y esfuerzo que el profesor Jorge Blanco le puso a ese proyecto. Con una importante contribución de los docentes de esta Facultad, y con el apoyo de la Facultad de Ciencias mediante la contribución de docentes para el dictado de algunos cursos de dicha Licenciatura. Este proyecto fue un sueño por el que el profesor Jorge Blanco trabajó incansablemente y nos involucró a los docentes que en ese entonces trabajábamos en el Instituto de Estadística.

Con el correr de los años, la Licenciatura dió sus frutos y comenzaron a egresar licenciados en Estadística, todos ellos deben realizar para culminar su grado lo que se titula "pasantía "que puede tomar diversas modalidades, desde la práctica laboral en una empresa o institución hasta la ejecución de un proyecto de investigación llevado adelante por el estudiante y coordinado por un docente orientador. En todos los casos se elabora un informe final de esta tarea.

Hoy la Licenciatura de Estadística se ha transformado; tiene un nuevo plan de estudios y la colaboración de otros servicios de la Universidad se han ampliado. Queremos dedicar esta recopilación a quien fue un impulsor fundamental de este proyecto el profesor Jorge Blanco, proyecto que ha dado una gran variedad de frutos, este es uno más.

El documento es una obra colectiva y contiene una recopilación de trabajos de jóvenes investigadores en el área de Estadística. Son cuatro trabajos que permiten conocer algunos temas de interés para los jóvenes estadísticos. Los autores son todos egresados de la Licenciatura de Estadística y cuatro de ellos actualmente son investigadores del Instituto de Estadística.

Los compiladores de este trabajo vemos esta publicación como una oportunidad para difundir el trabajo en el contexto de apoyos que la Comisión Sectorial de Investigación Científica abre cada año.

En él se selecciona un pequeño conjunto de trabajos que han venido realizando estos jóvenes, de modo de mostrar las áreas de interés en las que han incursionado así como

el resultado de la búsqueda, perfeccionamiento y aplicación de diferentes metodologías estadísticas.

El trabajo se divide en cuatro capítulos autocontenidos. En el primero se presenta el trabajo de Moreno titulado "Clustering robusto basado en modelos". En este documento el autor analiza y compara distintas técnicas de clustering basadas en modelos probabilísticos con el objetivo de encontrar el procedimiento más eficiente y estable entre el conjunto analizado, considerando diferentes tipologías de datos, en presencia de outliers. Se prueba este enfoque en la conformación de grupos para un conjunto de inmuebles de Montevideo y Canelones a partir del precio de ellos, los metros cuadrados construidos y del terreno en el que se aloja el inmueble.

En el siguiente capítulo, se presenta el trabajo de Coimbra y Antía titulado "Tratamiento de la no respuesta en encuestas de panel en el caso de poblaciones finitas", cuyo objetivo es abordar una forma de tratamiento de la no respuesta en un tipo específico de encuestas por muestreo, las encuestas de panel.

En el capítulo tres, en una línea próxima al trabajo de Moreno, se presenta el trabajo de Massa cuyo título es "Efecto de valores faltantes en estudios longitudinales en adultos mayores". Es ente trabajo se aborda la forma de solucionar el problema del abandono de los sujetos incluidos en estudios denominados Mini Mental State Examination (MMSE), se centra en el uso de modelos conjuntos donde el tiempo de sobrevida de los sujetos y los resultados del MMSE están relacionados.

Finalmente, en el último capítulo, se presenta el trabajo de Vernazza en el que se analizan las propiedades psicométricas de un instrumento de medición de la satisfacción estudiantil en nuestra Facultad de Ciencias Económicas y de Administración.

Profesor Agregado Ramón Álvarez-Vaz
Profesora Agregada Silvia Rodríguez-Collazo
Compiladores
Facultad de Ciencias Económicas y de Administración
Departamento de Métodos Cuantitativos
Instituto de Estadística
Julio 2015

Capítulo 1

Clustering robusto basado en modelos, por Leonardo Moreno

Resumen[1]

Una gran variedad de técnicas o procedimientos de clustering han surgido en los últimos años en distintas ramas de la investigación estadística.

Sin embargo, la mayoría se basa en conjeturas heurísticas que carecen del debido rigor científico, lo que conlleva a conclusiones "oscuras" y posiblemente ficticias. Brindar procedimientos consistentes, que sean eficientes en distintos paradigmas de datos –inclusive en presencia de ruido– es un problema de elevada complejidad.

Algunas principales líneas recientes de investigación, tales como (Krishnan y McLachlan, 1997), (Ritter y Gallegos, 2005) (Ritter y Gallegos, 2009), (Matrán et al., 2008), abordan el problema de clustering desde una óptica basada en modelos y en presencia de outliers, intentando brindar respuestas formales en este ámbito.

El objetivo del trabajo es introducir, analizar, comparar e implementar distintas técnicas robustas de clustering basadas en modelos probabilísticos. La finalidad es poder discernir qué procedimiento es más eficiente y estable frente a distintas tipologías de datos, pudiendo así obtener conclusiones más precisas sobre los grupos obtenidos.

El análisis se realizó en primera instancia sobre distintos patrones de datos simulados. Posteriormente se aplicaron las técnicas a un conjunto de inmuebles de una determinada inmobiliaria de Montevideo, con el objetivo de determinar grupos de propiedades según sus características y valor de venta. La metodología permite detectar valores atípicos, que pueden ser posibles ofertas del mercado inmobiliario.

Palabras Claves: Cluster,k-medias, mezcla de distribuciones, outliers, robustez, trimming.

1. Resumen del trabajo de pasantía realizado en conjunto con Rodrigo Gadea,con la tutoría del Dr. Marco Scavino, para la obtención del grado de la Licenciatura en Estadística.

1.1. Introducción

La clasificación, o categorización, es uno de los procesos en los que se sustenta el aprendizaje en los seres humanos. Esta habilidad es tomada como sinónimo de inteligencia por algunas teorías psicológicas y cognitivas, incluso midiéndola a través de problemas de clasificación (como lo son las pruebas de cociente intelectual).

La primera definición considerada científica de clasificación se le reconoce a Aristóteles (384 AC - 322 AC), quien en sus Tópicos propone los Cinco Predicables para describir la lógica de la clasificación. Una definición más reciente del problema es dada por John Stuat Mill (1806 - 1873), quien la definía como:

> La clasificación es la conjunción, actual o ideal, de aquellos que son similares, y la separación de aquellos que no lo son; el propósito de esta conjunción es primariamente:
>
> 1. el facilitar las operaciones en la mente concibiendo claramente y reteniendo en la memoria el carácter de los objetos en cuestión,
> 2. el desentrañar leyes de correlación de propiedades de unión y circunstancias, y
> 3. el habilitar el registro de los mismos de forma de que sean referenciados convenientemente.

La clasificación estadística, o matemática, es la realización de este proceso mediante herramientas estadístico-matemáticas.

Existen dos escenarios bien diferenciados dentro de la clasificación estadística, para los cuales han sido desarrollados métodos y teorías casi independientes: cuando se tiene un conjunto de objetos de referencia de donde inferir la clase o categoría, y cuando no se tiene referencia alguna sobre las posibles categorías. En el primer escenario se suele llamar "Análisis de Regresión" o "Análisis de Clasificación" según la naturaleza de la variable del análisis en cuestión, mientras que en el segundo se suele hablar de "Análisis de Grupos" o "Clustering". Desde el ámbito del reconocimiento de patrones y desde el de la computación, se les refiere a estos escenarios como técnicas de clasificación o aprendizaje supervisado o técnicas de clasificación o aprendizaje no supervisado, haciendo énfasis en tener o no referencias para realizar el mismo proceso.

Este trabajo se propone introducir, analizar, comparar e implementar distintas técnicas de Clustering.

La finalidad es poder discernir qué procedimiento es más eficiente y estable frente a distintas tipologías de datos, pudiendo así obtener conclusiones más precisas sobre los grupos obtenidos. En particular, se centrará en técnicas robustas basadas en modelos probabilísticos, evaluando cuando los algoritmos pueden seguir clasificando o reconociendo correctamente los grupos en las observaciones cuando estas se encuentran contaminadas, fenómeno cada vez más presente en los conjuntos de datos actuales.

El Análisis de Grupos ha tenido un gran desarrollo en los últimos años, probablemente inducido en gran medida por el desarrollo de herramientas computacionales más potentes.

Muchas veces se observa al Clustering como una colección de técnicas mayoritariamente heurísticas para particionar datos multivariados. Este percepción se apoya en el hecho de que la mayoría de las técnicas de Clustering no son explícitamente basadas en un modelo probabilístico. Esto podría "llevar al investigador inocente a creer que él o ella no hicieron ningún supuesto en absoluto, y por ende los resultados son objetivos" (Flury, 1997). Sin embargo, esa objetividad está lejos de la realidad en tanto la mayoría de las veces los resultados del Clustering están fuertemente afectados por el método elegido y su performance es muy dependiente del modelo probabilístico subyacente asumido. Por ejemplo, cuando se usa k-Medias, se debe tener en cuenta que el método está diseñado para construir grupos esféricos de aproximadamente igual tamaño, y por lo tanto, este método no es confiable cuando los grupos que se buscan se alejan fuertemente de este supuesto.

Entonces, para comprender los métodos de Clustering y decidir cuál de estos métodos se deberían aplicar a un caso particular, es de interés el determinar modelos apropiados y desarrollar métodos especialmente diseñados para esos modelos.

Otra problemática de interés actual es el impacto negativo y distorsionante que tienen los outliers en los procedimientos de Clustering. Sin embargo, sin especificar un modelo no es claro qué se entiende por una observación siguiendo un comportamiento "anómalo". Por esta razón, el presente trabajo toma como punto de partida al Clustering basados en modelos. Por ejemplo, no es claro cuando un conjunto de observaciones muy dispersas puede ser visto como un grupo o como meramente ruido de fondo a ser eliminado.

Adicionalmente, no es obvio si un pequeño grupo de outliers muy 'juntos' deberían ser considerados como un grupo propio en vez de un fenómeno de contaminación.

Por estas razones, se tratarán procedimientos de Clustering en un paradigma "model-based" bajo fenómenos de contaminación en la muestra, tanto desde el punto de vista de la consistencia como de la robustez.

En esta instancia, al comienzo de la investigación, caben las interrogantes de "¿Existe alguna razón para favorecer un algoritmo sobre otro?", "¿Existe algún procedimiento universal, que sea el 'mejor', frente a cualquier estructura de los datos?", "¿Existe algún resultado fundamental que se cumpla sin importar la inteligencia del diseñador, el número y distribución de los patrones, y la naturaleza de la tarea de clasificación?" Y, en caso afirmativo, ¿Por qué?.

Estas preguntas conciernen a las cimientos del reconocimiento de patrones estadístico.

Duda y Hart en su trabajo Pattern classification (Hart et al., 2001) concluye que independientemente del problema, sin realizar algún supuesto, ningún método de clasificación de patrones es inherentemente superior a algún otro, aún a la asignación aleatoria.

Los pilares de está afirmación son los teoremas de "No hay almuerzo gratis" y el teorema del "Patito feo" que hablan sobre la carencia de superioridad inherente de cualquier clasificador y que frente la ausencia de supuestos no existe una "mejor" representación de las características, y que aun la noción de similaridad entre patrones depende implícitamente en los supuestos, los cuales pueden o no ser correctos.

El teorema "No hay almuerzo gratis" justifica el escepticismo acerca de estudios que se proponen demostrar la superioridad de un algoritmo particular de aprendizaje o reconocimiento sobre el resto.

El teorema del "Patito feo"[2] fuerza a reconocer que aun la aparentemente simple noción de similaridad entre patrones está fundamentalmente basada en supuestos implícitos acerca del dominio del problema.

Como consecuencia, en el momento de comparar las diversas técnicas presentadas siempre se debe tener claramente explicitados los supuestos sobre la tipología de los datos y que las conclusiones extraídas no se podrán extender a otras formas de modelado de datos (de aquí la elección de la perspectiva basada en modelos en el trabajo). Como se

2. Nombrado por la famosa historia de Hans Christian Andersen, El patito feo. Se debe a que el teorema muestra que, si todas las cosas son iguales, un patito feo es tan similar a un cisne como dos cisnes son entre ellos.

intenta proporcionar herramientas de Clustering en diversas situaciones que se comporten de forma estable frente a distintos tipos de perturbaciones sobre el modelo de los datos, es necesario presentar variadas técnicas y enfoques que funcionen de forma más eficiente una con respecto a la otra en diferentes paradigmas.

1.2. Algunos procedimientos de Clustering

1.2.1. K-Medias

Uno de los procedimientos más populares para determinar clusters en un conjunto de datos es el método de k-Medias (k-Means). Si bien los precursores de este algoritmo fueron MacQueen en el año 1967 (McQueen, 1967) y Hartigan en el año 1978 (Hartigan, 1978), Pollard en sus trabajos de 1981 (Pollard, 1981) y 1982 (Pollard, 1982) prueba la consistencia fuerte del método y su distribución asintótica respectivamente.

Se comienza primero por describir el procedimiento a través de estos trabajos y luego para calcular el centro de cada clúster se implementará una métrica robusta de la forma $\frac{d}{1+d}$, siendo d la distancia euclideana, lo cuál convertirá al algoritmo en estable frente a la presencia de outliers. En lugar de tomar como centro de clúster la media –punto que minimiza la suma cuadrática de las distancias euclídeas– se toma el punto que minimiza la suma de las distancias robustas $\frac{d}{1+d}$. Además la misma prueba de Pollard servirá para probar su consistencia.

El procedimiento de clustering por k-Medias prescribe un criterio para particionar un conjunto de puntos en k grupos: para dividir los puntos $x_1, x_2, \ldots, x_n$ en $\mathbb{R}^s$ se debe elegir los centros de los clusters, $a_1, a_2, \ldots, a_k$, de forma de minimizar la función

$$W_n = \frac{1}{n} \sum_{l=1}^{n} \min_{1 \leq j \leq k} \|x_l - a_j\|^2,$$

donde $\|\cdot\|$ denota la norma euclídea, para entonces asignar cada x_l a su centro de clúster más cercano. De esta forma, cada centro a_l adquiere un subconjunto C_l de puntos x como su clúster asociado. La media de los puntos en C_l debe ser igual a a_l, de otra forma, W_n podría ser disminuida mediante el remplazo de a_l por la media del clúster, en primera instancia, y entonces reasignar algunos de los x's a sus nuevos centros. Este criterio es, entonces, equivalente al de minimizar la suma de los cuadrados entre los clusters.

Se asume que $\{x_1, x_2, \ldots, x_n\}$ es una muestra de observaciones independientes de alguna distribución P. Pollard brinda condiciones que aseguran la convergencia casi segura

de los centros de los clusters cuando el tamaño de la muestra aumenta, generalizando uno de los resultados de (Hartigan, 1978), quien lo probó para dos clúster.

(McQueen, 1967) obtuvo resultados de consistencia débil para el algoritmo de k-Medias que distribuye puntos secuencialmente entre k clusters. Con este algoritmo, los centros no son escogidos para minimizar W_n; en su lugar, cada x_n es asignado al clúster con el centro más cercano, entonces ese centro es movido a la media del clúster modificado.

Debido a las dificultades que pueden surgir de las ambigüedades en el etiquetado de los puntos $x_1, x_2, \ldots, x_n$ y los centros $a_1, a_2, \ldots, a_k$, es ventajoso el considerar W_n como una función del conjunto de centros de los clusters y de la medida empírica P_n obtenida de la muestra por asignarle masa n^{-1} a cada uno de los $x_1, x_2, \ldots, x_n$. El problema es entonces, el de minimizar

$$W(A, P_n) := \int \min_{a \in A} \|x - a\|^2 P_n(dx)$$

sobre todas las posibles elecciones del conjunto A conteniendo k (o menos) puntos. Para cada A fijo, una ley fuerte de los grandes números (LFGN) muestra que:

$$W(A, P_n) \to W(A, P) := \int \min_{a \in A} \|x - a\|^2 P(dx), \quad \text{c.s.}$$

Puede esperarse que A_n, el conjunto de centros de clusters óptimo para la muestras de tamaño n, debería estar cerca de $\bar{A}$, el conjunto de centros que minimizan $W(\cdot, P)$, siempre que $\bar{A}$ este determinado únicamente. Por tanto, existe un etiquetado $a_{n1}, a_{n2}, \ldots, a_{nk}$ de puntos en A_n, y un etiquetado $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_k$ de puntos en $\bar{A}$, de forma que $a_{nl} \to \bar{a}_l$ c.s. Este enfoque también evita problemas con la posible coincidencia de dos de los centros de clusters.

En la práctica, encontrar un A en el cual $W(\cdot, P_n)$ alcanza su mínimo global involucra una cantidad prohibitiva de cálculos. Sin embargo, existen algoritmos eficientes para encontrar particiones localmente óptimas de los puntos muestrales en k clusters.

El método de prueba de la convergencia de los centros está basado en la aplicación repetida de la LFGN; el argumento se aplica a casi todos los puntos muestrales ω.

La prueba se aplicará a un criterio de clustering más general. Por ejemplo, los centros de los clusters pueden ser escogidos para minimizar una cantidad basada en desviaciones absolutas,

$$\int min_{a\in A}\|x-a\|P_n(dx),$$

o aun un criterio con atractivo de robustez,

$$\int min_{a\in A}\|x-a\|\wedge 1 P_n(dx).$$

El teorema incluye tales posibilidades: una función monótona creciente $\phi(\|x-a\|)$ de las desviaciones $\|x-a\|$ puede ser usada en definir una suma de desviaciones entre clusters, lo cuál será utilizado en este trabajo.

Resultados Asintóticos

Sean $x_1, x_2, \ldots, x_n$ variables aleatorias independientes en $\mathbb{R}^s$ con distribución común P. Sea P_n la correspondiente medida empírica. La muestra $\{x_1, x_2, \ldots, x_n\}$ va a ser dividida en k clusters mediante minimizar una suma de desviaciones entre clusters, y se puede probar un resultado de consistencia sobre los centros de los clusters.

Para cada medida de probabilidad Q en $\mathbb{R}^s$ y cada subconjunto (finito) A de $\mathbb{R}^s$ se define

$$\Phi(A,Q) := \int min_{a\in A}\phi(\|x-a\|)Q(dx).$$

y

$$m_k(Q) := \text{ínf}\{\Phi(A,Q) : A \text{ contiene } k \text{ o menos puntos}\}.$$

Para un k dado, el conjunto $A_n = A_n(k)$ de centros de clusters muestrales óptimos será elegido para satisfacer $\Phi(A_n, P_n) = m_k(P_n)$; los centros de clusters poblacionales $\bar{A} = \bar{A}(k)$ satisfacen $\Phi(\bar{A}, P) = m_k(P)$. El objetivo es mostrar que $A_n \to \bar{A}$, c.s.

La convergencia de conjuntos debería ser tomada como la convergencia determinada por la métrica de Hausdorff $H(\cdot,\cdot)$, la cual está definida para subconjuntos compactos A, B de $\mathbb{R}^s$ por $H(A,B) < \delta$ si y solo si todo punto de A está entre una distancia (euclídea) δ de al menos un punto de B, y viceversa. Suponer que A contiene exactamente k puntos distintos, y que δ es elegido menor a la mitad de la distancia mínima entre puntos de A. Entonces si B es cualquier conjunto de k o menos puntos para el cual $H(A,B) < \delta$, él debe contener exactamente k puntos distintos, cada uno de ellos se encuentra a una distancia no mayor δ de un punto únicamente determinado en A.

La convergencia casi segura de A_n en el sentido de Hausdorff podría, entonces, ser traducida a una convergencia casi segura de los centros de los clusters bajo un adecuado etiquetamiento.

Para que los procedimientos aquí descritos tengan sentido, la función ϕ debe satisfacer algunas condiciones de regularidad. Se precisa tener una ϕ continua y no decreciente, con $\phi(0) = 0$. De forma de controlar el crecimiento de ϕ en las colas, asumir que existe alguna constante λ tal que $\phi(2r) \leq \lambda\phi(r)$ para todo $r > 0$. En tanto $\int \phi(\|x\|)P(dx)$ sea finito, esto asegura que $\Phi(A, P)$ es finito para cada A: para cada $a \in \mathbb{R}^s$,

$$\int \phi(\|x - a\|)P(dx) \leq \int \phi(\|x\| + \|a\|)P(dx) \leq$$

$$\leq \phi(2\|a\|) + \int_{\|x\| \geq \|a\|} \phi(2\|x\|)P(dx) \leq \phi(2\|a\|) + \lambda \int \phi(\|x\|)P(dx).$$

Estos supuestos sobre ϕ son necesarios en el desarrollo de la teoría.

Teorema 1 (Consistencia de los centros). Suponer que $\int \phi(\|x\|)P(dx) < \infty$ y que para $j = 1, 2, \ldots, k$ existe un único conjunto $\bar{A}(j)$ para el cual $\Phi(\bar{A}(j), P) = m_j(P)$. Entonces $A_n \to \bar{A}(k) c.s.$, y $\Phi(A_n, P_n) \to m_k(P)$ c.s.

Teorema 2 (Ley Fuerte de los Grandes Números). Sea $\mathcal{G}$ la familia de funciones P-integrables en $\mathbb{R}^2$ de la forma $g_A(x) := \text{mín}_{a \in A}\, \phi(\|x - a\|)$, donde A varía sobre todos los subconjuntos de $\mathcal{E}_k$ conteniendo k o menos puntos.

$$\sup_{g \in \mathcal{G}} \left| \int g\, dP_n - \int g\, dP \right| \to 0 \text{ c.s.} \tag{1.1}$$

Una condición suficiente para que ((1.1)) se cumpla es: para cada $\epsilon > 0$ existe una clase finita $\mathcal{G}_\epsilon$ de funciones tales que para cada $g \in \mathcal{G}$ existen funciones $\dot{g}$, $\bar{g} \in \mathcal{G}_\epsilon$ con $\dot{g} \leq g \leq \bar{g}$ y $\int (\bar{g} - \dot{g})\, dP < \epsilon$. (La prueba usa la LFGN aplicada a cada función en la clase numerable $\mathcal{G}_{1/2} \cup \mathcal{G}_{1/3} \cup \mathcal{G}_{1/4} \cup \cdots$, junto con la cota $\int (\bar{g} - g_0)\, dP + \text{máx}\{|\int \bar{g}\, dP_n - \int \bar{g}\, dP|, |\int \dot{g}\, dP_n - \int \dot{g}\, dP|\}$ para $|\int g\, dP_n - \int g\, dP|$.)

Notar que los argumentos no dependen realmente del espacio muestral subyacente, estando $\mathbb{R}^s$ equipado con su norma usual, cualquier espacio métrico para el cual todas las bolas cerradas son compactas serviría. Por ejemplo, si se toma la métrica $d'(x, y) = \frac{d(x,y)}{1+d(x,y)}$ siendo d la métrica euclideana, la convergencia se cumple.

Se realiza la descomposición de $W_n(a) = n^{-1} \sum_{i=1}^n \text{mín}_{1 \leq j \leq k} \|x_i - a_j\|^2$ en dos componentes que pueden ser expresadas en función de la medida empírica P_n y del proceso

empírico asociado $X_n(.) = n^{1/2}(P_n(.) - P(.))$. Para cualquier vector $a = [a_1, a_2, \ldots, a_k] \in \mathbb{R}^{kd}$ y para cualquier $x \in \mathbb{R}^d$ se define

$$\phi(x, a) = \min_{1 \leq j \leq k} \|x - a_j\|^2,$$

entonces

$$W_n(a) = P_n\phi(\cdot, a) = P\phi(\cdot, a) + n^{-1/2} X_n \phi(\cdot, a).$$

La componente $P\phi(\cdot, a)$ conocida popularmente como suma de cuadrados entre clúster, se denotará por $W(a)$.

Pollard, en su trabajo de 1982, demuestra propiedades asintóticas de normalidad para la sucesión de centros.

Una variante robusta de K-Medias

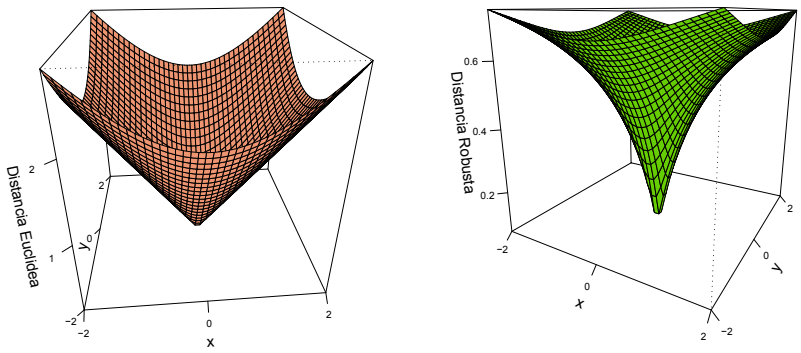


Figura 1.1: Funciones de distancias Euclídea (izquierda) y Robusta (derecha) al origen

Es sencillo probar que $\phi = \frac{\psi}{1+\psi}$ es una función monótona creciente si $\psi \geq 0$ y una distancia si ψ lo es. Tomando a ψ como la distancia euclídea, la distancia ϕ verifica las hipótesis necesarias para los trabajo de Pollard [81] y [82]. La elección de esta distancia acotada se debe a su sencillez y que crece a tasa decreciente con la distancia euclídea: observaciones muy alejadas de la restantes verán reducido su impacto en el cálculo de los centros de los clusters, mientras que mantiene el orden inducido por la distancia euclídea con respecto al centro.

Para ejemplificar el efecto de la contaminación en el algoritmo k-Medias original, se lo implementó independiente de la distancia a utilizar en la determinación del centro de cada cluster. Esta se efectuó con las distancias euclídea y robusta ψ, sobre simulaciones con

distintos tipos de contaminación: local y global. Debido a que no se tiene una expresión analítica para obtener el centro óptimo de cada clúster bajo una distancia cualquiera, se optó por realizar la optimización mediante métodos numéricos.

Estudio de Simulación

Se estudia la performance de las distintas variantes del algoritmo k-Medias y su estabilidad frente a la presencia de contaminación. Para ello se simularán distintos escenarios, en los cuales se considera ruido local entre los clusters y fuera de ellos, así como también ruido global.

En todos los casos se consideran 2 grupos conformados con 100 datos cada uno y 20 observaciones atípicas. Se comparará en estos escenarios tres técnicas de Clustering ya descritas:

- k-Medias con 3 grupos, donde el grupo más pequeño determina los outliers.
- k-Medias con 2 grupos, donde luego de conformar los grupos se determinan cuáles son los outliers tomando como criterio la distancia de Mahalanobis a el centro al que fue asociado, podando el 10 % de las observaciones.
- La variante robusta de k-Medias, luego de determinar los grupos se procede a podar igual que en k-Medias con 2 grupos.

A efectos de analizar la performance se realizan 150 repeticiones de cada técnica, y en cada una de estas se calcula el porcentaje de observaciones bien clasificadas. Se entiende que una observación es bien clasificado si es asociada al grupo del que fue simulado (Grupo A, Grupo B o Grupo R / Outlier).

Para estos 150 porcentajes se realiza los diagramas de caja correspondiente a cada técnica.

Se comienza por analizar el problema en los distintos tipos de escenarios y se derivan conclusiones de los mismos.

Primer Escenario: Ruido Global. A los efectos de analizar este primer escenario se simulan 220 datos en $\mathbb{R}^2$.

El primer grupo de 100 observaciones proviene de una distribución normal bivariante con vector de medias $(4, 0)$ y matriz de varianzas y covarianzas identidad,

mientras que el otro grupo también está conformado por 100 datos de una distribución normal bivariante con vector de medias $(0, 4)$ y matriz de varianzas y covarianzas identidad. EL papel de ruido global lo jugarán 20 datos que se simulan uniformemente en el cuadrado $[-10, 10]^2$.

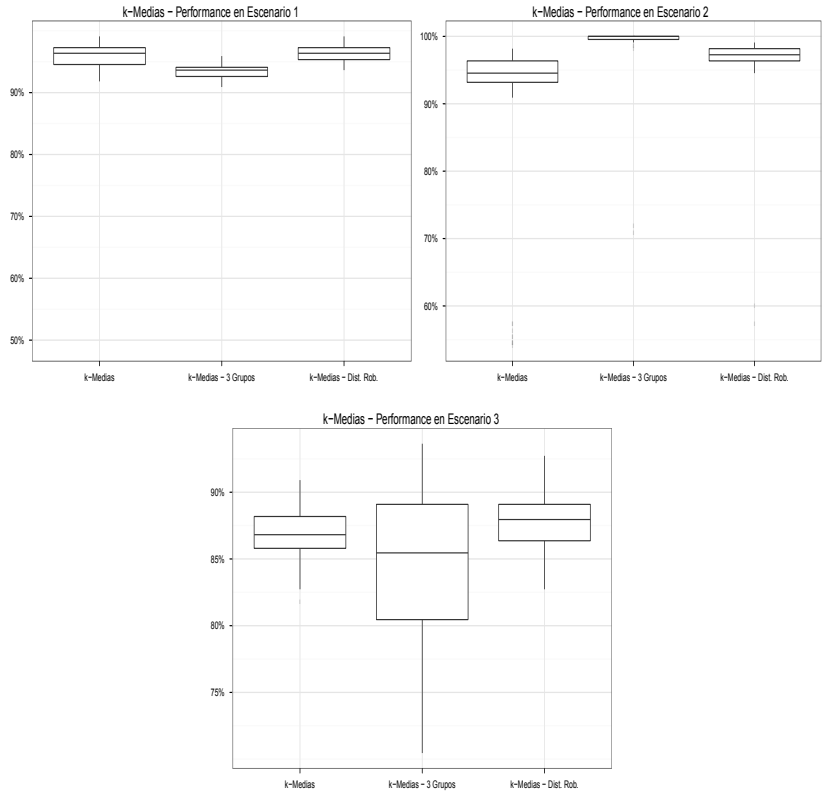


Figura 1.2: Diagramas de cajas del porcentaje de observaciones bien clasificadas por cada algoritmo.

Como se puede apreciar en la figura 1.2, los tres algoritmos tienen una alta eficiencia. Este ruido, al ser global y uniforme no produce sesgos considerables en la estimación de los centros de los cluster.

De todas formas, la variante robusta produce en general mejores resultados con una variabilidad similar.

Como en este escenario el grupo de los outliers no se presenta en un grupo claramente definido k-Medias con 3 grupos es claramente deficiente.

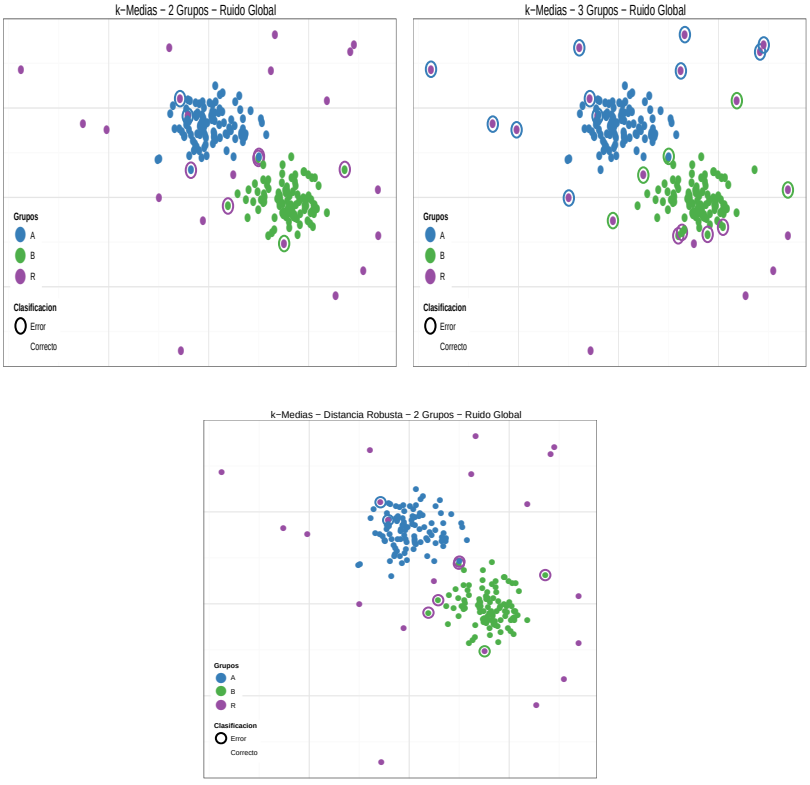


Figura 1.3: Clasificación de variantes de k-Medias en 2 grupos con ruido global

Segundo Escenario: Ruido Local Alejado. Para este segundo escenario se mantienen las simulaciones respecto a los grupos, 220 datos en $\mathbb{R}^2$, 100 de estos provenientes de una distribución normal bivariante con vector de medias $(4, 0)$ y matriz de varianzas y covarianzas identidad, 100 datos de una distribución normal bivariante con vector de medias $(0, 4)$ y matriz de varianzas y covarianzas identidad. Pero los 20 datos que jugarán al papel de ruido local se simulan uniformemente en el cuadrado $[-5, 0]^2$, ubicándose en el cuadrante inferior izquierdo del escenario.

Se puede observar que –como era de esperar en este caso– al estar el grupo de outliers claramente diferenciado de los grupos, k-Medias con 3 grupos es el algoritmo que mejor clasifica.

Sin embargo, la variante robusta de k-Medias se mantiene estable frente a estas tipologías, manteniendo en el 50 % central de las repeticiones un porcentaje de aciertos entre el 96 % y el 98 % (ver (1.2)).

La incidencia de este grupo local de outliers es alta sobre los centros de los clusters del algoritmo de k-Medias con 2 grupos, problema que se logra amortiguar con la variante robusta.

Tercer Escenario: Ruido Local entre Grupos En este tercer y último escenario a analizar se mantienen la cantidad de simulaciones respecto a los grupos, 220 datos en $\mathbb{R}^2$, 100 de estos provenientes de una distribución Normal bivariante con vector de medias $(4, 0)$ y matriz de varianzas y covarianzas identidad, 100 datos de una distribución Normal bivariante con vector de medias $(0, 4)$ y matriz de varianzas y covarianzas identidad. Pero el ruido local se simula uniformemente en el cuadrado $[0{,}5, 3{,}5]^2$, ubicándose entre los centros de ambas normales.

Se genera así un escenario donde no es nada trivial delimitar los grupos, así como tampoco es sencilla la identificaciones de los outliers.

Se esperaba que el algoritmo de k-Medias con tres grupo siguiera brindando la mejor clasificación. Sin embargo, esta tipología de outliers atrae los centros hacia la parte central del escenario, distorsionando de esta forma el algoritmo mencionado.

Observando la Figura (1.2) se puede ver como el porcentaje de aciertos disminuye considerablemente en los tres métodos, y la mayor eficiencia la presenta la variante robusta del k-Medias.

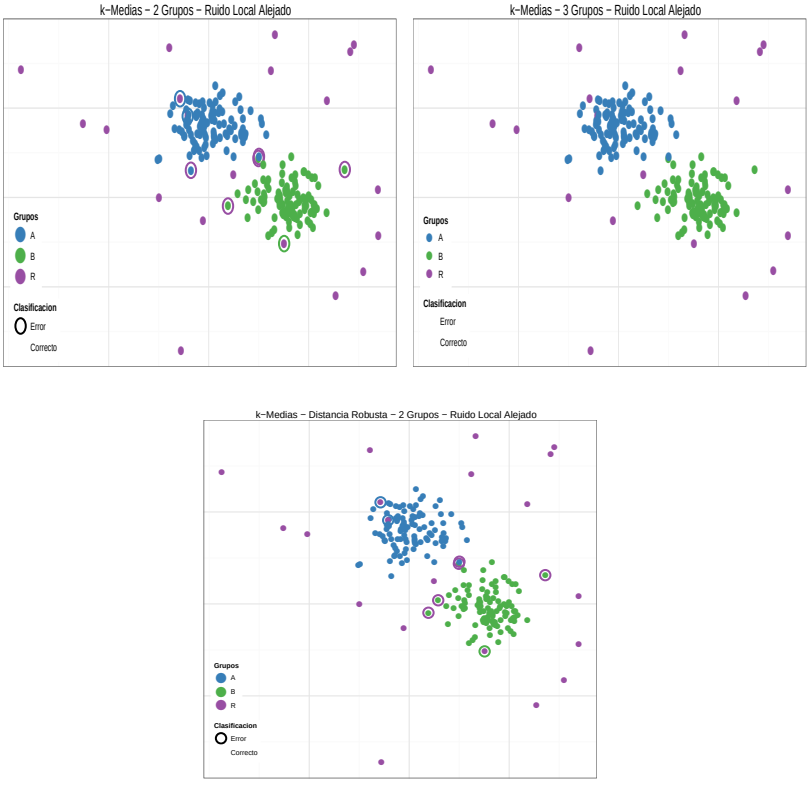


Figura 1.4: Clasificación de Algoritmo de k-Medias variante robusta en 2 grupos con ruido local alejado.

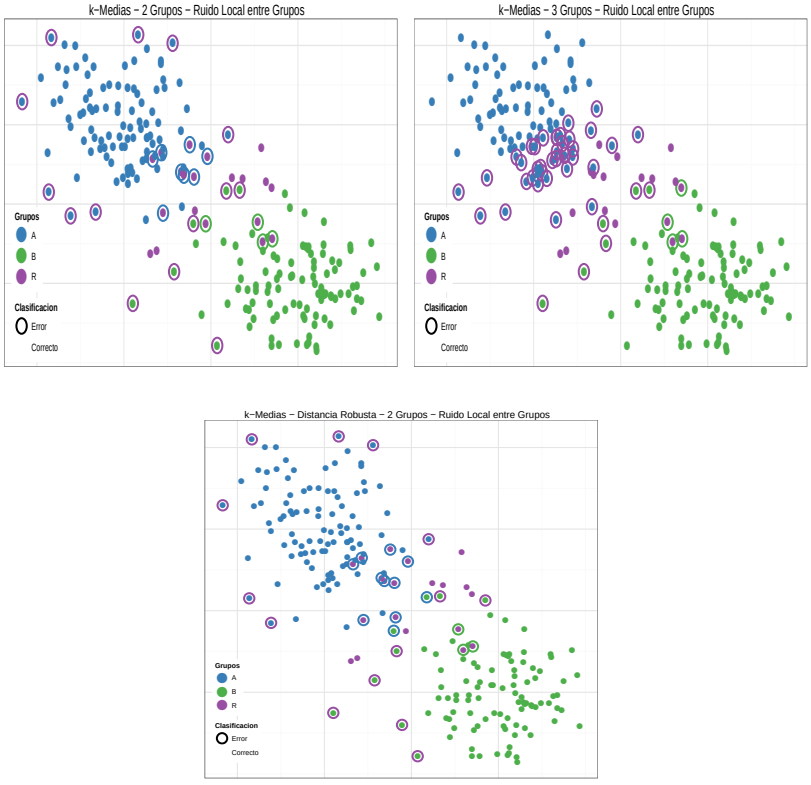


Figura 1.5: Clasificación de variantes de k-Medias en 2 grupos con Ruido Local entre Grupos.

Conclusiones

Los 3 métodos presentan ventajas y desventajas según el escenario de observaciones.

No obstante, la variante robusta es la que presenta mayor estabilidad frente a distintas tipologías de contaminación, sin ver su eficiencia comprometida en cuanto a lo que a clasificación se refiere.

Por tanto, como en general no se sabe de qué forma o múltiple formas se van a presentar los datos anómalos en la práctica, es aconsejable clasificar mediante un algoritmo que no sea altamente distorsionado por las diferentes variedades de grupos de outliers.

Si se cuenta con información acerca de que el ruido es local, es útil modelar estos como un nuevo grupo de menor tamaño y usar k-Medias sin tener que robustecer la métrica.

1.2.2. Mezcla de distribuciones t.

La mezcla de distribuciones para el modelado de datos tiene sus años en la historia de la estadística, ya (Pearson, 1894) utilizaba mezclas de gaussianas univariadas para la modelación de datos. (Wolfe, 1970) y (Day, 1969) en 1969 comenzaron el estudio de las estimaciones de los parámetros de la mezcla de forma eficiente. Actualmente el modelado mediante una mezcla finita de distribuciones tiene aplicaciones en varias ramas de la estadística. Un trabajo actual de (Melnykov, 2010) es un buen compendio de estos procedimientos.

Si bien la mezcla de normales tiene un extenso uso en modelaciones estadísticas, en particular para procedimientos de cluster, los parámetros de la mezcla de normales son muy sensibles a outliers.

Una alternativa para enfrentar este problema es dotar a las distribuciones de la mezcla de colas más pesadas, como son las distribuciones t, que soporten a estos outliers sin distorsionar en forma severa la estimación de los parámetros. Está propuesta es realizada por McLachlan y Peel en el 2000 (Peel y McLachlan, 2000).

Una manera de modelar los potenciales outliers es a través de una mezcla de dos densidades normales:

$$(1-\epsilon)\phi(y_j;\mu,\Sigma)+\epsilon\phi(y_j;\mu,k\Sigma).$$

Este modelo de mezclas lo podemos escribir de la siguiente manera:

$$\int \phi(y_j;\mu,\Sigma/u)dH(u).$$

Siendo H la distribución de una probabilidad con masa $(1-\epsilon)$ en el punto $u=1$ y con masa ϵ en el punto $u=\frac{1}{k}$. Si se sustituye la distribución de H por una chi-cuadrado con ν obtenemos una distribución de Student con parámetro de posición μ, con una matriz definida positiva Σ y ν grados de libertad,

$$f(y_j;\mu,\Sigma,\nu)=\frac{\Gamma(\frac{\nu+p}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})\{1+\delta(y_j;\mu,\Sigma)/\nu\}^{\frac{\nu+p}{2}}}$$

donde

$$\delta(y_j;\mu,\Sigma)=(y_j-\mu)^T\Sigma^{-1}(y_j-\mu),$$

denota el cuadrado de la distancia de Mahalanobis entre y_j y μ. Si $\nu>1$, μ es la media de Y_j, y si $\nu>2$ entonces $\nu(\nu-2)^{-1}\Sigma$ es la matriz de covarianzas.

Asumiendo la presencia de outliers, se modela mediante la mezcla de un número g de distribuciones t y se estiman los parámetros a través del algoritmo EM.

Se considera la estimación máximo verosímil para una mezcla de g-componentes de distribuciones t dada por

$$f(y_j;\Psi)=\sum_{i=1}^{g}\pi_i f(y_j;\mu_i,\Sigma_i,\nu_i),$$

siendo

$$\Psi=(\pi_1,\ldots,\pi_{g-1},\xi^T,\nu^T)^T,$$

donde $\nu=(\nu_1,\ldots,\nu_g)^T$ y $\boldsymbol{\xi}$ es el vector que contiene las g medias y los elementos de las g matrices de covarianzas.

El vector de datos completos está dado por

$$y_c=(y^T,Z_1^T,\ldots,Z_n^T,u_1,\ldots,u_n)^T,$$

donde $y=(y_1^T,\ldots,y_n^T)^T$ denota el vector de datos observados, $(z_1,\ldots,z_n)$ son los vectores que etiquetan el origen de $(y_1,\ldots,y_n)$ respectivamente, $z_{ij}=(z_j)_i$ es uno o cero, acorde si y_j pertenece o no a la i-ésima componente. En la caracterización a partir de las t distribuciones es también conveniente introducir en el vector de datos completos otros datos faltantes $(u_1,\ldots,u_n)$ definidos para $z_{ij}=1$ dado,

$$Y_j|u_j,z_{ij}=1\sim N(\mu_i,\Sigma_i/u_j),$$

$$U_j|z_{ij} = 1 \sim Gamma(\frac{1}{2}\nu_i, \frac{1}{2}\nu_i).$$

Dados $(z_1, \ldots, z_n)$ las v.a $(U_1, \ldots, U_n)$ son independientes. La verosimilitud de datos completos $L_c(\Psi)$ puede ser factorizada como el producto de las densidades de Z_j, por las densidades condicionales de U_j dadas las z_j y por las condicionales de Y_j dadas u_j y las z_j. Por tanto se puede escribir

$$\log L_c(\Psi) = \log L_{1c}(\pi) + \log L_{2c}(\nu) + \log L_{3c}(\xi),$$

donde

$$\log L_{1c}(\pi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log \pi_i,$$

$$\log L_{2c}(\nu) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{-\log\Gamma(\frac{1}{2}\nu_i) + \frac{1}{2}\nu_i \log(\frac{1}{2}\nu_i) + \frac{1}{2}\nu_i(\log(u_j) - u_j) - \log(u_j)\},$$

$$\log L_{3c}(\xi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{-\frac{1}{2}p\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}u_j(y_j - \mu_i)^T\Sigma_i^{-1}(y_j - \mu_i)\},$$

siendo $\pi = (\pi_1, \ldots, \pi_g)^T$ y $\xi = (\theta_1^T, \ldots, \theta_g^T)^T$ donde θ_i contiene a μ_i y a todos los elementos de Σ_i

El paso E en la $(k+1)$-ésima iteración del algoritmo EM requiere el calculo de $Q(\Psi; \Psi^(k))$, la actual esperanza condicional del logaritmo de la función de verosimilitud completa $\log L_c(\Psi)$,

El paso M en la $(k+1)$-ésima iteración de el algoritmo EM, $\pi^{(k+1)}$, $\xi^{(k+1)}$ y $\nu^{(k+1)}$ son computados independientemente uno de otros. La solución para $\pi^{(k+1)}$ y $\theta^{(k+1)}$ existen en forma cerrada.

Estudio de Simulación

Se evaluará la performance de la clasificación de mezclas de normales y mezclas de t de Student con distintos grados de libertad (4 y 12 grados de libertad respectivamente) mediante la simulación de un escenario 150 veces.

En todos los casos se determina el grupo de outliers podando el 15 % de las observaciones con menor verosimilitud en la distribución de la mezcla.

Para cada escenario se simulan 350 observaciones. Estas provienen de:

- 100 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0,5 \\ 0,5 & 0,5 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -0,5 \\ -0,5 & 0,5 \end{pmatrix}\right]$
- 50 observaciones uniformes en el cuadrado $[-10, 10]^2$

Las observaciones provenientes de las distribuciones normales determinan 3 grupos de 100 observaciones cada uno, mientras que las 50 observaciones uniformes conforman el ruido global.

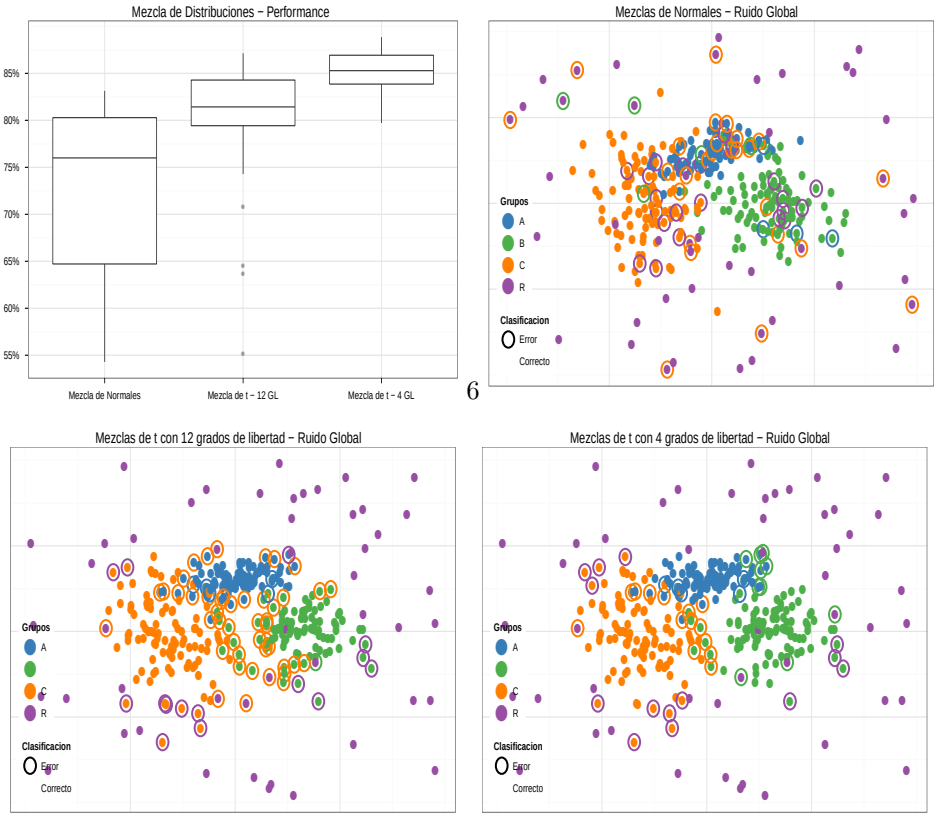
Para poder evaluar de manera correcta qué algoritmo clasifica de forma más efectiva, al igual que en el capítulo anterior, en cada simulación del escenario se computa el porcentaje de datos bien clasificados por cada técnica. Se realizan los diagramas de caja para estos porcentajes en cada caso (ver figura (1.6)).

Analizando estos diagramas se puede observar cómo las colas pesadas de la distribución de Student con 4 grados de libertad soporta de forma más estable a los outliers. Al aumentar los grados de libertad en la distribución de Student, esta pierde peso en sus colas y se asemeja a una normal, bajando el porcentaje de datos bien clasificados de forma brusca.

1.2.3. Podado en Cluster

El método propuesto por Garcia-Escudero, Gordaliza, Matran y Mayo-Iscar (Matrán et al., 2008), (Gordaliza et al., 2010), se trabaja con clusters de distinto peso y dispersión, admitiendo un proporción α de outliers.

El análisis que se desarrolla es el de podar las observaciones menos confiables, el cual no es para nada trivial, debido a que no existen direcciones privilegiadas en la búsqueda y que muchas veces es necesario eliminar observaciones "puente" entre los cluster.



6

Figura 1.6: Diagramas de caja del porcentaje de observaciones bien clasificadas por cada algoritmo y clasificación mediante diferentes mezclas

Si bien se han introducido métodos de penalización y poda en el método de k-Medias ellos muestran poseer mejores resultados en términos de robustez, además se levanta el supuesto implícito de que la matriz de covarianza es la misma y esférica para los grupos en el algoritmo de k-Medias.

Se afronta el problema desde una perspectiva diferente al capítulo anterior, en lugar de dotar a las distribuciones de colas más pesadas para que los outliers tengan un menor impacto sobre las estimaciones, los poda, partiendo de que son valores anómalos, que no provienen del modelo.

Se considera la presencia de una proporción α de outliers. La función de verosimilitud para el conjunto de datos $x_1, \ldots, x_n$ en este caso es:

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma)\right]\left[\prod_{i\notin R} g_{\Psi_i}(x_i)\right], \tag{1.2}$$

con $R = \cup_{j=1}^{k} R_j$ y $\#R = [n(1-\alpha)]$.

El parámetro k denota el número total de grupos, R_j contiene los índices de las observaciones "regulares" asignadas al grupo j y $f(\cdot;\mu;\Sigma)$ es la función de densidad de una distribución normal p-variada con media μ y matriz de covarianza Σ, mientras que las g_{Ψ_i} son alguna función de densidad en $\mathbb{R}^p$.

Si se elige $\Sigma = \sigma^2 I$, entonces se está realizando el método de k-Medias podadas. (Ritter y Gallegos, 2005) mostraron que la maximización se reduce a la consideración de la parte regular de las observaciones bajo algunos supuestos razonables para las g_{Ψ_i}s siempre y cuando las observaciones "no regulares" pueden ser vistas meramente como "ruido". El problema de maximizar esta verosimilitud es costoso computacionalmente. Para alivianar este problema es donde participa el algoritmo EM condicional.

El supuesto de igualdad de matrices de covarianza para los grupos puede ser restrictivo en muchos contextos, y sería un supuesto a levantar. Desafortunadamente, este problema de Clustering robusto es notablemente complejo. Gallegos y Ritter (2005) mantienen el supuesto de igualdad de matrices de covarianza y levantan el supuesto de esfericidad de las matrices de covarianzas, modelo que llaman spurious-outliers model (supuesto de igualdad que eliminan en el 2009).

Actualmente se han encontrado respuestas parciales, en general, imponiendo restricciones sobre las distintas matrices de covarianzas (se admite una moderada diferencia en las dispersiones). Es fácil ver la no acotación de la funcion objetivo perseguida, como

cada punto de los datos hace surgir una singularidad en el borde del espacio paramétrico. Si se utilizan métodos no restringidos frecuentemente se encuentran clusters conteniendo unos pocos puntos, ya sea muy juntos o casi estando en un espacio de menor dimensión, y la aplicación de algún tipo de restricción permitiría obtener particiones más interesantes o informativas.

Como forma de poner restricciones al problema, (Gallegos, 2002) propone normalizar las covarianzas para que tengan un determinante de unidad cuando se computan las distancias de Mahalanobis en el paso de "concentración". Esto sirve para evitar el efecto pernicioso de las diferentes escalas y beneficiarse de la lógica detrás del algoritmo Fast-MCD.

El procedimiento de Gallegos funciona adecuadamente cuando los grupos tiene escalas similares, pero claudica cuando escalas de grupos muy diferentes están involucradas. Normalizar las covarianzas para tener un determinante de unidad puede ser muy restrictivo y, seguramente, tales restricciones fuertes no se necesitan siempre. Aún más, parece también adecuado el incorporar restricciones directamente en la definición del problema en vez de aparecer (artificialmente) en el algoritmo.

Por tanto serán planteadas restricciones para el problema de Clustering robusto heterogéneo, incorporado a través de una restricción en el cociente de los valores propios, donde c será una constante que controlará la fuerza de la restricción planteada.

La introducción de algunos términos π_js de pesos serán considerados para tratar con grupos de distintos pesos, lo que hace el problema más general pero más duro de ser trabajado.

(Ritter y Gallegos, 2009) propone restricciones sobre las matrices de varianzas y covarianzas a partir del orden de *Löwner* las cuales son llamadas restricciones HDBT (en referencia histórica a sus creadores Hathaway, Dennis, Beale y Thompson).

Si denotamos $V_1, V_2, \ldots, V_g$ matrices de varianzas y covarianzas que cumplen:

$$V_j \succeq cV_l, \quad 1 \leq j, l \geq g, \tag{1.3}$$

para alguna constante $c > 0$. Donde el símbolo $\succeq$ establece un orden entre las matrices semidefinidas positivas y c es necesariamente acotada entre $(0, 1]$. Si $c = 1$ estamos en el caso de homosedasticidad.

Se define la proporción HDBT de la g-upla $V = (V_1, V_2, \ldots, V_g)$ al máximo valor de c que verifiquen las restricciones ((1.3)). Se puede observar que

$$r_{HDBT}(V) = \text{máx}\,\{c/V_j \succeq cV_l \quad j,l\} = \min_{j,l,k} \lambda_k(V_l^{-1/2} V_j V_l^{-1/2}) \tag{1.4}$$

donde $\lambda_1(A), \ldots, \lambda_d(A)$ denotan los valores propios de la matriz de A.

(Ritter y Gallegos, 2009) demuestran que las restricciones HDBT son suficientes para asegurar la existencia del máximo de lo que llaman criterio del determinante trimeado (TDC).

Sean $x_1, \ldots, x_n$ los datos disponibles en algún espacio p-dimensional. Sea $f(x; \mu; \Sigma)$ la densidad de una distribución normal de la forma

$$f(x; \mu; \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-(x-\mu)\prime \Sigma^{-1}(x-\mu)/2),$$

Denotamos una medida de probabilidad P actuando sobre una función f por $Pf(\cdot) = \int f(x) dP(x)$.

Se comienza modificando el modelo de "outlier espurio" considerado en (Ritter y Gallegos, 2005).

Primero, como se mencionó antes, se consideran diferentes matrices de dispersión Σ_is como en Gallegos (2001, 2002). Se asume la presencia de algunos pesos subyacentes, π_js con $\sum_{j=1}^{k} \pi_j = 1$ asociados a las distribuciones del conjunto de observaciones "regulares". Esto lleva a la maximización de

$$\left[\prod_{j=1}^{k} \prod_{i \in R_j} \pi_j f(x_i; \mu_j, \Sigma_j) \right] \left[\prod_{i \notin R} g_{\Psi_i}(x_i) \right], \tag{1.5}$$

con $R = \cup_{j=1}^{k} R_j$ y $\#R = n - [n\alpha]$. Adicionalmente, las restricciones sobre los valores propios de las matrices Σ_j serán introducidos mas tarde para evitar singularidades. Si las g_Ψ's satisfacen la condición

$$\arg\max_{\mathcal{R}} \max_{\mu_j, \Sigma_j} \prod_{j=1}^{k} \prod_{i \in R_j} \pi_j f(x_i; \mu_j, \Sigma_j) \subseteq \arg\max_{\mathcal{R}} \prod_{i \notin \cup_{j=1}^{k} R_j} \max_{\Psi_i} g_{\Psi_i}(x_i),$$

donde $\mathcal{R}$ denota el conjunto de todas las particiones de índices $1, \ldots, n$ en k grupos de observaciones regulares, R, y un grupo conteniendo las no regulares, con $\#R = n - [n\alpha]$. Esta condición se cumple bajo algunos supuestos razonables para las g_{Ψ_i}s siempre y cuando las observaciones "irregulares" sean vistas como mero "ruido".

Se usarán algunas funciones de asignación, z_js, diciendo a cual clase todo punto x en $\mathbb{R}^p$ es asignado (no solo las observaciones de la muestra, x_i's son clasificadas). Se utiliza un enfoque 0-1 "seco" donde x es asignado a la clase j si $z_j(x) = 1$ o es podada si $z_0(x) = 1$.

Con estas funciones, asumiendo que las g_{Ψ_i}'s pueden ser omitidas, podemos ver nuevamente el problema en (1.5) a la maximización de

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{k}\pi_j^{z_j(x_i)}f(x_i;\mu_j,\Sigma_j)^{z_j(x_i)}\right],$$

siendo z_j las funciones 0-1 definidas en todo el espacio de la muestra verificando $\sum_{j=0}^{k} z_j(x_i) = 1$ y $\sum_{i=1}^{n} z_0(x_i) = [n\alpha]$.

Tomando logaritmos para simplificar la expresión se obtiene la formulación del problema.

Problema de Clustering Robusto. Dada una medida de probabilidad P, se busca la maximización de

$$P\left[\sum_{j=1}^{k} z_j(\cdot)(\log\pi_j + \log f(\cdot,\mu_j,\Sigma_j))\right], \tag{1.6}$$

realizada en términos de las funciones de asignación:

$$z_j : \mathbb{R}^p \to \{0,1\}, \text{ de forma que } \sum_{j=0}^{k} z_j = 1 \text{ y } Pz_0(\cdot) = \alpha,$$

y los parámetros $\theta = (\pi_1,\ldots,\pi_k,\mu_1,\ldots,\Sigma_1,\ldots,\Sigma_k)$ correspondientes a los pesos $\pi_j \in [0,1]$ con $\sum_{j=1}^{k}\pi_j = 1$, vectores de medias $\mu_j \in \mathbb{R}^p$ y matrices de $p \times p$ simétricas semidefinidas positivas Σ_j, con $j = 1,\ldots,k$.

Si P_n denota la medida empírica, $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. Si se remplaza P por P_n en el problema previo para recuperar el problema original de la muestra (notar que, quizás, $P_n z_0(\cdot) = \alpha$ no puede ser exactamente alcanzado).

Se introducen restricciones de valores propios a las matrices de covarianzas que permite evitar las singularidades introducidas por la posibilidad de Σ_js muy diferentes, mediante el control del cociente entre el máximo y el mínimo de los valores propios de esas matrices:

(ER) Restricciones sobre el Cociente de Valores Propios. Se fija una constante $c \geq 1$ de forma que

$$M_n/m_n \leq c$$

para

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\Sigma_j) \text{ y } m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\Sigma_j)$$

donde $\lambda_l(\Sigma_j)$ son los valores propios de las matrices Σ_j, $j = 1, \ldots, k$ y $l = 1, \ldots, p$.

Se denota por Θ_c al conjunto constituido por los θs que cumplen la condición ER para un c dado.

Notar que, la restricción más fuerte posible surge de establecer $c = 1$. En este caso en particular, el método propuesto puede ser visto como un procedimiento de k-Medias podadas con pesos. Sin embargo, la ventaja principal de este enfoque yace en el hecho de que el parámetro c permite alcanzar cierta (controlada) libertad en como se puede manejar las diferentes dispersiones de los grupos.

La condición se cumple trivialmente si la distribución P subyacente es continua o si es una medida empírica P_n correspondiente a una muestra de una distribución absolutamente continua (para n suficientemente grande): la distribución P no está concentrada en k puntos después de remover una masa de probabilidad igual a α.

Dado $\theta \in \Theta_c$, se considera alguna función discriminante definida como

$$D_j(x; \theta) = \pi f(x; \mu_j, \Sigma_j),$$

y

$$D(x; \theta) = \max\{D_1(x; \theta), \ldots, D_k(x; \theta)\}.$$

Estas funciones sirven para determinar las observaciones más "outliers". Para una elección fija de θ, cuanto $D(x; \theta)$ más chica sea para un x dado, más se lo considerara "outlier".

Usando las definiciones previas, para un θ dado y una medida de probabilidad P, se define,

$$G(\cdot; \theta, P) : u \in \mathbb{R} \to P\left[I_{[0,u]}(D(\cdot; \theta))\right], \tag{1.7}$$

y

$$R(\theta, P) := G^{-1}(\alpha; \theta, P) = \inf_u \{G(u; \theta, P) \geq \alpha\}$$

(notar que si X es una variable aleatoria con distribución dada por P entonces $R(\theta, P)$ es el α-cuantil de la variable aleatoria $D(X;\theta)$).

Con esta notación, se tiene la siguiente caracterización para las funciones de los z_js:

Se asigna x a la clase j con el valor más alto de la función discriminante, $D_j(x;\theta)$, o x es podado cuando todos los $D_j(x;\theta)$s (y consecuentemente $D(x;\theta)$) son mas chicos que $R(\theta, P)$. Una regla para romper empates en los valores de las funciones discriminantes también es también necesaria. Por ejemplo, se podría aplicar el orden lexicógrafo.

Existencia y Consistencia

Teorema 3 (Existencia). Si (2.3) se cumple para la medida de probabilidad P, entonces existe algún $\theta \in \Theta_c$ de forma tal que el máximo de (2.6) bajo las restricción ER es alcanzado.

Dada $\{x_n\}_{n=1}^{\infty}$ una muestra aleatoria i.i.d. de la distribución de probabilidad subyacente (desconocida) P, sea $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1, \ldots, \pi_k^n, \mu_1^n, \ldots, \mu_k^n, \Sigma_1^n, \ldots, \Sigma_k^n)\}_{n=1}^{\infty} \subset \Theta_c$ la sucesión de los estimadores de la muestra obtenidos por resolver el problema para las medidas empíricas $\{P_n\}_{n=1}^{\infty}$ con la restricción de valores propios definida por ER para una constante fija $c \geq 1$.

La sección 2.2 muestra que tal secuencia siempre existe, para un n suficientemente grande siempre y cuando P es una distribución absolutamente continua verificando (2.3). Notar que aunque notación similar a la aplicada en la sección previa será usada, aquí el indice n indicara la dependencia en una muestra aleatoria de tamaño n para B.

Se observará primero que existe un conjunto compacto $K \subset \Theta_c$ tal que $\theta_n \in K$ para n suficientemente grande con probabilidad 1.

Teorema 4 (Consistencia). Asumir que P tiene una función de densidad estrictamente positiva y que θ_0 es el único máximo, bajo la restricción ER. Si $\theta_n \in \Theta_c$ denota la versión de la muestra del estimador basado en la medida empírica P_n, entonces $\theta_n \to \theta_0$ casi seguramente.

Notar que la condición de unicidad es necesaria para establecer el resultado de consistencia.

Desafortunadamente, esta propiedad no siempre se cumple. Por ejemplo, pensar en una mixtura simétrica P en la recta real con dos modas bien separadas, un nivel alto de podado y $k = 1$.

La propiedad de unicidad era ya era necesaria para establecer el mismo resultado de consistencia para k-Medias podadas y, aun en este caso mas simple, el enunciado de los resultados generales de unicidad eran difíciles (ver Observación 4.1 en García–Escudero et al. 1999).

Sin embargo, como en el problema de las k-Medias podadas, se cree que es bastante raro el encontrar una distribución donde esta unicidad falle, cuando se trata con datos "razonables" para el agrupamiento y cuando los parámetros k y α han sido propiamente escogidos.

El Algoritmo

El problema empírico presentado tiene obviamente una complejidad computacional muy alta. Un algoritmo exacto parece no ser factible aun para tamaños de muestra moderados. Entonces la existencia de un algoritmo adecuado para resolver aproximadamente el problema de la muestra puede ser tan importante como el procedimiento en sí mismo.

El algoritmo TCLUST es un algoritmo basado en el principio EM, planteado para buscar soluciones aproximadas. El EM es el método usual para obtener una solución al problema de la mezcla de verosimilitudes (Dempster et al. 1997). Aquí, se sigue un enfoque "seco" donde cada punto es asignado únicamente a un cluster. Las restricciones sobre los valores propios serán incorporadas a través del algoritmo de Dykstra (1983).

El algoritmo TCLUST puede ser descrito de la siguiente forma:

1. Seleccionar valores aleatorios para los centros m_j^0s, las matrices de covarianzas S_j^0s y los pesos de los grupos p_{j^0}s para $j = 1, \ldots, k$.

2. Desde el $\theta^l = (p_1^l, \ldots, p_k^l, m_1^l, \ldots, m_k^l, S_1^l, \ldots, S_k^l)$ retornado por la iteración previa:

   a) Obtener $d_i = D(x_{i,\theta^l})$ para las observaciones $\{x_1, \ldots, x_n\}$ y mantener el conjunto H teniendo las $[n(1-\alpha)]$ observaciones con las mas grandes d_is.

   b) Dividir H en $H = \{H_1, \ldots, H_k\}$ con $H_j = \{x_i \in H : D_j(x_i, \theta^l) = D(x_i\theta^l)\}$.

   c) Obtener el número de datos n_j en H_j, su media y matriz de covarianzas muestrales, m_j y S_j, $j = 1, .., k$.

   d) Considere la descomposición en valores propios de $S_j = U_j' D_j U_j$ donde U_j es una matriz ortogonal y $D_j = diag(\Lambda_j)$ es una matriz diagonal (con los elementos en la diagonal dados por el vector Λ_j). Si el vector entero de valores

propios $\Lambda = (\Lambda_{1,\ldots,\Lambda_k})$ no satisface la restricción de valores propios, obtener un nuevo vector $\tilde{\Lambda} = (\tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_k)$ a través del algoritmo de Dykstra que obedezca la restricción ER y que $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$ sea lo mas chico posible. Λ^{-1} denota el vector compuesto por los inversos de los elementos del vector Λ. Notar que la restricción ER para Λ corresponde a la misma restricción ER aplicada a Λ^{-1}.

e) Actualizar θ^{l+1} usando:

- $p_j^{l+1} \leftarrow n_j/[n(1-\alpha)]$
- $m_j^{l+1} \leftarrow m_j$
- $S_j^{l+1} \leftarrow U_j' \tilde{D}_j U_j$ y $\tilde{D}_j = diag(\tilde{\Lambda}_j)^{-1}$

3. Realizar F iteraciones del proceso descrito en el paso 2 (valores moderados para F son usualmente suficientes) y computar la función de evaluación $L(\theta^F; P_n)$.

4. Obtener valores de partida aleatorios (e.g. comenzar desde el paso 1) varias veces, mantener las soluciones que llevan a valores mínimos de $L(\theta^F; P_n)$ e iterar completamente sobre ellos para elegir el mejor.

Las probabilidades "a posteriori" computadas (paso E), $D_j(x_i, \theta^l) = p_j f(x_i; m_j, S_j)$, son convertidas a una clasificación discreta donde se deja sin asignar la proporción α de observaciones las cuales son las más difíciles de clasificar. Es fácil de ver que esto lleva a una asignación óptima.

Después se obtiene un nuevo θ^{l+1} por maximizar (paso M) la esperanza condicional una vez que todas las observaciones no podadas han sido asignadas a los grupos. La proposición 3 garantiza que el algoritmo presentado puede ser usado para realizar esta maximización.

Notar que la obtención de las matrices de dispersión óptimas se descompone en la búsqueda de los correspondientes valores y vectores propios óptimos. Para cada elección de valores propios, la elección de los mejores vectores propios surge simplemente de los vectores propios unitarios de la matriz de covarianzas muestral de las observaciones asignadas a cada grupo. Esta descomposición es de alguna forma similar a la considerada en la propuesta de Gallegos, donde las "formas" y las "escalas" son tratadas de forma separada.

Si se ve a $D(x_i, \theta^l)$ como medida inversa de atípico para la observación x_i con respecto a la elección de θ^l, entonces el paso 2 puede ser visto como cierto tipo de pasos

de “concentración”. Garcia-Escudero y Gordaliza (2006) analizan otros intentos para extender el principio del paso de “concentración” a la configuración de Clustering robusto heterogéneos.

Recordar que la esquema de inicialización aleatoria (paso 1) y el refinamiento final (paso 4) eran muy importantes en el algoritmo Fast-MCD. Para inicializar el procedimiento en el paso 1, se ha visto que simplemente elegir k puntos de la muestra para los centros, k matrices identidad para las matrices de covarianzas y los mismos pesos para los grupos (iguala $1/k$) provee una un punto de partida razonable en la mayoría de los casos.

Con respecto a la restricción de valores propios, se podría necesitar $\Lambda = (\Lambda_1, \ldots, \Lambda_k)$ con $\Lambda_j = (\lambda_{1,j}, \ldots, \lambda_{p,j})$ pertenecientes al cono $\mathcal{C}$, donde

$$\mathcal{C} = \{(\Lambda_1, \ldots, \Lambda_k) \in \mathcal{R}^{p \times q} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0 \text{ para todo } (u,v) \neq (r,s)\}. \tag{1.8}$$

Si $\Lambda \in \mathcal{C}$, se necesita remplazar L^{-1} por $\hat{\Lambda} \in \mathcal{C}$ con $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$ mínimo. El algoritmo de Dykstra sirve para resolver aproximadamente ese problema, donde además de la lógica detrás del Fast-MCD (Rousseeuw y van Driessen 1999) y detrás de algoritmo de k-Medias podadas (Garcia-Escudero et al 2003), también subyacerán mínimos cuadrados con restricciones cuando $\mathcal{C}$ es la intersección de varios conos cerrados convexos mediante el reordenamiento a proyecciones iterativas en los conos individuales.

Notar que $\mathcal{C}$ puede ser visto como la intersección de los conos

$$\mathcal{C}_h = \{(\Lambda_1, \ldots, \Lambda_k) \in \mathcal{R}^{p \times q} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0\}, h = (u,v,r,s)\},$$

y las proyecciones en los conos $\mathcal{C}_h$ son rápidas de obtener. Entonces un número fijo de proyecciones individuales pueden ser realizadas reteniendo la mejor solución alcanzada después de esas iteraciones y satisfaciendo las restricciones. Alternativamente, soluciones basadas en programación cuadrática pueden ser utilizadas (ver, Goldfarb e Idnani (1983)).

El siguiente resultado sirve para formalizar lo apropiado del algoritmo TCLUST:

Teorema 5. Si los conjuntos $H_j = \{x_i : z_j(x_i) = 1\}, j = 1, \ldots, k$, son mantenidos fijos, el máximo de (2.2) para $P = P_n$ puede ser obtenido a través de los siguientes pasos:

1. Fijada μ_j y Σ_j, la mejor elección de π_j es $\pi_j = n_j/[n(1-\alpha)]$, donde $n_j = \#H_j$.

2. Fijada Σ_j y los valores óptimos para π_j dados en (1), la mejor elección para μ_j es la media muestral m_j de las observaciones en H_j.

3. Fijado los valores propios para la matriz Σ_j y los valores óptimos dados en (1) y (2), la mejor elección para el conjunto de vectores propios son los vectores propios unitarios de la matriz de covarianza S_j de las observaciones en H_j.

4. Con las selecciones óptimas hechas en (1), (2), y (3), la mejor elección para los valores propios corresponde a la proyección del vector conteniendo los inverso de los valores propios en el cono $\mathcal{C}$ en (3.1).

Estudio de Simulación

Se considera el mismo escenario del capítulo anterior:

- 100 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0,5 \\ 0,5 & 0,5 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -0,5 \\ -0,5 & 0,5 \end{pmatrix}\right]$
- 50 observaciones uniformes en el cuadrado $[-10, 10]^2$

Se compara ahora la performance del algoritmo TCLUST contra el de Mezcla de Normales. De la misma forma que se viene trabajando se realizan 150 repeticiones y se estudia el porcentaje de datos correctamente clasificados .

Se puede apreciar que la poda de datos contribuye a una mejor eficiencia del algoritmo en presencia de ruido global. Se podría observar también que si el ruido es local aumenta en forma considerable las diferencias entre la performance de un algoritmo y otro si se sigue modelando con 3 grupos.

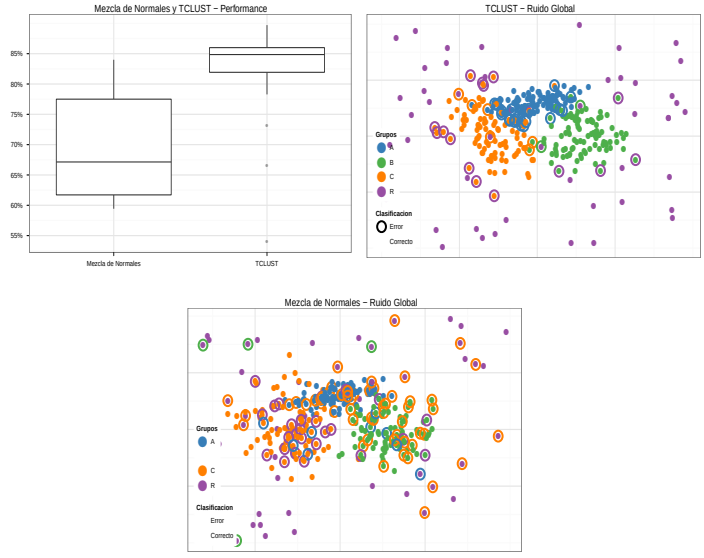


Figura 1.7: Diagramas de caja del porcentaje de observaciones bien clasificadas por cada algoritmo y clasificación mediante mezcla de normales y TCLUST.

1.3. Comparación mediante simulación de los métodos robustos

Se analizarán los tres algoritmos robustos propuestos según su desempeño bajo diversos escenarios.

Se realizarán modificaciones sobre la forma de los grupos (esférico o elíptico), sobre el tamaño de éstos (grupos de igual o distinto tamaño) y sobre la forma del ruido (global o local), variando el porcentaje de contaminación introducido.

Los escenarios a analizar son los siguientes:

Escenario 1: Se consideran tres grupos esféricos de igual tamaño y ruido global uniformemente distribuido con un porcentaje de contaminación del 10 %.

Escenario 2: Tres grupos esféricos de igual tamaño y ruido global uniformemente distribuido con un porcentaje de contaminación del 25 %. Se mantiene el modelado del caso anterior pero con una mayor cantidad de ruido.

Escenario 3: Se consideran tres grupos esféricos de igual tamaño y ruido uniforme sesgado hacia una de los cuadrantes. El porcentaje de contaminación es del 20 %.

Escenario 4: Se varía los supuestos sobre la distribución de los grupos. Se consideran grupos elípticos de distinto tamaño. El ruido es global y uniformemente distribuido, con un porcentaje de contaminación del 10 %.

En cada uno de los escenarios anteriormente se realizan 150 repeticiones y se estudia el porcentaje de datos correctamente clasificados por:

- Variante Robusta de k-Medias
- Mezcla de Distribuciones t
- El algoritmo TCLUST

1.3.1. Escenario 1

Se considera en este caso 500 observaciones bivariadas, tres grupos conformados cada uno por 150 observaciones y un 10 % de ruido global distribuido uniformemente:

- 150 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 150 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 150 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 50 observaciones uniformes en el cuadrado $[-10, 10]^2$

Como se puede apreciar en los respectivos diagramas de caja en la figura (1.8), quien clasifica mejor es el algoritmo mediante Mezcla de distribuciones t.

El poco ruido global es soportado por las colas pesadas de la distribuciones t, lo que hace posible un estimación eficiente de los centros de los clusters.

Sin embargo la variante robusta de k-Medias presenta una eficiencia no tanto menor, pero con una mayor variabilidad.

Es de hacer notar la alta variabilidad de algoritmo TCLUST. Posiblemente la poda incorrecta de algunas observaciones en cada simulación produce este efecto.

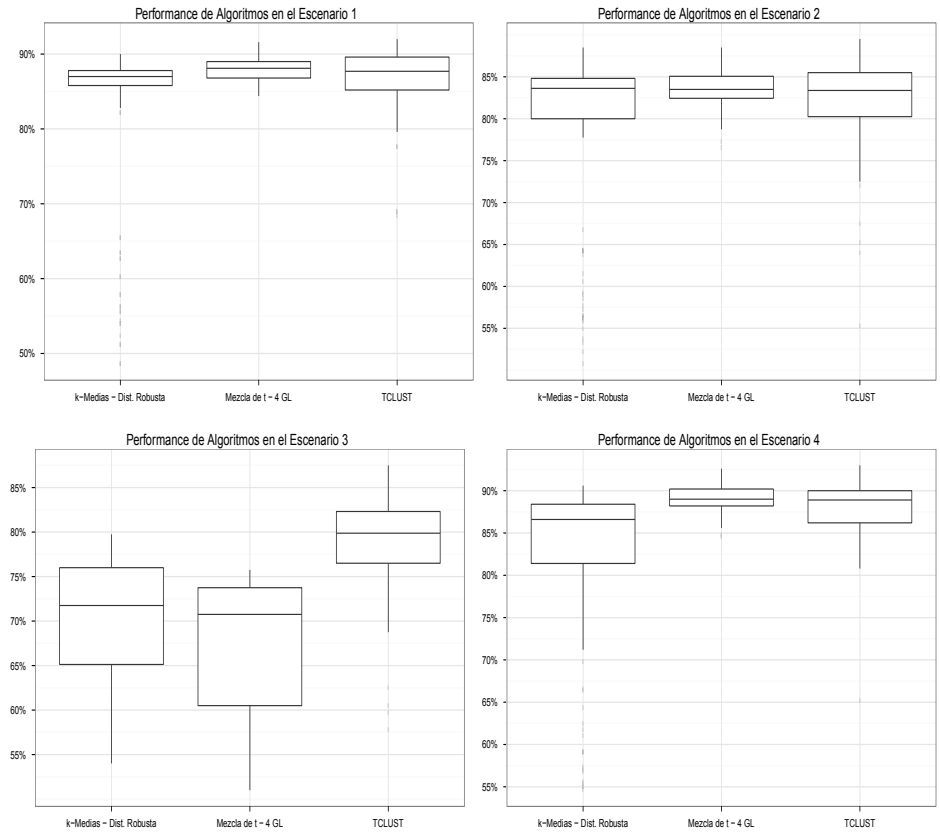


Figura 1.8: Diagramas de cajas del porcentaje de observaciones bien clasificadas por cada algoritmo en cada uno de los escenarios

1.3.2. Escenario 2

Se consideran 400 datos bivariados, se mantienen los tres grupos pero ahora con 100 observaciones en cada grupo y aumenta el porcentaje de ruido a un 25 %:

- 100 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones uniformes en el cuadrado $[-10, 10]^2$

Bajo esta tipología de datos, como se aprecia en los diagramas de caja expuestos en la figura (1.8), la mezcla de distribuciones t baja notoriamente su performance respecto al escenario anterior, debido al aumento en el número de outliers.

Aunque disminuyendo los grados de libertad de las distribuciones t se mejora la eficiencia, las colas de la distribución no soportan el alto número de outliers, produciendo una estimación errónea de los centros de los clusters.

La variante robusta de k-Medias, al partir de una métrica acotada entre 0 y 1, el alto porcentaje de outliers no afecta en gran medida la estimación de los centros de los grupos.

Al igual que en el escenario anterior se observa como el algoritmo TCLUST presenta una alta variabilidad frente a una contaminación global.

1.3.3. Escenario 3

En este escenario se mantienen los supuestos del caso anterior pero se sesga los datos atípicos al cuadrado $[0, 10] \times [0, 10]$:

- 100 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$

- 100 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones uniformes en el cuadrado $[0, 10]^2$

Se observa en la figura (1.8) como esta tipología de outliers impacta negativamente sobre la mezcla de distribuciones, pero la variante robusta de k-Medias se mantiene casi invariante frente a este cambio.

Al igual que en el ejemplo anterior el TCLUST presenta una alta dispersión, pero en casi todas las simulaciones un mayor eficiencia que la mezcla de distribuciones t.

1.3.4. Escenario 4

Se levanta ahora el supuesto de distribuciones esféricas y de igual tamaño de los grupos . Al igual que en los casos anteriores, se simulan normales bivariadas pero con diferentes matrices de varianzas y covarianzas. El ruido, como en el escenario 1, es global de un 10 % del total de la muestra, distribuido uniformemente en el cuadrado $[-10, 10] \times [-10, 10]$

- 150 observaciones de una $N\left[\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 & 0,5 \\ 0,5 & 0,5 \end{pmatrix}\right]$
- 200 observaciones de una $N\left[\begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}\right]$
- 100 observaciones de una $N\left[\begin{pmatrix} -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -0,5 \\ 0,5 & 5 \end{pmatrix}\right]$
- 50 observaciones uniformes en el cuadrado $[-10, 10]^2$

Es visible en los diagramas de caja representados en la figura (1.8) como la variante robusta del algoritmo de k-Medias es el que peor clasifica en estos casos.

Este algoritmo está diseñado para detectar grupos esféricos y de igual tamaño. Cuando estos supuestos se levantan, el algoritmo pierde eficiencia. Si bien los restantes algoritmos clasifican de forma semejante, el algoritmo de distribuciones t presenta menor variabilidad.

Como se pudo comprobar vía simulación, la performance de los algoritmos tratados presentan una alta dependencia respecto a la distribución de los datos, así como también a la tipología de los outliers.

Cuando la contaminación se presenta en un bajo porcentaje, en forma global y uniforme, la mezcla de t parece ser la mejor opción, puesto que las colas pesadas soportan estos valores y no es necesario la poda.

Cuando la contaminación es elevada o sesgada respecto a la distribución de los datos, si los grupos son esféricos y de igual tamaño la variante robusta de k-Medias es el que clasifica mejor.

Sin embargo, si los grupos presentan formas elípticas y el ruido toma formas menos previsibles se observa como el TCLUST supera a los algoritmos anteriores, teniendo una elevada precisión con respecto a los datos bien clasificados en cada simulación.

1.4. Datos reales

Con el objetivo de identificar observaciones atípicas, se aplicaron las técnicas a un conjunto de datos provisto por una inmobiliaria de Montevideo.

Este cuenta con 190 observaciones que representan propiedades que estuvieron o están a la venta en la inmobiliaria en el último año, de las cuales se tienen las siguientes variables:

Variable	Descripción
`id`	Identificador numérico de la Propiedad
`zona`	Nombre de la zona en que se encuentra la Propiedad
`precio`	Precio en dólares de la Propiedad
m^2 `const`	Metros cuadrados construidos en la Propiedad
m^2 `terreno`	Metros cuadrados de la Propiedad

Tabla 1.1: Conjunto de datos de propiedades de Montevideo y Canelones

De acuerdo a lo conversado con la gerencia de la inmobiliaria, esta percibe 4 estratos dentro del conjunto de datos. El objetivo del análisis será delimitar esos estratos e identificar datos atípicos con el fin de evaluar la política de precios empleada.

1.4.1. Descripción de los datos

El conjunto de datos contiene observaciones de 7 zonas. Se puede apreciar una correlación lineal positiva entre los metros de la propiedad y su precio en todas las zonas. Dicha relación lineal varía su pendiente de acuerdo a la zona. También se verifica dicha relación entre precio y metros cuadrados construídos.

		Precio (Miles)		Metros		Metros Const.	
Zona	Obs.	Media	Desvío	Media	Desvío	Media	Desvío
Barra de Carrasco	10	119.10	44.43	660.20	186.87	157.90	46.21
Carrasco	51	359.69	243.65	811.69	671.11	299.45	266.45
Carrasco Norte	24	153.83	96.76	878.08	708.77	202.38	130.03
Malvín	12	204.58	70.40	559.67	264.89	267.75	120.39
Parque Miramar	46	197.46	281.78	778.35	875.16	198.93	104.37
Punta Gorda	42	256.79	145.40	620.26	289.58	260.36	131.36
Shangrilá	5	94.00	54.70	903.40	290.19	141.60	36.54

Mediante una inspección visual primaria, podrían existir datos atípicos en las zonas de Carrasco, Carrasco Norte, Parque Miramar y Punta Gorda.

En las zonas de Barra de Carrasco y Malvín, las gráficas sugieren que el tamaño del terreno y sus metros cuadrados construidos no contribuyen al precio de la propiedad. Dicha relación (independencia) parece ser más débil en Shangrilá.

En Punta Gorda, y en menor medida en Carrasco Norte, los metros construidos parecen tener un gran impacto en el precio de la propiedad.

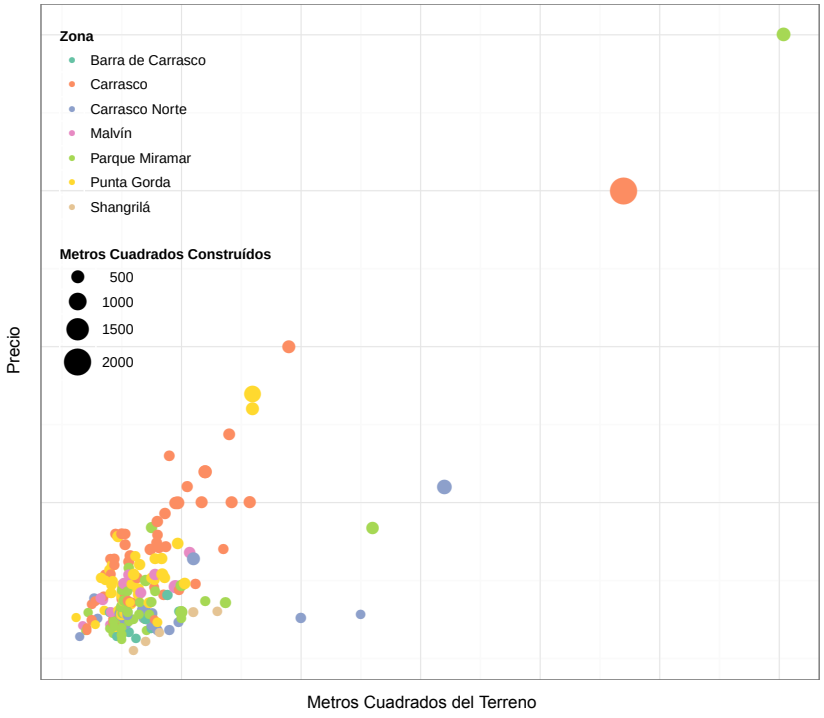


Figura 1.10: Conjunto de Datos

1.4.2. Algoritmo TCLUST

Se aplicó el algoritmo TCLUST al conjunto de datos para encontrar 4 grupos con una poda del 10 %.

Como se puede apreciar en la tabla (1.2), el grupo `A` se encuentra caracterizado por pocas propiedades de más alto valor, con terrenos y edificaciones más grandes.

		Precio (Miles)		Metros		Metros Const.	
Zona	Obs.	Media	Desvío	Media	Desvío	Media	Desvío
R	19	576.84	498.73	1926.53	1409.86	493.00	410.99
A	7	528.57	48.80	1211.86	224.22	452.86	47.16
B	40	321.12	61.92	678.27	183.74	297.32	44.89
C	106	146.43	51.36	509.97	160.14	160.92	48.01
D	18	166.28	44.25	882.50	224.10	234.89	55.29

Tabla 1.2: Descriptiva de los grupos identificados por TCLUST

El grupo `B` se caracteriza por propiedades del 60 % del valor –en promedio– que las del grupo `A`, con terrenos de la mitad de tamaño pero la mitad del terreno se encuentra edificado en vez de un cuarto del primer grupo.

El grupo `C` es el más numeroso y está compuesto por las propiedades más pequeñas, de menor valor y con menos metros cuadrados construidos.

El grupo `D` se trata de 18 propiedades cuyo precio promedio es un 15 % mayor a las del grupo `C`, pero el tamaño del terreno es un 70 % superior y los metros construidos un 40 % mayor.

En cuanto a la composición de zonas de los grupos (figura (1.14)), el grupo `A` –propiedades de mayor valor– está compuesto únicamente por propiedades en Carrasco.

El grupo `B` está conformado en su mayoría por propiedades en Punta Gorda y Carrasco, con algunas de Parque Miramar y Malvín.

El grupo `C` –el más numeroso– tiene propiedades de todas las zonas, pero mayoritariamente de Parque Mirmar y Punta Gorda, Carrasco Norte y Carrasco en menor medida.

El grupo `D` contiene propiedades de todos las zonas pese a su pequeño tamaño.

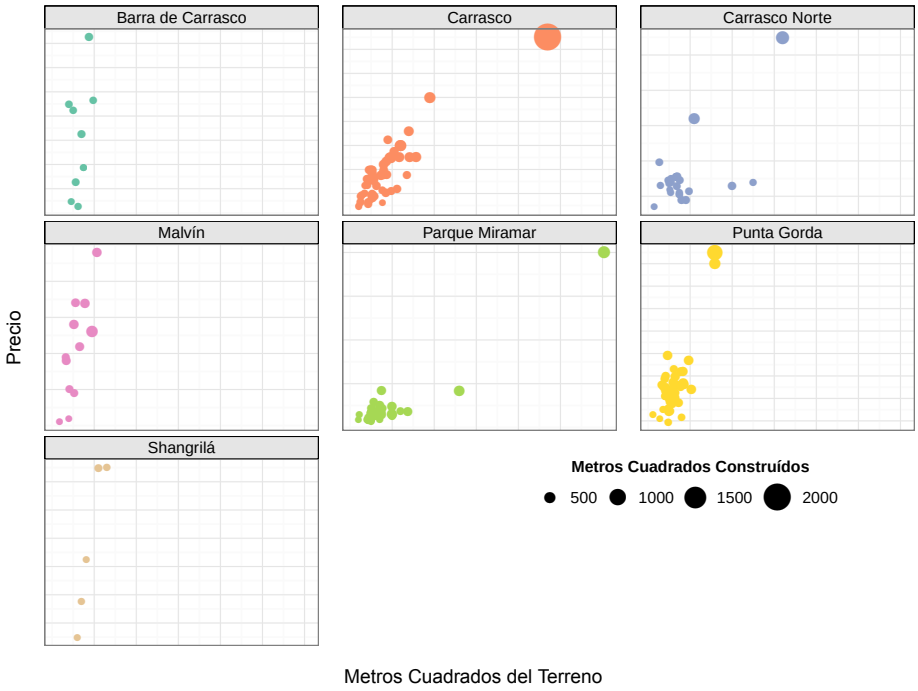


Figura 1.11: Metros cuadrados del terreno contra precio, Tamaño por metros cuadrados construidos. Escala de precio variable.

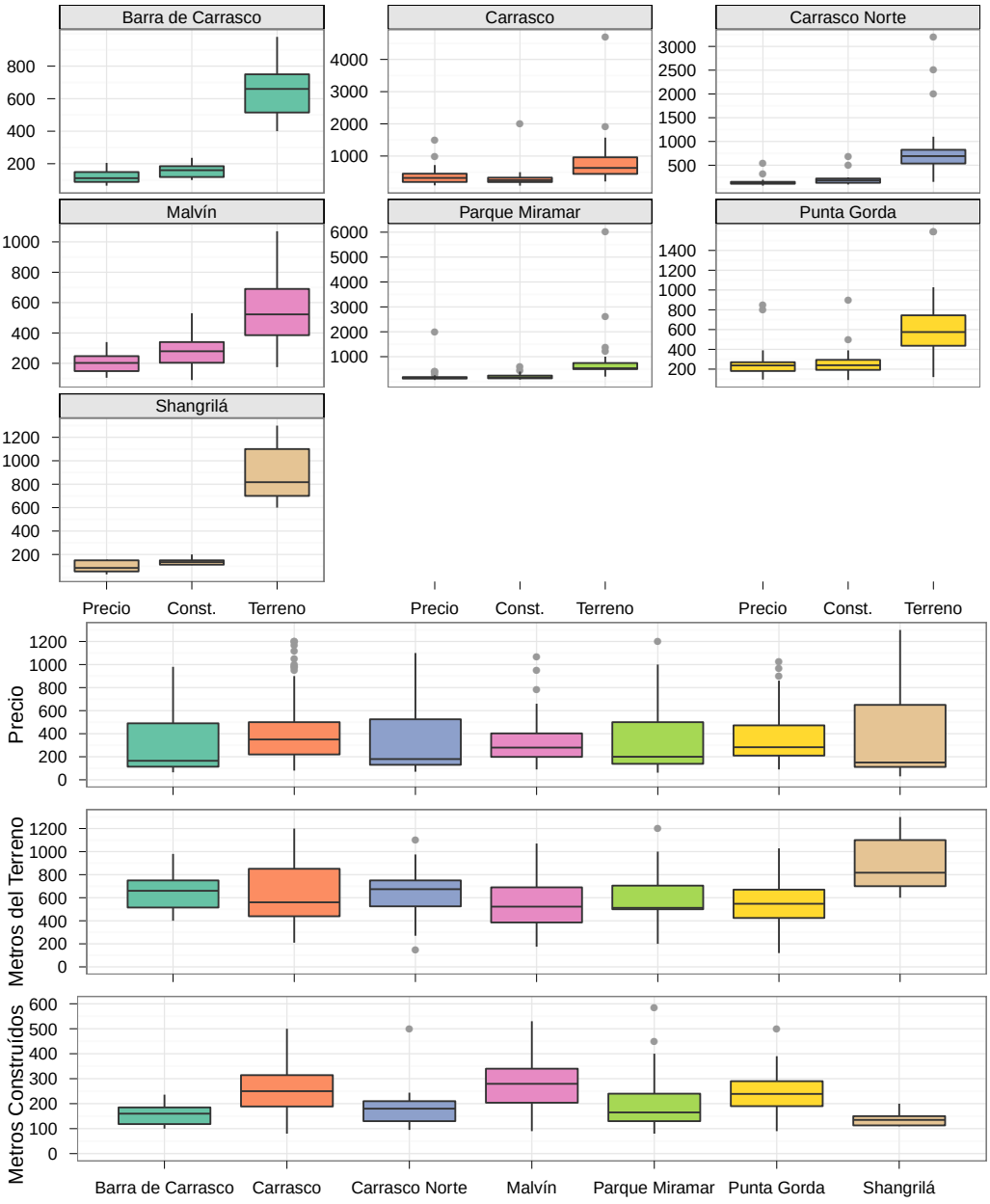


Figura 1.12: Diagrama de Cajas por Zona

Tanto Shangrilá como Carrasco Norte y Barra de Carrasco solamente tienen propiedades en los grupos `C` y `D` (sin tomar en cuenta los atípicos).

En cuanto a los datos atípicos, el grupo `R`, están integrados principalmente por propiedades en Carrasco, Parque Miramar y Carrasco Norte. Shangrilá y Barra de Carrasco no presentan atípicos detectados por TCLUST.

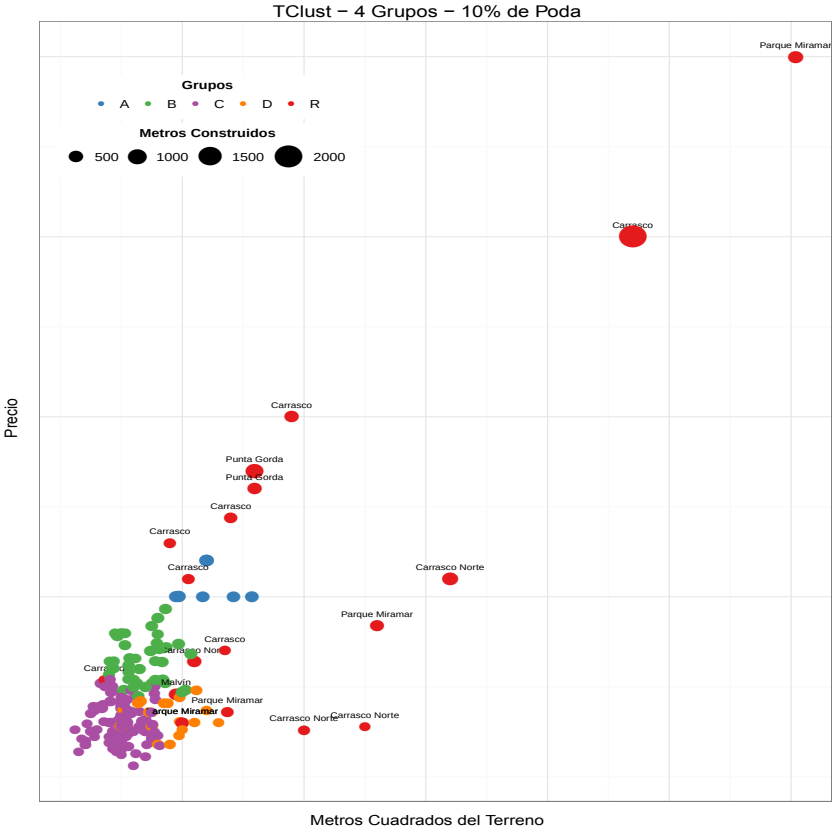


Figura 1.13: Grupos identificados por TCLUST

1.4.3. Algorítmo EMMIX

Se aplicó el algoritmo EMMIX al conjunto de datos para encontrar 4 grupos con una poda del 10 %.

Como se puede apreciar en la tabla (1.3), el grupo `C` identificado por EMMIX es, salvo por una observación, igual al grupo `A` identificado por TCLUST. Son las propiedades de más alto valor, con terrenos y edificaciones más grandes.

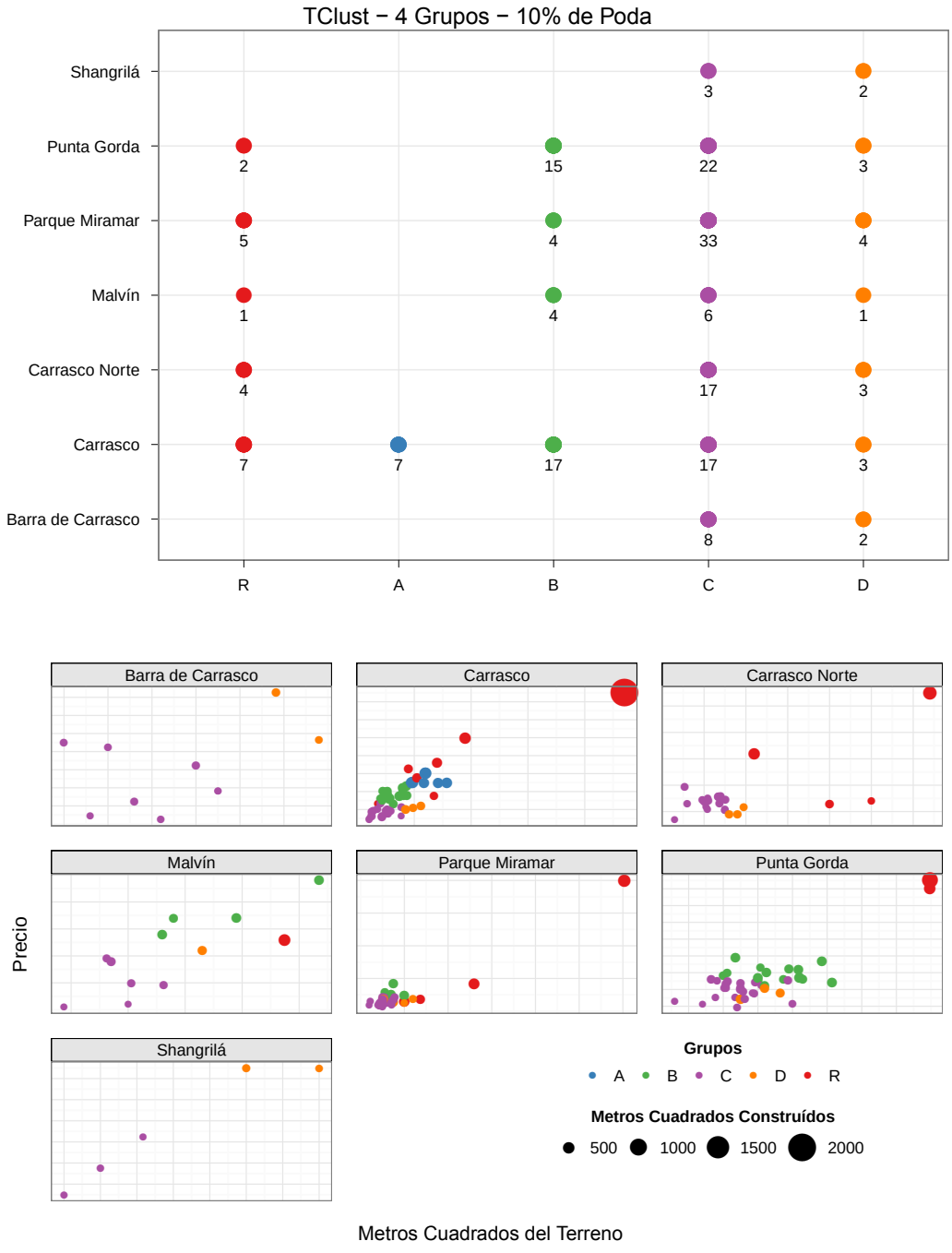


Figura 1.14: Grupos identificados por TCLUST - Zonas por Grupos

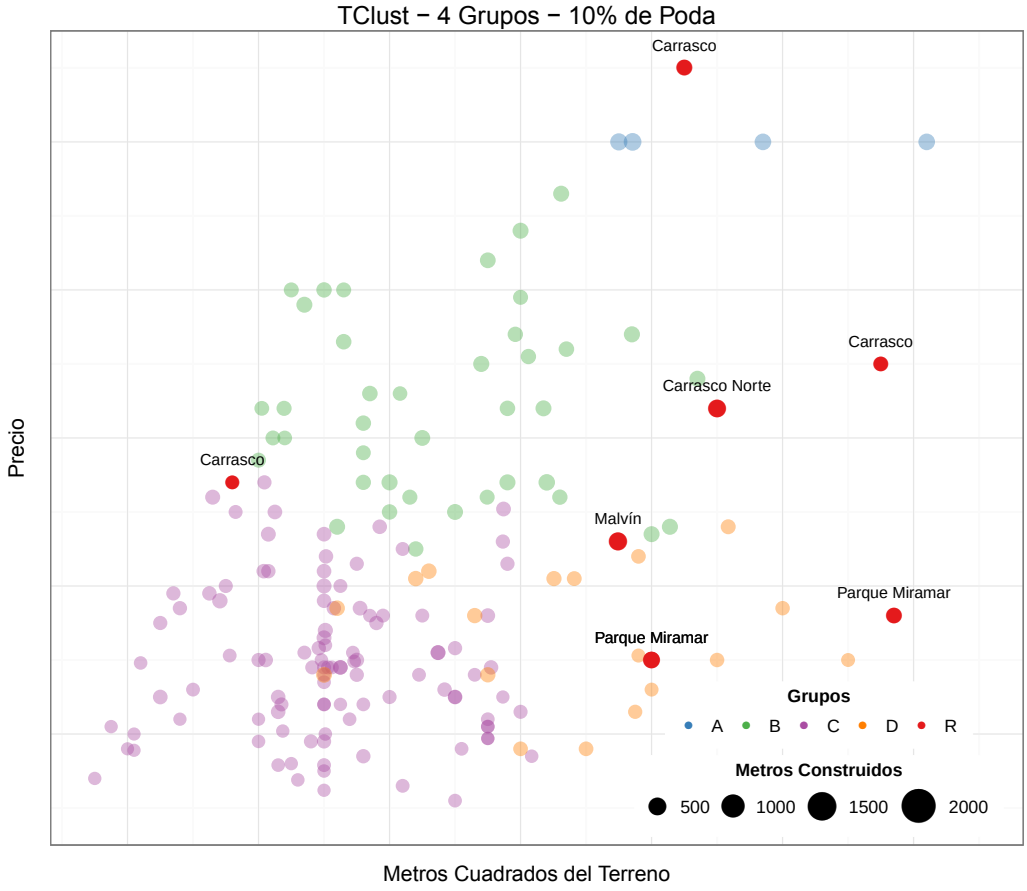


Figura 1.15: Datos atípicos "interiores" identificados por TCLUST

		Precio (Miles)		Metros		Metros Const.	
Zona	Obs.	Media	Desvío	Media	Desvío	Media	Desvío
A	52	292.35	78.14	679.69	195.69	293.25	42.04
B	66	170.82	45.66	573.48	201.28	193.38	28.72
C	6	533.33	51.64	1252.00	216.31	445.00	46.37
D	47	109.66	31.64	509.34	196.82	116.09	19.02
R	19	588.95	493.64	1974.47	1368.87	504.05	407.84

Tabla 1.3: Descriptiva de los grupos identificados por EMMIX

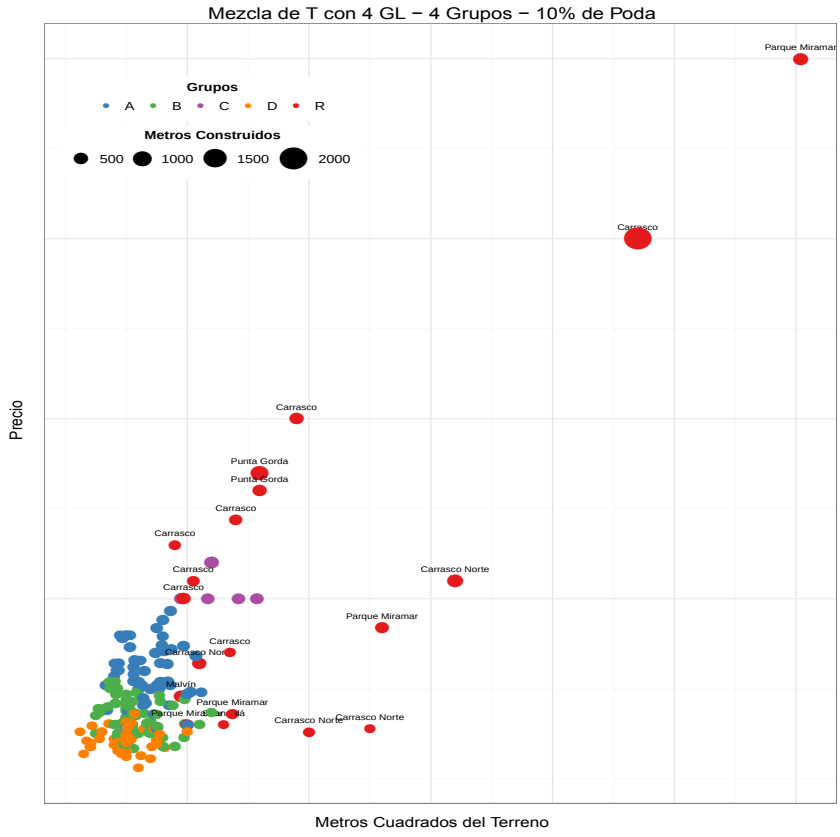


Figura 1.16: Grupos identificados por EMMIX

El grupo `A` es análogo al grupo `B` identificado por TCLUST: propiedades del 60 % del valor –en promedio– que las del grupo `A`, con terrenos de la mitad de tamaño pero la mitad del terreno se encuentra edificado.

El grupo `B` es el más numeroso, análogo al grupo `D` de TCLUST, con propiedades un 70 % más caras que el grupo `D` (del EMMIX), terrenos un poco más grandes pero el triple de metros cuadrados construidos. Presenta el valor del metro cuadrado construido (sin considerar el tamaño del terreno) más bajo de los grupos, aproximadamente USD 880.

El grupo `D` se trata de las propiedades con menor valor, tamaño del terreno y metros construidos. Tiene el mismo valor del metro cuadrado construido (sin considerar el tamaño del terreno) que el grupo `A`, aproximadamente USD 1000.

En cuanto a la composición de zonas de los grupos (figura (1.17)), si bien las proporciones varían, la composición es la misma que la de los grupos identificados por TCLUST.

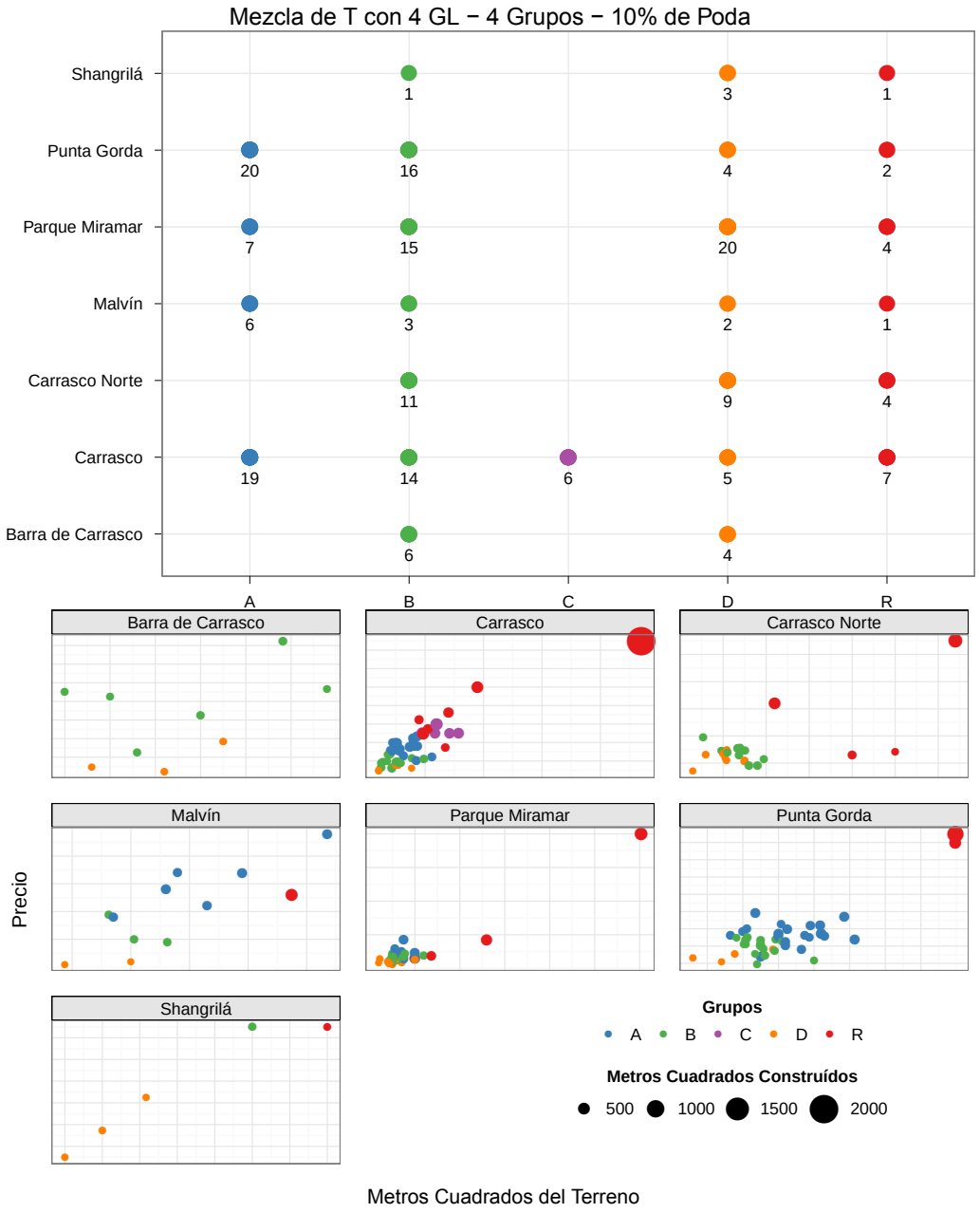


Figura 1.17: Grupos identificados por EMMIX - Zonas por Grupos

Lo mismo es para los datos atípicos, únicamente introduciendo a Shangrilá con una observación y detectando otra observación de Carrasco.

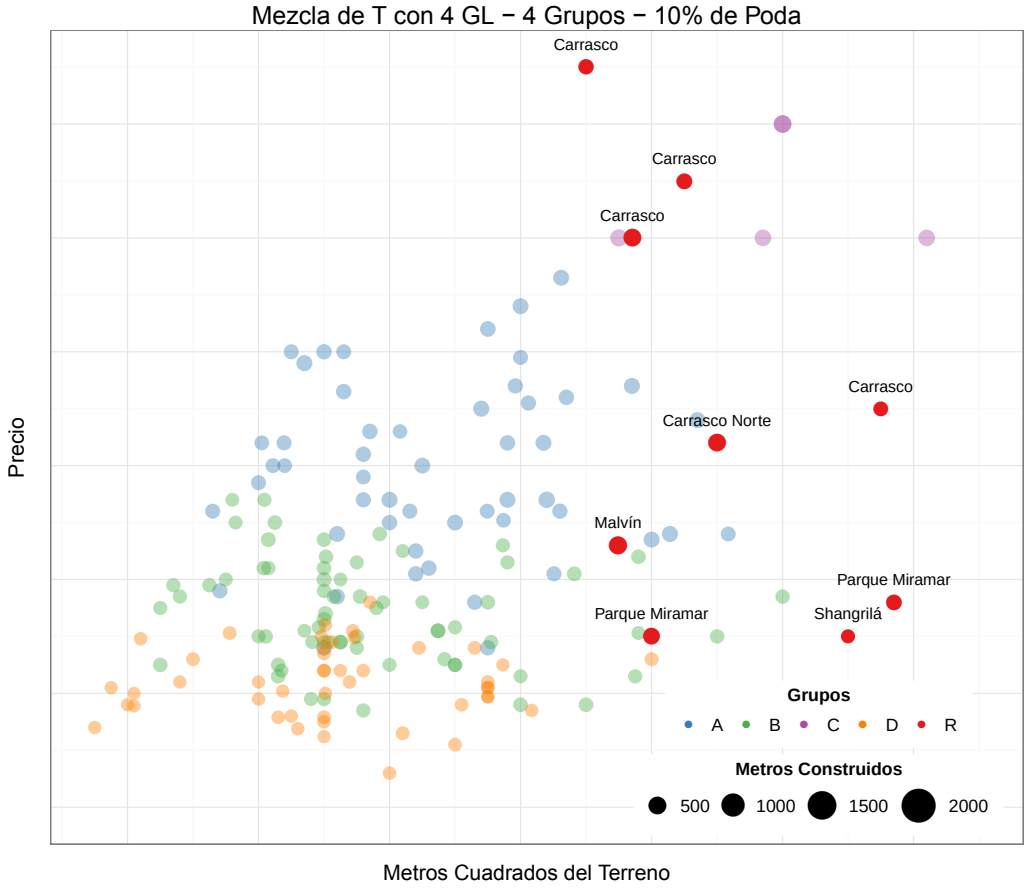


Figura 1.18: Datos atípicos "interiores" identificados por EMMIX

1.4.4. Algorítmo de K-Means Robusto

Se aplicó el algoritmo K-Means Robusto al conjunto de datos para encontrar 4 grupos con una poda del 10 %.

Los grupos identificados por K-Means Robusto son los más esféricos en comparación con TCLUST y EMMIX.

El grupo `A` es análogo al grupo `A` identificado por TCLUST y `B` por EMMIX, las propiedades de mayor valor. Sin embargo, este engloba la mayoría de los datos declarados como atípicos por los otros algoritmos.

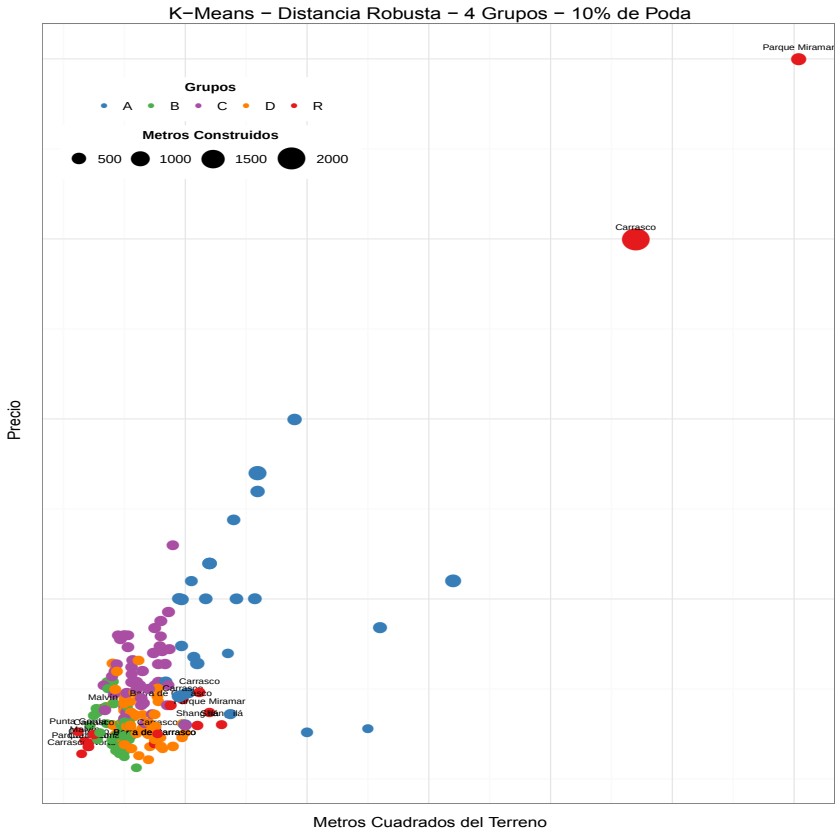


Figura 1.19: Grupos identificados por K-Means Robusto

		Precio (Miles)		Metros		Metros Const.	
Zona	Obs.	Media	Desvío	Media	Desvío	Media	Desvío
A	26	444.04	229.20	1422.96	586.35	423.69	147.95
B	44	139.45	52.33	446.05	93.19	134.89	27.30
C	49	286.43	101.75	637.73	161.91	287.24	35.67
D	52	162.13	63.17	635.73	153.39	184.71	33.15
R	19	309.11	516.87	1117.37	1567.70	267.58	435.38

Tabla 1.4: Descriptiva de los grupos identificados por K-Means Robusto

Los grupos `B`, el de segundo menor precio promedio, presenta la misma relación con el grupo `D` (propiedades de menor valor), precio levemente mayor pero sustancial incremento en el tamaño del terreno y sus metros construidos.

El grupo `C` se trata de propiedades con un precio 70 % mayor que las del grupo `D`, mismo tamaño del terreno pero con un 50 % más de metros construidos.

En cuanto a la composición de zonas de los grupos, la estructura es más difusa.

A excepción de Shangrilá y Barra de Carrasco, todos los grupos tienen propiedades de todas las zonas, inclusive los detectados como atípicos.

1.4.5. Conclusiones

Tanto TCLUST como EMMIX detectan grupos muy similares, tanto en composición de zonas como características de las propiedades. Esta estructura es detectada por K-Means Robusto pero de forma más difusa, lo que hace suponer que dicha estructura es razonable para el problema.

Los grupos identificados por TCLUST son los más elípticos y de distinto tamaño, mientras que los de K-Means Robusto son los más esféricos y de tamaño similar.

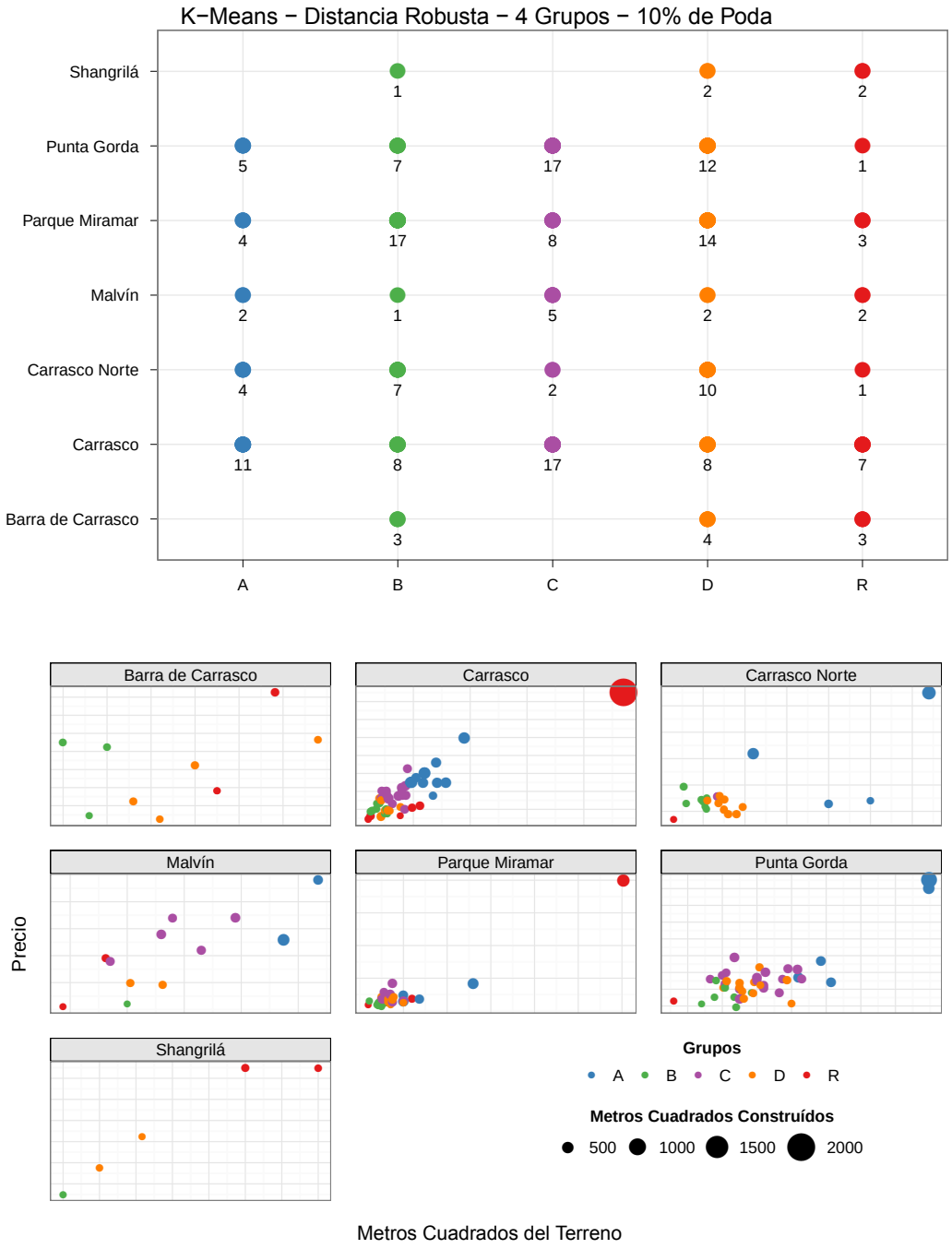


Figura 1.20: Grupos identificados por K-Means Robusto - Zonas por Grupos

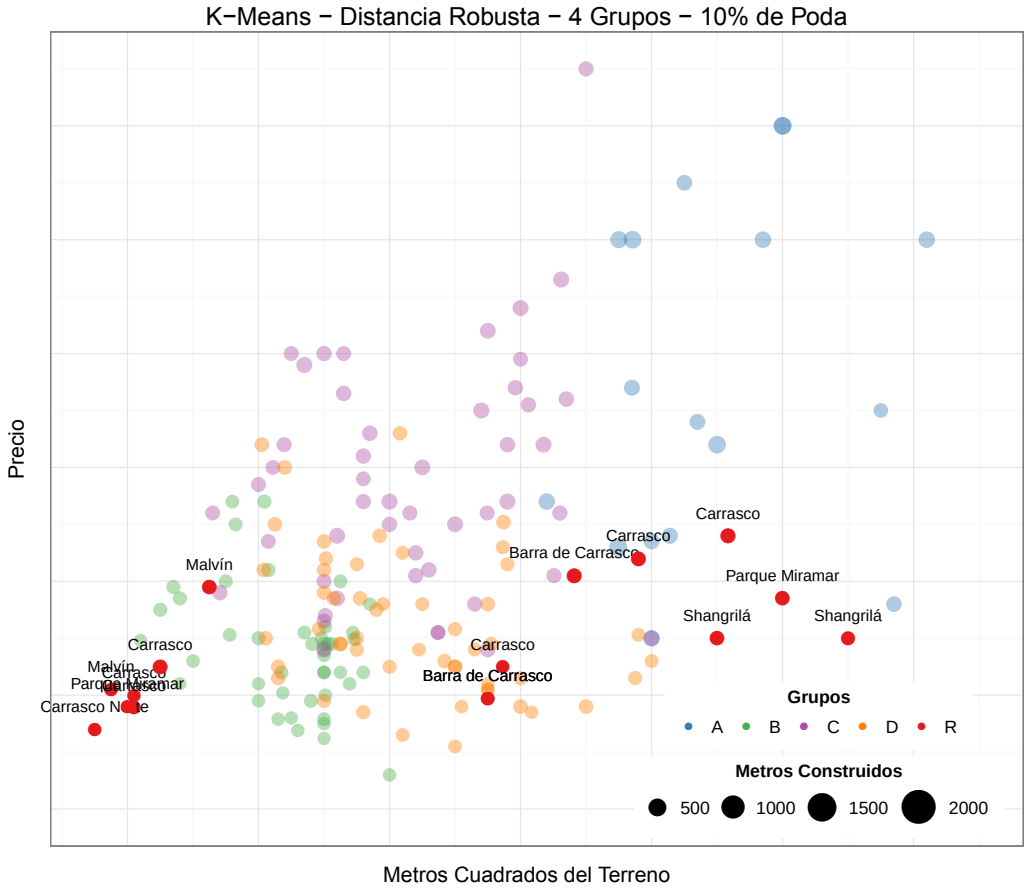


Figura 1.21: Datos atípicos "interiores" identificados por K-Means Robusto

Esto se debe al alto valor utilizado para la restricción del cociente de los valores propios de las matrices de varianzas (en este caso 90). Dicho parámetro permite identificar grupos de estructura más heterogénea entre sí que el algoritmo EMMIX.

Consultando con la gerencia de la inmobiliaria, ella confirma la estructura detectada (características y relaciones entre grupos) a partir de su conocimiento del campo.

La gerencia de la inmobiliaria se notó sorprendida por los grupos `B` y `D` identificados por TCLUST, ya que ella capta un conjunto de propiedades que consideran muy interesantes.

Los outliers detectados por TCLUST y EMMIX son muy similares, mientras que K-Means Robusto detecta en las "colas" de los datos porque dedica un grupo a la mayoría de los detectados por las otras técnicas.

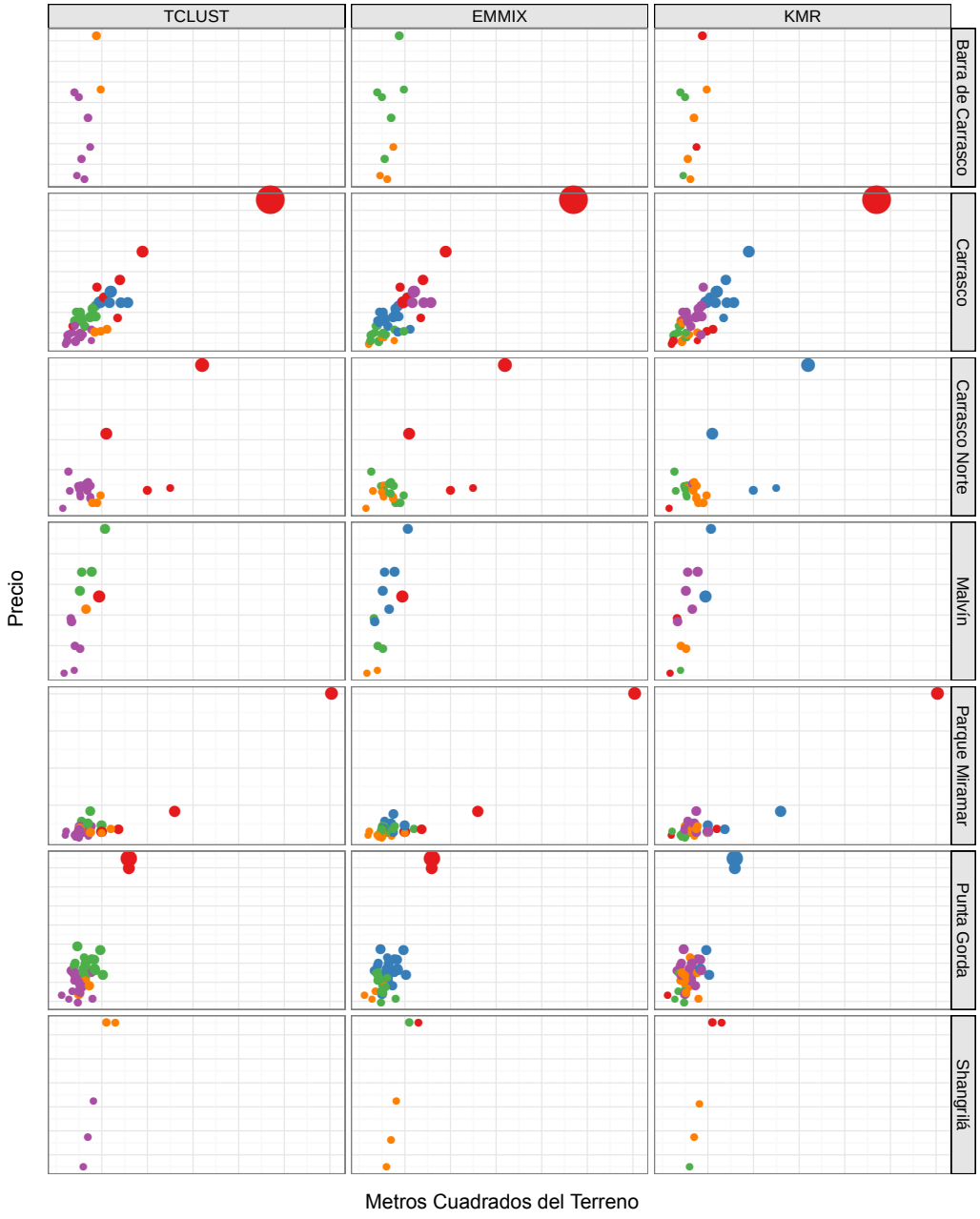


Figura 1.22: Comparación de Grupos por Zonas - Colores no alineados excepto atípicos

Dicho grupo, "atípicos exteriores", se trata de propiedades que por su ubicación, calidad de la construcción y tamaño, su precio es fijado más arbitrariamente ya que son

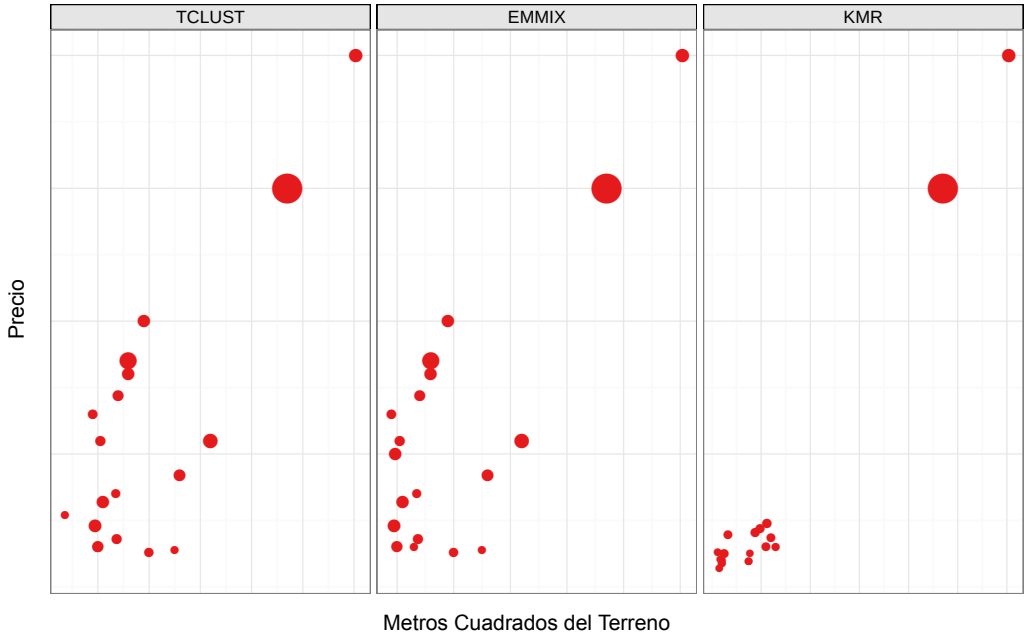


Figura 1.23: Comparación de Atípicos por Técnica

"menos comparables" con las restantes.

Por lo tanto, es de mayor interés estudiar los atípicos "interiores", es decir, aquellos que no surgen de la inspección visual ya que se encuentran con más "profundidad" en la nube de puntos.

En la figura (1.24) se visualizan los atípicos "interiores" detectados por al menos dos técnicas, donde se pueden apreciar 3 grupos dentro de ellos.

El primer grupo, integrado por las propiedades 116, 119, 390 y 291, se tratan de propiedades de muy bajo valor para el tamaño de su terreno. Este puede derivarse de la inspección visual y se tratan de propiedades con ubicaciones poco deseadas y construcciones que necesitan importantes reparaciones.

Un segundo grupo está conformado por las observaciones 405 y 105 en la zona de Carrasco (las observaciones 109, 268, 150 y 287 se consideran "exteriores"). Dichas observaciones son un "punto medio" entre las "exteriores" y las "comparables".

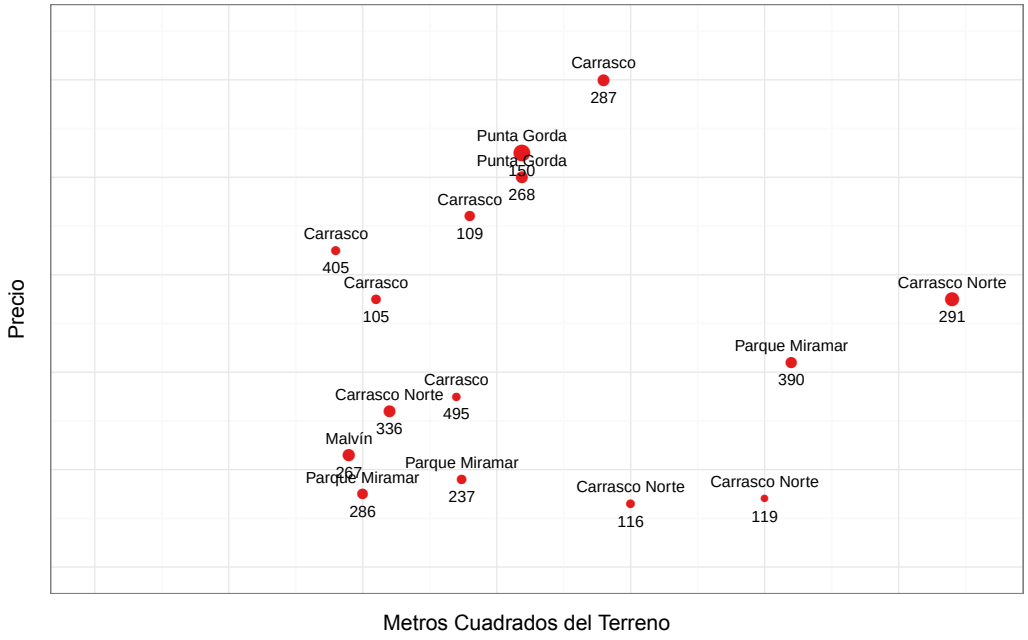


Figura 1.24: Atípicos "interiores" detectados por al menos dos técnicas

El tercer grupo es el de mayor interés, ya que son los atípicos con mayor profundidad en la nube.

Con la excepción de la propiedad 237 de Parque Miramar, todas ellas han sido señaladas como excelentes ofertas. En especial, la propiedad 267 en Malvín, lamentablemente ya vendida.

ID	Zona	Precio (Miles)	Metros Const.	Metros del Terreno
105	Carrasco	550	319	1050
109	Carrasco	720	380	1400
110	Carrasco	1500	2000	4700
287	Carrasco	1000	500	1899
405	Carrasco	650	270	900
495	Carrasco	350	221	1350
116	Carrasco Norte	130	244	2000
119	Carrasco Norte	140	140	2500
291	Carrasco Norte	550	680	3200
336	Carrasco Norte	320	500	1100
267	Malvín	230	530	949
237	Parque Miramar	180	310	1370
279	Parque Miramar	2000	583	6038
286	Parque Miramar	150	400	1000
390	Parque Miramar	420	450	2600
150	Punta Gorda	850	900	1594
268	Punta Gorda	800	500	1594

Tabla 1.5: Detalle de atípicos detectados por al menos dos técnicas

1.5. Comentarios finales.

En el presente trabajo se intentó introducir el concepto de robustez en Clustering a partir de diferentes metodologías.

A partir de la elección de un modelo, se define para el problema lo que se considera "típico" y "atípico" a través de una forma de captarlo.

Como no hay un algoritmo mejor que otro per-se, lo mismo sucede con la robustez: los algortimos son robustos en cierto sentido bajo determinado modelo. Esto se intentó demostrar mediante los estudios de simulación.

A partir de técnicas robustas se puede detectar mejor lo "típico", como se mostró en el capítulo de aplicación a datos reales. A su vez, los datos "atípicos" tienen un interés en sí mismos, como la identificación de propiedades que constituyen buenas ofertas.

El tamaño de los conjuntos de datos ha crecido a órdenes donde el suponer que fueron generados por un único proceso –y que a su vez es conocido perfectamente– es, cuando menos, extremadamente poco realista.

La robustez es entonces, prácticamente, una necesidad en la actualidad.

Una posible investigación a futuro –que no fue tratada en el trabajo monográfico por motivos de tiempo y extensión– es la de obtener clusters robustos mediante el uso de cópulas.

Los métodos desarrollados en este trabajo consideran la matriz de dispersión como la "fuerza motriz" del análisis. Entonces, se asume que toda la información acerca de la dependencia entre los componentes del vector aleatorio está contenida en la matriz de covarianzas.

(Jajuga, 2005) propone un enfoque alternativo a los métodos clásicos. En vez de analizar conjuntamente los parámetros de escala y dependencia, dados en la matriz de varianzas y covarianzas, el análisis se realiza separadamente para los parámetros de escala (a través del análisis univariado), y para los parámetros de dependencia.

La importancia del análisis de cópulas es la de permitir levantar el supuesto de distribuciones elípticas, supuesto necesario de los métodos anteriores.

El objetivo consistiría en construir métodos robustos de clustering, modelando a través de cópulas y siendo sensible a fenómenos de contaminación. Se podría a través del trabajo de (Mendes et al., 2007), crear un nuevo algoritmo de Clustering basado en cópulas que sea estable frente a perturbaciones.

Bibliografía

Day, N. (1969). Estimating the components of a mixture of two normal distributions. Biometrika, 56(3):463–474.

Flury, B. (1997). A First Course in Multivariate Statistics. Springer.

Gallegos, M. T. (2002). A survey of sampling from contaminated distributions. En Classification, Clustering and Data Analysis: Recent advances and application, pp. 247–255. K. Jajuga, A.Sokolowski, and H.H. Bock eds.

Gordaliza, A., Matrán, C., Cuesta-Albertos, J. A., y Mayo-Iscar, A. (2010). Exploring the number of groups in robust model-based clustering. Statistics and Computing, 20:1–29.

Hart, P. E., Stork, D. G., y Duda, R. O. (2001). Pattern classification. Wiley.

Hartigan, J. (1978). Asymptotic distributions for clustering criteria. Annals of Statistics, 6:117–131.

Jajuga, K. (2005). Model-based clustering: Discussion on some approaches. En Data Analysis and Decision Support, pp. 73–81. Springer.

Krishnan, T. y McLachlan, G. J. (1997). The EM Algorithm and Extensions. John Wiley & Sons.

Matrán, C., Mayo-Iscar, A., y Cuesta-Albertos, J. A. (2008). Trimming and likelihood: Robust location and dispersion estiamtion in the elliptical model. The Annals of Statistics, 36(5):2284–2318.

McQueen (1967). Soma methods for classification and analysis of multivariate observation. Computer and Chemitry, 4:257–272.

Melnykov, V. (2010). Finite mixture models and model-based clustering. Statistics Surveys, 4:80–116.

Mendes, B. V. M., Melo, E. F. L. D., y Nelsen, R. B. (2007). Models, copulas and applications. Communications in Statistics Simulation and Computation, 36:997–1017.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. Philosophical Transactions of the Royal Society, 185:71–110.

Peel, D. y McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing, 10:339–348.

Pollard, D. (1981). Strong consistency of k-means clustering. Annals of Probability, 9(1):135–140.

Pollard, D. (1982). A central limit theorem for k-means clustering. Annals of Probability, 10(4):919–926.

Ritter, G. y Gallegos, M. T. (2005). A robust method for cluster analysis. Annals of Statistics, 33:347–380.

Ritter, G. y Gallegos, M. T. (2009). Trimmed ml estimation of contaminated mixtures. The Indian Journal of Statistics, 71-A(2):164–220.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5:329–350.

Capítulo 2

Tratamiento de la no respuesta en encuestas de panel en el caso de poblaciones finitas, por Ana Coimbra

Resumen[1]

Este trabajo pretende mostrar cómo tratar la no respuesta en un caso particular de las encuestas por muestreo como son las encuestas de panel. Estas refieren a estudios basados en observaciones repetidas efectuadas sobre las mismas unidades de muestreo: personas, hogares, empresas, etc. La muestra es extraída por única vez al inicio del estudio y todas las unidades seleccionadas serán visitadas a lo largo de la duración del panel.

Los distintos momentos del tiempo en los que las encuestas son llevadas a cabo se denominan "olas"; la duración del panel y el período entre olas son definidos en la etapa del diseño de la encuesta.

La medición periódica de elementos permite realizar un seguimiento de la población objetivo, logrando captar su dinámica en el tiempo. Los resultados particulares en cada instancia de medición (estimaciones transversales) pueden ser obtenidos sin perjuicio de lo anterior y, aunque no sea el objetivo principal de las encuestas de panel, suelen ser de interés en sí mismos.

Un problema usual en las encuestas de panel es la mortalidad de unidades a lo largo del tiempo. En términos generales, esta dificultad puede pensarse como un problema de no respuesta, definida como la imposibilidad de obtener toda o alguna información para una o más de las unidades seleccionadas en la muestra. Este es un fenómeno presente en la mayoría de las encuestas por muestreo y es imprescindible su tratamiento para evitar sesgos en las estimaciones. La inclusión del factor "tiempo" en las encuestas de panel provoca un agravamiento del problema de no respuesta respecto a las encuestas cross-section, reflejado en reducciones considerables en el "tamaño de muestra" período

1. Resumen del trabajo de pasantía realizado en conjunto con Margarita Antía, con la tutoría de Juan José Goyeneche, Guillermo Zoppolo, para la obtención del grado de la Licenciatura en Estadística.

a período debido a la movilidad, el fallecimiento y otros factores (como la pérdida de cooperación de unidades) que resultan en el "agotamiento" del panel. Otro efecto causado por la inclusión del factor "tiempo" es la potencial pérdida de representatividad de la muestra para inferir resultados transversales en olas posteriores a la primera.

2.1. Estimador de cambio bajo condiciones ideales

Las condiciones ideales en cualquier tipo de encuesta por muestreo están regidas por la obtención de respuesta completa de las unidades muestreadas a partir de un marco muestral perfecto acorde a la población objetivo, sin presencia de errores de medición. En las encuestas por panel, dichas condiciones requieren el supuesto adicional de que la población objetivo sea fija en el tiempo. Adicionalmente se requiere obtener respuesta completa de todas las unidades en todas las olas. Esto es, s es una muestra aleatoria de la población U de N individuos, tomada bajo un diseño $p(s)$[2] de tamaño n_s el cual genera probabilidades de inclusión π_k.

Bajo estos supuestos, el estimador de cambios de la variable de interés es :

$$\begin{aligned}\hat{A}_{j,j+h} &= \hat{(t_{j+h} - t_j)} \\ &= \sum_s \left(\frac{y_{(j+h)\,k} - y_{j\,k}}{\pi_k} \right) \\ &= \sum_s \frac{a_{(j,j+h)\,k}}{\pi_k} \end{aligned} \tag{2.1}$$

donde $a_{(j,j+h)\,k}$ es cambio individual del elemento k en la variable de interés entre las olas j y $j+h$.

Los estimadores $\hat{A}_{j,j+h}$ tienen la forma de estimadores π (Horvitz-Thompson), por lo tanto, también comparten sus propiedades. La varianza de $\hat{A}_{j,j+h}$ también puede expresarse como

$$Var_{p(s)}\left(\hat{A}_{j,j+h}\right) = Var(\hat{t}_j) + Var(\hat{t}_{j+h}) - 2Cov(\hat{t}_j, \hat{t}_{j+h}) \tag{2.2}$$

y dado que si estos totales están calculados a partir de las mismas unidades, se espera que los totales estimados en olas sucesivas estén correlacionados positivamente resultando en una varianza pequeña del estimador de diferencias.

2. La elección del diseño a utilizar depende del objeto de estudio y no de la utilización de paneles.

Una alternativa a la utilización de paneles es la estimación de diferencias de la variable de interés utilizando totales estimados mediante encuestas cross-section en los momentos j y $j+h$. En este caso el estimador de diferencias también será insesgado pero su varianza será mayor que en el caso anterior, ya que la independencia entre muestras en una y otra instancia determina que el tercer término de (2.2) sea cero.

Este es un aspecto importante en la justificación de utilización de paneles frente a encuestas cross-section para medir cambios.

2.2. Calibración como tratamiento de la no respuesta

El estimador propuesto tiene la cualidad de ser insesgado y de sencillo cálculo, ya que es un estimador π. Para su desarrollo, se partió de supuestos muy restrictivos rara vez presentes en la práctica, a saber: la existencia de respuesta perfecta y población fija en el tiempo, reflejados en ponderadores constantes en el tiempo para cada elemento e iguales al inverso de su probabilidad de inclusión en la muestra.

El levantamiento del supuesto de respuesta perfecta plantea la necesidad de estudiar cus son los efectos que provoca la imposibilidad de obtener toda o alguna información para todas o algunas unidades seleccionadas en la muestra.

Si la no respuesta se presentara de manera completamente aleatoria, el único inconveniente al que se enfrenta el investigador es la reducción del tamaño de muestra y un consiguiente aumento en la varianza de las estimaciones, que podría ser fácilmente contrarrestado mediante un "sobremuestreo" (fijando un tamaño de muestra mayor en la etapa de diseño). De esta manera el único efecto negativo de la no respuesta sería el incremento en la carga administrativa y los costos de recolección de datos. En la práctica, la situación anterior sería una "feliz" casualidad. Las unidades que no contestan "normalmente" difieren en algunas características de aquellas que sí lo hacen, y el sesgo introducido en las estimaciones por esta causa constituye el obstáculo más importante por corregir. Frente a la pérdida del insesgamiento de los estimadores, el incremento de su varianza es un disturbio menor: en presencia de sesgo significativo, un intervalo de confianza calculado estará centrado en un valor erróneo y no se logra el nivel de confianza requerido.

Sea $\hat{t}_y$ el estimador de t_y cuando hay respuesta completa, o sea cuando el conjunto de respondentes r coincide con la muestra s; sea $\hat{t}_{y_{NR}}$ el estimador de t_y en presencia de no respuesta.

El error de estimación de $\hat{t}_{y_{NR}}$ puede expresarse mediante un término que representa el error muestral y otro que representa el error por no respuesta.

$$Error = \hat{t}_{y_{NR}} - t_y = (\hat{t}_y - t_y) + (\hat{t}_{y_{NR}} - \hat{t}_y) \tag{2.3}$$

Siendo

- $\hat{t}_y - t_y$ el error muestral (el error que surge por elegir y observar una muestra, en vez de observar toda la población).
- $\hat{t}_{y_{NR}} - \hat{t}_y$ el error por no respuesta (error que surge por la no existencia de respuesta completa).

El sesgo total de $\hat{t}_{y_{NR}}$ se obtiene calculando el valor esperado bajo los mecanismos de selección y de respuesta de los dos componentes de error previamente definidos:

$$\begin{aligned} B_{pq}(\hat{t}_{y_{NR}}) &= B_{SAM} + B_{NR} \\ &= E(\hat{t}_{y_{NR}} - \hat{t}_y) + E(\hat{t}_y - t_y) \end{aligned} \tag{2.4}$$

El término B_{SAM} (sesgo muestral) es cero o irrelevante para la mayoría de los propósitos prácticos, por lo tanto el sesgo de $\hat{t}_{y_{NR}}$ se convierte casi enteramente en el sesgo por no respuesta. Esto evidencia la necesidad de realizar algún tipo de tratamiento en la etapa de estimación.

En general se distinguen dos tipos de no respuesta: no respuesta al ítem, que refiere a faltantes en la respuesta para un ítem en particular del formulario debido a omisión (tanto del entrevistador como del entrevistado) o negativa del encuestado a contestar; y no repuesta de la unidad, que se da cuando la unidad seleccionada para ser entrevistada no es encontrada o se rehúsa a participar en la encuesta.

Las técnicas dominantes en la literatura actual para el tratamiento de la no respuesta son la calibración y la imputación (Särndal y Lundström, 2005). Usualmente la calibración predomina en el tratamiento para el caso de no respuesta de unidades; mientras que la imputación es más extensamente aplicada en los problemas de no respuesta de ítems. La primera es una estrategia global, tratando todas las variables de forma simultánea, mientras que la segunda es particular, específica de cada variable. La decisión acerca de utilizar una u otra no es obvia, y depende de distintos factores como lo son: la cantidad

y el número de olas, el tipo de análisis a llevarse a cabo, la disponibilidad de variables auxiliares con poder predictivo de los valores faltantes y el costo de implementar los procedimientos.

No obstante lo anterior, se han ensayado soluciones que aplican ambas técnicas en conjunto (Deville y Särndal, 1994).

La imputación es el procedimiento a través del cual los valores faltantes en una o más variables de estudio se completan con sustitutos. Los valores perdidos en la base de datos se reemplazan por los valores "plausibles" dando como resultado una matriz completa de valores. Existen varios métodos de imputación que básicamente difieren en como definen "plausible", pero la mayoría coinciden en la necesidad de utilización de información auxiliar.

La calibración, o más precisamente el uso de estimadores calibrados, se basa fuertemente en el uso de información auxiliar tanto a nivel poblacional como a nivel de la muestra original. Su creciente popularidad puede explicarse porque no se basa en la especificación de un modelo de no respuesta, brinda un enfoque unificado dentro de la teoría del muestreo de poblaciones finitas, es computacionalmente sencilla de implementar y generaliza otras técnicas del tratamiento de la no respuesta como la post-estratificación, el raking y algunos casos de los ajustes basados en la teoría del muestreo en dos fases. El enfoque de calibración abarca a una familia de estimadores $\hat{t}_{y_W}$ cuyos miembros corresponden a diferentes inputs de información.

Es deseable que los estimadores afectados por la no respuesta sean útiles para estimar los totales de las variables de interés, que no estén sesgados y que tengan varianza reducida. El sistema de ponderadores que surge del enfoque de calibración verifica también, que al ser aplicado a las variables auxiliares reproduce el input de información auxiliar. La idea central de los estimadores calibrados es sencilla, consiste en modificar los ponderadores originales de la muestra minimizando alguna función de distancia entre dichos ponderadores y los ponderadores finales (o calibrados) y de manera que estos últimos estimen sin error algunas cantidades conocidas con los datos de los respondentes.

Sea w_k el peso calibrado para $k \,\epsilon\, r$, luego, el estimador de $t_y = \sum_U y_k$ es:

$$\hat{t}_{y_W} = \sum_r w_k \, y_k \tag{2.5}$$

Se busca el conjunto de valores w_k para todo $k \epsilon r$ que satisfaga la ecuación de calibración:

$$\mathbf{X} = \sum_r w_k \, \mathbf{x}_k \tag{2.6}$$

Se dice que estos pesos w_k están calibrados al input de información $\mathbf{X}$, ya que cuando se aplican al vector auxiliar $\mathbf{x}_k$ reproducen exactamente la información dada en $\mathbf{X}$.

Como resultado de la selección de la muestra, a cada elemento k le corresponde el peso $d_k = \frac{1}{\pi_k}$. En presencia de no respuesta, $\sum_r d_k \, y_k$ subestima $\sum_U y_k$, en una magnitud $\sum_{s-r} d_k \, y_k$ (en el caso que la variable de interés tome solamente valores positivos). Es por esto que los d_k deben ser modificados. Se buscarán nuevos pesos que sean mayores que d_k al menos para la mayoría de los respondentes, de manera de compensar la pérdida de unidades. Los nuevos ponderadores $w_k = d_k \nu_k$ se obtienen "aumentando" los pesos originales mediante el factor ν_k, que reflejará las características individuales conocidas de los elementos $k \epsilon r$ (resumidas en el vector $\mathbf{x}_k$), y puede pensarse como una función del vector auxiliar $\nu_k = F(\lambda' \mathbf{x}_k)$, donde λ es un vector de la misma dimensión que $\mathbf{x}_k$ y se determinará para que se verifique la ecuación de calibración.

La clave para una calibración exitosa es el uso de información auxiliar poderosa; permitiendo reducir tanto sesgo como la varianza. La efectividad del estimador de calibración para controlar el sesgo ocasionado por la no respuesta dependerá de propiedades del vector auxiliar. (Särndal y Lundström, 2005) realizaron un estudio de Simulación Monte Carlo a través del cual se obtiene evidencia empírica de la fuerte relación existente entre el sesgo del estimador de calibración que proviene de la no respuesta y la información auxiliar utilizada para calibrar.Este será menor cuanto más estrecha sea la relación entre la información auxiliar y la probabilidad de respuesta o la variable de interés. Si el sesgo es modesto el intervalo de confianza será válido y la probabilidad de cobertura será cercana al nivel de confianza requerido. Limitar el sesgo de las estimaciones en presencia de no respuesta se tornará en la mayor precupación, la minimización de la varianza pasará a segundo plano ya que de nada sirve que un estimador presente varianza chica cuando está fuertemente sesgado.

2.3. Estimadores calibrados en encuestas de panel

La no respuesta de unidades en una ola es una forma de no respuesta parcial particular al muestreo por paneles, generando distintos patrones de respuesta a lo largo del estudio, como muestra la siguiente tabla. Algunos miembros de la muestra pueden abandonar la encuesta en cierta ola y perderse para el resto del estudio (desertores); mientras que otros pueden perderse en una ola, y volver al panel en alguna de las siguientes (respondentes episódicos).

Patrón	Estado de respuesta	Ola 1	Ola 2	Ola 3	Ola 4	Ola 5
1	Respondentes	x	x	x	x	x
2	No	x	x	x	x	-
3	Respondentes	x	x	x	-	-
4	por	x	x	-	-	-
5	Desgaste	x	-	-	-	-
6	No	x	x	-	x	x
7	Respondentes	x	-	-	x	x
8	Episódicos	x	-	-	-	x

Ref: x: respuesta, -: no respuesta

La no respuesta en panel (manifestada bajo los patrones de no respuesta por desgaste y no respuesta episódica) genera en cada ola, un conjunto de respondentes r_i, todos incluidos en la muestra s. Esto requerirá el cálculo de ponderadores calibrados particulares a cada individuo respondente en cada ola.

2.3.1. Estimación transversal

Para la estimación transversal se calcularán los ponderadores de las unidades respondentes en cada ola. El estimador calibrado del total correspondiente a la ola i, $\hat{t}_{y_{Wi}}$ se define por:

$$\hat{t}_{y_{Wi}} = \sum_{r_i} w_{k_i}\, y_{k_i} \tag{2.7}$$

siendo w_{k_i} los ponderadores calibrados del elemento k respondente en la ola i, y y_{k_i} el valor de la variable de interés para este individuo en dicha ola.

2.3.2. Estimación longitudinal

Adicionalmente, para las estimaciones longitudinales será necesario el cálculo de un nuevo conjunto de ponderadores aplicables únicamente a las unidades respondentes en todas las instancias de las que se quiere medir el cambio: respondentes simultáneos.

Cuando el patrón de respuesta admite únicamente no respuesta por desgaste, los respondentes de i-ésima ola también fueron respondentes en las olas anteriores $i-1$, $i-2$, ...,1, por lo tanto, los cambios solamente podrán ser medidos para las unidades respondentes en la ola más reciente. En la siguiente figura puede verse la representación gráfica de un estudio de panel de tres olas con patrón de respuesta por desgaste. A modo de ejemplo, si el interés radica en la estimación de cambios de una variable entre las olas 1 y 3, se calibrarán los cambios individuales para cada respondente de la ola 3.

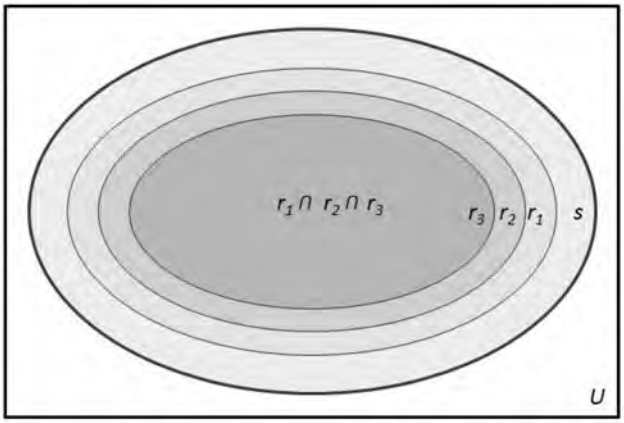


Figura 2.1: Conjunto de respondentes con patrón de respuesta: desgaste

Frente a un patrón de respuesta episódica, los cambios entre dos olas también serán medidos en los respondentes simultáneos, pero en este caso, este conjunto no necesariamente coincide con la ola más reciente de las sujetas a medición. Siguiendo el mismo ejemplo, la estimación de cambios entre la primera y la tercera ola se realizará en base al conjunto marcado en la figura (2).

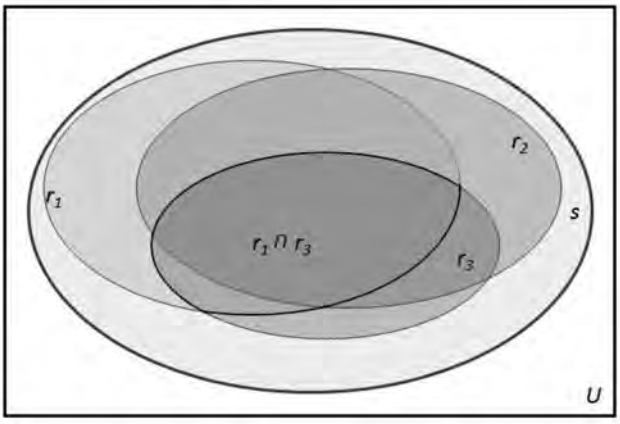


Figura 2.2: Conjunto de respondentes con patrón de respuesta: episódica

Luego de definir el conjunto de respondentes simultáneos a las olas j y $j+h$, la estimación de las diferencias de la variable y entre dichas olas se obtiene mediante la siguiente fórmula:

$$\begin{aligned} \hat{A}_{W_{j,j+h}} &= \sum_{r_j \cap r_{j+h}} (y_{(j+h)\,k} - y_{j\,k})\, w_{(j,j+h)\,k;} \\ &= \sum_{r_j \cap r_{j+h}} a_{(j,j+h)\,k}\, w_{(j,j+h)\,k} \end{aligned} \tag{2.8}$$

Donde:

- $w_{(j,j+h)\,k}$ son los ponderadores obtenidos mediante algún método de calibración para las unidades respondentes en las olas j y $j+h$, y
- $a_{(j,j+h)\,k} = y_{(j+h)\,k} - y_{j\,k}$ es el cambio en la variable de interés y entre estas olas para cada elemento respondente $k \in \{r_j \cap r_{j+h}\}$.

La aplicación de la fórmula (8) en los patrones de respuesta episódicos y por desgaste difiere únicamente en la definición del conjunto de respondentes simultáneos $r_j \cap r_{j+h}$.

2.4. Aplicación: Las damas perdidas

Una aplicación concreta del uso de ponderadores calibrados para encuestas de panel se realizó para la "Encuesta sobre Situaciones Familiares y Desempeños Sociales en

Montevideo y rea Metropolitana", llevada a cabo por un equipo de investigadores de la Universidad de la República (Facultad de Ciencias Económicas y de Administración , Instituto de Economía y de la Facultad de Ciencias Sociales, Departamento de Economía y Programa de Población).

Este panel consiste en dos olas, la primera fue realizada entre marzo y octubre de 2001 a una muestra de 1806 mujeres y la segunda ola se realizó en el año 2008, logrando recontactar a 828 de ellas. El conjunto de respondentes simultáneos para estimaciones longitudinales está entonces compuesto por las 828 mujeres recontactadas en 2008.

El método de calibración aplicado fue el raking o calibración en las marginales. Para el cálculo de los ponderadores se utilizó el programa (R Core Team, 2008) con el paquete (Lumley, 2009) y la función rake. Los insumos necesarios para calcular estos nuevos ponderadores son el diseño que genera los expansores originales (estratificado por nivel socioeconómico) y los totales marginales sobre los cuales se calibrará, que en esta aplicación corresponden a la edad y nivel educativo estimados a partir de la Encuesta Nacional de Hogares Ampliada del año 2006 (ENHA-2006).

2.4.1. Obtención de ponderadores

Los ponderadores a ser utilizados en la estimación longitudinal de cambios entre olas y en la estimación transversal de totales no podrán ser los mismos, ya que estas estimaciones estarán basadas en distintas unidades. Los cambios longitudinales solamente podrán ser estimados utilizando las mediciones efectivas sobre las 828 mujeres que fueron encuestadas en las dos olas del estudio, mientras que para las estimaciones transversales se cuenta con información sobre el total de mujeres entrevistadas (1229 mujeres). Por este motivo se obtendrán dos conjuntos de ponderadores, utilizando el raking como técnica de calibración.

Ponderadores para estimaciones longitudinales

Entre la instancia inicial en 2001 y la segunda ola en 2008, la imposibilidad de contacto de algunas mujeres (fallecimiento, movilidad, etc.) y la negativa de la mujer seleccionada a seguir participando genera la pérdida de 978 mujeres, provocando el desgaste del panel. Las 1806 mujeres entrevistadas en la primera instancia representan bien a la población de mujeres del 2001 con las características ya mencionadas. Si fuera posible medir los cambios entre olas en ciertas variables en cada una de estas mujeres, estos también serían

representativos de la población objetivo 2008. La no respuesta imposibilita dicha medición para todas las unidades de la muestra, y debe compensarse. Este objetivo se puede lograr siguiendo distintas estrategias: calculando ponderadores calibrados para las 828 mujeres respondentes para que representen a las 1806 iniciales (que a su vez, representan la población objetivo 2001), o calibrar los ponderadores de las 828 respondentes para que representen a la población objetivo 2008 de forma directa (las mujeres de 32 a 61 años de Gran Montevideo).

Se opta por seguir la segunda estrategia.Como ya se dijo, el procedimiento de calibración utilizado para el cálculo de estos ponderadores es el raking. Del pool de variables disponibles se deben seleccionar aquellas que serán utilizadas como información auxiliar para el cálculo de los nuevos ponderadores. Algunas de las características deseables para las variables auxiliares son su estabilidad en el tiempo y que permitan una buena caracterización de la población objeto de estudio. Para el caso particular de este estudio de panel, se descarta la utilización de variables provenientes del cuestionario de 2001 como variables auxiliares, ya que ellas serán objeto de la medición de cambios longitudinales. Las variables finalmente elegidas para la calibración son la edad y nivel educativo, al entenderse que verifican las condiciones mencionadas. Los totales poblacionales correspondientes se estiman a partir de la ENHA 2006.

Raking

Para el cálculo de los ponderadores se utilizó el programa (R Core Team, 2008) con el paquete (Lumley, 2009) y la función rake. Los insumos necesarios para calcular estos nuevos ponderadores son el diseño que genera los expansores originales y los totales marginales sobre los cuales se calibrará. En las tablas siguientes se presenta dicha información.

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.

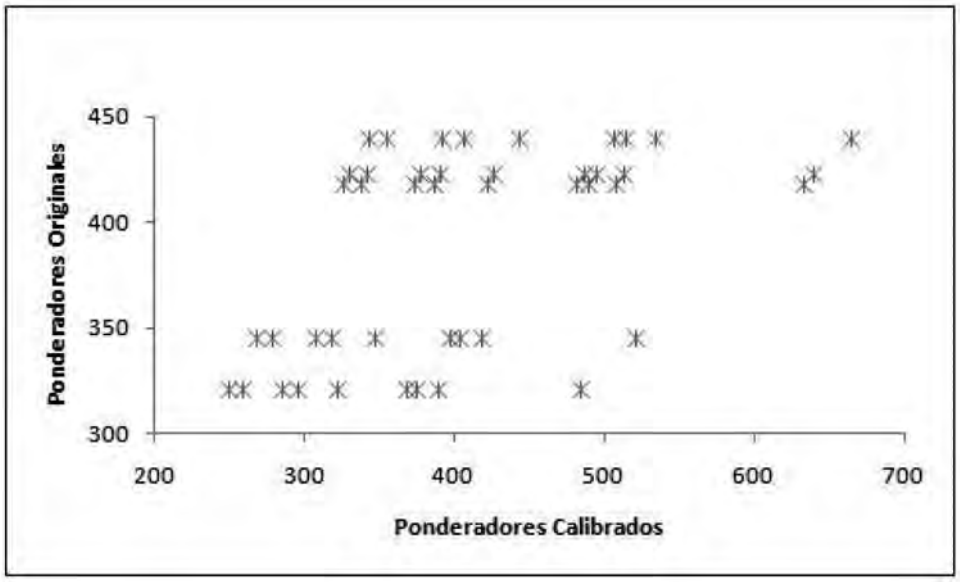
Estrato	Tot. Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	46924	111	533,738
MVD Medio Bajo	68596	199	344,703
MVD Medio Alto	76267	238	320,449
MVD Alto	57736	138	418,377
Periferia	62418	142	439,563
Total	311941	828	

Tabla 2.1: Totales poblacionales, muestrales y expansores originales por Estrato

Nivel Educativo	Frecuencia
Primaria	83112
Secundaria	147627
Terciaria	81202
Total	311941

Edad	Frecuencia
32 a 40 años	94225
41 a 50 años	112040
51 a 61 años	105676
Total	311941

Tabla 2.2: Totales poblacionales de las variables auxiliares



En el gráfico puede notarse que los ponderadores dentro de cada estrato (representados por los conjuntos de puntos alineados de forma paralela al eje de las abscisas) dejan de ser iguales. Dentro de cada estrato, hay nueve pesos diferentes, determinados por la interacción entre tramo de edad y nivel educativo, utilizados en el raking. En la siguiente tabla se presentan los totales muestrales que dan origen a 45 nuevos ponderadores.

Estrato y Nivel Educativo/Edad	32 a 40	41 a 50	51 a 61
Mvd Bajo - Primaria	9	17	20
Mvd Bajo - Secundaria	22	18	14
Mvd Bajo - Terciaria	4	2	5
Mvd Medio Bajo - Primaria	8	13	22
Mvd Medio Bajo - Secundaria	33	46	38
Mvd Medio Bajo - Terciaria	13	11	15
Mvd Medio Alto - Primaria	2	5	16
Mvd Medio Alto - Secundaria	27	49	45
Mvd Medio Alto - Terciaria	26	45	23
Mvd Alto - Primaria	1	1	4
Mvd Alto - Secundaria	9	17	19
Mvd Alto - Terciaria	17	36	34
Periferia - Primaria	10	20	21
Periferia - Secundaria	22	26	20
Periferia - Terciaria	11	10	2

Tabla 2.3: Totales muestrales según edad por estrato y nivel educativo

De la tabla surge la justificación de la utilización del raking frente a la post estratificación, ya que existen celdas de la clasificación estrato, nivel educativo y edad con muy pocas observaciones. Frente a la alternativa de colapsar categorías para luego aplicar la post estratificación para la obtención de ponderadores, se opta por la aplicación directa del raking.

Estimador de cambios longitudinales

Las diferencias para la variable y entre 2001 y 2008 se estiman por $\hat{A}_{W_{2001,2008}}$:

$$\begin{aligned}\hat{A}_{W_{2001,2008}} &= \sum_{k=1}^{828} (y_{2008\,k} - y_{2001\,k})\, w_{Ck;} \\ &= \sum_{k=1}^{828} a_{(2001,2008)\,k}\, w_{Ck} \end{aligned} \tag{2.9}$$

donde $a_{(2001,2008)\,k} = (y_{2008\,k} - y_{2001\,k})$ representa la diferencia de los valores de la

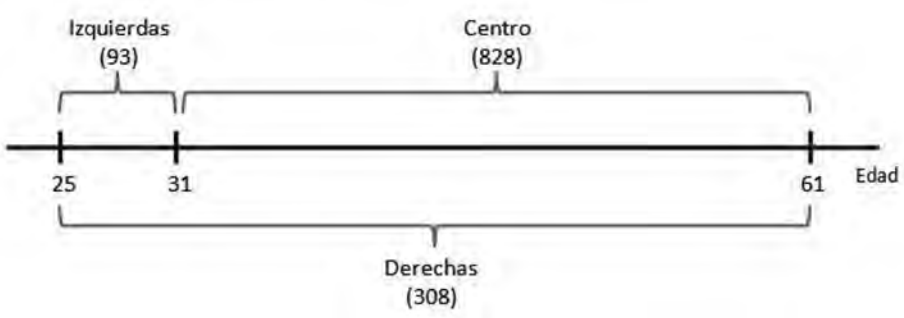
variable de interés y medidos en 2008 y 2001 para las 828 mujeres de la muestra original entrevistadas en 2001 y en 2008, y w_{Ck} el ponderador que surge de la aplicación del método de calibración raking para este grupo de mujeres.

Ponderadores para estimaciones transversales

Las estimaciones transversales para el 2008 se calcularán a partir de las respuestas dadas por las 1229 mujeres respondentes: 828 provenientes del panel original, 93 mujeres de edades entre 25 y 31 años (las "rejuvenecedoras" del panel) y 308 de 25 a 61 años de edad (las "ampliadoras" del panel). El objetivo de lograr que las 1229 mujeres representen bien a las mujeres de Gran Montevideo con edades entre 25 y 61 años en 2008 requiere el cálculo por separado de ponderadores en cada uno de estos tres grupos, ya que las mujeres que integran cada grupo provienen de muestras distintas. La libertad de elección de ponderadores iniciales $d_{\alpha k}$ mencionada en la sección (5.5.1) no contempla el caso en que las unidades provengan de muestras distintas. Es por esto que en cada uno de estos grupos se calcularán los nuevos ponderadores con el método raking para luego combinar los resultados. Las variables auxiliares a ser utilizadas en el procedimiento de calibración seleccionado para cada uno de estos grupos serán edad y nivel educativo, de igual manera que en la sección anterior.

En la siguiente figura se representa el rango de edad de las 1229 mujeres relevadas en 2008 según la muestra de la que provienen: centro, derechas e izquierdas.

Figura 2.3: Representación gráfica de las edades de las 1229 mujeres relevadas en 2008 según muestra de origen



Raking Centro

Los nuevos ponderadores para estas 828 mujeres centro son los mismos que fueron calculados en la parte anterior.

Raking Izquierdas

En este caso la única variable auxiliar para calibrar es nivel educativo dado que todas estas mujeres tienen entre 25 a 31 años, que corresponde a una única categoría de la variable edad. Los ponderadores originales y los totales poblacionales se presentan en la tabla ?? y la tabla 2.5.

Estrato	Tot.Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	13931	14	995,071
MVD Medio Bajo	18622	24	775,917
MVD Medio Alto	21677	20	1083,850
MVD Alto	14050	10	1405,000
Periferia	15804	25	632,160
Total	84084	93	

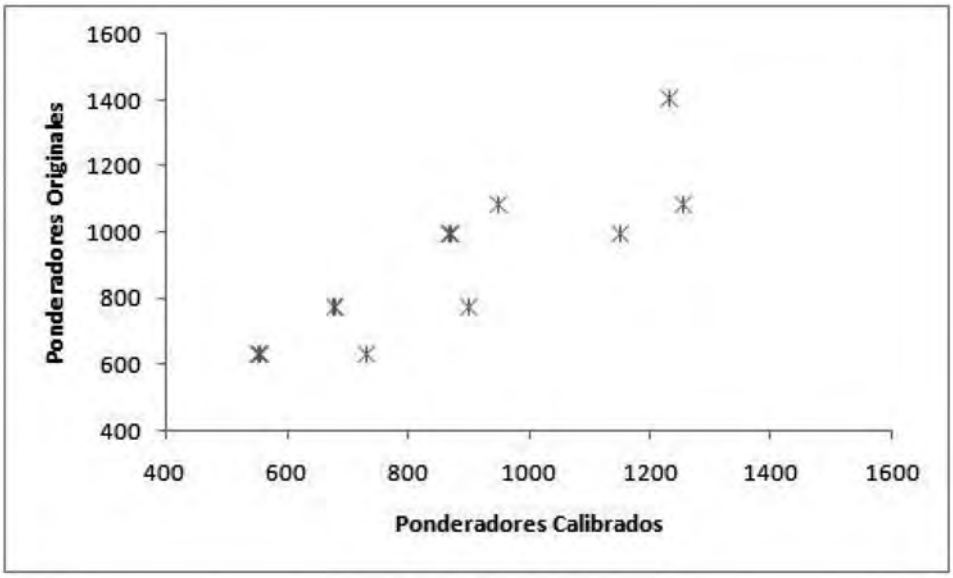
Tabla 2.4: Totales poblacionales, muestrales y expansores originales según estrato

Nivel Educativo	Frecuencia
Primaria	12804
Secundaria	42852
Terciaria	28428
Total	84084

Tabla 2.5: Totales poblacionales de la variable auxiliar

En este caso el método raking genera los mismos resultados que el método de post estratificación, ya que calibrar en las marginales de una única variable auxiliar es equivalente a calibrar en las celdas.

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.



En este caso hay tres ponderadores por estrato, las tres categorías de nivel educativo, excepto para los estratos Montevideo Medio Alto y Alto; esto último se debe a que en esta muestra no hay mujeres con solamente primaria completa en estos estratos, como se muestra en la siguiente tabla.

Estrato/Nivel Educativo	Primaria	Secundaria	Terciaria	Total
MVD Bajo	4	9	1	14
MVD Medio Bajo	4	14	6	24
MVD Medio Alto	0	8	12	20
MVD Alto	0	2	8	10
Periferia	12	9	4	25
Total	20	42	31	93

Tabla 2.6: Totales muestrales por nivel educativo según estrato

Raking derechas

Los insumos necesarios para calcular estos nuevos ponderadores son el diseño que genera los expansores originales y los totales marginales sobre los cuales se calibrará, de igual manera a las partes anteriores, pero con la excepción de que la variable auxiliar edad tiene ahora cuatro categorías.

Estrato	Tot. Poblacionales	Tot. Muestrales	Expansor Original
MVD Bajo	60855	39	1560,385
MVD Medio Bajo	87218	81	1076,765
MVD Medio Alto	97944	84	1166,000
MVD Alto	71786	47	1527,362
Periferia	78222	57	1372,316
Total	396025	308	

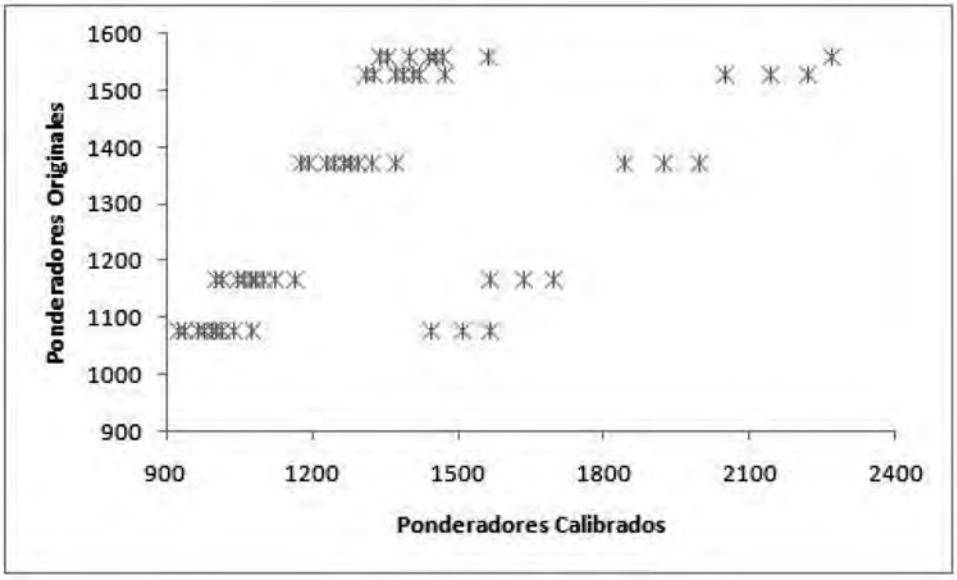
Tabla 2.7: Totales poblacionales, muestrales y expansores originales según estrato

Nivel Educativo	Frecuencia
Primaria	95916
Secundaria	190479
Terciaria	109630
Total	396025

Edad	Frecuencia
25 a 31 años	84084
32 a 40 años	94225
41 a 50 años	112040
51 a 61 años	105676
Total	396025

Tabla 2.8: Totales poblacionales de las variables auxiliares

En el siguiente gráfico se presenta la modificación de los expansores originales obtenidos a partir del raking.



La interacción entre las variables nivel educativo (3 categorías) y edad (4 categorías) para cada uno de los cinco estratos debería generar 60 pesos diferentes. En los hechos, se pueden distinguir 54 ponderadores diferentes debido a que no todas las celdas correspondientes a la interacción de las variables auxiliares contienen observaciones, como se puede ver en la tabla 2.9.

Estrato y Nivel Educativo/Edad	25 a 31	32 a 40	41 a 50	51 a 61
Mvd Bajo - Primaria	4	6	4	3
Mvd Bajo - Secundaria	4	7	10	0
Mvd Bajo - Terciaria	0	1	0	0
Mvd Medio Bajo - Primaria	5	5	4	6
Mvd Medio Bajo - Secundaria	15	14	9	7
Mvd Medio Bajo - Terciaria	4	5	4	3
Mvd Medio Alto - Primaria	1	3	1	4
Mvd Medio Alto - Secundaria	8	7	21	7
Mvd Medio Alto - Terciaria	7	11	8	6
Mvd Alto - Primaria	0	2	0	2
Mvd Alto - Secundaria	4	5	6	2
Mvd Alto - Terciaria	6	1	8	11
Periferia - Primaria	2	4	10	2
Periferia - Secundaria	9	10	9	5
Periferia - Terciaria	1	2	2	1

Tabla 2.9: Totales muestrales por edad según por estrato y nivel educativo

Ponderadores combinados

Para cada una de las 1229 mujeres se calculó su respectivo ponderador, con relación a la submuestra a la que pertenece. Las 828 mujeres del centro representan al total de mujeres de Gran Montevideo con edades entre 32 y 61 años (311941 mujeres); las 93 mujeres de la izquierda representan a las 84.084 mujeres dentro de la franja etaria de 25 a 31 años; y las 308 de la derecha a aquellas de edades entre 25 y 61 años, cuyo total asciende a 396.025 mujeres. Si para realizar cálculos de totales se utilizaran los ponderadores obtenidos en cada uno de los grupos de manera directa, dichos totales se estarían sobreestimando. De hecho, se estaría estimando el total correspondiente a una población compuesta por el doble de las mujeres existentes en Gran Montevideo de edades entre 25 y 61 años.

Es por este motivo que los ponderadores calculados en las partes anteriores deben utilizarse de forma combinada, para lograr la estimación sobre el total efectivo de mujeres de dichas características: 396.025 mujeres.

Las 1229 mujeres entrevistadas en 2008 se clasifican en cuatro subpoblaciones con relación a su edad (mayor o menor a 31 años) y muestra de procedencia (centro, izquierdas, derechas), como muestra la Figura 2.4

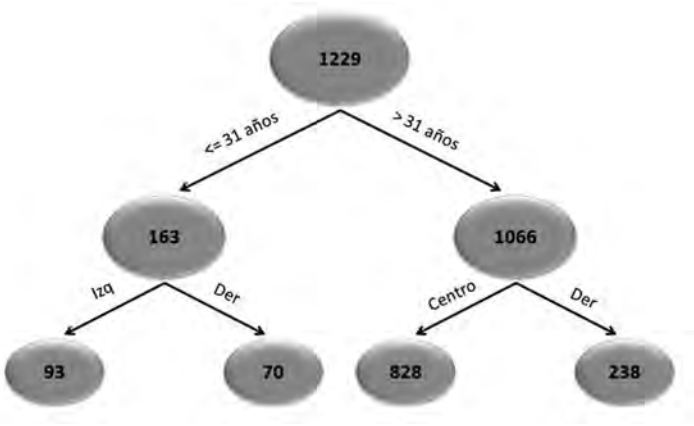


Figura 2.4: Clasificación de las mujeres entrevistadas en 2008

El primer nivel de división (corte por edad) permite identificar dos grandes grupos: 163 mujeres de edades entre 25 y 31 años que deberán representar a 84.084 mujeres, y 1066 mayores a 31 años, que deberán representar a 311.941 mujeres existentes en dicha franja etaria. El segundo corte queda determinado por la muestra de la que provienen cada una de las mujeres: las 163 mujeres menores a 32 años pueden pertenecer a la muestra de izquierdas o derechas, y las 1066 mayores de 31 pueden ser parte del panel original o de la muestra de derechas.

La combinación de resultados sigue la misma lógica que la figura: en primer lugar, se particiona la muestra total de acuerdo a la edad (mayores o menores de 31 años) formando dos grandes grupos. Dentro de cada grupo los nuevos ponderadores combinados tendrán en cuenta la muestra de la que provienen a través de la relación que existe entre la cantidad de mujeres en esta última partición y la cantidad de mujeres en el grupo etario que corresponda. Por ejemplo, los pesos obtenidos a través del raking para las 828 mujeres provenientes del panel original se multiplicarán por 828/1066, la proporción de mujeres del panel que corresponde a la franja etaria establecida. Considerando que las

varianzas de los estimadores son del orden de $\frac{1}{n_i}$ siendo n_i el tamaño de cada una de las cuatro particiones, se combinan los ponderadores resultantes del raking de manera que las mujeres provenientes de subgrupos más grandes adquieran mayor importancia en el análisis.

Los ponderadores combinados a ser utilizados en las estimaciones transversales resultan ser:

$$w_{comb_k} = \begin{cases} \frac{93}{163} w_{Ik} & k \in \text{muestra de izquierdas} \\ \frac{70}{163} w_{Dk} & k \in \text{muestra de derechas con edades entre 25 y 31 años} \\ \frac{238}{1066} w_{Dk} & k \in \text{muestra de derechas con edades entre 32 y 61 años} \\ \frac{828}{1066} w_{Ck} & k \in \text{panel original} \end{cases}$$

Estos ponderadores combinados estiman sin error la cantidad total de mujeres pertenecientes a la población objetivo estimada por la ENHA 2006, que se considera equivalente a la población 2008.

$$\begin{aligned} \hat{N} &= \sum_{k=1}^{1229} w_{comb_k} = \\ &= \sum_{k=1}^{93} \frac{93}{163} w_{Ik} + \sum_{k=1}^{70} \frac{70}{163} w_{Dk} + \sum_{k=1}^{238} \frac{238}{1066} w_{Dk} + \sum_{k=1}^{828} \frac{828}{1066} w_{Ck} \\ &= 47974,3 + 36109,7 + 69645,4 + 242295,6 \\ &= 396025 = N \end{aligned}$$

Estimación de totales transversales

La estimación transversal de los totales de interés debe realizarse utilizando los ponderadores combinados de acuerdo a la siguiente fórmula:

$$\begin{aligned} \hat{t}_{y_{Wcomb}} &= \sum_{k=1}^{1229} w_{comb_k} y_k \\ &= \sum_{k=1}^{93} \frac{93}{163} w_{Ik} y_k + \sum_{k=1}^{70} \frac{70}{163} w_{Dk} y_k + \sum_{k=1}^{238} \frac{238}{1066} w_{Dk} y_k + \sum_{k=1}^{828} \frac{828}{1066} w_{Ck} y_k \end{aligned} \quad (2.10)$$

2.5. Conclusiones

La medición de cambios entre distintas instancias en el tiempo es el principal objetivo de las encuestas de panel. Se presentaron estimadores del cambio de las variables de interés que, bajo el supuesto de respuesta perfecta, son insesgados para estimar el cambio total en la población objetivo. Para su desarrollo, se partió de supuestos muy restrictivos rara vez presentes en la práctica: la existencia de respuesta perfecta y población fija en el tiempo, reflejados en ponderadores constantes en el tiempo para cada elemento e iguales al inverso de su probabilidad de inclusión en la muestra.

La no respuesta es un fenómeno presente en la mayoría de las encuestas por muestreo y es necesario su tratamiento para evitar sesgos en las estimaciones. En las encuestas de panel la inclusión del factor tiempo provoca un agravamiento del problema de no respuesta, reflejado en reducciones considerables en el tamaño de muestra período a período. Si la no respuesta se presentara de manera completamente aleatoria, el único incoveniente al que se enfrenta el investigador resulta en la reducción del tamaño de muestra y su respectivo aumento en la varianza de las estimaciones, pero las unidades que no contestan suelen diferir de aquellas que sí lo hacen, y el sesgo introducido en las estimaciones por esta causa constituye el obstáculo más importante por corregir.

Aun cuando se invierta esfuerzo en intentar alcanzar la mayor tasa de respuesta posible, la no respuesta existe y es deber de los investigadores realizar algún tratamiento en la etapa de análisis de datos para controlar el sesgo introducido a las estimaciones por su causa. Como método de tratamiento se descartan la sustitución y submuestreo de no respondentes y los métodos basados en la quasi-ramdomnization por entender que no son las opciones preferibles en encuestas de panel. La imputación como forma de tratar la no respuesta de unidades en la ola también se descarta por generar una fabricación masiva de datos, que pueden distorisionar las asociaciones entre las va- riables que representan el principal objetivo del panel. Es preferible entonces utilizar una estrategia global como es la calibración cuando es la unidad la que no provee de respuesta. De todas maneras la imputación es el tratamiento elegido como manera de compensar la no respuesta en los ítems, previo a la calibración por la no respuesta de unidades.

La idea central de los estimadores calibrados es sencilla, a partir de información au- xiliar se modifican los ponderadores originales de la muestra minimizando alguna función de distancia entre dichos ponderadores y los ponderadores finales (o calibrados)

y de manera que estos últimos estimen sin error totales poblacionales conocidos de las variables auxiliares que asisten al procedimiento de calibración.

El estimador calibrado corresponde en realidad a una familia entera de estimadores que dependen de formulaciones diferentes del vector auxiliar y de la función de distancia.

La efectividad del estimador de calibración para controlar el sesgo ocasionado por la no respuesta dependerá de propiedades del vector auxiliar. Se obtiene una expresión del sesgo aproximado que demuestra que este será menor cuanto más estrecha sea la relación entre la información auxiliar y la probabilidad de respuesta o la variable de interés. También se propone una formulación para la estimación de la varianza del estimador calibrado que podrá ser utilizada en la construcción de intervalos de confianza. Si el sesgo es modesto el intervalo de confianza será válido y la probabilidad de cobertura será cercana al nivel de confianza especificado.

El estimador de cambios entre olas propuesto para el caso de respuesta perfecta se extiende utilizando los ponderadores calibrados calculados para las unidades respondentes. De esta manera, los cambios entre las olas de interés medido en los respondentes simultáneos ponderado por los pesos calibrados representarán los cambios de la población objetivo.

Bibliografía

Deville, J. y Särndal, C. (1994). Variance estimation for the regression imputed Horvitz Thompson estimator. Journal of Official Statistics, 10(381-394).

Lumley, T. (2009). survey: analysis of complex survey samples R package version 3.11-2.

R Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Särndal, C. y Lundström, S. (2005). Estimation in Surveys with Nonresponse. New York: John Wiley and Sons, Ltd.

Capítulo 3

Efecto de valores faltantes en estudios longitudinales en adultos mayores, por Fernando Massa

Resumen[1]

Este proyecto pretende modelar el deterioro de la función cognitiva global de un conjunto de adultos mayores a lo largo del tiempo. En cuanto al deterioro cognitivo, existen muchas herramientas para cuantificarlo, en este trabajo se optó por utilizar los resultados del Mini Mental State Examination ($MMSE$), medido a lo largo del tiempo, en intervalos de dos años. Una de las desventajas de los diversos estudios longitudinales es que, a lo largo del período de seguimiento, algunos de los individuos desertan del estudio por diversas causas. En el caso de investigaciones sobre deterioro cognitivo, la población objetivo está compuesta por ancianos, por este motivo, el fallecimiento es la causa más importante de deserción.

En numerosos estudios se ha comprobado que el abandono de los sujetos, además de reducir el tamaño muestral, genera importantes sesgos en los resultados de los análisis estadísticos. Debido a esto, en la literatura han surgido diversos enfoques que intentan solucionar este inconveniente. En esta tesis se le presta especial atención a los modelos conjuntos, los cuales parten de la base de que el tiempo de sobrevida de los sujetos y sus resultados de $MMSE$ están relacionados.

Como resultados generales, se vio que los valores del $MMSE$ se ven afectados tanto por la edad de los indivduos al inicio del estudio, como por la educación. Al incluir el análisis de sobrevida como parte del proceso de deterioro cognitivo, se "corrigieron" los valores de los coeficientes estimados mejorando la precisión de las estimaciones.

Palabras claves: Datos faltantes, datos longitudinales, deterioro cognitivo, modelo conjunto.

1. Resumen del trabajo de tesis, con la tutoría de la Dra. Graciela Muñiz Terra para la obtención del posgrado de la Maestría en Ingeniería Matemética.

3.1. Introducción

Naciones Unidas ha estimado que en el año 2050 el numero de individuos mayores de 60 años superará los 2 billones y que por primera vez, la proporción de la poblacion de adultos mayores superará a la de niños menores de 14 años. Las consecuencias de este cambio en la pirámide poblacional son vastas desde el punto de vista social e individual. Algunos países están tomado previsiones al respecto con el fin de reducir el impacto en las estructuras sociales.

A nivel individual, la pérdida de la salud física y de la capacidad cognitiva son los mayores desafíos que los adultos mayores enfrentan. Entender estos procesos es entonces fundamental, y para eso, los estudios longitudinales representan una excelente oportunidad. Científicamente hay varias preguntas de interés, siendo algunas de ellas ¿Cuál es el cambio marginal de la poblacion total a lo largo el tiempo? ¿Cuál es el cambio experimentado a nivel individual? ¿Cuáles son los principales factores de riesgo que afectan la pérdida de la capacidad cognitiva? ¿Son los mismos para individuos en el segmento mas jóven (75-79 años), 80-84 y los más ancianos (89+)? ¿Cuáles de estos factores son modificables? ¿Cuál es el efecto de haber sufrido un accidente cardiovascular, o de estar clínicamente deprimido en la salud mental? ¿Existe un efecto de los años de educación en la pérdida de la capacidad cognitiva? ¿Personas con mayor escolaridad declinan a la misma velocidad que los menos educados? Estas últimas interrogantes son de suma importancia debido a que la educación es un factor modificable por el propio individuo (Lenehan et al., 2015).

Cuando buscamos respuestas a estas preguntas, nos enfrentamos con uno de los mayores problemas de los estudios longitudinales de individuos mayores: el gran número de observaciones perdidas. Entender las razones por las cuales los individuos cesan su participación en los estudios es fundamental para modelar adecuadamente el cambio experimentado por estos individuos, ya sea a nivel individual como marginal. Por ejemplo, algunos individuos cesan su participación debido a condiciones de salud física, otros dejarán de participar por razones totalmente ajenas a su situacion de salud (mudanza a otra zona del país), otros estarán muy impedidos mentalmente como para comprender las preguntas que se les hacen en los tests y finalmente otros morirán durante el desarrollo del estudio.

Ignorar la información aportada por algunos de estos casos quizá no afecte los estimadores (y por ende las conclusiones del estudio), pero ciertamente ignorar la información aportada por otros sesgará los resultados masivamente.

3.2. Antecedentes

En los últimos 20 años se han formado diversas líneas de investigación dentro del campo del estudio de la evolución del envejecimiento, más concretamente, en lo que refiere al deterioro cognitivo, algunas plicaciones de estas técnicas han tocado temas como mortalidad, memoria, deterioro y habilidad motora.

El trabajo de (Arbeev et al., 2014) presenta una comparación entre las virtudes de los modelos conjuntos y los modelos de procesos estocásticos (así como algunas modificaciones de este último) para capturar la variación biológica en los patrones longitudinales y predecir las tasas de mortalidad tanto de individuos como de poblaciones. En (Terrera et al., 2011) se presta particular atención a los cambios en la memoria y como su deterioro podría presentar una aceleración en su trayectoria. Uno de los principales hallazgos fue que el la tasa de decaimiento de la memoria consiste en un predictor de la proximidad del fallecimiento. En (Ghisletta, 2008) se trata con la velocidad de percepción y la fluidez verbal de ancianos y cómo estas variables se asocian con el riesgo de fallecimiento, habiendo controlado por la edad al inicio del estudio, género, salud general, estado socioeconómico. El trabajo de (Hughes et al., 1997) se encarga de estudiar los factores que predicen el deterioroen la habilidad manual de los ancianos.

Fuera del ámbito del envejecimiento, existe una vasta literatura dedicada a la descripción de los avances en el área de modelos conjuntos. En la gran mayoría de los casos, está dedicada al análisis de biomarcadores relacionados con la evolución del virus del síndrome de inmunodeficiencia adquirida (SIDA). En este entorno se destacan los trabajos de (Self y Pawitan, 1992) y de (DeGruttola y Tu, 1994). El primero considera el desarrollo de un modelo conjunto para la descripción de la evolución del número de células de los biomarcadores T4 y T8 así como el tiempo desde la seroconversión hasta el diagnóstico de SIDA. El segundo se encarga de analizar la incidencia de diversos factores en la evolución del conteo de linfocitos CD4 y su asociación con el tiempo de sobrevida. En este trabajo, los autores adoptan la metodología propuesta por (Wu y Carroll, 1988) incluyendo

efectos aleatorios para modelar la asociación entre el proceso longitudinal y el proceso de sobrevida. En contrapartida a los casos anteriores, (Faucett y Thomas, 1996) se valen de métodos Bayesianos, más concretamente Gibbs-Sampling para estimar la distribución a posteriori de los parámetros del modelo conjunto. En última instancia es muy valioso destacar el trabajo de (Wulfsohn y Tsiatis, 1997) quienes sientan las bases del que hoy se considera el modelo conjunto standard. En su paper, los autores proponen el uso del algoritmo Expectation-Maximization (EM) (Dempster et al., 1977) asegurando que su método es superior a los anteriores ya que no se basa en una estimación en dos etapas y no maximiza la porción de la verosimilitud del modelo de sobrevida de manera utilizando los valores observados del biomarcador (contaminados por el error de medición).

Pese a no existir antecedentes en el Uruguay, estos modelos podrían ser de especial utilidad debido a que Uruguay es uno de los países más envejecidos de la región (Cabella y Pellegrino, 2009). En este sentido, estudiar los posibles determinantes que intervienen en el proceso de envejecimiento de los ancianos uruguayos sería un punto a tener en cuenta para futuras líneas de investigación ya que tener un mejor entendimiento de proceso de deterioro podría permitir una mejor planificación económia/familiar. En este sentido, dado que estos modelos permiten realizar predicciones (tanto de la trayectoria de las variables asociadas al deterioro como de la probabilidad de fallecimiento) esto permitiría al sistema de salud/seguridad social planificar con más información.

Otras posiblees aplicaciones de estos modelos, ya fuera del ámbito del deterioro cognitivo, serían en el contexto de los trasplantes de órganos (algunos biomarcadores podrían predecir la probabilidad de rechazo del órgano y el tiempo de sobrevida de este y del paciente) o para estudiar la evolución (y las causas) del abandono de los niños al sistema educativo.

3.3. Objetivos

Este trabajo se llevó a cabo con la finalidad de aplicar las técnicas referentes al modelado conjunto de datos longitudinales y de sobrevida. Para ello se utilizaron los datos del estudio "Origins of Variance in the Old-old: Octogenarian Twins" (OCTO-Twin Study). Este consistía de 351 parejas de mellizos de 80 años o más en el año 1991 en Suecia. El período de seguimiento de los individuos consistió de 4 visitas (posteriores a la evaluación inicial) en períodos de 2 años donde se recabaron datos sobre memoria,

capacidad funcional, y salud entre otras mediciones. El presente estudio se focalizó sobre la evolución del resultado del "Mini-Mental State Examination" ($MMSE$). Se trata de un cuestionario con un puntaje máximo de 30 puntos desarrollado por Folstein (Folstein et al., 1975) que pretende determinar el deterioro cognitivo y detectar la demencia.
El objetivo general que se persigue en este trabajo es el de estimar la velocidad de cambio del $MMSE$ y determinar los factores que puedan alterarla, prestando especial atención al proceso de fallecimiento de los individuos ya que este puede sesgar los resultados.

3.4. Deterioro cognitivo

Todas las personas desarrollan cierto grado de deterioro en sus funciones cognitivas a medida que transcurre el tiempo. El impedimento cognitivo es un término clínico utilizado para describir una condición asociada a problemas de la función cognitiva como lo son pensar, recordar, razonar, problemas en el lenguaje, en la atención o en ciertas actividades visuales o espaciales. Las personas que lo sufren suelen tener mayores problemas en el "día a día" en actividades relacionadas principalmente con la memoria, pero dichos problemas no son lo suficientemente graves como para diagnosticar un caso de demencia.
En estas personas, el proceso de deterioro es más pronunciado que en personas que envejecen de manera natural. En el caso particular de la pérdida de memoria, esto puede indicar un primer signo del desarrollo de demencia o incluso Alzheimer. En pacientes que presentan signos de deterioro en otras actividades cognitivas, esto puede resultar en el desarrollo de otras enfermedades como demencia vascular, demencia fronto-temporal o demencia de cuerpos de Lewy.
En cuanto a la detección del impedimento en sí, existen diversas alternativas, uno de los cuales es el $MMSE$. Pese a que este test no constituye una herramienta de diagnóstico, si es un instrumento que permite detectar y estimar cuantitativamente la severidad del deterioro cognitivo y sus cambios a lo largo del tiempo. La prueba en sí es sencilla y de rápida aplicación pero su aplicación es limitada ya que presenta ciertos sesgos en tanto que no logra captar pérdidas leves en la memoria de sujetos con alto nivel educativo. La prueba se compone de varias secciones, las cuales pretenden evaluar la orientación, memoria de corto plazo y lenguaje. El resultado de la prueba es un puntaje (con máximo de 30 puntos), valores superiores a 25 sugieren un estado normal en la persona, valores entre 21 y 24 sugieren un deterioro leve de las funciones cognitivas, valores entre 10 y

20 son signos de un deterioro moderado mientras que valores menores a 10 puntos son característicos de personas con un deterioro severo.

3.5. Análisis de datos longitudinales

Un estudio longitudinal es aquel en el cual las variables de interés se registran en múltiples instancias sobre un conjunto de individuos. La principal caracterísitca de esta manera de investigar es que permite caracterizar el cambio y los factores que lo determinan a través del tiempo. Sin embargo, la metodología de análisis clásico presenta limitaciones en este tipo de estudios debido a que, dada la naturaleza secuencial del relevamiento de datos, existe correlación entre las observaciones de un mismo individuo. No obstante, el hecho de que las medidas se registren de manera secuencial acarrea el inconveniente de que las mediciones de un mismo individuo suelen presentar correlación (generalmente positiva) por lo tanto requieren de un tratamiento específico que permita realizar inferencias válidas. De todas maneras, al hacer un balance entre "ventajas" y "desventajas", el resultado neto es que los datos longitudinales proveen de una mayor cantidad de información para cada individuo que un análisis de tipo "cross-section".
El modelo presentado utilizado para analizar estudios longitudinales suele adoptar la siguiente formulación:

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \epsilon_i \\ b_i \sim N(0, D) \\ \epsilon \sim N(0, \Sigma) \end{cases} \qquad (3.1)$$

En este caso Y_i es un vector con las n_i observaciones del i-ésimo individuo, X_i es la matriz que contiene las covariables que afectan los coeficientes poblacionales, Z_i contiene las covariables de la parte aleatoria, β y b representan los efectos fijos y aleatorios respectivamente y por último ϵ_i es el vector de errores, de la misma dimensión de Y_i que pueden o no presentar varianzas heterogéneas y covarianzas nulas para un mismo individuo. Vale aclarar que uno de los supuestos de partida de estos modelos es que se asume independencia entre las variables no observables ϵ y b así como también se supone independencia entre los vectores de respuesta de distintos individuos. Desde el punto de vista del modelo esto implica que $Cov(\epsilon_i, \epsilon_{i'}) = 0 \quad \forall \quad i \neq i'$.
Es sencillo notar que al condicionar en los efectos aleatorios, la distribución del perfil del sujeto i es normal con vector de medias $X_i\beta + Z_i b_i$ y varianza Σ.

Algunas de las ventajas que postula el modelo descrito en (3.1) para analizar patrones longitudinales es que no presenta inconvenientes cuando diferentes individuos tienen distinto número de observaciones y que es capaz de manejar adecuadamente el hecho de que, por lo general, en estudios longitudinales las mediciones no suelen llevarse a cabo para toda la muestra de individuos en los mismos momentos. En cuanto a las ventajas del modelo propiamente dicho, permite la predicción de la trayectoria futura de cada individuo en particular así como de la población en su conjunto. Sin embargo, postulado de esta manera, el modelo también permite realizar la predicción de los valores del individuo luego de fallecimiento.

3.5.1. Inferencia

Según lo expuesto en el apartado anterior, el vector con las mediciones del i-ésimo individuo se distribuye

$$Y_i \sim N\left(X_i\beta, Z_i^{'}DZ_i + \Sigma\right) \tag{3.2}$$

Las inferencias realizadas bajo este modelo marginal no necesariamente asumen la presencia de efectos aleatorios, pese a que estos se hayan usado para modelar la heterogeneidad entre los individuos. Antes de plantear la función de verosimilitud de la muestra de n individuos, es conveniente particionar el vector de parámetros (como en (Verbeke y Molenberghs, 2000)) en dos conjuntos, de esta forma, este el vector θ contendrá todos los parámetros del modelo, siendo β el vector de parámetros asociados a la media del vector Y_i y γ el vector cuyas componentes intervienen en los elementos de D y Σ. De esta manera la función de log-verosimilitud será la indicada en la ecuación (3.3).

$$\mathcal{L}_{ML}(\theta) \propto \sum_{i=1}^{n}\left\{\frac{1}{2}\left|V_i(\gamma)\right| - \frac{1}{2}(Y_i - X_i\beta)^{'}V_i^{-1}(\gamma)(Y_i - X_i\beta)\right\} \tag{3.3}$$

Al maximizar esta función con respecto a γ y β se obtienen los estimadores de máxima verosimilitud "marginal". Sin embargo estos estimadores no poseen una forma cerrada, por lo cual un procedimiento es obtener el estimador de mínimos cuadrados generalizados de β como:

$$\hat{\beta}(\gamma) = \left(\sum_{i=1}^{n} X_i V_i^{-1}(\gamma) X_i\right)^{-1} \sum_{i=1}^{n} X_i V_i(\gamma) Y_i \tag{3.4}$$

De esta manera, suponiendo conocido el vector de parámetros de covarianza, es posible estimar el vector β, no obstante para esto se hace necesario estimar al vector γ, proceso que suele llevarse a cabo mediante métodos numéricos. Es necesario agregar que la

estimación de los parámetros de covarianza puede llevarse a cabo maximizando la verosimilitud o la versión restringida de esta última, a estos estimadores se los llama $REML$ por su sigla en inglés Restricted Maximum Likelihood. La ventaja de estos estimadores es que, a diferencia de la versión de máxima verosimilitud, son insesgados.
Al ser derivados bajo este planteamiento, tanto los estimadores obtenidos por ML como por $REML$ gozan de propiedades deseables como consistencia, normalidad asintótica y eficiencia (Richardson y Welsh, 2008).

3.5.2. Inferencia sobre β

A la hora de realizar inferencias sobre los elementos del vector β existen diferentes procedimientos. Los más comunes son realizar intervalos de confianza, o llevar a cabo pruebas de hipótesis. En este último caso existen diversos enfoques. Aquí se presentarán las dos metodologías más comunes que son las pruebas basadas en los estadísticos de Wald y de cociente de verosimilitud.

- Pruebas basadas en el estadístico de Wald.

  Antes de pasar al cuerpo de la prueba, vale la pena notar que el vector de estimaciones $\hat{\beta}(\gamma)$ estimado por mínimos cuadrados generalizados (asumiendo que se conocen los parámetros contenidos en γ) se distribuye normal con la siguiente media y matriz de covarianzas:

  $$\hat{\beta}(\gamma) \sim N\left(\beta, X^{'}V^{-1}(\gamma)X\right) \tag{3.5}$$

  donde $V(\gamma) = Z^{'}DZ + \Sigma$.

  Trabajando sobre esta base, se pueden construir pruebas de hipótesis basadas en restricciones lineales del vector β de la siguiente manera:

  $\mathrm{H}_0) L\beta = \underline{c}$

  $H_1) L\beta \neq \underline{c}$

  Donde L es una matriz de contrastes (lineales) que indica las restricciones que se desea poner a prueba. El estadístico de prueba es consecuencia de la normalidad del vector $L\beta$ y su forma es:

  $$F = \frac{(\hat{\beta} - \beta)^{'} L^{'} \left[L(X^{'}V^{-1}(\gamma)X)^{-1}L^{'}\right] L(\hat{\beta} - \beta)}{rango(L)} \tag{3.6}$$

Bajo el cumplimiento de la hipótesis nula, el estadístico (3.6) sigue una distribución F con grados de libertad del numerador equivalentes al rango de la matriz L, mientras que los grados de libertad del denominador deben ser estimados a partir de los datos (véase (Waseem, 2007)). En el caso especial de que se quisiera contrastar:

$H_0)\ \beta_k = 0$

$H_1)\ \beta_k \neq 0$

basta con utilizar $L = (0, 0, \ldots, 1, \ldots, 0, 0)$ y $c = 0$. Donde el 1 ocupa el lugar k-ésimo del vector fila L. En dicho caso, la raíz cuadrada del estadístico sigue una distribución de Student cuyos grados de libertad son los correspondientes a los del denominador del estadístico F (los cuales suelen ser estimados por el procedimiento de Satterthwaite, Welch, Kenward-Roger, etc.).

- Pruebas basadas en el estadístico de cociente de verosimilitud (LRT).

  En el caso del estadístico de cociente de verosimilitud, es necesario re-estimar los parámetros del modelo bajo el cumplimiento de la hipótesis nula (modelo que suele denominarse "reducido" o "restringido"), lo cual puede resultar costoso si la estimación es computacionalmente demandante. Otros dos aspectos a tener en cuenta son que el método de estimación debe ser máxima verosimilitud y que se utilicen las mismas observaciones en la estimación de modelo "completo" y del modelo "reducido". Este último punto es de especial importancia en los casos en que algunas covariables presenten datos faltantes en distintas observaciones. El procedimiento dictamina la comparación de las log-verosmilitudes de ambos modelos ($\mathcal{L}_{completo}(\theta)$ y $\mathcal{L}_{reducido}(\theta)$), y se espera que cuanto mayor sea la diferencia entre ambos, mayor es la evidencia en contra de la hipótesis nula.

  Finalmente el estadístico de prueba se construye de la siguiente manera:

$$LRT = -2(\mathcal{L}_{reducido}(\theta) - \mathcal{L}_{completo}(\theta)) \tag{3.7}$$

  La distribución del estadístico (3.7) se aproxima (asintóticamente) a una χ^2 con tantos grados de libertad como restricciones se hayan impuesto sobre el modelo reducido. Un uso adicional de este procedimiento (aunque computacionalmete es aún más demandante) es la construcción de intervalos de confianza mediante el "perfilado" de la verosimilitud.

Inferencia sobre γ

Pese a que en la gran mayoría de las situaciones, el interés de estos modelos recae sobre las inferencias realizadas sobre la media, modelar los parámetros correspondientes a la covarianza de manera adecuada es igual de importante, no solo para realizar una descripción más rica del fenómeno bajo estudio (modelando la variación intra-individual) sino también, para obtener inferencias válidas de los elementos del vector β.
La interrogante más común a la hora de realizar inferencias sobre estos parámetros es la siguiente:

$$\begin{aligned} H_0) \quad & \sigma_e = 0 \\ H_1) \quad & \sigma_e > 0 \end{aligned} \tag{3.8}$$

La literatura sobre este tipo de pruebas es amplia y existen diversos procedimientos que permiten extraer conclusiones sobre este planteo. A continuación se exponen las alternativas más comunes.

- Pruebas basadas en el estadístico de Wald.

  Para los parámetros de covarianza existen problemas al utilizar el estadístico de Wald (problema que se acentúa en muestras pequeñas) debido a que, cuando la varianza es cercana a cero, las inferencias realizadas con este procedimiento se encuentran muy cerca del "borde" del espacio paramétrico.

- Pruebas basadas en el estadístico de cociente de verosimilitud (LRT).

  Mientras que el uso de este tipo de estadísticos no plantea mayores problemas cuando se utiliza sobre elementos de β, su uso para contrastar la hipótesis de ausencia de heterogeneidad intra-sujetos ($\sigma_e = 0$), requiere ciertos ajustes para realizar inferencias adecuadas.

  El estadístico de prueba es este que el que se presentó en el apartado de inferencias sobre β sin embargo su distribución es asintóticamente χ^2 solo si se cumplen ciertas condiciones, una de las cuales es que H_0 no se encuentre sobre el "borde" del espacio paramétrico, ya que en dicho caso, este estadístico padece los mismo problemas que el de Wald. Sin embargo, el estadístico LRT es capaz de realizar el contraste indicado en (3.8) si se realiza un ajuste sobre la distribución del estadístico. El procedimiento sugerido por (Stram y Lee, 1994) intenta determinar secuencialmente cuál es la especificación correcta de los efectos aleatorios contenidos en la matriz

D. El procedimiento sugiere llevar a cabo la siguiente prueba sobre una matriz D de dimensiones $(q+k)\times(q+k)$:

$$H_0) D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix}$$

$$H_1) D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$$

De esta manera se plantea que el q-ésimo efecto aleatorio tiene varianza igual a cero (y por ende covarianzas iguales a cero). Como se describió anteriormente se estima el modelo completo y el modelo asumiendo H_0 como cierta, para luego construirse el estadístico de prueba. Pese a que, en primera instancia, se compararía este valor con el valor correspondiente a una distribución χ^2 con tantos grados de libertad como parámetros iguales a cero, estudios de simulación han demostrado que este procedimiento es demasiado conservador. Por lo tanto se propone que la distribución del estadístico no es χ^2 sino que es una mezcla de estas distribuciones, según se indica en (3.9).

$$P(LRT \geq x) = 0{,}5P(\chi^2_q \geq x) + 0{,}5P(\chi^2_{q+k} \geq x) \tag{3.9}$$

También es importante aclarar que, a diferencia del apartado que concierne a la inferencia sobre β, la discusión anterior también es válida utilizando $REML$ en vez de ML, de hecho el estadístico basado en $REML$ presenta niveles de rechazo de H_0 levemente más cercanos al nivel nominal.

3.6. Análisis de datos de sobrevida

El análisis de sobrevida comprende un conjunto de métodos para estudiar la duración hasta que suceda un evento de interés. El caso más común se da en la biología: el fallecimiento. Sin embargo, sus aplicaciones pueden surgir en diversas áreas, algunos ejemplos son el análisis del tiempo de estadía en un hospital, la duración de una huelga, etc. Es más, pese a que el nombre sugiere la intervención del tiempo, en el campo de ensayo de materiales, estas técnicas son útiles para responder preguntas como: ¿Cuánta carga resistirá tal o cual material hasta fallar? La teoría en la que se basa este tipo de análisis asume eventos que se puedan definir adecuadamente en momentos específicos,

en este contexto el fallecimiento constituye un "evento" definido sin ambigüedades en el sentido de que un solo evento sucede en cada individuo. Pese a que hay casos que relajan este supuesto y permiten múltiples "eventos" por individuo, este trabajo dedica especial atención al primer caso.

3.6.1. Definiciones

Antes de introducir los elementos básicos con las que se trabajará en este apartado, es necesario definir la variable aleatoria de interés, a la que se llamará T. En el caso del análisis de sobrevida dicha variable aleatoria debe ser estríctamente positiva, medida a partir de un punto de origen y hasta un final (que se denominará evento) bien determinados y con una cierta escala. Acorde a estos lineamientos es que puede surgir que en algunos estudios, algunos individuos entren al estudio de manera tardía, se pierdan durante el período de seguimiento, o que incluso al final del estudio, aún no hallan experimentado el evento que define su tiempo de sobrevida. Todos estos casos son distintos tipos de censura (el primero a la izquierda y los últimos dos a la derecha) y deben ser tenidos en cuenta de manera adecuada en el análisis.

En este trabajo solo se trabajó con situaciones de censura a la derecha, para hacer frente a esta situación se introdujo la siguiente notación. Sea $T = \text{mín}(T^*, C)$, donde T es el tiempo efectivamente observado, T^* es el tiempo de sobrevida (sea este observado por el investigador o no) y C es el tiempo hasta el momento de la censura. Adicionalmente se introduce la variable δ que indica si el dato de un individuo ha sido observado ($\delta = 1$) o si corresponde a una censura ($\delta = 0$). Estas cantidades son de particular utilidad al construir la función de verosimilitud de una muestra.

El objetivo primordial de estudio en este tipo de análisis es la llamada función de sobrevida:

$$S(t) = P(T > t) \tag{3.10}$$

En la gran mayoría de las aplicaciones se supone que $S(0) = 1$, lo cual indica que la probabilidad de sobrevivir al inicio es cierta. Es trivial notar que si $u > t$ entonces $S(u) \leq S(t)$.

Adicionalmente a la función de sobrevida, también se suele trabajar con la función de riesgo $h(t)$, la cual denota la probabilidad instantánea de que un individuo experimente un evento, condicional a haber sobrevivido t unidades de tiempo.

$$h(t) = \lim_{dt \to \infty} \frac{P(t \leq T < t + dt | T > t)}{dt} = \frac{-S^{'}(t)}{S(t)} \tag{3.11}$$

La función de riesgo (también llamada "fuerza de la mortalidad") es no negativa en todo el recorrido de la variable aleatoria pero a diferencia de la función de sobrevida puede ser creciente, decreciente o ni siquiera ser monótona, de esta manera representa una cantidad mucho mas flexible a la hora de modelar la variable T.
Otra manera de vincular estas dos funciones es la siguiente:

$$S(t) = e^{-\int_{-\infty}^{t} h(s)ds} \tag{3.12}$$

A partir de esta ecuación se define la función de riesgo acumulado $\Lambda(t) = \int^t h(s)ds$. Una última relación que se puede obtener entre $f(t)$, $S(t)$ y $h(t)$ es la siguiente:

$$f(t) = h(t)S(t) \tag{3.13}$$

Utilizando las relaciones indicadas en (3.12)) y (3.13), es posible construir la función de verosimilitud de una muestra aleatoria simple. La contribución a la verosimilitud de un individuo cuyo evento ha sido observado es $f(t_i)$ mientras que en el caso de un individuo cuyo tiempo corresponde a una censura, su contribución es $S(t_i)$. De esta manera, la función de verosimilitud es:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \tag{3.14}$$

que haciendo uso de las relaciones (3.12) y (3.13), se puede expresar del siguiente modo:

$$L = \prod_{i=1}^{n} h(t_i)^{\delta_i} e^{-\int^{t_i} h(s)ds} \tag{3.15}$$

Esta última expresión resulta de mayor practicidad debido a que solo requiere expresar la función de riesgo, la cual (como se mencionó anteriormente) permite mayor flexibilidad que $f(t)$ o $S(t)$.

En los modelos de regresión para datos de sobrevida, el punto de atención se centra en el proceso de envejecimiento al que se someten los objetos que se siguen a través del

tiempo. A diferencia de los modelos convencionales de regresión, donde la componente sistemática modela cambios en la media de la distribución, en los modelos de sobrevida, dicho componente suele ser utilizado para describir la función de riesgo. Adicionalmente se debe tener en cuenta la posible presencia de censura entre los datos. En los siguientes apartados se describen dos metodologías que plantean, desde diferentes perspectivas, modelos capaces de describir variaciones en la función de riesgo.

A continuación se presentan el modelo semiparamétrico de (Cox, 1972) y el modelo paramétrico de Weibull. Estos parten de la base de especificar la función de riesgo del iésimo individuo de la siguiente manera:

$$h(t; x_i) = h_0(t)c(x_i\beta) \tag{3.16}$$

Donde ambas funciones deben ser elegidas de manera que $h(t; x_i) > 0$. Nótese que cuando $c(x\beta) = 1$ entonces $h(t; x_i) = h_0(t)$, por lo cual esta última suele ser llamada función de riesgo de referencia. La característica más sobresaliente de estos modelos (por la cual adoptan su nombre) se basa en el cociente de la función de riesgo de dos individuos. Este cociente, denominado cociente de riesgos (CR) se calcula de la siguiente manera:

$$CR(t; x_1; x_2) = \frac{\lambda_0(t)c(x_1\beta)}{\lambda_0(t)c(x_2\beta)} = \frac{c(x_1\beta)}{c(x_2\beta)} \tag{3.17}$$

aquí se puede ver que el riesgo de un individuo con respecto a otro es una función independiente del tiempo (al menos en el caso de que las covariables sean fijas). Por este motivo suelen llamarse modelos de riesgo proporcional. Ambas metodologías se basan en el supuesto de "riesgos proporcionales" y se diferencian en el tratamiento de la función de riesgo de referencia. Mientras que el modelo Cox deja sin especificar al componente $h_0(t)$, el modelo Weibull lo especifica incluyendo un parámetro más a la distribución.

Modelo semiparamétrico

Tal como se estableció anteriormente, en los modelo de sobrevida, la clave se encuentra en especificar la función de riesgo. De manera más concisa, se puntualizó que el modelo de Cox asume que dicha función adopta la siguiente especificación:

$$h(t; x_i) = h_0(t)e^{x\beta} \tag{3.18}$$

dejando sin especificar la función $h_0(t)$, debido a esta elección en particular, la literatura clasifica a este modelo como "semiparamétrico".

Al calcular el logaritmo de la función de riesgo presentada en (3.18) se puede ver que:

$$\log h(t; x_i) = \log h_0(t) e^{x\beta} = \log h_0(t) + x\beta = \eta(t) + x\beta \tag{3.19}$$

función que sigue las mismas reglas básicas de los modelos lineales. Por este motivo, la codificación de las variables y la inclusión de interacciones se realizan de la misma manera. Y la interpretación de los coeficientes refleja el impacto de aumentar en una unidad cada covariable sobre el logaritmo del riesgo en el momento t. En cuanto al CR, es fácil notar que bajo esta especificación, adopta la siguiente forma:

$$CR(t; x_1, x_2) = e^{\beta(x_1 - x_2)} \tag{3.20}$$

En particular si una de las covariables indica la pertenencia de un individuo a un grupo o tratamiento, e^{β} se interpreta como el riesgo de experimentar el evento de interés con respecto a un individuo que pertenezca al grupo complentario.

Hasta aquí se ha prestado especial atención a la función de riesgo, pero en la práctica, la función de supervivencia juega un rol más importante debido a es más sencilla de interpretar y permite una visualización más clara del fenómeno bajo estudio. Gracias a las fórmulas presentadas en la sección 3.6.1 se puede ver que la función de sobrevida del i-ésimo individuo es:

$$S(t; x_i) = e^{-\int_0^t h_0(u) e^{x\beta}} = e^{-e^{x\beta} \int_0^t h_0(u)} = S_0(t)^{exp(x\beta)} \tag{3.21}$$

Modelo paramétrico

En el apartado anterior se presentó el caso en el que la función de riesgo de referencia se deja sin especificar. Se plantearán ahora las características correspondientes al caso de que la distribución de los tiempos de sobrevida es específica en su totalidad. Este caso merece especial atención debido a que en las ocasiones en que esta familia de modelos proporcionan un buen ajuste, las estimaciones que proporcionana suelen ser más precisas. La representación clásica de estos modelos es mediante la siguiente ecuación:

$$\log T_i = x_i\beta + \sigma\epsilon_i \tag{3.22}$$

donde β y σ son parámetros a estimar y ϵ_i el es error que introduce aleatoriedad al modelo, cuya distribución se asume conocida. Elecciones típicas para esta distribución suelen ser la logística, que implica una distribución log-logística para T y la distribución estandar de valores extremos, la cual acarrea la distribución Weibull para T. Para describir la

función de riesgo del modelo Weibull, existen diversas alternativas en la literatura, la presentada aquí corresponde al caso de definir $\lambda = \frac{1}{\sigma}$, de esta manera:

$$h(t; x_i, \theta, \lambda) = \lambda t^{\lambda-1} e^{-\lambda(\theta_0 + x_i\theta)} \tag{3.23}$$

siendo esta la representación correspondiente a tiempo acelerado. Sin embargo, al definir $\beta = -\lambda\theta$ se obtiene la representación de riesgos proporcionales:

$$h(t; x_i, \theta, \lambda) = \lambda e^{-\lambda\theta_0} t^{\lambda-1} e^{-x_i\beta} \tag{3.24}$$

donde $\lambda e^{\beta_0} t^{\lambda-1}$ no es más que la función de riesgo de referencia. Pese a que σ está asociado a la varianza de la distribución, la literatura se refiere el como "parámetro de escala" mientras que se refiere a λ como parámetro de forma. En última instancia, la función de sobrevida de T adopta la siguiente parametrización:

$$S(t; x_i, \beta, \lambda) = e^{-t^{\lambda} e^{\lambda x_i \beta}} \tag{3.25}$$

Bajo esta parametrización y análogo al significado que β tiene para el cociente de riesgos, e^{θ} representa el cociente de percentiles del tiempo de sobrevida para individuos con distino valor en una covariable.

3.6.2. Inferencia

En cuanto a la estimación de estos modelos, se debe hacer la distinción entre el caso paramétrico y el semiparamétrico debido a que en este último, la dependencia de la verosimilitud de la función de riesgo, plantea un inconveniente importante. Es por esto que Cox planteó la función de verosimilitud parcial como una manera de eludir este obstáculo y estimar los parámetros que modifican la función de riesgo de referencia sin tener que estimar la función propiamente dicha.

Estimación

En el caso del modelo de Cox, la idea detrás del método de verosimilitud parcial está en plantear la probabilidad de la ocurrencia de un evento en el momento t_i dado que alguno de los individuos experimenta dicho evento. Esto es:

$$\mathcal{PL}(\beta) = \prod_{i=1}^{k} \frac{h_0(t_i) e^{x_i\beta}}{\sum_{j \in R(t_i)} h_0(t_j) e^{x_j\beta}} = \prod_{i=1}^{n} \frac{e^{x_i\beta}}{\sum e^{x_i\beta}} \tag{3.26}$$

Donde se debe tener en cuenta que la suma planteada en el denominador abarca a los individuos cuya sobrevida supera el momento t_i.

Finalmente, el vector β es estimado mediante técnicas numéricas de optimización.

El caso del modelo paramétrico (más concretamente el modelo Weibull) la función de verosimilitud es la siguiente:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \underbrace{(\lambda e^{\beta_0} t_i^{\lambda-1} e^{-x_i\beta})}_{h_0(t_i)}^{\delta_i} \underbrace{e^{-t^\lambda e^{\lambda x_i \beta}}}_{S(t_i)} \tag{3.27}$$

que también es maximizada numéricamente.
En ambos casos, la varianza de los estimadores es aproximada utilizando la diagonal de la inversa de la matriz de información de Fisher, denotada por $I(\beta)$ (negativa de la matriz hessiana de la verosimilitud (parcial)).

Inferencia sobre β

A la hora de realizar pruebas de hipótesis o intervalos de confianza sobre las estimaciones de los coeficientes de regresión, se pueden utilizar las mismas herramientas presentadas en el apartado de inferencia de la sección de datos longitudinales. Por lo cual, para llevar a cabo la siguiente prueba de hipótesis:

$H_0) \underline{\beta} = \underline{\beta_0}$
$H_0) \underline{\beta} \neq \underline{\beta_0}$

siendo $\underline{\beta}$ un elemento o un conjunto de elementos del vector β, es válido utilizar estadísticos de Wald, de cociente de verosimilitud o verosimilitud perfil.

Estimación de la función de riesgo de referencia

En el ámbito del modelo de Cox (una vez que se han estimado los coeficientes de regresión) puede surgir interés en estimar la función de riesgo de referencia $h_0(t)$. A partir de dicha estimación es que se construyen curvas de sobrevida posiblemente ajustadas por los valores de las covariables. En este sentido, la literatura describe dos alternativas, el estimador de (Breslow, 1972) y el propuesto por (Kalbfleisch y Prentice, 1973).
El estimador de Breslow surge de maximizar la verosimilitud completa con respecto a $h_0(t)$ remplazando β por su estimación $\hat{\beta}$.

De esta manera el estimador es:

$$\hat{h}_{0,B}(t_i) = \frac{1}{\sum\limits_{j \in R_{(i)}} exp(\hat{\beta} x_i)} \tag{3.28}$$

el cual lleva a la siguiente estimación de la función de sobrevida de referencia.

$$\hat{S}_{0,B}(t_i) = \prod_{i|t_i<t} \left[1 - \frac{1}{\sum\limits_{j \in R_{(i)}} exp(\hat{\beta} x_i)} \right] \tag{3.29}$$

La gran desventaja de este estimador es que puede llegar a adoptar valores negativos. En dichos casos se suele sustituir su valor por cero.

El estimador propuesto por Kalbfleisch y Prentice parte del supuesto de que $S_0(t)$ es una función decreciente que presenta saltos en los tiempos observados $t_1 < t_2 < \ldots, t_n$. Al sustituir $S(t|x)$ por $S_0(t)^{x\beta}$ dentro de la verosimilitud y maximizarla con respecto de dicha cantidad se obtiene el siguiente estimador:

$$\hat{S}_{0,KP}(t_i) = \prod_{i|t_i<t} \left[1 - \frac{exp(\hat{\beta} x_i)}{\sum\limits_{j \in R_{(i)}} exp(\hat{\beta} x_i)} \right]^{exp(-\hat{\beta} x_i)} \tag{3.30}$$

Este estimador tiene la desventaja de que no se puede generalizar al caso de que haya tiempos iguales.

3.7. Datos faltantes

Los estudios longitudinales suelen encontrarse con el problema de datos faltantes (missing data), esto significa que las variables de interés pueden no ser registradas en los individuos en algunas de las ocasiones planificadas según el diseño del estudio. Esto puede suceder por diversos motivos, ya sea que los individuos no son encontrados en alguna ocasión en particular, por eventos adversos (enfermedades), por razones administrativas, falta de cooperación y finalmente, los sujetos pueden abandonar el estudio a partir de un cierto momento, este caso especial de datos faltantes al que la literatura de análisis de datos longitudinales le ha prestado especial atención es la deserción (drop-out). (Verbeke y Molenberghs, 2000) detallan cuatro razones por las cuales se debe recalcar el análisis de la deserción en este tipo de estudios.

1. La clasificación del proceso generador de la falta de datos tiene una interpretación mucho más sencilla que en otras formas de pérdida de datos.

2. Formular modelos en el contexto de la deserción es más sencillo.

3. Gran parte de la literatura referente a datos faltantes en el entorno del análisis de datos longitudinales se restringe a este caso particular.

4. La deserción es, por amplio margen, el caso de dato faltante más común en estudios longitudinales.

3.7.1. Procesos generadores de datos faltantes

En el apartado anterior se mencionó que una de las principales debilidades de los procedimientos allí descritos es que no aprovechan la información contenida en los datos faltantes y su posible relación con la variable bajo estudio, con el fin de ahondar en esta relación resulta necesario introducir los tres mecanismos de datos faltantes mencionados por (Rubin, 1976). Sin embargo, en primer lugar se mencionarán algunas definiciones que permitirán allarar el camino hacia la clasificación de Rubin.

Sea r_i un vector de componentes r_{ij} una variable indicatriz que adopta el valor 1 cuando el individuo i es observado en la ocasión j y 0 cuando no es observado en dicha ocasión. La definición de este vector es necesaria para particionar el vector de observaciones y_i de la forma (y_i^O, y_i^M), siendo y_i^O el sub-vector asociado a las componentes de r_i donde se registran unos (datos observados), mientras que y_i^M es el sub-vector correspondiente a los ceros de r_i (datos faltantes). Se puntualizó que en estudios longitudinales es común que los datos faltantes se den a través de la deserción de los individuos, por esto, también suele definirse la variable d_i que indica el número de mediciones efectivamente observadas del sujeto i.

Para definir la clasificación propiamente dicha es importante aclarar que la misma refiere a la siguiente factorización de la distribución conjunta de r_i y y_i:

$$p(r_i, y_i|x_i, \theta) = p(y_i|x_i, \theta)p(r_i|y_i, x_i, \theta) \tag{3.31}$$

Los mecanismos generadores de datos faltantes se refieren al modelos probabilístico detrás del vínculo entre el vector de datos faltantes r_i y la variable de respuesta y_i.

La idea de la clasificación de estos mecanismos es la factorización de la distribución condicional de r_i dado el vector (y_i^O, y_i^M).

Datos perdidos completamente al azar (Missing Completely at Random (MCAR))

En este caso la factorización de la ecuación (3.31) es la siguiente:

$$p(r_i|y_i^O, y_i^M, \theta) = p(r_i|\theta) \tag{3.32}$$

Este mecanismo asume que la distribución de r_i no guarda relación alguna con la variable de respuesta y_i, definiendo así, una situación de independencia entre los vectores r_i y y_i. La característica más importante de $MCAR$ es que el sub-vector y_i^O puede ser considerado como una muestra aleatoria de los datos completos Y_i, lo cual implica que los datos observados pueden ser considerados como una muestra de la población bajo estudio. De esta manera, el resultado de asumir $MCAR$, es que las inferencias realizadas utilizando los datos disponibles son válidas para toda la población sin tener en cuenta en ningún momento el proceso generador de datos faltantes.

Datos perdidos al azar (Missing at Random (MAR))

En este caso la factorización de la ecuación (3.31) es la siguiente:

$$p(r_i|y_i^O, y_i^M, \theta) = p(r_i|y_i^O, \theta) \tag{3.33}$$

En este caso r_i solo se asume independiente de las mediciones no observadas y_i^M dada la información contenida en y_i^O. En esta factorización, al permitir que r_i dependa de los valores observados y_i^O se está incurriendo en un supuesto menos restrictivo que el postulado bajo $MCAR$. Dado que la única diferencia entre los mecanismos $MCAR$ y MAR es la dependencia de r_i en los valores observados, uno podría poner a prueba la hipótesis de que suponer $MCAR$ es razonable contra la alternativa de que MAR sea el enfoque adecuado (véase (Hedeker y Gibbons, 2006)). Sin embargo, debido al hecho de que el mecanismo de datos faltantes depende de y_i^O, la distribución de y_i^O no es la misma que la de Y_i por lo cual, los datos observados no pueden considerarse una muestra aleatoria de los datos completos y por ende tampoco conforman un muestra aleatoria de la población como si es el caso de $MCAR$. Sin embargo la distribución de los valores faltantes y_i^M condicionada a los valores observados y_i^O concide con su contraparte

poblacional. De esta manera, los valores faltantes pueden ser predichos de una manera válida utilizando los datos observados asumiendo un modelo correctamente especificado para el vector (y_i^O, y_i^M). De esta manera surge que los análisis llevados a cabo en un contexto MAR pueden proveer inferencias válidas aun si se ignora la contribución a la verosimilitud de r_i.

Datos perdidos no al azar (Missing not at Random (MNAR))

Finalmente, cuando la distribución condicional de r_i también incluye a y_i^M (o al menos alguno de sus elementos) el mecanismo es llamado $MNAR$. Al igual que en el caso MAR, bajo $MNAR$ los datos observados no constituyen una muestra de la población objetivo. Por otro lado, al contrario de MAR, la distribución predictiva de y_i^M condicional a y_i^O no coincide con la poblacional ya que adicionalmente depende de $p(y_i\,|r_i)$. Por este motivo, la correcta especificación del proceso de pérdida de datos es crucial y debe ser incluida en la verosimilitud. De esta manera resulta claro que el mecanismo de pérdida más complejo para trabajar es $MNAR$, sin embargo no plantea supuestos restrictivos ni poco reales. Cuando los datos longitudinales surgen de un mecanismo $MNAR$, la validez del proceso inferencial está ligada a que se modeló adecuadamente la distribución conjunta de y_i y r_i. En la literatura se pueden distringuir tres tipos de modelos.

1. Modelos de selección.

   Esta familia de modelos se caracteriza por realizar la siguiente factorización de (3.31):

   $$p(y_i, r_i|\theta) = p(y_i|\theta_y)p(r_i|y_i, \theta_r) \tag{3.34}$$

   Estos modelos fueron introducidos por (Heckman, 1976) en la literatura econométrica y su nomenclatura se basa en que mediante la distribución de r_i condicional a la trayectoria descrita por y_i se puede pensar que cada individuo selecciona probabilísticamente si deserta del estudio o si continúa en él. Su uso en el análisis de datos longitudinales se debe principalmente, al trabajo de Diggle y Kenward (1994).

2. Modelos de mezcla de patrones.

   En estos modelos propuestos por (Little, 1993), la distribución conjunta de y_i y r_i adopta la siguiente forma:

   $$p(y_i, r_i|\theta) = p(y_i|r_i, \theta_y)p(r_i|\theta_r) \tag{3.35}$$

Como puede verse, estos modelos se basan en la factorización opuesta a la de los "modelos de selección". Tal como lo indica su nombre, estos modelos permiten modelar la distribución de y_i de distinta manera en cada patrón de datos faltantes. En el caso de que los datos faltantes solo sean deserciones, cada patrón indicaría el momento de la deserción de cada individuo. De esta manera, la distribución marginal de y_i corresponde a una mezcla probabilística con pesos dados por la distribución marginal de cada patrón.

3. Modelos de parámetros compartidos.

   Por último, los "modelos de parámetros compartidos" (véase Wu y Bailey, 1988) se basan en la idea de que existe un proceso latente (descrito a través de efectos aleatorios) que dictamina los valores observados en y_i y r_i.

$$p(y_i, r_i|\theta) = \int p(y_i|b_i, \theta_y)p(r_i|b_i, \theta_r)p(b_i\theta_b)db_i \tag{3.36}$$

   Así, para un valor dado de los efectos aleatorios, los procesos de pérdida y de medición se consideran independientes.

3.8. Análisis conjunto de datos longitudinales y de sobrevida

En el capítulo Antecedentes se mencionaron varias investigaciones donde se generaron tanto datos longitudinales (mediante la repetición de ciertas mediciones) como datos de tiempo hasta cierto evento. La gran mayoría del trabajo mencionado en dicho apartado se basa en casos particulares donde metodologías específicas se desarrollaron para cada caso debido a que cada uno de ellos perseguía objetivos ligeramente diferentes. Sin emabrgo (Henderson et al., 2000) resume los objetivos de estos y otros estudios en tres categorías:

1. Ajustar las inferencias referentes al proceso longitudinal de manera de permitir una posible dependencia del proceso de sobrevida.

2. Ajustar la distribución hasta el tiempo de falla en función de las variaciones del proceso longitudinal.

3. Caracterizar la evolución conjunta del ambos procesos.

La metodología que se presenta en este apartado trata de cumplir con estos tres objetivos partiendo de la base de la maximización de la verosimilitud conjunta entre los dos procesos involucrados. Esta estrategia de estudio resulta mucho más eficiente a la hora de llevar a cabo los análisis ya que los datos del proceso longitudinal se usan simultáneamente con los del proceso de sobrevida. En este sentido es de esperar que se obtengan estimaciones más precisas de la intensidad de la asociación entre ambos procesos.

3.8.1. Formulación del modelo

Tal como se mencionó anteriormente, la idea principal detrás de los modelos conjuntos, es la de acoplar el modelo de sobrevida con el modelo longitudinal. El objetivo específico de esta familia de modelos es al de medir la asociación entre el nivel de la variable registrada longitudinalmente (libre de ruido) con el riesgo de experimentar el evento de interés en cada momento del tiempo, teniendo en cuenta al mismo tiempo, la posible influencia de un conjunto de covariables. Para una mejor descripción matemática, (Rizopoulos, 2012) propone utilizar la siguiente denominación del modelo longitudinal:

$$y_i(t) = m_i(t) + \epsilon_i(t) \tag{3.37}$$

Donde $m_i(t)$ es la estimación del nivel de la variable longitudinal en el momento t para el iésimo individuo y $\epsilon_i(t)$ es un componente de ruido que se desea remover de la medición $y_i(t)$. El punto de partida del modelo conjuto serán los modelos presentados en los capítulos 3.5 y 3.6, por lo cual este capítulo utilizará la misma notación. La información longitudinal es recolectada intermitentemente y con un componente de error. Por este motivo, la tarea del submodelo longitudinal consiste en estimar el valor del proceso libre de ruido $m_i(t)$ para cualquier valor de t. Con este fin se especifican los componentes del modelo longitudinal presentado en (3.37) presentado en el capítulo 3.5:

$$\begin{cases} y_i(t) = m_i(t) + \epsilon_i(t) \\ m_i(t) = x_{Li}(t)\beta_L + z_i(t)b_i \\ b_i \sim N\left(0, D\right) \\ \epsilon(t) \sim N\left(0, \sigma^2\right) \end{cases} \tag{3.38}$$

Este sencillo modelo longitudinal es capaz de separar el ruido incluido en las mediciones del verdadero valor el proceso y reconstruir completamente la evolución de $m_i(t)$ gracias a la parametrización temporal impuesta en las covariables $x(t)$ y $z(t)$.

Por otro lado, la parametrización del submodelo de sobrevida incluirá el historial del proceso $m(t)$ de la siguiente manera:

$$h_i(t|m_i(t), x_{Si}) = h_0(t)e^{x_{Si}\beta_S + \alpha m_i(t)} \tag{3.39}$$

Nótese cómo se optó por diferenciar las covaraibles x_S y x_L. Las primeras refieren a las involucradas en el submodelo longitudinal mientras que las segundas son las intervinientes en el submodelo de sobrevida. La misma convención se utilizó en el vector de efectos fijos β_S y β_L.

Bajo esta nueva especificación, el parámetro α cuantifica el efecto del verdadero valor (no observado) del proceso longitudinal sobre el riesgo de experimentar el evento en el momento t. Debe notarse como bajo esta parametrización, el riesgo de experimentar el evento solo depende del valor actual de $m(t)$. Sin embargo esto no se cumple para la función de sobrevida, véase que en dicho caso:

$$S_i(t|m_i(t), x_{Si}) = e^{\int_0^t h_0(s)exp\{x_{Si}(s)\beta_S + \alpha m_i(s)\}\, ds} \tag{3.40}$$

lo que implica que la función de sobrevida depende de toda la historia de $m(t)$. Bajo esta especificación, solo la función de riesgo depende del valor de $m(t)$ en el instante t, pero este supuesto puede no ser muy realista. Tenga en cuenta el caso en el que el riesgo del evento no solo depende del valor puntual de $m(t)$ sino también de la tendencia experimentada por este último o por el historial completo de las fluctuaciones experimentadas por este. En dichos casos, es sensato pensar en la siguiente parametrización del submodelo de sobrevida:

$$\begin{aligned} h_i(t|m_i(t), x_{Si}) &= h_0(t)e^{x_{Si}\beta_S + \alpha_0 m_i(t) + \alpha_1 f_1(m_i(t)) + \alpha_2 f_2(m_i(t))} \\ f_1(m_i(t)) &= m_i^{'}(t) \\ f_2(m_i(t)) &= \textstyle\int^t m_i(s)ds \end{aligned} \tag{3.41}$$

De incluir estos términos, los coeficientes α_1 y α_2 miden la intensidad de la asociación entre el riesgo y la pendiente/historial de la verdadera trayectoria longitudinal.

Finalmente se debe agregar que en el contexto del modelo conjunto, no es conveniente dejar sin especificar la función de riesgo de base. Esto se debe a que, como notó (Hsieh et al., 2006), esta elección puede acarrear una subestimación de los errores estándar de las estimaciones, que puede llevar a inferencias equívocas. Por este motivo, en el caso de que

el submodelo de sobrevida no tenga una distribución específica (por ejemplo Weibull) es preferible especificar algún tipo de parametrización alternativa para $h_0(t)$. La más comunmente utilizada es la siguiente función de riesgo en escalera (RE):

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I\left(v_{q-1} < t_q\right) \tag{3.42}$$

donde $0 = v_0 < v_1 < v_2 < \ldots, v_Q$ es un conjunto de nodos (tipicamente especificados por el investigador o equiespaciados entre el menor y el mayor valor de los eventos) que dividen la escala temporal y ξ_q indica el valor de la función de riesgo en el intervalo $(v_{q-1}, v_q]$. Nótese como al incrementar el número de nodos, aumenta la flexibilidad de esta función.

3.8.2. Estimación

El método de máxima verosimilitud[2] consiste en la búsqueda del modo de la log-verosimilitud de la distribución conjunta de $\{Y_i, T_i\}$. Para definir esta distribución conjunta (Rizopoulos, 2012) asume que el vector de efectos aleatorios b_i está presente tanto en el proceso longitudinal como en el de sobrevida. De este modo, al asumir que los dos procesos involucrados se acoplan mediante los efectos aleatorios, se postula el supuesto de independencia de las variables Y_i y T_i condicional al valor de b_i. Así, la distribución conjunta se factoriza de la siguiente manera:

$$\begin{aligned} \mathrm{p}(\mathrm{T}_i, Y_i | b_i; \theta) &= p(T_i, | b_i; \theta) p(Y_i | b_i; \theta) \\ &= p(T_i, | b_i; \theta) \prod_{j=1}^{n_i} p(y_i(t_{ij}) | b_i; \theta) \end{aligned} \tag{3.43}$$

Bajo la formulación (3.43) se asume que condicional a los datos, la censura y los tiempos de visita son independientes de los tiempos observados y de las futuras mediciones del proceso longitudinal. A efectos prácticos esto significa que el hecho de que un sujeto abandone el estudio depende solo de su "historial" en el estudio.

2. Debido a que el modelo conjunto contiene un submodelo de riesgo proporcional, el método puede ser semiparamétrico (en el caso del modelo de Cox) o paramétrico (en el caso Weibull).

La contribución del i-ésimo sujeto a la log-verosimilitud será:

$$\begin{aligned}\log p(T_i, Y_i; \theta) &= \log \int p(T_i, |b_i; \theta) p(Y_i | b_i; \theta) db_i \\ &= \log p(T_i, |b_i; \theta_t) \left[\prod_{j=1}^{n_i} p(y_i(t_{ij}) | b_i; \theta_y) \right] p(b_i | \theta_b) db_i \end{aligned} \tag{3.44}$$

donde θ_t son los parámetros de sobrevida, θ_y son los parámetros de efectos fijos del sub-modelo longitudinal y θ_b son los parámetros de covarianza del sub-modelo longitudinal. En última instancia se definen cada uno de los componentes de la ecuación (3.44) de la siguiente manera:

$$\begin{aligned} p(T_i, |b_i; \theta_t) &= [h_0(T_i) exp\{x_{Si}\beta_S + \alpha m_i(T_i)\}]^{\delta_i} \\ &\quad \times \exp\left(\int_0^{T_i} h_0(s) \exp\{x_{Si}\beta_S + \alpha m_i(s)\} ds \right) \\ \prod_{j=1}^{n_i} p(y_i(t_{ij}) | b_i; \theta_y) &= (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp\left\{ -\frac{\sum_j (y_i(t_{ij}) - x_{Li}(t_{ij})\beta_L - z_i(t_{ij}) b_i)^2}{2\sigma^2} \right\} \\ p(b_i | \theta_b) &= (2\pi)^{-\frac{q_b}{2}} |D|^{-\frac{1}{2}} \exp\left\{ -b_i' D^{-1} b_i \right\} \end{aligned} \tag{3.45}$$

La maximización de la función de log-verosimilitud (3.44) se realiza mediante métodos numéricos, los usados más frecuentemente son el algoritmo EM (Dempster et al., 1977) o variantes del método de Newton-Raphson (Nocedal y Wright, 2006). Sin embargo, el primero es el más utilizado, dado que convenientemente trata a los efectos aleatorios como "datos faltantes" y provee predicciones de los mismos como un resultado adicional del proceso de estiamción y que en el paso M, algunos de los estimadores tienen una expresión cerrada.

3.8.3. Inferencia

Pruebas de hipótesis

Una vez estimado el modelo mediante máxima verosimilitud, los mismos procedimientos descritos en las secciones 3.5.1 y 3.6.2 se pueden aplicar, tanto el estadístico LRT como las pruebas de Wald se aplican del mismo modo, ya sea re-estimando el modelo bajo las restricciones que se deseen poner a prueba y comparando los valores de la verosimilitud (en el caso LRT) o comparando las estimaciones de los coeficientes de regresión con sus errores estandar (en el contexto del estadístico de Wald). El único parámetro extra que aparece en estos modelos es α. Tal vez la hipótesis más importante a contrastar sea:

H_0) $\alpha = 0$

H_1) $\alpha \neq 0$

donde la hipótesis nula implica la ausencia de asociación entre los procesos longitudinal y de sobrevida.

El caso de los elementos de la matriz D (parámetros correspondientes a los efecos aleatorios), los estadísticos LRT se utilizan con los mismos ciudados descritos en la sección 3.5.1.

Intervalos de confianza

Los intervalos de confianza de los parámetros del modelo conjunto (β_S y β_L) son calculados mediante la aproximación asintótica de los estimadores a la distribución normal en la que se basa el estadístico de Wald. De esta manera, los intervalos se construyen utilizando las estimaciones y los errores estándar ($\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \hat{s.e.}(\hat{\theta})$). Del mismo modo se pueden construir intervalos para la media del proceso longitudinal, es decir:

$$X\hat{\beta} \pm z_{1-\frac{\alpha}{2}} \left\{ diag \left[X \hat{var}(\hat{\beta}) X' \right] \right\}^{\frac{1}{2}} \tag{3.46}$$

Predicción de efectos aleatorios

Hasta este punto nos hemos enfocado en la ralización de inferencias sobre los efectos fijos del modelo conjunto. Pese a que los efectos aleatorios b_i fueron introducidos al modelo con el propósito de tener en cuenta la heterogeneidad entre los perfiles individuales de los individuos, puede que el énfasis del estudio esté sobre la predicción individual de futuras mediciones del proceso longitudinal, para lo cual la predicción de los efectos aleatorios es de vital importancia. Debido al carácter aleatorio de estas cantidades, lo más natural es predecirlas bajo un enfoque bayesiano (Rizopoulos, 2012, cap. 4.5). De esta manera, la distribución a posteriori de los efectos aleatorios será:

$$P(b_i|T_i, Y_i; \theta) = \frac{p(T_i|b_i; \theta) p(Y_i|b_i; \theta) p(b_i; \theta)}{p(T_i, Y_i; \theta)} \tag{3.47}$$

Contrario a lo que sucede en el caso de modelos mixtos gaussianos, la distribución en (3.47) no pertenece a la familia gaussiana y ni siquiera posee una "forma cerrada" por lo cual, su estudio es realizado mediante métodos numéricos. Sin embargo debe realizarse la aclaración que en el caso de que el número de mediciones por individuo (n_i) tiende a infinito, esta distribución a posteriori tiende a una distribución normal.

Típicamente, suele caracterizarse esta distribución o bien mediante su media o mediante su modo.

$$\begin{aligned} \bar{b_i} &= \int b_i p(b_i|T_i, Y_i; \theta) db_i \\ \hat{b_i} &= \arg\max_b \log p(b_i|T_i, Y_i; \theta) \end{aligned} \tag{3.48}$$

La elección del modo frente a la media responde a lo indicado anteriormente, debido a que la distribución a posteriori de los efectos aleatorios no necesariamente es normal, es posible que presente asimetrías, por lo cual, la media puede no resultar en la mejor medida de resumen. En cuanto a la variabilidad de estas mediciones, es común aproximar su varianza mediante las fórmulas indicadas en la ecuación (3.49).

$$\begin{aligned} var(b_i) &= \int (b_i - \bar{b_i})^2 p(b_i|T_i, Y_i; \theta) db_i \\ H_i &= \left\{ -\frac{\partial^2 \log p(b_i|T_i, Y_i; \theta)}{\partial b' \partial b}|_{b=\hat{b_i}} \right\}^{-1} \end{aligned} \tag{3.49}$$

Tanto para el cálculo de las cantidades en (3.48) como en (3.49) se sustituye a θ por $\hat{\theta}$.

3.8.4. Utilidad en el caso de datos faltantes

El investigador supone que la ocurrencia de un evento implica la discontinuación o al menos un cambio en la distribución de la variable longitudinal. Debido a este enunciado es que se pueden conectar las ideas sobre mecanismos generadores de datos faltantes de la sección 3.7.1 con la teoría detrás del modelo conjunto.

Es importante notar que dada la formulación del modelo conjunto, el submodelo mixto asume una distribución para todas las posibles mediciones de la variable medida repetidamente, esto significa que (al menos en principio) la distribución del vector compelto Y_i es válida tanto para los valores observados como para los faltantes. En la sección 3.7.1 se descompuso al vector Y_i en dos componentes, y_i^O y y_i^M. En este caso en particular $y_i^O = \{y_i(t_{ij}) : t_{ij} < T_i^*, j = 1, 2, \ldots, n_i\}$ representa las mediciones efectivamente observadas del individuo i, mientras que $y_i^M = \{y_i(t_{ij}) : t_{ij} \geq T_i^*, j = 1, 2, \ldots, n_i\}$ contiene las mediciones que se habrían observado del mismo individuo si no hubiese experimentado el evento. Bajo estas definiciones se puede explicitar el mecanisno de pérdida de datos, como la distribución de la variable T^* condicional a los elementos y_i^O y y_i^M.

$$\begin{aligned} p(T_i^*|y_i^O, y_i^M; \theta) &= \int p(T_i^*, b_i|y_i^O, y_i^M; \theta) db_i \\ &= \int p(T_i^*|b_i, y_i^O, y_i^M; \theta) p(b_i|y_i^O, y_i^M; \theta)) db_i \\ &= \int p(T_i^*|b_i; \theta) p(b_i|y_i^O, y_i^M; \theta)) db_i \end{aligned} \tag{3.50}$$

Donde la simplificación del último paso se debe al supuesto de independencia condicional entre T y Y (véase el apartado de estimación). Finalmente, se puede observar cómo el tiempo hasta el evento depende tanto de y_i^O como de y_i^M a través de la distribución a posteriori de los efectos aleatorios. Por este motivo, el modelo conjunto pertenece a la familia de modelos $MNAR$.

3.8.5. Análisis de sensibilidad

Uno de los problemas más importantes a la hora de evaluar los resultados del modelo conjunto reside en el hecho de que no es posible poner a prueba el supuesto MAR respecto de $MNAR$. En este sentido, la única manera práctica de evaluar los supuestos en este ámbito es a través de un análisis de sensibilidad (Diggle et al., 2007).
En primer lugar, debe hacerse mención a los distintos tipos de análisis de sensibilidad que pueden llevarse a cabo. En el análisis de sensibilidad global se investiga una amplia clase de modelos donde la idea está en determinar cuánto hay que alejarse de los supuestos del modelo bajo estudio para que las inferencias cambien. Por otro lado, el análisis de sensibilidad local se basa en evaluar el cambio en las inferencias en un "vecindario" cercano a los supuestos del modelo bajo estudio (Daniels, 2008).

Por lo general, el primer paso a la hora de llevar a cabo este análisis es comparar las estimaciones del modelo $MNAR$ con su equivalente MAR (sensibilidad global), que en este contexto implica comparar los resultados del modelo conjunto con el modelo mixto. Luego, la propuesta consiste en evaluar la estimación del vector β_L mediante un indicador de sensibilidad (local) al supuesto de ignorabilidad, este indicador es el $ISNI$ por su sigla en inglés. El propósito de este índice es cuantificar que tanto varían las estimaciones del submodelo longitudinal al alejarse de la hipótesis de ignorabilidad del mecanismo de pérdida. La idea detrás del $ISNI$ es medir el cambio en los parámetros al alejarse del supuesto de MAR ($\alpha = 0$), esto es:

$$ISNI = \frac{\partial}{\partial \alpha} \hat{\beta(\alpha)} \mid_{\alpha=0} \quad (3.51)$$

En (Rizopoulos, 2012) se aclara que al no poder calcularse analíticamente el valor exacto del $ISNI$, se propone una aproximación de segundo orden de la log-verosimilitud del modelo conjunto. A partir de la misma se deriva el siguiente estimador:

$$ISNI \approx -\left\{ \frac{\partial^2}{\partial \beta^T \partial \beta} \ell(\theta)|_{\theta=\theta^{(0)}} \right\}^{-1} \left\{ \frac{\partial^2}{\partial \beta^T \partial \alpha} \ell(\theta)|_{\theta=\theta^{(0)}} \right\} \quad (3.52)$$

Siendo $\theta^{(0)}$ el vector de todos los parámetros del modelo conjunto estimado bajo el supuesto de MAR. De esta manera, el cálculo indicado en (3.52) requiere la hessiana del modelo bajo $\alpha = 0$[3]. Una vez calculado el indicador, algunos autores poponen normalizarlo con respecto a la magnitud de la estimación o con respecto a la estimación del desvío del parámetro correspondiente. De esta manera se logran $ISNI$s relativos que permiten evaluar sensibilidad sin tener en cuenta la unidad de medida de las variables. De esta manera Troxel considera que valores del $INSI$ (relativo al desvío estándar) mayores que uno, indican cierta sensibilidad al mecanismo de pérdida. La otra alternativa, propuesta por (Viviani, 2012) considera que valores mayores a 0.5 indician cierta sensibilidad en el parámetro.

3.9. Aplicación a un estudio longitudinal: "Origins of Variance in the Oldest-Old: Octogenarian Twins" ($OCTO-Twin$)

En este trabajo se presenta una aplicación sobre los datos del estidio "Origins of Variance in the Oldest-Old: Octogenarian Twins" ($OCTO - Twin$). Este se compone de 351 parejas de mellizos, de 80 años o más seleccionados del registro sueco de mellizos. El estudio tenía como propósito investigar las causas de las diferencias individuales en este conjunto de personas ancianas en características que iban desde salud, capacidad funcional, funcionamiento cognitivo y bienestar psicológico entre otras.

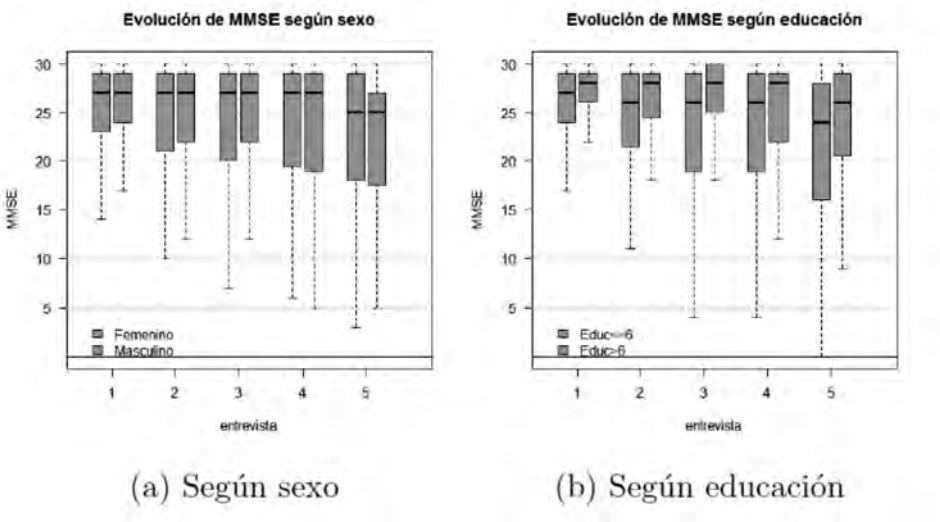
El estudio comenzó en 1991 y la cohorte involucrada fue estudiada hasta 1999. En dicho período se realizaron entrevistas cada 2 años aproximadamente. Aquí se presentan datos donde cada participante presenta información parcial o completa sobre su estado cognitivo. Esta información fue relacionada con el sexo, la escolaridad y la edad de los individuos con el fin de determinar diferentes patrones de deterioro cognitvo.

3.10. Análisis exploratorio inicial

Al inicio del estudio, la muestra fue de 702 individuos (234 hombres y 468 mujeres) con edades entre 79 y 98 años con una mediana de 86 años.

3. En el caso del que el modelo incluya tanto un término con α_1 como otro con α_2, se calculó la hessiana bajo la restricción $\alpha_1 = \alpha_2 = 0$.

En cuanto a la educación, se detectó que la mayoría tenía 6 años de educación, por lo que se optó por dividir la muestra en 2 grupos; “6 años de educación o menos” y “más de 6 años”. Como parte de el análisis descriptivo se investigó la posible asociación del



(a) Según sexo (b) Según educación

Figura 3.1: Distribución del MMSE

deterioro cognitivo de estas personas con el sexo y el nivel educativo. Se construyeron gráficos de caja para cada visita diferenciando cada subpoblacion, (figuras 3.1).

En cuanto al mecanismo de pérdida de datos, en la figura 3.2 se puede observar como el paso del tiempo (medido en el número de visitas por sujeto) tuvo un impacto considerable en la información disponible.

Para finalizar la descripción inicial de los datos se presenta la tabla 3.1, el cual contiene el promedio de $MMSE$ de los individuos agrupados según el número total de visitas en cada una de las evaluaciones que se les realizaron. Así mismo se expone la edad promedio de cada grupo al inicio del estudio para facilitar la comparación. En la tabla se puede ver que pese a que las personas que solo lograron ser evaluadas una vez eran las más jóvenes, su $MMSE$ era el más bajo, por lo cual ya presentaban un deterioro avanzado.

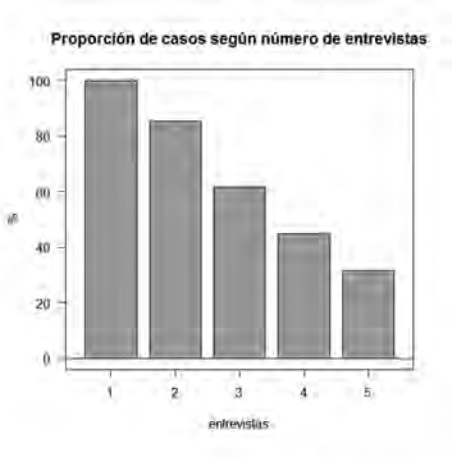


Figura 3.2: Disminución del tamaño muestral según número de entrevistas

	Total de visitas				
	1	2	3	4	5
primera	19.98	23.80	24.60	24.29	27.10
segunda	–	19.36	22.36	23.51	26.79
tercera	–	–	18.92	20.95	26.35
cuarta	–	–	–	17.10	24.95
quinta	–	–	–	–	22.00
Edad	84.62	85.03	85.60	86.66	86.82
n	102	166	119	93	222

Tabla 3.1: Estimaciones e indicadores de modelos longitudinales

3.11. Estrategia de Análisis y Estimación

La estimación del modelo conjunto requiere de tres insumos fundamentales que son la especificación de un modelo longitudinal que describa la trayectoria de la medida bajo estudio (en este caso el $MMSE$), un modelo de sobrevida que ajuste la probabilidad de deserción del estudio de cada individuo (en este caso debido al fallecimiento) y un proceso que vincule ambos modelos. Para la elaboración de este modelo conjunto, se optó por seguir los siguientes pasos:

1. Construcción del modelo longitudinal.
   En esta etapa se planteó la construcción del un modelo que describa el crecimiento (o decrecimiento) del $MMSE$ a través de una trayectoria lineal o cuadrática medida a partir del comienzo del estudio y ajustó por la edad en dicho momento.

2. Construcción de un modelo de sobrevida.
   En esta etapa se trató de elaborar un modelo para los tiempos de permanencia en el estudio de los individuos. El evento a modelar fue el fallecimiento, considerando como "censurados" a aquellos casos correspondientes a los individuos que sobrevivieron todo el período de seguimiento.

3. Construcción del modelo conjunto.
   La última instancia fue la correspondiente a la estimación conjunta, que permitió establecer los determinantes de la velocidad del deterioro cognitivo, teniendo en cuenta los factores que incidían sobre la deserción de los individuos. Finalmente se llevó a cabo un análisis de sensibilidad de los parámetros ante los supuestos de Missing at Random y Missing not at Random

Antes de finalizar el apartado se deben realizar dos consideraciones. Se aclara que el tiempo se midió en décadas a partir del inicio de estudio, ya que de esta forma las estimaciones de las varianzas de los efectos aleatorios se alejaron del borde del espacio paramétrico. La segunda consideración, es que dado que en algunas instancias no se contaba con el dato pertinente al $MMSE$ o a la educación de los participantes, se debió reducir el tamaño muestral. Finalmente se trabajó con 2125 ocasiones relevadas sobre 688 individuos. El análisis de los datos fue llevado a cabo en el software de uso libre R (R Core Team, 2013).

3.12. Resultados

3.12.1. Estimación del submodelo longitudinal

El modelo de partida utilizado para modelar la variación temporal del $MMSE$ fue el siguiente:

$$\begin{aligned}
MMSE_{ij} &= \beta_{0i} + \beta_{1i}t_j + \beta_{2i}t_j^2 + \epsilon_{ij} \qquad j = 1, \ldots, n_i \qquad i = 1, \ldots, n \\
\beta_{0i} &= \beta_{00} + \beta_{01}Sexo_i + \beta_{02}(Edad_i - 82) + \beta_{03}Educ_i + b_{0i} \\
\beta_{1i} &= \beta_{10} + \beta_{11}Sexo_i + \beta_{12}(Edad_i - 82) + \beta_{13}Educ_i + b_{1i} \\
\beta_{2i} &= \beta_{20} + \beta_{21}Sexo_i + \beta_{22}(Edad_i - 82) + \beta_{23}Educ_i + b_{2i} \\
\underline{b_i} &\sim N_3(0, D) \\
\epsilon_{ij} &\sim N(0, \sigma^2)
\end{aligned}$$

Donde $(Edad - 82)$ es el valor centrado de la covariable que indica la edad del individuo al momento del inicio del estudio. De esta manera β_{00} indica el valor de $MMSE$ de una mujer de 82 años con 6 años de educación o menos. A partir de este modelo de base (modelo 1) se ensayaron distintas alternativas. La estimación se llevó a cabo a través de la librería nlme (Pinheiro et al., 2013). En la tabla ?? se presentan los resultados de modelos con efectos aleatorios independientes (modelo 2), un modelo sin la incidencia del sexo (modelo 3), un modelo sin efectos fijos en el término cuadrático y el modelo final (modelo 4).

El cuanto a los efectos aleatorios, el modelo 1 especificó una estructura general, con varianzas distintas para cada efecto y covarianzas libres. Luego se testeó la posibilidad de que las covarianzas fueran iguales a cero (comparando las verosimilitudes de los modelos 1 y 2) comprobándose de que el supuesto de efectos aleatorios independientes era demasiado restrictivo. Luego se puso a prueba el efecto aleatorio correspondiente al término cuadrático era significativo, rechazando la hipótesis de varianza nula. Para esta prueba se comparó el valor del estadistico LRT con el cuantil 95 de una distribución $\chi^2_{2:3}$[4].
Finalmente, se optó por trabajar con el modelo 4 ya que este poseía una estructura de efectos aleatorios adecuada, efectos fijos significativos y menores valores de BIC y AIC. La interpretación de sus coeficientes indican que al entrar al estudio, una persona de 82 años (sin importar su sexo) con menos de 6 años de educación tiene un valor promedio de $MMSE$ de 25.47 puntos, con más de 6 años tendría 26.75 puntos.

4. Se denota como $\chi^2_{2:3}$ a la mezcla de dos distribuciones χ^2 con 2 y 3 grados de libertad respectivamente.

		Modelo 1	Modelo 2	Modelo 3	Modelo 4
efecto	parámetro	EMV	EMV	EMV	EMV
Constante					
constante	β_{00}	25,27**	25,56**	25,52**	25,47**
Sexo Masculino	β_{01}	-0.15	-0.16		
(Edad - 82)	β_{02}	−0,26**	−0,26**	−0,26**	−0,24**
Educ>6	β_{03}	1,28**	1,30**	1,28**	1,28**
Tiempo					
constante	β_{10}	−6,41**	−5,81**	−6,26**	−3,58*
Sexo Masculino	β_{11}	0.37	1.39		
(Edad - 82)	β_{12}	−0,61*	-0.44	−0,60*	−0,75**
Educ>6	β_{13}	3.72	3.51	3.67	2,89*
Tiempo^2					
constante	β_{20}	9,12**	7,27*	7,34*	
Sexo Masculino	β_{21}	-4.02	−4,70*		
Edad - 82	β_{22}	-0.65	−0,88**	-0.60	
Educ>6	β_{23}	-1.40	-0.61	-1.24	
Covarianza de b_i					
var(b_{i1})	d_{11}	17.04	20.93	17.04	16.99
var(b_{i2})	d_{22}	297.45	181.55	297.17	297.89
var(b_{i3})	d_{33}	205.05	24.40	206.76	200.32
cor(b_{i1},b_{i2})	$\frac{d_{12}}{\sqrt{d_{11}d_{22}}}$	0.62	0	0.62	0.61
cor(b_{i1},b_{i3})	$\frac{d_{13}}{\sqrt{d_{11}d_{33}}}$	-0.52	0	-0.53	-0.50
cor(b_{i2},b_{i3})	$\frac{d_{23}}{\sqrt{d_{11}d_{33}}}$	-0.75	0	-0.75	-0.74
Varianza residual					
var(ϵ_{ij})	σ^2	6,81	7.13	7.18	7.23
$L(y\|\theta)$		-6269.28	-6324.0	-6271.80	-6274.28
AIC		12576.56	12680.00	12575.61	12574.55
BIC		12684.15	12770.60	12666.22	12648.17

Tabla 3.2: Estimaciones e indicadores de modelos longitudinales

* significativo al 5 % ** significativo al 1 %

Personas mayores de 82 años comienzan el estudio con 0.24 puntos menos por cada año que supere los 82 y 0.24 mas por cada año menos.
Al comparar personas de baja y alta educación (sin tener en cuenta el sexo ni la edad inicial) se ve que el $MMSE$ de los de baja educación desciende 3,58 puntos por década, mientras que en los de mayor educación, este puntaje solo se decrementa 0,69 unidades. En cuanto a la edad al inicio, se nota que cuanto mayor es la persona, menor es su puntaje inicial de $MMSE$ y más pronunciada es la pendiente que indica la velocidad de su deterioro (parámetro asociado β_{12}).
En cuanto a los elementos de la matriz D se puede ver que cuando el efecto aleatorio de la constante es elevado, la caída es más pronunciada (correlaciones negativas con los efectos asociados a la pendiente y al término cuadrático). En la figura 3.3 se representan las trayectorias del $MMSE$ de personas con diferentes edades y años de educación.

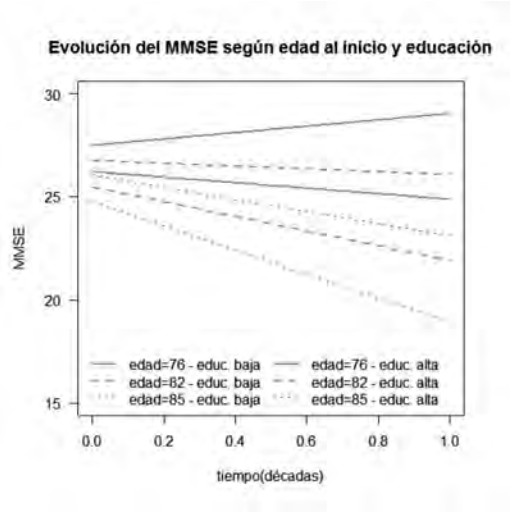


Figura 3.3: Evolución del $MMSE$ según modelo 4

3.12.2. Estimación del sub-modelo de sobrevida

En esta etapa se analizó el componente que describe el tiempo transcurrido por cada individuo hasta que se dejan de observar sus valores de $MMSE$, ya sea debido al fallecimeinto o a la censura. La figura 3.4 muestra la sobrevida de general de todos los inidividuos. Tanto esta figura, como todos los análisis llevados a cabo en este apartado fueron realizados utilizando la librería survival (Therneau, 2012).

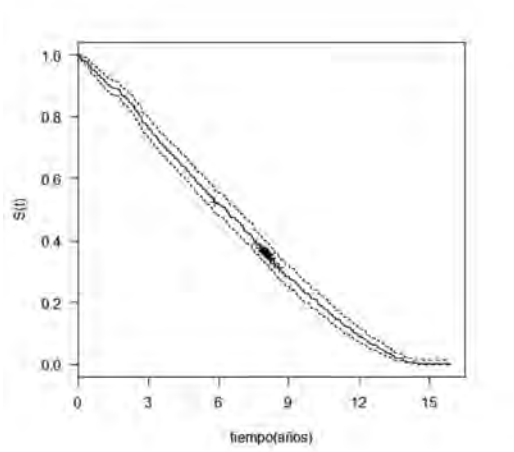


Figura 3.4: Sobrevida de los individuos

Luego se procedió a determinar si existían diferentes comportamientos en cuanto a la sobrevida entre los grupos generados por sexo y educación. Se llegó a la conclusión de que hombres y mujeres presentan diferencias en cuanto al tiempo hasta el fallecimiento sin encontrarse diferencias entre los 2 grupos educativos (véase tabla 3.3).

grupos	χ^2	g.l.	p-valor
Sexo	11.7	1	<0.001
educación	0.2	1	0.622

Tabla 3.3: Exploración de diferencias en $S(t)$ mediante la rueba del rango logarítmico

Luego se estimaron distintos modelos de sobrevida que permitieran cuantificar las diferencias entre los distintos grupos. En esta etapa se consideró además, la inclusión de la edad al inicio del estudio y se adicionó la variable educación con el fin de comprobar si ella incidía en la sobrevida, al controlar por la edad al inicio del estudio. A continuación se detalla la función de riesgo de los modelos estimados.

$$h_i(t) = h_0(t)^{\gamma_1 Sexo_i + \gamma_2 (Edad_i - 82) + \gamma_3 Educ_i}$$

En el caso del modelo de Cox, la función $h_0(t)$ (riesgo de referencia, en este caso mujeres de 82 años con baja educación) se dejó sin especificar, mientras que en el modelo Weibull

se especificó mediante un parámetro de escala y otro de forma. Finalmente, se descartó incluir la educación en los modelos y se obtuvieron las estimaciones que se detallan en la tabla 3.4.

Los resultados indican que los hombres tienen un riesgo menor de fallecimiento (30 % menos según el modelo de Cox y 24 % menos según el modelo paramétrico) mientras que por cada año que el individuo excede de 82 al comienzo del estudio, su riesgo de fallecimiento aumenta un 9 % según el modelo semiparamétrico y un 7.6 % según el modelo paramétrico. A continuación se puso a prueba el supuesto de riesgos proporcionales del modelo de Cox sin rechazarse la hipótesis nula. Por último se presentan curvas de sobrevida para individuos de ambos sexos con distintas edades al inicio bajo el modelo Weibull (figura 3.5).

	parámetro	Cox	WeibullLos parámetros de este modelo están su versión PH y no AI
Sexo Masculino	γ_1	-0.35 (<0.001)	-0.26 (0.001)
(Edad - 82)	γ_2	0.09 (<0.001)	0.07 (<0.001)
escala	p	-	0.62 (<0.001)
forma	λ	-	0.66 (<0.001)

Tabla 3.4: Modelos de regresión

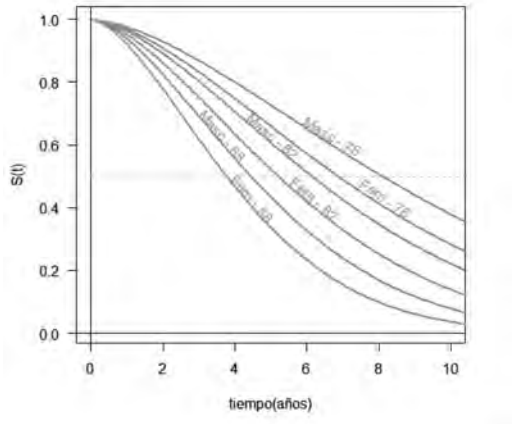


Figura 3.5: Curvas ajustadas según sexo y edad al inicio del estudio

3.12.3. Estimación del modelo conjunto

En esta etapa se investigó si las estimaciones del modelo longitudinal podían verse "alteradas" al adicionar al análisis la información contenida en el proceso de sobrevida. Para ello se ligaron los modelos finales de las secciones 3.12.1 y 3.12.2 de la siguiente manera:

$$\begin{aligned}
h_i(t) &= h_0(t)\exp\{\gamma_1 Sexo_i + \gamma_2(Edad_i - 82) + \alpha_1 f_1(m_i(t)) + \alpha_2 f_2(m_i(t))\} \\
MMSE_{ij} &= m_i(t_j) + \epsilon_{ij} \qquad j = 1,\ldots,n_i \qquad i = 1,\ldots,n \\
m_i(t_j) &= \beta_{0i} + \beta_{1i}t_j + \beta_{2i}t_j^2 \\
\beta_{0i} &= \beta_{00} + \beta_{01}(Edad_i - 82) + \beta_{02}Educ_i + b_{0i} \\
\beta_{1i} &= \beta_{10} + \beta_{11}(Edad_i - 82) + \beta_{12}Educ_i + b_{1i} \\
\beta_{2i} &= b_{2i} \\
\underline{b_i} &\sim N_3(0, D) \\
\epsilon_{ij} &\sim N(0, \sigma^2)
\end{aligned}$$

Es claro como bajo esta formulación la relación entre ambos sub-modelos se da a través de los efectos aleatorios b_i. A partir de esta formulación se estimaron distintas alternativas de modelos conjuntos. Todas las estimaciones concernientes a esta etapa fueron llevadas a cabo utilizando la librería JM (Rizopoulos, 2010). Se comprobó que el vínculo entre ambos modelos fue significativo en todas las especificaciones mientras que de las tres alternativas planteadas para $f_2(m(t))$, la que logró mejores resultados fue $m_i^{'}(t)$. Para cada especificación se comprobó que el modelo RE arrojó un BIC menor a su equivalente al modelo Weibull. En la tabla 3.5 se presentan los resultados del modelo longitudinal (modelo 1), el modelo conjunto (modelo 2), el modelo que incluye el efecto de la pendiente (modelo 3) y el modelo que incluye el efecto acumulado (modelo 4).

A través de estas opciones se pretende estudiar la sensibilidad de las estimaciones al proceso de pérdida de los datos. Puede notarse como el modelado conjunto no parece afectar la constante, sin embargo altera el efecto de la edad y la educación tanto en el nivel de base como en el efecto del tiempo.

3.12.4. Análisis de sensibilidad

En última instancia se trató de cuantificar el efecto del modelado conjunto en el cambio de los parámetros del modelo longitudinal. A continuación se presenta el valor del $ISNI$ y dos modificaciones del mismo para los tres modelos $MNAR$ considerados.

efecto	parámetro	EMV IC 95 %	EMV IC 95 %	EMV IC 95 %	EMV IC 95 %
Sub-modelo longitudinal		Modelo 1	Modelo 2	Modelo 3	Modelo 4
Constante					
constante	β_{00}	25.27 (24.69 ; 25.85)	24.95 (23.96 ; 25.94)	25.46 (24.47 ; 26.45)	25.44 (25.07 ; 25.81)
(Edad - 82)	β_{02}	-0.22 (-0.35 ; -0.09)	-0.21 (-0.34 ; -0.07)	-0.27 (-0.4 ; -0.14)	-0.27 (-0.33 ; -0.2)
Educ>6	β_{03}	1.32 (0.53 ; 2.11)	1.6 (0.54 ; 2.67)	1.51 (0.45 ; 2.57)	1.54 (1.06 ; 2.02)
Tiempo					
constante	β_{10}	-4.59 (-7.58 ; -1.6)	-5.74 (-9.6 ; -1.87)	-2.25 (-6.12 ; 1.61)	-2.15 (-4.41 ; 0.11)
(Edad - 82)	β_{12}	-0.77 (-1.01 ; -0.53)	-0.75 (-0.97 ; -0.53)	-0.99 (-1.21 ; -0.77)	-0.99 (-1.20 ; -0.78)
Educ>6	β_{13}	3.05 (0.44 ; 5.67)	1.31 (-1.26 ; 3.88)	4.28 (1.71 ; 6.84)	4.39 (2.11 ; 6.67)
Sub-modelo de sobrevida					
sexo	γ_1		-0.39 (-0.57 ; -0.22)	-0.37 (-0.54 ; -0.2)	-0.39 (-0.56 ; -0.22)
(Edad - 82)	γ_2		0.04 (0.01 ; 0.06)	0.03 (0.01 ; 0.06)	0.04 (0.01 ; 0.07)
$m_i(t)$	α_1		-0.04 (-0.06 ; -0.03)	-0.03 (-0.04 ; -0.02)	-0.08 (-0.09 ; -0.06)
$m_i^{'}(t)$	α_2			-0.01 (-0.39 ; 0.36)	0.09 (0.06 ; 0.12)

Tabla 3.5: Estimaciones e indicadores de modelos longitudinales

La primera alternativa consiste en dividir el valor del indicador entre el error estándar de cada parámetro (estimado bajo MAR). Rizopoulos considera que valores mayores a uno indican gran sensibilidad. La segunda alternativa también es un índice relativo, pero en este caso, relativo al tamaño del coeficiente estimado bajo MAR.

La tabla 3.6 presenta los valores calculados para cada parámetro bajo cada especificación de $f_2(m(t))$ en el modelo con función de riesgo RE. En dicha tabla se puede apreciar que el modelo 2 ($\alpha_2 = 0$) presenta coeficientes muy sensibles salvo el correspondiente al efecto de la edad al inicio en la constante. Esta situación cambia al considerar modelos donde el riesgo de fallecimiento se ve afectado por el efecto de la pendiente de la evolución del $MMSE$ (modelo 3) o el efecto acumulado del $MMSE$ (modelo 4). En estos casos, todas las variables presentan una alta sensibilidad al supuesto MAR, por lo cual se concluye que el acoplamiento de ambos modelos es capaz de producir alteraciones en las estimaciones que hacen que el modelo reproduzca una situación más cercana a la realidad.

	parámetro	ISNI	$\frac{ISNI}{s.e.\beta_j}$	$\frac{ISNI}{\beta_j}$
Modelo 2				
constante	β_{00}	26.43	89.10	1.05
(Edad - 82)	β_{01}	-0.01	-0.22	0.06
Educ6	β_{02}	-28.44	-70.68	-21.57
tiempo	β_{10}	54.26	35.51	-11.82
(Edad - 82) (tiempo)	β_{11}	1.10	8.95	-1.43
Educ6 (tiempo)	β_{12}	-71.50	-53.66	-23.41
Modelo 3				
constante	β_{00}	116.41	392.48	4.61
(Edad - 82)	β_{01}	10.88	168.13	-49.27
Educ6	β_{02}	-136.76	-339.81	-103.71
tiempo	β_{10}	65.13	42.64	-14.19
(Edad - 82) (tiempo)	β_{11}	16.08	130.47	-20.91
Educ6 (tiempo)	β_{12}	-267.85	-201.01	-87.70
Modelo 4				
constante	β_{00}	17.34	58.46	0.69
(Edad - 82)	β_{01}	4.55	70.32	-20.61
Educ6	β_{02}	-22.42	-55.72	-17.01
tiempo	β_{10}	4.49	2.94	-0.98
(Edad - 82) (tiempo)	β_{11}	3.00	24.34	-3.90
Educ6 (tiempo)	β_{12}	-34.67	-26.02	-11.35

Tabla 3.6: Estimaciones del ISNI y sus modificaciones

3.13. Conclusiones

A través de este análisis se llegó a la conclusión de que la evolución del $MMSE$ se ve afectada tanto por la edad como por la educación de las personas. Al incluir el análisis de sobrevida como parte del proceso de deterioro cognitivo, se "corrigieron" los valores de los coeficientes estimados en el modelo especificado bajo MAR. Se observó que la sobrevida de los hombres es inferior a la de las mujeres y (como era de esperarse) que a mayor edad al inicio, mayor es el riesgo de fallecimiento. Adicionalmente se observó que valores bajos de $MMSE$ incrementan el riesgo de fallecimiento. Lo mismo sucede con la pendiente en la evolución del $MMSE$, al disminuir la misma (avance del deterioro) el riesgo de fallecimiento aumenta aún más. En cuanto al modelo longitudinal se observó que la trayectoria del $MMSE$, entre los individuos con educación baja, decrece más lentamente que lo indicado en principio por el modelo MAR mientras que la educación tiene un efecto mayor al pensado originalmente.

Bibliografía

Arbeev, K., Akushevich, I., Kulminski, A., Ukraintseva, S., y Yashin, A. (2014). Joint analyses of longitudinal and time-to-event data in research on aging: Implications for predicting health and survival. Frontiers in Public Health.

Breslow, N. (1972). Discussion following "Regression models and life tables'' by D. R. Cox. Journal of the Royal Statistical Society, B(34):187–220.

Cabella, W. y Pellegrino, A. (2009). La seguridad social en el Uruguay. Contribuciones a su historia, caplo El envejecimiento de la población uruguaya y la transición estructural de las edades, pp. 89–114. AFAP-BROU , Montevideo.

Cox, D. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, 34(2):187–220.

Daniels, M. anda Hogan, J. (2008). Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman and Hall/CRC.

DeGruttola, V. y Tu, X. (1994). Modeling progression of cd-4 lymphocyte count and its relationship to survival time. Biometrics, 50:1003–1014.

Dempster, A., Laird, N., y Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, 39(1):1–38.

Diggle, P., Farewell, D., y Henderson, R. (2007). Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. Journal of the Royal Statistical Society, Series C, 56:499–550.

Diggle, P. y Kenward, M. (1994). Informative dropout in longitudinal data analysis. ournal of the Royal Statistical Society, Series C, 43:49–93.

Faucett, C. y Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A gibbs sampling approach. Statistics in Medicine, 15(15):1663–1685.

Folstein, M., Folstein, S., y McHugh, P. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12(3):189–198.

Ghisletta, P. (2008). Application of a joint multivariate longitudinal-survival analysis to examine the terminal decline hypothesis in the swiss interdisciplinary longitudinal study on the oldest old. The Journals of Gerontology, 63(3):185–192.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement, 5:475 – 492.

Hedeker, D. y Gibbons, R. (2006). Longitudinal Data Analysis. Wiley Series in Probability and Statistics. Wiley.

Henderson, H., Diggle, P., y Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. Biostatistics, 1(4):465–480.

Hsieh, F., Tseng, Y.-K., y Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. Biometrics, 62:1037–1043.

Hughes, S., Gibbs, J., Dunlop, D., Edelman, P., Singer, R., y Chang, R. (1997). Predictors of decline in manual performance in older adults. Journal of the American Geriatrics Society, 45(8):905–910.

Kalbfleisch, J. y Prentice, R. (1973). Marginal likelihoods based on cox's regression and life model. Biometrika, 60:267–278.

Lenehan, M., Summers, M., Saunders, N., Summers, J., y Vickers, J. (2015). Relationship between education and age-related cognitive decline: a review of recent research. Psychogeriatrics, 15(2):154–162.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association, 88:125–134.

Nocedal, J. y Wright, S. J. (2006). Numerical Optimization. Springer, New York, 2nd edici

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., y R Core Team (2013). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-108.

R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Richardson, A. y Welsh, A. (2008). Asymptotic properties of restricted maximum likelihood (reml) estimates for hierarichal mixed linear models. Australian Journal of Statistics, 36(1):31–43.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. Journal of Statistical Software, 35(9):1–33.

Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman & Hall/CRC Biostatistics Series. Chapman and Hall/CRC.

Rubin, D. (1976). Inference and missing data. Biometrika, 63:581–592.

Self, S. y Pawitan, Y. (1992). Modeling a marker of disease progression and onset of disease. Methodological Issues. Birkher Boston.

Stram, D. y Lee, J. (1994). Variance components testing in the longitudinal mixed effects model. Biometrics, 50(4):1171–1177.

Terrera, G., Piccinin, A., Johansson, B., Matthews, F., y Hofer, S. (2011). Joint modeling of longitudinal change and survival: An investigation of the association between change in memory scores and death. GeroPsych, 24(4):177–185.

Therneau, T. (2012). A Package for Survival Analysis in S. R package version 2.37-2.

Verbeke, G. y Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. Springer, 1 edici

Viviani, S. (2012). Mixed effect joint models for longitudinal responses with dropout: estimation and sensitivity issues. Tesis doctoral, Sapienza, Università di Roma.

Waseem, S. (2007). Kenward-Roger Approximate F Test for Fixed Effects in Mixed Linear Models. Tesis doctoral, Oregon State University.

Wu, M. y Bailey, K. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. Statistics in Medicine, 5:337–346.

Wu, M. y Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics, 44:175–188.

Wulfsohn, M. y Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. Biometrics, 53:330–339.

Capítulo 4

Evaluación de un instrumento de medición del nivel de satisfacción estudiantil a través de la aplicación de Structural Equation Modelling (SEM), por Elena Vernazza

Resumen [1]

En este trabajo se estudian las propiedades psicométricas de un instrumento propuesto para medir la satisfacción estudiantil para los cursos superiores de la Universidad de Beira Interior (Portugal), para luego ver los resultados que surgen de aplicarlo para el caso de la Facultad de Ciencias Económicas y de Administración, Udelar (Uruguay).

El indicador propuesto para medir el nivel de satisfacción estudiantil considera relaciones de causa-efecto entre algunas variables que son consideradas como "antecedentes" y otras como "consecuencia"de la satisfacción. En el primer conjunto de variables se encuentran las expectativas de los alumnos, la imagen que tienen de la facultad, la calidad de la enseñanza y servicios, y el valor percibido, mientras que como "consecuencias"de la satisfacción se encuentran la lealtad hacia la institución y el impacto en el boca a boca.

Los datos utilizados para la aplicación presentada en este trabajo provienen de una encuesta realizada por la Cátedra de Metodología de la Investigación de FCCEEyA, en conjunto con el IESTA, aplicada sobre una muestra probabilística de estudiantes de la facultad, en el año 2009.

1. Resumen del trabajo de pasantía, con la tutoría de Ramón Álvarez-Vaz, Danny Freira para la obtención del grado de la Licenciatura en Estadística.

El cuestionario aplicado, presenta 9 bloques de preguntas; el primero contiene las variables que permitirán realizar una caracterización de los estudiantes en función de características sociodemográficas. Por otra parte, las variables pertenecientes a los bloques A - H presentan las variables del modelo ECSI (European Customer Satisfaction Index) y serán las utilizadas como insumos para el cálculo del índice de satisfacción estudiantil.

Los resultados presentados surgen en primer lugar, de una comparación directa con los resultados obtenidos para el caso portugués. Por otra parte, se presentan resultados teniendo en cuenta solo a los estudiantes de la FCCEEyA, considerándolos, en primera instancia, sin distinciones y luego diferenciándolos por sexo y carrera.

Palabras claves: Análisis factorial, índice de satisfacción, modelos de ecuaciones estructurales.

4.1. Introducción

Conocer el nivel de satisfacción de los clientes, con determinado servicio al cual acceden, resulta fundamental ya que en función de este conocimiento se podrán tomar decisiones que tengan como objetivo primordial mantener ó mejorar, en caso de que sea necesario, aquellos aspectos que se entiende determinan la "satisfacción". Para llevar a cabo esto, será necesario encontrar un mecanismo que permita medir, de forma objetiva y sin ambigüedades, este concepto.

En este orden es que, en 1989, surge en Suecia la propuesta de elaborar un método uniforme para medir la satisfacción del cliente y, a partir de los resultados, poder evaluar la calidad de los servicios. Esto da lugar al Swedish Customer Satisfaction Barometer (SCSB), primer índice elaborado para medir la satisfacción del consumidor, el cual considera la calidad percibida (que recoge tanto las expectativas de los clientes como el valor percibido), la satisfacción (como consecuencia de las percepciones de los clientes) y el comportamiento futuro (reflejado en reclamos y lealtad) (García et al., 2006). Tomando este índice como referencia, e introduciendo algunas modificaciones y adaptaciones, surgen otros, entre los que se destacan: el American Customer Satisfaction Index (ACSI),

elaborado en 1994 por la Universidad de Michigan y la ASQ (American Society for Quality), el Norwegian Customer Satisfaction Barometer (NCSB), el cual agrega la noción de imagen como antecedente de la satisfacción, y el European Customer Satisfaction Index (ECSI) el cual es coordinado por la EOQ (European Organitation for Quality) y se presenta como una variación del ACSI. Éste último mantiene el concepto de imagen, como "causa" de la satisfacción, y en cuanto a los comportamientos futuros, sólo maneja la idea de lealtad. Todos estos índices definen a la satisfacción como una evaluación general de los actos de consumo y entienden que la lealtad debe ser el comportamiento futuro más perseguido por quienes brindan los servicios (García et al., 2006).

Tal como establecen Alves y Raposo (Alves y Raposo, 2007b)[2] tanto el SCSB, como los demás instrumentos generados a partir de modificaciones y adaptaciones de este, dejan de lado la idea de la "satisfacción" como un concepto estático y comienzan a tratarla como parte de un sistema amplio conformado por una serie diversa de interacciones. Es por esto que se entiende que estos índices no solo permitirán cuantificar el nivel de satisfacción de los clientes, sino que también lograrán poner de manifiesto cómo se genera dicho nivel de satisfacción o insatisfacción (Dermanov y Eklof, 2001).

En resumen, podría decirse que los índices de satisfacción del cliente, se encuentran dentro de un sistema que establece relaciones de causa y efecto, que parten desde los antecedentes y llegan hasta las consecuencias de la satisfacción.

En este trabajo, se vincula el concepto de satisfacción descrito previamente, con la educación universitaria, para lo cual se toma como punto de partida lo propuesto por Alves y Raposo (Alves y Raposo, 2004), quienes plantean: "Sólo con la satisfacción de los alumnos se podrá alcanzar el éxito escolar, la permanencia de los estudiantes en la institución y, sobre todo, la formación de una valoración positiva boca a boca. En este sentido, es extremamente importante encontrar formas fiables de medir la satisfacción del alumno en la enseñanza universitaria, permitiendo así a las instituciones de enseñanza conocer su realidad, compararla con la de los otros competidores y analizarla a lo largo del tiempo".

2. Citando a Wilton y Nicosia (1986). "Emerging paradigms for the study of consumer satisfaction".

En función de esto, se considerará a los estudiantes de los cursos superiores de la Facultad de Ciencias Económicas y de Administración, FCCEEyA (Udelar), como "clientes" y se entenderá que el "servicio" que se les brinda es el de la educación de nivel terciario. Tal como establecen Blanco y Blanco (Blanco y Blanco Peck, 2007) [3] no se pueden dejar de lado los valores y metas de la Universidad como institución, es decir, no debe perderse la visión humana de los estudiantes que forman parte de ella. Si esto se logra, se evitará considerar al modelo de educación como un modelo industrial, donde se estaría considerando a los estudiantes como simples productos del sistema.

Lograr conocer la dimensión de la satisfacción de los estudiantes con la facultad a la cual concurren, permitirá identificar aspectos tanto positivos como negativos, siendo estos últimos fundamentales a la hora de determinar estrategias de mejora tanto de la educación, como de los demás servicios.

En este trabajo, la información necesaria para poder evaluar y entender por un lado, qué conceptos se asocian a la satisfacción y por otro, cómo se establecen la interrelaciones entre estos conceptos, se obtiene a través de la aplicación de un cuestionario formado por una serie de bloques de preguntas que se corresponden con lo propuesto en el modelo ECSI; sobre este instrumento y con la ayuda del análisis factorial (AF) y, más precisamente, de los modelos de ecuaciones estructurales (MES) que se presentan en la sección y 4.2.1, se logra poner de manifiesto los componentes de la satisfacción.

El presente trabajo se estructura en cuatro secciones. En primera instancia, en la sección 4.2 se presenta la metodología utilizada donde se hace referencia a los principales aspectos del análisis factorial y se pone especial énfasis en la presentación de los modelos de ecuaciones estructurales. Esta sección culmina con la presentación teórica del índice de satisfacción calculado en este trabajo.
En la tercera sección se exponen los principales resultados y por último, la sección 4.4, presenta las principales conclusiones obtenidas.

3. Citando a Gaitán y López, (1999). "La calidad, nueva función en la Universidad Venezolana"

4.1.1. Objetivos

El objetivo general de este trabajo es evaluar un instrumento de medición del nivel de satisfacción estudiantil a través de la aplicación de modelos de ecuaciones estructurales. Para esto, se estudian las propiedades psicométricas de un instrumento propuesto para medir la satisfacción estudiantil para los cursos superiores de la Universidad de Beira Interior (Portugal), para ver los resultados que surgen de aplicarlo para el caso de la FCCEEyA (Udelar).

En función de este, surgen los siguientes objetivos específicos:

- Medir la satisfacción estudiantil con los cursos de la FCCEEyA, a través de la aplicación de modelos de ecuaciones estructurales, a partir de considerar variables que resulten ser causa o consecuencia de esta.
- Determinar, a partir de una comparación directa, si existen diferencias entre el modelo propuesto para el caso portugués y el caso de la FCCEEyA.
- Determinar si existen diferencias, al considerar tanto el sexo de los estudiantes como la carrera a la que están inscriptos.

4.2. Metodología

4.2.1. Modelos de ecuaciones estructurales

Este tipo de modelos pueden ser vistos, fundamentalmente, de dos maneras. Por un lado, pueden ser enmarcados en el ámbito de los modelos de regresión, con ciertas particularidades que los diferencian de los modelos de regresión clásicos y, por otro, pueden ser vistos como una técnica de análisis factorial que permite establecer relaciones entre los factores.

De modo simplificado, podría entenderse que en los modelos de ecuaciones estructurales se presentan relaciones causales entre, por un lado, un conjunto de variables observables y por otro, variables tanto observables como no observables.

A partir de esto, y recordando además que estos modelos se presentan en el contexto del análisis factorial confirmatorio, es que resulta fundamental establecer de forma clara el concepto de causalidad, ya que justamente es esta relación la que este tipo de modelos intentan confirmar.

Tomando como referencia lo propuesto por Casas Guillen (Casas Guillén, 2002)[4] podríamos decir que existe una relación de causalidad entre la variable X y la variable Y y, más precisamente, que X causa a Y si cada vez que sucede X, sucede Y, y nunca se da Y sin que previamente se haya dado X.

Los modelos de ecuaciones estructurales presentan la particularidad de que una variable puede ser causada por otra variable del sistema y a la vez, dentro del mismo modelo, ser causa de otra variable.

Como se explicitó previamente, existen fundamentalmente dos tipos de relaciones presentadas en los modelos de ecuaciones estructurales. Por un lado, se establecen relaciones entre variables no observadas y, por otro, relaciones entre estas variables, y variables observadas, lo que da lugar a dos submodelos: modelo estructural y modelo de medida, respectivamente
Si bien la presentación de estos, se plantea en términos de ecuaciones, por lo general, se agrega una representación gráfica que permite visualizar mejor las relaciones entre variables.

Previo a presentar los pasos a seguir al trabajar con este tipo de modelos, se establecen las características más relevantes de las variables que los conforman. Tanto las variables observadas como las no observadas, pueden ser de naturaleza endógena o exógena. En lo que refiere a variables latentes, de aquí en adelante se entenderá por variable endógena, a aquella variable cuyas causas están presentes en el modelo, estas estarán siempre acompañadas de un término de error/perturbación. Por otra parte, una variable será tratada como exógena cuando no reciba efecto alguno de ninguna de las variables que forman parte del modelo. Estas serán manejadas como "libres de error". En cuanto a las variables observadas, se dirá que una variables es exógena, cuando las variables que la causan sean variable latentes exógenas, mientras que se estará frente a una variable endógena cuando las causas de esta sean de naturaleza también endógena.

4. Citando a Bisquerra, R. (1989), en "Introducción conceptual al análisis multivariable". Vol. II, PPU, Barcelona.

Por último se presentan los errores, variables aleatorias no observables que recogen aquellos efectos sobre las variables dependientes, que el modelo no logra captar.

Una vez establecidas las principales características que pueden presentar las variables que se manejarán de aquí en adelante, y las posibles relaciones que podrán existir entre ellas, se presentan los detalles de los MES.

Especificación: Modelo estructural

El modelo estructural es el submodelo, dentro de los MES, que captura las relaciones existentes entre las variables no observables, también denominadas variables latentes, constructos o factores.[5]

En formato matricial, podemos representar estos modelos de la siguiente manera:

$$\beta\eta = \Gamma\xi + \zeta \quad \Rightarrow \quad \eta = B\eta + \Gamma\xi + \zeta \tag{4.1}$$

donde, considerando un modelo con m variables latentes endógenas y k variables latentes exógenas, se tiene:

- β (beta) matriz, de dimensión mxm, de pesos β que determinan la relación entre dos variables latentes endógenas,
- η (eta) vector, de dimensión mx1, de variables latentes endógenas,
- Γ (gamma) matriz, de dimensión mxk, de pesos γ que determinan la relación entre una variable endógena y una exógena, ambas latentes,
- ξ (xi) vector, de dimensión kx1, de variables latentes exógenas,
- ζ (zeta) vector, de dimensión mx1, de términos de error/perturbación.

También forman parte de este modelo, las matrices Φ (phi) y Ψ (psi), que representan la matriz de correlaciones entre las variables latentes exógenas (ξ) y la matriz de correlaciones entre los errores de las variables latentes endógenas (ζ), respectivamente.

5. De aquí en adelante, utilizaremos cualquiera de estos términos indistintamente.

A modo de ejemplo, se consideran tres variables latentes endógenas η_1, η_2, η_3 y una variable latente exógena ξ_1 y lo que se desea es confirmar que: ξ_1 y η_2 causan a η_1, ξ_1 y η_3 causan a η_2, y ξ_1y η_1 causan a η_3. La representación analítica del modelo es la siguiente:

$$\begin{cases} \eta_1 = \gamma_1\xi_1 + \beta_1\eta_2 + \zeta_1 \\ \eta_2 = \gamma_2\xi_1 + \beta_2\eta_3 + \zeta_2 \\ \eta_3 = \gamma_3\xi_1 + \beta_3\eta_1 + \zeta_3 \end{cases} \tag{4.2}$$

Especificación: Modelo de medida

En el modelo de medida se establecen las relaciones que existen entre los factores y las variables observables. Se presentan, por separado, las relaciones entre las variables exógenas y las endógenas, lo que determina dos submodelos. La expresión matricial para el modelo de medida, para las variables exógenas, queda determinado por:

$$X = \Lambda_x\xi + \delta \tag{4.3}$$

donde, considerando un modelo con k variables latentes y q variables observables, se tiene:

- X vector, de dimensión qx1, de variables observables,
- Λ_x (lambda) matriz, de dimensión $q$xk, de pesos λ que determinan la relación entre cada x y cada ξ,
- ξ (xi) vector, de dimensión kx1, de variables latentes exógenas,
- δ (delta) vector, de dimensión qx1, de términos de error/perturbación.

La matriz Θ_δ también forma parte de este submodelo. Esta es la matriz de covarianzas entre los errores de las variables exógenas observadas (δ). Los errores δ se suponen incorrelacionados, por lo que la matriz Θ_δ resulta una matriz diagonal.

Para las variables endógenas, la expresión matricial para el modelo de medida, es la siguiente:

$$Y = \Lambda_y \eta + \epsilon \tag{4.4}$$

donde, considerando un modelo con m variables latentes y p variables observables, se tiene:

- Y vector, de dimensión px1, de variables observables,
- Λ_y (lambda) matriz, de dimensión pxm, de pesos λ que determinan la relación entre cada y y cada η,
- η (eta) vector, de dimensión mx1, de variables latentes endógenas,
- ϵ (epsilon) vector, de dimensión px1, de términos de error/perturbación.

La matriz Θ_ϵ también forma parte de este submodelo. Esta es la matriz de covarianzas entre los errores de las variables endógenas observadas (ϵ). Los errores ϵ se suponen incorrelacionados, por lo que la matriz Θ_ϵ resulta una matriz diagonal.

Continuando con el ejemplo presentado para el modelo estructural, se agregan dos variables exógenas x_1, x_2 y seis endógenas $y_1 ... y_6$, todas observables. Se supone que x_1 y x_2 son causadas por ξ_1 , y_1 y y_2 por η_1, y_3 y y_4 por η_2 y y_5 y y_6 por η_3. Las ecuaciones que determinan este modelo son:

$$\begin{cases} x_1 = \lambda_1 \xi_1 + \delta_1 \\ x_2 = \lambda_2 \xi_1 + \delta_2 \end{cases} \tag{4.5}$$

$$\begin{cases} y_1 = \lambda_3 \eta_1 + \epsilon_1 \\ y_2 = \lambda_4 \eta_1 + \epsilon_2 \\ y_3 = \lambda_5 \eta_2 + \epsilon_3 \\ y_4 = \lambda_6 \eta_2 + \epsilon_4 \\ y_5 = \lambda_7 \eta_3 + \epsilon_5 \\ y_6 = \lambda_8 \eta_3 + \epsilon_6 \end{cases} \tag{4.6}$$

Identificación

Kline (Kline, 2011) entiende que un modelo está "identificado" si es posible obtener una estimación única para cada uno de los parámetros involucrados en el modelo.

Esta condición resulta necesaria pero, por lo general, no suficiente. El modelo resultará completamente identificado, cuando tanto el submodelo estructural como el de medida lo estén, para lo cual será necesario establecer restricciones sobre algunos de los parámetros a estimar en ambos submodelos.

Estimación

Una vez que finaliza el proceso de identificación del modelo se pasa a la etapa de estimación del modelo, lo cual implica obtener una estimación puntual para cada uno de los parámetros involucrados en los submodelos de medida y estructural. Tres de los métodos más utilizados son:

- Máxima verosimilitud (MV).
- Mínimos cuadrados generalizados (MCG).
- Mínimos cuadrados parciales (MCP).

Los resultados presentados en este trabajo se obtienen estimando por MV.

Evaluación

Para poder evaluar si tanto la especificación como la estimación del modelo han sido adecuadas, resulta conveniente establecer a qué refiere el concepto de ajuste en el ámbito de los modelos de ecuaciones estructurales.

Tal como plantea Ruiz (s.f.) el concepto de ajuste de un modelo puede ser resumido en la siguiente hipótesis: "si el modelo es correcto y conociéramos los parámetros del modelo estructural, la matriz de covarianzas poblacional podría ser reproducida exactamente a partir de la combinación de los parámetros del modelo". A partir de esto se establece que la hipótesis fundamental a testear para determinar la bondad de ajuste del modelo es:

$$H_0)\ \Sigma = \Sigma(\theta) \tag{4.7}$$

donde: Σ es la matriz de varianzas y covarianzas poblacionales y $\Sigma(\theta)$ es la matriz de varianzas y covarianzas que deriva del modelo (denominada matriz de varianzas y covarianzas implícita), considerando que el vector θ contiene a todos los parámetros involucrados en el modelo.

La noción de ajuste entendida de esta forma hace referencia a los modelos de ecuaciones estructurales desde una perspectiva global, es decir, considerando que se está trabajando con un único modelo. Sin embargo, podría decirse que el proceso de evaluación consta de tres etapas: evaluación del modelo de medida, del estructural, y una evaluación global, considerando ambos submodelos como un único modelo.

La evaluación del modelo de medida consiste en en testear si las variables incorporadas resultan significativas. Evaluar el submodelo estructural, desde el enfoque de carer confirmatorio, requiere determinar si las relaciones establecidas, en la etapa de especificación del modelo efectivamente, se confirman. Para esto se realizan las pruebas de significación correspondientes sobre los parámetros estimados. Una vez que finalizan los procesos de evaluación de los dos submodelos, se pasa a realizar una evaluación de ambos en conjunto, considerándolos como un nico modelo. Esta evaluación se realiza a través de la comparación de la matriz de varianzas y covarianzas observadas y la que resulta del modelo estimado.

4.2.2. Índice de satisfacción

El índice de satisfacción (IS) presentado en este trabajo es calculado a partir de lo establecido en "The American Customer Satisfaction Index: Nature, Purpose, and Findings" (Claes et al., 1996), donde se propone la siguiente fórmula de cálculo para el índice ACSI:

$$ACSI = \frac{E(S) - Min(S)}{Max(S) - Min(S)} \tag{4.8}$$

donde:

- S es la variable latente referente a la satisfacción
- $E(S) = \sum_i w_i \bar{y_i}$
- $Min(S) = \sum_i w_i Min(y_i)$

- $Max(S) = \sum_i w_i Max(y_i)$
- y_i es la i-ésima variable observada sobre la que satura el constructo S
- w_i es la estimación no estandarizada del peso con el que satura la i-ésima variable observada sobre el constructo S.

Tal como surge de la fórmula, este índice está acotado entre 0 (situación de total insatisfacción) y 1 (situación de total satisfacción).

Considerando que los índices de satisfacción se distinguen fundamentalmente por las variables que consideran y por las relaciones que se establecen entre estas, es que se entiende que la fórmula de cálculo establecida en (4.8), puede ser extendida al índice de satisfacción estudiantil que se presentará en este trabajo.

4.3. Resultados

4.3.1. Datos

Diseño muestral

La aplicación que se presentará en este trabajo fue realizada sobre los datos obtenidos mediante la aplicación de un cuestionario sobre una muestra probabilística a estudiantes de los cursos superiores de la FCCEEyA, en el año 2009.

La muestra fue seleccionada a partir de un marco muestral que se construyó a partir de las inscripciones a cursos de FCCEEyA en 2009. El diseño muestral usado fue estratificado por conglomerados en 2 etapas y presentó las siguientes características: en una primera instancia se formaron 6 estratos que corresponden a cada uno de los 5 años en los en los que podía estar cada estudiante en el 2009 (aproximadamente). Adicionalmente, se propone un 6to estrato para un grupo reducido de materias que corresponden únicamente a la Licenciatura en Administración. Una vez conformados los estratos, se determina que la muestra total se repartirá en forma proporcional a la matrícula de cada estrato.
Al tener definidas las unidades de muestreo, se procede a seleccionar la muestra, proceso que presentó las siguientes etapas:

1. Se sortean los grupos prácticos de cada materia en cada estrato con probabilidad proporcional a la matrícula de cada grupo (conglomerado).

2. Mediante muestreo aleatorio simple (MAS), se seleccionan la misma cantidad de estudiantes en cada grupo seleccionado en la primera etapa. La cantidad de estudiantes de cada grupo es la misma en los 6 estratos.

La muestra finalmente queda conformada por estudiantes que provienen de 60 grupos prácticos (repartidos en forma proporcional en los 6 estratos). Se sortean 12 estudiantes por grupo, lo que determina un tamaño de muestra de 720 estudiantes.

La siguiente tabla muestra como quedan repartidos los 60 grupos prácticos en los 6 estratos.

Estrato	1	2	3	4	5	6	Total
# grupos prácticos	21	15	9	9	4	2	60

Tabla 4.1: Cantidad de grupos prácticos por estrato

Con la muestra seleccionada, se procedió a realizar el relevamiento de los datos el cual culminó con 647 encuestas realizadas, dejando en evidencia que no fue posible acceder a los 720 estudiantes originalmente estipulados, quedando determinada entonces una tasa de cobertura de la muestra de $647/720 = 90\,\%$.
En función de esto, al momento de calcular los expansores, lo primero que se hace es analizar el 10 % de estudiantes que quedó sin encuestar, con el objetivo de evaluar si se podía pensar que estos eran una muestra aleatoria de los 720 estudiantes originales, descartando de esta manera un sesgo de selección. Considerando como variables fundamentales el estrato, la edad y el sexo de los estudiantes, se constató que estas no estaban asociadas a ese 10 % que quedó sin encuestar, es decir que ninguno de esos 3 atributos estaban sub o sobre representados. Otros dos aspectos a tener en cuenta previo al cálculo de los expansores son los siguientes: por un lado se debe tener en cuenta la existencia de multiplicidad en el marco muestral debido a que hay un número diferente de matrículas correspondientes a cada estudiante, lo que impacta en la probabilidad de selección ya que la unidades primarias de muestreo son conglomerados de matrículas y no de estudiantes, es decir, hay estudiantes que están repetidos y pueden ser encontrados en más de una materia.

Por último, debe ser tenido en cuenta el hecho de que la distribución por sexo y edad presente en la muestra definitiva no es la distribución poblacional, lo cual genera la necesidad de aplicar un proceso de calibración mediante postestratificación.

Cuestionario utilizado

El cuestionario, aplicado sobre la muestra seleccionada, a partir del cual se obtuvieron los datos que resultan el insumo fundamental para el trabajo aquí presentado, resulta de una adaptación del cuestionario utilizado por los investigadores Alves y Raposo de la Universidad de Beira Interior (Portugal). Este presenta la siguiente estructura: un primer bloque, claramente diferenciado de los demás, que contiene algunas variables de carácter sociodemográfico, como sexo, edad y algunas otras variables que caracterizan al estudiante dentro del ámbito de la facultad, como año de ingreso, año y cantidad de materias en curso, entre otras. Los restantes 8 bloques de preguntas (presentados como bloque A hasta bloque H) presentan todos la misma estructura, se plantea una pregunta general que determina la esencia del bloque y a partir de ella, se establecen una serie de afirmaciones sobre las cuales el estudiante deberá expresar su posición, utilizando una escala Likert que toma valores en el intervalo [1 - 10], donde 1 indicará la mayor discrepancia con lo planteado en la pregunta y 10 el mayor acuerdo.

Los bloques A a H presentan las siguientes características:

- Bloque A - Contiene 12 afirmaciones referentes a las expectativas de los estudiantes, previo ingreso a facultad.
- Bloque B - Consta de 6 afirmaciones vinculadas a la imagen que tienen los estudiantes sobre la facultad.
- Bloque C - Conformado por 9 afirmaciones asociadas a la calidad del servicio que brinda la facultad.
- Bloque D - Contiene 9 afirmaciones asociadas a la calidad de los servicios que brinda la facultad con respecto a la biblioteca, bedelía y cafetería, entre otros.
- Bloque E - Conformado por las mismas 9 afirmaciones que el bloque C, pero asociadas a necesidades/deseos actuales.
- Bloque F - Presenta 7 afirmaciones que indagan sobre el valor percibido.

- Bloque G - Contiene 6 afirmaciones que refieren a la satisfacción de los estudiantes con la facultad.

- Bloque H - Conformado por 5 preguntas que pueden dividirse en 2 subgrupos, las 3 primeras referentes a la lealtad de los estudiantes con la facultad, y las 2 últimas asociadas al boca a boca que se genera entre los estudiantes.

En este trabajo, los bloques D y E no serán considerados, lo que da lugar a 45 variables observables, que son las que se utilizarán para la aplicación del AF y en particular de los MES.

4.3.2. Modelos propuestos

En esta sección se presentan los cuatro modelos que formaron parte del proceso de búsqueda de un modelo que permita alcanzar los objetivos fijados. Estos cuatro modelos serán comparados fundamentalmente en cuanto a ajuste considerando el modelo de medida y el estructural como un único modelo, a partir de esta comparación tratará de establecerse cuesulta el mejor.
Una vez que esto se logre, se expondrán las principales características del modelo seleccionado, se interpretarán los parámetros estimados y, por último, se calculará el índice de satisfacción estudiantil en los cursos de educación superior de la FCCEEyA.

La presentación de los modelos formulados a continuación consistirá en exponer la especificación, identificación, estimación, evaluación, reespecificación (si fuera necesaria) e interpretación para cada uno de ellos, a excepción de la última etapa que se presentará sólo para el modelo seleccionado.

Previo a entrar en detalles en cada uno de los modelos, se establecen algunas generalidades que involucrarán a todos los modelos propuestos.
Lo primero a resaltar refiere directamente a los datos, y a cómo fueron obtenidos. Tal como se especificó en secciones anteriores, el análisis de modelos que proponen relaciones de causalidad, planteados en términos de MES, depende fundamentalmente de la especificación de la matriz de varianzas y covarianzas, la que varía sustancialmente si se toma en cuenta que los datos no provienen de una muestra generada mediante muestreo aleatorio simple (MAS).

El hecho de considerar datos generados mediante algún diseño muestral complejo obliga a hacer correcciones mediante incrementos de la varianza (Stapleton, 2008).

Por otra parte, considerando en detalle las etapas que conforman el modelado a través de los MES, lo primero a considerar es la etapa de "especificación". En cuanto al submodelo estructural se seguirán fundamentalmente dos propuestas: en algunos modelos se seguirá lo establecido por el ECSI (ver figura 4.1), mientras que en otros se considerará que las relaciones entre las variables latentes son las mismas que las encontradas para la educación superior de Portugal (Alves y Raposo, 2007b) (ver figura 4.2). En lo que refiere al submodelo de medida, la decisión sobre cuáles variables fue tomada en función de un análisis de consistencia interna realizado, dentro de cada bloque, sobre todas las variables del cuestionario. En algunos casos se hizo especial énfasis sobre aquellas variables que resultaron significativas en la investigación de los investigadores portugueses.

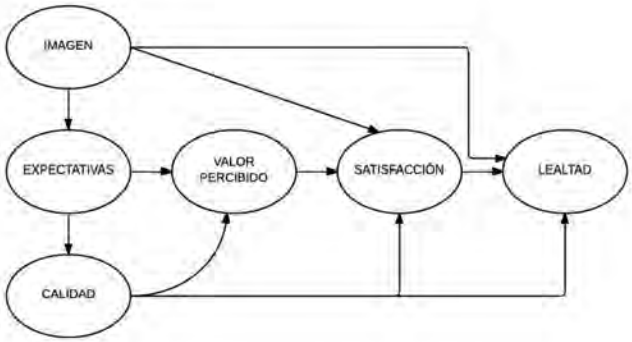


Figura 4.1: Modelo estructural ECSI

En cuanto a la identificabilidad del modelo, las restricciones que resultan comunes a los 4 modelos, impuestas para que estos resulten identificados, son:

- Cada variable latente satura en al menos 2 variables observadas.
- Por cada variable latente, existe un λ fijado igual a 1.

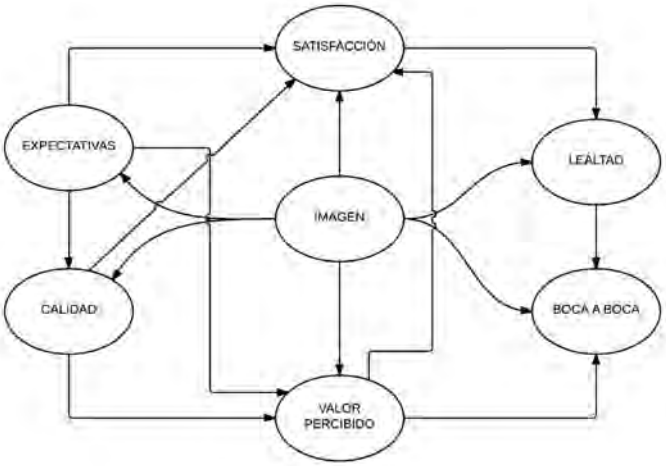


Figura 4.2: Modelo estructural portugués

- Para cada par de variables (X_i, X_j), sobre las cuales satura la misma variable latente, se tiene $cor(\delta_i, \delta_j) = 0$.
- Para cada par de variables (Y_i, Y_j), sobre las cuales satura la misma variable latente, se tiene $cor(\epsilon_i, \epsilon_j) = 0$.
- Las varianza de las variables latentes es fijada igual a 1.

Otra de las etapas que resulta común a todos los modelos, en cuanto a lo metodológico, es la etapa de "estimación"; en todos los casos, los parámetros serán estimados por máxima verosimilitud.

Modelo portugués

El primer resultado a ser presentado en esta sección es el que surge de replicar el modelo que encuentran los investigadores portugueses, para medir la satisfacción de los estudiantes con la educación superior de su país. Dicha investigación toma como referencia las relaciones que establece el índice ECSI, modificando algunas e incorporando el concepto del boca a boca, el cual es tomado como "consecuencia"de la satisfacción.
Las variables a considerar en este modelo son las que se presentan en la tabla 4.2.

Variable	Descripción	Tipo
E	Expectativas	Lat. endógena
C	Calidad	Lat. endógena
VP	Valor percibido	Lat. endógena
S	Satisfacción	Lat. endógena
L	Lealtad	Lat. endógena
BB	Boca a boca	Lat. endógena
I	Imagen	Lat. exógena
EXP_2	Buena preparación para la carrera	Obser. endógena
EXP_3	Capacidad y conocimiento de los docentes	Obser. endógena
Q_1	Calidad global de la enseñanza	Obser. endógena
Q_2	Nivel de conocimiento de los docentes	Obser. endógena
Q_5	Contenido de los cursos	Obser. endógena
S_1	Satisfacción global	Obser. endógena
S_2	Correspondencia con las expectativas	Obser. endógena
S_3	Correspondencia con deseos/necesidades	Obser. endógena
L_1	Volvería a elegir esta facultad	Obser. endógena
L_2	Elegiría esta facultad para carreras de posgrado	Obser.endógena
BB_4	Es una facultad de la cual los egresados se enorgullecen	Obser. endógena
BB_5	Recomendaría esta facultad a un amigo	Obser. endógena
VP_1	Estudiar en esta facultad me ayudará a conseguir un buen empleo	Obser. endógena
VP_2	Mi carrera en esta facultad es una buena inversión	Obser. endógena
VP_5	Empleadores interesados en contratar estudiantes de esta facultad	Obser. endógena
IM_1	Buena Universidad para estudiar	Obser. exógena
IM_2	Facultad innovadora y con visión al futuro	Obser. exógena
IM_4	Facultad que da una buena preparación a sus estudiantes	Obser. exógena

Tabla 4.2: Variables consideradas en el modelo portugués

Las variables observadas que se presentan en la tabla 4.2, y que serán utilizadas acá para plantear un primer modelo, son las que resultan significativas en el trabajo de Alves y Raposo (Alves y Raposo, 2007a).

Modelo estructural

Este modelo propone las siguientes relaciones entre las variables no observadas: en cuanto al constructo imagen (I) este es causa de todos los demás constructos, mientras que las causas de este no están presentes en el sistema. A partir de esto, se entiende que la variable no observable imagen es de naturaleza exógena. En lo que refiere al constructo satisfacción (S), se propone considerar como causas de este, además de la imagen, todas las demás variables, menos lealtad (L) la cual se entiende es una consecuencia de la satisfacción, y el boca a boca (BB). Por otra parte, se desea confirmar que las expectativas son causa directa, además de la satisfacción, del valor percibido y de la calidad, siendo esta última también causa del valor percibido. Por último, en este modelo se buscará confirmar que tanto el valor percibido como la lealtad causan el boca a boca.

Modelo de medida

En cuanto a las relaciones que existen entre las variables observadas y no observadas, estas quedan determinadas en el modelo de medida.

En lo que refiere a las variables exógenas, es decir aquellas variables que conforman el constructo imagen, las relaciones encontradas en los estudios de los investigadores portugueses (Alves y Raposo, 2004), (Alves y Raposo, 2007a), (Alves y Raposo, 2007b) involucran a las variables IM_1,IM_2 e IM_4, presentadas en la tabla 4.2.
Las restantes 15 variables conforman el submodelo, dentro del modelo de medida, que refiere a las variables endógenas.

A continuación se presentan algunos indicadores de bondad de ajuste de estos submodelos considerados como un único modelo y se comparan con los obtenidos por los investigadores portugueses (Alves y Raposo, 2007a), (Alves y Raposo, 2007b).

Índice	FCCEEyA - Uruguay	Portugal
NFI	0.892	0.960
NNFI	0.887	0.958
CFI	0.906	–
RMSEA	0.098	0.065
SRMR	0.094	–

Tabla 4.3: Comparación de los IBJ: Uruguay vs. Portugal

En términos generales, considerando los indicadores presentados en la tabla 4.3, podría decirse que el modelo que se logra estimar para la educación en Portugal ajusta mejor que el estimado con los datos de FCCEEyA. De todas formas, este último también presenta un buen ajuste.

La existencia de una diferencia en el ajuste de ambos modelos a favor, si se quiere, de los portugueses resulta coherente con el hecho de que estos presentan este modelo como su "mejor"modelo, luego de descartar otros, mientras que en el caso de Uruguay, en esta instancia, el modelo se toma como dado y no como parte de un proceso de selección.

Luego de evaluar el ajuste de los dos modelos, se presentan las estimaciones de los parámetros para ambos, y se comparan a través del cociente entre ellos (ver tabla 4.4).

Los parámetros que no aparecen en la tabla fueron fijados en 1, lo que introduce restricciones al modelo estimado. Estas restricciones son las mismas para ambos modelos. En cuanto a las estimaciones, cabe destacar que hay solamente 3 relaciones que en el modelo estimado para Uruguay no resultan significativas con un nivel $\alpha = 0{,}05$[6]. Estas son las determinadas por los parámetros: β_2 (E→VP), β_4 (E→S) y γ_6 (I→BB) indicadas, en la tabla 4.4, con (*).

Tanto las relaciones presentes en este modelo, como las estimaciones, pueden verse gráficamente en la figura 4.3.

6. Considerando el estadístico de contraste $t \sim t_{n-1}$, bajo supuesto de normalidad. En la investigación portuguesa este supuesto sí se verifica, aunque no se presentan resultados con respecto a este tópico.

Parámetro	Estimación Uruguay	Estimación Portugal	Comparación de modelos (Ratios)
β_1	0.31	0.12	2.66
$\beta_2(*)$	0.06	0.10	0.61
β_3	0.32	0.25	1.28
$\beta_4(*)$	-0.04	-0.12	0.33
β_5	0.33	0.16	2.08
β_6	0.33	0.37	0.87
β_7	0.75	0.89	0.84
β_8	0.23	0.21	1.12
β_9	0.60	0.61	0.98
γ_1	0.81	0.54	1.50
γ_2	0.79	0.78	1.01
γ_3	0.72	0.58	1.23
γ_4	0.39	0.43	0.90
γ_5	0.48	0.34	1.42
$\gamma_6(*)$	0.17	0.32	0.54
λ_1	1.21	0.98	1.24
λ_2	1.20	0.96	1.25
λ_4	1.04	0.90	1.16
λ_7	0.77	0.82	0.93
λ_8	0.83	0.87	0.96
λ_{10}	0.97	1.02	0.95
λ_{11}	0.89	1.04	0.86
λ_{13}	1.02	0.93	1.09
λ_{14}	0.64	0.84	0.76
λ_{17}	0.89	1.02	0.87
λ_{18}	0.80	1.00	0.80

Tabla 4.4: Comparación de los coeficientes estimados: Uruguay vs. Portugal

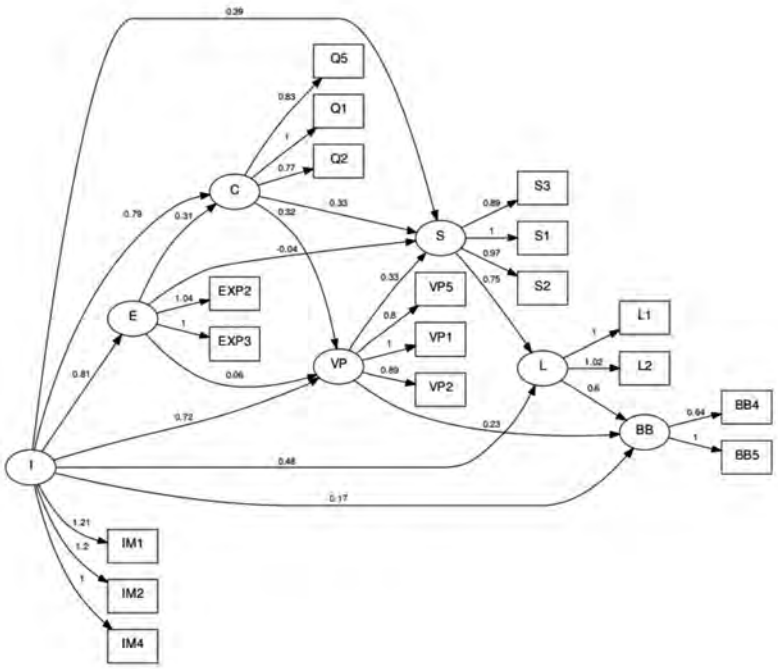


Figura 4.3: Diagrama de senderos modelo portugués

Como última etapa en la comparación directa con la investigación portuguesa, se reporta el índice de satisfacción para el caso de la FCCEEyA. Recordando lo establecido en la ecuación (4.8), la fórmula de cálculo de este índice será la siguiente:

$$IS = \frac{\sum_i w_i \bar{y}_i - \sum_i w_i}{10 \sum_i w_i - \sum_i w_i} * 100 = \frac{\sum_i w_i \bar{y}_i - \sum_i w_i}{9 \sum_i w_i} * 100 \tag{4.9}$$

Variable Medida (y_i)	Estimación no estandarizada Uruguay (w_i)	Media variable medida ($\bar{y}_i$)	$w_i \bar{y}_i$
S_1	1.00	7.45	7.45
S_2	0.967	7.03	6.79
S_3	0.891	6.70	5.96
$Total$	2.858	21.17	20.231

Tabla 4.5: Índice de satisfacción para Uruguay

Utilizando el mismo modelo que en Portugal el IS vale en el caso de la FCCEEyA 67.5 %, el cual supera al reportado por Alves y Raposo (Alves y Raposo, 2007b) (54 %).

Modelo UP

En el proceso de buscar un modelo que permita determinar cuál es el nivel de satisfacción de los estudiantes con la facultad a la cual concurren, la primera propuesta es aquella que intenta confirmar las relaciones existentes entre las variables latentes que fueron determinadas para el caso portugués, es decir, el mismo modelo estructural ya presentado, pero plantear un modelo de medida que incluya todas las variables del cuestionario.

Al analizar la consistencia interna de cada uno de los bloques de preguntas, se tiene que, al retirar determinadas variables, el α de Cronbach del bloque al cual pertenecían, se mantenía o aumentaba. Estas variables eran:

- La probabilidad de que las cosas pudieran ser diferentes de lo esperado (EXP_{11}), del bloque expectativas.
- Es una facultad muy comprometida con la comunidad (IM_5), del bloque imagen.
- Es una facultad que los empleadores valoran (IM_6), del bloque imagen.
- Considero que el contenido de las asignaturas se aplica en su mayoría a la vida práctica (VP_4), del bloque valor percibido.
- Imagine una facultad perfecta en todos los aspectos. A qué distancia colocará esta facultad de ese ideal? (S_4), del bloque satisfacción.

De aquí en más, estas 5 variables no serán tenidas en cuenta.

En lo que refiere al modelo de medida, si bien al plantear este modelo se sabe que no resultará bueno, o por lo menos no resultará el mejor, la intención es estimarlo de todas formas con el objetivo de dimensionar cuántas y cuáles variables resultan significativas para tomarlas como punto de partida del proceso de búsqueda de un modelo que sí resulte "bueno".

Para esto, se procede a estimar el modelo y como puede verse en la tabla 4.6, efectivamente se verifica que no presenta un buen ajuste.

Índice	Modelo UP
NFI	0.718
NNFI	0.732
CFI	0.749
RMSEA	0.102
SRMR	0.143

Tabla 4.6: Índices de bondad de ajuste - modelo UP

En cuanto a las estimaciones de los parámetros, los valores y los resultados de los contrastes para determinar si estos resultan significativamente distintos de cero o no, se tiene que todas las variables resultan significativas al 5 %. En función de esto, resulta necesario cambiar de estrategia para proceder a la selección de un conjunto reducido de variables candidatas a formar parte del modelo definitivo.

Modelo ECSI

Como alternativa al modelo anterior, (UP), se propone considerar un modelo de ecuaciones estructurales cuyo modelo estructural proponga confirmar las relaciones que establece el ECSI, y un modelo de medida que considere las mismas variables que el modelo anterior a excepción de las variables del bloque boca a boca ya que el ECSI no considera este constructo. De todas formas, estas variables son solo 2, por lo que tampoco se estará ganando demasiado en cuanto a la reducción de la cantidad de variables.

Tal como ocurría con el modelo anterior, no se espera que este modelo resulte el definitivo para, a través de él, medir el nivel de satisfacción estudiantil en FCCEEyA, ni siquiera que resulte un buen modelo, pero sí que permita identificar aquellas variables que resultan significativas, para continuar trabajando sobre estas.
En la tabla 4.7, puede verse que el ajuste de este modelo tampoco resulta bueno, incluso ajusta peor que el anterior.

Índice	Modelo ECSI
NFI	0.702
NNFI	0.715
CFI	0.731
RMSEA	0.107
SRMR	0.164

Tabla 4.7: Índices de bondad de ajuste - modelo ECSI

Los resultados obtenidos al estimar este modelo, indican que todas las variables resultan significativas al 5 %.
Este modelo es descartado considerando el pobre ajuste que presenta, sumado a que de este no surge una posible alternativa de un modelo más parsimonioso.

Modelo ECSI2

Como alternativa al modelo inmediato anterior, (ECSI), se propone un modelo que mantiene intacto el submodelo estructural e introduce algunas modificaciones en lo que refiere al submodelo de medida. Este contendrá las variables observadas presentadas en la tabla 4.2 pero además se incorporarán aquellas variables que se entienden "importantes"dentro de su bloque, haciendo referencia a que al quitarlas, la consistencia interna de este disminuye. Algunas de estas, como las que corresponden a los bloques de imagen y lealtad, ya formaban parte de las variables consideradas. De los demás bloques se incorporan las siguientes variables:

- El contenido del curso (EXP_6), del bloque expectativas.
- El ambiente académico (Q_4), del bloque calidad.
- ?'Cuál es su grado de felicidad por haber elegido esta facultad?
  (S_6), del bloque satisfacción
- Teniendo en cuenta que la FCCEEyA es pública, considero que recibo un servicio de calidad (VP_3), del bloque valor percibido.

El modelo estimado considerando, en lugar de 18, 20 variables observadas presenta los siguientes indicadores de bondad de ajuste (ver tabla 4.8):

Índice	Modelo ECSI2
NFI	0.850
NNFI	0.847
CFI	0.866
RMSEA	0.109
SRMR	0.137

Tabla 4.8: Índices de bondad de ajuste - modelo ECSI2

Tomando en cuenta los IBJ del modelo ECSI2 y comparándolos con los valores de estos para los dos modelos antes presentados, puede verse un incremento de la calidad de ajuste del modelo, con índices de ajuste incremental que en este caso presentan valores en el entorno del 0.85, superando los valores cercanos al 0.71 de los modelos anteriores.

Tanto los valores de las estimaciones para cada uno de los parámetros, como los p-valores asociados a cada prueba de significación dan cuenta de que todas las relaciones presentes en él, pueden confirmarse y en lo que refiere al modelo de medida, se observa que las variables latentes, efectivamente saturan sobre las variables observadas presentes en el modelo.

Modelo UP2

Por último, se presenta un modelo que toma los modelos anteriores (UP2 y ECSI2) y los combina de la siguiente manera: toma el modelo estructural presentado en las subsecciones 4.3.2 y 4.3.2, es decir el propuesto por investigadores portugueses y, en cuanto al modelo de medida toma lo presentado en las subsecciones 4.3.2 y 4.3.2, lo cual implica considerar las variables del bloque “boca a boca”, y las 4 variables incorporadas en la subsección 4.3.2. En función de esto, los modelos a considerar en esta sección propondrán por un lado confirmar determinadas relaciones entre los 7 constructos no observables (modelo estructural) y, por otro, ver sobre qué variables observables (considerando 22 de las 45 posibles) saturan dichos constructos (modelo de medida).

La presentación de este modelo consistirá en plantear su especificación, verificar que esté correctamente identificado, presentar una estimación puntual para cada uno de los parámetros involucrados y a partir de esto, evaluar la calidad de ajuste del modelo estimado.

Especificación - Modelo estructural

Las principales hipótesis a ser confirmadas en el modelo estructural aquí presentado son las siguientes: en cuanto a la variable imagen, se propone testear si la imagen que tienen los estudiantes de la facultad a la cual asisten, impacta sobre sus expectativas, sobre como perciben la calidad del servicio que reciben, sobre el valor que creen que tiene estudiar en dicha facultad, sobre lo que dicen/piensan sobre esta, sobre el nivel global de satisfacción que sienten los estudiantes con la facultad y sobre si la volverían a elegir. Cabe recordar que tal como se planteó en la subsección 4.3.2 la variable imagen es una variable exógena, por lo que sus causas no están presentes en el modelo.

En lo que refiere a las variables calidad y expectativas, se propone confirmar si efectivamente ambas se presentan como causantes tanto de la satisfacción global de los estudiantes con las facultad, como del valor que estos entienden que les aporta estudiar en esta. También se propone testear si las expectativas que tienen los estudiantes sobre la facultad, influyen directamente sobre cómo perciben la calidad del servicio que reciben. La otra variable que se propone como determinante de la satisfacción es el valor percibido, es decir, el valor que los estudiantes sienten que les aporta estudiar en la FCCEEyA lo cual podría ser causa, también, de lo que se denomina el "boca a boca", es decir, lo que los estudiantes dicen sobre la facultad en la que estudian. Por último, se intentará confirmar si efectivamente la lealtad y el boca a boca son consecuencia de la satisfacción. La lealtad se entiende como una consecuencia directa, mientras que el boca a boca es una consecuencia indirecta, ya que este es causada por la lealtad.

Estas hipótesis pueden ser presentadas de forma analítica, a través de las siguientes ecuaciones:

$$\begin{cases} E = \gamma_1 I + \zeta_1 \\ C = \gamma_2 I + \beta_1 E + \zeta_2 \\ VP = \gamma_3 I + \beta_2 E + \beta_3 C + \zeta_3 \\ S = \gamma_4 I + \beta_4 E + \beta_5 C + \beta_6 VP + \zeta_4 \\ L = \gamma_5 I + \beta_7 S + \zeta_5 \\ BB = \gamma_6 I + \beta_8 VP + \beta_9 L + \zeta_6 \end{cases} \tag{4.10}$$

Las cuales generan la siguiente representación mediante matrices:

$$\beta\eta = \gamma\xi + \zeta$$

donde cada una de las matrices involucradas tiene el siguiente formato:

$$\beta = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -\beta_1 & 1 & 0 & 0 & 0 & 0 \\ -\beta_2 & -\beta_3 & 1 & 0 & 0 & 0 \\ -\beta_4 & -\beta_5 & -\beta_6 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\beta_7 & 1 & 0 \\ 0 & 0 & -\beta_8 & 0 & -\beta_9 & 1 \end{pmatrix} \quad \eta = \begin{pmatrix} E \\ C \\ VP \\ S \\ L \\ BB \end{pmatrix} \quad \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \\ \gamma_6 \end{pmatrix} \quad \xi = I \quad \zeta = \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \\ \zeta_5 \\ \zeta_6 \end{pmatrix}$$

Especificación - Modelo de medida

Lo que diferencia al submodelo de medida presentado en esta subsección, del presentado en la subsección 4.3.2, es la incorporación de 4 variables observadas en el modelo de medida de las variables endógenas. El modelo para las variables exógenas resulta idéntico al propuesto por los investigadores portugueses, el cual se presenta en las siguientes ecuaciones:

$$\begin{cases} IM_1 = \lambda_1 I + \delta_1 \\ IM_2 = \lambda_2 I + \delta_2 \\ IM_4 = \lambda_3 I + \delta_3 \end{cases} \tag{4.11}$$

que llevan a la siguiente representación matricial:

$$X = \Lambda_X \xi + \delta$$

donde cada una de las matrices tiene el siguiente formato:

$$X = \begin{pmatrix} IM_1 \\ IM_2 \\ IM_4 \end{pmatrix} \qquad \Lambda_X = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \qquad \xi = I \qquad \delta = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}$$

En cuanto a las variables endógenas, la incorporación de las 4 variables queda determinada a través de las siguientes ecuaciones:

$$\begin{cases} EXP_2 = \lambda_4 E + \epsilon_1 \\ EXP_3 = \lambda_5 E + \epsilon_2 \\ EXP_6 = \lambda_6 E + \epsilon_3 \\ Q_1 = \lambda_7 C + \epsilon_4 \\ Q_2 = \lambda_8 C + \epsilon_5 \\ Q_4 = \lambda_9 C + \epsilon_6 \\ Q_5 = \lambda_{10} C + \epsilon_7 \end{cases} \quad \begin{cases} S_1 = \lambda_{11} S + \epsilon_8 \\ S_2 = \lambda_{12} S + \epsilon_9 \\ S_3 = \lambda_{13} S + \epsilon_1 0 \\ S_6 = \lambda_{14} S + \epsilon_1 1 \\ L_1 = \lambda_{15} L + \epsilon_{12} \\ L_2 = \lambda_{16} L + \epsilon_{13} \end{cases} \quad \begin{cases} BB_4 = \lambda_{17} BB + \epsilon_{14} \\ BB_5 = \lambda_{18} BB + \epsilon_{15} \\ VP_1 = \lambda_{19} VP + \epsilon_{16} \\ VP_2 = \lambda_{20} VP + \epsilon_{17} \\ VP_3 = \lambda_{21} VP + \epsilon_{18} \\ VP_5 = \lambda_{22} VP + \epsilon_{19} \end{cases} \tag{4.12}$$

La representación matricial de este submodelo es:

$$Y = \Lambda_Y \eta + \epsilon \tag{4.13}$$

$$Y = \begin{pmatrix} EXP_2 \\ EXP_3 \\ EXP_6 \\ Q_1 \\ Q_2 \\ Q_4 \\ Q_5 \\ S_1 \\ S_2 \\ S_3 \\ S_6 \\ L_1 \\ L_2 \\ BB_4 \\ BB_5 \\ VP_1 \\ VP_2 \\ VP_3 \\ VP_5 \end{pmatrix} \quad \Lambda_Y = \begin{pmatrix} \lambda_4 & 0 & 0 & 0 & 0 & 0 \\ \lambda_5 & 0 & 0 & 0 & 0 & 0 \\ \lambda_6 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_7 & 0 & 0 & 0 & 0 \\ 0 & \lambda_8 & 0 & 0 & 0 & 0 \\ 0 & \lambda_9 & 0 & 0 & 0 & 0 \\ 0 & \lambda_{10} & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_{11} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{12} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{13} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{14} & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{15} & 0 & 0 \\ 0 & 0 & 0 & \lambda_{16} & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{17} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{18} & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{19} \\ 0 & 0 & 0 & 0 & 0 & \lambda_{20} \\ 0 & 0 & 0 & 0 & 0 & \lambda_{21} \\ 0 & 0 & 0 & 0 & 0 & \lambda_{22} \end{pmatrix} \quad \eta = \begin{pmatrix} E \\ C \\ VP \\ S \\ L \\ BB \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \\ \epsilon_{10} \\ \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{15} \\ \epsilon_{16} \\ \epsilon_{17} \\ \epsilon_{18} \\ \epsilon_{19} \end{pmatrix}$$

Antes de pasar a la etapa de estimación, debe corroborarse que el modelo formado por los submodelos aquí especificados esté correctamente identificado, para lo cual se utilizará la "regla de conteo" presentada en la subsección 4.2.1, la que establece que el modelo está identificado si la cantidad de parámetros a estimar en el modelo (t) es menor a la cantidad de elementos no redundantes en la matriz de varianzas y covarianzas (r), lo que

genera que el número de grados de libertad (df) del modelo, sea positivo. En este caso, $r = 253$, $t = 52$ y, por lo tanto, $df = 201$, lo que permite establecer que el modelo está identificado. En lo que refiere a las demás restricciones necesarias para la identificación del modelo, sólo resta aclarar que los parámetros fijados en 1 son: $\lambda_3, \lambda_5, \lambda_7, \lambda_{11}, \lambda_{15}, \lambda_{18}$ y λ_{19}, lo que quiere decir que se fijan las mismas restricciones que fijaron los investigadores portugueses (Alves y Raposo, 2007a).

En cuanto a la etapa que consiste en estimar el modelo, cabe recordar que el método utilizado es el de máxima verosimilitud. Bajo este método, se estima un modelo que presenta los siguientes indicadores de bondad de ajuste:

Índice	Modelo UP2
NFI	0.870
NNFI	0.870
CFI	0.887
RMSEA	0.097
SRMR	0.093

Tabla 4.9: Índices de bondad de ajuste - modelo UP2

En cuanto a la evaluación del modelo en términos generales, a partir de la tabla 4.9, puede verse que el ajuste mejora con respecto a los modelos anteriores, con índices de ajuste incremental que superan el 0.85 e índices de ajuste global inferiores a 0.1. De todas formas, no puede dejarse de lado el hecho de que estos indicadores son calculados bajo el supuesto de normalidad de las variables, supuesto que aquí no se verifica.

En lo que refiere a la evaluación de los dos submodelos que conforman el modelo global, en la tabla 4.10 pueden consultarse las estimaciones de los parámetros (factores de carga estandarizados y no estandarizados), junto al p-valor asociado a la prueba de significación (siempre asumiendo que se cumple el supuesto de normalidad multivariada de los datos). Puede verse que en el modelo estructural, hay tres relaciones que, al 5 % de significación, no pueden ser confirmadas, estas son: "Una de las causas del valor que perciben los estudiantes de la FCCEEyA, refiere a las expectativas con las cuales ingresan a esta"(β_2), "Las expectativas con las cuales un estudiante ingresa a facultad, tiene efectos directos sobre la satisfacción que sienta dicho estudiante con el servicio brindado por la

facultad"(β_4), y "La imagen que tiene el estudiante de la facultad, influye directamente sobre lo que este diga/piense, sobre la facultad"(γ_6).

Tanto las relaciones que se presentan en este modelo, como las estimaciones de los parámetros, pueden verse gráficamente en la figura 4.4.

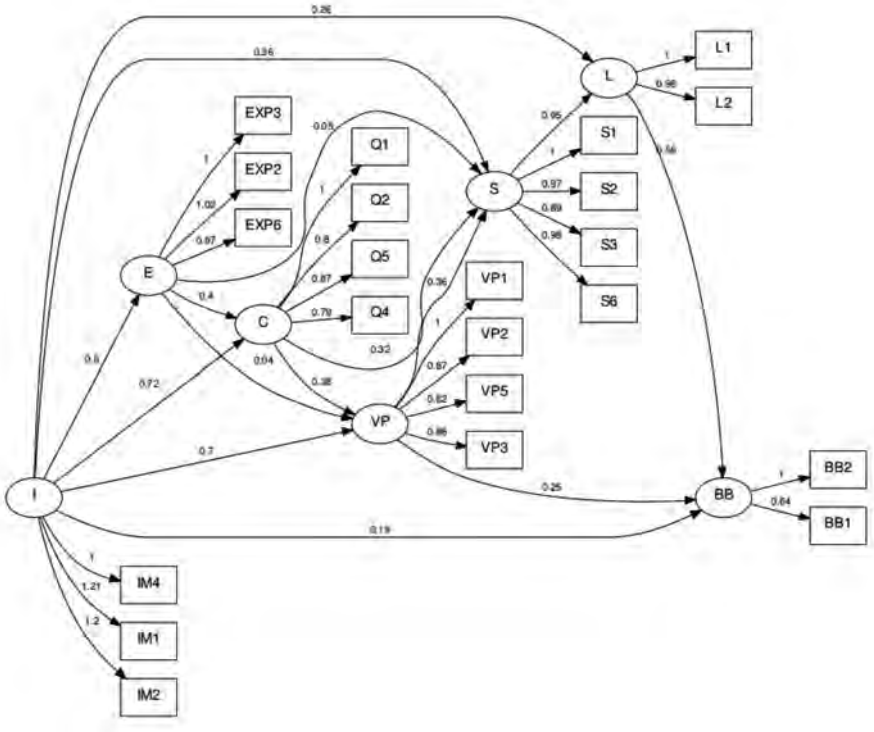


Figura 4.4: Diagrama de senderos - modelo UP2

Otra medida que resulta de interés al momento de evaluar el modelo estructural es la proporción de varianza de cada variable latente que logra ser explicada por las variables, también latentes, que la causan. Esta proporción queda determinada a partir del coeficiente de determinación R^2.

Parámetro	Estimación no estandarizada	Std Error	z value	Pr(> |z|)	Estimación estandarizada	Relación
λ_1	1.21	0.05	25.14	0.00	0.84	$IM_1 \leftarrow$ I
λ_2	1.20	0.06	19.74	0.00	0.72	$IM_2 \leftarrow$ I
λ_3	1	-	-	-	0.72	$IM_4 \leftarrow$ I
λ_4	1.02	0.04	23.64	0.00	0.83	$EXP_2 \leftarrow$ E
λ_5	1	-	-	-	0.79	$EXP_3 \leftarrow$ E
λ_6	0.97	0.04	23.09	0.00	0.81	$EXP_6 \leftarrow$ E
λ_7	1	-	-	-	0.90	$Q_1 \leftarrow$ C
λ_8	0.80	0.03	24.04	0.00	0.76	$Q_2 \leftarrow$ C
λ_9	0.79	0.04	21.89	0.00	0.72	$Q_4 \leftarrow$ C
λ_{10}	0.87	0.03	28.23	0.00	0.83	$Q_5 \leftarrow$ C
λ_{11}	1	-	-	-	0.93	$S_1 \leftarrow$ S
λ_{12}	0.97	0.03	38.03	0.00	0.90	$S_2 \leftarrow$ S
λ_{13}	0.89	0.03	32.03	0.00	0.84	$S_3 \leftarrow$ S
λ_{14}	0.98	0.03	30.71	0.00	0.83	$S_6 \leftarrow$ S
λ_{15}	1	-	-	-	0.86	$L_1 \leftarrow$ L
λ_{16}	0.98	0.05	21.25	0.00	0.75	$L_2 \leftarrow$ L
λ_{17}	0.64	0.03	22.12	0.00	0.72	$BB_4 \leftarrow$ BB
λ_{18}	1	-	-	-	0.96	$BB_5 \leftarrow$ BB
λ_{19}	1	-	-	-	0.82	$VP_1 \leftarrow$ VP
λ_{20}	0.87	0.03	25.62	0.00	0.83	$VP_2 \leftarrow$ VP
λ_{21}	0.88	0.04	24.38	0.00	0.80	$VP_3 \leftarrow$ VP
λ_{22}	0.82	0.04	22.22	0.00	0.76	$VP_5 \leftarrow$ VP
γ_1	0.80	0.06	13.30	0.00	0.63	E $\leftarrow$ I
γ_2	0.72	0.08	9.27	0.00	0.48	C $\leftarrow$ I
β_1	0.40	0.06	6.64	0.00	0.34	C $\leftarrow$ E
γ_3	0.70	0.09	7.40	0.00	0.45	VP $\leftarrow$ I
β_2	0.04	0.06	0.62	0.54	0.03	VP $\leftarrow$ E
β_3	0.38	0.06	6.37	0.00	0.36	VP $\leftarrow$ C
γ_4	0.36	0.10	3.68	0.00	0.23	S $\leftarrow$ I
β_4	-0.05	0.06	-0.77	0.44	-0.04	S $\leftarrow$ E
β_5	0.32	0.06	5.40	0.00	0.30	S $\leftarrow$ C
β_6	0.36	0.06	6.46	0.00	0.36	S $\leftarrow$ VP
γ_5	0.26	0.09	2.80	0.01	0.13	L $\leftarrow$ I
β_7	0.95	0.06	15.87	0.00	0.76	L $\leftarrow$ S
γ_6	0.19	0.10	1.95	0.05	0.10	BB $\leftarrow$ I
β_8	0.25	0.06	4.25	0.00	0.21	BB $\leftarrow$ VP
β_9	0.56	0.04	12.83	0.00	0.60	BB $\leftarrow$ L

Tabla 4.10: Estimaciones - modelo UP2

Variable	R^2
E	0.39
C	0.55
S	0.59
VP	0.59
L	0.73
BB	0.70

Tabla 4.11: R^2 para el modelo estructural

A partir de la tabla 4.11 puede concluirse que lealtad es el constructo que mejor queda explicado por las variables que lo preceden (imagen y satisfacción), las que logran explicar un 73 % de la variabilidad total del constructo. En segundo lugar se encuentra aquella variable que logra captar lo que los estudiantes piensan o dicen sobre la facultad, es decir, el boca a boca, el 70 % de la varianza total de este constructo queda explicado por las variables valor percibido, imagen y lealtad. En el otro extremo, se encuentran las expectativas, las cuales se entiende quedan determinadas por la imagen que tienen los estudiantes sobre la facultad, sin embargo esta parece no resultar suficiente, ya que solo logra captar un 39 % de la variabilidad total.

En cuanto al modelo de medida, lo primero que cabe destacar es que, tal como puede verse en la tabla 4.10, los parámetros estimados resultan todos significativamente (al 5 %) distintos de 0. Por otra parte se debe recordar este modelo se encuentra en el ámbito del análisis factorial, donde el concepto de comunalidad resulta fundamental. Este concepto refiere a la proporción de varianza original de cada variable observada, que queda explicada por el factor que satura sobre ella.

Para el bloque que contiene las variables referentes a las expectativas que tienen los estudiantes sobre la facultad, se tiene que la proporción de esta que queda explicada por el factor común es la que se presenta en la tabla 4.12, donde puede verse que el constructo expectativas logra explicar más del 60 % de la varianza original de cada una de las 3 variables.

Variable	Comunalidad
EXP_2	0.68
EXP_3	0.63
EXP_6	0.66

Tabla 4.12: Comunalidades del bloque expectativas

En cuanto a las variables que determinan la imagen que tienen los estudiantes de la FCCEEyA sobre esta, puede verse que las comunalidades toman valores entre 0.5 y 0.7 (ver tabla 4.13). La variable que mejor queda explicada por el constructo imagen es aquella que refiere a la visión general que tienen los estudiantes sobre la Universidad, como lugar donde estudiar (IM_1)

Variable	Comunalidad
IM_1	0.71
IM_2	0.51
IM_4	0.52

Tabla 4.13: Comunalidades del bloque imagen

Al considerar las variables que conforman el bloque referente a la evaluación que hacen los estudiantes sobre la calidad del servicio brindado por la facultad, la tabla 4.14 muestra que la variable observada que mejor queda explicada por la variable latente calidad es la que refiere a la calidad global de enseñanza (Q_1), ya que el 82 % de su varianza original es captada por el constructo. En el otro extremo, se encuentra la afirmación que hace referencia al ambiente académico (Q_4), donde el factor logra explicar el 52 % de la varianza original.

En lo que refiere al bloque de preguntas específicas sobre satisfacción, puede verse que este constructo logra captar una cantidad importante (entre un 60 % y un 87 %) de la varianza original de las variables observadas que lo conforman. Tal como se ve en la tabla 4.15, la variable que mejor queda explicada es la que refiere al grado en el que la facultad atendió las expectativas de los estudiantes (S_2).

Considerando aquellas variables que son entendidas como causantes de la satisfacción, sólo resta presentar el bloque de preguntas que refiere al valor percibido.

Variable	Comunalidad
Q_1	0.82
Q_2	0.58
Q_4	0.52
Q_5	0.68

Tabla 4.14: Comunalidades del bloque calidad

Variable	Comunalidad
S_1	0.60
S_2	0.87
S_3	0.80
S_6	0.70

Tabla 4.15: Comunalidades del bloque satisfacción

Tal como puede verse en la tabla 4.16, las cuatro variables observadas, aquí consideradas, que conforman el constructo valor percibido quedan bien explicadas por este.

Variable	Comunalidad
V_1	0.68
V_2	0.69
V_3	0.65
V_5	0.58

Tabla 4.16: Comunalidades del bloque valor percibido

En cuanto a las variables consideradas como consecuencias de la satisfacción, en la tabla 4.17, se presentan aquellas que conforman el constructo lealtad, donde se ve que la variable que mejor queda explicada por este es "Si tuviera que decidir nuevamente, volvería a elegir esta facultad" (L_1).

Variable	Comunalidad
L_1	0.73
L_2	0.56

Tabla 4.17: Comunalidades del bloque lealtad

Por último, en la tabla 4.18, puede verse que la varianza de aquellas variables que forman el constructo boca a boca resulta captada en gran proporción por este factor, sobre todo para la variable "Recomendaría esta facultad a un amigo" (BB_5).

Variable	Comunalidad
BB_4	0.52
BB_5	0.93

Tabla 4.18: Comunalidades del bloque boca a boca

Índice de satisfacción estudiantil - FCCEEyA

Considerando las estimaciones no estandarizadas de los factores de carga que vinculan el factor satisfacción con cuatro de las variables observadas que conforman este constructo, se calcula el índice de satisfacción estudiantil para la FCCEEyA, a partir de la fórmula propuesta por Fornell et al. (Claes et al., 1996).

A partir de lo expuesto en la tabla 4.19, el IS para la FCCEEyA es 69 %.

De esta forma concluye la exposición de los principales resultados obtenidos al considerar a todos los estudiantes por igual, sin ninguna distinción.

Variable Medida (y_i)	Estimación no estandarizada Uruguay (w_i)	Media variable medida ($\bar{y}_i$)	$w_i\bar{y}_i$
S_1	1.00	7.45	7.45
S_2	0.97	7.03	6.81
S_3	0.89	6.70	5.96
S_6	0.98	7.53	7.37
$Total$	3.84	-	27.59

Tabla 4.19: Índice de satisfacción estudiantil FCCEEyA - modelo UP2

Modelo UP2 - Considerando el sexo de los estudiantes

A continuación se presentan los principales resultados obtenidos al considerar dos grupos de estudiantes, en función de la variable sexo. La especificación del modelo, así como el método de estimación y las restricciones impuestas sobre cada uno de los submodelos, son las mismas que las consideradas en el modelo UP2.

Observando la tabla 4.20, los principales resultados a destacar son: con respecto al modelo de medida, tal como sucede al considerar a todos los estudiantes sin distinciones, todos los parámetros resultan significativamente distintos de 0, lo cual indica que las variables latentes efectivamente saturan sobre las variables observadas presentes en el modelo. En cuanto al modelo estructural cabe destacar que tanto para mujeres como para hombres, existen 3 relaciones que no logran ser confirmadas. En el caso de los hombres, estas relaciones son las mismas que no pueden ser confirmadas al considerar a todos los estudiantes juntos. Sin embargo, para el caso de las mujeres, dos de las relaciones que no se pueden confirmar también coinciden, pero la tercera no.

Como complemento a estos resultados se presenta, a continuación, el IS calculado por sexo. En las tablas 4.21 y 4.22 se encuentran los insumos necesarios para el cálculo.

El nivel de satisfacción estudiantil por sexo es de 70 % y 67 % para mujeres y hombres, respectivamente.

Si en lugar de considerar el sexo como variable auxiliar, se considera la carrera a la cual está inscripto el estudiante, se tienen los resultados presentados a continuación.

Parámetro	Estimación Mujeres	Std Error Mujeres	$Pr(>\|z\|)$ Mujeres	Estimación Hombres	Std Error Hombres	$Pr(>\|z\|)$ Hombres	Relación
λ_1	1.16	0.06	0.00	1.26	0.08	0.00	$IM_1 \leftarrow$ I
λ_2	1.16	0.07	0.00	1.24	0.09	0.00	$IM_2 \leftarrow$ I
λ_3	1	-	-	1	-	-	$IM_4 \leftarrow$ I
λ_4	1.03	0.06	0.00	0.99	0.06	0.00	$EXP_2 \leftarrow$ E
λ_5	1	-	-	1	-	-	$EXP_3 \leftarrow$ E
λ_6	0.96	0.05	0.00	0.97	0.06	0.00	$EXP_6 \leftarrow$ E
λ_7	1	-	-	1	-	-	$Q_1 \leftarrow$ C
λ_8	0.82	0.04	0.00	0.79	0.05	0.00	$Q_2 \leftarrow$ C
λ_9	0.78	0.05	0.00	0.82	0.05	0.00	$Q_4 \leftarrow$ C
λ_{10}	0.86	0.04	0.00	0.88	0.05	0.00	$Q_5 \leftarrow$ C
λ_{11}	1	-	-	1	-	-	$S_1 \leftarrow$ S
λ_{12}	1.00	0.03	0.00	0.91	0.04	0.00	$S_2 \leftarrow$ S
λ_{13}	0.88	0.04	0.00	0.89	0.04	0.00	$S_3 \leftarrow$ S
λ_{14}	1.04	0.04	0.00	0.91	0.05	0.00	$S_6 \leftarrow$ S
λ_{15}	1	-	-	1	-	-	$L_1 \leftarrow$ L
λ_{16}	0.98	0.05	0.00	0.93	0.07	0.00	$L_2 \leftarrow$ L
λ_{17}	0.65	0.04	0.00	0.61	0.04	0.00	$BB_4 \leftarrow$ BB
λ_{18}	1	-	-	1	-	-	$BB_5 \leftarrow$ BB
λ_{19}	1	-	-	1	-	-	$VP_1 \leftarrow$ VP
λ_{20}	0.93	0.04	0.00	0.80	0.05	0.00	$VP_2 \leftarrow$ VP
λ_{21}	0.84	0.04	0.00	0.93	0.06	0.00	$VP_3 \leftarrow$ VP
λ_{22}	0.80	0.05	0.00	0.84	0.05	0.00	$VP_5 \leftarrow$ VP
γ_1	0.74	0.07	0.00	0.90	0.09	0.00	E $\leftarrow$ I
γ_2	0.71	0.09	0.00	0.71	0.12	0.00	C $\leftarrow$I
β_1	0.39	0.07	0.00	0.40	0.09	0.00	C $\leftarrow$E
γ_3	0.81	0.12	0.00	0.60	0.15	0.00	VP $\leftarrow$I
β_2	-0.01	0.08	0.85	0.08	0.10	0.39	VP $\leftarrow$E
β_3	0.35	0.07	0.00	0.42	0.09	0.00	VP $\leftarrow$C
γ_4	0.34	0.12	0.01	0.37	0.15	0.01	S $\leftarrow$I
β_4	0.01	0.07	0.93	-0.11	0.09	0.23	S $\leftarrow$E
β_5	0.23	0.07	0.00	0.44	0.09	0.00	S $\leftarrow$C
β_6	0.43	0.07	0.00	0.28	0.08	0.00	S $\leftarrow$VP
γ_5	0.06	0.12	0.62	0.47	0.14	0.00	L $\leftarrow$I
β_7	1.11	0.07	0.00	0.80	0.09	0.00	L $\leftarrow$S
γ_6	0.37	0.12	0.00	-0.02	0.15	0.87	BB $\leftarrow$I
β_8	0.20	0.08	0.01	0.29	0.08	0.00	BB $\leftarrow$VP
β_9	0.49	0.05	0.00	0.67	0.07	0.00	BB $\leftarrow$L

Tabla 4.20: Estimaciones no estandarizadas por sexo - modelo UP2

Variable Medida (y_i)	Estimación (w_i)	Media variable ($\bar{y}_i$)	$w_i\bar{y}_i$
S_1	1.00	7.53	7.53
S_2	1.00	7.13	7.13
S_3	0.88	6.83	6.01
S_6	1.03	7.65	7.88
$Total$	3.91	-	28.55

Tabla 4.21: Índice de satisfacción estudiantil - modelo UP2 - Mujeres

Variable Medida (y_i)	Estimación (w_i)	Media variable ($\bar{y}_i$)	$w_i\bar{y}_i$
S_1	1.00	7.34	7.34
S_2	0.91	6.89	6.27
S_3	0.89	6.53	5.81
S_6	0.91	7.36	6.70
$Total$	3.71	-	26.12

Tabla 4.22: Índice de satisfacción estudiantil - modelo UP2 - Hombres

Modelo UP2 - Considerando la carrera de los estudiantes

De los 647 estudiantes que conforman la muestra, 500 están inscriptos a la carrera de Contador, mientras que los restantes 147 se encuentran repartidos entres las otras carreras de facultad. Es justamente esta la forma en la que serán tratados los estudiantes en este apartado, es decir la variable carrera tomará solo dos valores: “Contadores” y “Otras”.

En la tabla 4.23, se presentan los parámetros estimados diferenciando a los estudiantes por carrera. En lo que refiere al modelo de medida, todas los parámetros resultan significativamente distintos de 0, lo cual permite afirmar que las variables latentes efectivamente saturan sobre las variables observadas presentes en el modelo. En cuanto al modelo estructural lo más interesante a resaltar es que para el caso de los estudiantes de la carrera Contador, resulta imposible confirmar cuatro de las relaciones propuestas. Tres de estas coinciden con las que no podían confirmarse ni para el modelo conjunto, ni para el modelo que consideraba solo a los hombres, mientras que la cuarta coincide con aquella que sí se confirmaba en el modelo conjunto, pero no lo hacía al considerar

solo mujeres. En cuanto a los estudiantes de las restantes carreras, puede verse que seis de las relaciones establecidas en el modelo estructural, no pueden ser confirmadas al 5 % de significación. Tres de estas coinciden con las que no pueden se confirman para los Contadores, mientras que las otras tres no coinciden con ninguno de los escenarios antes planteados.

El último resultado a ser presentado es aquel que indica cuál es el nivel de satisfacción de los estudiantes, considerando la carrera a la cual están inscriptos. Los elementos necesarios para calcular el IS por carrera, se presentan en las tablas 4.24 y 4.25.

A partir de estos datos, puede verificarse que el nivel de satisfacción estudiantil prácticamente no difiere en función de la carrera de los estudiantes. Los estudiantes de la carrera de contador tienen un nivel de satisfacción del 69 % con la FCCEEyA, mientras que los estudiantes de las demás carreras, presentan un nivel de satisfacción un punto porcentual menor.

4.4. Conclusiones

En este trabajo, se relaciona el concepto de "satisfacción" del cliente, con la educación universitaria. En función de esto, se propone estudiar las propiedades psicométricas de un instrumento de medida propuesto para medir el nivel de satisfacción estudiantil en los cursos superiores de la Universidad de Beira Interior (Portugal), para ver si este resulta adecuado para el caso de la FCCEEyA (Uruguay). Planteado este objetivo general, se busca modelizar la satisfacción a través de la aplicación de modelos de ecuaciones estructurales, lo que implica por un lado, determinar cómo se relacionan un número reducido de factores, con una cantidad mayor de variables observadas (modelo de medida) y, por otro, tratar de confirmar algunas relaciones entre dichos factores (modelo estructural).

Como punto de partida se propone un modelo que contiene 40 variables observadas, las 45 variables del cuestionario candidatas a ser consideradas en este trabajo, menos aquellas que al ser eliminadas, producen un aumento en la consistencia interna de su bloque.

Parámetro	Estimación Contadores	Std Error Contadores	$Pr(>\|z\|)$ Contadores	Estimación Otras	Std Error Otras	$Pr(>\|z\|)$ Otras	Relación
λ_1	1.24	0.05	0.00	1.10	0.10	0.00	$IM_1 \leftarrow$ I
λ_2	1.21	0.06	0.00	1.19	0.14	0.00	$IM_2 \leftarrow$ I
λ_3	1	-	-	1	-	-	$IM_4 \leftarrow$ I
λ_4	0.97	0.05	0.00	1.18	0.08	0.00	$EXP_2 \leftarrow$ E
λ_5	1	-	-	1	-	-	$EXP_3 \leftarrow$ E
λ_6	0.94	0.05	0.00	1.12	0.09	0.00	$EXP_6 \leftarrow$ E
λ_7	1	-	-	1	-	-	$Q_1 \leftarrow$ C
λ_8	0.80	0.04	0.00	0.80	0.07	0.00	$Q_2 \leftarrow$ C
λ_9	0.80	0.04	0.00	0.82	0.08	0.00	$Q_4 \leftarrow$ C
λ_{10}	0.86	0.03	0.00	0.90	0.08	0.00	$Q_5 \leftarrow$ C
λ_{11}	1	-	-	1	-	-	$S_1 \leftarrow$ S
λ_{12}	0.98	0.03	0.00	0.92	0.05	0.00	$S_2 \leftarrow$ S
λ_{13}	0.88	0.03	0.00	0.93	0.05	0.00	$S_3 \leftarrow$ S
λ_{14}	1.01	0.03	0.00	0.89	0.06	0.00	$S_6 \leftarrow$ S
λ_{15}	1	-	-	1	-	-	$L_1 \leftarrow$ L
λ_{16}	0.98	0.04	0.00	0.91	0.12	0.00	$L_2 \leftarrow$ L
λ_{17}	0.66	0.03	0.00	0.56	0.07	0.00	$BB_4 \leftarrow$ BB
λ_{18}	1	-	-	1	-	-	$BB_5 \leftarrow$ BB
λ_{19}	1	-	-	1	-	-	$VP_1 \leftarrow$ VP
λ_{20}	0.86	0.03	0.00	0.91	0.08	0.00	$VP_2 \leftarrow$ VP
λ_{21}	0.87	0.04	0.00	0.96	0.09	0.00	$VP_3 \leftarrow$ VP
λ_{22}	0.81	0.04	0.00	0.85	0.08	0.00	$VP_5 \leftarrow$ VP
γ_1	0.85	0.07	0.00	0.65	0.11	0.00	E $\leftarrow$ I
γ_2	0.68	0.09	0.00	0.80	0.14	0.00	C $\leftarrow$I
β_1	0.43	0.07	0.00	0.31	0.11	0.01	C $\leftarrow$E
γ_3	0.73	0.10	0.00	0.57	0.19	0.00	VP $\leftarrow$I
β_2	0.03	0.07	0.69	0.13	0.12	0.27	VP $\leftarrow$E
β_3	0.40	0.06	0.00	0.32	0.12	0.01	VP $\leftarrow$C
γ_4	0.38	0.11	0.00	0.36	0.19	0.06	S $\leftarrow$I
β_4	-0.05	0.07	0.47	0.01	0.11	0.91	S $\leftarrow$E
β_5	0.31	0.06	0.00	0.18	0.11	0.11	S $\leftarrow$C
β_6	0.34	0.06	0.00	0.49	0.11	0.00	S $\leftarrow$VP
γ_5	0.07	0.10	0.48	0.91	0.21	0.00	L $\leftarrow$I
β_7	1.11	0.07	0.00	0.44	0.12	0.00	L $\leftarrow$S
γ_6	0.11	0.10	0.27	0.46	0.25	0.06	BB $\leftarrow$I
β_8	0.33	0.06	0.00	0.10	0.12	0.37	BB $\leftarrow$VP
β_9	0.54	0.04	0.00	0.49	0.13	0.00	BB $\leftarrow$L

Tabla 4.23: Estimaciones no estandarizadas por carrera - modelo UP2

Variable Medida (y_i)	Estimación (w_i)	Media variable ($\bar{y}_i$)	$w_i\bar{y}_i$
S_1	1.00	7.45	7.45
S_2	0.98	7.05	6.90
S_3	0.88	6.71	5.90
S_6	1.01	7.55	7.62
$Total$	3.87	-	27.89

Tabla 4.24: Índice de satisfacción estudiantil - modelo UP2 - Contadores

Variable Medida (y_i)	Estimación (w_i)	Media variable ($\bar{y}_i$)	$w_i\bar{y}_i$
S_1	1.00	7.44	7.44
S_2	0.92	6.95	6.39
S_3	0.93	6.68	6.21
S_6	0.89	7.45	6.63
$Total$	3.74	-	26.68

Tabla 4.25: Índice de satisfacción estudiantil - modelo UP2 - Otras

Una vez que se determina que estas serán las variables que conformarán el modelo de medida, se proponen 2 alternativas en cuanto al modelo estructural: las relaciones propuestas por los investigadores portugueses (Alves y Raposo, 2007a) (modelo UP), y las que se establecen en el ECSI (modelo ECSI).
El ajuste global de ambos modelos resulta "pobre", por lo que son descartados.

Como alternativa a estos, surgen dos nuevos modelos, que consideran la misma estructura de relaciones entre los factores pero menos cantidad de variables observadas. Por un lado, el modelo ECSI2, que considera el modelo estructural del modelo ECSI y 20 variables observadas y, por otro, el modelo UP2 que propone el mismo modelo estructural que el modelo UP y 22 variables observadas. Este último es el que mejor ajuste presenta.

4.4.1. Sobre la comparación Uruguay-Portugal

Los resultados encontrados al replicar el modelo exacto propuesto para la Universidad de Beira Interior (Portugal) obligan, primero que nada, a destacar las diferencias

metodológicas que existen entre ambos modelos. En primer lugar, y retomando lo ya planteado, los investigadores portugueses afirman (aunque no reportan resultados, ni indican como la testean) en su artículo "Conceptual Model of Student Satisfaction in Higher Education", que las variables observadas que forman parte de su modelo, sí provienen de una distribución normal multivariada. Por otra parte, el tamaño de muestra para el caso portugués es 4 veces mayor al utilizado para el caso de la FCCEEyA y se puede suponer, además, que en el caso Portugués se están considerando datos muy heterogéneos, ya que estos provienen de 13 facultades de diferentes campos de la educación (Alves y Raposo, 2007a).

En cuanto a los diseños muestrales utilizados, para el caso de la FCCEEyA se usó un diseño muestral complejo, el cual determinaba el manejo de pesos autoponderados, lo cual finalmente no resultó posible, ya que existió la necesidad de calibrar y trabajar con multiplicidad. Para el caso de Portugal, el diseño muestral proponía tener un número fijo de estudiantes por Universidad (250) pero esto no resulta (para 2 Universidades el tamaño se redujo sensiblemente).
De todas formas, si bien originalmente ambos diseños difieren, los resultados presentados en ambos casos se obtienen bajo el supuesto de un muestreo aleatorio simple.

En lo que refiere a los modelos estimados, los resultados obtenidos resultan similares. Las 18 variables que resultan significativas para el modelo de Portugal y que conforman el modelo de medida definitivo para la educación de dicho país, también resultan significativas para el caso uruguayo de la FCCEEyA. La principal diferencia entre ambos modelos se da en el modelo estructural, donde 3 de las relaciones determinadas en el caso portugués, no se confirman para el caso de Uruguay. Estas son: "Las expectativas que tienen los estudiantes sobre la facultad, influyen directamente sobre el valor percibido y sobre la satisfacción", y "La imagen que tienen los estudiantes de la facultad, determinan lo que estos piensan y dicen sobre ella".
En cuanto al índice de satisfacción de los estudiantes, utilizando el mismo modelo y las mismas variables para su cálculo, este indica que el nivel de satisfacción estudiantil es superior en el caso de la FCCEEyA.

4.4.2. Sobre los demás modelos

En cuanto a los modelos planteados luego de presentar aquel que replica exactamente el modelo portugués, y antes de encontrar aquel modelo considerado como definitivo para el caso uruguayo, el principal objetivo al plantearlos era tratar de reducir dimensiones, por lo que la mayor atención estuvo centrada en el modelo de medida. A partir de que se constata que esta disminución no resulta posible, se descarta la opción de reducir dimensiones con este mecanismo y se seleccionan, utilizando otro método, las variables observadas a considerar. Estas resultan ser 22, 18 de las cuales coinciden con las utilizadas por los investigadores portugueses y 4 más que surgen de tener en cuenta la consistencia interna dentro de cada bloque.

Una vez que se determina que estas 22 variables observadas serán las consideradas en el modelo de medida, solo resta establecer cuáles serán las relaciones que se intentarán confirmar en el modelo estructural. En este trabajo las opciones manejadas fueron dos; por un lado, se testearon las relaciones establecidas en el modelo europeo de satisfacción del cliente, que dan lugar al cálculo del ECSI, y por otro, las relaciones propuestas para el caso portugués. El modelo global, que incorpora a este último como modelo estructural, es el que presenta mejor ajuste para el caso de la FCCEEyA.

4.4.3. Consideraciones sobre el modelo seleccionado

Modelo general

Las estimaciones de los parámetros que conforman este modelo, el definitivo para el caso aquí estudiado, se presentan en la tabla 4.10. En lo que refiere al modelo estructural, se concluye que de las 15 relaciones propuestas, hay 3 que no pueden ser confirmadas. Estas son: “El valor percibido está determinado directamente por las expectativas de los estudiantes”, “Las expectativas de los estudiantes influyen sobre el nivel de satisfacción de estos”, siempre haciendo referencia a los servicios brindados por facultad, y “El boca a boca que se genera entre los estudiantes, se crea a partir de la imagen que estos tienen sobre la facultad”.

Considerando las relaciones que sí se confirman se puede concluir que de las variables que causan la percepción que tienen los estudiantes sobre la calidad de los servicios brin-

dados por facultad, la que tiene mayor peso es la imagen. Esta es, también, la variable que tiene mayor peso en la determinación del valor percibido. En lo que refiere al constructo satisfacción, la variable que más influye en la determinación de esta, es el valor percibido.
Por otra parte, el nivel de satisfacción resulta ser causa directa de la lealtad la cual, a la vez, tiene un efecto directo sobre el boca a boca que se genera entre los estudiantes.

Teniendo en cuenta todos estos aspectos, se calcula un índice de satisfacción estudiantil, el cual considera, para su cálculo, las variables observadas sobre las que satura la variable latente satisfacción. Este determina que el nivel de satisfacción estudiantil en los cursos superiores de FCCEEyA es de 69 %.

Consideraciones sobre el modelo que considera el sexo de los estudiantes

Al considerar los estudiantes en función de su sexo los resultados obtenidos permiten concluir, en primera instancia, que en cuanto al modelo de medida todas las variables incluidas en el modelo resultan significativas (al 5 %), lo que indica que las variables latentes efectivamente saturan sobre las variables observadas que la conforman.

En cuanto al modelo estructural, y las relaciones propuestas en él, para el caso de los hombres resulta imposible confirmar las mismas 3 relaciones que no se confirman a nivel general. Para el caso de las mujeres, dos de las relaciones que no se confirman, coinciden con las que no se confirman a nivel general; estas son las que determinan que las expectativas que tienen los estudiantes sobre la facultad, influyen directamente sobre el valor percibido y sobre la satisfacción. Además, para el caso de las mujeres no existe evidencia estadística que permita afirmar que la imagen que estas tienen sobre la facultad influye directamente sobre la lealtad.

Al medir el nivel de satisfacción por sexo, se constató que este resulta apenas superior en las mujeres que en los hombres, con valores de 70 % y 67 % respectivamente.

Consideraciones sobre el modelo que considera la carrera de los estudiantes

Cuando la distinción propuesta entre estudiantes, es en función de la carrera a la cual están inscriptos, los resultados en cuanto al modelo de medida son los mismos que a nivel global, es decir, que todas las variables incluidas en el modelo resultan significativas al 5 %, por lo que puede concluirse que las variables latentes efectivamente saturan sobre las variables observadas con las que se relaciona.

Al evaluar el modelo estructural estimado, se encuentran las mayores diferencias entre los estudiantes de distintas carreras. Para el caso de aquellos estudiantes que conforman el grupo "Contadores", se concluye que cuatro de las relaciones propuestas no logran ser confirmadas, si se considera $\alpha = 0{,}05$. De estas, tres coinciden con las que no se confirman a nivel general, mientras que la cuarta es aquella que propone que la lealtad de los estudiantes es causada por la imagen que estos tenían sobre la facultad.
En cuanto a los estudiantes inscriptos a otras carreras, la cantidad de relaciones que no logran ser confirmadas se incrementa en tres, con respecto a aquellas que no se confirman a nivel general. Estas tres, que no coinciden con ninguno de los escenarios antes presentados, son: "La imagen y la calidad son causas directas de la satisfacción" y "El valor percibido influye sobre el boca a boca que se genera entre los estudiantes".

El cálculo del índice de satisfacción estudiantil para los estudiantes en función de la carrera a la cual están inscriptos indica que este difiere, apenas en una unidad porcentual, entre aquellos que estudian para ser contadores y los demás (69 % y 68 %, respectivamente).

4.4.4. Consideraciones generales

En cuanto al objetivo principal de este trabajo, este apuntaba fundamentalmente a la evaluación de un instrumento de medida para determinar el nivel se satisfacción estudiantil para los cursos de educación superior de la FCCEEyA. Esto implicaba llevar a cabo la modelización de la satisfacción a través de la aplicación de modelos de ecuaciones estructurales, a partir de los cuales se generan nuevos objetivos.

Considerando el modelo de medida, y su desarrollo dentro del análisis factorial, la intención era que este sirviera para reducir dimensiones, sin embargo esto no resultó en ninguno de los escenarios planteados.
En lo que refiere al modelo estructural, el objetivo perseguido al plantearlo era ver si determinadas relaciones, tomadas tanto del ECSI como de las investigaciones portuguesas, se confirmaban para el caso de la FCCEEyA. De esto surgen conclusiones que apuntan por un lado, a la comparación directa con, por ejemplo, la Universidad de Beira Interior, las cuales establecen que existen diferencias en cómo se elabora el constructo satisfacción en ambos casos. Al considerar solo el caso de los estudiantes de la FCCEEyA y compararlos por sexo y por carrera, también surgen diferencias relevantes, en cuanto a cómo se entiende y determina la satisfacción.

Todos los resultados obtenidos en este trabajo, presentados en secciones previas, están basados en el supuesto de distribución multinormal de las variables observadas, y por la no consideración del diseño que generó la muestra que dio lugar a los datos aquí utilizados. Es por esto que las conclusiones, presentadas en esta sección, también están determinadas por estos dos aspectos.
En cuanto a la normalidad de las variables observadas, esta fue testeada a través de los estadísticos de simetría y kurtosis propuestos por Mardia (Kankainen et al., 2004), a partir de los cuales se rechazó la hipótesis de existencia de normalidad multivariada. Este resultado es el esperado considerando, por un lado, el tamaño de muestra y por otro, el hecho de que las variables observadas son variables discretas que toman valores en el intervalo [1 - 10], por lo que la normalidad nunca podría resultar más que una aproximación.

La violación de este supuesto, fundamental dentro del análisis factorial, influye sobre las estimaciones de los parámetros involucrados en el modelo, afectando directamente las decisiones que se tomen a partir de estas. De todas formas, en este trabajo se asume que sí existe multinormalidad de las variables y se presentan resultados, y por ende conclusiones, respaldados en este supuesto.

Bibliografía

Alves, H. y Raposo, M. (2004). La medición de la satisfacción en la enseñanza universitaria: El ejemplo de la Universidade da Beira Interior. International Review on Public and Nonprofit Marketing, 1(1):73–88.

Alves, H. y Raposo, M. (2007a). Conceptual model of student satisfaction in higher education. Total Quality Management, Vol. 18,(5):571–588.

Alves, H. y Raposo, M. (2007b). Student satisfaction index in portuguese public higher education. The Service Industries Journal, 27(6):795–808.

Blanco, R. J. y Blanco Peck, R. (2007). La medición de la calidad de servicios en la educación universitaria. Cuaderno de Investigación en la Educación, 22(7):121–136.

Casas Guillén, M. (2002). Los modelos de ecuaciones estructurales y su aplicación en el indice europeo de satisfacción del cliente. Reporte técnico, Facultad de Económicas, Universidad San Pablo, CEU.

Claes, F., Michael D., J., Eugene W., A., Jaesung, C., y Bryant, B. E. (1996). The american customer satisfaction index: Nature, purpose, and findings. Journal of Marketing, 60:7–18.

Dermanov, V. y Eklof, J. (2001). Using aggregate customer satisfaction index - challenges and problems on comparison with special reference to russia. Total Quality Management, 12(7-8).

García, A., Elena, M., Domínguez, C., y Jesús, A. (2006). Indices nacionales de satisfacción: Una vista general. En IV Congreso de Metodología de Encuestas. Pamplona, 20, 21 y 22 de septiembre de 200. Universidad Complutense de Madrid.

Kankainen, A., Taskinen, S., y Oja, H. (2004). On mardia's tests of multinormality. En Hubert, M., Pison, G., Struyf, A., y Van Aelst, S., editores, Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology, pp. 153–164. Birkhäuser Basel.

Kline, R. (2011). Principles and Practice of Structural Equation Modeling. The Guilford Press.

Stapleton, L. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. Structural Equation Modeling, 15(2):183–210.

Esta obra colectiva recopila el trabajo de cuatro jóvenes investigadores en el área de Estadística, donde se puede encontrar variedad de enfoques tanto metodológicos como de aplicaciones. La lectura de este libro los hará recorrer algunos enfoques específicos para abordar temas sociales, médicos, de política universitaria o de interés inmobiliario.

En el primer capítulo se comparan técnicas de clustering basadas en modelos probabilísticos, recorriendo diferentes tipologías de datos y considerando la posible presencia de outliers. Estas técnicas se aplican a un conjunto de inmuebles de Montevideo y Canelones con el objetivo de agrupar propiedades. En el segundo capítulo se aborda una forma de tratamiento de la no respuesta en encuestas de panel. Esa metodología se aplica a un conjunto de mujeres que fueron parte de la encuesta sobre situaciones familiares y desempeños sociales de Montevideo y área metropolitana realizada en 2001 y 2008.
El estudio del deterioro de la función cognitiva global en los adultos mayores es un tema que ocupa cada vez más la agenda de investigación. Mediante el uso de estudio Mini Mental State Examination como herramienta para cuantificar el deterioro cognitivo, se aborda el problema del abandono de los sujetos en el panel. El trabajo se centra en el uso de modelos donde el tiempo de sobrevida de los sujetos y los resultados del MMSE están relacionados.
En el último capítulo la autora redimensiona el concepto de satisfacción, abandonando su connotación estática. Aplica esta conceptualización a la educación universitaria mediante la aplicación de análisis factorial y modelos estructurales para evaluar el instrumento seleccionado. Las propiedades del instrumento de medición de la satisfacción estudiantil se analizan sobre una muestra de estudiantes de la Facultad de Ciencias Económicas y Administración.

ISBN: 978-9974-0-1550-0