Cuantificación del riesgo global del asegurado para mejorar la tarificación

Alemar Elaine Padilla Barreto

Tesis presentada para obtener el título de Doctora en Estadística e Investigación Operativa por la Universidad Politécnica de Cataluña

Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas y Estadística
Universidad Politécnica de Cataluña

Departamento de Econometría, Estadística y Economía Aplicada
Facultad de Economía y Empresa
Universidad de Barcelona

Directoras:
Montserrat Guillen Estany
Catalina Bolancé Losilla

Tutor:
Josep Ginebra

Barcelona, Marzo 2019

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAT DE BARCELONA

Cuantificación del riesgo global del asegurado para mejorar la tarificación

Alemar Elaine Padilla Barreto

Agradecimientos

Esta Tesis es el resultado de tres años de formación académica, profesional y personal. Tres años durante los cuales adquirí innumerables conocimientos teóricos, prácticos y personales. Todas y cada una de las experiencias vividas han sido absolutamente enriquecedoras y sin lugar a dudas no hubiesen sido posibles sin el acompañamiento de todos los que han formado parte de esta aventura maravillosa.

«El objetivo del Plan de Doctorados Industriales es contribuir a la competitividad y la internacionalización de la industria catalana, reforzar los instrumentos para captar el talento que genera el país y situar a los futuros doctores en condiciones de desarrollar proyectos de I+D+i en una empresa. El elemento esencial del proceso de doctorado industrial es el proyecto de investigación de la empresa o institución en la que el doctorando desarrolla su formación investigadora, en colaboración con una universidad o centro de investigación, y que es objeto de una tesis doctoral. Los doctorados industriales actúan, de este modo, como puentes de transferencia de conocimiento y contribuyen a estrechar las relaciones entre el tejido industrial de Cataluña y las universidades y los centros de investigación.» Agradezco la financiación recibida del programa de Doctorados Industriales de la Secretaría de Universidades e Investigación de la Generalitat de Cataluña y de la Agencia de Gestión de Ayudas Universitarias y de Investigación (AGAUR).

En primer lugar quiero agradecer a mis directoras Cati y Montse, por su profesionalidad, dedicación, apoyo incondicional y sobre todo su gran corazón, son las mejores!. Ellas han sido mis guías y centinelas durante todos estos años de trabajo. Desde el primer momento y con mucha claridad estructuraron lo que sería nuestra hoja de ruta, plantearon la estrategia que seguiríamos y fijaron los tiempos precisos para que pudiésemos cumplir con nuestros objetivos. Gracias por enseñarme a centrarme en lo importante, definir estrategias y ser lo suficientemente flexible y versátil como para poder avanzar continuamente y materializar todos nuestros esfuerzos. Gracias por su infinita paciencia y por toda la confianza que depositaron en mí. Trabajar con ustedes ha sido una de las experiencias más gratificantes de mi vida, siempre les estaré infinitamente agradecida.

También quiero agradecer a la Profesora Mercedes Ayuso por su apoyo incondicional desde el primer momento, por toda su confianza, enseñanzas y por todas las oportunidades que me ha brindado.

Muchas gracias a la Profesora Manuela Alcañiz por todo su cariño y todas las conversaciones tan interesantes y enriquecedoras que tuvimos durante todos estos años.

Gracias a toda la familia del Riskcenter, por todas las interesantes, divertidas y a veces polémicas conversaciones. En especial gracias a Miguel, Estefanía, Luis, Helena, Ana, Oscar, Rosina y María. Ramón gracias también por tu apoyo. Coloma e Isabel, gracias por siempre estar allí y ayudarme con lo que hiciera falta.

Un doctorado dirigido entre dos universidades es posible cuando se tiene el apoyo indicado. Gracias a mi tutor Josep Ginebra por todas sus recomendaciones y a Carme Macias por ayudarme con todas las gestiones durante todo este tiempo.

Mi experiencia profesional adquirida en el sector seguros ha sido fundamental en el resultado final de este trabajo. Así que, agradezco a todas las personas que me apoyaron y con las que tuve el gusto de trabajar en Zurich.

Asentarse fuera de su tierra natal es una experiencia sin lugar a dudas retadora, que se hace más llevadera con la compañía de amigos maravillosos. Gracias a todos mis amigos tanto en Venezuela como en España por formar parte de esta experiencia de vida. Llevo su esencia junto a mi a donde quiera que voy.

Marcin llegaste a mi vida para llenarla de amor incondicional, momentos inolvidables y un sin fin de aprendizajes, gracias por existir. Gracias por tu infinita paciencia, por ser mi soporte, mi fortaleza, por siempre animarme y ser mi compañero inseparable durante toda esta aventura. A tu lado todo es especial.

Quiero agradecer a toda mi hermosa familia, especialmente a mis papás Morelia y Cirilo, a mi tía Elena, a mis hermanas Alejandra y Elaine, a mi prima Fabiola y a mi sobrino Unai, mi angelito precioso gracias por enseñarme tanto en tan poco tiempo. Gracias por existir y por todo su amor y apoyo incondicional, llenan mi vida de alegría e ilusión día tras día, sin ustedes nada de esto sería posible. Desde Caracas hasta el País Vasco y de vuelta a Barcelona me enseñan que las distancias se acortan cuando existe el amor.

Por último gracias a mi abuela María Catalina, eres el mejor ejemplo de que el amor y las enseñanzas viven más allá del tiempo. Siempre estas presente en mi mente y mi corazón.

Resumen

La primera contribución de esta Tesis es la creación de una nueva medida de riesgo que permite clasificar a los clientes en función del riesgo que éstos aportan a la compañía. Esta medida definida a partir de la información agregada a nivel cliente es fácil de entender, implementar y además, tiene en cuenta la propensión a la renovación, la siniestralidad y la vinculación existente entre las diferentes pólizas contratadas por el asegurado.

Posteriormente nos centramos en uno de los factores que afectan a la medida de riesgo definida, concretamente en el análisis de la propensión a la renovación de las pólizas, lo que supone la segunda contribución del trabajo. En primer lugar, desde un punto de vista univariante y para dos líneas de negocio diferentes, que se asumen independientes la una de la otra - Hogar y Auto -, se evalúa la capacidad predictiva de diversos modelos de deserción alternativos en función de distintos criterios de ajuste (se denominan "modelos de deserción" a aquellos que permiten evaluar la probabilidad de que el cliente cancele sus pólizas y abandone la compañía). Se establece un análisis comparativo entre un modelo clásico de regresión univariante y dos modelos de aprendizaje de máquinas o *machine learning*. Los criterios de ajuste propuestos en esta parte representan una herramienta sencilla y práctica para la elección del *threshold* a partir del cual es posible clasificar a los clientes según su propensión a la renovación o no de sus pólizas.

La modelización conjunta de la deserción (en más de un tipo de póliza) puede ser considerada como nuestra tercera contribución. En ésta parte analizamos la propensión a la renovación de las pólizas de clientes con dos tipos de riesgos asegurados simultáneamente. Es decir, tenemos en cuenta la dependencia existente entre líneas de negocio y mostramos el efecto que supone la existencia de dependencia sobre las decisiones de renovación en ambos tipos de póliza -Hogar y Auto-.

Nuestra cuarta contribución plantea diversas reflexiones sobre el cambio que supone para las compañías aseguradoras la digitalización de la información. Además, planteamos una aplicación práctica donde, a partir del uso de modelos multinomiales, se obtiene la

propensión simultánea a la renovación o no de las pólizas que un mismo cliente tiene aseguradas con la misma compañía.

Las contribuciones siguientes corresponden a la segunda parte de la Tesis y están directamente relacionadas con la cuantificación del riesgo en finanzas y seguros. En este sentido, una primera aportación consiste en analizar el efecto que tiene sobre la estimación del valor en riesgo la selección del modelo de dependencia. En la misma línea, nuestra otra aportación en esta parte analiza las debilidades y fortalezas de los métodos no paramétricos en la cuantificación del riesgo.

Abstract

The first contribution of this Thesis is the creation of a new risk measure that allows to classify the clients according to the risk that they bring to the company. This measure defined from the aggregate information at the client level is easy to understand, implement and also takes into account the propensity to renew, the claims and the relationship between the different policies subscribed by the insured.

Subsequently we focus on one of the factors that affect the risk measure defined previously, specifically in the analysis of the propensity to renewal, which is the second contribution of this work. First, from an univariate point of view and for two different lines of business, which are assumed independent of each other - Home and Motor -, the predictive capacity of different alternative churn prediction models is evaluated according to different adjustment criteria ("churn models") are those that allow evaluating the probability that clients cancel their policies and leave the company). A comparative analysis is established between a classic univariate regression model and two machine learning models. The adjustment criteria proposed in this part represent a simple and practical tool in order to choose the threshold from which it is possible to classify clients according to their propensity for renewal.

The joint modeling for customer lapses (in more than one type of policy) can be considered as our third contribution. In this part we analyze the propensity to renew the policies of customers with two types of insured risks simultaneously. That is, we take into account the existence of dependence between lines of business and show how the existence of dependence affects decisions about renewal in both types of policy - Home and Motor.

Our fourth contribution raises several reflections about the change that supposes for the insurance companies the digitalization of the information. In addition, we propose a practical application where, from the use of multinomial models, we obtain the simultaneous propensity to renew or not the policies that the same client has underwritten in the same company.

The following contributions correspond to the second part of the Thesis and are directly related to the risk quantification in finance and insurance. In this sense, a first contribution consists of analyzing the effect that the selection of the dependency model has on the estimation of value at risk. In the same vein, the second contribution in this part analyzes the weaknesses and strengths of nonparametric methods in the quantification of risk.

Resum

La primera contribució d'aquesta Tesi és la creació d'una nova mesura de risc que permet classificar als clients en funció del risc que aquests aporten a la companyia. Aquesta mesura definida a partir de la informació agregada a nivell client és fàcil d'entendre, implementar i a més, té en compte la propensió a la renovació, la sinistralitat i la vinculació existent entre les diferents pòlisses contractades per l'assegurat.

Posteriorment ens centrem en un dels factors que afecten la mesura de risc definida, concretament en l'anàlisi de la propensió a la renovació de les pòlisses, fet que suposa la segona contribució del treball. En primer lloc, des d'un punt de vista univariant i per a dues línies de negoci diferents, que s'assumeixen independents l'una de l'altra - Llar i Auto -, s'avalua la capacitat predictiva de diversos models de deserció alternatius en funció de diferents criteris de ajust (s'anomenen "models de deserció" a aquells que permeten avaluar la probabilitat que el client cancel•li les seves pòlisses i abandoni la companyia). S'estableix una anàlisi comparativa entre un model clàssic de regressió univariant i dos models d'aprenentatge de màquines o *machine learning*. Els criteris d'ajust proposats en aquesta part representen una eina senzilla i pràctica per a l'elecció del nivell o threshold a partir del qual és possible classificar els clients segons la seva propensió a la renovació o no de les seves pòlisses.

La modelització conjunta de la deserció (en més d'un tipus de pòlissa) pot ser considerada com la nostra tercera contribució. En aquesta part analitzem la propensió a la renovació de les pòlisses de clients amb dos tipus de riscos assegurats simultàniament. és a dir, tenim en compte la dependència que hi ha entre línies de negoci i mostrem l'efecte que suposa l'existència de dependència sobre les decisions de renovació en tots dos tipus de pòlissa -Llar i Auto-.

La nostra quarta contribució planteja diverses reflexions sobre el canvi que suposa per a les companyies asseguradores la digitalització de la informació. A més, plantegem una aplicació pràctica on, a partir de l'ús de models multinomials, s'obté la propensió

simultània a la renovació o no de les pòlisses que un mateix client té assegurades amb la mateixa companyia.

Les contribucions següents corresponen a la segona part de la Tesi i estan directament relacionades amb la quantificació del risc en finances i assegurances. En aquest sentit, una primera aportació consisteix a analitzar l'efecte que té sobre l'estimació del valor en risc la selecció del model de dependència. En la mateixa línia, la nostra altra aportació en aquesta part analitza les debilitats i fortaleses dels mètodes no paramètrics en la quantificació del risc.

Índice general

Lista de figuras

Lista de tablas

Artículos

Los capítulos que conforman ésta Tesis Doctoral han sido parcialmente publicados en los siguientes trabajos:

- Bolancé, C., Alemany, R., y Padilla-Barreto, A. E. (2018). Impact of D-vine structure on risk estimation. Journal of Risk, 20(5):1-32.
- Padilla-Barreto, A. E. (2018). Joint modelling for customer lapses in the insurance sector. En Sarabia, J.M., Prieto, F., y Guillen, M., editores, Contributions to Risk Analysis: RISK 2018, pp. 219-226. Cuadernos de la Fundación Mapfre, Madrid.
- Padilla-Barreto, A. E., Guillen, M., y Bolancé Losilla, C. (2017). Big-data analytics en seguros. Anales del Instituto de Actuarios Españoles, 23:1-19.
- Alemany, R., Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016). Combining parametric and non-parametric methods to compute value-at-risk. Economic Computation & Economic Cybernetics Studies & Research, 50(4):61-74.
- Padilla-Barreto, A. E., Bolancé, C., y Guillen, M. (2016). Cuantificación del riesgo para la tarificación en seguros de automóvil. Anales del Instituto de Actuarios Españoles, 22:1-24.
- Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016). Predicting defection in non-life motor and home insurance. Lectures on Modeling and Simulation: a selection from AMSE 2016, 2:107-120.
- Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016). Predicting probability of customer churn in insurance, en León, R. et al. (ed.) Modelling and Simulation in Engineering Economics and Management MS 2016. Lecture Notes in Business Information Processing, 254:82-91. Springer International Publishing, Cham.

Los co-autores de dichos trabajos, conocen y están de acuerdo en que sean utilizados en ésta Tesis. En la Sección 1.4 de la Introducción se detalla el contenido y capítulos a los que corresponde cada publicación.

Capítulo 1

Introducción

1.1. Antecedentes

El objetivo general de esta Tesis es analizar una medida de riesgo global, que tenga en cuenta las dependencias entre los comportamientos del asegurado en sus distintas pólizas contratadas y permita cuantificar la contribución individual de cada asegurado en el riesgo de pérdida o ganancia de una compañía de seguros.

El sector de los seguros está compuesto por dos representantes básicos: los tomadores y los aseguradores. Los tomadores son los clientes, quienes suscriben pólizas de seguro con las compañías, introduciendo de esta manera sus riesgos dentro de un grupo de riesgo compartido. Los aseguradores son los propietarios de las compañías de seguros y se preocupan por las situaciones adversas en las que las pérdidas de la compañía puedan exceder el valor esperado.

Desde un punto de vista económico, el proceso de producción de un contrato de seguro sigue lo que se conoce como un "ciclo invertido". El precio tiene que ser fijado antes de que el coste del producto se dé a conocer. Por ejemplo, en los procesos de producción industrial, en primer lugar, el costo de la creación, fabricación y venta de un producto es conocido y, luego, de acuerdo con este coste y según la demanda existente, el precio se fija posteriormente. Por su parte, en el sector de los seguros, el precio final de la prima depende, entre otras cosas, de la ocurrencia de accidentes, lo cual no puede ser determinado a priori.

En términos estadísticos, el problema es bastante simple. El resultado de un contrato de seguro es o una pérdida o un beneficio, dependiendo de si el asegurado sufre un accidente o no y dependiendo de la severidad del accidente. Si el accidente ocurre y

está cubierto por el contrato, entonces la compensación recibida por el asegurado tiene que ser pagada por la compañía de seguros y esta cantidad puede ser mucho mayor que la prima recibida al inicio del contrato.

Hay muchas fuentes que estudian los fundamentos de los seguros (ver, por ejemplo, Williams y Heins, 1985; Dionne et al., 2000; Reavis, 2012; Kunreuther et al., 2013), de hecho este campo de investigación es tan amplio que constituye una disciplina en sí misma (ver Newhouse et al., 1993, donde se detallan los resultados de uno de los proyectos de investigación más emblemáticos en seguros de salud, "The RAND Health Insurance Study"). Además, de la existencia de un gran componente financiero en la gestión de una empresa de seguros, el concepto de seguro tiene fundamentos estadísticos, esencialmente porque todo el proceso se basa en la predicción de un resultado incierto. No hay duda de que los recientes avances en el análisis de grandes cantidades de datos han cambiado la forma en la que en el sector seguros se aborda el problema de encontrar el precio correcto de un contrato de seguros. El precio tiene que ser justo con el cliente, por lo que debe tener en cuenta las características específicas del contrato y del asegurado. Además, el precio también debe ser suficiente para garantizar que la empresa permanecerá solvente. También, se debe tener en cuenta que las compañías de seguros invierten su riqueza en el mercado financiero, pero por lo general son inversores muy conservadores, de manera que prefieren los bonos y bienes inmuebles en lugar de las acciones altamente volátiles. Algunos productos de seguros también pueden proteger a los clientes de las dificultades financieras, por ejemplo, el seguro de crédito cubre a una empresa que pueda tener clientes que no pagan sus facturas. Otro ejemplo de ello son los planes de pensiones destinados a proporcionar a los clientes un flujo estable de ingresos económicos durante su retiro.

Las compañías de seguros son instituciones altamente reguladas en todo el mundo. Una compañía de seguros no puede vender productos a menos que haya sido autorizada por el supervisor. En España esta supervisión se lleva a cabo por la Dirección General de Seguros y Fondos de Pensiones, una agencia oficial que depende del Ministerio de Economía. También, hay un reglamento europeo que es aplicable en España desde enero de 2016. Esta directiva es conocida como Solvencia II y establece normas sobre los requisitos legales que se establecen con el fin de permitir que las empresas aseguradoras operen en el territorio europeo.

La necesidad de cuantificar el riesgo justifica el enfoque de esta tesis. Una vez que la compañía de seguros establece una tarifa, el riesgo de pérdida asociado a la cartera de seguros de la compañía necesita ser cuantificado. Las pérdidas también deben considerar los posibles cambios futuros en las primas, teniendo en cuenta el riesgo de accidentes

y el riesgo de deserción. El proceso de cuantificación de riesgo ha de ser continuo y revisado periódicamente. Al modelar las pérdidas esperadas y la probabilidad de deserción para cada cliente, el análisis estadístico clásico se basa en la información proporcionada sobre el precio que ha sido pagado por el cliente. Sin embargo, si el precio cambia, las pérdidas y ganancias cambian según corresponda y las probabilidades de permanecer en la empresa también cambian. Un nuevo enfoque podría consistir en estimar el precio, la pérdida esperada y la probabilidad de deserción de forma simultánea.

Cuantificación de riesgo

Las compañías se preocupan, entre otras cosas, por el riesgo de pérdida, especialmente para los seguros de no vida. Esto por lo general se denomina *riesgo de suscripción* y se define como el riesgo de que las primas recibidas no sean suficientes para cubrir las pérdidas que se producen más tarde, una vez que todos los demás factores se han cubierto. En general, se considera una perspectiva de un año, es decir, las decisiones sobre la forma en que las primas y las pérdidas son analizadas tienen en cuenta un período de doce meses.

Para simplificar el concepto de pérdidas y ganancias en un contrato de seguro anual, consideremos un cliente que compra un seguro de automóvil por un año. La compañía de seguros tiene que cubrir los gastos administrativos, debido a las exigencias reglamentarias, la publicidad y los sistemas de Tecnología de Información (*Information technology* - IT). En otras palabras, una empresa tiene que tomar parte de la prima para cubrir los gastos generales que se derivan de las operaciones ordinarias. Al mismo tiempo, otra parte de la prima cobrada puede ser invertida en el mercado financiero y puede producir retornos a la empresa antes de que se necesiten los recursos financieros para pagar la compensación a ese cliente en particular. Una aseguradora que vende seguros de coche es probable que distribuya miles de contratos anuales y tenga que pagar por las compensaciones a aquellos clientes que sufran un accidente.

En este punto, se entiende fácilmente que la fijación de precios adecuada es esencial para garantizar la estabilidad y la solidez financiera de una compañía de seguros. El concepto de fijación de precios se refiere a sólo una parte de toda la producción de un contrato de seguro, específicamente al establecimiento de un modelo que se utiliza para calcular el precio del contrato, usualmente llamado “prima”, que un cliente determinado, con características dadas tiene que pagar para conseguir el contrato. El precio depende de las características del contrato (es decir, por ejemplo no es lo mismo comprar una póliza de seguro para un coche pequeño que para un coche grande, así como tampoco es lo mismo contratar un seguro que cubra cualquier riesgo a uno que cubra un riesgo básico). También depende de las características del cliente (por ejemplo, un conductor joven

paga un precio más alto que un conductor experimentado, pues es bien sabido que en general el riesgo de tener un accidente de coche es más alto en el primero que en el segundo). Por lo tanto, en todo lo que sigue, vamos a suponer que el contrato de seguro ya está diseñado y que no tenemos que tener en cuenta información adicional, como los gastos generales de la empresa, nos limitaremos a concentrarnos en la parte del precio que tiene que ver con las circunstancias alrededor del objeto que se está asegurando y la persona que asegura ese objeto. En éste trabajo, nos concentraremos específicamente en las pólizas de seguros para automóviles y para viviendas.

En la evaluación global del riesgo de un contrato de seguro, también sabemos que la lealtad del cliente es, de lejos, una de las prioridades más importantes en la mayoría de las compañías de seguros, debido a que cada día los asegurados deciden abandonar la empresa y cambiarse a la competencia. Cuando un cliente decide cancelar un contrato de seguro para buscar otra compañía, la aseguradora original pierde la oportunidad de generar beneficios futuros. Esta parte del riesgo se conoce como *riesgo de caída de cartera* y también se considera explícitamente en la regulación de Solvencia II para las compañías aseguradoras.

Hasta ahora sólo unos pocos autores han considerado la suscripción y el riesgo de deserción juntos, a pesar de estar bien conectados. De hecho, tal como se evidencia en el trabajo de Guelman y Guillen (2014) es un hecho que cuanto mayor sea el precio de la prima a pagar, mayor será la probabilidad de que el cliente abandone la compañía, incluso si el efecto es heterogéneo en función de algunas covariables observadas. Por el contrario, un precio bajo es un incentivo para que un cliente mantenga la vigencia de su contrato con la empresa, pero si el precio es excesivamente bajo para demasiados clientes, entonces el flujo de ingresos puede ser insuficiente para cubrir las pérdidas esperadas y la empresa que disminuya demasiado los precios puede estar en riesgo de volverse insolvente.

En este contexto es fundamental analizar de manera conjunta ambos comportamientos, los asociados con las reclamaciones (incluyendo su severidad) y la probabilidad de comprar o renovar un producto de seguros. Algunos ejemplos de este tipo de análisis se muestran en (Thuring et al., 2012, 2013).

El sector seguros

El sector de seguros es el mayor inversor institucional en la Unión Europea, lo que contribuye a su crecimiento y desarrollo económico, entre muchas otras cosas. [1] Este

[1] Insurance Europe (2016). European insurance - key facts. https://www.insuranceeurope.eu/sites/default/files/attachments/European%20Insurance%20-%20Key%20Facts%20-%20August% 202016.pdf. Consultado: 2017-03-24.

sector está compuesto por las compañías de seguros que pueden ser clasificadas por el tipo de pólizas de seguros que venden. En general, las aseguradoras pueden comercializar seguros de vida, no vida o ambas cosas. Los productos de vida se refieren a los contratos de seguro que compensan a los sobrevivientes en caso de un accidente o los productos que proveen pagos mientras que el cliente está vivo, tales como lo son las rentas vitalicias (pensiones). Por su parte, los productos de no vida son los contratos para proteger objetos, propiedades o pasivos. Un pasivo se refiere a la responsabilidad, la deuda u obligación legal que surge de una situación dada.

En esta tesis, nos centraremos en los seguros de no vida, específicamente en líneas personales. Dos de los productos más importantes en líneas personales son: el producto de auto, en donde se encuentra el mayor mercado de seguros de no vida, y el producto de hogar. En conjunto, estas dos líneas de negocio representan más del 50 % de todas las primas de no vida. De manera que, comprender en detalle las características de estos dos productos es crucial para las compañías de seguros que venden este tipo de pólizas.

Un aspecto a tener en cuenta es que el seguro de automóvil con una cobertura de responsabilidad civil es obligatorio en casi todo el mundo, debido a que los conductores pueden causar accidentes y ocasionar daños materiales y/o corporales a terceros. El seguro protege a otras personas y sus bienes. La idea es que en caso de accidente, una tercera persona inocente sufriría como resultado de la culpa de un conductor y no sería capaz de compensar las pérdidas sufridas. Por esta razón, dada su obligatoriedad, el número de asegurados en una compañía de una cierta dimensión puede alcanzar una cifra de centenares de miles de pólizas.

Una idea interesante es la sugerida por Gourieroux y Jasiak (2011). En una compañía de seguros, un asegurado es un individuo que asume riesgos y se protege así mismo. Pero al mismo tiempo, es visto como riesgoso, debido a que otros individuos y la propia empresa pueden verlo como alguien que puede generar pérdidas.

1.2. Revisión de la literatura

La mayoría de las contribuciones en seguros de no vida, centran su atención en el área de tarificación (*pricing*) y reservas (*reserving*), ambos enfoques son claves a nivel empresarial, debido a que sus resultados impactan directamente sobre la rentabilidad y estabilidad de la compañía. En *pricing* la modelización de la siniestralidad esperada es fundamental para la previsión de riesgos futuros; mientras que en *reserving* es esencial

el cálculo de las provisiones necesarias para dar respuesta a las reclamaciones pendientes. Por otra parte, algunas publicaciones (ver por ejemplo Crosby y Stephens, 1987; Fornell, 1992; Guillen et al., 2008) también enfocan el análisis de los seguros pero desde la perspectiva del cliente, de manera que tienen en cuenta aspectos como su satisfacción, fidelidad y rotación dentro de la cartera. En éste trabajo nos centraremos en la cuantificación del riesgo individual de los asegurados combinando ambas perspectivas.

En el sector seguros, la modelización estadística juega un papel estelar. En la práctica, sus aplicaciones son diversas y su uso tiene un carácter multidisciplinar (ver Frees et al., 2014, para profundizar en la modelización estadística a nivel actuarial). En la actualidad, los modelos lineales generalizados son los modelos por excelencia utilizados a nivel actuarial. El término "*generalized linear model*"(GLM) fue introducido por primera vez por Nelder y Wedderburn (1972) y más tarde extendido por McCullagh y Nelder (1983) en su libro "*Generalized Linear Models*", donde ilustraron mediante ejemplos algunas áreas de aplicación. Desde entonces y hasta la actualidad, su uso continua vigente en ciencias actuariales. Haberman y Renshaw (1996) demuestran la versatilidad de los GLM en el sector seguros, a través de diversas aplicaciones prácticas como: modelización de la supervivencia, modelos de múltiples estados en seguros de salud, ajuste de distribuciones de pérdidas vinculadas a la severidad de los siniestros, clasificación de riesgo, entre otras. Por su parte, De Jong et al. (2008) preocupados por la lenta aceptación y comprensión de los GLM a nivel actuarial, aportan una literatura rica en ejemplos prácticos, donde los análisis han sido realizados en base a datos reales. De forma similar, Fahrmeir y Tutz (2013) ponen de manifiesto la utilidad de los modelos lineales generalizados desde una perspectiva multivariante.

Otra vertiente de la modelización estadística es la que se orienta hacia el análisis de riesgos individuales. En éste sentido, podemos mencionar el trabajo de Almer (1963), quien emplea la teoría de riesgos individuales para el análisis de la siniestralidad en seguros. También, Cossette et al. (2002) proponen dos estructuras de dependencia distintas para la modelización de riesgos correlacionados, en el contexto del modelo de riesgo individual. Desde el enfoque de *pricing*, podemos hacer referencia a Bailey y Simon (1960) quienes proponen diversos métodos basados en el test chi-cuadrado para la estimación de parámetros en un modelo de clasificación de riesgos. Con este mismo objetivo, Hsiao et al. (1990) sugieren la utilización de modelos estadísticos flexibles como: el modelo Tobit de una sola ecuación, el modelo en dos partes y el modelo de ecuaciones simultáneas para la estimación de primas puras en seguros. Por otra parte, autores como Brown y Gottlieb (2007) y Werner y Modlin (2010) son un referente en *ratemaking*, es decir, en el cálculo de precios.

Desde un punto de vista estadístico, tradicionalmente el precio del seguro se ha basado en la modelización de la frecuencia de los accidentes ocurridos en el período contractual y en la severidad de las cantidades reclamadas. Tal como lo exponen Frees et al. (2014), algunos de los motivos por los cuales las compañías aseguradoras dedican parte de sus recursos al análisis de la frecuencia y la severidad de las reclamaciones son:

- Las condiciones contractuales, respecto a la franquicia y límites de coberturas impuestos.
- Las variables que explican el apetito de riesgo de los asegurados pueden diferir entre los modelos de frecuencia y los modelos de severidad.
- El almacenamiento de bases de datos diferenciadas, por una parte con información acerca del asegurado, su(s) póliza(s) y el(los) objeto(s) de riesgo y por otra con información detallada acerca de las reclamaciones ocurridas.
- Los requerimientos exigidos por los organismos reguladores a las entidades aseguradoras, en cuanto al reporte del número de reclamaciones y sus respectivos costes.

Comúnmente, la modelización de las reclamaciones se realiza bajo el contexto clásico de los modelos lineales generalizados (ver De Jong et al., 2008). Además, estos modelos permiten incorporar la denominada tarificación a posteriori, en esta línea existen diversas contribuciones que analizan de forma alternativa la distribución del número de siniestros en el contexto de los GLMs. Por ejemplo, Pinquet et al. (2000), Bolancé et al. (2003b) y Bolancé et al. (2008b) utilizan un modelo de Poisson mixto para el diseño de sistemas de bonus malus en grandes compañías aseguradoras. De forma similar, Denuit et al. (2007) dedican su obra al análisis del número de siniestros en el contexto de los seguros de auto basándose en diferentes distribuciones que toman como base la distribución de Poisson como: la distribución binomial negativa, la distribución Poisson-Inversa Gaussiana y la distribución Poisson-LogNormal.

Una característica intrínseca de la frecuencia y la severidad de las reclamaciones, es el excesivo número de ceros presentes, debido a la ausencia de accidentes reportados en un porcentaje de las pólizas, lo cual justifica el uso de modelos que permitan combinar distribuciones discretas y continuas como las analizadas en Frees et al. (2013) y Frees et al. (2016). Otras contribuciones como las de Boucher et al. (2007) y Boucher et al. (2009), emplean modelos de Poisson cero inflados (*zero-inflated Poisson models*) enfocados al análisis exclusivo del número de siniestros de auto reportados al asegurador. Publicaciones más recientes, abordan la modelización de la siniestralidad teniendo en

cuenta las relaciones existentes entre la frecuencia y el coste de las reclamaciones (ver Frees et al., 2014, 2016).

Desde la perspectiva del cliente, aspectos como su fidelización y retención son factores determinantes para la mayoría de las compañías de seguros, dado el impacto que tienen sobre el riesgo asumido por la empresa (Guillen et al., 2008) y el éxito del negocio (Reichheld y Teal, 2001). Elevados niveles de retención favorecen, entre otras cosas, a la creación de ventajas competitivas para la empresa, mientras que, niveles bajos de retención son un indicador de la firme convicción del cliente de buscar una opción en el mercado, donde sus necesidades sean mejor valoradas (ver Reichheld y Teal, 2001, para una explicación detallada sobre el efecto de la lealtad de los clientes en todo tipo de negocios). En general, éste es un tema álgido, al que las aseguradoras dedican una gran cantidad de tiempo y recursos, pero en el que solo muy pocas han conseguido cambios profundos (ver Reichheld y Teal, 2001, donde se menciona el caso de éxito de Northwestern Mutual, "*the policyholder's company*"). Diversos estudios han sido desarrollados en torno a este tema. Por ejemplo, un aumento desacertado en la tarifa de los asegurados podría impactar directamente sobre la rentabilidad global de la cartera, si provoca que un gran número de clientes decidan anular una o más pólizas y buscar alternativas con la competencia, de modo que la modelización del efecto de la variación de la prima sobre la retención sería decisivo (ver Guelman y Guillen, 2014). Por su parte, Crosby y Stephens (1987) estudian el efecto de las relaciones cliente - mediador sobre la satisfacción, retención y los precios en el contexto de los seguros de vida. Algunas contribuciones interesantes, cuyas ideas son extrapolables al sector asegurador, son las que tratan el tema de la fidelización en clientes en el contexto de la banca minorista. Un ejemplo de ello es el trabajo de Hallowell (1996) quien tras estudiar la doble relación existente entre satisfacción versus lealtad y fidelización frente a rentabilidad, concluye que la satisfacción del cliente guarda relación con la rentabilidad de la compañía y resalta la importancia de identificar clientes objetivo, cuyas necesidades sean más susceptibles de ser cubiertas por la propia compañía que por su competidores de forma rentable, o bien la propuesta de Beerli et al. (2004) quienes utilizan un modelo de ecuaciones estructurales para demostrar cómo la satisfacción y los costes de cambiarse de compañía (*switching costs*) influyen directamente en la fidelización de los clientes. De forma similar, el trabajo de Guelman et al. (2012) presenta un procedimiento basado en *random forests* para la identificación de aquellos clientes que probablemente responderán positivamente a una actividad de retención, mientras que la contribución de Fang et al. (2016) propone un enfoque para predecir la rentabilidad de los clientes en una compañía de seguros, considerando también los fondos de reservas vinculados a obligaciones pendientes de cumplir.

La bibliografía existente sobre el estudio de los comportamientos de fidelización y deserción de clientes en diferentes áreas es amplia. Por ejemplo, Bolton et al. (2000) analizan los datos sobre clientes pertenecientes a una compañía de servicios financieros que ofrece un programa de recompensa por lealtad, con el objetivo de determinar las condiciones bajo las cuales dichos programas tendrán un efecto positivo sobre la respuesta de los clientes. En el estudio de Verhoef (2003) se evidencia cómo el compromiso afectivo de la compañía con sus clientes junto a incentivos económicos derivados de programas de fidelización afectan positivamente la retención de los clientes. Por otra parte, algunos ejemplos centrados en los procesos de modelización y/o predicción son los expuestos en las contribuciones de Smith et al. (2000), Mozer et al. (2000), Yeo et al. (2001), Au et al. (2003), Neslin et al. (2006), Kumar y Garg (2013), Günther et al. (2014),Vafeiadis et al. (2015), Ekinci y Duman (2015) y Jahromi et al. (2016).

Las contribuciones publicadas sobre la rotación de clientes generalmente se concentran en una sola línea de negocio. Esto significa que los análisis son realizados, por ejemplo sólo sobre la cartera de automóvil o sólo sobre la cartera de hogar, pero no respecto a ambas. Por lo tanto, emplean modelos estadísticos separados para cada tipo de seguro (ver Bolancé et al., 2016b, para un análisis comparativo entre diferentes medidas de rendimiento para evaluar modelos de deserción de clientes en seguros de automóviles). Existen contribuciones que abordan el tema de la dependencia dentro de una misma línea de negocio, las más frecuentes hogar o auto (véase, por ejemplo Brockett et al., 2008; Bermúdez et al., 2013; Avanzi et al., 2016; Jeong et al., 2018). Sin embargo, no hemos encontrado trabajos que analicen información sobre los asegurados con contratos en más de una línea de negocio. Por tanto, entre otros, en el sector de los seguros español, hasta el momento no es posible constatar si los análisis de tarificación, reservas o a nivel de clientes se están llevando a cabo teniendo en cuenta la dependencia entre líneas de negocio tanto a nivel agregado como individual. Por tanto, el análisis de la dependencia entre líneas de negocio supone un tema inexplorado.

Otro análisis de interés sería la modelización conjunta de los costes y la retención de clientes, destacamos que la mayoría de los artículos sólo estudian los modelos de fijación de precios o modelos de retención, pero no los dos a la vez. Sólo existen unas pocas excepciones como es el caso de Guelman y Guillen (2014) o Thuring et al. (2013), donde se mencionan el concepto de elasticidad de los precios, que significa el cambio de la demanda de seguro en función de los cambios de precios. En esta tesis tampoco abordamos la modelización conjunta de los costes y la retención. Sin embargo, nos centramos en la modelización conjunta de la retención en dos líneas de negocio distintas, quedando la incorporación de los costes en el modelo multivariante como futura línea de investigación.

1.3. Datos

El diseño de una base de datos con información referida a dos líneas de negocio distintas, en nuestro caso hogar y auto no fue sencillo. Su creación se realizó a partir de un "*Data Lake*" (un repositorio en el que se almacenan todos los datos disponibles independientemente de que estén estructurados o no) que contenía toda la información acerca de los asegurados de la compañía. Aunque su uso supuso cierta agilidad y flexibilidad en cuanto a la disponibilidad de las variables utilizadas, también requirió de tiempo para la definición, homogeneización y combinación de los datos provenientes de distintas fuentes de información.

Los datos han sido proporcionados por una compañía de seguros internacional con implantación en España e incluyen información relevante desagregada a nivel individual, por pólizas y reclamaciones. Los datos son ilustrativos y están compuestos por tres archivos diferentes: los dos primeros son datos de corte transversal que corresponden a registros de pólizas de auto y hogar; el tercero es un archivo constituido por la totalidad de las reclamaciones registradas durante el período 2010-2015.

Las bases de datos de póliza proveen características acerca de la póliza contratada, el asegurado y el objeto de riesgo. Cada registro es una foto estática de dicha información, de modo que cualquier cambio sufrido por la póliza a priori queda excluido de este estudio. Además, por cada cliente existirán tantos registros en estas bases de datos como vehículos y/o viviendas aseguradas con la compañía. Cada conjunto de datos incluye una **variable clave identificativa del cliente**, de manera que los asegurados que han contratado ambos tipos de productos pueden ser vinculados entre las dos bases de datos [2]. El conjunto de datos inicial está constituido por 3925221 pólizas de auto, 835415 pólizas de hogar y 2562659 clientes. El 76 % de los clientes ha comprado sólo pólizas de auto, el 15 % sólo pólizas de hogar y aproximadamente el 9 % de los clientes tiene contratadas ambos tipos de pólizas con la compañía. Por su parte, la base de datos de reclamaciones está compuesta por 6990044 registros de siniestros vinculados a las pólizas de auto y hogar ocurridos durante los últimos cinco años. Una reclamación suele tener más de un registro si afecta a diferentes garantías.

El tratamiento inicial de los datos de pólizas ha sido realizado en tres fases. La primera fase corresponde con la segmentación de las bases de datos por tipo de cliente. Para cada línea de negocio se han creado dos nuevas bases de datos diferentes. La primera,

[2] Nótese que dicha clave solo sirve para localizar las pólizas pertenecientes a un mismo cliente. La base de datos está anonimizada completamente

corresponde con clientes que únicamente han contratado una o más pólizas del mismo tipo de producto y la segunda está compuesta por clientes que tienen contratadas pólizas de ambos productos simultáneamente. La segunda fase está centrada en la depuración de dichas bases de datos. Para los propósitos de nuestro estudio se ha realizado un primer filtro genérico en que se han tenido en cuenta solo tipo de personas físicas, agentes y corredores como tipos de mediadores, vehículos turismo de uso particular (en el caso del producto de auto) y han quedado excluidas todas las pólizas que hayan sido sometidas a procesos de saneamiento por mal rendimiento por parte de la compañía. A continuación, se ha realizado el tratamiento individualizado de cada variable, en el que se eliminan o se corrigen valores ausentes y posibles errores de transcripción. Tras el proceso de construcción y depuración de las bases de datos, seleccionamos una muestra aleatoria del 10 % en cada una de ellas.

A continuación se presentarán los detalles del tratamiento realizado a las bases de datos de pólizas y reclamaciones. La descripción de los datos estará dividida en dos partes. Primero se expondrán los descriptivos a nivel de pólizas y reclamaciones y seguidamente, se presentarán los descriptivos a nivel de cliente.

1.3.1. Pólizas de Auto

La base de datos original está compuesta por 3925221 pólizas vinculadas a 2167194 clientes. Aquí, cada póliza corresponde con un único vehículo asegurado por el tomador, de manera que no hay duplicidad de pólizas.

En la Tabla 1.1 se presentan las variables proporcionadas por la compañía, las cuales han sido clasificadas por el tipo de información que aportan respecto al asegurado, los conductores, la póliza, el vehículo y la exposición del cliente al riesgo.

Tal como mencionamos al inicio de esta sección, hemos segmentado la base de datos de auto en dos archivos diferentes, uno con clientes que tienen contratadas exclusivamente pólizas de auto con la compañía y que denotaremos como “DB Auto” y otro con asegurados que tienen contratadas pólizas de hogar y auto simultáneamente y que denominaremos “DB Hogar/Auto”.

El análisis descriptivo inicial por variables evidenció la existencia de errores, incongruencias y valores perdidos, de modo que la limpieza de los datos supuso un esfuerzo importante al inicio del estudio.

Seleccionamos los tomadores y primeros conductores con edades comprendidas entre los 18 y 90 años; en España los conductores tienden a ser longevos y en la actualidad,

aunque exista un límite de edad mínimo para poder conducir, no se establece un límite máximo. Fijamos en tres el número máximo de pólizas en vigor por cliente, dado que aproximadamente el 98 % de los clientes tenían a lo sumo tres pólizas de auto contratadas. Entendemos que los clientes con más de tres pólizas responden a situaciones especiales, autónomos o pequeñas empresas que podrían distorsionar las conclusiones. Excluimos todas aquellas pólizas cuya forma de pago fuese única, pues están vinculadas a la contratación de pólizas temporales. También, tuvimos en cuenta vehículos cuya potencia estuviese comprendida entre 45 CV y 300 CV. Por último, se eliminaron los valores extremos o incongruentes observados en la variable antigüedad del vehículo. En cuanto al tratamiento de valores ausentes, hemos decidido utilizar el análisis de casos completos (para más detalles, ver Gelman y Hill, 2006), de forma que los casos con información faltante son excluidos del estudio. En este sentido, no se consideraron los registros con valores perdidos de las variables sexo, garantías, relación peso potencia y prima anterior, pues solo se tuvieron en cuenta pólizas que como mínimo hubieran renovado una vez. Este paso redujo el tamaño de nuestra muestra a 793481 pólizas (donde 696487 registros corresponden con la base de datos DB Auto y 96994 con DB Hogar/Auto) y 612013 clientes (donde 547211 de los registros conforman la base de datos DB Auto y 64802 la base de datos DB Hogar/Auto).

TABLA 1.1: Descripción de las variables - Pólizas de Auto

Relacionado con	**Variable**	**Descripción**
Asegurado	client_sex	Sexo ("Mujer", "Varón")
	client_age	Edad
	client_mcurrentpol	Otras pólizas de auto en vigor
	pol_other	Pólizas de otros ramos en vigor (1= Sí, 0= No)
Conductor	firstdriver_age	Primer conductor - edad
	firstdriver_agelicense	Primer conductor - antigüedad del carnet
	veh_seconddriver	Segundo conductor (Sí , No)
Póliza	policy	Número de póliza (clave enmascarada)
	client_id	Identificador del cliente (clave enmascarada)
	pol_status	Estado de la póliza (A= anulada, V= vigente)
	pol_startdate	Fecha de entrada en vigencia de la póliza
	veh_guilty	Nº de siniestros (FC= con culpa del asegurado)
	veh_load_claims	Carga siniestral (Coste total de los siniestros tanto abiertos como cerrados de la póliza)
	client_tpremiumspaid	Suma de las primas pagadas por las pólizas de auto tanto en vigor como anuladas
	pol_guarantees	Garantías (TRCF, TRSF, T, O)*
	pol_newpremium	Prima actual
	pol_lastrenewal	Prima anterior
	mediator_type	Tipo de mediador (A: Agente exclusivo, C: Corredor de seguros)
	mediator_currentpol	Mediador - pólizas en vigor
	mediator_cancelpol	Mediador - pólizas anuladas
	pol_supplements	Suplementos
	pol_diffpremium	Variación de prima
	pol_malus	Descuentos - Malus
	pol_surcharge	Recargos
	pol_malus_last	Malus (anterior)
	pol_surcharge_last	Recargo (anterior)
	pol_age	Antigüedad de la póliza
	pol_waytopay	Forma de pago (A: Anual, S: Semestral, T: Trimestral)
Vehículo	veh_use	Uso del vehículo
	veh_class	Tipo de vehículo
	veh_age	Antigüedad del vehículo
	veh_power	Potencia
	veh_weightpower	Relación peso potencia
	veh_seats	Nº de plazas
	veh_fueltype	Tipo de combustible (G: Gasolina, D: Diesel, Otro: otros)
Exposición	exp	Exposición 2010-2015. Tiempo de vigencia

Fuente: Base de datos propia, 2015. * TRCF/TRSF= Todo riesgo con/sin franquicia, T= Terceros, O= Otros

1.3.2. Pólizas de Hogar

La base de datos original de hogar está constituida por 835415 registros de pólizas correspondientes a un total de 616968 clientes. Esta base de datos resulta menos extensa que la base de datos de auto, tanto en número de pólizas como de clientes.

Análogo al caso de auto, la base de datos de hogar ha sido dividida en dos archivos distintos, uno con pólizas de clientes que exclusivamente han contratado una o más pólizas de hogar con la compañía y que hemos definido como DB Hogar y otro con pólizas vinculadas a clientes que tienen contratadas una o más pólizas de hogar y auto simultáneamente y que denominaremos DB Hogar/Auto.

En la Tabla 1.2 se exponen algunas de las variables proporcionadas por la compañía de seguros y que han sido organizadas en función de su vinculación con el asegurado, la póliza, la vivienda y la exposición del objeto de riesgo.

La depuración de los datos en general resultó más sencilla que la realizada con el producto de auto. En este caso nos centramos en clientes con edades comprendidas entre los 18 y 95 años. Decidimos trabajar con clientes que a lo sumo hubiesen contratado dos pólizas de este producto, dado que el 96 % de los tomadores presentaban esta característica. Similar al caso de auto, excluimos todas aquellas pólizas cuya forma de pago fuese única, pues están vinculadas a la contratación de pólizas temporales. En cuanto al tratamiento de valores perdidos, se han excluido los registros con valores ausentes en las variables sexo y prima anterior. Este paso redujo el tamaño de nuestra muestra a 196615 pólizas (donde 118768 registros corresponden con la base de datos DB Hogar y 77847 con DB Hogar/Auto) y 169159 clientes (donde 104357 de los registros conforman la base de datos DB Hogar y 64802 la base de datos DB Hogar/Auto).

TABLA 1.2: Descripción de las variables - Pólizas de Hogar

Relacionado con	**Variable**	**Descripción**
Asegurado	client_sex	Sexo (“Mujer”, “Varón”)
	client_age	Edad
	client_hcurrentpol	Otras pólizas de hogar en vigor
	pol_other	Pólizas de otros ramos en vigor (1= Sí, 0= No)
Póliza	policy	Número de póliza (clave enmascarada)
	client_id	Identificador del cliente (clave enmascarada)
	pol_status	Estado de la póliza (A= anulada, V= vigente)
	pol_startdate	Fecha de entrada en vigencia de la póliza
	home_guilty	Nº de siniestros (FC= con culpa del asegurado)
	pol_newpremium	Prima actual
	pol_lastrenewal	Prima anterior
	mediator_type	Tipo de mediador (A: Agente exclusivo, C: Corredor de seguros)
	mediator_currentpol	Mediador - pólizas en vigor
	mediator_cancelpol	Mediador - pólizas anuladas
	pol_supplements	Suplementos
	pol_diffpremium	Variación de prima
	pol_malus	Malus
	pol_surcharge	Recargos
	pol_malus_last	Malus (anterior)
	pol_surcharge_last	Recargo (anterior)
	pol_age	Antiguedad de la póliza
	pol_waytopay	Forma de pago (A: Anual, S: Semestral, T: Trimestral)
Vivienda	pol_capi_continent	Continente
	pol_capi_content	Contenido
	home_type	Tipo de vivienda (A, B, C, D, E, F)
Exposición	exp	Exposición 2010-2015. Tiempo de vigencia

Fuente: Base de datos propia, 2015.

1.3.3. Reclamaciones

Tal como mencionamos al inicio de esta sección, disponemos de una base de datos con 6990044 registros de siniestros vinculados con pólizas de auto y hogar. En esta base de datos los costes de los siniestros se encuentran desagregados por garantías y coberturas, de modo que en la mayoría de los casos existirá más de una fila por siniestro. La Tabla 1.3 muestra algunas de las variables que componen la información disponible para cada reclamación.

Similar al caso de auto y hogar, se realizó una depuración de la base de datos de reclamaciones. El primer filtro utilizado ha sido la elección de siniestros ocurridos en el período 2010-2015 donde el asegurado figurase como culpable, lo cual redujo considerablemente el número de siniestros a 3104675. Además, solo se han considerado determinados subproductos dentro de todos los disponibles y para cada línea de negocio lo cual ha reducido la muestra a 2476754 registros.

Con el objetivo de acotar y simplificar el análisis, se decidió seleccionar aquellas garantías que fuesen comunes al momento de la contratación de las pólizas. Teniendo en cuenta que en ambas líneas de negocio existen garantías obligatorias, se escogieron como referente para el análisis de la siniestralidad de los clientes. Este paso, redujo la muestra a 757147 registros.

En el caso de auto es obligatorio asegurar la responsabilidad civil (RC) derivada del uso de un vehículo automotor implicado en un accidente tanto en España como en el resto de los países que conforman la Unión Europea, de manera que todas las pólizas de auto tendrán contratada esta garantía. De forma adicional, también se han tenido en cuenta las pólizas que hubiesen contratado la garantía de responsabilidad civil suplementaria dada su relación con la garantía RC obligatoria.

Respecto a las pólizas de hogar, los clientes siempre deben contratar la garantía de continente o de contenido. Con frecuencia, la garantía de continente es contratada cuando el tomador es el dueño del inmueble, mientras que la de contenido es la opción preferida por los tomadores que pagan un alquiler. Sin embargo, es frecuente encontrar tomadores que optan por contratar ambas garantías y asignar un mayor capital a aquella cuyas coberturas se adapten mejor a su perfil de propietario o inquilino. Adicionalmente, en el caso de las pólizas de hogar se ha decidido incluir también dentro del análisis la garantía de daños agua, dado que ha sido contratada por el 99 % de los clientes durante los últimos 8 años.

Por último, en el caso de auto las coberturas vinculadas a la garantía de responsabilidad civil son las siguientes: daños materiales, daños corporales, invalidez permanente, invalidez temporal, deceso y atención médico farmacéutica. Mientras que en el caso de las pólizas de hogar, las coberturas observadas están relacionadas con: incendio, robo, daños eléctricos, extensión de garantías y daños agua.

TABLA 1.3: Descripción de las variables - Reclamaciones

Relacionado con	**Variable**	**Descripción**
Siniestro	claim	Número de siniestro (clave enmascarada)
	policy	Número de póliza (clave enmascarada)
	guarantee	Garantías contratadas en la póliza
	coverage	Coberturas contratadas en la póliza
	ocurrence_date	Fecha de ocurrencia del siniestro
	branch	Producto (“Hogar”, “Auto”)
	cost	Coste del siniestro
	status	Situación del siniestro (P=pendiente, R=rehabilitado, T=terminado)
	fault	Culpa (A=asegurado, C=contrario)

Fuente: Base de datos propia, 2015.

1.3.4. Clientes

En esta parte se aborda el análisis descriptivo de los datos desde la perspectiva del cliente, dado que hasta ahora el tratamiento y descripción de las bases de datos había sido realizado a nivel de póliza.

Para la creación de la base de datos única de clientes primero combinamos la información de las bases de datos DB Hogar/Auto descritas en las Sub-secciones 1.3.1 y 1.3.2 con la información agregada a nivel póliza de la base de datos de reclamaciones definida en la Sección 1.3.3. La combinación de las bases de datos resultantes a partir de la variable identificativa del cliente, nos permitió finalmente definir una base de datos con 64802 registros con información de ambas líneas de negocio agregada a nivel cliente. Ésta será nuestro referente para los análisis que realizaremos en los capítulos 4 y 5.

En la Tabla 1.4 se describen las variables agregadas vinculadas con la siniestralidad de los asegurados.

TABLA 1.4: Descripción de las variables - Agregadas

LoB	Variable	Descripción
Hogar	npol_home_ag	# de pólizas de hogar contratadas
	nclaims_home_ag	Frecuencia - # de siniestros
	cost_home_ag	Severidad - coste de los siniestros
Auto	npol_motor_ag	# de pólizas de auto contratadas
	nclaims_motor_ag	Frecuencia - # de siniestros
	cost_motor_ag	Severidad - coste de los siniestros

Fuente: Base de datos propia, 2015.

Para realizar los análisis que se muestran en los Capítulos 2, 3, 4 y 5 de esta Tesis seleccionamos el 10 % de los clientes de auto, el 10 % de los clientes de hogar y el 10 % de los clientes que tienen pólizas en ambas líneas. A todos se les asignan todas sus pólizas y en algunos casos las reclamaciones durante el período de análisis 2010-2015.

1.3.5. Modelización

Desde la perspectiva analítica, en las aplicaciones de los capítulos 3, 4 y 5 decidimos dividir las muestras correspondientes, en dos partes: 70 % para entrenamiento (*training set*) y 30 % para prueba (*test set*), de modo que podamos evaluar el ajuste de los modelos "a posteriori" o con datos que no han sido utilizados en la estimación o entrenamiento del mismo. Respecto a dicha división, no existe una forma única que permita decidir cómo ha de realizarse la partición de los datos de cara a la creación de las bases de datos de entrenamiento y prueba. Sin embargo, lo que sí es cierto es que se han de tener en cuenta ciertas consideraciones relacionadas con la elección de ambas submuestras y la evaluación del modelo, para garantizar que los resultados sean generalizables y, a su vez, que el modelo sea extrapolable. En este sentido, existen factores importantes como el tamaño de la muestra y la dispersión de las submuestras de prueba y de entrenamiento, que no deberían ser muy distintas (ver Myatt, 2007; Dobbin y Simon, 2011). De todos modos, parece haber un consenso no explícito que ha hecho que la partición entrenamiento-prueba equivalente a 70 %-30 % sea la más utilizada en la práctica.

1.4. Objetivos

En el sector seguros, el comportamiento en sí mismo de los clientes es heterogéneo, de manera que las contribuciones de cada asegurado al riesgo agregado de la compañía son diferentes. Dichas contribuciones dependerían, entre otras cosas, de las características del objeto u objetos de riesgo asegurado(s), las características personales de cada cliente y también de su comportamiento observado a lo largo del tiempo.

Una aseguradora podría estar interesada en determinar el índice de riesgo individual con el fin de ser capaz de clasificar a los clientes asegurados en función del riesgo que generan a la empresa. Esta medida podría utilizarse en los modelos de fijación de precios y de retención. Además, este concepto podría conducir a la integración de los modelos de tarificación y de retención dentro de un solo enfoque.

El objetivo general de esta tesis es modelizar de forma conjunta o desde una perspectiva multivariante, por un lado, la cuantía de las reclamaciones realizadas en las distintas pólizas y, por otro, las probabilidades de cancelación de los contratos. Ello nos permitirá diseñar métodos que sean capaces de generar una puntuación de riesgo para cada asegurado teniendo en cuenta que este puede haber suscrito una o más pólizas con la compañía, ya no sólo en un mismo producto sino en productos distintos. Nuestra hipótesis es la existencia de dependencia entre las pólizas de un mismo cliente en términos de su siniestralidad y de su propensión a la renovación. Con este objetivo en la primera parte de la tesis se presentan los siguientes análisis:

1. En el Capítulo 2 se propone una medida para cuantificar el riesgo individual del asegurado que, a priori, tiene en cuenta los costes de las reclamaciones y las probabilidades de cancelación de las pólizas. Esta medida se ha utilizado para la cuantificación del riesgo en el seguro de auto, pero puede generalizarse para cuantificar el riesgo conjunto en más de una línea de negocio. En la última parte de este capítulo presentamos una aplicación simplificada, en la que se supone que el individuo no cancelará la(s) póliza(s) y se cuantifica el riesgo a partir de la modelización univariante o multivariante de los costes en la(s) póliza(s) de auto contratada(s) por un mismo cliente. El contenido de este capítulo está publicado en Padilla-Barreto et al. (2016).

2. En el tercer Capítulo de la tesis se modeliza la deserción del asegurado para las líneas de negocio de auto y hogar de forma independiente utilizando, para ello, diferentes métodos clásicos y de *machine learning*. También, se definen criterios de

selección del mejor modelo predictivo en cada caso. Este capítulo se ha elaborado a partir de los artículos: Bolancé et al. (2016a,b).

3. En el Capítulo 4 utilizamos modelos bivariantes, basados en la distribución normal y en las cópulas Gaussiana y t-Student, para estimar la probabilidad conjunta y las probabilidades condicionadas de renovar las pólizas de auto y hogar. Parte del contenido de este capítulo actualmente está publicado en Padilla-Barreto (2018) y algunas partes se están preparando para una nueva publicación.

4. En el Capítulo 5 analizamos el papel que en la actualidad deberían tener el *business analytics* y el *big data* en las empresas aseguradoras. En este capítulo se presenta un primer ejemplo de modelización conjunta de la probabilidad de cancelación de las pólizas de hogar y auto. El contenido de este capítulo está publicado en Padilla-Barreto et al. (2017).

La segunda parte de la tesis tiene un carácter más metodológico relacionado con la construcción de cópulas multivariantes y la estimación no paramétrica del Valor en Riesgo . Esta parte, está formada por los Capítulos 6 y 7. En el Capítulo 6 se analiza el efecto sobre el valor del Valor en Riesgo y el Valor en Riesgo Condicional de la selección del *Vine* en la construcción del *pair-copula*. Por el contrario, en el Capítulo 7, con un enfoque univariante, se comparan distintos métodos no paramétricos utilizados en la estimación del Valor en Riesgo. Ambos capítulos proponen formas alternativas para el análisis y la cuantificación del riesgo en seguros. El Capítulo 6 está publicado en Bolancé et al. (2018a) y el Capítulo 7 en Alemany et al. (2016). Finalmente, en el Capítulo 8 se resumen las principales conclusiones de la Tesis.

Parte I

Riesgo individual del asegurado

Capítulo 2

Riesgo global del asegurado para una línea de negocio

2.1. Introducción

La misión principal de las compañías de seguros es gestionar el riesgo de pérdidas que les transfieren los asegurados, especialmente en lo que concierne a la severidad de los accidentes en los seguros de no-vida. El riesgo de suscripción se define como la posibilidad de que las primas recibidas no sean suficientes para cubrir las pérdidas que se producen por la declaración de siniestros. En general, y debido a la influencia de la normativa regulatoria, se establece una perspectiva de un año, lo que significa que las decisiones sobre la forma cómo se analizan las primas y las pérdidas consideran un período de doce meses.

Para simplificar el concepto de pérdidas y ganancias en un contrato de seguro, consideremos un cliente que compra un seguro de automóvil por un año. La compañía de seguros tiene que cubrir los gastos de administración, los márgenes de solvencia para atender a las exigencias normativas, la publicidad, los sistemas de comercialización y de gestión, incluyendo la tecnología informática. Una aseguradora que vende seguros para vehículos de motor tendrá miles de contratos de un año de duración y deberá pagar por las compensaciones a aquellos clientes que tienen un accidente cubierto por su correspondiente póliza.

El concepto de *pricing* se refiere a la fijación de precios y por lo tanto es tan sólo una parte de la producción de un contrato de seguro. En este punto, se entiende que las estrategias de cálculo de precios son esenciales para garantizar la estabilidad y la

solidez financiera de una compañía de seguros. El precio depende de las características del contrato y del cliente. Por lo tanto, en todo lo que sigue, vamos a suponer que el contrato de seguro ya está diseñado y que no tenemos que contabilizar información adicional, como los gastos generales de la empresa. Nos limitaremos a concentrarnos en la parte del precio que tiene que ver con las circunstancias alrededor del objeto asegurado, el tomador y el asegurado que cubre la póliza. Vamos a considerar sólo el seguro de automóviles, como un ejemplo concreto de las pólizas más habituales, dentro del ramo de los seguros generales, a nivel de las familias españolas.

En una evaluación global del riesgo de un contrato de seguro, y aparte del precio de la prima, se debe tener en cuenta la vinculación del cliente con su entidad. Dicha actitud es, de lejos, una de las prioridades más importantes en la mayoría de las compañías de seguros. Cuando un cliente decide cancelar un contrato de seguro para buscar otra compañía, la entidad original pierde la oportunidad de generar beneficios en los ejercicios futuros. Esta parte del riesgo se conoce como riesgo de caída de cartera y se considera explícitamente en la regulación de las exigencias de solvencia de las compañías aseguradoras en Europa.

Hasta ahora sólo unos pocos autores han considerado los riesgos derivados de la suscripción y la caída de cartera conjuntamente, a pesar del hecho de que están muy conectados. Como se demuestra en el trabajo de Guelman y Guillen (2014) entre otros, cuanto mayor es el precio, mayor es la probabilidad de que el cliente abandone la entidad, lo contrario ocurre si el precio disminuye. Además, se ha analizado que el efecto es heterogéneo en función de algunas variables observadas. Por ejemplo, en el artículo de Guelman y Guillen (2014) se estudia un caso del mercado canadiense y se encuentra que los asegurados jóvenes tienen una mayor sensibilidad al precio, por lo que reaccionan antes que los asegurados mayores con la cancelación de su póliza si se produce un aumento de la prima. Este análisis supone que a jóvenes y mayores se les aplica el mismo incremento de prima y el resto de circunstancias del contrato son iguales entre ambos tipos de clientes (ver, por ejemplo, otro trabajo con conclusiones similares para el mercado escandinavo Bolancé et al., 2018b)

Una prima baja es un incentivo para que un cliente permanezca vinculado a su compañía. Sin embargo, si una entidad decidiera aplicar una determinada política de descuentos de forma masiva destinada a un excesivo número de clientes, entonces el flujo de ingresos podría llegar a ser insuficiente para cubrir las pérdidas y la entidad aseguradora podría sufrir tensiones de solvencia, e incluso antes de llegar a este extremo, ver incrementados significativamente sus requerimientos de capital.

Así pues, el precio relativo a la naturaleza del riesgo y la renovación de la póliza son dos fenómenos que no pueden desvincularse. En otras palabras, en el análisis del riesgo para la tarificación es fundamental analizar conjuntamente el comportamiento asociado al riesgo de suscripción, viendo las reclamaciones esperadas e incluyendo su severidad y, a la vez, el comportamiento derivado del riesgo de caída o, lo que es lo mismo, la probabilidad de "renovar la póliza de seguro". Algunos ejemplos de este tipo de análisis se muestran en Thuring et al. (2012) y Thuring et al. (2013).

Partiendo de la idea de evaluar la fijación de precios y retención del cliente simultáneamente, en este artículo se pretende proponer una medida del riesgo para cuantificar la aportación global de cada asegurado a las pérdidas que pueden generar en una cartera, sabiendo que el cliente puede haber contratado una o más pólizas con la empresa, posiblemente de distintos ramos. La puntuación depende de las características personales y también del comportamiento observado a lo largo del tiempo para cada cliente y sus correspondientes pólizas. La idea de la modelización conjunta está inspirada en el hecho de que los clientes podrían querer abarcar más de un riesgo a la vez, ya sea por sí mismos o para otros miembros de su familia.

Para el análisis de los modelos de fijación de precios y de retención de clientes, en forma separada, suponemos que las dos variables dependientes (pérdidas agregadas y renovación de la póliza) son independientes entre sí, pero están influenciadas por variables exógenas observadas que pueden aparecer en ambos modelos.

En los antecedentes analizados no hemos hallado estudios de casos en los que se integren los modelos de fijación de precios y modelos de retención de pólizas para dos líneas de negocio diferentes (por ejemplo, automóviles y seguro del hogar). Algunos artículos sí se centran en el valor de un cliente en seguros, pero no estudian una puntuación de riesgo. No parece que se haya estudiado qué medida de riesgo es la más adecuada y la forma de modelarla. El caso más simple es utilizar la medida basada en el cuantil o Valor en Riesgo (VaR), pero esta no es la única medida posible.

Al final del proceso de modelización propuesto en este Capítulo, utilizamos como medida para cuantificar el riesgo individual de cada uno de los asegurados, el VaR, pero es posible emplear otras medidas como el valor en riesgo de la cola (Artzner et al. (1999)), u otro tipo de medidas también basadas en los cuantiles de las pérdidas (véase Dowd y Blake, 2006) o incluso aquellas que aprovechan, por ejemplo, la existencia de una nueva familia de medidas de riesgo más generales como las GlueVar (Belles-Sampera et al., 2014).

2.2. Datos

La base de datos utilizada es ilustrativa y se deriva de la muestra definida en 1.3.1. En este sentido es importante dejar claro que, la base de datos corresponde con una muestra del 10 % de clientes que, de forma exclusiva, **tenían pólizas de auto en vigor** en el momento de la extracción de la base de datos. Además, también incorporamos los costes de los siniestros ocurridos durante el período 2010-2014. En este análisis se ha eliminado el año 2015 dado que existían expedientes de siniestros no cerrados de los cuales se desconocía el coste. Finalmente, obtenemos una base de datos con 18656 pólizas y 17212 clientes, cuyos descriptivos se exponen en las Tablas A.1 y A.2 presentadas en el Apéndice A.

Los archivos de pólizas proporcionan características acerca de la póliza, el cliente y el vehículo asegurado. El conjunto de datos de reclamaciones facilita un registro de cada siniestro si este se produjo durante el período de observación. Cada conjunto de datos contiene una variable de identificación de póliza, lo que permite enlazar las dos bases de datos de forma simultánea, por lo que el resultado será un conjunto más grande de datos relacionados entre sí. Algunas variables utilizadas en los modelos posteriores se presentan en la Tabla 2.1.

En las Tablas A.3 y A.4 del Apéndice A se exponen los descriptivos básicos para esta muestra, de las variables agregadas a nivel cliente que mencionamos en la Sub-sección 1.3.4 y de las variables categóricas vinculadas al asegurado en función de la frecuencia y la siniestralidad. En el caso de auto aproximadamente el 20 % de los clientes de la muestra han sufrido al menos un siniestro durante los últimos cinco años.

El proceso de modelización se inicia vinculando la base de datos de las pólizas y su correspondiente siniestralidad. En cada renovación se analiza si el tomador renovó o no cada uno de los contratos de auto que tenía vigentes, o en su caso, si suscribió algún contrato nuevo.

2.3. Planteamiento de los modelos

2.3.1. Probabilidades de renovación

Consideramos una anualidad concreta. Sea Y_{ij} una variable aleatoria de respuesta binaria, cuya observación y_{ij} corresponde a la decisión del tomador i sobre la renovación

TABLA 2.1: Algunos factores de riesgo utilizados en el estudio.

Relacionado con	**Variable (Descripción)**
Asegurado	Sexo (“Mujer”, “Varón”) Pólizas de otros ramos en vigor (0 = Si no tiene otras pólizas, 1= Si tiene otras pólizas)
Conductor	Primer conductor - edad Primer conductor - antiguedad del carnet
Pólizas	Estado de la póliza (A= anulada, V= vigente) Garantías (TRCF, TRSF, T, O)* Prima actual Prima anterior Descuentos - Malus
Vehículo	Antigüedad del vehículo Potencia Relación peso potencia
Siniestros	Fecha de ocurrencia del siniestro Culpa (A=asegurado, C=contrario) Frecuencia - # de siniestros Severidad - coste de los siniestros

Fuente: Muestra de asegurados en el ramo de autos 2015. *
TRCF/TRSF= Todo riesgo con/sin franquicia, T= Terceros,
O= Otros

(valor igual a 1) o no (valor igual a 0) de la póliza j ese año. La variable toma el valor $y_{ij} = 1$ con probabilidad p_{ij} y el valor $y_{ij} = 0$ con probabilidad $(1 - p_{ij})$, donde $p_{ij} \in (0,1)$ y se interpreta como la probabilidad de renovar el contrato j-ésimo para el cliente i-ésimo.

Nuestro objetivo es modelizar la probabilidad de renovación en función de las k variables que conforman el vector columna $\mathbf{X}_{ij}$, donde el primer elemento es igual a 1 y el resto de los elementos contiene información sobre las características observables relacionadas con los asegurados y sus pólizas.

2.3.2. Frecuencia y severidad

En ciencias actuariales la modelización de las pérdidas o el coste del conjunto de las reclamaciones que se asocian a una anualidad de una póliza se suele dividir en dos

componentes: la frecuencia, que corresponde a la ocurrencia de un determinado número de accidentes declarados, y la cantidad total pagada al asegurado en compensación por cada uno de ellos, que se denomina la gravedad o la severidad del siniestro. Con frecuencia, el análisis del número de siniestros se realiza desde una perspectiva de modelo de frecuencia - severidad. En este sentido, las distribuciones de ambos factores se pueden estimar de manera aislada, tomando el número de siniestros por un lado y su severidad por otro, o bien se pueden estudiar simultáneamente como veremos a continuación (para más detalles ver Frees, 2009; Frees et al., 2013, 2016).

2.3.3. Modelización del coste agregado de los siniestros

El riesgo de accidentes vinculados a una póliza de un determinado producto se descompone por tipologías de coberturas, por lo que es habitual sumarlas para ofrecer paquetes a los clientes con diferentes tipos de garantías. Habitualmente, cada garantía tiene una o más coberturas asociadas, por lo que, cuando se produce un accidente, los gastos que se generan se clasifican en dichas sub-categorías de riesgo.

Al analizar la información de una cartera de pólizas de seguros, los casos se dividen en dos grupos; por un lado, se encuentran aquellas pólizas que no han sufrido ningún siniestro y por lo tanto no se ha efectuado ningún pago de reclamaciones y, por otro lado, se tienen los gastos positivos asociados con el pago de la indemnización correspondiente en las pólizas que sí han generado siniestralidad para la entidad aseguradora. Nótese que en algunas ocasiones y debido a la existencia de convenios entre entidades, la declaración de un siniestro puede acarrear un cobro recibido de otra entidad, por lo que la pérdida asociada resulta ser negativa. Para simplificar la exposición no se consideran estos casos y por lo tanto todas las cuantías son positivas.

En este trabajo únicamente analizamos siniestros que tienen que ver con las coberturas de responsabilidad civil para terceros (lesiones corporales, daños a la propiedad, atención médica y farmacéutica) y consideramos sólo los casos en los que los asegurados tienen la culpa del siniestro declarado. En este sentido, decidimos agregar cada reclamación de información por medio de la suma de las coberturas afectadas con el fin de definir la severidad agregada.

En lo que sigue vamos a emplear la siguiente notación relativa al asegurado i-ésimo y su póliza j-ésima. Sea N_{ij} el número de siniestros declarados durante una anualidad y L_{ij} la suma de las pérdidas causadas por dichos siniestros asociados a la mencionada póliza j.

2.3.4. Modelos para variables con truncamiento inferior en el cero

Como se acaba de exponer, una de las características de las variables relacionadas con el número de siniestros y su severidad es la presencia de un elevado número de ceros debido a la ausencia de accidentes declarados en algunas (generalmente la mayor parte) de las pólizas. Para tratar este comportamiento aleatorio se necesitan modelos que permitan manejar una mezcla de distribución discreta y continua, como por ejemplo los modelos: Tobit (ver Lin y Grace, 2007), Tweedie (véase Jørgensen y Paes De Souza, 1994) y los modelos en dos partes (para más detalles ver Frees et al., 2016).

Comenzamos nuestro análisis mediante la modelización de la cuantía agregada de los siniestros, es decir, ajustando los modelos para los valores positivos de la variable aleatoria "pérdidas" o también el ajuste de una mezcla utilizando un modelo frecuencia-cuantía conocido como el modelo basado en la distribución Tweedie. Dicho modelo predictivo permite que se pueda efectuar una estimación de cuál será la pérdida esperada que va a generar una póliza dadas las circunstancias conocidas del contrato (tipo, coberturas, etc.), del asegurado (sexo, edad, etc.) y del objeto asegurado si es vehículo, su antigüedad y potencia, por ejemplo y si es una vivienda, su superficie y su tipología, etc.

2.4. Definición del riesgo del tomador de las pólizas

A partir de la mezcla de las probabilidades de renovación y el coste agregado de las reclamaciones de una póliza, se define la siguiente variable aleatoria:

$$Y_{ij}^{*} = \begin{cases} L_{ij} - \rho_{ij}, & \text{si } y_{ij} = 1 \\ b \cdot \rho_{ij}, & \text{si } y_{ij} = 0 \end{cases} \tag{2.4.1}$$

donde, para el i-ésimo cliente, L_{ij} es la pérdida de la póliza j-ésima y el valor y_{ij} indica que se ha renovado la póliza si vale 1, o bien si ha sido cancelada si vale 0. Además, se ha incorporado la cuantía de la prima devengada por el asegurado, que se denota por ρ_{ij}. Además, el factor denotado por b, denota una proporción de coste que puede asociarse a la pérdida generada por la cancelación de una póliza en relación a la cuantía de la prima que hubiera pagado.

Hasta el momento, tenemos una forma de cuantificar el riesgo individual por póliza para cada cliente a partir de la expresión anterior que denota la pérdida que se genera

según se renueve o se cancele la póliza. Sin embargo, como hemos mencionado en la introducción, estamos interesados en proponer una nueva medida de riesgo que permita cuantificar el riesgo asegurado por el cliente como una puntuación total, teniendo en cuenta que el mismo individuo podría tener una o varias pólizas en la misma entidad aunque sean de diferentes ramos.

La idea anterior nos lleva a definir el "riesgo global del asegurado (RgA)" a partir de la siguiente variable aleatoria:

$$RgA_i = \sum_{j=1}^{m_i} Y_{ij}^* \tag{2.4.2}$$

siendo m_i el número total de pólizas del asegurado i-ésimo.

En este sentido, el concepto RgA permite reflejar la hipótesis de que la experiencia observada en las pérdidas de las pólizas de automóviles pertenecientes al mismo tomador está correlacionada. En general, los miembros de una familia, para los que el tomador del seguro suele ser el cabeza de familia, comparten características comunes y probablemente tienen actitudes de riesgo similares, por lo que no es de extrañar que se comporten de manera similar.

Aunque sea de forma indirecta, la expresión del RgA indica la satisfacción del cliente a través de la posibilidad de renovación o cancelación, que tiene una relación directa con los flujos de ingresos futuros (ver Fornell, 1992) y puede ser afectada por muchas circunstancias exógenas (ver Fornell y Wernerfelt, 1988, para más detalles). De hecho, varios estudios combinados con la experiencia de casos reales muestran que un cliente insatisfecho puede cancelar varias pólizas de forma simultánea, por lo tanto, lo que afecta a una póliza puede afectar al resto de las pólizas que tenga contratadas el mismo asegurado, e incluso su familia.

El riesgo global del asegurado RgA representa una forma de analizar el valor del cliente y permite cuantificarlo, por lo que es un índice que puede resultar clave para mejorar las decisiones de gestión, ya que permite a la compañía entender, clasificar y ser capaz de cuidar a los clientes más valiosos, lo que aumenta inevitablemente la rentabilidad de su negocio.

Bajo este escenario, se aborda el análisis de RgA desde dos perspectivas diferentes. La primera de ellas considera la independencia entre las pólizas contratadas en los ramos de automóviles y hogar, y la segunda consiste en una modelización conjunta de su comportamiento conjunto.

Por otra parte, a priori, consideramos que la probabilidad de renovación de cada póliza es conocida a través de algún modelo de retención de clientes que ya exista en la entidad. Además, el porcentaje de beneficio b se puede predeterminar. En consecuencia, se puede establecer el comportamiento aleatorio del RgA simplificando las componentes estocásticas de las pérdidas (L_{ij}, variable aleatoria que toma valores no-negativos) y la renovación (Y_{ij}, variable aleatoria binaria que únicamente admite dos posibilidades: renovar o cancelar).

Sin embargo, el reto del modelo RgA es diseñar una optimización de precios para las primas con todos los componentes juntos, lo que plantea un problema de optimización que no es fácil de resolver. En este caso, el uso de la metodología de las cópulas que proponemos en el Capítulo 6 de la segunda parte de la tesis, puede ser una alternativa apropiada.

2.4.1. Análisis del riesgo global asegurado (RgA)

Deseamos estudiar la distribución de RgA_i para cada asegurado i, suponiendo que a priori son conocidos los valores de p_{ij} (probabilidad de renovación) y b (porcentaje de la prima que se utiliza para valorar el gasto que supone la cancelación de una póliza). Vamos a suponer también que la prima fijada para el cliente i-ésimo y la póliza j-ésima es conocida (ρ_{ij}), y que se sabe cuántas pólizas (m_i) posee cada asegurado. Por lo tanto, bajo este supuesto y como mencionábamos antes, las únicas componentes que son aleatorias son por un lado L_{ij}, es decir, la pérdida generada por la póliza j-ésima del cliente i-ésimo si éste renueva, y el hecho de que se renueve ($y_{ij} = 1$) o no ($y_{ij} = 0$).

Mediante los modelos predictivos planteados en la Sección 3.2 (ver también Bolancé et al., 2016b), se puede tener una aproximación al comportamiento que tienen tanto las pérdidas L_{ij}, como la renovación Y_{ij}, en función de los factores conocidos de cada cliente.

Proposición 1. La esperanza y varianza matemáticas de RgA_i son, respectivamente,

$$\sum_{j=1}^{m_i} \left[(E(L_{ij}) - \rho_{ij}) \mathrm{p}_{ij} + b \cdot \rho_{ij} (1 - \mathrm{p}_{ij}) \right] \tag{2.4.3}$$

y,

$$\sum_{j=1}^{m_i} \left\{ V[L_{ij}] + [E(L_{ij}) - \rho_{ij} - b\rho_{ij}]^2 \right\} \mathrm{p}_{ij} (1 - \mathrm{p}_{ij}). \tag{2.4.4}$$

Demostración. Podemos calcular la esperanza matemática de la variable RgA_i:

$$E\left[RgA_i\right] = E\left[\sum_{j=1}^{m_i} Y_{ij}^*\right] = \sum_{j=1}^{m_i} E\left[Y_{ij}^*\right] = \sum_{j=1}^{m_i} \left[(E(L_{ij}) - \rho_{ij})\mathrm{p}_{ij} + b \cdot \rho_{ij}(1-\mathrm{p}_{ij})\right]. \quad (2.4.5)$$

Además, la varianza de la variable RgA_i es:

$$V\left[RgA_i\right] = V\left[\sum_{j=1}^{m_i} Y_{ij}^*\right] = \begin{cases} \sum_{j=1}^{m_i} V\left[Y_{ij}^*\right], & \text{si } Y_{ij}^* \text{ son independientes} \\ \sum_{j=1}^{m_i} (V\left[Y_{ij}^*\right] + 2\sum_{j'>j} Cov\left[Y_{ij}^*, Y_{ij'}^*\right]), & \text{si no lo son} \end{cases} \quad (2.4.6)$$

Para obtener la varianza hay que utilizar la expresión de la esperanza de la esperanza condicionada.

$$V\left[Y_{ij}^*\right] = E\left[E((Y_{ij}^* - E(Y_{ij}^*))^2 | L_{ij})\right] \quad (2.4.7)$$

Tomando la parte correspondiente a la esperanza condicionada,

$$E((Y_{ij}^* - E(Y_{ij}^*))^2 | L_{ij}) = V(Y_{ij}^* | L_{ij}) = (L_{ij} - \rho_{ij} - b\rho_{ij})^2 \mathrm{p}_{ij}(1-\mathrm{p}_{ij}). \quad (2.4.8)$$

Por lo tanto,

$$V\left[Y_{ij}^*\right] = E\left[(L_{ij} - \rho_{ij} - b\rho_{ij})^2\right] \mathrm{p}_{ij}(1-\mathrm{p}_{ij}) = \left\{V\left[L_{ij}\right] + \left[E(L_{ij}) - \rho_{ij} - b\rho_{ij}\right]^2\right\} \mathrm{p}_{ij}(1-\mathrm{p}_{ij}). \quad (2.4.9)$$

Finalmente,

$$V\left[RgA_i\right] = \sum_{j=1}^{m_i} \left\{V\left[L_{ij}\right] + \left[E(L_{ij}) - \rho_{ij} - b\rho_{ij}\right]^2\right\} \mathrm{p}_{ij}(1-\mathrm{p}_{ij}). \quad (2.4.10)$$

□

2.4.2. Cálculo del valor en riesgo del riesgo global del asegurado (RgA)

Como mencionamos anteriormente, estamos interesados en cuantificar el Valor en Riesgo (*VaR*) para cada cliente. Simplificando el problema, de modo que se supone que la probabilidad de cancelación es cero, por lo que la póliza siempre se renueva y la prima es fija y conocida, necesitamos centrarnos en el modelo:

$$VaR(RgA_i) = VaR(\sum_{j=1}^{m_i} Y_{ij}^*) \approx VaR(\sum_{j=1}^{m_i} L_{ij}). \tag{2.4.11}$$

El VaR es una de las medidas de riesgo más conocidas basada en la distribución de la pérdida que es igual al cuantil a un nivel de confianza α, es decir, para un tomador del seguro i y un contrato j, si $F_{L_{ij}}^{-1}()$ es la inversa de la función de distribución de la cuantía L_{ij} tenemos:

$$VaR(L_{ij}) = F_{L_{ij}}^{-1}(\alpha). \tag{2.4.12}$$

Sin embargo, como no se puede garantizar la subaditividad, entonces según Artzner et al. (1999) no podemos asegurar que siempre se cumpla que:

$$VaR(\sum_{j=1}^{m_i} L_{ij}) \leq \sum_{j=1}^{m_i} VaR(L_{ij}). \tag{2.4.13}$$

La propiedad anterior es un aspecto que analizaremos posteriormente en el apartado de modelización, teniendo en cuenta que las pérdidas agregadas pueden analizarse bajo el supuesto de independencia o dependencia entre las diferentes pólizas para el mismo asegurado.

Sea $S_i = L_{i1} + \cdots + L_{im_i}$ la suma de los m_i costes que ha generado el cliente i, a causa de los accidentes en los que haya incurrido. Empezamos a calcular el valor en riesgo suponiendo que los costes L_{ij} son variables aleatorias independientes e idénticamente distribuidas (i.i.d.). En este caso tenemos que deducir la distribución de la suma de variables i.i.d. y después calcular el cuantil de dicha distribución.

Supongamos que $m_i = 2$, de modo que L_{ij} y $L_{ij'}$, $j, j' = 1...m_i$ representan a dos pérdidas aleatorias idénticamente distribuidas, con respectivas funciones de densidad iguales a f_{ij} y $f_{ij'}$, entonces, la densidad de la suma de las dos pérdidas $(L_{ij} + L_{ij'})$ es:

$$f_{S_i}(x) = \int_0^{+\infty} f_{i1}(x-v) f_{i2}(v) dv. \tag{2.4.14}$$

Cuando el número de pérdidas es superior a 2 ($m_i > 2$) entonces se aplica la expresión anterior recursivamente.

Si suponemos que las pérdidas derivadas de las diferentes pólizas que pertenecen a un mismo cliente son dependientes, la densidad de probabilidad de la suma de variables aleatorias se puede calcular a partir de la función de densidad multivariante. Seguidamente, a partir de la densidad podemos deducir la función de distribución acumulativa y el valor en riesgo. Por lo tanto, podemos suponer una distribución multivariante de las pérdidas (Klugman et al., 2012) o una cópula con marginales univariantes dadas y estimar el riesgo asociado a la suma de las pérdidas. Hay varias distribuciones que pueden ser buenas candidatas para la modelización de la cuantía de los siniestros, como por ejemplo las utilizadas en este trabajo: Inversa Gaussiana, Log-Normal y Tweedie. Es de señalar que la estimación del valor en riesgo es más complicada cuando se usan distribuciones distintas de la Log-normal (Guillen et al., 2013), aunque existen expresiones analíticas en algunos casos.

2.5. Modelización del riesgo global del asegurado

La Tabla 2.2 muestra las variables y su descripción en términos de si la póliza presenta alguna reclamación o no y de la misma manera, en la Tabla 2.3 se exponen los parámetros estimados para los tres modelos estimados en este trabajo: Inversa Gaussiana, Tweedie y Log-normal.

TABLA 2.2: Descripción de las variables dependiendo de la existencia de siniestros declarados

Variable	Total	Sin siniestros	Con siniestros
Sexo			
Mujer	31.23 %	13.37 %	17.86 %
Hombre	68.77 %	31.84 %	36.93 %
Edad conductor	53	54	53
Antigüedad del carnet	30	30	29
Potencia	110.1	108.9	111.07
Peso-Potencia	477	475.83	478.05
Antigüedad del vehículo	11	11	10

Fuente: Elaboración propia. Excepto los valores en porcentaje, el resto de descriptivos muestra la media.

TABLA 2.3: Estimación de los parámetros para tres modelos de cuantía de los siniestros.

	Inv. Gaussian			Tweedie			Log-Normal		
Variable	Estimate	t-value		Estimate	t-value		Estimate	t-value	
Constante	7.384	24.091	***	6.925	22.802	***	5.961	75.447	***
Sexo	0.06	0.54		0.012	0.112		0.028	0.99	
Edad conductor	-0.009	-1.226		-0.008	-1.034		0	-0.077	
Antigüedad permiso	0.01	1.212		0.007	0.817		0.002	0.956	
Potencia	0.002	1.608		0.003	2.532	*	0.004	11.598	***
Antigüedad Vehículo	-0.036	-3.591	***	-0.043	-4.171	***	-0.041	-15.185	***
Peso - Potencia	0	-1.211		0	-1.455		0	-2.598	**

Fuente: Elaboración propia. Máxima verosimilitud considerando el tiempo de exposición según la vigencia de cada póliza. *** Indica una significación del 1 %, ** (5 %) y * (10 %), respectivamente.

2.5.1. Valor en riesgo individual

A partir de los resultados anteriores presentamos un ejemplo del cálculo del VaR asociado a la pérdida global del asegurado y, en su caso, el VaR asociado a las pérdidas ocasionadas por cada una de las pólizas contratadas. Supondremos una distribución Log-Normal para las pérdidas ocasionadas por la siniestralidad en cada una de las pólizas (L_{ij}).

Un modo sencillo de ajustar una distribución Log-Normal a las variables aleatorias L_{ij} es transformar dichas variables en logaritmos, de modo que podemos definir la variable $S_i = ln(L_{i1}) + \cdots + ln(L_{im_i})$, que equivale a una suma de normales que a su vez es Normal. Es decir, se sabe que la distribución del logaritmo de una variable Log-Normal es Normal y además, como la distribución de la suma de variables normales también es Normal es sencillo obtener los valores de VaR de la variable S_i suponiendo correlación entre las variables $ln\left(L_{ij}\right)$, a dicho valor en riesgo lo denominamos "VaR0.99(multi)" y asume que existe dependencia entre las siniestralidades de las distintas pólizas que tiene contratadas el asegurado i. Además, calculamos los valores del VaR asociados a cada variable $ln\left(L_{ij}\right)$, los cuales denominamos VaR0.99, y representan el riesgo asociado a las pérdidas por siniestralidad del asegurado i en la póliza j.

La Tabla 2.4 muestra un ejemplo de los resultados para el VaR0.99 y el VaR0.99(multi). Como puede observarse en la Tabla 2.4, los asegurados que tienen más de una póliza tienen un valor en riesgo agregado (multi) que es inferior al que resulta de sumar el valor en riesgo individual de cada póliza. En este caso, y aunque no siempre sea así en el caso de VaR, se cumple la propiedad de subaditividad de la medida del riesgo.

Si tenemos la certeza de que el asegurado renovará su póliza, los valores en la columna VaR0.99(multi) son una aproximación del riesgo global del asegurado, definido anteriormente como $\text{VaR}(RgA_i)$.

Una aplicación futura del enfoque presentado en este estudio es la cuantificación del riesgo global de cada asegurado, teniendo en cuenta que éste puede tener una o más pólizas contratadas, ya no sólo en un mismo producto sino en productos diferentes, y que existe una probabilidad p_{ij} inferior a la unidad de que el individuo renueve cada una de sus pólizas contratadas.

2.6. Conclusiones

Una compañía aseguradora puede estar interesada en un índice de riesgo individual con el fin de ser capaz de clasificar a los clientes que conforman su cartera de asegurados. Ello permite clasificarlos en función del riesgo que generan para la entidad. Por otra parte, dicha medida puede ser utilizada para mejorar las estrategias y los modelos de precios y, en consecuencia, la política de retención. Idealmente, este concepto puede generar una forma de integrar los modelos de fijación de precios y de fidelización en un solo enfoque. Dicha integración facilitará la cuantificación del riesgo contemplando

TABLA 2.4: Ejemplo de estimación del valor en riesgo del cliente

Cliente	Póliza	Log(Costes)	Valores estimados	Exposición	VaR0.99	VaR0.99 (multi)
947	63625	7.21	6.09	0.6	5.65	5.65
699	61501	4.62	5.98	0.8	7.45	7.45
653	49646	4.77	5.91	1	9.24	9.24
4035	57714	8.00	6.73	0.8	8.05	8.05
5795	70368	5.56	5.80	0.4	3.65	3.65
6329	58317	7.81	6.25	0.8	7.67	7.67
8433	32219	5.55	6.33	1	9.66	9.66
9796	53602	6.78	5.81	1	9.14	9.14
1985	47658	7.72	6.14	1	9.47	9.47
2236	45631	5.09	6.08	1	9.42	9.42
2411	49760	7.13	5.84	1	9.17	9.17
2521	18866	8.70	6.24	1	9.58	9.58
3407	41855	4.31	6.14	1	9.47	9.47
3600	57180	6.78	5.86	0.8	7.35	7.35
4705	44366	7.35	6.32	1	9.65	9.65
4892	56925	4.10	5.86	0.8	7.35	7.35
6879	**12235**	**8.37**	**6.62**	**1**	**9.95**	**16.62**
6879	**59853**	**8.37**	**7.17**	**0.8**	**8.40**	**16.62**
8976	72142	6.78	5.75	0.4	3.63	3.63
9320	54224	8.41	6.11	1	9.44	9.44
9343	93583	5.15	5.91	1	9.24	9.24
1191	**88106**	**5.10**	**5.83**	**1**	**9.16**	**16.79**
1191	**50941**	**4.05**	**6.25**	**1**	**9.58**	**16.79**

Fuente: Elaboración propia. Sombreados, los clientes con dos pólizas diferentes.

simultáneamente suscripción y caída de cartera, lo cual es un factor clave para las compañías de seguros que centran sus esfuerzos en la gestión completa de los riesgos.

Sin lugar a dudas, es razonable pensar que precio y renovación son dos conceptos que están relacionados entre sí, debido a la presencia de factores no observables que afectan a ambos. Si se desea modelizar este comportamiento estocástico completo, se puede utilizar un modelo conjunto. Podemos considerar, por ejemplo, un modelo lineal generalizado para la severidad de los siniestros, un modelo lineal generalizado para el número de reclamaciones y un modelo lineal generalizado para el resultado binario sobre la renovación o no de la póliza, sabiendo que los parámetros de asociación entre dichas tres magnitudes pueden ser incluidos en el modelo simultáneo. Alternativamente, podemos utilizar una cópula con diferentes distribuciones marginales en función del tipo de variable, lo cual permitiría incorporar de un modo sencillo la dependencia entre las pérdidas de los diferentes contratos y las probabilidades de renovación de los mismos. Esta será la vía futura de la investigación presentada.

En resumen, la aproximación presentada permite analizar una estructura de datos de alta complejidad con el fin de explicar y predecir el comportamiento de pérdida de clientes y, en especial, su riesgo asociado, teniendo en cuenta la relación entre la siniestralidad de las pólizas contratadas.

Capítulo 3

Modelos de deserción para una línea de negocio

3.1. Introducción

En este capítulo, se evalúa el poder predictivo de cuatro modelos de deserción o cancelación [1] para dos carteras de pólizas, una de auto y otra de hogar. Nuestra comparación analiza los resultados de una regresión logística, un árbol condicional, una red neuronal y una máquina de vectores de soporte.

La satisfacción de los asegurados representa una de las prioridades más importantes para las aseguradoras, pues garantiza la estabilidad de su cuota de mercado empresarial. De hecho, las compañías de seguros necesitan mantener a sus clientes satisfechos, para así poder aumentar las posibilidades de que las pólizas contratadas sean renovadas en la fecha de vencimiento. Diferentes investigadores argumentan que retener a un cliente es menos costoso que conseguir uno nuevo (ver Fornell y Wernerfelt, 1987, 1988; Zeithaml et al., 1996), lo cual representa una estrategia rentable para la compañía. Por otra parte, no existe información en los registros de la compañía vinculada a los nuevos clientes, mientras que este conocimiento si está disponible en el caso de los clientes preexistentes. La información personal de los asegurados es valiosa y se usa intensamente junto al análisis de datos para predecir el comportamiento futuro esperado del riesgo asumido (en Thuring et al., 2013; Verhoef, 2003, se muestran algunos ejemplos sobre este tema). Por su parte, la información histórica permite calcular un precio justo para la cobertura del seguro. Además, los clientes más antiguos tienden a estar más comprometidos con la compañía (Thuring et al., 2012) y su retención le consume menos tiempo

[1] Son modelos utilizados para la predicción de la cancelación de una póliza por parte de un cliente.

a ésta en comparación con el tiempo que se requiere para atraer nuevos clientes (ver Keaveney, 1995; Brockett et al., 2008).

Cada día una parte de los asegurados deciden dejar su compañía y cambiarse a la competencia. En este sentido, la satisfacción del cliente tiene un vínculo directo con los flujos de ingreso futuros (Fornell, 1992) y puede estar afectada entre otras cosas por no invertir en atender las quejas de los clientes insatisfechos (Fornell y Wernerfelt, 1988), por incrementos en los precios de las pólizas, por la gestión deficiente de los requerimientos del cliente y los retrasos en los pagos de las reclamaciones, entre otros. Por tanto, identificar los motivos por los cuales un cliente decide cancelar su(s) póliza(s), así como determinar los factores que explican el comportamiento de cambio es un aspecto clave para la correcta gestión y estrategia de una compañía de seguros (Guillen et al., 2012). La necesidad de implementar acciones a corto plazo orientadas a mejorar la satisfacción de los clientes, revertir la intención de abandonar la compañía e incrementar la rentabilidad de la misma, es el motivo por el que se hace necesario el diseño de modelos predictivos que permitan identificar la propensión de los clientes a la cancelación de sus pólizas (ver Guillen et al., 2008; Guelman y Guillen, 2014; Guelman et al., 2014, 2015a,b).

Las compañías de seguros generalmente enfocan su actividad económica a través de lo que se conoce como **líneas de negocios** (*Lines of business* - LoB). Cada LoB tiene características específicas, tales como: el objeto asegurado, diversos factores de riesgo, coberturas, modelos de precios, estrategias específicas para los diferentes tipos de clientes, entre otros. En este sentido, dada la naturaleza del tipo de seguro, las empresas establecen recursos especializados a nivel individual y personalizado para analizar las circunstancias y resultados obtenidos por cada LoB.

Nuestro objetivo es contribuir con la literatura actual y ayudar a cerrar la brecha existente en la modelización de la retención de clientes de una compañía aseguradora. En general, las contribuciones disponibles relacionadas con la deserción de clientes en seguros no necesariamente comparan los criterios utilizados para escoger entre diferentes alternativas de modelización. Además, hasta donde sabemos, ningún trabajo anterior compara la capacidad predictiva de diferentes modelos de predicción entre dos líneas de negocios distintas. En este capítulo, proponemos medidas de ajuste y valores umbrales (*cut-offs*) que pueden guiar en la práctica para decidir cuál es el mejor modelo predictivo en su contexto particular.

En nuestro caso de estudio, nos centramos en los seguros de no vida, que son suscritos para proteger objetos o propiedades, y para cubrir responsabilidades derivadas de un accidente. Específicamente, optamos por focalizarnos en dos de las líneas personales

más importantes: auto, que abarca la mayor parte del mercado de seguros de no vida y hogar, cada una de las cuales representa una línea de negocio con características intrínsecas. Por ejemplo, el porcentaje de pólizas canceladas cada año en cada uno de estos tipos de seguros es claramente distinto.

En general, el % de cancelación de pólizas de auto es superior que el observado con las pólizas de hogar. La existencia de un mercado amplio de seguros de auto y la competencia feroz entre aseguradoras en contraposición a lo que ocurre en los seguros de hogar, son algunos de los motivos que explican las diferencias entre las cancelaciones de cada tipo de póliza.

Por lo tanto, los consumidores tienen la facilidad de encontrar oportunidades más convenientes, adecuadas a sus preferencias en el momento de buscar un seguro de auto (en el seguro del automóvil, la prima es fundamental, ver, por ejemplo, Yeo et al., 2001; Pinquet et al., 2001; Bolancé et al., 2003b, 2008b; Guillen et al., 2013). Por otro lado, en el seguro del hogar, la competencia entre las aseguradoras es menor que en el seguro del automóvil y, por lo tanto, los consumidores parecen perder su interés por cambiar de una compañía a otra.

En el caso de estudio que presentamos a continuación, estamos interesados en estimar las probabilidades de renovación de las pólizas de auto y de hogar desde una perspectiva de clasificación. Para ello, seleccionamos el modelo de regresión logística como modelo base para establecer los análisis comparativos posteriores. Éste es un modelo predictivo clásico ampliamente utilizado a nivel actuarial, cuya implementación y análisis de resultados es familiar. También, se escogen los árboles de decisión por lo versátiles que son a nivel empresarial, dada su facilidad para reflejar las interacciones entre variables y por ende la segmentación de grupos de riesgo. Ambos modelos son conocidos como "*white-box models*" dado que permiten la interpretación de los parámetros del modelo. Por último, seleccionamos dos modelos de tipo "*black-box*" - las redes neuronales y la máquina de vectores de soporte -, reconocidos por tener un poder discriminante significativamente mejor que el de los "*white-box models*" (ver Dreiseitl y Ohno-Machado, 2002, para más detalles). Estos métodos permiten clasificar cada nuevo cliente a partir de su probabilidad de pertenecer al grupo de los que renuevan su(s) póliza(s) o bien el de los que deciden cancelarla(s). De hecho, consideramos dos grupos diferenciados en el algoritmo de clasificación: los que se quedan y los que se van.

Conducimos nuestro análisis comparando la curva ROC (*receiver operating characteristic curve*) obtenida con cada método. La curva ROC es una de las herramientas más comunes para evaluar y comparar modelos de clasificación (Fawcett, 2006). Por ejemplo, la contribución de Guillen (2014) presenta una explicación detallada de la curva

ROC como herramienta gráfica para evaluar la bondad de ajuste de los modelos predictivos. Sin embargo, cuando las curvas ROC se cruzan unas con otras una o más veces, la selección del mejor método a partir de esta comparación puede generar resultados ambiguos (Gigliarano et al., 2014). Nuestra principal contribución es la búsqueda de diferentes criterios de ajuste, de modo que podamos comparar los resultados para cada uno de los métodos utilizados para la predicción.

En este capítulo consideramos tres tipos de criterios, dos de ellos basados en puntos de corte óptimos y otro basado en el área bajo la curva ROC. Queremos escoger el mejor modelo de acuerdo a los objetivos específicos de la compañía, apoyándonos en estos criterios. Este hecho tiene una implicación directa en la habilidad predictiva de los modelos, es decir, diferentes criterios proporcionan diferentes preferencias hacia los modelos.

Utilizamos dos bases de datos proporcionadas por una compañía de seguros, con información relevante sobre las pólizas contratadas. El tratamiento de los datos ha sido realizado con SAS y para la implementación de los modelos y de las medidas de rendimiento se ha utilizado el lenguaje R.

Este capítulo está organizado como sigue: en la Sección 3.2 se describen brevemente los métodos utilizados para estimar la probabilidad de pertenecer a cada grupo y se presentan las medidas discriminatorias basadas en puntos de corte y la curva ROC. En la Sección 3.3 se describen los datos y el tratamiento realizado. En la Sección 3.4 se expone el análisis comparativo y la discusión de los resultados. Y finalmente, en la Sección 3.5 resumimos las conclusiones obtenidas de los resultados que se presentan en el capítulo.

3.2. Modelos para la predicción de la deserción

Sea y_i una observación de la variable aleatoria Y_i cuyos valores pertenecen al conjunto $\{0,1\}$ y proporcionan información acerca de la decisión que toma el i-ésimo cliente en el momento de la renovación de su póliza. En caso de que el cliente decida mantener la vigencia de su póliza, el valor observado será igual a 1 y en caso de que decida cancelarla, el valor observado será igual a 0. Además, sea $\mathbf{X}_i^{'} = (X_{i0}, X_{i1}, ..., X_{ik})$ un vector de k covariables, todas ellas asociadas con la i-ésima póliza. El número total de pólizas es denotado por N.

Los modelos de predicción de la deserción se diseñan desde una perspectiva de clasificación binaria, de modo que, los casos se asignan a una de las dos clases $Y_i =$

{Renueva (se queda), No renueva (se va)} sobre la base de sus características observadas y representadas en el vector de covariables $\mathbf{X}_i$. Los clasificadores discriminan los casos a partir de las clases definidas anteriormente. En nuestro caso, usamos clasificadores blandos, es decir, los métodos utilizados asignan a cada caso un *score* en vez de clasificarlo directamente como una renovación o un abandono (Stripling et al., 2015; Liu et al., 2011). Estos métodos estiman la probabilidad de clase p_i, la cual denota la propensión a la renovación de la i-ésima póliza. Posteriormente, es posible definir las clases predichas al comparar p_i con diferentes puntos de corte de clasificación $t \in [0,1]$. Cada una de estas comparaciones produce una matriz de confusión como la expuesta en la Tabla 3.1.

TABLA 3.1: Matriz de confusión para un punto de corte dado t.

		Predicho	
		No renueva	Renueva
Real	No renueva	Verdadero negativo	Falso positivo
	Renueva	Falso negativo	Verdadero positivo

Fuente: Diseño propio.

De la Tabla 3.1 se pueden definir las siguientes medidas para evaluar el rendimiento predictivo de los modelos de clasificación que serán implementados más adelante.

- La sensibilidad es una medida de la proporción de clientes que habiendo renovado sus pólizas han sido correctamente clasificados por el modelo.
- La especificidad es una medida de la proporción de clientes que habiendo cancelado sus pólizas efectivamente han sido clasificados como desertores.
- La precisión mide la proporción total de clientes que han sido clasificados de forma correcta.

A continuación, presentamos cuatro técnicas populares usadas para problemas de clasificación binaria: regresión logística (RL), árboles condicionales (AC), redes neuronales (RD) y máquina de vector de soporte (MVS). Estos métodos permiten generar una puntuación entre [0,100], de modo que es posible clasificar cada cliente basado en dicha puntuación. De hecho, la puntuación en sí misma es sólo una transformación de la probabilidad estimada obtenida con cada método.

3.2.1. Regresión logística

El modelo de regresión logística (*Logistic regression*) es un modelo lineal generalizado de respuesta binaria ampliamente utilizado a nivel actuarial (ver McCullagh y Nelder, 1989, para más detalles sobre los modelos lineales generalizados y sus aplicaciones), cuya especificación tiene en cuenta tres elementos claves: una variable aleatoria Y, que toma valor y=1 en caso de renovación del contrato o valor y=0 en caso de anulación; un predictor lineal sistemático (η=$\mathbf{X}^{'}\beta$), donde $\mathbf{X}^{'} = (X_0, X_1, ..., X_k)$ y $\beta^{'} = (\beta_0, \beta_1, ..., \beta_k)$ es un vector de parámetros a estimar por máxima verosimilitud, y una función de enlace (g) que relaciona el valor esperado de la variable respuesta con el predictor lineal η. Bajo el modelo de regresión logística se asume que la probabilidad de renovación es:

$$P(Y = 1|\mathbf{X}) = p = \frac{1}{(1 + \exp(-\mathbf{X}^{'}\beta))}. \tag{3.2.1}$$

Una de sus principales ventajas es que es un modelo conocido, cuyos resultados son fáciles de entender, explicar e implementar. Así mismo, resulta bastante informativo dado que apriori aporta pistas sobre la influencia de las variables en la modelización, lo cual es ideal en el momento de decidir la configuración final del modelo. Al mismo tiempo sirve como referente, en el supuesto de que se deseen comparar resultados utilizando modelos más complejos.

Entre sus principales limitaciones se encuentran su forma funcional cerrada, la dificultad de extrapolar los resultados obtenidos en caso de que exista sobre-ajuste (*overfitting*) lo cual ocurre cuando se introducen demasiados parámetros en el modelo y la complejidad en la interpretación de los coeficientes estimados en los casos en los que existen interacciones entre covariables.

3.2.2. Árboles condicionales

Los árboles condicionales (*CTREE-Conditional Tree*) son un tipo de árbol de decisión en los que la selección de las variables se realiza en dos fases, primero se formula una hipótesis global de independencia en términos de m hipótesis parciales, siendo m el número de condiciones. Es decir, se evalúa si existe dependencia entre la variable respuesta Y y cada una de las variables explicativas en $\mathbf{X}^{'}$ y en caso de no poder rechazar la hipótesis nula de independencia planteada se detiene el proceso recursivo. En contraposición, si la hipótesis global de independencia es rechazada, el siguiente paso es medir el nivel de asociación entre la variable dependiente y cada una de las variables

explicativas, lo cual permite generar nuevas divisiones del árbol de manera secuencial (ver Hothorn et al., 2006, para más detalles).

Análogos a los árboles de decisión clásicos, los árboles condicionales se caracterizan por ser poderosas herramientas de clasificación y visualización, además de ser útiles en situaciones en las que el objetivo es agrupar segmentos de clientes, identificar características de un grupo o toma de decisiones de negocio (ver por ejemplo Guelman et al., 2014, para más detalles sobre la implementación de árboles condicionales en la venta cruzada de pólizas de seguros). Otra de sus ventajas es su versatilidad en el caso de que existan relaciones no lineales y en el uso de variables numéricas y categóricas de forma simultánea (ver Friedl y Brodley, 1997), por lo tanto, las transformaciones en intervalos no son necesarias.

Entre sus principales desventajas se encuentran la poca robustez de los resultados en caso de que los datos de entrenamiento difieran mucho de los datos de prueba (ver Kim et al., 2005). Así mismo, se ha de tener en cuenta que los árboles de decisión proveen resultados que son difíciles de interpretar cuando las variables explicativas tienen escalas diferentes.

3.2.3. Redes Neuronales

Las redes neuronales (*Neural Networks*) son métodos complejos de procesamiento de información, inspirados en asociaciones neuronales biológicas, compuestas por neuronas interconectadas cuyos enlaces de comunicación están ponderados (véase Hastie, 1998). Usualmente son vistas como especies de "cajas negras", dado que los resultados son poco intuitivos y, además, aportan poca información acerca de la influencia de las variables explicativas sobre la variable respuesta (dicho análisis deberá realizarse a posteriori). Entre sus principales ventajas destaca su habilidad para detectar relaciones complejas no lineales entre las variables explicativas y la variable respuesta y su capacidad para aprender de tales relaciones. En este sentido, los pesos dentro de la red se ajustan de forma gradual durante la fase de entrenamiento con el objetivo de reducir las diferencias entre el valor real y el valor predicho de la variable respuesta (para más detalles ver Tu, 1996, quien expone las ventajas y desventajas entre el uso de redes neuronales y regresión logística en predicciones médicas). Así mismo, el proceso de modelización no requiere de la especificación de ningún modelo a priori, a diferencia de otras técnicas de modelización no lineal (para más detalles ver Livingstone et al., 1997) y pueden ser entrenadas para generar resultados óptimos a partir de los inputs, por ejemplo minimizar falsos positivos.

Entre sus debilidades más importantes se sitúan la sensibilidad de los resultados respecto al ajuste de los parámetros que controlan el algoritmo; su poca versatilidad en el caso de que se presenten correlaciones extremas entre las variables explicativas utilizadas y, finalmente, su propensión al sobreajuste.

3.2.4. Máquinas de vectores de soporte

Los métodos de máquinas de vector soporte (SVM-*Support Vector Machine*) fueron introducidos por primera vez por Boser et al. (1992) y Cortes y Vapnik (1995), y representan un método de predicción alternativo y comúnmente utilizado en problemas de clasificación, análisis de regresión y detección de valores extremos. Su principal característica es que los datos son mapeados en un espacio de dimensión superior, en donde las clases de la variable respuesta se separan mediante un hiperplano de división óptimo (ver Suykens y Vandewalle, 1999; Meyer y Wien, 2001; Hsu et al., 2003; Hornik et al., 2006; Meyer et al., 2012). Algunas de las ventajas de este método es que el uso de funciones núcleo (*kernel*) flexibiliza la elección del borde de separación entre clases, permitiendo que este tenga una forma no lineal. Así mismo, los resultados son fácilmente extrapolables a otros datos distintos a los de la muestra y es práctico en términos de generación de *scores*.

Por otra parte, entre sus principales limitaciones destacan la velocidad y el tiempo de ejecución de los algoritmos durante las fases de entrenamiento y prueba, las dificultades en cuanto a la interpretación de los resultados y el sesgo del hiperplano óptimo en caso de que el número de individuos en cada clase esté desbalanceado, tal como se evidencia en la contribución de Tao et al. (2006).

3.2.5. Criterios de ajuste de la predicción

Con el objetivo de evaluar la capacidad predictiva de los modelos en la predicción de la deserción proponemos tres criterios diferentes basados en: (C1) sensibilidad más especificidad y (C2) exactitud, ambos vinculados a un punto de corte específico (*threshold* o *cut-off*) y (C3) el área bajo la curva (AUC). Estas medidas resumen en un simple valor escalar el desempeño de cada modelo. Además, verdaderos y falsos positivos son denotados como VP y FP, mientras que verdaderos y falsos negativos son denotados como VN y FN, respectivamente. Así,

$C1 = \max(\text{ sensibilidad} + \text{especificidad}) = \max(\frac{VP}{VP+FN} + \frac{VN}{VN+FP})$
$C2 = \max(\text{exactitud}) = \max(\frac{VP+VN}{VP+VN+FP+FN})$
$C3 = \text{AUC}.$

El área bajo la curva (AUC - *Area under the curve*) es una medida bien conocida por evaluar el poder discriminante de los modelos predictivos. Un área por debajo de la diagonal en el espacio ROC se asocia a un modelo de clasificación aleatoria puro, por lo que a mayor AUC, mejor será el modelo.

3.3. Datos

En este capítulo se utilizan dos sub-muestras, extraídas de las bases de datos originales, con información acerca de las pólizas de auto y hogar contratadas por los clientes. Las bases de datos son independientes a nivel cliente, es decir, no comparten clientes comunes. Por tanto aquellos clientes con una o más pólizas de auto y hogar contratadas de forma simultánea se encuentran excluidos de este estudio. Tal como mencionamos en 1.3.1 y 1.3.2 los clientes que han contratado exclusivamente pólizas de auto o de hogar con la compañía constituyen las bases de datos "DB Auto" y "DB Hogar" respectivamente.

Además, los registros contienen variables relacionadas con la póliza y el cliente como las expuestas en la Tabla 3.2 y variables específicas vinculadas al objeto de riesgo como se muestra en las Tablas 3.3 y 3.4, las cuales se espera ayuden a explicar la deserción de los clientes. Por último, la situación de la póliza desde el inicio de la relación contractual hasta el final es fundamental, es por ello, que las variables con información acerca de la fecha de inicio, anulación y el estatus de cada póliza también son factores claves durante todo el análisis.

TABLA 3.2: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de auto y hogar

Relacionado con	**Variable (Descripción)**
Tomador	Sexo, edad, otras pólizas de auto en vigor, pólizas de otros ramos en vigor, suma de las primas pagadas
Póliza	Estado de la póliza (A= anulada, V= vigente), garantías (TRCF, TRSF, T, O), variación de prima, descuentos - malus, tipo de mediador (A: Agente exclusivo, C= Corredor de seguros)

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de auto y hogar, 2015.

TABLA 3.3: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de auto

Relacionado con	Variable (Descripción)
Objeto de riesgo	Tipo de vehículo, primer conductor - edad, segundo conductor (Si , No), carga siniestral, potencia, nº de plazas

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de auto, 2015.

TABLA 3.4: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de hogar

Relacionado con	Variable (Descripción)
Objeto de riesgo	Tipo de vivienda, continente, contenido

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de hogar, 2015.

El estudio es llevado a cabo en las dos fases usuales, una relacionada con el pre-procesamiento de los datos llamada filtrado y otra vinculada directamente con el proceso de modelización. El pre-procesamiento de los datos es un proceso largo pero necesario previo a la selección de las variables y la implementación del modelo, que incluye entre otras cosas: identificación de valores atípicos, tratamiento de valores perdidos, transformación y/o creación de nuevas variables. Para el proceso de modelización se han implementado las siguientes reglas: realizar una selección manual de las características que deben formar parte de la modelización; para evitar problemas de endogeneidad descartar todas aquellas variables que a priori aporten información acerca de la posible decisión del asegurado respecto a la renovación (Por ejemplo: devolución del último recibo, cancelación de otras pólizas); evitar variables que estén interrelacionadas con otras y descartar predictores innecesarios.

En este estudio se extraen dos sub-muestras del 10 % - una para cada LoB - a partir de las bases de datos DB Auto y DB Hogar definidas en las sub-secciones 1.3.1 y 1.3.2 respectivamente. Un aspecto importante es que aunque utilizamos bases de datos de pólizas, el muestreo fue realizado a nivel cliente. En el tratamiento de los datos, previo a la extracción de las muestras, se tuvieron en cuenta todas aquellas pólizas iniciadas hasta el mes de Diciembre del año 2014 y cuya anulación se produjo durante el año 2015. Este paso redujo el tamaño de nuestras bases de datos de auto y hogar a 197554 y 54781pólizas respectivamente. Además, con el objetivo de utilizar variables que aportaran mayor valor predictivo a los modelos, decidimos crear nuevas variables a partir de algunas ya existentes, como por ejemplo: carga siniestral media (costes total de los

siniestros / número de siniestros con factor culpa del asegurado); ingresos del cliente (monto total de primas pagadas por el cliente / exposición) y ratio de anulaciones del mediador. En este punto, el conjunto de datos de auto está constituido por un 80 % de pólizas en vigor frente a un 20 % de pólizas canceladas, mientras que la base de datos de hogar está compuesta por un 90 % de pólizas en vigor respecto a un 10 % de pólizas anuladas. Esto da lugar a un total de 19751 pólizas de auto y 5954 de hogar.

Finalmente, tal como mencionamos en la Sub-sección 1.3.5 hemos dividido la muestra en dos partes: 70 % para entrenamiento (*training set*) y 30 % para prueba (*test set*). A partir de los resultados de la modelización, generamos para ambos conjuntos de entrenamiento y prueba una matriz de confusión para cada punto de corte posible desde 0 hasta 1 y calculamos cada uno de los criterios propuestos en la Sub-sección 3.2.5. Además, de calcular el área bajo la curva (AUC) tanto para los datos de entrenamiento como para los de prueba.

3.4. Resultados

En esta sección se presentan los resultados de la modelización en función de los diferentes criterios de desempeño propuestos y teniendo en cuenta la muestra de prueba. En la Figura 3.1 se visualizan cada una de las curvas ROC asociadas a los diferentes clasificadores para cada una de las LoB. En el caso de auto, podemos observar múltiples intersecciones, de manera que, no se evidencia una curva que domine a las otras. Por su parte, en el caso de los seguros hogar, aunque también existen múltiples intersecciones es posible observar que la curva ROC vinculada a la máquina de vectores de soporte supera con diferencia al resto de clasificadores. Por cada intersección entre las curvas ROC, tendremos diferentes niveles de falsas alarmas donde un clasificador supere a los otros. En general, la probabilidad aproximada de clasificación errónea se encuentra entre el 25 % y 26 %.

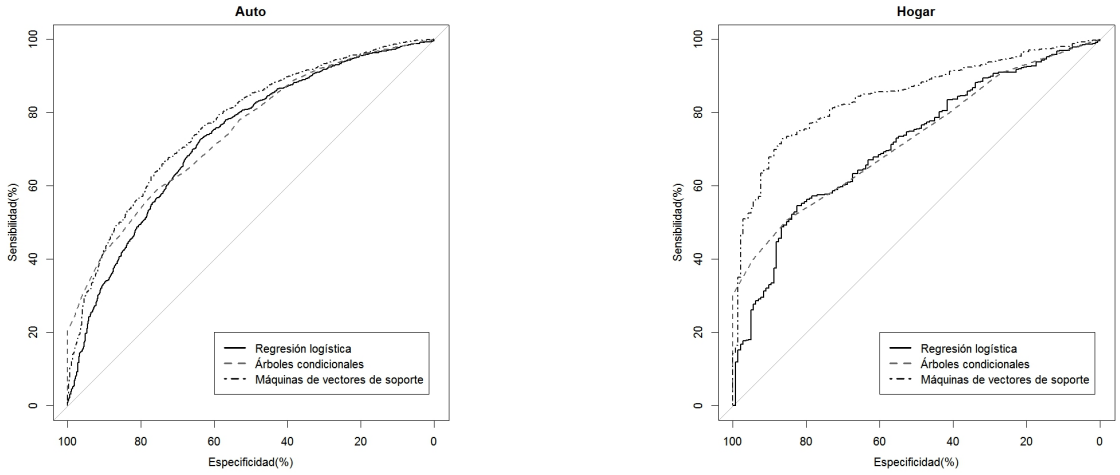


FIGURA 3.1: Curvas ROC para cada LoB y cada método.

Las Tablas 3.5 y 3.6 muestran los resultados para las bases de datos de seguros de auto y hogar, respectivamente. Para cada método obtenemos el umbral t, el cual maximiza C1 y C2. Así mismo, calculamos el área bajo la curva ROC (C3) la cual no depende del valor umbral t. En los seguros de auto, en general se observan resultados similares para cada una de las medidas de rendimiento utilizadas, aunque encontramos ligeras diferencias con respecto al modelo de máquina de vectores de soporte. Finalmente, la mayor área debajo de la curva corresponde al método de máquina de vectores de soporte, seguido por el árbol condicional y el método de regresión logística. En el caso de los seguros de hogar, donde el porcentaje de asegurados que renuevan su póliza supera el 90 %, existen diferencias importantes entre la máquina de vectores de soporte y el resto de métodos. Específicamente, en la Tabla 3.6, es interesante observar cómo los resultados de C1 y C3 para la máquina de vectores de soporte mejoran considerablemente a los obtenidos en la regresión logística y el árbol condicional; sin embargo, los resultados del C2 siguen siendo los mismos, es decir, el porcentaje global de clasificación correcta es similar para todos los métodos.

Una vez aplicados los modelos predictivos y basándonos en la probabilidad de abandono estimada para todos los casos, debemos decidir sobre qué grupos de clientes se debe hacer foco en el momento de implementar acciones específicas para aumentar las probabilidades de que permanezcan en la empresa aquellos clientes que interesan. Para ello, la selección óptima de un umbral es el siguiente paso clave a considerar. Las últimas dos filas de las Tablas 3.5 y 3.6 nos dan posibles alternativas sobre la elección de dicho punto de corte. Aquí, es posible observar intervalos que indican que existe más de un umbral óptimo, es decir, todos los valores que están dentro de cada intervalo alcanzan el mismo criterio de optimalidad.

TABLA 3.5: Criterios de ajuste para cada modelo en los datos de prueba - Auto.

		Modelos		
		Regresión logística	**Árbol condicional**	**Máquina de vectores de soporte**
Criterios	C1	1.36	1.35	1.4
	C2(%)	79	79	80
	C3(%)	73	75	77
Umbral Óptimo	C1	0.77	0.78	0.96
	C2	0.56	[0.39,0.54]	0.81

Fuente: Cálculos propios.

TABLA 3.6: Criterios de ajuste para cada modelo en los datos de prueba - Hogar.

		Modelos		
		Regresión logística	**Árbol condicional**	**Máquina de vectores de soporte**
Criterios	C1	1.35	1.36	1.53
	C2(%)	92	92	92
	C3(%)	71	73	85
Umbral Óptimo	C1	0.92	[0.89,0.93]	0.97
	C2	[0.01,0.37]	[0.01,0.76]	[0.01,0.88]

Fuente: Cálculos propios.

3.5. Conclusiones

En la sección anterior utilizamos tres métodos diferentes para estimar la propensión a la anulación de un grupo de clientes con pólizas de auto u hogar contratadas con la compañía. Estas probabilidades proporcionan un primer modelo de puntuación en cada caso, de forma que es posible identificar grupos de clientes con mayor propensión al abandono e implementar acciones de negocio dirigidas a segmentos de clientes específicos. La selección del mejor modelo no es una elección trivial, sino que depende de las preferencias del asegurador.

La preferencia por un modelo u otro debe estar vinculada a la selección del criterio que proporcione una clasificación lo más equilibrada posible entre los clientes que se quedan y los que abandonan la empresa, pero el criterio no es necesariamente único.

Los criterios de rendimiento tienen uno o más umbrales asociados que pueden servir como guía para la clasificación de los clientes. En este sentido, la selección del umbral debe realizarse con cuidado y debe compararse con la sensibilidad y especificidad obtenidas en cada caso, de lo contrario, esto llevaría a obtener clasificaciones poco acertadas. El hecho de que los clientes con pólizas de hogar contratadas tengan tasas de renovación superiores al 90 % afecta a los resultados obtenidos con los criterios C1 y C2. En este sentido, en el caso de C1, el único método que encuentra un porcentaje igual de sensibilidad y especificidad es el SVM. Además, el porcentaje máximo de clasificación correcta (C2) obtenido con cada uno de los tres modelos se alcanza cuando la sensibilidad es exactamente igual al 100 % y la especificidad es igual a 0 %, de modo que si elegimos los umbrales asociados con la maximización de C2, entonces, todos los clientes se clasificarían en el grupo de renovación y, por lo tanto, se esperaría que todos continúen la relación contractual con la empresa para el año siguiente, lo cual no tiene especial sentido.

Futuras investigaciones deberían incluir diferentes métricas que permitan establecer nuevas comparaciones y evaluar los resultados de los modelos predictivos. Sugerimos probar diferentes intervalos temporales de renovación, con el fin de adaptar el análisis anterior a las características de cada línea de negocio. Así mismo, la estacionalidad es un aspecto fundamental a tener en cuenta en el momento de definir los intervalos de renovación de las pólizas de seguro, ya que los clientes tienden a variar su comportamiento en meses específicos (por ejemplo, antes de las vacaciones, festivos, etc.). Además, sería necesario un monitoreo regular de los resultados y de los umbrales utilizados.

Capítulo 4

Modelización conjunta de la deserción para la cuantificación del riesgo global

4.1. Introducción

Los individuos generalmente buscan proteger lo que valoran (familia, salud, negocios, coche, moto, etc.), lo cual podría incluir la cobertura de uno o más riesgos de forma simultánea. Según la oferta aseguradora de cada compañía, éstas podrían vender más de un tipo de póliza a una misma persona y con ello satisfacer la necesidad global de protección existente.

Los asegurados con múltiples pólizas de seguros contratadas en la misma compañía pertenecen a un grupo específico de riesgo, cuyos perfiles no han sido estudiados en profundidad. Por otra parte, los miembros de una misma familia comparten, entre otras cosas, características comunes y probablemente sus preferencias y actitudes frente al riesgo resulten similares. Así mismo, toman decisiones consensuadas, por lo que no es sorprendente que las decisiones que tomen sobre alguna de sus pólizas termine afectando al resto de pólizas contratadas (ver Brockett et al., 2008; Guillen et al., 2012, 2008, para obtener más información). Por ejemplo, un cliente puede haber contratado una póliza de seguro de automóvil para su cónyuge y otra para su hijo o hija y, al mismo tiempo, puede haber suscrito una póliza de seguro de hogar. Por lo tanto, la renovación de las pólizas podría ser una decisión acordada entre uno o más individuos que pertenecen a la misma unidad familiar, lo cual establece una relación instantánea entre las líneas de negocio.

Las contribuciones existentes acerca de la deserción de clientes en seguros, generalmente se centran en una línea de negocio (*Line of Business* - LoB). Esto significa que los análisis se realizan teniendo en cuenta la información de las pólizas de auto u hogar, pero no la de ambas LoB simultáneamente. De aquí, que se construyan modelos estadísticos diferenciados por LoB (Auto u Hogar) (por ejemplo, ver Bolancé et al., 2016b, donde se expone un análisis comparativo entre diferentes medidas de rendimiento con el objeto de evaluar diversos modelos de deserción de clientes en seguros de auto). Solo un pequeño número de contribuciones abordan de alguna forma el tema de la dependencia entre LoB. Por ejemplo, Brockett et al. (2008) analizan el tiempo del que dispone el asegurador para retener a un cliente con múltiples tipos de pólizas contratadas con la compañía, una vez que éste decide cancelar su primera póliza y así evitar que se lleve el resto de pólizas a la competencia. Por su parte, Bermúdez et al. (2013) utilizando datos históricos agregados pertenecientes al mercado de seguros de no vida en España, comparan los resultados obtenidos al estimar el requerimiento de capital de solvencia (*the solvency capital requirement* - SCR) utilizando un modelo estándar y un modelo interno, esto es, primero bajo una perspectiva de independencia y a continuación teniendo en cuenta la dependencia entre las doce líneas de negocio en las que se segmentan las obligaciones de los seguros de no vida según el quinto estudio de impacto cuantitativo (*the fifth Quantitative Impact Study* - QIS-5). Mientras que, Avanzi et al. (2016) establecen algunas consideraciones metodológicas para analizar la correlación existente entre segmentos de negocio en seguros y resaltan cómo la estimación precisa de los beneficios de la diversificación vinculados a las estructuras de dependencia entre LoB es crucial, primero para la gestión eficiente del capital y segundo para la solvencia de la compañía de seguros. Sin embargo, se desconoce la existencia de publicaciones que aborden el tema de la dependencia entre las líneas de negocio de auto y hogar, tanto para el análisis de la deserción como para el cálculo del valor en riesgo de los clientes. Del mismo modo, en el sector seguros, no hay evidencia que indique que los análisis realizados en cuanto a tarificación, cálculo de reservas o a nivel cliente se estén llevando a cabo teniendo en cuenta la dependencia entre líneas de negocio, ya sea a nivel agregado o a nivel individual.

Focalizándonos en los seguros de no-vida, específicamente en dos de las líneas de negocio personales más importantes, auto - la línea de negocios de seguros de no vida más grande - y hogar, las cuales según *European insurance - key facts* [1] representan cerca

[1] Insurance Europe (2016). European insurance - key facts. https://www.insuranceeurope.eu/sites/default/files/attachments/European%20Insurance%20-%20Key%20Facts%20-%20August%202016.pdf. Consultado: 2017-03-24

del 50 % de todas las primas de no-vida, estamos interesados en modelizar el comportamiento de abandono del asegurado pero desde una perspectiva conjunta, es decir, teniendo en cuenta dependencia entre líneas de negocio.

Contribuimos con la literatura existente al explorar la influencia que tienen algunos factores de riesgo sobre la probabilidad de renovación de las pólizas de seguros de un tipo específico de asegurados. También estudiamos la existencia de dependencia entre LoBs, comparamos el ajuste de un modelo clásico como el probit univariante respecto al ajuste de un modelo probit bivariante [2] y el enfoque cópulas, y estudiamos los efectos marginales resultantes. Para este propósito, proponemos el área bajo la curva ROC (AUC) como el criterio para la evaluación de la capacidad predictiva de los diferentes modelos.

4.2. Modelos predictivos

En esta aplicación, proponemos dos modelos probit diferentes y un enfoque de cópulas para estimar las probabilidades de que un cliente abandone la compañía. Los modelos probit son modelos lineales generalizados, ampliamente reconocidos por su habilidad para relacionar la probabilidad de ocurrencia "p" de un evento con un grupo de variables explicativas (ver McCullagh y Nelder, 1983, para una explicación detallada acerca de los modelos lineales generalizados). Por su parte, una cópula es una función que permite modelizar distribuciones multivariantes teniendo en cuenta una estructura de dependencia específica (ver Frees y Valdez, 1998).

Nuestro análisis comparativo se lleva a cabo desde dos perspectivas diferentes. Primero, analizamos la respuesta del modelo de regresión probit de forma individual, es decir, bajo el supuesto de independencia entre LoB. Segundo, bajo el supuesto de dependencia combinamos la información de ambas LoB, por un lado a través de un modelo modelo probit bivariante y por otro bajo un enfoque de cópulas, para seguidamente deducir las probabilidades marginales y condicionales en cada caso.

4.2.1. Modelo probit univariante

El modelo probit es, junto al logit, uno de los modelos de respuesta binaria más populares (ver Aldrich y Nelson, 1984; Ashford y Sowden, 1970, para más detalles acerca de

[2] Dado que estamos interesados en el estudio de los perfiles de riesgo de clientes que han suscrito ambos tipos de pólizas - auto y hogar - y cuyas decisiones acerca de la renovación de las mismas se presume que están interrelacionadas, proponemos el modelo probit bivariante, el cual permite modelizar dos variables respuestas correlacionadas entre sí.

los modelos probit). En esta aplicación, se propone un modelo probit univariante como referencia para el análisis comparativo realizado más adelante.

4.2.1.1. Especificación del modelo

Sea Y_i la variable aleatoria respuesta asociada al individuo i, que puede tomar solo dos valores posibles, $y_i = 1$ para denotar la ocurrencia del evento, que en este caso se refiere a la renovación de la póliza del cliente, y $y_i = 0$ en caso contrario. Asumiendo que la variable respuesta Y_i sigue una distribución de Bernoulli con probabilidad p_i, el modelo de regresión probit especifica que

$$Pr(Y_i = 1|\mathbf{x}_i) = p_i = \Phi(\mathbf{x}_i^{'}\beta) = \int_{-\infty}^{\mathbf{x}_i^{'}\beta} \frac{1}{\sqrt{2\pi}} \exp\left[-(\frac{t^2}{2})\right] dt, \tag{4.2.1}$$

y la inversa de esta relación, conocida como función *link*, la cual expresa al predictor lineal $\mathbf{x}_i^{'}\beta$ como una función de p_i, está definida como

$$\Phi^{-1}(p_i) = \mathbf{x}_i^{'}\beta + \varepsilon_i, \tag{4.2.2}$$

donde Φ es la función de distribución Normal estándar, $\mathbf{x}_i$ es el vector de covariables formado por un conjunto de características individuales, β es el vector de parámetros a ser estimados por máxima verosimilitud y ε_i es el término de error que se distribuye como una Normal$(0,1)$. En este punto, a partir de la 4.2.2 podemos definir nuestro modelo probit en términos de la variable latente Y_i^*, esto es,

$$Y_i^* = \Phi^{-1}(p_i) = \mathbf{x}_i^{'}\beta + \varepsilon_i,\ Y_i = I(Y_i^* > 0) \tag{4.2.3}$$

donde $I(\bullet)$ es la función indicatriz que toma valor 1 si la condición entre los paréntesis es verdadera y 0 en caso contrario.

En este estudio, ya sea que se trate del modelo de auto o del modelo de hogar, la respuesta predicha p_i corresponde con la probabilidad de que el asegurado i renueve su póliza de auto o bien su póliza de hogar con la compañía y de forma análoga $(1 - p_i)$ representa su propensión particular al abandono o a la no renovación de dichas pólizas en cada caso.

4.2.2. Modelo probit bivariante

4.2.2.1. Especificación del modelo

El modelo probit bivariante fue introducido por Ashford y Sowden (1970) y es un caso particular del modelo de respuesta cualitativa multivariante, definido a partir de la distribución de probabilidad conjunta de dos o más variables dependientes discretas (ver Amemiya, 1985, para más detalles).

Sean $\{y_{ia}, \mathbf{x}_{ia}\}$ y $\{y_{ih}, \mathbf{x}_{ih}\}$ dos conjuntos de valores observados vinculados al i-ésimo cliente, donde y_{ia} y y_{ih} son observaciones que corresponden con las variables dependientes binarias de interés Y_{ia} y Y_{ih}, que hacen referencia a la ocurrencia de una cancelación en las LoB de auto (a) y hogar (h). Sean $\mathbf{x}_{ia}, a = 1 \dots p$ y $\mathbf{x}_{ih}, h = 1 \dots q$ dos vectores de factores de riesgo conocidos y vinculados también a cada LoB. De forma análoga a la Sección 4.2.1.1 cada una de las variables respuesta toma valor 1 cuando el cliente renueva su póliza ó 0 cuando la póliza es cancelada.

Análogo al modelo probit univariante, la especificación del modelo probit bivariante puede expresarse en términos de variables latentes (ver Greene, 2003, para más detalles sobre un modelo de regresión latente), esto es:

$$Y_{ia}^{*} = \Phi^{-1}(p_{ia}) = \mathbf{x}_{ia}^{'}\beta_a + \varepsilon_{ia},\ Y_{ia} = I(Y_{ia}^{*} > 0) \tag{4.2.4}$$

$$Y_{ih}^{*} = \Phi^{-1}(p_{ih}) = \mathbf{x}_{ih}^{'}\beta_h + \varepsilon_{ih},\ Y_{ih} = I(Y_{ih}^{*} > 0). \tag{4.2.5}$$

Aquí, la distribución de probabilidad conjunta de las variables dependientes discretas Y_{ia} y Y_{ih} $\forall i = 1, \dots, n$, puede ser descrita por una tabla de probabilidades de cuatro celdas, tal como la mostrada en la Tabla 4.1, donde

$$p_{jk} = Pr(Y_{ia} = j, Y_{ih} = k) = \Phi_{\rho}(\mathbf{x}_{ia}^{'}\beta_a + \varepsilon_{ia}, \mathbf{x}_{ih}^{'}\beta_h + \varepsilon_{ih}),\quad j,k = \{0,1\} \tag{4.2.6}$$

representa la función de probabilidad conjunta, dada por una función de distribución normal bivariante con media cero, varianzas unitarias y correlación ρ. Es decir, en el modelo probit bivariante se asume que

$$(\varepsilon_{im}, \varepsilon_{ih}) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

En este sentido, es importante destacar que las variables dependientes Y_{ia}^* y Y_{ih}^* deberían estar correlacionadas a través del parámetro de correlación de los errores ρ, de lo contrario, dichos modelos podrían estimarse de forma independiente.

Una vez definido el modelo probit, la estimación de los vectores de parámetros desconocidos β_m, β_h y ρ se realiza por máxima verosimilitud teniendo en cuenta toda la información de ambas LoB simultáneamente. Como resultado, obtenemos dos vectores de probabilidades estimadas, es decir, las distribuciones marginales de cada LoB,

TABLA 4.1: Distribución conjunta de las variables aleatorias binarias

		Y_{ih}	
		1	0
Y_{ia}	1	p_{11}	p_{10}
	0	p_{01}	p_{00}

4.2.2.2. Probabilidad condicional

Denominamos probabilidad condicional de $Y_{ia}|Y_{ih}$, como la probabilidad de que el cliente i renove su póliza de auto teniendo en cuenta la decisión que ha tomado acerca de la renovación de su póliza de hogar. Dicha probabilidad queda definida como:

$$Pr(Y_{ia} = j|Y_{ih} = k, \mu_c, \sigma_c) = \Phi_\rho(\mathbf{x}_{ia}^{'}\beta_a|\mathbf{x}_{ih}^{'}\beta_h, \mu_c, \sigma_c),\ j = k = \{0,1\} \tag{4.2.7}$$

donde μ_c representa la **esperanza condicional** del individuo i dada por

$$E(Y_{ia}|Y_{ih}) = \mu_{ic} = \mu_{ia} + \sigma_{ah}\sigma_{hh}^{-1}(\mathbf{x}_h^{'}\beta_h - \mu_{ih}), \tag{4.2.8}$$

y σ_c representa la **covarianza condicional** expresada como

$$var(Y_{ia}|Y_{ih}) = \sigma_c = \sigma_{aa} - \sigma_{ah}\sigma_{hh}^{-1}\sigma_{ha}, \tag{4.2.9}$$

siendo σ_{aa}, σ_{ah}, σ_{hh} y σ_{ha} las respectivas varianzas y covarianzas de las variables latentes correspondientes.

A partir del modelo probit bivariante, podemos establecer la distribución de la probabilidad condicional de $Y_{ia}|Y_{ih} \sim N(\rho \mathbf{x}_{ih}^{'}\beta_h, 1-\rho^2)$ a ser estimada. De forma análoga, se define la probabilidad condicional de $Y_{ih}|Y_{ia}$.

4.3. Modelización de la distribución bivariante mediante cópulas

Las cópulas representan un enfoque diferente y flexible para modelar distribuciones multivariantes de mayor dimensión y comprender la relación entre los diferentes factores de riesgo (ver Joe, 1997; Nelsen, 2007), para una introducción a las cópulas).

Abe Sklar demuestra mediante su famoso Teorema de Sklar (ver Sklar, 1959, 1973, para más detalles), que dada F una función de distribución conjunta con marginales $F_1, \ldots, F_k$ asociadas al vector aleatorio k-dimensional $\mathbf{Y}$, existe una función de distribución apropiada $C : [0,1]^k \longrightarrow [0,1]$ denominada "cópula", con marginales uniformes, tal que se cumple:

$$F(y_1, \ldots, y_k) = C(F_1(y_1), \ldots, F_k(y_k)), \tag{4.3.1}$$

esto es, una función de distribución multivariante definida en el cubo unitario k-dimensional $[0,1]$ que permite separar la estructura de dependencia del comportamiento marginal.

Dado que nuestra aplicación se centra en la modelización de la dependencia entre dos líneas de negocio, simplificaremos la expresión presentada en 4.3.1 a nuestro caso particular bivariante, es decir,

$$F(y_{ia}, y_{ih}) = C_\theta(F_a(y_{ia}), F_h(y_{ih})) \tag{4.3.2}$$

donde y_{ia} y y_{ih} corresponden con la ocurrencia de una cancelación en las líneas de negocio de auto (a) y hogar (h) como mencionamos en la Sección 4.2.2, F_a y F_h son las funciones de distribución marginales y θ es el parámetro de cópula que describe la dependencia entre las variables aleatorias Y_{ia} y Y_{ih}.

4.3.1. Modelización de la dependencia con variables discretas

La modelización de la distribución conjunta de variables aleatorias dependientes discretas supone un nivel de dificultad añadido, pues la estimación de los parámetros del modelo se debe realizar teniendo en cuenta las diferentes combinaciones de valores posibles que las variables dicotómicas pueden asumir.

Uno de los primeros trabajos en tener en cuenta esta problemática y proveer un procedimiento para la estimación de los parámetros teniendo en cuenta las correlaciones existentes entre variables aleatorias dependientes discretas fue el desarrollado por Zellner y Lee (1965). Allí se plantea una aplicación donde se evalúa la relación entre la decisión de compra de un bien (compra o no compra) y la decisión de usar o no un crédito a plazos; y comprueba cómo la estimación conjunta permite obtener estimadores con varianzas asintóticas más pequeñas que aquellas obtenidas obviando la correlación existente. Lee (1983) propuso un modelo de regresión censurado de dos ecuaciones donde las funciones de distribución de los términos de perturbación son absolutamente continuas y están correlacionadas. En su trabajo, se sugiere que para que la especificación de dicho modelo esté completa, se debe elegir una distribución bivariante apropiada que tenga en cuenta las distribuciones marginales especificadas. Por su parte, Ophem (1999) generaliza el método de Lee (1983) para variables discretas. Específicamente, presenta un modelo simultáneo donde la estimación de los parámetros se realiza teniendo en cuenta dos variables aleatorias discretas correlacionadas cuyas distribuciones univariantes son conocidas. La estimación de los parámetros se realiza especificando el logaritmo de la función de verosimilitud en función de distribuciones normales bivariantes aplicadas sobre las distribuciones marginales originales. Los resultados obtenidos en este estudio, permiten evidenciar cómo aunque la magnitud de los coeficientes estimados no es muy distinta bajo el supuesto o no de correlación entre las variables dependientes, los errores estándar si se ven afectados pues son considerablemente mayores cuando no se tiene en cuenta tal correlación. Finalmente, Winkelmann (2009) presenta un enfoque similar al que tenemos en cuenta en este Capítulo, donde bajo el supuesto de distribuciones marginales probit se induce dependencia no normal entre dos variables discretas utilizando cópulas, es decir, integra el enfoque cópulas en el modelo de respuesta binaria probit bivariante.

4.3.2. Cópulas utilizadas

En este capítulo utilizaremos las cópulas Gaussiana y t-Student, bien conocidas por ser las representantes más importantes de la familia de cópulas elípticas. Su expresión se

deriva de funciones de distribución multivariante a través del Teorema de Sklar y su estructura coincide con funciones de distribución conocidas como lo son la Normal y la t-Student (ver Fang et al., 1990; Embrechts et al., 2001, para una explicación detallada acerca de cópulas elípticas).

Una de las características más importantes de este tipo de cópulas es su simetría y dependencia del valor θ asociado a la correlación entre las marginales, aunque difieren entre otras cosas, en la fuerza de dependencia en las colas de la distribución bivariante. En este sentido, la Cópula Gaussiana supone que la dependencia en las colas es 0, mientras que la cópula t de Student sí que admite dependencia en las colas.

Es importante destacar que el parámetro θ en la cópula Gaussiana coincide con la correlación lineal en la distribución normal. Por su parte, la cópula t-Student depende de un parámetro adicional ν, que equivale a los grados de libertad y proporciona información acerca de la fuerza de la dependencia entre las variables aleatorias consideradas, en nuestro caso, Y_{ia} y Y_{ih}. En este sentido, la cópula t de Student permite la modelización conjunta de eventos extremos ya sea en ambas colas de la distribución bivariante o en ninguna de ellas.

4.3.3. Función de verosimilitud para el modelo de respuesta binaria bivariante

En esta parte, se define la función de verosimilitud para la estimación del parámetro óptimo de dependencia θ. Tal como mencionamos en la Sección 4.2.2.1, el modelo probit bivariante definido a partir de las probabilidades marginales 4.2.4 y 4.2.5 , tiene una distribución conjunta genérica descrita en función de las siguientes probabilidades:

$$p_{00} = P(Y_{ia} = 0, Y_{ih} = 0 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = P(\varepsilon_{ia} \leq -\mathbf{x}_{ia}^{'}\beta_a, \varepsilon_{ih} \leq -\mathbf{x}_{ih}^{'}\beta_h), \tag{4.3.3}$$

$$p_{10} = P(Y_{ia} = 1, Y_{ih} = 0 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = P(\varepsilon_{ia} > -\mathbf{x}_{ia}^{'}\beta_a, \varepsilon_{ih} \leq -\mathbf{x}_{ih}^{'}\beta_h), \tag{4.3.4}$$

$$p_{01} = P(Y_{ia} = 0, Y_{ih} = 1 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = P(\varepsilon_{ia} \leq -\mathbf{x}_{ia}^{'}\beta_a, \varepsilon_{ih} > -\mathbf{x}_{ih}^{'}\beta_h), \tag{4.3.5}$$

$$p_{11} = P(Y_{ia} = 1, Y_{ih} = 1 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = P(\varepsilon_{ia} > -\mathbf{x}_{ia}^{'}\beta_a, \varepsilon_{ih} > -\mathbf{x}_{ih}^{'}\beta_h). \tag{4.3.6}$$

Tal como mencionamos en la Sub-sección 4.3.1, bajo el enfoque de cópulas y teniendo en cuenta marginales probit, las probabilidades anteriores se reescriben de la siguiente forma:

$$p_{00} = P(Y_{ia} = 0, Y_{ih} = 0 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = C_\theta(1 - \Phi(\mathbf{x}_{ia}^{'}\beta_a), 1 - \Phi(\mathbf{x}_{ih}^{'}\beta_h)), \quad (4.3.7)$$

$$p_{10} = P(Y_{ia} = 1, Y_{ih} = 0 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = C_\theta(1, 1 - \Phi(\mathbf{x}_{ih}^{'}\beta_h)) - C_\theta(1 - \Phi(\mathbf{x}_{ia}^{'}\beta_a), 1 - \Phi(\mathbf{x}_{ih}^{'}\beta_h)), \quad (4.3.8)$$

$$p_{01} = P(Y_{ia} = 0, Y_{ih} = 1 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = C_\theta(1 - \Phi(\mathbf{x}_{ia}^{'}\beta_a), 1) - C_\theta(1 - \Phi(\mathbf{x}_{ia}^{'}\beta_a), 1 - \Phi(\mathbf{x}_{ih}^{'}\beta_h)), \quad (4.3.9)$$

$$p_{11} = P(Y_{ia} = 1, Y_{ih} = 1 | \mathbf{x}_{ia}, \mathbf{x}_{ih}) = 1 - C_\theta(1 - \Phi(\mathbf{x}_{ia}^{'}\beta_a), 1) - C_\theta(1, 1 - \Phi(\mathbf{x}_{ih}^{'}\beta_h)) + p_{00}. \quad (4.3.10)$$

Para la definición de estas expresiones se ha tenido en cuenta la propiedad de simetría respecto a la media de la distribución normal y por otra parte la definición de probabilidad conjunta.

Función de verosimilitud

Asumiendo que se tiene una muestra independiente de N individuos $(y_{ia}, y_{ih}, \mathbf{x}_{ia}, \mathbf{x}_{ih})$, el logaritmo de la función de verosimilitud $L(y_a, y_h, \theta)$ para el caso discreto, se define como:

$$\begin{aligned} log(L(\theta)) = \sum log(&P(y_{ia} = 1, y_{ih} = 1) \times P(y_{ia} = 1, y_{ih} = 0) \\ &\times P(y_{ia} = 0, y_{ih} = 1) \times P(y_{ia} = 0, y_{ih} = 0)). \end{aligned} \quad (4.3.11)$$

La estimación del parámetro óptimo de cópula θ se realiza por el método "L-BFGS-B" de la función *optim* del software R, el cual está basado en el algoritmo "*quasi-Newton*" y es reconocido por ser útil para resolver problemas de optimización no lineal con restricciones simples en las variables (ver Byrd et al., 1995, para más detalles acerca de su implementación).

Matriz Hessiana para el ajuste de los errores estándar

Tal como mencionamos en la Sección 4.2.1, para cada LoB se define un modelo univariante a partir del cual se estiman las probabilidades de renovación de cada póliza. En este sentido, la estimación de los vectores de parámetros β_a para el caso de auto o β_h

para el caso de hogar, se realiza de forma independiente, es decir, se omite la correlación existente entre las LoB. Este supuesto de independencia entre LoB conlleva una estimación errónea de los errores estándar.

Una forma de re-calcular los errores estándar, de forma correcta, bajo la hipótesis de dependencia es a través de la inversa de la matrix Hessiana **H**, la cual contiene información acerca de la varianza y covarianza de los parámetros del modelo. Dicha matriz está compuesta por las derivadas parciales de segundo orden del logaritmo de la función de verosimilitud evaluadas en el vector de parámetros estimados.

Asumiendo que la estimación de los parámetros en las marginales realmente es insesgada y que por tanto es muy similar a la obtenida con la estimación correcta y además teniendo en cuenta el parámetro óptimo de cópula θ calculado anteriormente, definimos $\gamma = (\widehat{\beta}_h, \widehat{\beta}_a, \widehat{\theta})$ como el vector de parámetros estimados donde se evalúa la matriz **H**.

$$H(\gamma) = \begin{pmatrix} \frac{\partial^2}{\partial \gamma_1^2} & \frac{\partial^2}{\partial \gamma_1 \partial \gamma_2} & \cdots & \frac{\partial^2}{\partial \gamma_1 \partial \gamma_p} \\ \frac{\partial^2}{\partial \gamma_2 \partial \gamma_1} & \frac{\partial^2}{\partial \gamma_2^2} & \cdots & \frac{\partial^2}{\partial \gamma_2 \partial \gamma_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \gamma_p \partial \gamma_1} & \frac{\partial^2}{\partial \gamma_p \partial \gamma_2} & \cdots & \frac{\partial^2}{\partial \gamma_p^2} \end{pmatrix} \tag{4.3.12}$$

Una vez definida la matriz **H**, el siguiente paso es especificar lo que comúnmente se denomina "matriz de información" (*information matrix*), definida como

$$\Im(\gamma) = -E[H(\gamma)], \tag{4.3.13}$$

y cuya inversa equivale a la matriz de varianzas y covarianzas,

$$var(\gamma) = Inv(\Im(\gamma)). \tag{4.3.14}$$

Finalmente, la raíz cuadrada de los elementos de la diagonal de la matriz de varianzas y covarianzas son los estimadores de los errores estándar.

Tal como veremos más adelante, el cálculo de los errores estándar de los parámetros estimados del modelo de cópula permitirá evaluar los resultados de los modelos implementados, utilizando cópulas para medir la dependencia.

4.3.4. Probabilidad condicional con cópulas

Análogamente a la Sección 4.2.2.2, estamos interesados en la estimación de la probabilidades condicionales, pero ahora utilizando la teoría de cópulas.

La probabilidad condicional de y_{ia} dado y_{ih} definida por Darsow et al. (1992) y por Joe (1996), representa el caso más simple y se define como

$$F(y_{ia}|y_{ih}) = P(Y_{ia} \leq y_{ia}|Y_{ih} = y_{ih};\theta) = \frac{\partial C_\theta(y_{ia}, y_{ih})}{\partial y_{ih}}. \tag{4.3.15}$$

4.4. Datos

En este capítulo, analizamos una muestra de 22106 asegurados, quienes han contratado dos pólizas de seguro con la compañía, en dos líneas de negocio distintas. Tal como describimos en la Sección 1.3.4, se crea una base de datos con información agregada a nivel cliente, a partir de tres bases de datos complementarias. El primer y el segundo conjunto de datos contienen información acerca de las pólizas, el cliente y el objeto de riesgo, tanto para el ramo de auto como para el ramo de hogar. Por su parte, la tercera base de datos incluye la información siniestral de cada cliente.

Similar al tratamiento de los datos realizado en el Capítulo 3 tuvimos en cuenta todas aquellas pólizas iniciadas hasta el mes de Diciembre del año 2014 y cuya anulación se produjo durante el año 2015. Este paso redujo el tamaño de nuestra base de datos de clientes de 64802 a 29132 registros. Además para simplificar el análisis, en esta aplicación sólo consideramos clientes con una única póliza contratada en cada LoB, es decir, los clientes con más de una póliza contratada para cada LoB han sido excluidos del estudio. Este último paso, redujo la muestra a un total de 22106 registros tal como como especificamos al inicio de esta Sección. Finalmente, tal como se mencionó en la Sección 1.3.5 dividimos la muestra en dos partes, en una proporción de 70 % para entrenamiento y 30 % para prueba, de modo que fuera posible evaluar la capacidad predictiva de los modelos a posteriori.

Las estadísticas descriptivas de las variables que conforman la base de datos de clientes se exponen en las Tablas B.1, B.2, B.3 y B.4 del Apéndice B. Así mismo, la descripción de las variables presentadas en nuestro análisis se detallan en la Sección 1.3 de esta Tesis.

4.5. Resultados

Intuitivamente, podemos anticipar que existe correlación entre las LoB. Sin embargo, buscamos evidencias cuantitativas que lo confirmen. Un primer análisis consiste en identificar la proporción de pólizas suscritas por los clientes en cada línea de negocio. La Tabla 4.2 muestra el porcentaje global de pólizas contratadas en una o ambas LoB.

TABLA 4.2: Pólizas contratadas por los clientes. Descripción en porcentajes de clientes sobre el total

		Número de pólizas de Auto			
		0	**1**	**2**	**3**
Nº de pólizas de hogar	**0**	-	62.47 %	4.58 %	0.21 %
	1	19.52 %	9.04 %	1.27 %	0.09 %
	2	1.64 %	1.03 %	0.14 %	0.01 %

La venta de seguros agrupada no es una práctica común en España. Sin embargo, es posible observar que aproximadamente el 11.5 % de los asegurados de una misma compañía tienen contratadas pólizas de auto y hogar de forma simultánea.

Con el objetivo de contrastar que existe correlación entre las cancelaciones de cada línea de negocio, se estiman las probabilidades de renovación a través de un modelo probit bivariante y mediante cópulas, y se examina por una parte el coeficiente de correlación resultante ρ y el parámetro de cópula θ.

En las Tablas 4.3 y 4.4 se presentan los resultados de las estimaciones correspondientes a los modelos probit univariante y bivariante para ambas líneas de negocio. En este último obtenemos un ρ estimado igual a -0.286, el cual es significativo al 99.9 %. Este resultado indica que los factores no incluidos en la modelización inducen correlación negativa, es decir que indicarían una propensión opuesta a renovar la póliza de auto y hogar. Este resultado sugiere que tiene sentido modelizar ambas líneas de negocio simultáneamente. En este sentido, si la correlación existente no fuese tomada en cuenta, la estimación de la precisión de los parámetros sería incorrecta.

En las Tablas 4.3 y 4.4 a primera vista se observa que, la más alta concentración de efectos significativos está relacionada con las características de la póliza y además la mayoría de estos impactos son significativos en el caso de los seguros de auto. Un aspecto interesante es que los coeficientes de las variables: pol_other (Pólizas de otros

TABLA 4.3: Resultados del modelo probit univariante

		LoB			
		Hogar		**Auto**	
	Variables	**Coeficientes**	**Nivel de significación**	**Coeficientes**	**Nivel de significación**
Comunes	(Intercept)	1.121	**	2.464	***
	client_sex - M	0.027	-	0.156	***
	client_age	0.001	-	0.000	-
	pol_lastrenewal	-0.006	-	-0.001	***
	CAI_relativ	-0.002	***	-0.001	-
	pol_other - 0	-1.135	***	-1.280	***
	mediator_currentpol	0.004	***	0.002	**
	mediator_type - A	0.088	*	0.239	***
	pol_supplements	-0.153	***	-0.091	***
	pol_diffpremium	0.001	.	0.001	-
	pol_malus	0.003	-	0.003	***
	pol_surcharge	0.000	-	-0.003	***
	pol_malus last	0.000	-	-0.003	***
	pol_surcharge last	-0.002	.	0.003	***
	pol_waytopay - A	0.162	-	-0.156	.
	pol_waytopay - S	0.032	-	-0.062	-
	pol_age	0.032	***	0.020	***
	mediator_cancelpol	-0.282	*	-0.032	-
Hogar	home_type - A	0.538	.		
	home_type - B	0.373	-		
	home_type - C	0.288	-		
	home_type - D	0.470	.		
	pol_capi_continent	0.000	-		
	pol_capi_content	0.003	**		
Auto	pol_guarantees - O			-0.512	***
	pol_guarantees - T			-0.522	***
	pol_guarantees - TRCF			-0.195	*
	firstdriver_agelicense			0.003	-
	veh_power			0.001	*
	veh_weightpower			0.000	-
	veh_seats - 2			-0.058	-
	veh_seats - 4			-0.168	.
	veh_seats - 5			-0.130	*
	veh_fueltype - D			-0.337	-
	veh_fueltype - G			-0.397	-
	veh_seconddriver - No			-0.077	-

Fuente: Cálculos propios. *** Indica una significación del 1 %, ** (5 %) y * (10 %), respectivamente.

TABLA 4.4: Resultados del modelo probit bivariante

		LoB			
		Hogar		**Auto**	
	Variables	**Coeficientes**	**Nivel de significación**	**Coeficientes**	**Nivel de significación**
Comunes	(Intercept)	1.095	***	2.398	***
	client_sex - M	0.026	-	0.158	***
	client_age	0.001	-	0.000	-
	pol_lastrenewal	-0.007	-	-0.001	***
	CAI_relativ	-0.002	***	-0.002	-
	pol_other - 0	-1.534	***	-1.725	***
	mediator_currentpol	0.004	***	0.002	***
	mediator_type - A	0.076	**	0.233	***
	pol_supplements	-0.149	***	-0.090	***
	pol_diffpremium	0.001	**	0.001	-
	pol_malus	0.003	-	0.003	***
	pol_surcharge	0.000	-	-0.003	***
	pol_malus last	-0.001	-	-0.003	***
	pol_surcharge last	-0.002	*	0.003	***
	pol_waytopay - A	0.185	-	-0.161	*
	pol_waytopay - S	0.057	-	-0.060	-
	pol_age	0.031	***	0.020	***
	mediator_cancelpol	-0.270	**	-0.017	-
Hogar	home_type - A	0.556	**		
	home_type - B	0.404	-		
	home_type - C	0.315	-		
	home_type - D	0.501	**		
	pol_capi_continent	0.000	-		
	pol_capi_content	0.003	***		
Auto	pol_guarantees - O			-0.510	***
	pol_guarantees - T			-0.525	***
	pol_guarantees - TRCF			-0.200	***
	firstdriver_agelicense			0.003	-
	veh_power			0.001	**
	veh_weightpower			0.000	-
	veh_seats - 2			-0.061	-
	veh_seats - 4			-0.158	*
	veh_seats - 5			-0.118	**
	veh_fueltype - D			-0.252	-
	veh_fueltype - G			-0.311	-
	veh_seconddriver - No			-0.073	-

Fuente: Cálculos propios. *** Indica una significación del 1 %, ** (5 %) y * (10 %), respectivamente.

ramos en vigor), mediator_currentpol (Mediador - pólizas en vigor), mediator_type (Tipo de mediador), pol_supplements (Suplementos), pol_surcharge_last (Recargo (anterior)) y pol_age (Antiguedad de la póliza), tienen un efecto significativo en ambas LoB y ambos modelos. También, se observa que a excepción del coeficiente de la variable pol_surcharge_last (Recargo (anterior)), todos los coeficientes contribuyen o no a la propensión de la renovación de las pólizas de cada LoB en la misma dirección.

La variable "Otras pólizas", nos da información acerca de si el cliente tiene otro tipo de riesgos asegurados con la compañía, es decir, si ha contratado pólizas en otras líneas de negocio, por ejemplo, una póliza de vida, un seguro de salud o si tiene asegurado su negocio. La categoría "0" de esta variable indica que el cliente no tiene otras pólizas contratadas en otras líneas de negocio. El parámetro estimado además de ser significativo es negativo, por tanto tiene sentido que el hecho de no tener una mayor vinculación con la compañía afecte de forma negativa la propensión a la renovación de sus pólizas de auto y hogar.

Los *suplementos* podrían entre otras cosas involucrar pagos adicionales respecto al fijado por la compañía en el momento de la suscripción del riesgo, por ejemplo: la contratación de garantías o coberturas adicionales, los cuales encarecerían el precio de la prima pagada por el cliente. Por otra parte, en la línea de negocio de auto, se observa que los coeficientes estimados para las diferentes categorías de la variable pol_guarantees son negativos, lo cual sugiere que independientemente de la elección de una categoría u otra, existe un efecto contraproducente sobre la probabilidad de renovación, lo cual tendría sentido si se tiene en cuenta que un cambio en las coberturas podría aumentar el precio de la prima.

La segunda parte de nuestro análisis consiste en comparar las curvas ROC obtenidas a partir de las probabilidades estimadas para cada modelo e implementadas en la muestra de test. En la Figura 4.1 se exponen las curvas ROC relacionadas con la muestra de prueba para cada LoB, cuyas probabilidades han sido obtenidas a partir de las expresiones (4.2.1), (4.2.4), (4.2.5) y (4.2.7), descritas anteriormente en la Sección 4.2. Para cada LoB se observa cómo la curva ROC vinculada a las probabilidades condicionales supera a las curvas ROC vinculadas al modelo probit univariante y a las distribuciones marginales del modelo probit bivariante.

Por su parte, en la Tabla 4.5 se presentan las diferentes áreas bajo las curvas ROC obtenidas para cada modelo. Aquí, es posible observar cómo el área bajo las curvas aumenta una vez que se combina la información de cada LoB a nivel cliente.

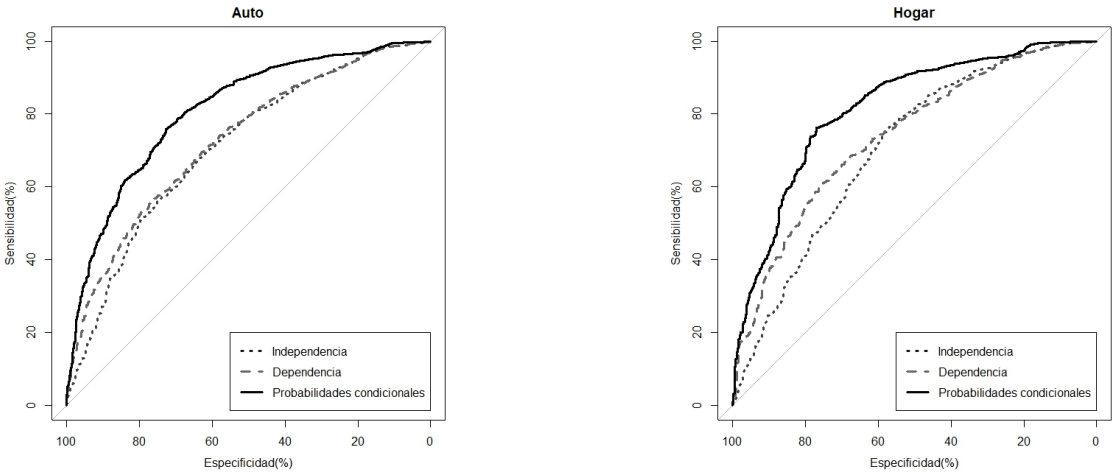


FIGURA 4.1: Áreas bajo las curvas ROC en función de la distribución utilizada.

TABLA 4.5: Resultados de los modelos probit. Áreas bajo las curvas ROC obtenidas a partir del modelo probit univariante (Independencia), las marginales del modelo probit bivariante (Dependencia) y las probabilidades condicionales definidas a partir del modelo probit bivariante

	Independencia	**Dependencia**	**Probabilidades condicionales**
Auto	70.92 %	72.67 %	81.00 %
Hogar	70.43 %	73.83 %	81.57 %

4.5.1. Resultados bajo el enfoque cópulas

En la tercera parte de nuestro análisis nos centramos en el enfoque cópulas. En las Figuras 4.2 y 4.3 presentamos dos tablas con los *p-values* obtenidos para cada parámetro estimado tras el re-cálculo de los errores estándar siguiendo la metodología presentada en la Sección 4.3

FIGURA 4.2: Comparativa de la significación de los *p-values* para cada modelo - Hogar

	Variables	Probit Univariante	Copulas		Probit Bivariante
			Gauss_cop	t-student	
Comunes	(Intercept)	0.002	0.000	0.000	0.002
	client_sex - M	0.486	0.013	0.015	0.498
	client_age	0.322	0.001	0.001	0.423
	pol_lastrenewal	0.830	0.213	0.222	0.800
	CAI_relativ	0.000	0.000	0.000	0.000
	pol_other - 0	0.000	0.000	0.000	0.000
	mediator_currentpol	0.000	0.000	0.000	0.001
	mediator_type - A	0.017	0.000	0.000	0.037
	pol_supplements	0.000	0.000	0.000	0.000
	pol_diffpremium	0.051	0.000	0.000	0.056
	pol_malus	0.360	0.006	0.015	0.317
	pol_surcharge	0.697	0.058	0.061	0.692
	pol_malus_last	0.909	0.381	0.398	0.850
	pol_surcharge_last	0.068	0.000	0.000	0.076
	pol_waytopay - A	0.447	0.008	0.011	0.369
	pol_waytopay - S	0.883	0.322	0.327	0.789
	pol_age	0.000	0.000	0.000	0.000
	mediator_cancelpol	0.028	0.000	0.000	0.031
Hogar	home_type - A	0.057	0.000	0.000	0.038
	home_type - B	0.155	0.000	0.000	0.102
	home_type - C	0.307	0.001	0.003	0.237
	home_type - D	0.074	0.000	0.000	0.044
	pol_capi_continent	0.221	0.000	0.000	0.199
	pol_capi_content	0.003	0.000	0.000	0.004

FIGURA 4.3: Comparativa de la significación de los *p-values* para cada modelo - Auto

	Variables	Probit Univariante	Copulas		Probit Bivariante
			Gauss_cop	t-student	
Comunes	(Intercept)	0.000	0.000	0.000	0.000
	client_sex - M	0.000	0.000	0.000	0.000
	client_age	0.922	0.387	0.399	0.983
	pol_lastrenewal	0.000	0.000	0.000	0.000
	CAI_relativ	0.137	0.001	0.006	0.113
	pol_other - 0	0.000	0.000	0.000	0.000
	mediator_currentpol	0.007	0.000	0.000	0.009
	mediator_type - A	0.000	0.000	0.000	0.000
	pol_supplements	0.000	0.000	0.000	0.000
	pol_diffpremium	0.257	0.001	0.002	0.268
	pol_malus	0.000	0.000	0.000	0.000
	pol_surcharge	0.000	0.000	0.000	0.000
	pol_malus_last	0.000	0.000	0.000	0.000
	pol_surcharge_last	0.000	0.000	0.000	0.000
	pol_waytopay - A	0.098	0.000	0.000	0.083
	pol_waytopay - S	0.533	0.031	0.041	0.537
	pol_age	0.000	0.000	0.000	0.000
	mediator_cancelpol	0.779	0.212	0.234	0.880
Auto	pol_guarantees - O	0.000	0.000	0.000	0.000
	pol_guarantees - T	0.000	0.000	0.000	0.000
	pol_guarantees - TRCF	0.028	0.000	0.000	0.022
	firstdriver_agelicense	0.218	0.000	0.001	0.215
	veh_power	0.033	0.000	0.000	0.033
	veh_weightpower	0.956	0.431	0.436	0.982
	veh_seats - 2	0.768	0.184	0.206	0.754
	veh_seats - 4	0.054	0.000	0.000	0.068
	veh_seats - 5	0.031	0.000	0.000	0.047
	veh_fueltype - D	0.224	0.000	0.000	0.352
	veh_fueltype - G	0.154	0.000	0.000	0.252
	veh_seconddriver - No	0.117	0.000	0.000	0.130

Las celdas grises están asociadas a los parámetros estimados que son significativamente distintos de 0 al 1 % y 5 % para cada uno de los modelos trabajados. Se observa que aunque existen parámetros significativos comunes a todos los modelos, el enfoque cópulas permite identificar nuevos efectos significativos hasta el momento indetectables por los modelos probit, dado que el enfoque cópulas tiene en cuenta directamente la dependencia existente entre las marginales, a diferencia del modelo probit bivariante donde la dependencia viene dada por la correlación entre los errores de los modelos de regresión asociados a las variables latentes.

Por último, en las Figuras 4.4 y 4.5 presentamos los resultados de las estimaciones de las probabilidades condicionales bajo el enfoque cópulas. En todos los casos se observa, cómo los resultados vinculados a las probabilidades condicionales tanto para los modelos probit bivariante como para el caso de las cópulas Gaussiana y t-Student suponen una mejora en las estimaciones de las probabilidades estimadas respecto al uso de un modelo univariante.

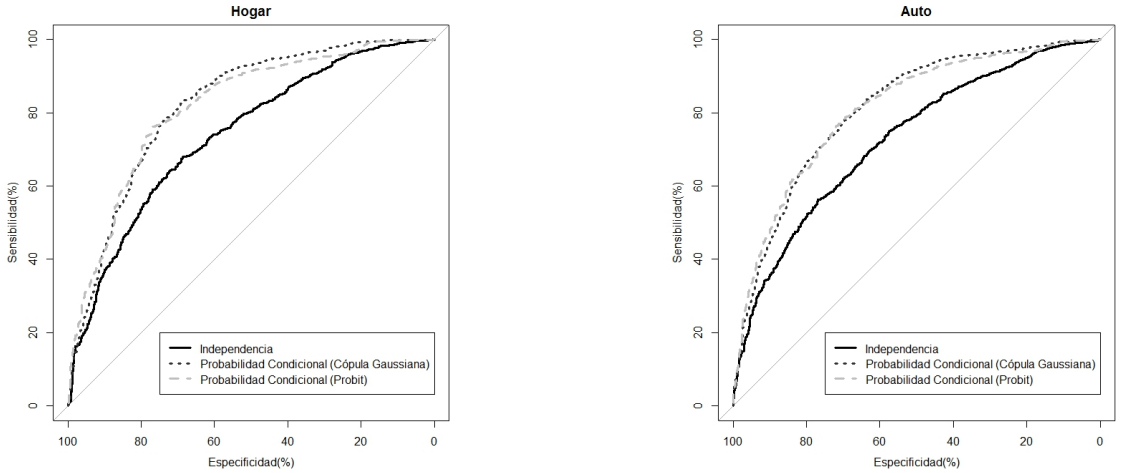


FIGURA 4.4: Áreas bajo las curva ROC - cópula Gaussiana.

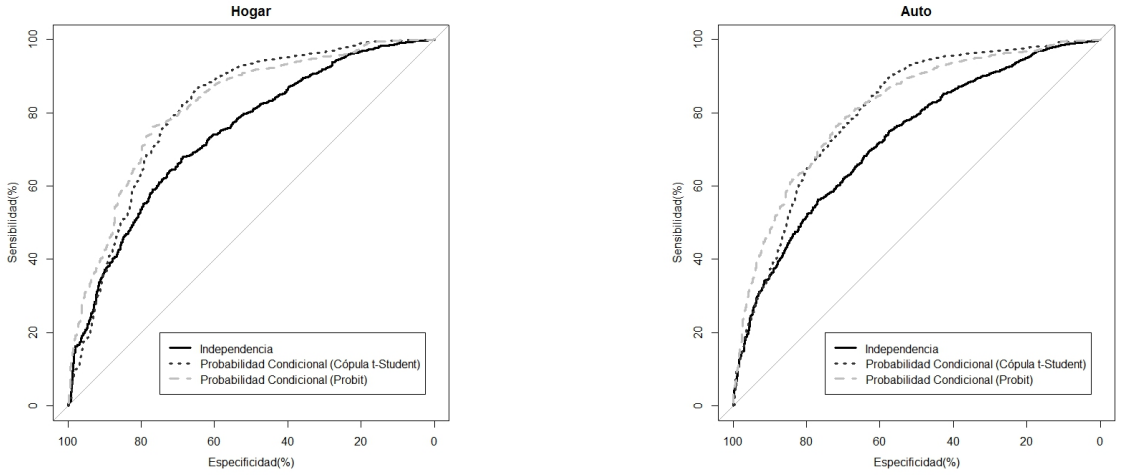


FIGURA 4.5: Áreas bajo las curvas ROC - cópula t-Student.

Por su parte, en la Tabla 4.6 se presentan las diferentes áreas bajo las curvas ROC obtenidas para cada distribución. Aquí, es evidente cómo la capacidad predictiva de los modelos mejora una vez que se incorpora la dependencia entre LoB, específicamente la clasificación mejora en unos 8 puntos básicos (por ejemplo, pasa de 73.78 % a 82.01 % en el caso de hogar).

TABLA 4.6: Áreas bajo las curvas ROC obtenidas a partir del modelo probit univariante (Independencia) y las probabilidades condicionales definidas a partir del modelo probit bivariante, la cópula Gaussiana y la cópula t-Student

		Probabilidades condicionales		
	Independencia	**Probit Bivariante**	**Cópula Gaussiana**	**Cópula t-Student**
Auto	72.64 %	81.00 %	81.11 %	80.18 %
Hogar	73.78 %	81.57 %	82.01 %	80.63 %

4.6. Conclusiones

El análisis realizado en la sección anterior, contempla la modelización de la deserción de clientes bajo el supuesto de independencia, en primer lugar, y a continuación bajo el supuesto de dependencia entre LoB considerando varios modelos bivariantes: probit bivariante, cópulas gaussiana y t-student. Tal como se evidenció, la inclusión de mayor información acerca del cliente y los riesgos que tiene contratados permite mejorar la predicción de las probabilidades de renovación de la póliza.

El efecto de las variables sobre la propensión a la renovación sugiere entre otras cosas, que la decisión del cliente se encuentra realmente influenciada por la experiencia que tiene con el resto de las pólizas contratadas. Así mismo, el análisis conjunto permitió descubrir efectos comunes que afectan la renovación de ambos tipos de pólizas y que hasta el momento eran desconocidos. Por otra parte, la correlación entre los errores de las variables latentes en el modelo probit bivariante es un indicador de la existencia de dependencia entre LoB. En este sentido, existen elementos no observables e indescriptibles que generan dependencia. El signo negativo del coeficiente ρ evidencia que una subida en la probabilidad de renovación de una de las pólizas podría tener un efecto contraproducente sobre la renovación de la otra póliza contratada.

Luego, las probabilidades condicionales estimadas a partir de los resultados del modelo probit son significativamente mejores en ambos casos (auto y hogar), por tanto, deja en evidencia la importancia que tiene el considerar la decisión que el cliente ha tomado a priori acerca de la renovación de una de sus pólizas al estimar la probabilidad de renovación de su otra póliza. De esta forma, el modelo probit bivariante representan una alternativa para las compañías de seguros interesadas en la modelización de las cancelaciones de pólizas de clientes con más de un tipo de riesgo asegurado.

El enfoque cópulas, es una alternativa diferente al modelo probit, que también tiene en cuenta la dependencia existente entre LoB. El análisis comparativo realizado entre la significación de los parámetros estimados con los modelos probit vs el enfoque cópulas nos permitió identificar parámetros significativos que aparentemente no lo eran. Así mismo, los resultados obtenidos tras el cálculo de las probabilidad condicionales representan una evidencia adicional a favor de la necesidad del análisis del cliente desde una perspectiva distinta a la actual.

Ambos enfoques son útiles para examinar segmentos de clientes desde una perspectiva global, es decir, teniendo en cuenta toda su información individual de forma simultánea y por ende la dependencia existente.

Este Capítulo resalta la importancia de trabajar con riesgos agrupados a nivel cliente en el sector seguros. Identificar y analizar posibles fuentes de dependencia entre grupos de riesgo debería convertirse en una práctica común en las compañías aseguradoras, debido a las ventajas que esto provee a nivel de negocio. Primero, la compañía sería capaz de analizar a sus clientes de forma global, más que como individuos que pertenecen a una línea de negocios concreta de una cartera de pólizas. Segundo, el análisis de interrelaciones hasta ahora desconocidas permitiría poner en práctica estrategias de negocio, tarificación e incentivos más eficaces para segmentos de clientes específicos.

Capítulo 5

Big-data Analytics en seguros

5.1. Introducción

En la actualidad, la sociedad se adapta a un entorno en el que todo su dinamismo queda registrado en cuestión de segundos. Vivimos en una era en la que la mayoría de las fuentes de información están digitalizadas. Redes sociales, páginas web, *smartphones*, dispositivos telemáticos, entre otros, son los responsables de nutrir grandes sistemas de información. El cambio es constante, así como lo es la demanda de información. Por un lado, los ciudadanos necesitan saber más, de ahí su urgencia por estar conectados y, por otro, las organizaciones necesitan obtener cierta ventaja de la información disponible, lo cual supone descubrir aspectos - hasta ahora desconocidos - vinculados con el comportamiento de sus clientes, socios, riesgos, costes y operaciones, así como de la sociedad en general.

La transformación ha sido vigorosa y, sin lugar a dudas, es sinónimo de reinvención. Más allá de las distancias, los individuos, las empresas, los países y los continentes en su totalidad han tenido que adaptarse a los nuevos desafíos en la manera de entender y analizar los datos. Las posibilidades son innumerables y las ventajas se encuentran en manos de aquellas compañías capaces de adaptarse rápidamente. La era del *big-data* llegó para quedarse y el sector asegurador, como era de esperar, se ha visto afectado por esta vorágine en primera persona. Por ejemplo, la posibilidad de conseguir más información cada vez más detallada de los asegurados plantea, entre otros, cambios en el tratamiento de la información, modelos de tarificación, estrategias de venta y canales de gestión.

La acumulación masiva de datos ha sido una práctica habitual en las compañías aseguradoras. Los ficheros contienen información individualizada y longitudinal para cada línea de negocio. El sistema de información es tan complejo, que actualmente los clientes pueden interactuar a través de distintos canales de mediación como: agentes y corredores, callcenters, redes sociales, internet y bancaseguros. En este sentido, el aprovechamiento de los datos suele segmentarse y adaptarse a las necesidades y requerimientos de cada departamento. Aunque en muchos casos sea frecuente el uso de variables comunes, como por ejemplo la edad, el sexo, el tipo de póliza, la prima pagada y los tipos de descuentos, la realidad es que cada departamento incorpora variables específicas vinculadas a sus objetivos de negocio, y sus responsables suelen tener autonomía sobre los análisis que realizan y trabajan bajo una cultura de independencia interdepartamental.

Hasta ahora, los análisis a nivel interno, en general, han estado orientados al reporte (*reporting*), control y seguimiento de indicadores. Son pocas las empresas que han sido capaces de establecer, en la práctica, verdaderas sinergias entre departamentos con el fin de dar respuestas rápidas a las necesidades del entorno, cada vez más dinámico, en el que nos encontramos y sobre el cual tenemos infinidad de información. En general, el tiempo ha sido considerado como un recurso limitado, donde apremia lo urgente y lo usual es procrastinar los cambios que son necesarios. Sin embargo, el reciente crecimiento digital ha desafiado la forma de entender el marco teórico de los seguros, haciendo necesario que las entidades aseguradoras se redefinan a todo nivel.

La forma de analizar los datos ha cambiado y con ello una parte de la estructura interna del negocio. Internamente, el primer cambio ha sido a nivel cultural, es decir, entender que existe una necesidad no cubierta y hasta hace algunos años inexplorada, para la cual se justifica que se destinen parte de los recursos de la compañía y, entre otras acciones, se invierta tiempo y dinero en el diseño y ejecución de un plan estratégico basado en una cultura de datos "*data driven*" (ver McAfee et al., 2012, para reflexiones adicionales sobre lo que implica la revolución empresarial en torno al *big-data*). El segundo cambio ha sido actuar en consecuencia, siendo uno de los primeros pasos la actualización de los sistemas informáticos; nueva tecnología, formación de los equipos de trabajo, actualización de procesos o familiarización con el uso de datos disponibles. Y, de forma casi simultánea, también se ha producido la integración de la "ciencia de los datos" en los nuevos modelos que estudian los seguros (véase Guillen, 2016).

Todo esto ha derivado en la creación de equipos de trabajo, con perfiles interdisciplinares y especializados, capaces de explotar la gran cantidad de datos disponibles y, a su vez, aportar valor añadido al resto de áreas de la compañía y por ende al negocio. Los

denominados equipos de *big-data analytics*, *analytics* o *advanced analytics* son cada vez más habituales dentro de las compañías de seguros. Su principal objetivo es la aplicación de técnicas analíticas avanzadas sobre grandes volúmenes de datos (ver Russom et al., 2011).

Nuestra principal contribución en este Capítulo es la de dar a conocer, desde un punto de vista empresarial y técnico, lo que debería ser un Departamento de *analytics* dentro de una compañía de seguros. Para ello, estableceremos algunas reflexiones sobre este tema y expondremos un caso de éxito basado en el análisis de la retención de clientes, donde se vincula la información de dos líneas de negocio de forma simultánea.

5.2. Visión empresarial del departamento de *Analytics*

En esta sección exponemos a grandes rasgos lo que consideramos deberían ser las responsabilidades de un Departamento de *analytics*, su posición dentro de la estructura organizativa de la empresa, las ventajas de su creación y los retos que supone contar con un departamento de tal envergadura.

5.2.1. Responsabilidades

Entre las principales responsabilidades de un Departamento de *analytics* se encuentran:

- Responder a las interrogantes de negocio planteadas a nivel directivo
- Ofrecer soluciones de negocio rápidas a iniciativas basadas en datos
- Identificar nuevas áreas de oportunidad
- Dar soporte al resto de departamentos de la compañía

5.2.2. Estructura organizacional

El Departamento de *analytics* funciona desde una perspectiva global del negocio, por ello debería depender directamente de la presidencia de la compañía, lo cual le proporcionaría autonomía sobre sus análisis y propuestas. La dependencia de otro departamento no está exenta de riesgos, pues su desarrollo, en general, quedaría supeditado a

los criterios e intereses de un área concreta de la compañía, aun cuando esta última se encuentre alineada con las directrices generales de la empresa.

Por otra parte, el Departamento de *analytics* necesita trabajar en paralelo con el departamento de tecnologías de la información (*Information Technology*) pues, en general, es este último el responsable de proporcionar las bases de datos necesarias para su posterior análisis, además de disponer de las fuentes y recursos para su procesamiento inicial.

5.2.3. Roles analíticos

La importancia del rol de las personas dentro de un programa de *analytics* es un aspecto que resalta [1]. Dicho autor, centra su énfasis en una gobernanza clara, el patrocinio necesario a nivel ejecutivo y el debido acceso la capacitación y los recursos necesarios. En este sentido, establece las principales funciones organizativas, lo cual, supone la creación de equipos de trabajo específicos e interconectados, capaces de desarrollar e implementar modelos analíticos operativos. Entre los equipos claves considerados se encuentran:

- Desarrollo analítico
- Arquitectura técnica
- Análisis de negocio
- Gestión del cambio
- Biblioteca de datos de *analytics*

Aunque no existe un consenso acerca de la estructura que debería tener un equipo de *analytics*, se espera que esté conformado por áreas funcionales diferenciadas según el alcance del trabajo que realizan y los roles de las personas que los integran.

5.2.4. Ventajas

Entre las principales ventajas de la creación de un Departamento de *analytics* se encuentran:

[1] Harrington, E. (2014, September 8). Building an analytics team for your organization part I. http://iianalytics.com/research/building-an-analytics-team-for-your-organization-part-i [Consultado: 2017-07-07]

- La rapidez con la que se podrían abordar transacciones de negocio específicas.
- Generación automática y disponibilidad inmediata de diferentes tipos de informes.
- Monitorización en tiempo real de los indicadores de negocio.
- Seguimiento efectivo de resultados, derivados de implementaciones operativas.
- Implementación de reglas para la detección de fraudes en tiempo real.
- Mejor segmentación de los riesgos.
- Análisis de dependencias entre tipos de productos, características de los clientes, etc.
- Detección anticipada de abandonos.
- Ofrecer pólizas personalizadas, adaptadas a las características de los consumidores.
- Tarificación de productos con mayor precisión (ver Swedloff, 2014, para más detalles sobre sus implicaciones) .

5.2.5. Fortalezas y debilidades

La incorporación de un Departamento de *analytics*, favorece la adopción de nuevas estrategias de negocio basadas en datos. Igualmente, ayuda con el establecimiento de sinergias entre los distintos departamentos y facilita el intercambio de información entre equipos de trabajo. Además, beneficia la capacidad de reconocer el tipo de datos que le es útil a cada departamento, sus *outputs* complementan los análisis existentes y sirven como puente para la transición desde las pólizas clásicas a las pólizas telemáticas, sobre todo en los seguros de automóvil.

En contraposición, Russom et al. (2011) menciona que los perfiles técnicos y habilidades inadecuadas, representan las principales barreras para el análisis de grandes volúmenes de datos. Así mismo, en caso de que el equipo de *analytics* no sea adecuadamente respaldado termina relegado y, por lo tanto, sus proyectos son vistos y acogidos por sus compañeros como acciones aisladas e injustificadas, en vez de relevantes para la compañía. Por otra parte, las posibles restricciones de acceso a los datos, junto a los retrasos que pudiesen generarse en la entrega de los mismos, debido a la dependencia de un proveedor de los datos (con frecuencia IT- *Information Technology*), le restan dinamismo a los proyectos y posteriores análisis del Departamento de *analytics*.

5.3. *Analytics* en seguros de no vida

A continuación, presentaremos un ejemplo de la aplicación de *big-data* en seguros. El objetivo de este ejemplo es el estudio de la retención de clientes con pólizas contratadas simultáneamente en dos líneas de negocio distintas: seguro de autos y seguro de hogar. Para ello, se evalúa el ajuste de cuatro modelos predictivos en base a dos criterios diferentes. Este ejemplo es complementario al análisis realizado en Bolancé et al. (2016a), donde los clientes exclusivamente habían contratado un seguro de auto o un seguro de hogar.

El software utilizado para el diseño de los modelos ha sido R y para el tratamiento de las bases de datos R y SAS.

5.3.1. Datos

En esta parte se analiza la misma muestra aleatoria de 22106 clientes descrita en 4.4, formada por clientes que habían decidido asegurar dos riesgos diferentes, concretamente auto y hogar, con la misma compañía. A efectos de este estudio sólo consideraremos clientes con una póliza de auto y una póliza de hogar, estos se corresponden con el 75.89 % de los clientes que tienen contratadas pólizas en ambas líneas; el resto son clientes con más de una póliza contratada en una o en las dos líneas de negocio analizadas.

En las Tablas 5.1, 5.2 y 5.3 se describe la información utilizada en el ajuste y/o entrenamiento de los distintos modelos. La variable dependiente puede referirse a tres tipos de decisión por parte del cliente:

1. Si el cliente decide renovar o no su póliza de hogar, al margen de lo que haga con la póliza de autos.
2. Si el cliente decide renovar o no su póliza de autos, al margen de lo que haga con la póliza de hogar
3. Si el cliente decide renovar tanto su póliza de autos como su póliza de hogar o, por el contrario, no renovar una o ninguna

Entre las variables explicativas de la decisión de renovar existen tres grupos, las que se utilizan tanto para explicar la renovación de la póliza de hogar como de autos (Tabla

5.1), las que se utilizan para modelizar la renovación o no de la póliza de autos (Tabla 5.2) y las que se utilizan para modelizar la renovación o no de la póliza de hogar (Tabla 5.3). Para explicar la renovación conjunta de ambas pólizas se utilizan todas las variables descritas en las Tablas 5.1, 5.2 y 5.3.

TABLA 5.1: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de auto y hogar

Relacionado con	**Variable (Descripción)**
Tomador	Sexo, edad, pólizas de otros ramos en vigor, suma de las primas pagadas
Póliza	Antigüedad de la póliza, prima anterior, variación de prima, descuentos - malus, recargos, forma de pago (A: Anual, S: Semestral, T: Trimestral), Tipo de mediador (A: Agente exclusivo, C= Corredor de seguros), suplementos, recargos, Ratio de anulaciones del mediador, Estado de la póliza (A= anulada, V= vigente),

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de auto y hogar, 2015.

TABLA 5.2: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de auto

Relacionado con	**Variable (Descripción)**
Objeto de riesgo	Tipo de vehículo, primer conductor - edad, segundo conductor (Si , No), carga siniestral, potencia, relación peso potencia, nº de plazas

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de auto, 2015.

TABLA 5.3: Algunas variables comunes utilizadas en el caso de estudio sobre la modelización de la deserción en seguros de hogar

Relacionado con	**Variable (Descripción)**
Objeto de riesgo	Tipo de vivienda (A, B, C, D, E, F), continente, contenido

Fuente: Conjunto de datos propios para estudiar la pérdida de clientes en los seguros de hogar, 2015.

5.3.2. Modelos

Los modelos de clasificación utilizados en este estudio son los mismos que describimos en la Sección 3.2, es decir,

- Regresión logística (*Logistic regression*)
- Árboles de decisión condicionales (*CTREE-Conditional Tree*)
- Maquina vector soporte (*SVM-Support Vector Machine*)
- Redes neuronales (*NN-Neuronal Network*)

Es importante mencionar que el modelo de regresión logística, los árboles condicionales, las máquinas de vectores de soporte y las redes neuronales son utilizados en el contexto del aprendizaje supervisado. Posteriormente, el resultado numérico obtenido en cada caso es la probabilidad de renovación de cada cliente basado en sus características personales, las particularidades de la póliza y las del objeto de riesgo. Es importante resaltar que además de modelizar la probabilidad de renovación de las pólizas de hogar y auto de forma independiente, también hemos querido presentar un modelo en el que se incluyen todas las variables tanto de hogar como de auto, cuya variable respuesta nos da información acerca de la renovación simultánea de ambas pólizas, es decir, obtendremos la propensión (probabilidad) de cada cliente a la renovación de sus dos pólizas durante el mismo periodo.

5.3.3. Medidas de para evaluar la capacidad predictiva de los modelos

Los métodos utilizados estiman la probabilidad de clase p_i, es decir, la probabilidad de que la i-ésima póliza sea retenida, lo cual permite definir las clases predichas al comparar p_i con diferentes puntos de corte de clasificación $t \in [0,1]$. Cada una de estas comparaciones produce una matriz de confusión, a través de la cual es posible determinar la proporción de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN) para cada modelo.

Los criterios utilizados para evaluar el desempeño de los modelos predictivos corresponden con dos de los presentados en la Sub-sección 3.2.5, concretamente,

$\text{C1} = \max(\text{ sensibilidad} + \text{especificidad}) = \max(\frac{VP}{VP+FN} + \frac{VN}{VN+FP})$
$\text{C2} = \text{AUC}$

5.3.4. Resultados

A continuación, en las Figuras 5.1 y 5.2 se muestran los resultados de las curvas ROC asociadas a cada uno de los modelos propuestos y para cada una de las líneas de negocio -hogar y auto- analizadas. La Figura 5.1 se corresponde con las curvas ROC de los modelos para el ajuste de la retención en el seguro de hogar. En dicha figura se observan diversos cruces entre curvas y en general un mejor ajuste para los árboles condicionales y la máquina de vectores de soporte. En la Figura 5.2, asociada a la retención en el seguro de autos, las curvas ROC de los distintos modelos presentan un comportamiento más homogéneo, excepto al inicio de las curvas ROC donde se aprecia que los resultados del árbol condicional son un poco mejores.

FIGURA 5.1: Curvas ROC para cada LoB y cada método - Hogar

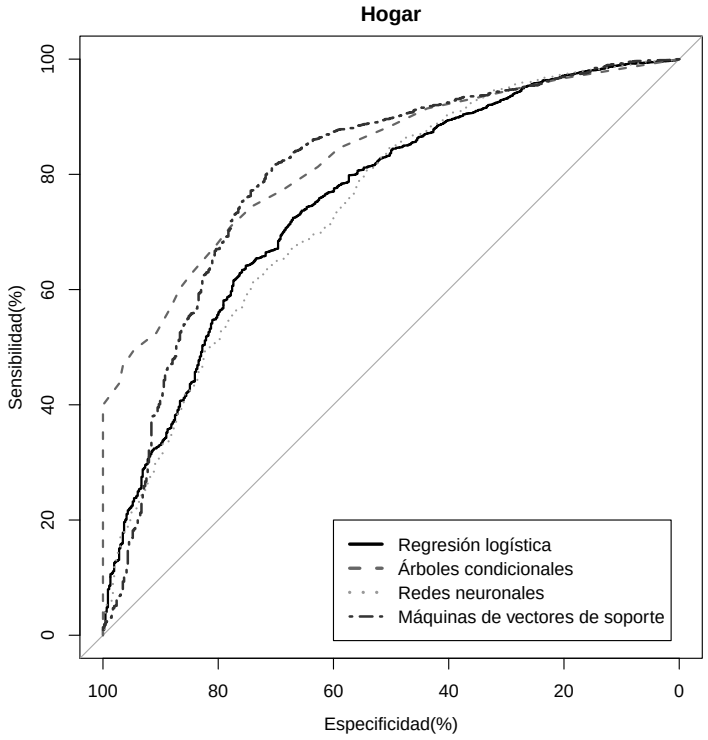
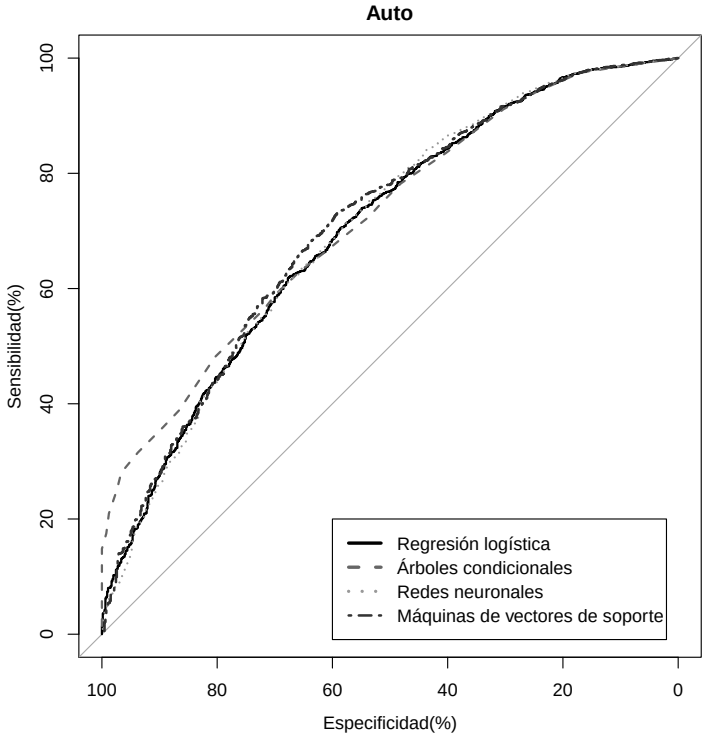


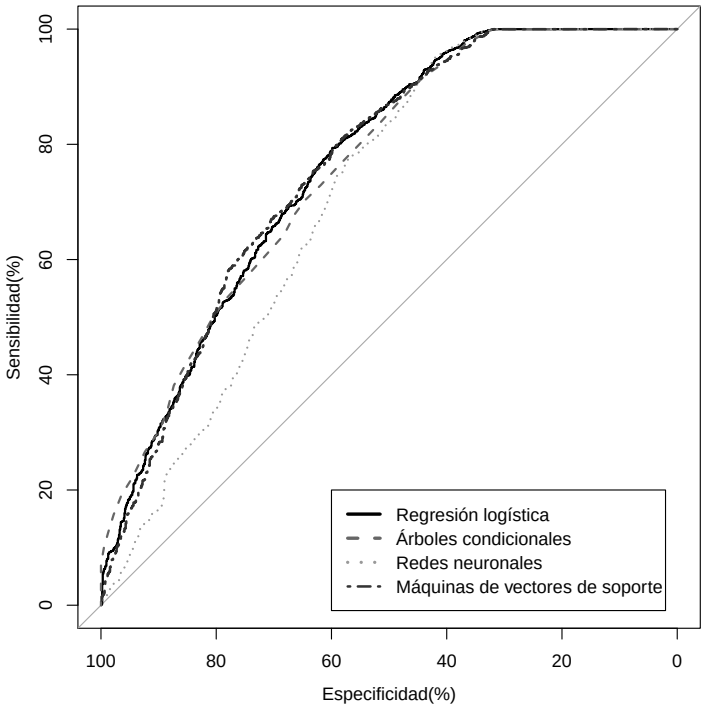
FIGURA 5.2: Curvas ROC para cada LoB y cada método - Auto



El hecho de que las curvas se crucen implica que no hay un modelo que domine o mejore completamente al resto. Es decir, en función de la especificidad deseada el modelo óptimo podría variar.

Hasta aquí la propensión de los clientes a renovar es vista desde la perspectiva de hogar y desde la perspectiva de auto, por separado. Sin embargo, la Figura 5.3 muestra los resultados de la propensión a la renovación de ambas pólizas (hogar y auto) de los mismos clientes en función de todas sus características, es decir, los modelos han sido definidos de manera que se tuviesen en cuenta todas las variables de hogar y de auto de forma simultánea. En esta figura podemos observar múltiples intersecciones, y aunque no existe una curva que domine sobre el resto, los resultados de la regresión logística, CTREE y SVM son mejores que los obtenidos con la red neuronal. Por cada intersección tendremos diferentes niveles de falsas alarmas donde un clasificador supera a las otras.

FIGURA 5.3: Curvas ROC para cada LoB y cada método - Hogar y Auto



En las Tablas 5.4, 5.5 y 5.6 se muestran los resultados de C1 y C2 utilizando el conjunto de datos de prueba. En general, el valor de los criterios varía según la línea de negocio. En el caso de hogar el árbol condicional es el que tiene un mejor ajuste, mientras que en auto el SVM y CTREE son los mejores. Cuando la información de ambas líneas de negocio es fusionada en un solo modelo los resultados de los modelos que pudieron ajustarse son muy parecidos, excepto para la red neuronal. Además, de forma similar a lo ocurrido en los resultados del Capítulo 3, en la Tabla 5.5, es posible observar un intervalo que indica que existe más de un umbral óptimo, es decir, en los valores 0.89 y 0.90 el modelo alcanza el mismo criterio de optimalidad.

TABLA 5.4: Criterios de desempeño para cada modelo en los datos de prueba de Hogar.

		Modelos			
		Regresión logística	**Árbol condicional**	**Máquina de vectores de soporte**	**Redes Neuronales**
Criterios	C1	1.39	1.49	1.21	1.35
	C2(%)	75	83	80	74
Umbral Óptimo	C1	0.94	0.92	0.97	0.91

Fuente: Cálculos propios.

TABLA 5.5: Criterios de desempeño para cada modelo en los datos de prueba de Auto.

		Modelos			
		Regresión logística	**Árbol condicional**	**Máquina de vectores de soporte**	**Redes Neuronales**
Criterios	C1	1.29	1.29	1.32	1.29
	C2(%)	70	72	71	70
Umbral Óptimo	C1	0.89	[0.89-0.90]	0.97	0.89

Fuente: Cálculos propios.

TABLA 5.6: Criterios de desempeño para cada modelo en los datos de prueba de Hogar y Auto.

		Modelos			
		Regresión logística	**Árbol condicional**	**Máquina de vectores de soporte**	**Redes Neuronales**
Criterios	C1	1.39	1.36	1.38	1.37
	C2(%)	76	76	76	71
Umbral Óptimo	C1	0.87	0.78	0.96	0.83

Fuente: Cálculos propios.

Teniendo en cuenta el criterio C1 y considerando el modelo que mejor ajusta en cada caso, hemos construido las matrices de confusión utilizando el "Umbral óptimo" que

aparece en la última fila de las Tablas 5.4, 5.5 y 5.6. Dicho " Umbral óptimo" equivale al valor de la probabilidad a partir del cual se predice que el individuo renovará su póliza. El objetivo es el de visualizar desde otra perspectiva el poder predictivo de cada modelo. A partir de los resultados de la Tabla 5.7 se observa que el porcentaje de aciertos sobre la renovación o no del seguro de hogar es del 71.45 %. Por su parte, en la Tabla 5.8 se observa un porcentaje de acierto en la renovación de los seguros de auto es del 66.12 %. Mientras que para el caso conjunto (hogar y auto) Tabla 5.9 obtenemos un 75.66 % de casos bien clasificados.

TABLA 5.7: Matriz de confusión del modelo óptimo (CTREE) en el seguro de hogar. Umbral óptimo c=0.92.

		Predicho	
		No renueva	Renueva
Real	No renueva	361	103
	Renueva	1789	4375

Fuente: Diseño propio.

TABLA 5.8: Matriz de confusión del modelo óptimo (SVM) en el seguro de autos. Umbral óptimo c=0.97.

		Predicho	
		No renueva	Renueva
Real	No renueva	431	226
	Renueva	2019	3952

Fuente: Diseño propio.

TABLA 5.9: Matriz de confusión del modelo óptimo (regresión logística) en ambos ramos simultáneamente. Umbral óptimo c=0.87.

		Predicho	
		No renueva	Renueva
Real	No renueva	577	380
	Renueva	1233	4438

Fuente: Diseño propio.

Con el objetivo de analizar si la capacidad predictiva de los modelos es homogénea en función de grupos de asegurados, hemos calculado los porcentajes de aciertos para hombres y mujeres. Los resultados no muestran diferencias entre dichos porcentajes en los modelos para el seguro de autos y en los modelos para ambos ramos. Por el contrario, el porcentaje de aciertos es mayor para los hombres en los modelos para el seguro de hogar, en la práctica, este último resultado implica que la compañía podría considerar que las mujeres se comportan de un modo distinto al esperado en un mayor número de casos.

Esta aplicación, es un ejemplo claro de cómo a nivel empresarial es posible sacar partido de la información disponible y a los *outputs* generados por el Departamento de *analytics*. Identificar la propensión simultánea de un cliente a la renovación o no de sus pólizas es un indicador que aporta valor a la compañía. Por una parte el Departamento de cliente, sería capaz de tomar acciones preventivas ante posibles anulaciones o por el contrario cuidar de aquellos clientes buenos que tienen más de una póliza contratada con la compañía. Luego, a nivel de tarificación sería posible la calibración de los modelos predictivos en función de una o más variables que sirvan como indicadores de la propensión a la renovación de los clientes para cada cartera. Además las acciones del Departamento de tarificación y el Departamento de cliente estarían alineadas en el sentido de que ambos tendrían una misma estrategia para el grupo de riesgo identificado. Por su parte, a nivel directivo, las decisiones estratégicas de negocio tendrían en cuenta un colectivo específico de clientes a los que se les debe tratar de forma diferenciada y por ende el negocio en sí mismo sería capaz de reorientarse de forma más asertiva.

5.4. Conclusiones

La transformación requiere de un cambio de paradigma. El seguimiento de indicadores clásico, el informe (*reporting*) y la toma final de decisiones basada en la intuición no son suficientes. Cada vez es más popular la frase “los datos hablan por sí solos”, y el éxito está en manos de las compañías que decidan evolucionar y “escucharlos”.

Todos los empleados de la compañía deben entender los motivos y los beneficios de la creación de un Departamento de *analytics*. La aceptación favorece la colaboración y con ello resulta más fácil establecer sinergias entre el resto de departamentos y el equipo de *analytics*. El fin último es aportar valor, así que cuanto mayor sea la disposición a compartir los detalles operativos de cada área, mayor será el *input* que recibirá el Departamento de *analytics* y mayores los beneficios globales tanto para la entidad como para los asegurados.

Los cambios que supone la creación de un equipo de *Analytics* dentro de una compañía de seguros vs. la rapidez con la que se espera tener resultados, se encuentran desfasados temporalmente, es decir, resulta casi imposible tener un Departamento de *analytics* si no se dispone de la tecnología, los recursos humanos y la experiencia necesaria. De ahí que sea necesario que los equipos de trabajo dispongan del perfil técnico específico y los líderes se encuentren familiarizados con el sector asegurador, además de tener conocimientos sólidos en *big-data analytics*. En este sentido, una alternativa inicial es la de contar con asesores expertos externos, para, en un primer momento, liderar los proyectos, capacitar a los equipos internos y realizar la transferencia del conocimiento necesario.

Por otra parte, y como comentamos al inicio, cada departamento suele disponer de ficheros de datos específicos para las actividades de análisis e informes usuales. En este sentido, urge la democratización de la información. La capacidad de las compañías de centralizar el acceso a los datos en un único lugar, de modo que cualquier departamento tenga acceso transversal a los datos de otros departamentos, supone un verdadero reto, donde el primer beneficiado sería el Departamento de *Analytics*.

Cuando una compañía disponga de más de un equipo de *Analytics*, por ejemplo, porque cuenta con diferentes sedes en diversas partes del mundo, la relación debe ser absolutamente cercana. Aunque resulte evidente, la realidad es que no lo es. En este sentido, han de compartir información, reciclar ideas e intercambiar casos de éxito y fracaso. Se trata de poder trabajar alineados, re-aprovechar la experiencia existente, facilitar la estandarización de procesos y la creación de modelos corporativos fácilmente reproducibles.

Por último, el caso de estudio presentado resume algunos de los beneficios que le aportaría al negocio implementar *Big Data Analytics* dentro de sus procesos. Por una parte, sería posible implementar métodos predictivos alternativos a los convencionales con el objetivo de complementar y mejorar la modelización realizada con los métodos clásicos. Así mismo, este tipo de análisis serviría como punto de partida para el análisis de las dependencias entre diferentes grupos de riesgo, teniendo en cuenta información de más de un departamento de forma simultánea. Tal y como se ha evidenciado, la combinación de información del cliente a nivel global permite mejorar la capacidad predictivas de los modelos de retención. Por lo tanto, teniendo en cuenta que podrían incorporarse más líneas de negocio al análisis y con ello más información, es muy recomendable que los modelos utilizados dentro de las aseguradoras puedan escalar a otro nivel de uso más general que el de los informes específicos para cada ramo. Establecer en las entidades un equipo exclusivamente dedicado y capacitado para llevar a cabo análisis

de tal envergadura, sin lugar a dudas, marcaría la diferencia entre el pasado y el futuro del sector asegurador.

Parte II

Ejemplos de metodologías alternativas para la cuantificación del riesgo univariante y multivariante

Capítulo 6

Impacto de la estructura D-vine en la estimación del riesgo

6.1. Introducción

El Valor en Riesgo (*Value-at-Risk* o VaR) es la medida más popular propuesta por el regulador de la industria financiera y aseguradora para la cuantificación del riesgo. La estimación del VaR ha planteado y seguirá planteando diferentes desafíos en el contexto del análisis del riesgo. Cuando las pérdidas han sido generadas por un conjunto de factores de riesgo dependientes, debemos tener en cuenta esta dependencia para estimar el VaR. Además, incluimos el Valor en Riesgo Condicional (*Conditional Value at Risk* o *CVaR*) en nuestro análisis.

Tenemos dos estrategias para incluir la dependencia. Podemos usar ya sea una distribución multivariante para ajustar el comportamiento marginal de los factores de riesgo y su dependencia conjuntamente o, alternativamente, podemos usar distribuciones univariantes para modelizar cada factor de riesgo y utilizar una cópula para modelizar la dependencia. Sin embargo, cuando se usan cópulas existen algunas restricciones cuando el número de dimensiones o de factores de riesgo es superior a dos, es decir, excepto para la cópula Gaussiana, la generalización de la cópula bivariante a la cópula multivariante asume ciertas restricciones acerca de la dependencia entre pares de factores de riesgo. Por ejemplo, por un lado, al usar cópulas elípticas como la t de Student necesitamos suponer que los grados de libertad son los mismos para todos los pares. Por otro lado, para estimar la cópula arquimediana se supone que el valor del parámetro de la cópula es el mismo para todos los pares (para una introducción a las cópulas ver, por ejemplo, Nelsen, 2006).

La descomposición pair-copula de una distribución multivariante es una forma flexible para modelar relaciones de dependencia entre pares de variables considerando diferentes intensidades de dependencia o incluso estructuras de dependencia. Sin embargo, tal flexibilidad podría conducir a ciertas ineficiencias y sesgos en la estimación del riesgo. En respuesta a ello, en este Capítulo analizamos el efecto de la selección de pares, mejor conocidos como "*vines*" (ver, por ejemplo, Bedford y Cooke, 2002), sobre la estimación del riesgo. En particular, realizamos un estudio sobre la selección del *Drawable vine (D-vine)*, aunque un análisis similar puede llevarse a cabo con el *Canonical vine (C-vine)* o a partir de la representación más general, el *Regular vine (R-vine)*.

La literatura sobre *vine copula* es cada vez más popular. De hecho, existe una amplia variedad de aplicaciones financieras, no solo relacionadas con la estimación del VaR y CVaR, sino también en otros contextos como, por ejemplo, optimización de carteras y estrategias de negociación (ver, por ejemplo, Weiß y Supper, 2013; Low et al., 2016; Brechmann y Czado, 2013; Rad et al., 2016; Nikoloulopoulos et al., 2012; Low et al., 2013). En este aspecto, realizar análisis estadísticos tales como los presentados en este trabajo son fundamentales.

Nuestra contribución está compuesta por tres resultados principales. Primero, mostramos cómo diferentes descomposiciones *pair-copula* proveen diferentes valores para el VaR y CVaR y analizamos las magnitudes de estas diferencias. El segundo resultado está relacionado con la selección de la descomposición óptima, donde se analizan diferentes criterios en función de si se asume que la cópula es conocida o no. Finalmente, definimos un algoritmo específico para la selección del *D-vine* óptimo.

Tal y como indicamos en el párrafo anterior, nuestro objetivo estadístico es analizar los efectos del uso de diferentes *D-vines* para estimar el VaR y el CVaR. Relacionado con esto, la mayor o menor dependencia entre las variables sería una característica fundamental. Para ejemplificar nuestros resultados, modelizamos las pérdidas obtenidas a partir de los rendimientos filtrados vinculados a cuatro carteras de acciones, cuyas diferencias yacen en su grado de diversificación, aunque el mismo análisis se podría llevar a cabo con datos de otra naturaleza. En la discusión de los resultados al final de este capítulo se comenta la posible utilización de esta metodología en el análisis de la cartera de seguros de hogar y auto. Diseñamos las cuatro carteras para que una mayor diversificación implique una menor dependencia y una menor diversificación, una mayor dependencia entre los rendimientos filtrados. Además, para tener un número suficiente de *D-vines* diferentes, las carteras están compuestas por seis rendimientos de acciones que cotizan en dólares. Para completar el análisis, y comparar los resultados con distintas dimensiones, también incluimos un estudio de simulación que nos permite

comparar las propiedades estadísticas de los diferentes criterios en torno a la selección de la descomposición óptima.

En este contexto, utilizamos el *D-vine* para la descomposición *pair-copula*, debido a, que no tenemos información que justifique ninguna relación jerárquica entre los activos. Nuestro estudio es similar los realizados por Aas et al. (2009) y Min y Czado (2010), respectivamente, los cuales analizan la estructura *D-vine* aplicada a los rendimientos financieros filtrados.

Cuando utilizamos el *D-vine*, la descomposición *pair-copula* está relacionada con la elección del orden más adecuado de los factores de riesgo al inicio del proceso. En este sentido, es posible utilizar diversos criterios para la ordenación de las variables. Por ejemplo, en el caso de la cópula t'Student, Aas et al. (2009) estiman los grados de libertad para cada par de variables y seleccionan el orden de las mismas empezando por los pares con menores grados de libertad y terminando con aquellos que tienen mayores grados de libertad; es decir, de mayor a menor dependencia en la cola inferior y superior. Dissmann et al. (2013) basándose en la Tau de Kendall, usan un método para la selección del *R-vine* que denominan algoritmo del *"árbol de máxima expansión"*; típicamente, este algoritmo se describe como árbol de expansión mínima (algoritmo de Prim). Righi y Ceretta (2013, 2015) también usan la matriz de dependencia de la Tau de Kendall para determinar el orden de los factores de riesgo en el diseño del *D-vine*. Un aspecto común en la mayoría de las aplicaciones es la flexibilidad con la que se puede elegir este orden, debido a la ausencia de una regla de decisión específica; por esta razón, es necesario analizar el efecto de la selección del orden de las variables en el vector multivariante para definir el *D-vine*, con el objetivo de estimar el riesgo de pérdida.

Así, para las carteras consideradas realizamos dos tipos de análisis. Primero, estimamos el VaR para todos los ordenes posibles y analizamos su dispersión. Segundo, determinamos cuál es el VaR y el CVaR obtenido usando diferentes criterios para la selección del orden en el *D-vine*. En general, para una cópula dada, la Tau de Kendall, la Rho de Spearman, y si existe, la dependencia en la cola pueden utilizarse para definir el orden de las variables en el *D-vine*. Estos criterios son evaluados suponiendo que la cópula es conocida o desconocida. Para la estimación de cada VaR y CVaR utilizamos el método de Monte Carlo (ver McNeil et al., 2015, Capítulo 2).

En la actualidad, el uso de la descomposición *pair-copula* es un tópico relevante que está ganando cada vez más seguidores en diferentes líneas de investigación, especialmente en la industria financiera. Recientemente, Weiß y Scheffer (2015) propusieron el uso de diferentes cópulas para los diferentes pares de una descomposición *pair-copula*, con

el fin de pronosticar el VaR de carteras financieras. Por su parte, Min y Czado (2014) analizaron las propiedades del estimador de máxima verosimilitud de los parámetros de cópula asociados con cada par, utilizando pseudo-datos. Este estimador se denomina estimador de pseudo-máxima verosimilitud o, como dicen los autores, estimador semi-paramétrico de máxima verosimilitud, dado que la distribución empírica multiplicada por $T/(T+1)$ (donde T es el tamaño de la muestra) se utiliza para las marginales. En dicho trabajo, los autores muestran una aplicación sobre los logaritmos de los rendimientos diarios de los tipos de cambio (ver Bolancé et al., 2014, para comparar estimadores semi-paramétricos de máxima verosimilitud usando diferentes métodos no paramétricos para distribuciones marginales).

Para estimar la distribución multivariante utilizando la descomposición *pair-copula*, se deben tener observaciones independientes e idénticamente distribuidas. Por tanto, de manera similar a los artículos citados en los párrafos anteriores, utilizamos los modelos ARMA(P,Q)-GARCH(p,q) para filtrar nuestros datos, es decir, tenemos en cuenta los residuos de los modelos ajustados para estimar los parámetros de cópula bajo el enfoque de la descomposición *pair-copula*.

6.2. Cuantificación del riesgo de la cartera

Bedford y Cooke (2002) definen el concepto de *vine* and *R-vine*. Específicamente, podemos ver que un *R-vine* es una secuencia de árboles que representan la estructura de dependencia de un vector multivariante de variables aleatorias continuas, dada una factorización de la función de densidad multivariante. El *D-vine* es un caso particular del *R-vine*. En el Apéndice C del final de esta Tesis, se describe el caso particular del *D-vine* con seis variables, el cual analizaremos en la parte empírica de este trabajo.

En la práctica, el orden en el nodo T_1 del *D-vine* (ver el desarollo en el Apéndice C) afecta a los resultados de la estimación y, por lo tanto a los riesgos estimados (ver Min y Czado, 2010, para el desarollo). Por ello, analizamos cómo usando órdenes distintos se obtienen diferentes estimaciones del VaR y CVaR.

Para el análisis de la dispersión de los VaRs y CVaRs estimados, es decir, la precisión de los riesgos estimados, específicamente para el caso de seis dimensiones, obtenemos $\frac{6!}{2} = 360$ órdenes posibles y sus VaRs y CVaRs estimados usando cada *D-vine* resultante. El procedimiento, basado en una simulación de Monte Carlo, se describe posteriormente.

Para cuantificar el riesgo de una cartera necesitamos calcular una medida de los rendimientos para cada activo. Por tanto, el interés se centra en utilizar una medida que

nos permita reflejar los cambios relativos en los precios, por ello para cada acción j, definimos la variable *log-return* en el momento t, es decir, $R_{jt} = log(\frac{P_{jt}}{P_{jt-1}})$, $t = 1,...,T$, donde T es el número total de puntos temporales observados y donde P_{jt} es el precio del activo j en el momento t. Estamos interesados en modelizar la dependencia entre los cambios aleatorios de los rendimientos a lo largo del tiempo. Por esta razón, antes de empezar nuestro análisis utilizando *pair-copula* hemos filtrado los rendimientos de las acciones de la siguiente manera. Asumimos que los rendimientos R_{jt} están generados por un modelo de serie temporal ARMA(P,Q)-GARCH(p,q) que puede expresarse como:

$$\begin{aligned} R_{jt} &= \mu_j + \sum_{i=1}^{P} \phi_{ji} R_{jt-i} - \sum_{i=1}^{Q} \psi_{ji} e_{jt-i} + e_{jt}, \\ e_{jt} &= \sigma_{jt} x_{jt}, \\ \sigma_{jt}^2 &= \alpha_{j0} + \sum_{i=1}^{p} \alpha_{ji} e_{jt-i}^2 + \sum_{i=1}^{q} \beta_{ji} \sigma_{jt-i}^2. \end{aligned} \tag{6.2.1}$$

Los logaritmos de los rendimientos filtrados son x_{jt}, $t = 1,...,T$ y podemos suponer que estos son T valores independientes e idénticamente distribuidos procedentes de la variable aleatoria X_j.

Supongamos una cartera formada por 6 activos y sea $(X_{(1)},\ldots,X_{(6)})$ un vector multivariante de seis variables aleatorias continuas que representan el logaritmo de los rendimientos filtrados, donde los paréntesis indican el orden dado, sus distribuciones marginales son $F_{(1)},\ldots,F_{(6)}$ y la función de distribución multivariante es F. En nuestro caso, estamos interesados en estimar el VaR y el CVaR, con el nivel de confianza α, de la variable aleatoria $L = -V_0 \cdot (w_1 \cdot X_1 + \ldots + w_6 \cdot X_6)$ (ver McNeil et al., 2015, Capítulo 2), donde L es la variable de pérdida linealizada asociada a una cartera de acciones, w_j es el peso de la acción j en la cartera y V_0 es el valor de inversión inicial que podemos suponer igual a 1.

El *VaR* con nivel de confianza α para una variable aleatoria de pérdida continua L, puede ser definido como:

$$VaR_\alpha(L) = \text{ínf}\{l, F_L(l) \geq \alpha\} = F_L^{-1}(\alpha), \tag{6.2.2}$$

donde L tiene función de densidad de probabilidad f_L y función de distribución acumulada F_L. Dado un VaR, el CVaR es (ver Denuit et al., 2005):

$$CVaR_\alpha(L) = E\left(L - VaR_\alpha(L) \,|L > VaR_\alpha(L)\right). \tag{6.2.3}$$

Asumiendo que $(X_{(1)t},...,X_{(6)t})$, $t = 1,...,T$, denota un muestra hexadimensional de T observaciones independientes e idénticamente distribuidas provenientes de un vector aleatorio $(X_{(1)},\ldots,X_{(6)})$, para cada posible orden $o = 1,...,\frac{6!}{2} = 360$, utilizamos el método de Monte Carlo para estimar el $VaR^o_\alpha(L)$ y el $CVaR^o_\alpha(L)$. A continuación, se describen los pasos a seguir:

PASO 1: Obtener los pseudo-datos:

$$U_{jt} = \frac{T}{T+1}\hat{F}_{j_T}(X_{jt}),\ j = 1,....,6, \tag{6.2.4}$$

donde $\hat{F}_{j_T}(x) = \frac{1}{T}\sum_{t=1}^{T} I(X_{jt} \leq x)$ es la función de distribución empírica y $I(\bullet)$ es la función indicatriz que toma valor 1 si la condición entre paréntesis es verdadera y 0 en caso contrario.

PASO 2: Utilizar los pseudo-datos $(U_{(1)t},...,U_{(6)t})$, $t = 1,\ldots,T$, para calcular el logaritmo de la pseudo-verosimilitud asociado a la cópula multivariante, con el fin de estimar los parámetros de dependencia que maximizan la versosimilitud parcial. El logaritmo de la pseudo-verosimilitud (ver Genest et al., 1995; Min y Czado, 2014, para una revisión de los procedimientos de estimación basados en pseudo-datos) es:

$$ln\,(L(\theta)) = \sum_{t=1}^{T} ln\{c_\theta(U_{(1)t},...,U_{(6)t})\}, \tag{6.2.5}$$

donde c_θ es la densidad *pair-copula* y θ es un vector con $6(6-1)/2 = 15$ parámetros de cópula que dependen del orden seleccionado para los factores de riesgo, es decir

$$\begin{aligned}\theta = (&\theta_{(1)(2)},\theta_{(2)(3)},\theta_{(3)(4)},\theta_{(4)(5)},\theta_{(5)(6)},\\ &\theta_{(1)(3)|(2)},\theta_{(2)(4)|(3)},\theta_{(3)(5)|(4)},\theta_{(4)(6)|(5)},\\ &\theta_{(1)(4)|(2)(3)},\theta_{(2)(5)|(3)(4)},\theta_{(3)(6)|(4)(5)},\\ &\theta_{(1)(5)|(2)(3)(4)},\theta_{(2)(6)|(3)(4)(5)},\\ &\theta_{(1)(6)|(2)(3)(4)(5)}).\end{aligned} \tag{6.2.6}$$

Para evaluar la bondad del ajuste de la descomposición *pair-copula* estimada, dada la cópula bivariante utilizada para todos los pares, utilizamos los estadísticos basados en el método PIT (*Probability Integral Transform*), propuesto y descrito en detalle en Aas et al. (2009).

PASO 3: Simulando los vectores $(\tilde{U}_{(1)s},...,\tilde{U}_{(6)s})$, $s=1,\ldots,S$, a partir de la cópula estimada, donde S es el número de vectores hexadimensionales simulados, utilizamos el paquete CDVine de R. El método de simulación implementado en este paquete de R es descrito en Brechmann et al. (2013).

PASO 4: Los factores de riesgo simulados se calculan como $\tilde{X}_{(1)s} = \hat{F}^{-1}_{(1)}(\tilde{U}_{(1)s}),...,\tilde{X}_{(6)s} = \hat{F}^{-1}_{(6)}(\tilde{U}_{(6)s},\ s=1,\ldots,S$, donde $\hat{F}^{-1}_{(j)}$ denota la inversa de la función de distribución marginal acumulada estimada asociada a la variable $X_{(j)}$. Para las funciones de distribución marginales estimamos $j=1,\ldots,6$ distribuciones normales univariantes con parámetros μ_j^{Normal} y σ_j^{Normal} o 6 distribuciones t-Student univariantes con υ grados de libertad y parámetros $\mu_j^{Student}$ y $\sigma_j^{Student}$. En todos los casos, para la estimación de $\hat{F}_{(j)}$ utilizamos el método de máxima verosimilitud. Estudios recientes como el de Christoffersen et al. (2013) y Oh y Patton (2013) muestran que la distribución de los *log-return* son asimétricos y poseen colas gruesas (ver Fernández y Steel, 1998). En este sentido, realizamos diferentes contrastes de asimetría de las distribuciones marginales de los rendimientos filtrados (ver Boos, 1982) y, como indicaremos en la Sección 6.3, no rechazamos la hipótesis nula de simetría.

PASO 5: Finalmente, las pérdidas linealizadas simuladas se obtienen como $\tilde{L}_s = -V_0 \cdot \left(w_{(1)} \cdot \tilde{X}_{(1)s} + \ldots + w_{(6)} \cdot \tilde{X}_{(6)s}\right)$, $s=1,\ldots,S$ y, finalmente, estimamos el $VaR_\alpha(L)$ y el $CVaR_\alpha(L)$ empíricamente una vez que disponemos de un gran número de S datos simulados. En nuestro ejemplo numérico de la Sección 6.3, donde el objetivo es analizar la sensibilidad del VaR y el CVaR para la selección de D-vine, el cálculo de los pesos es irrelevante, por lo tanto, por simplicidad, asumimos $w_{(j)} = 1/6, \forall j = 1,\ldots,6$.

La estimación empírica del VaR para el orden o es:

$$\widehat{VaR}^o_\alpha(L)_S = \text{ínf}\left\{l, \widehat{F}^o_{L_S}(l) \geq \alpha\right\}, \tag{6.2.7}$$

donde $\widehat{F}^o_{L_S}$ es la distribución empírica de la variable pérdida, dado el orden o y que se obtiene a partir de las S pérdidas simuladas. La estimación empírica del CVaR ($\widehat{CVaR}^o_\alpha(L)_S$) es el promedio de las diferencias $\tilde{L}^o_s - \widehat{VaR}^o_\alpha(L)_S, \forall \tilde{L}^o_s > \widehat{VaR}^o_\alpha(L)_S$, donde $\tilde{L}^o_s$ representa las pérdidas simuladas usando el D-vine asociado con el orden o.

Una vez que hemos estimado el $\widehat{VaR}^o_\alpha(L)_S$ y el $\widehat{CVaR}^o_\alpha(L)_S$ para $o=1,...,\frac{6!}{2}=360$, con el objetivo de analizar la precisión en la estimación del riesgo, podemos calcular algunas medidas de dispersión basadas en momentos centrales, como por ejemplo la

desviación estándar o el coeficiente de variación. Alternativamente, también podemos calcular el rango (el máximo menos el mínimo) o el rango intercuartílico (el tercer cuartil menos el primer cuartil). En este punto, es importante tener en cuenta que el uso del procedimiento de Monte Carlo provoca cierta dispersión asociada con el proceso aleatorio en sí mismo. Para controlar esta dispersión espuria y para que los resultados sean comparables utilizamos la misma semilla inicial en cada generación aleatoria para cada descomposición *pair-copula*.

6.2.1. Criterio de selección del orden en el *D-vine*

Una manera sencilla de seleccionar el orden óptimo de los factores de riesgo en el *D-vine* consiste en maximizar el logaritmo de la pseudo-verosimilitud que se definió en (6.2.5), pero este criterio tiene algunos inconvenientes. En primer lugar, es necesario ajustar los parámetros de la cópula para todos las órdenes posibles con el fin de buscar aquellos que maximicen el logaritmo de la pseudo-verosimilitud. En segundo lugar, maximizar el logaritmo de la pseudo-verosimilitud no tiene que estar relacionado con la obtención de la mejor estructura de dependencia para estimar el riesgo; de hecho, el logaritmo de la pseudo-verosimilitud obtenido con diferentes órdenes nos permite estimar diferentes modelos condicionados dando más importancia a las observaciones con mayor densidad, lo que contradice el hecho de que el riesgo se asocia con las observaciones menos probables. Por esta razón, es necesario buscar diferentes criterios que nos permitan seleccionar el orden antes de estimar la cópula multivariante (ver de Melo Mendes et al., 2010, para un ejemplo financiero en seis dimensiones).

Dado que estamos interesados en estimar la estructura de dependencia, los criterios más naturales para la selección del orden, están basados en las medidas de dependencia relacionadas con las cópulas, es decir, la τ de Kendall, la ρ de Spearman y la dependencia en la cola inferior (izquierda) o superior (derecha) (λ_L y λ_U, respectivamente). Estas medidas de dependencia pueden ser definidas a partir de la cópula o empíricamente. Por esta razón, podemos decir que la selección del orden el *D-vine* se puede llevar a cabo con o sin la información de la cópula. Para estimar λ_L y λ_U empíricamente utilizamos la estimación no paramétrica propuesta por Schmidt y Stadtmüller (2006).

Para la selección de los pares, definimos un algoritmo que consiste en ordenar los nodos en el primer árbol (ver Figura 6.1) desde la dependencia máxima hasta la mínima, teniendo en cuenta la estructura *D-vine*. El procedimiento se define a continuación.

6.2.1.1. Procedimiento para la selección de los pares

Para la selección del *D-vine* se podría utilizar un algoritmo que encuentre la ruta más corta o más larga entre los nodos de un gráfico (ver, por ejemplo, Dissmann et al., 2013). Sin embargo, en nuestro contexto, este tipo de algoritmos se basa en maximizar la suma de las dependencias entre los pares de rendimientos, sin considerar la selección de un orden específico previo a la definición del *D-vine*. El procedimiento que proponemos está basado en la clasificación de los pares de rendimientos en función del grado de dependencia existente (de mayor a menor). En cada paso, se elige un nuevo par en función del par seleccionado en el paso anterior.

Sea D una matriz de dependencias que es simétrica y definida positiva y sea $d_{ij} = d_{ji}$, $i, j = 1, ...k$, la dependencia entre los rendimientos i y j, donde $d_{ij} = 1$ if $i = j$; el procedimiento en caso de no haber empates entre dependencias es:

PASO 1: El primer par $[(1),(2)]$ es aquel donde (el paréntesis indica el orden):

$$d_{(1)(2)} = Max_{i \neq j} \left(d_{ij} \right),$$

entonces el primer par, puede ser $[(1),(2)] = [i^*, j^*]$ o $[(1),(2)] = [j^*, i^*]$. Para poder seleccionar uno de los dos pares es necesario analizar los segundos posibles pares $[(2),(3)]$. Así:

$$d_{i^*(3)} = Max_{j \neq j^*} \left(d_{i^* j} \right)$$

y

$$d_{j^*(3)} = Max_{i \neq i^*} \left(d_{i j^*} \right),$$

entonces:

$$\text{si } d_{i^*(3)} > d_{j^*(3)} \text{ el primer par es } [(1),(2)] = [j^*, i^*]$$
$$\text{si } d_{i^*(3)} < d_{j^*(3)} \text{ el primer par es } [(1),(2)] = [i^*, j^*]$$

y eliminamos la fila i^* y la columna j^* de la matriz D. Denotamos como $D^{[-1]}$ a la matriz de dependencia con $k-1$ columnas y filas. En general, $D^{[-s]}$ es la matriz de dependencia con $k-s$ columnas y filas.

PASO 2 : Los siguientes pares $[(s+1)(s+2)]$ son seleccionados utilizando el criterio:

$$d^{-s}_{(s+1)(s+2)} = Max_{h \neq (s)} \left(d_{(s+1)h} \right) = d_{(s+1)h^*}, \tag{6.2.8}$$

donde $(s+2) = h^*$.

Si hay empates entre las dependencias, aplicamos el paso 1 para cada par inicial con la misma dependencia y seleccionamos los pares iniciales que proporcionan el mayor $d_{i^*(3)}$ o $d_{j^*(3)}$. Si existen empates entre las dependencias en el paso 2, similar al paso 1, aplicamos el criterio definido en (6.2.8) para cada par con la misma dependencia, así hasta el desempate. Este procedimiento es implementado en R por los autores.

6.2.2. Las cópulas analizadas

Comparamos los resultados obtenidos con las cópulas bivariantes t de Student, Gumbel, Clayton y Frank en la descomposición *pair-copula*. Para cada modelo multivariante utilizamos la misma cópula en todos los pares, aunque, como sugieren Weiß y Scheffer (2015), se podría utilizar la cópula que mejor ajusta en cada par. El uso de estas cuatro cópulas, nos permite tener en cuenta un amplio rango de estructuras de dependencia alternativas al modelo de referencia Gaussiano.

La cópula bivariante t de Student es una cópula implícita y elíptica que pertenece a la familia de cópulas de valor extremo (ver Bahraoui et al., 2014, para una revisión de la familia de cópulas de valor extremo y su inferencia). Su forma funcional es igual a la de la función de distribución univariante acumulada t de Student estándar, con υ grados de libertad y coeficiente de dependencia ρ. Esta cópula representa una estructura de dependencia simétrica y tiene colas más pesadas que las de la cópula Gaussiana. Adicionalmente, la cópula Gaussiana no presenta dependencia en las colas, mientras que la cópula t de Student tiene ambas, dependencia en la cola inferior y superior. Para un par dado, la cópula bivariante t de Student es:

$$C_{\rho_{12},\upsilon_{12}}(u_1,u_2)=\int_{-\infty}^{t_{\upsilon_{12}}^{-1}(u_1)}\int_{-\infty}^{t_{\upsilon_{12}}^{-1}(u_2)}\frac{1}{2\pi\sqrt{1-\rho_{12}^2}}\left(1+\frac{t_{\upsilon_{12}}^{-1}(s_1)^2+t_{\upsilon_{12}}^{-1}(s_2)^2-2\rho_{12}t_{\upsilon_{12}}^{-1}(s_1)t_{\upsilon_{12}}^{-1}(s_2)}{\upsilon_{12}(1-\rho_{12}^2)}\right)^{-\frac{\upsilon_{12}+2}{2}} dt_{\upsilon_{12}}^{-1}(s_1)dt_{\upsilon_{12}}^{-1}(s_2), \tag{6.2.9}$$

donde t_{υ}^{-1} es la función cuantil de la t de Student univariante.

De la familia de cópulas explícitas y Arquimedianas utilizamos las más populares: las cópulas de Gumbel (1960), Clayton (1978) y Frank (1979). Esta clase de cópulas bivariantes tiene una forma cerrada simple y su estructura depende únicamente de un único parámetro.

Cópula de Gumbel

La Gumbel es una cópula de valor extremo cuya forma funcional viene dada por:

$$C_{\theta_{12}}(u_1,u_2) = \exp(-[(-\ln(u_1))^{\theta_{12}} + (-\ln(u_2))^{\theta_{12}}]^{1/\theta_{12}}), \tag{6.2.10}$$

donde $\theta_{12} \in [1,+\infty)$ es el parámetro que controla la estructura de dependencia. La dependencia es perfecta cuando $\theta_{12} \to \infty$ y existe independencia cuando $\theta_{12} = 1$. Finalmente, para la cópula de Gumbel se sabe que $\lambda_L = 0$ y $\lambda_U = 2 - 2^{\frac{1}{\theta_{12}}}$, es decir, la cópula de Gumbel asume dependencia en la cola superior o derecha de la distribución.

Cópula de Clayton

La cópula de Clayton tiene una forma funcional igual a:

$$C_{\theta_{12}}(u_1,u_2) = (u_1^{-\theta_{12}} + u_2^{-\theta_{12}} - 1)^{-1/\theta_{12}}, \tag{6.2.11}$$

donde $\theta_{12} > 0$. En este caso, la estructura de dependencia perfecta se consigue cuando $\theta_{12} \to \infty$, mientras que la independencia se logra cuando $\theta_{12} \to 0$. En contraste con la cópula de Gumbel, la cópula de Clayton tiene dependencia en la cola inferior, de modo que $\lambda_L = 2^{-\frac{1}{\theta_{12}}}$ y $\lambda_U = 0$.

Cópula de Frank

La cópula de Frank se define a partir del parámetro $\theta_{12} \in (-\infty,0[\cup]0,+\infty)$, y su forma funcional, viene dada por la siguiente expresión:

$$C_{\theta_{12}}(u_1,u_2) = -\frac{1}{\theta_{12}} \ln(1 - \frac{(1-e^{\theta_{12}u_1})(1-e^{\theta_{12}u_2})}{1-e^{-\theta_{12}}}).$$

Esta cópula se caracteriza por presentar mayor dependencia en los cuantiles centrales. Los coeficientes de dependencia en las colas superior e inferior de la cópula de Frank son iguales a 0.

6.3. Resultado del análisis empírico usando datos financieros

Para estudiar el efecto de la selección del orden de los factores de riesgo en un *D-vine* en la cuantificación del riesgo, analizamos cuatro carteras (P1, P2, P3 y P4). La selección de las carteras se realiza cambiando las compañías que conforman cada una de ellas (vea la Tabla C.1 del Apéndice C, donde se detallan las empresas que conforman cada cartera). Las carteras más diversificadas (P1 y P2) están compuestas por los activos de las compañías que pertenecen, al menos, a cinco sectores económicos diferentes y, alternativamente, las carteras menos diversificadas (P3 y P4) están formadas por activos de compañías que pertenecen al mismo sector económico, o como mucho, de dos sectores diferentes. En nuestro caso, usamos solo acciones que cotizan en el mercado continuo pero los resultados pueden ser extrapolados directamente a portafolios compuestos por otros activos. Las diferencias en las diversificaciones del logaritmo de los rendimientos filtrados de las acciones o pérdidas que forman cada cartera, se evaluan tal y como describimos en la Sección 6.2.1 utilizando diferentes matrices de dependencia empírica para los logaritmos de los rendimientos filtrados - por ejemplo, la τ de Kendal y la ρ de Spearman - (vea la Tabla C.2 del Apéndice C donde se muestran la estimaciones de los modelos ARMA(P,Q)-GARCH(p,q)). Los logaritmos de los rendimientos se calculan utilizando los precios diarios desde Enero del 2011 hasta Diciembre del 2013, obtenidos a partir de Yahoo Finance. Todos los precios se expresan en dólares.

En las Figuras 6.1, 6.2, 6.3 y 6.4 se muestran los gráficos de dispersión de los rendimientos filtrados en las carteras P1, P2, P3 y P4, respectivamente. De las Figuras 6.2 y 6.4 eliminamos los tres gráficos que habíamos incluido en las Figuras 6.1 y 6.3. En general, se observa una forma elíptica que refleja dependencia entre los rendimientos filtrados. Esta forma elíptica es más pronunciada en las Figuras 6.3 y 6.4, que corresponden con las carteras menos diversificadas. También se observan formas bastante simétricas y algunos puntos extremos que reflejan la dependencia en la cola entre los rendimientos filtrados.

En la Tabla 6.1 se muestra la Tau de Kendall (debajo de la diagonal principal) y la ρ de Spearman (encima de la diagonal principal) para cada cartera. Posteriormente, en la Tabla 6.2 se presentan las dependencias de la cola superior e inferior también calculadas de forma empírica (R1, ..., R6 se refiere al logaritmo de los rendimientos filtrados). Además, se calcula el determinante (entre paréntesis) de cada matriz de dependencia, cuanto menor sea este determinante, mayor será la dependencia entre los rendimientos.

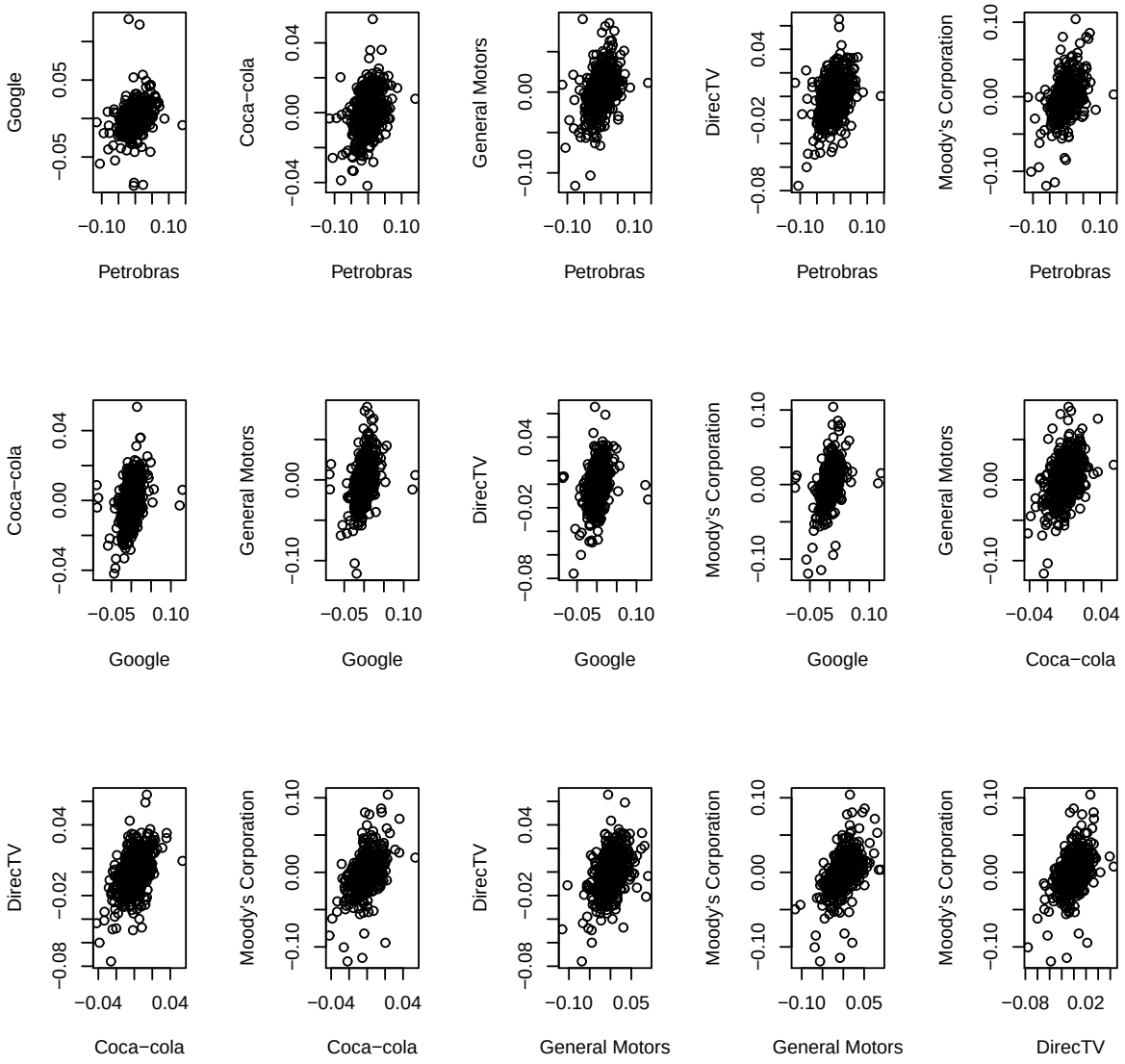


FIGURA 6.1: Gráfico de dispersión de los logaritmos de los rendimientos filtrados en P1.

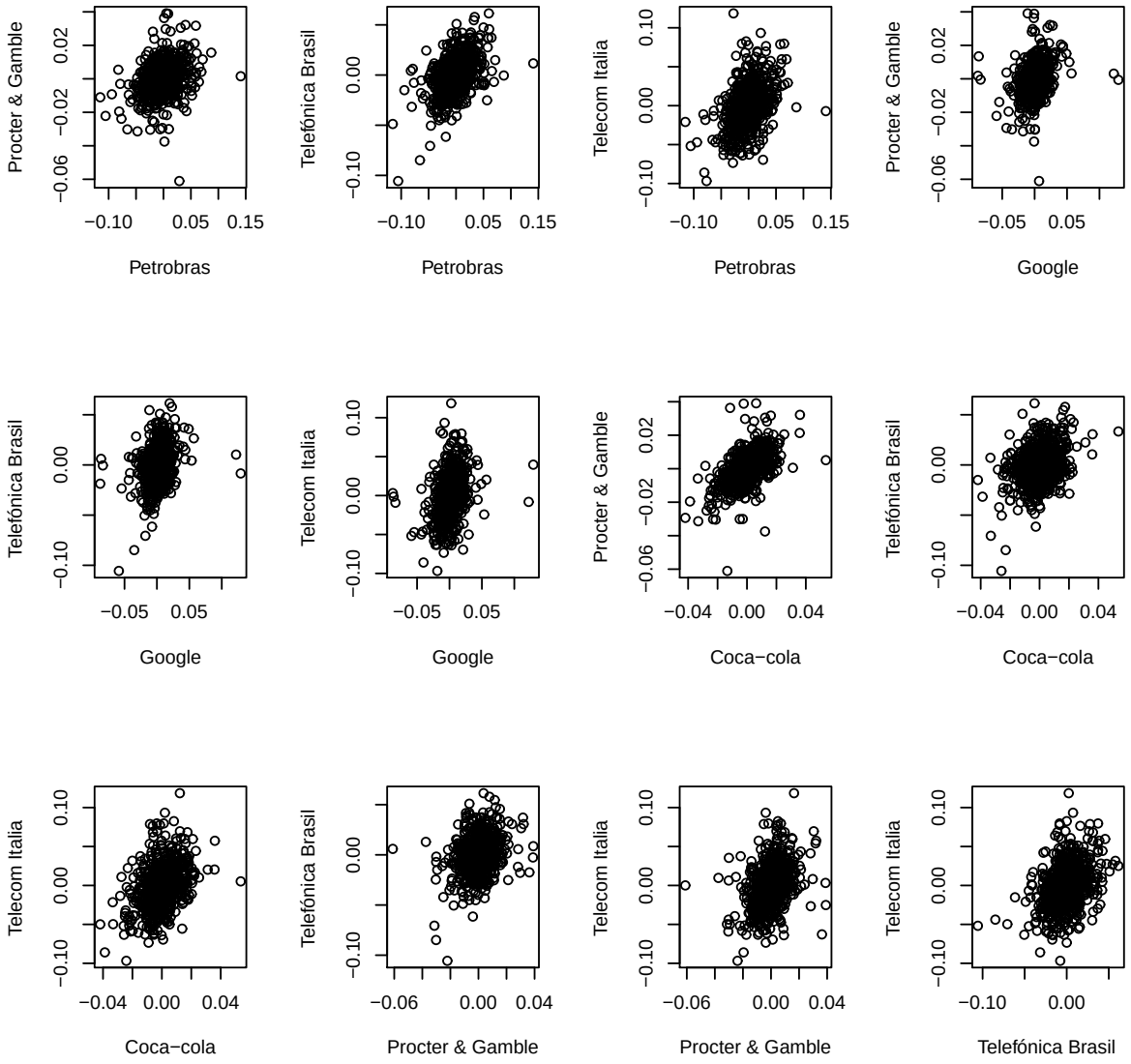


FIGURA 6.2: Gráfico de dispersión de los logaritmos de los rendimientos filtrados en P2 que no están incluidos en P1.

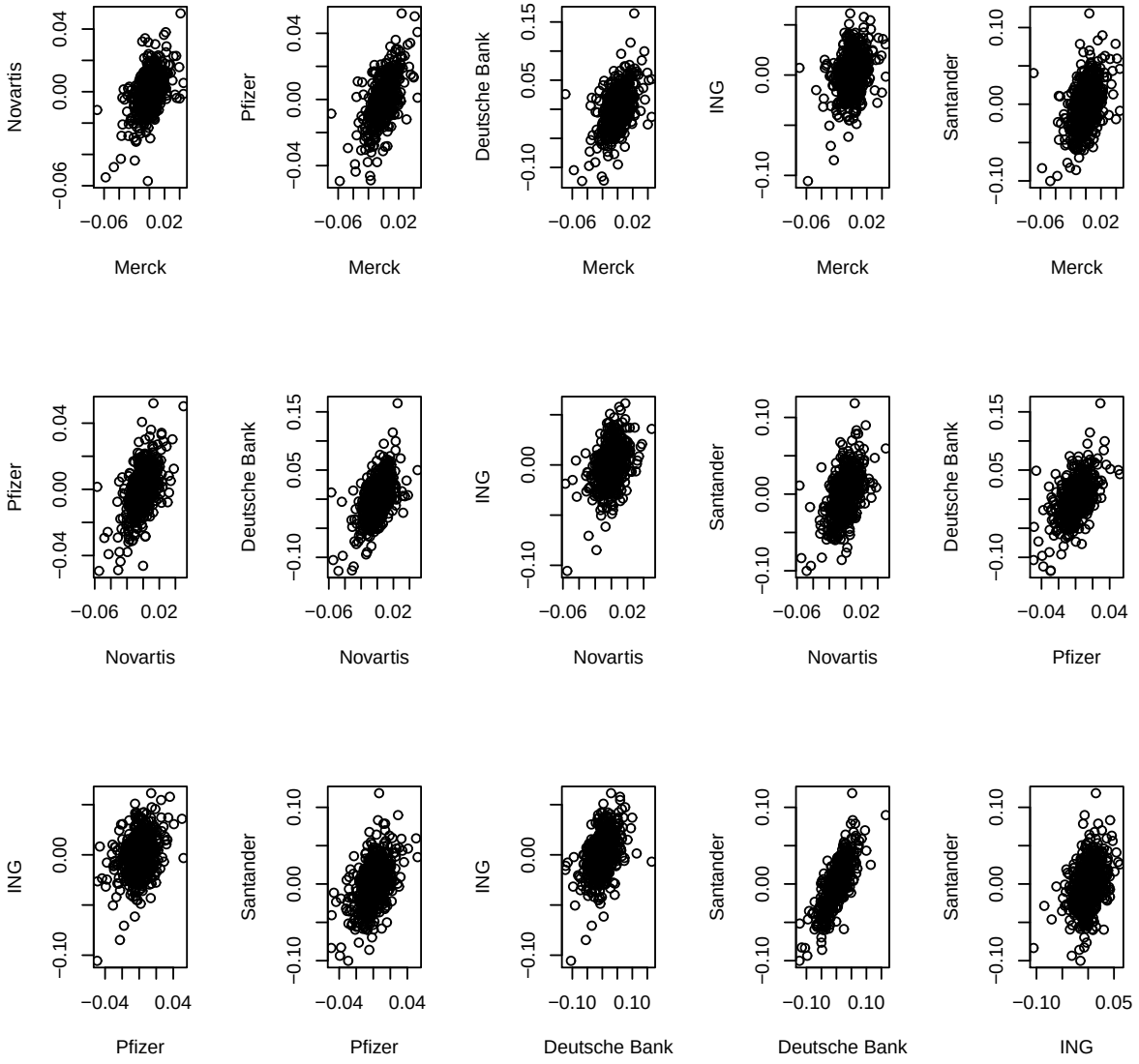


FIGURA 6.3: Gráfico de dispersión de los logaritmos de los rendimientos filtrados en P3.

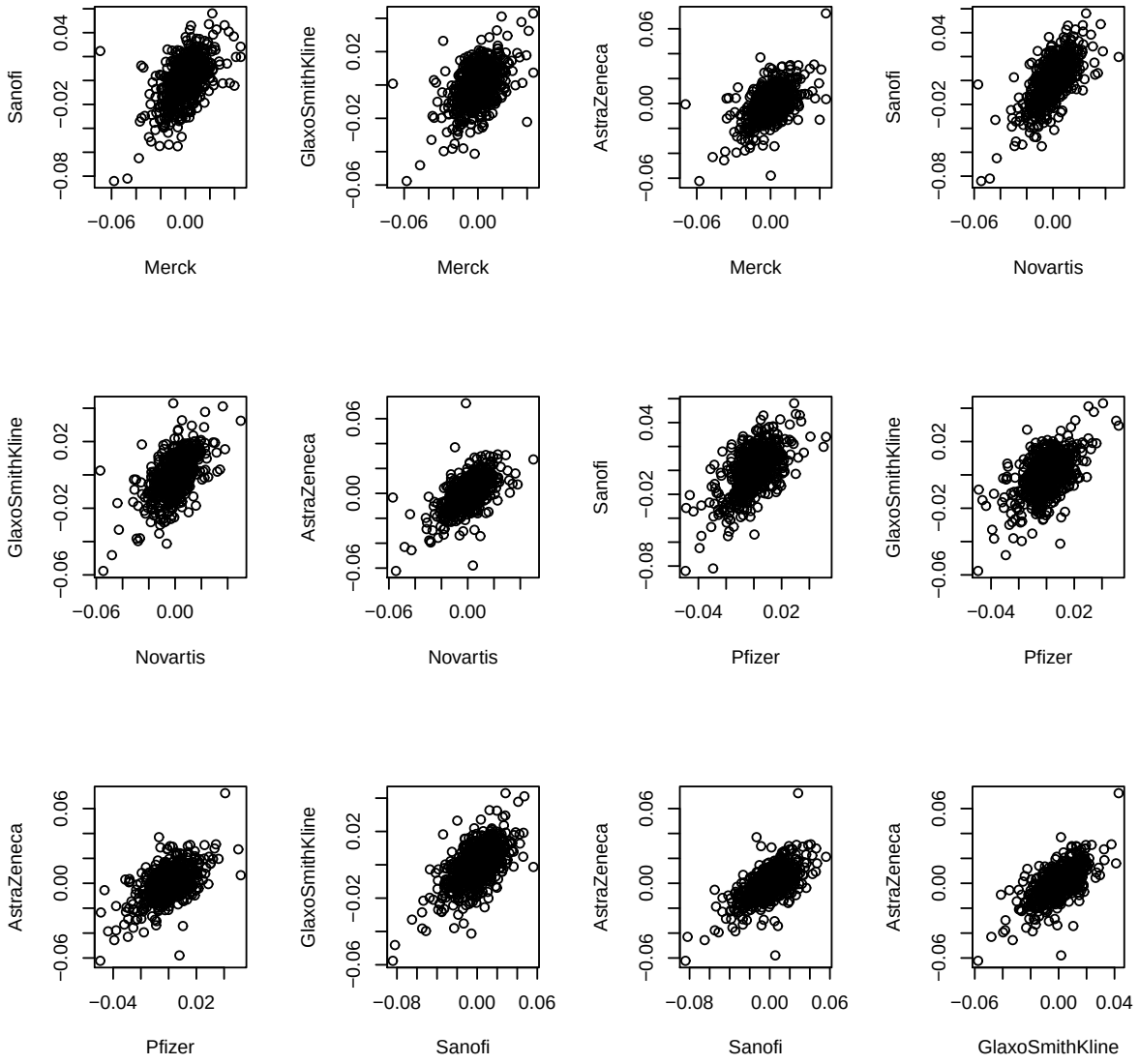


FIGURA 6.4: Gráfico de dispersión de los logaritmos de los rendimientos filtrados en P4 que no están incluidos en P3.

TABLA 6.1: Medidas de dependencia: ρ de Spearman empírica (encima de la diagonal) y la τ de Kendall empírica (debajo de la diagonal). P1, ..., P4 hacen referencia a la cartera y R1, ..., R6 se refieren a los logaritmos de los rendimientos filtrados.

P1

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.371	0.315	0.404	0.373	0.427	
R2	0.256	1.000	0.396	0.465	0.392	0.455	
R3	0.215	0.274	1.000	0.377	0.433	0.462	ρ
R4	0.278	0.323	0.261	1.000	0.390	0.538	*(0.19)*
R5	0.261	0.270	0.302	0.270	1.000	0.446	
R6	0.298	0.318	0.322	0.381	0.311	1.000	
			τ *(0.43)*				

P2

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.371	0.315	0.276	0.427	0.421	
R2	0.256	1.000	0.396	0.392	0.264	0.377	
R3	0.215	0.274	1.000	0.548	0.292	0.374	ρ
R4	0.189	0.270	0.389	1.000	0.250	0.374	*(0.25)*
R5	0.295	0.178	0.201	0.171	1.000	0.339	
R6	0.292	0.261	0.258	0.259	0.231	1.000	
			τ *(0.50)*				

P3

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.449	0.573	0.411	0.239	0.388	
R2	0.312	1.000	0.482	0.477	0.321	0.469	
R3	0.418	0.338	1.000	0.460	0.287	0.407	ρ
R4	0.289	0.335	0.324	1.000	0.388	0.766	*(0.11)*
R5	0.162	0.217	0.196	0.265	1.000	0.333	
R6	0.270	0.329	0.282	0.585	0.226	1.000	
			τ *(0.33)*				

P4

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.449	0.573	0.465	0.399	0.480	
R2	0.312	1.000	0.482	0.630	0.557	0.583	
R3	0.418	0.338	1.000	0.521	0.441	0.489	ρ
R4	0.327	0.456	0.373	1.000	0.554	0.622	*(0.071)*
R5	0.279	0.395	0.308	0.395	1.000	0.630	
R6	0.338	0.421	0.346	0.450	0.458	1.000	
			τ *(0.25)*				

TABLA 6.2: Medidas de dependencia: λ_L empírica (encima de la diagonal) y λ_U empírica (debajo de la diagonal). P1, ..., P4 hacen referencia a la cartera and R1, ..., R6 se refieren a los logaritmos de los rendimientos filtrados.

P1

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.222	0.333	0.370	0.333	0.370	
R2	0.222	1.000	0.333	0.296	0.259	0.370	
R3	0.111	0.185	1.000	0.333	0.370	0.407	λ_L
R4	0.074	0.148	0.148	1.000	0.296	0.333	*(0.34)*
R5	0.148	0.185	0.333	0.148	1.000	0.296	
R6	0.222	0.185	0.222	0.185	0.222	1.000	
			λ_U *(0.66)*				

P2

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.222	0.333	0.222	0.296	0.222	
R2	0.222	1.000	0.333	0.222	0.148	0.185	
R3	0.111	0.185	1.000	0.481	0.296	0.333	λ_L
R4	0.148	0.222	0.222	1.000	0.296	0.333	*(0.41)*
R5	0.259	0.185	0.185	0.074	1.000	0.222	
R6	0.185	0.111	0.111	0.148	0.148	1.000	
			λ_U *(0.70)*				

P3

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.370	0.333	0.444	0.259	0.259	
R2	0.222	1.000	0.444	0.444	0.185	0.333	
R3	0.407	0.296	1.000	0.370	0.185	0.259	λ_L
R4	0.370	0.185	0.296	1.000	0.259	0.370	*(0.34)*
R5	0.222	0.148	0.185	0.259	1.000	0.148	
R6	0.259	0.259	0.259	0.481	0.148	1.000	
			λ_U *(0.41)*				

P4

	R1	R2	R3	R4	R5	R6	
R1	1.000	0.370	0.333	0.370	0.259	0.407	
R2	0.222	1.000	0.444	0.444	0.333	0.407	
R3	0.407	0.296	1.000	0.370	0.296	0.481	λ_L
R4	0.296	0.296	0.259	1.000	0.519	0.481	*(0.21)*
R5	0.222	0.370	0.444	0.333	1.000	0.370	
R6	0.296	0.259	0.222	0.333	0.333	1.000	
			λ_U *(0.36)*				

Tras el ajuste de la descomposición *pair-copula* utilizando las cópulas t de Student, Gumbel, Clayton y Frank, el mejor resultado se obtiene con la cópula t de Student. La cópula de Clayton no pasa el test de bondad de ajuste *PIT* al 5 % de significación en ninguno de los casos, así que, se eliminan los resultados asociados a esta cópula. Para el resto de las cópulas, todos los *D-vines* pasan el test de *PIT* en las cuatro carteras y el mayor nivel de significación se obtiene con la cópula t de Student.

Para el ajuste de las marginales univariantes, se utilizan los rendimientos filtrados de cada activo y se contrasta la hipótesis de simetría utilizando el estadístico basado en el estimador Hodges-Lehmann (ver Boos, 1982). En todos los casos, no es posible rechazar la hipótesis nula de simetría. Dado estos resultados, se analiza el ajuste de las distribuciones Normal y t de Student, y se evidencia que la t de Student es la que mejor ajusta para todos los rendimientos filtrados.

6.3.1. Análisis de la dispersión

El análisis de la forma de las distribuciones del VaR y CVaR estimados se realiza considerando los 360 órdenes iniciales diferentes en el árbol T_1 del *D-vine* (lo cual, se traduce en 360 modelos de dependencia multivariantes distintos). Es decir, obtenemos el VaR y el CVaR para cada orden en el árbol *D-vine* y utilizamos el método de estimación núcleo de la función de densidad (ver Silverman, 1986). Este es un método no paramétrico que permite de un modo sencillo suavizar la forma de los histogramas y así facilitar la comparación entre las distribuciones del VaR y *CVaR* estimados.

En las Figuras 6.5, 6.6 and 6.7 se muestran las densidades obtenidas utilizando las cópulas t-Student, Frank y Gumbel, con distribuciones marginales t-Student. Es importante mencionar que, las cópulas t-Student y Frank asumen estructuras de dependencia simétricas entre los logaritmos de los rendimientos filtrados dos a dos. Mientras que, la cópula de Gumbel supone asimetría en las estructuras de dependencia y, en nuestro ejemplo, genera estimaciones con mayor riesgo.

Las diferencias entre las cópulas, también pueden observarse en las Tablas 6.3 y 6.4, donde se exponen la media (Mean), la desviación estándar (STD), el coeficiente de variación (CV), el rango intercuartílico (IQR) y el rango (Range) de las estimaciones del VaR y CVaR a niveles de confianza del 99 % y 99.5 %, utilizando 360 órdenes iniciales distintos. Después de analizar las diferentes medidas de dispersión, se observa que los resultados de la cópula de Gumbel difieren de los obtenidos con las cópulas t de Student y Frank.

Las densidades del VaR expuestas en las Figuras 6.5 y 6.6 (parte superior del gráfico), se corresponden con las cópulas t de Student y Frank, respectivamente. Se observa solo una moda principal y en el caso de la cópula de Frank, todas las distribuciones tienen más curtosis (comparar eje vertical), menor dispersión y centros más pequeños respecto a los observados en la cópula t de Student (ver también la Tabla 6.3). A partir de los resultados obtenidos utilizando las medidas de riesgo analizadas - el VaR para los niveles de confianza del 99 % y 99.5 % - es posible visualizar cómo para ambas cópulas las carteras pueden ser ordenadas de mayor a menor riesgo de la misma manera: P3, P1, P2 y P4. Por su parte, en los gráficos de las densidades del CVaR (parte inferior del gráfico) se observan, por una parte, muchas similitudes entre P1 y P3 y, por otra, cómo se reducen las diferencias entre P2 y P4.

Las densidades expuestas en la Figura 6.7 y que han sido obtenidas utilizando la cópula de Gumbel en los pares de la descomposición *pair-copula* de la distribución multivariante, presentan una forma bimodal, mayor dispersión de la distribución y un valor del

centro mayor que el obtenido utilizando la cópula t de Student. En general, los riesgos estimados a partir de la cópula de Gumbel son mucho mayores y las diferencias entre dichos riesgos utilizando diferentes órdenes iniciales son también mucho más grandes. Nuevamente, se observan similitudes entre las densidades de P1 y P3, y lo mismo ocurre con P2 y P4.

En general, se observa que las distribuciones de las estimaciones del VaR y CVaR con diferentes *D-vines* dependen de la cópula seleccionada y pueden tener formas distintas. En este sentido, resulta fundamental analizar los resultados proporcionados por los diferentes criterios de selección del *D-vine*.

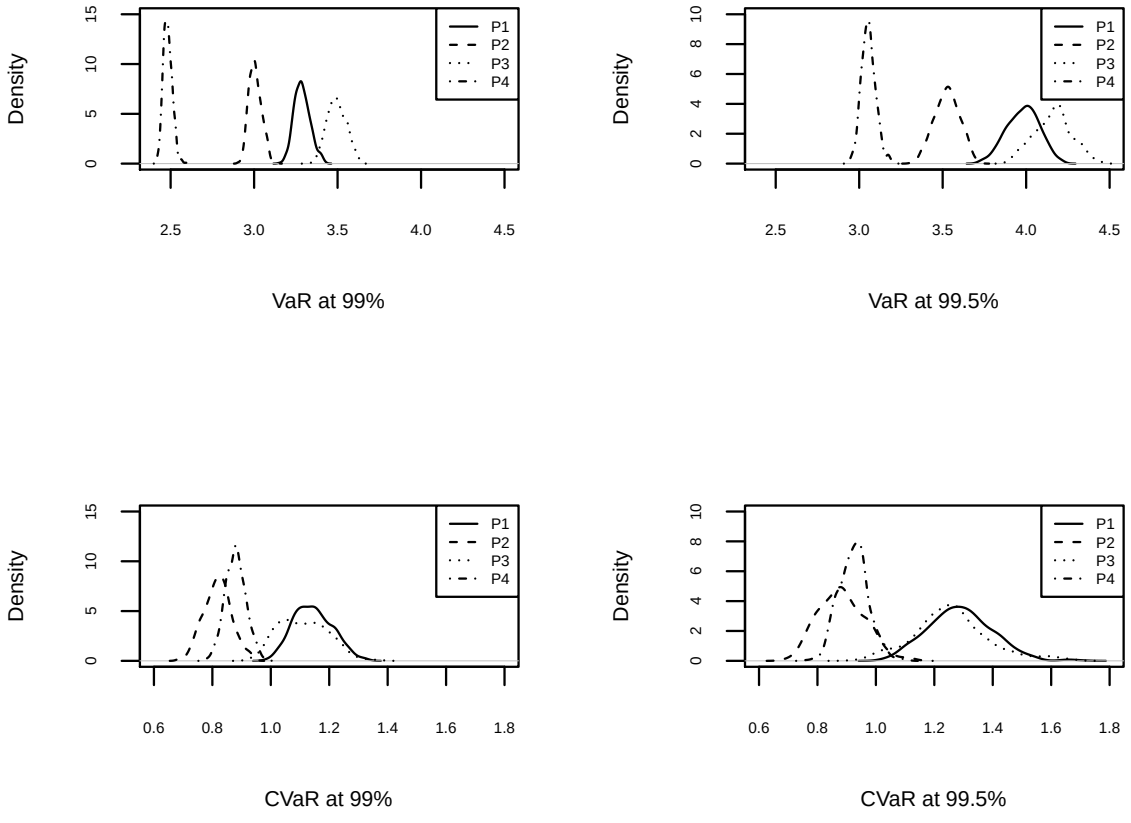


FIGURA 6.5: Estimación núcleo de la densidad del VaR (gráficos de la parte superior) y CVaR (gráficos de la parte inferior), teniendo en cuenta todos los posibles *D-vines* y utilizando la cópula t de Student en la descomposición *pair-copula*.

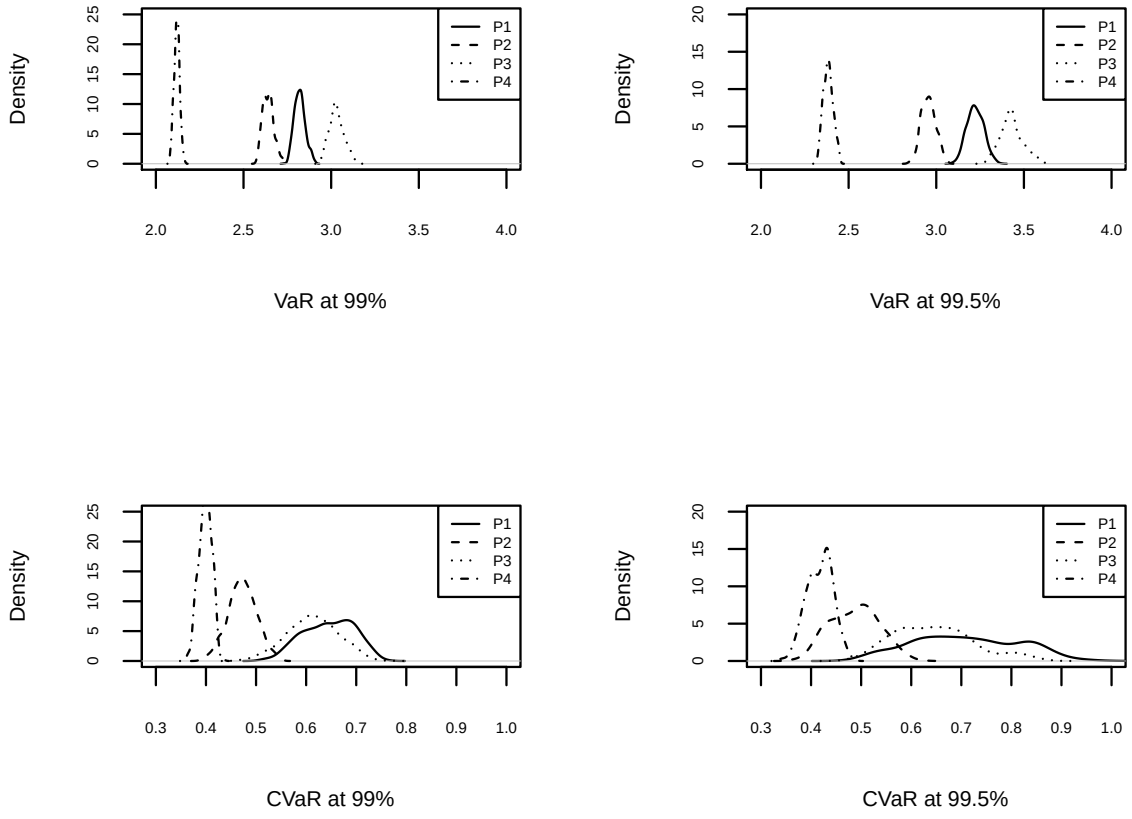


FIGURA 6.6: Estimación núcleo de la densidad del VaR (gráficos de la parte superior) y CVaR (gráficos de la parte inferior), teniendo en cuenta todos los posibles *D-vines* y utilizando la cópula de Frank en la descomposición *pair-copula*.

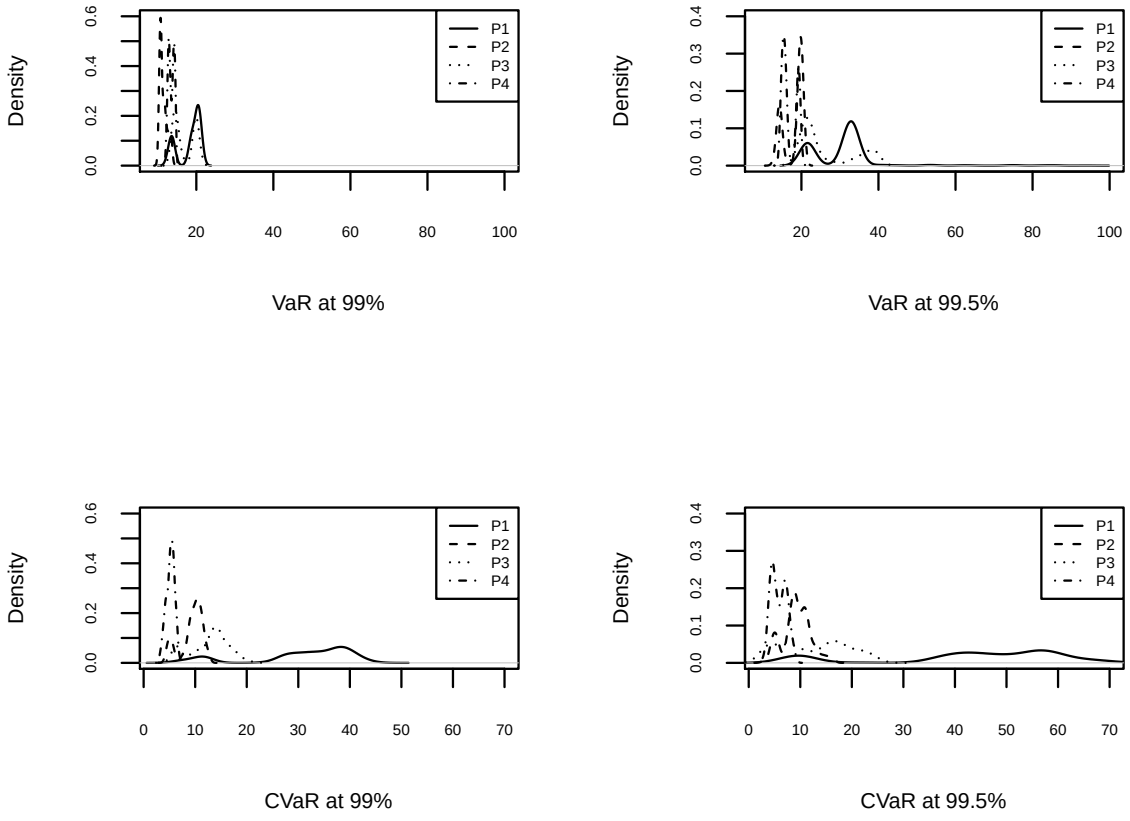


FIGURA 6.7: Estimación núcleo de la densidad del VaR (gráficos de la parte superior) y CVaR (gráficos de la parte inferior), teniendo en cuenta todos los posibles *D-vines* y utilizando la cópula de Gumbel en la descomposición *pair-copula*.

TABLA 6.3: Estadísticas descriptivas del VaR estimado para todos los órdenes: el VaR es multiplicado por 100.

Cartera	P1						P2					
	Student's t		Gumbel		Frank		Student's t		Gumbel		Frank	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Mean	3.287	3.981	18.286	30.514	2.820	3.223	3.006	3.526	11.449	17.975	2.641	2.956
STD	0.048	0.098	2.986	9.015	0.031	0.048	0.038	0.075	0.993	2.619	0.031	0.041
CV	0.015	0.025	0.163	0.295	0.011	0.015	0.013	0.021	0.087	0.146	0.012	0.014
IQR	0.066	0.140	6.256	11.228	0.041	0.068	0.055	0.105	1.388	5.349	0.043	0.058
Range	0.277	0.510	9.487	73.181	0.177	0.281	0.232	0.412	4.542	8.302	0.173	0.229
Cartera	P3						P4					
	Student's t		Gumbel		Frank		Student's t		Gumbel		Frank	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Mean	3.495	4.160	16.933	26.408	3.031	3.430	2.482	3.056	13.512	16.862	2.119	2.383
STD	0.056	0.111	2.733	7.177	0.042	0.064	0.027	0.043	0.742	1.879	0.016	0.027
CV	0.016	0.027	0.161	0.272	0.014	0.019	0.011	0.014	0.055	0.111	0.008	0.012
IQR	0.083	0.144	5.408	13.169	0.055	0.080	0.036	0.056	1.325	3.833	0.021	0.040
Range	0.308	0.561	8.854	23.810	0.214	0.338	0.173	0.269	2.815	5.475	0.092	0.136

TABLA 6.4: Estadísticas descriptivas del CVaR estimado para todos los órdenes: el CVaR es multiplicado por 100.

Cartera	P1						P2					
	Student's t		Gumbel		Frank		Student's t		Gumbel		Frank	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Mean	1.144	1.289	30.757	43.502	0.645	0.709	0.818	0.883	9.352	9.324	0.473	0.487
STD	0.066	0.110	10.138	17.728	0.051	0.106	0.047	0.080	2.174	2.680	0.029	0.049
CV	0.058	0.085	0.330	0.408	0.079	0.149	0.058	0.091	0.233	0.287	0.060	0.100
IQR	0.096	0.148	10.432	18.585	0.079	0.168	0.063	0.114	2.181	2.995	0.038	0.074
Range	0.342	0.687	39.541	68.379	0.249	0.519	0.255	0.452	8.877	13.996	0.157	0.264
Cartera	P3						P4					
	Student's t		Gumbel		Frank		Student's t		Gumbel		Frank	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Mean	1.111	1.258	12.799	12.789	0.613	0.645	0.879	0.924	5.189	5.815	0.398	0.418
STD	0.084	0.129	3.691	6.509	0.051	0.080	0.035	0.051	0.818	1.468	0.013	0.026
CV	0.075	0.103	0.288	0.509	0.083	0.124	0.040	0.055	0.158	0.252	0.033	0.062
IQR	0.128	0.144	4.808	10.690	0.071	0.112	0.049	0.065	1.173	2.492	0.018	0.038
Range	0.448	0.766	16.219	23.390	0.299	0.390	0.207	0.346	4.226	6.848	0.079	0.145

6.3.2. Selección de orden óptimo

En la Sección 6.2.1 describimos criterios alternativos para seleccionar el orden óptimo entre los 360 órdenes posibles en el D-vine de seis dimensiones. En esta sección, se exponen los resultados obtenidos con el mejor ajuste que, en este caso, es el de la cópula t de Student con las distribuciones marginales t de Student.

Primero, en la Tabla 6.5 mostramos el orden óptimo que se obtuvo usando los diferentes criterios estimados empíricamente (parte izquierda) y suponiendo una cópula t de Student (parte derecha). En el último caso agregamos el criterio basado en grados de libertad (cambiando Max por Min en el procedimiento definido en la Sub-sección 6.2.1.1) y el basado en la estructura de dependencia que maximiza la pseudo-verosimilitud. Segundo, en las Tablas 6.6 y 6.7 se exponen los VaR y CVaR estimados para los niveles de confianza del 99 % y 99.5 %, respectivamente, usando los diferentes órdenes citados en la Tabla 6.5. Los resultados muestran cómo los diferentes criterios proporcionan diferentes órdenes pero, en general, no hay muchas diferencias entre el VaR y el CVaR estimado. Aunque los órdenes difieren, los resultados del VaR y del CVaR estimados con criterios empíricos y basados en la cópula son, a primera vista, muy similares.

En las Tablas 6.6 y 6.7, debajo de las estimaciones del VaR y CVaR, para cada orden, incluimos los intervalos de confianza *bootstrap* (IC) a un nivel de confianza de 90 % (entre paréntesis), es decir, un 5 % de valores estimados se encuentran por encima del límite superior y un 5 % se encuentra por debajo del límite inferior. Los IC nos permiten evaluar en qué medida las diferencias entre los resultados obtenidos con diferentes órdenes son estadísticamente significativas. Para calcular los intervalos de confianza *bootstrap* extraemos 500 muestras con reemplazo del conjunto de datos asociado a cada cartera. Para cada muestra se estima el VaR y CVaR con los niveles de confianza de 99 % y 99.5 % utilizando el mismo orden seleccionado en la Tabla 6.5. Por lo tanto, no se considera la dispersión asociada a los criterios de selección, sino que ésta se evalúa en el estudio de simulación desarrollado en la Sección 6.4.

En general, observamos que los IC se superponen en mayor o menor medida. Solo hay un caso donde los IC no se superponen, y corresponde con la cartera P1 al comparar el VaR obtenido con el criterio λ_U estimado empíricamente, con el VaR estimado a partir del *D-vine* que maximiza la verosimilitud. Para todos los demás casos, podemos concluir que los valores de riesgo estimados con diferentes D-vines no son significativamente diferentes a un nivel del 5 % si la prueba es para una cola y a un nivel del 10 % si es para dos colas.

TABLA 6.5: Órdenes seleccionados.

	Criterio empírico				Criterio basado en la cópula t de Student					
Cartera	ρ	τ	λ_U	λ_L	ρ	τ	λ_U	λ_L	gdl	Verosimilitud
P1	4,6,3,5,2,1		5,3,6,1,2,4	6,3,5,1,4,2	6,4,2,3,5,1		1,5,3,6,4,2		1,5,3,4,6,2	6,1,5,3,4,2
P2	3,4,2,6,1,5		1,5,2,4,3,6	4,3,1,5,6,2	3,4,2,1,5,6		4,3,2,1,6,5			5,1,2,6,4,3
P3	4,6,2,3,1,5		4,6,1,3,2,5	4,1,2,3,6,5	4,6,2,3,1,5		4,6,1,3,2,5		1,3,4,6,2,5	5,4,6,3,1,2
P4	4,2,6,5,3,1	6,5,4,2,3,1	3,5,2,4,6,1	4,5,6,3,2,1	4,2,6,5,3,1		3,1,6,5,4,2		3,1,5,6,2,4	6,5,4,3,1,2

TABLA 6.6: VaR y CVaR estimados a un nivel del 99 % utilizando los diferentes órdenes seccionados en la Tabla 6.5: los intervalos de confianza están entre paréntesis.

VaR	Criterio empírico				Criterio basado en la cópula t de Student					
Cartera	ρ	τ	λ_U	λ_L	ρ	τ	λ_U	λ_L	gdl	Verosimilitud
P1	3.255 (3.014,3.399)		3.364 (3.257,3.644)	3.250 (3.017,3.372)	3.253 (3.005,3.373)		3.338 (3.180,3.563)		3.346 (3.207,3.588)	3.193 (2.856,3.245)
P2	3.071 (2.963,3.282)		2.957 (2.705,3.038)	3.017 (2.812,3.138)	3.034 (2.873,3.198)		3.030 (2.905,3.217)			3.026 (2.851,3.171)
P3	3.551 (3.538,3.908)		3.494 (3.441,3.797)	3.487 (3.430 3.805)	3.551 (3.538,3.908)		3.494 (3.424,3.794)		3.489 (3.410,3.758)	3.460 (3.262,3.628)
P4	2.473 (2.265,2.540)	2.499 (2.317,2.592)	2.464 (2.255,2.524)	2.533 (2.403,2.663)	2.473 (2.265,2.540)		2.501 (2.335,2.599)		2.522 (2.373,2.639)	2.518 (2.369,2.635)
CVaR	Criterio empírico				Criterio basado en la cópula t de Student					
Cartera	ρ	τ	λ_U	λ_L	ρ	τ	λ_U	λ_L	gdl	Verosimilitud
P1	1.210 (0.906,1.481)		1.052 (0.626,1.161)	1.260 (1.002,1.587)	1.211 (0.959,1.471)		1.120 (0.755,1.286)		1.156 (0.822,1.342)	1.270 (1.071,1.599)
P2	0.819 (0.591,0.944)		0.861 (0.713,1.043)	0.800 (0.551,0.895)	0.840 (0.622,0.973)		0.836 (0.634,0.987)			0.785 (0.527,0.879)
P3	1.066 (0.748,1.239)		1.125 (0.867,1.365)	1.103 (0.861,1.357)	1.066 (0.748,1.239)		1.125 (0.866,1.357)		1.085 (0.781,1.294)	1.093 (0.774,1.292)
P4	0.921 (0.740,1.125)	0.905 (0.708,1.093)	0.884 (0.683,1.070)	0.853 (0.630,1.000)	0.921 (0.740,1.125)		0.898 (0.689,1.087)		0.857 (0.604,1.006)	0.865 (0.645,1.021)

TABLA 6.7: VaR y CVaR estimados a un nivel del 99.5 % utilizando los diferentes órdenes seccionados en la Tabla 6.5: los intervalos de confianza están entre paréntesis.

VaR	Criterio empírico				Criterio basado en la cópula t de Student					
Cartera	ρ	τ	λ_U	λ_L	ρ	τ	λ_U	λ_L	gdl	Verosimilitud
P1	4.019 (3.753,4.317)		4.019 (3.776,4.348)	4.068 (3.856,4.415)	4.127 (3.997,4.512)		4.076 (3.856,4.413)		4.041 (3.788,4.357)	3.995 (3.673,4.248)
P2	3.613 (3.450,3.893)		3.533 (3.287,3.725)	3.495 (3.153,3.614)	3.641 (3.493,3.937)		3.655 (3.561,4.008)			3.497 (3.175,3.652)
P3	4.135 (3.940,4.488)		4.164 (4.011,4.575)	4.221 (4.146,4.704)	4.135 (3.940,4.488)		4.164 (3.998,4.546)		4.193 (4.054,4.575)	4.073 (3.703,4.233)
P4	3.068 (2.855,3.271)	3.099 (2.917,3.333)	3.046 (2.829,3.234)	3.038 (2.816,3.222)	3.068 (2.855,3.271)		3.050 (2.809,3.238)		3.062 (2.832,3.255)	3.087 (2.883 3.320)
CVaR	Criterio empírico				Criterio basado en la cópula t de Student					
Cartera	ρ	τ	λ_U	λ_L	ρ	τ	λ_U	λ_L	gdl	Verosimilitud
P1	1.315 (0.691,1.637)		1.207 (0.609,1.428)	1.355 (0.779,1.722)	1.239 (0.647,1.486)		1.145 (0.457,1.301)		1.286 (0.696,1.554)	1.408 (1.001,1.820)
P2	0.856 (0.442,1.018)		0.865 (0.563,1.063)	0.917 (0.599,1.138)	0.758 (0.270,0.822)		0.779 (0.309,0.861)			0.895 (0.565,1.098)
P3	1.292 (0.875 1.671)		1.276 (0.866,1.658)	1.162 (0.712,1.458)	1.292 (0.875,1.671)		1.276 (0.843,1.639)		1.136 (0.603,1.365)	1.303 (0.917,1.686)
P4	1.013 (0.695,1.277)	0.966 (0.601,1.183)	0.907 (0.482,1.088)	0.955 (0.603,1.195)	1.013 (0.695,1.277)		0.996 (0.624,1.250)		0.964 (0.551,1.188)	0.922 (0.524,1.097)

6.4. Estudio de simulación

En esta parte presentamos los resultados de un estudio de simulación. Por una parte, comparamos los errores cuadráticos medios (ECM) del VaR y CVaR obtenidos con los diferentes criterios de selección del *D-vine* descritos en la Sección 6.2.1, y por otra, analizamos la existencia de diferencias significativas entre estos criterios.

Específicamente, generamos 500 muestras de tamaño 500 a partir de cuatro descomposiciones *pair-copula*, dos con dimensión cuatro (dim = 4) y dos con dimensión seis (dim = 6), en total cuatro modelos, respectivamente denominados: Modelo 1 (dim = 4), Modelo 2 (dim = 6), Modelo 3 (dim = 4) y Modelo 4 (dim = 6). Se utiliza el *D-vine* basado en los órdenes dados, definidos como $(1,2,3,4)$ para dim = 4 y como $(1,2,3,4,5,6)$ para dim = 6. A su vez, para cada par se utiliza la cópula t de Student. Los parámetros teóricos utilizados en los modelos simulados se muestran en la Tabla C.3 del Apéndice C. Estos parámetros teóricos se definen en función de los obtenidos para las carteras analizadas en la Sección 6.3, es decir, suponiendo mayor (Modelo 3 y Modelo 4) o menor (Modelo 1 y Modelo 2) dependencia entre las variables aleatorias en el vector multivariante. Además, en la Tabla C.4 del Apéndice C se muestran los verdaderos VaR y CVaR, utilizados para obtener el ECM asociado con criterios de selección alternativos.

El estudio de simulación se lleva a cabo en cuatro pasos. Primero, a partir de cada uno de los cuatro modelos teóricos, se generan 500 muestras de tamaño 500, es decir, se definen cuatro conjuntos de datos simulados. Segundo, aplicamos el procedimiento de selección de pares descrito en la Sección 6.2.1.1 a cada muestra en cada conjunto de datos. Tercero, se estiman los parámetros para cada par del *D-vine* asociados con cada orden y utilizando la cópula t de Student. Por último, se utiliza el método de Monte Carlo para estimar el VaR y CVaR para cada muestra. Todo esto, suponiendo que las marginales siguen una distribución $\mathrm{U}(0,1)$.

El cálculo de los ECM del VaR_j y el $CVaR_j$ estimados a partir del modelo j, utilizando los k criterios se obtienen a partir de las siguientes expresiones:

$$\widehat{MSE_k}\left(\widehat{VaR_j}\right) = \frac{1}{500}\sum_{i=1}^{500}\left(VaR_{ij}^k - VaR_j\right))^2 \tag{6.4.1}$$

and

$$\widehat{MSE_k}\left(\widehat{CVaR_j}\right) = \frac{1}{500}\sum_{i=1}^{500}\left(CVaR_{ij}^k - CVaR_j\right))^2, \tag{6.4.2}$$

donde k se refiere, por una parte, a los criterios basados en ρ, τ, λ_U y λ_L empíricos y, por otra parte, a los criterios basados en ρ, τ, λ_U, λ_L y los grados de libertad estimados a partir de la cópula t de Student.

En la práctica, para eliminar el efecto del tamaño dados los diferentes valores teóricos, calculamos un ECM relativo, el cual se define como la raíz cuadrada del ECM dividido entre el VaR o CVaR teórico correspondiente. Los resultados para el VaR y el CVaR se muestran en la Tabla 6.8. En todos los casos se utiliza la inferencia estadística para comprobar el valor de la media estimada del VaR y el CVaR, obtenida a partir de las 500 muestras simuladas, es decir, contrastamos la hipótesis nula que supone que los valores estimados son iguales a los valores teóricos. En ambas tablas, subrayamos y marcamos en cursiva los casos donde no se rechaza la hipótesis nula.

TABLA 6.8: Raíz cuadrada del ECM dividido por los valores teóricos correspondientes del VaR y CVaR obtenidos con los diferentes criterios de selección de órdenes utilizando la cópula t de Student.

	Empirical Criterion															
	VaR								CVaR							
	τ		ρ		λ_U		λ_L		τ		ρ		λ_U		λ_L	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Model 1	0.0175	0.0145	0.0184	0.0137	0.0177	0.0144	0.0175	0.0144	0.2515	0.3599	0.2592	0.3510	0.2632	0.3551	0.2603	0.3539
Model 2	0.0106	0.0307	*0.0092*	0.0293	*0.0092*	0.0298	*0.0093*	0.0297	0.6942	0.4836	0.6945	0.4816	0.6856	0.4801	0.6891	0.4806
Model 3	0.0111	0.0131	0.0112	0.0130	0.0113	0.0131	0.0121	0.0121	0.2670	0.5157	0.2702	0.5117	0.2741	0.5143	0.2798	0.4927
Model 4	0.0135	0.0139	0.0134	0.0137	0.0133	0.0140	0.0135	0.0137	*0.2708*	0.4323	*0.2670*	0.4306	*0.2679*	0.4359	*0.2696*	0.4351
	Student's t Copula based Criterion															
	τ		ρ		λ_U, λ_L		d.f.		τ		ρ		λ_U,λ_L		d.f.	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Model 1	0.0186	0.0136	0.0186	0.0136	0.0169	0.0149	0.0187	0.0135	0.2606	0.3467	0.2606	0.3467	0.2586	0.3545	0.2593	0.3521
Model 2	*0.0093*	0.0297	*0.0093*	0.0297	0.0097	0.0305	0.0089	0.0285	0.6881	0.4796	0.6881	0.4796	0.6857	0.4812	0.6957	0.4792
Model 3	0.0115	0.0127	0.0115	0.0127	0.0115	0.0127	0.0127	0.0113	0.2755	0.5052	0.2755	0.5052	0.2764	0.5021	0.2818	*0.4714*
Model 4	*0.0134*	*0.0135*	*0.0134*	*0.0135*	0.0136	0.0137	0.0151	0.0121	*0.2681*	*0.4259*	*0.2681*	*0.4259*	*0.2715*	*0.4337*	0.2744	0.4077

Los casos no marcados son aquellos donde el VaR o el CVaR estimados están sesgados. Esto ocurre en el Modelo 1 con dim = 4 utilizando cualquiera de los criterios para seleccionar el *D-vine* y en el Modelo 3, también con dim = 4, excepto para el *CVaR* estimado al 99.5 % con el criterio basado en los grados de libertad de la t de Student. Para los modelos 2 y 4 con dim = 6 los resultados mejoran cuando suponemos mayor dependencia y utilizamos criterios basados en la cópula conocida. Sin embargo, en la práctica la cópula es desconocida y los resultados están basados en criterios estadísticos de bondad de ajuste.

Para evaluar el efecto sobre el ECM cuando utilizamos una cópula diferente a la verdadera, re-calculamos el ECM relativo para los modelos 3 y 4 considerando la cópula de Gumbel. Los resultados se exponen en la Tabla 6.9. En este caso, rechazamos la hipótesis nula de que los valores estimados son iguales a los valores teóricos del VaR y CVaR al 99 % y 99.5 %. El ECM del VaR estimado aumenta, especialmente para los niveles de confianza del 99.5 %. Por otra parte, los resultados para el CVaR muestran que para el nivel de confianza del 99 % el ECM relativo es similar al obtenido con la cópula t de Student, sin embargo, para el nivel de confianza del 99.5 % el ECM relativo aumenta considerablemente.

TABLA 6.9: Raíz cuadrada del ECM dividida por los valores teóricos correspondientes del VaR y CVaR, obtenidos a partir de los diferentes criterios de selección de órdenes utilizando la cópula de Gumbel.

	Empirical Criterion															
	VaR								CVaR							
	τ		ρ		λ_U		λ_L		τ		ρ		λ_U		λ_L	
	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %	99 %	99.5 %
Model 3	0.0129	0.0231	0.0129	0.0231	0.0126	0.0228	0.0126	0.0228	0.2472	0.8422	0.2473	0.8433	0.2426	0.8300	0.2417	0.8337
Model 4	0.0188	0.0313	0.0187	0.0313	0.0178	0.0304	0.0180	0.0305	0.2489	0.9158	0.2481	0.9157	0.2386	0.8792	0.2408	0.8848
	Gumbel Copula based Criterion															
	τ		ρ		λ_U		d.f.		τ		ρ		λ_U			
Model 3	0.0128	0.0230	0.0128	0.0230	0.0128	0.0230			0.2452	0.8345	0.2452	0.8345	0.2452	0.8345		
Model 4	0.0189	0.0313	0.0189	0.0313	0.0189	0.0313			0.2521	0.9146	0.2521	0.9146	0.2521	0.9146		

En aquellos casos en los que se rechaza la hipótesis nula que supone que los valores estimados son iguales a los valores teóricos, debemos asumir cierto sesgo en las estimaciones del riesgo, que puede ser positivo o negativo. En nuestro estudio de simulación, este sesgo es bastante pequeño para el VaR; no llega al 4 % del valor teórico en ningún caso. Sin embargo, en el caso del CVaR el sesgo es mucho más grande. Cuando el nivel de confianza es del 99.5 %, en los casos del Modelo 2 (dim = 6), éste puede exceder el 65 % del valor teórico.

6.5. Conclusiones

Hemos demostrado que el uso de diferentes órdenes en el primer árbol del *D-vine* proporcionan estimaciones distintas del VaR y CVaR. La mayor o menor dispersión de dichas estimaciones depende de la cópula seleccionada, es decir, el modelo de dependencia entre pares tiene un papel fundamental.

La dispersión del VaR y el CVaR estimados plantea el problema acerca de la selección del orden óptimo. La elección basada en la maximización de la pseudo-verosimilitud requiere de la estimación de todos los posibles *D-vines*. Además, esta estimación puede no ser la mejor si el objetivo es estimar el VaR o CVaR. Por tanto, es adecuado utilizar criterios alternativos basados en medidas de dependencia. En nuestro análisis empírico, en general, encontramos que al utilizar diferentes criterios obtenemos una descomposición *pair-copula* diferente; sin embargo, en las medidas estimadas del VaR y CVaR no se evidencian diferencias significativas una vez que se utiliza la cópula que mejor ajusta, en nuestro caso la cópula t de Student.

En el estudio de simulación, tras comparar los modelos con cuatro y seis dimensiones, obtuvimos mejores resultados para el modelo de seis dimensiones. En algunos casos, las estimaciones del VaR y CVaR están sesgadas y este sesgo puede aumentar considerablemente si la cópula no es la cópula verdadera. Debido al sesgo, en general, es aconsejable utilizar diferentes criterios para la selección del *D-vine* y evaluar las diferencias entre las diferentes estimaciones del riesgo.

6.6. Discusión

Tras la realización del contenido de este capítulo se pensó en la posible utilización de esta aproximación a la valoración al riesgo de asegurados cuando la dimensionalidad

viene dada por la existencia de varias pólizas de un mismo ramo o bien de varios ramos. La gran dificultad de cara a la implementación de los resultados de este capítulo es la falta de homogeneidad en el número de dimensiones.

Capítulo 7

Un enfoque de distribución-libre para la cuantificación del riesgo

7.1. Introducción

La cuantificación del riesgo, a menudo, se lleva a cabo en dos pasos: en primer lugar, se fija un nivel de tolerancia y, en segundo lugar, se calcula el Valor en Riesgo (*Value-at-Risk*). El valor en riesgo es la cantidad que no es superada por una pérdida individual con probabilidad igual al nivel de confianza correspondiente. Esta noción corresponde al concepto de cuantil extremo de una distribución estadística, es decir con un nivel de confianza cercano a 1.

En la práctica, los datos provenientes de pérdidas observadas conforman la base para realizar el análisis estadístico que conduce a la cuantificación del riesgo. Sin embargo, los gestores de riesgos necesitan basar dichos análisis en supuestos sobre el comportamiento estadístico de los datos existentes. La estimación del *VaR* desde un enfoque estadístico clásico puede realizarse de tres maneras: i) se puede usar la distribución estadística empírica de la pérdida o una versión suavizada, ii) se puede suponer una distribución Normal o t de Student, o iii) puede asumirse otra aproximación paramétrica (ver McNeil et al., 2005). El tamaño de la muestra es un factor clave para determinar el método de cálculo. En este sentido, el uso de la función de distribución empírica requiere de un tamaño de muestra mínimo. Por su parte, la aproximación normal proporciona una expresión sencilla para el valor en riesgo, pero por desgracia muchas pérdidas están

lejos de tener una distribución con forma normal o incluso con forma t de Student. Alternativamente, es posible encontrar una densidad paramétrica adecuada, a la cual deberían adaptarse los datos de pérdida (ver Klugman et al. (1998)). Además, hay que tener en cuenta que los métodos propuestos por Harrell y Davis (1982) y Sheather y Marron (1990) no son adecuados para estimar el *VaR* (cuantil extremo) cuando las distribuciones tienen fuerte asimetría a la derecha, tal como se mostró en Alemany et al. (2013).

Un enfoque no paramétrico, como la estimación núcleo clásica (*classical kernel estimation - CKE*), alisa la forma de la distribución empírica y "extrapola" su comportamiento cuando se trata de los extremos. Sin embargo, el número de observaciones muestrales en la cola derecha de la distribución es escaso, por esta razón, la estimación núcleo clásica no puede suavizar la forma de la distribución empírica y, por lo tanto, no puede extrapolar la forma de la distribución por encima del valor máximo observado en la muestra. Por esta razón, se propone un procedimiento de estimación en dos pasos. Primero, se ajusta un modelo paramétrico flexible. Segundo, se utiliza un método de estimación núcleo transformada (*transformed kernel estimation - TKE*), asegurando así que el resultado final sea asintóticamente óptimo y garantizando que la forma de la cola derecha se extrapole de forma eficiente y con el mínimo sesgo posible. El método de estimación núcleo transformada está basado en una transformación de los datos originales, de modo que los datos transformados sigan una distribución que pueda estimarse óptimamente con el *CKE*.

En este capítulo presentamos un sistema para cuantificar el riesgo y demostrar que es adecuado estimar cuantiles extremos de una distribución con asimetría positiva y con una cola derecha pesada. A modo de resumen, en las Figuras 7.1 y 7.2 se expone cómo debería implementarse un sistema básico de cuantificación de riesgos de forma clásica y con la metodología propuesta en este trabajo.

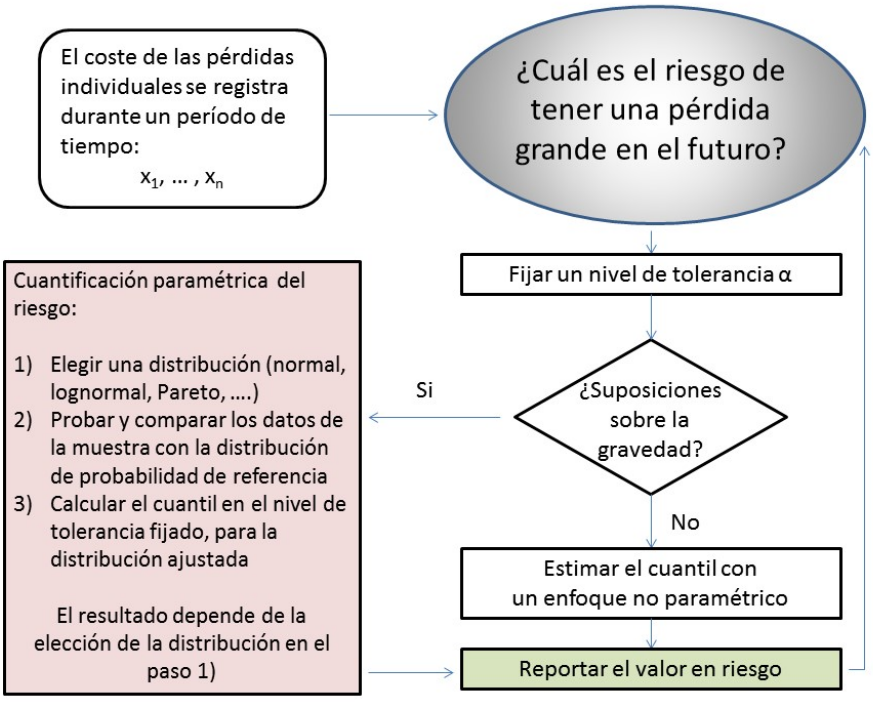


FIGURA 7.1: Descripción de un procedimiento clásico de cuantificación del riesgo

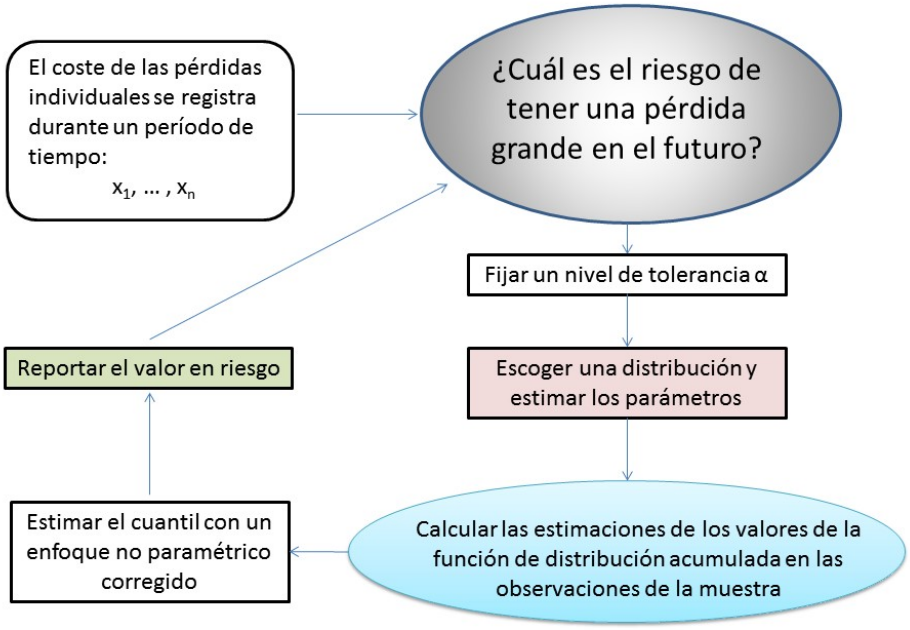


FIGURA 7.2: Sistema de cuantificación del riesgo propuesto, basado en un método no paramétrico

7.2. Notación

Sea X una variable aleatoria de pérdida, con función de distribución acumulada F_X. Por ejemplo, X puede referirse al coste de un fallo operacional o de un accidente. Cuanto mayor sea el coste, mayor será la gravedad del evento de pérdida. Tal como mencionamos anteriormente en la Sección 6.2, el valor en riesgo (*VaR*) también equivale a un cuantil extremo de F_X y se define como:

$$VaR_\alpha(X) = \text{ínf}\{x, F_X(x) \geq \alpha\} = F_X^{-1}(\alpha), \tag{7.2.1}$$

donde el nivel de confianza α es una probabilidad cercana a 1, de modo que calculamos un cuantil en la cola derecha de la distribución.

El VaR_α es el nivel de coste que una proporción α de pérdidas no supera. Así, una fracción de las pérdidas $(1-\alpha)$ excedería ese nivel.

Como estamos interesados en calcular el VaR_α, necesitamos establecer un supuesto respecto al comportamiento estocástico de las pérdidas, o como sugerimos, podríamos estimar F_X al margen de establecer supuestos acerca de la forma de la distribución.

7.3. Estimación no paramétrica del cuantil

7.3.1. Distribución empírica

La estimación del VaR_α es sencilla cuando la F_X en (7.2.1) se reemplaza por la distribución empírica:

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \tag{7.3.1}$$

donde $I(\bullet)$ es la función indicatriz que toma valores 1 ó 0. Si $I(\bullet) = 1$, la condición entre paréntesis es verdadera. Sustituyendo la distribución empírica en la expresión del *VaR* en (7.2.1) obtenemos:

$$VaR_\alpha(X) = \text{ínf}\left\{x, \widehat{F}_n(x) \geq \alpha\right\}. \tag{7.3.2}$$

El sesgo de la distribución empírica es cero y su varianza es:

$$\left(F_X\left(x\right)\left[1-F_X\left(x\right)\right]\right)/n.$$

La distribución empírica es sencilla y es un estimador insesgado de la cdf, pero no puede extrapolarse más allá del máximo observado en la muestra. Esto es particularmente problemático si la muestra no es demasiado grande y se sospecha que podría producirse una pérdida mayor que la pérdida máxima observada en la muestra de datos.

7.3.2. Métodos basados en la estimación núcleo clásica

La estimación núcleo clásica de la función de distribución F_X se obtiene mediante la integración de la estimación nucleo clásica de su función de densidad f_X, definida como:

$$\widehat{f}_X(x) = \frac{1}{nb}\sum_{i=1}^{n} k\left(\frac{x-X_i}{b}\right),$$

donde k es una función de densidad conocida como núcleo de la estimación. Algunos ejemplos de funciones núcleo muy comunes son la Epanechnikov y el núcleo Gaussiano (ver Silverman, 1986). El parámetro b es conocido como parámetro de alisamiento o ancho de banda, el cual controla la suavidad de la estimación. Cuanto mayor sea b, más suave será el resultado final.

Siendo K la función de distribución de k, el estimador núcleo de la función de distribución se obtiene con facilidad.

$$\begin{aligned}\widehat{F}_X(x) &= \int_{-\infty}^{x}\widehat{f}_X(u)du = \int_{-\infty}^{x}\tfrac{1}{nb}\textstyle\sum_{i=1}^{n} k\left(\frac{u-X_i}{b}\right)du \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-X_i}{b}} k\left(t\right)dt = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} K\left(\frac{x-X_i}{b}\right),\end{aligned} \tag{7.3.3}$$

para estimar el VaR_α se utiliza el método de Newton-Raphson para resolver la igualdad

$$\widehat{F}_X\left(\widehat{VaR}_\alpha\left(X\right)\right) = \alpha.$$

La estimación núcleo clásica de una cdf como la definida en (7.3.3) tiene muchas similitudes con la expresión de la "bien conocida" distribución empírica definida en (7.3.1). En la expresión (7.3.3) $K\left(\frac{x-X_i}{b}\right)$ debería reemplazarse por $I(X_i \leq x)$ para obtener (7.3.1). La principal diferencia entre (7.3.1) y (7.3.3) es que la función de distribución empírica solo utiliza datos por debajo de x para obtener la estimación de $F_X(x)$,

mientras que el estimador núcleo clásico de la función de distribución usa todos los datos por encima y por debajo de x, dando más peso a las observaciones que son menores que x y están más cercanas a ese punto. Reiss (1981) y Azzalini (1981) demostraron que, cuando $n \to \infty$, el error cuadrático medio (*Mean squared error - MSE*) de $\widehat{F}_X(x)$ puede aproximarse con:

$$E\left\{\widehat{F}_X(x) - F_X(x)\right\}^2 \sim \frac{F_X(x)[1-F_X(x)]}{n} - f_X(x)\frac{b}{n}\left(1 - \int_{-1}^{1} K^2(t)\,dt\right) + b^4\left(\frac{1}{2}f'_X(x)\int t^2 k(t)\,dt\right)^2. \tag{7.3.4}$$

Los primeros dos términos de (7.3.4) corresponden a la varianza asintótica y el tercer término es el sesgo asintótico al cuadrado. El estimador núcleo de la función de distribución es sesgado, sin embargo tiene menos varianza que la distribución empírica. Notamos que el sesgo tiende a cero si el tamaño de la muestra es grande.

El valor para el parámetro de suavizado b que minimiza (7.3.4) es:

$$b_x^* = \left(\frac{f_X(x)\int K(t)\,[1-K(t)]\,dt}{\left(f'_X(x)\int t^2 k(t)\,dt\right)^2}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}, \tag{7.3.5}$$

donde el subíndice x indica que el parámetro de suavizado es óptimo en este punto. Además, en Azzalini (1981) se demuestra que (7.3.5) también es óptimo cuando se calculan los cuantiles (es decir, el VaR_α). Sin embargo, en la práctica, calcular b_x^* no es sencillo pues depende del valor teórico de $f_X(x)$ y el cuantil x que también es desconocido.

Una alternativa al parámetro de suavizado expresado en (7.3.5) es utilizar la regla basada en la distribución normal (*rule-of-thumb*) propuesta en Silverman (1986). Sin embargo, dado que el objetivo de este capítulo es estimar un cuantil en la cola derecha de una distribución, Alemany et al. (2013) recomiendan calcular el parámetro de alisamiento que minimice asintóticamente el error cuadrado ponderado integrado (*weighted integrated squared error - WISE*), es decir:

$$WISE\left\{\widehat{F}_X\right\} = E\left\{\int\left[F_X(x) - \widehat{F}_X(x)\right]^2 x^2 dx\right\}.$$

El valor de b que minimiza al *WISE* asintóticamente es:

$$b^{**} = \left(\frac{\int f_X(x)\,x^2 dx \int K(t)\,[1-K(t)]\,dt}{\int\left[f'_X(x)\right]^2 x^2 dx\left(\int t^2 k(t)\,dt\right)^2}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \tag{7.3.6}$$

y cuando se reemplaza la densidad teórica f_X por la de la Normal, el parámetro de suavizado estimado es:

$$\hat{b}^{**} = \sigma_X^{\frac{5}{3}} \left(\frac{8}{3}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}. \tag{7.3.7}$$

Existen diversos métodos para calcular b. Por ejemplo, los métodos de validación cruzada y *plug-in* (ver, por ejemplo Bowman et al., 1998; Altman y Leger, 1995) son muy habituales. Sin embargo, estos métodos requieren un esfuerzo computacional considerable cuando se trabaja con una gran cantidad de datos y no proporcionan resultados aceptables cuando la distribución es asimétrica y tiene una cola derecha pesada.

7.3.3. Estimación núcleo transformada

La estimación núcleo transformada es mejor que la estimación de la densidad del núcleo clásica, cuando se estiman distribuciones con asimetría hacia la derecha (ver Bolancé et al., 2003a; Buch-Larsen et al., 2005; Bolancé et al., 2008a; Bolancé, 2010). Incluso, si se dispone de una muestra grande, el número de observaciones en la cola derecha suele ser escaso y las estimaciones no paramétricas estándares son ineficientes para estimar un cuantil extremo.

La estimación núcleo transformada está basada en la aplicación de una transformación a la variable original para que la variable transformada tenga una distribución simétrica. Una vez que la estimación núcleo clásica es implementada sobre los datos transformados, la inversa de la transformación los devuelve a la escala original.

Sea $T(\bullet)$ una transformación cóncava, donde $Y = T(X)$ y $Y_i = T(X_i)$, $i = 1 \ldots n$ son los datos transformados. La estimación núcleo transformada de la función de distribución original es:

$$\widehat{F}_X(x) = \widehat{F}_{T(X)}(T(x)) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{T(x) - T(X_i)}{b}\right) \tag{7.3.8}$$

donde b y K son como los definidos en la Sección 7.3.2.

Cuando estimamos el VaR_α necesitamos resolver la siguiente ecuación:

$$\widehat{F}_{T(X)}(T\left(\widehat{VaR}_\alpha(X)\right)) = \alpha$$

para encontrar $T(X)$ y seguidamente hallar la estimación del $\widehat{VaR}_\alpha$ utilizando la inversa de esta transformación.

El parámetro de suavizado en la estimación núcleo transformada de una función de distribución o cuantil es el mismo que el parámetro de suavizado en la estimación núcleo clásica de la cdf asociada a la variable transformada. Por tanto, podemos calcular el parámetro de ventana en (7.3.7) si reemplazamos σ_X por σ_Y.

Muchos estudios han propuesto transformaciones en el contexto de la estimación núcleo transformada de la función de densidad (vea Wand et al., 1991; Bolancé et al., 2003a; Buch-Larsen et al., 2005; Ruppert y Cline, 1994; Bolancé, 2010). Sin embargo, solo unos pocos estudios analizan la estimación núcleo transformada de la cdf y el cuantil (ver Alemany et al., 2013; Swanepoel y Van Graan, 2005). Estas transformaciones pueden clasificarse en aquellas que son una cdf y aquellas que no corresponden con una cdf específica. Además, también se pueden considerar las transformaciones basadas en una cdf no paramétrica.

Si $T(x)$ es una cdf paramétrica, la estimación núcleo transformada en (7.3.8) puede interpretarse como una estimación paramétrica con una corrección no paramétrica. En Alemany et al. (2013) se ha demostrado que el MSE de TKE es:

$$\begin{aligned} &= E\left\{\hat{F}_X(x) - F_X(x)\right\}^2 \\ &\sim \frac{F_X(x)[1-F_X(x)]}{n} - \frac{1}{T'(x)} f_X(x) \frac{b}{n}\left(1 - \int_{-1}^{1} K^2(t)\,dt\right) \\ &+ \frac{1}{T'(x)}\left(1 - \frac{\frac{f_X(x)}{f'_X(x)}}{\frac{T'(x)}{T''(x)}}\right)^2 \left[\frac{1}{2} f'_X(x) \int_{-1}^{1} t^2 k(t)\,dt\right]^2 b^4. \end{aligned} \tag{7.3.9}$$

En la expresión (7.3.9) es posible observar que cuando $T(x)$ es igual a la cdf teórica $F_X(x)$, el sesgo de la estimación núcleo transformada es cero. Además, si $T(x)$ es diferente de la cdf teórica $F_X(x)$, se observa que TKE es un estimador consistente, si se cumple que $nb \to 0$ cuando $n \to \infty$. Sin embargo, si utilizamos un modelo paramétrico y este no coincide con $F_X(x)$, el estimador paramétrico no es consistente.

El método de estimación núcleo doble transformada (*Double transformed kernel estimation - DTKE*) para estimar el cuantil fue propuesto por Alemany et al. (2013). Primero, se estiman los parámetros de una distribución paramétrica. En nuestro caso, proponemos elegir entre el siguiente conjunto de modelos paramétricos: exponencial, Weibull, Lognormal, distribución de Champernowne y sus generalizaciones. En esta

aplicación, se utiliza la cdf generalizada de la Champernowne [1] dado que proporciona el mejor ajuste para nuestro conjunto de datos reales. Por tanto, en primer lugar los datos son transformados con la cdf de esta distribución. En segundo lugar, los datos transformados se transforman nuevamente utilizando la cdf inversa de una distribución $Beta(3,3)$ definida en el dominio $[-1,1]$ (consulte Alemany et al., 2013; Bolancé et al., 2012, para una explicación más detallada y para los códigos de programación en SAS y R respectivamente) .

El enfoque de doble transformación está basado en el hecho de que la cdf de $Beta(3,3)$ puede ser estimada de forma óptima utilizando la estimación núcleo clásica (ver Terrell, 1990). Luego, dado que los datos doblemente transformados tienen una distribución que es similar a la distribución $Beta(3,3)$, es posible utilizar un parámetro de ventana asintóticamente óptimo para estimar el VaR_α. Los detalles sobre cómo se puede calcular este ancho de banda óptimo se encuentran en Alemany et al. (2013).

7.4. Datos

Analizamos una base de datos, que contiene una muestra de 5122 costes de siniestros vinculados a pólizas de auto. Esta es una base de datos estándar con información acerca de los costes derivados de una accidente automovilístico, es decir, una muestra grande y de cola pesada, con una gran cantidad de valores pequeños y algunos valores extremos grandes. La muestra representa el 10 % de todas las pérdidas que se declaran al departamento de siniestros de auto de la compañía.

Los datos originales se dividen en dos grupos: las reclamaciones de los tomadores con una edad inferior a los 30 años (es decir, tomadores más jóvenes) y las reclamaciones de los asegurados con 30 años de edad o más (asegurados mayores) en el momento de la ocurrencia del accidente que ocasionó la demanda de compensación. El primer grupo está constituido por 1061 observaciones, cuyos costes (en euros) se ubican en un intervalo que va de 1 a 126000 y el segundo grupo está formado por 4061 observaciones cuyos costes se sitúan en un intervalo que va de 1 a 17000. En la Tabla 7.1 se presentan algunos estadísticos descriptivos, donde es posible observar que las distribuciones de las pérdidas para ambos grupos de tomadores presentan asimetría hacia la derecha. Además se observa que la distribución de la severidad de las reclamaciones para los conductores

[1] Una distribución generalizada de Champernowne tiene la siguiente cdf:

$$T_X(x) = \left((x+c)^\gamma - c^\gamma\right) / \left((x+c)^\gamma + (M+c)^\gamma - 2c^\gamma\right). \; c,\gamma,M > 0 \;\; -c < x.$$

más jóvenes presenta una cola más pesada que la asociada con los conductores más mayores (ver Bolancé et al., 2003a).

Para cada grupo de datos de conductores jóvenes y mayores, se estima el VaR_α con α = 0.95 y α = 0.995. El *VaR* es necesario para determinar cuál de los dos grupos posee mayor riesgo en términos de la gravedad de los accidentes, de modo que pueda asignarse una prima mayor a dicho grupo, para ello, se implementan los siguientes métodos no paramétricos: i) La distribución empírica (Emp), como en la expresión (7.3.1), ii) la estimación núcleo clásica de una cdf, como se describe en la Sección 7.3.2 con una ventana basada en la minimización del WISE y iii) la estimación núcleo doble transformada de la cdf, como se describe en la Sección 7.3.3, con una ventana basada en la minimización del MSE en $x = VaR_\alpha$. Además, las funciones núcleo Epanechnikov se usan para obtener el CKE y el DTKE.

TABLA 7.1: Resumen de los costes de los siniestros de automóvil para conductores más jóvenes y mayores

	Jóvenes	Mayores	Todos
Número de observaciones	1061	4061	5122
Media	243.10	402.70	276.10
Mediana	66	68	67
Desviación estándar	3952.30	704.60	1905.50
Máximo	126000	17000	126000

Costes de los siniestros en unidades monetarias.

En la Tabla 7.2 se exponen los valores de las estimaciones del $VaR_{0.95}$ y $VaR_{0.995}$ utilizando las muestras originales. Para α = 0.95, todos los métodos producen valores estimados similares. Sin embargo, con α = 0.995 los resultados difieren de un método a otro. Observamos que, para los conductores mayores, la estimación núcleo clásica produce un resultado del $VaR_{0.995}$ similar a la del cuantil empírico y para los conductores más jóvenes proporciona estimaciones por encima de dicho cuantil empírico.

Los resultados de la Tabla 7.2 muestran que la estimación núcleo doble transformada no subestima el riesgo. Como era de esperar, es un método adecuado para "extrapolar el cuantil extremo" en las zonas de la distribución donde casi no hay información de muestra disponible. El $VaR_{0.995}$ estimado con este método es más alto que el cuantil empírico.

TABLA 7.2: Valor en riesgo estimado (VaR_α) con nivel de tolerancia α para los costes de los siniestros de automóvil

		Jóvenes	Mayores	Todos
α=0.95				
	Emp	1104.00	1000.00	1013.00
	CKE	1293.00	1055.33	1083.26
	DTKE	1257.33	1005.98	1048.51
α=0.995				
	Emp	5430.00	3000.00	4678.00
	CKE	5465.03	4040.40	4695.80
	DTKE	7586.27	4411.11	4864.08

En la Figura 7.3, representamos el VaR_α estimado para una malla de valores α entre 0.99 y 0.999, tanto para conductores jóvenes como para los mayores, todo ello utilizando la distribución empírica, la estimación núcleo clásica y la estimación núcleo doble transformada. Los gráficos de la Figura 7.3 muestran que tanto Emp como CKE son muy similares, es decir, en la zona donde los datos son escasos, la CKE no suaviza a la Emp. En ambos gráficos observamos que el DTKE es una versión más suave que la Emp y el CKE y, por lo tanto, permite la extrapolación del VaR_α más allá del máximo observado en la muestra con una curva suavizada.

La estimación núcleo doble transformada es, en este caso, el método más preciso para estimar cuantiles extremos, tal como se muestra en el enfoque *bootstrap* descrito en el Apéndice D. Por lo tanto, podemos concluir que el DTKE es un método no paramétrico que puede utilizarse para obtener estimaciones del riesgo a niveles de confianza cercanos a 1.

Es evidente que el riesgo de un accidente grave entre el grupo de asegurados más jóvenes es mayor que el estimado para los asegurados mayores.

En consecuencia, la asignación del riesgo debería ser proporcionalmente más alta para este grupo de edad más joven. En otras palabras, los conductores más jóvenes deberían pagar primas de seguro proporcionalmente más altas porque es más probable que se vean involucrados en accidentes graves.

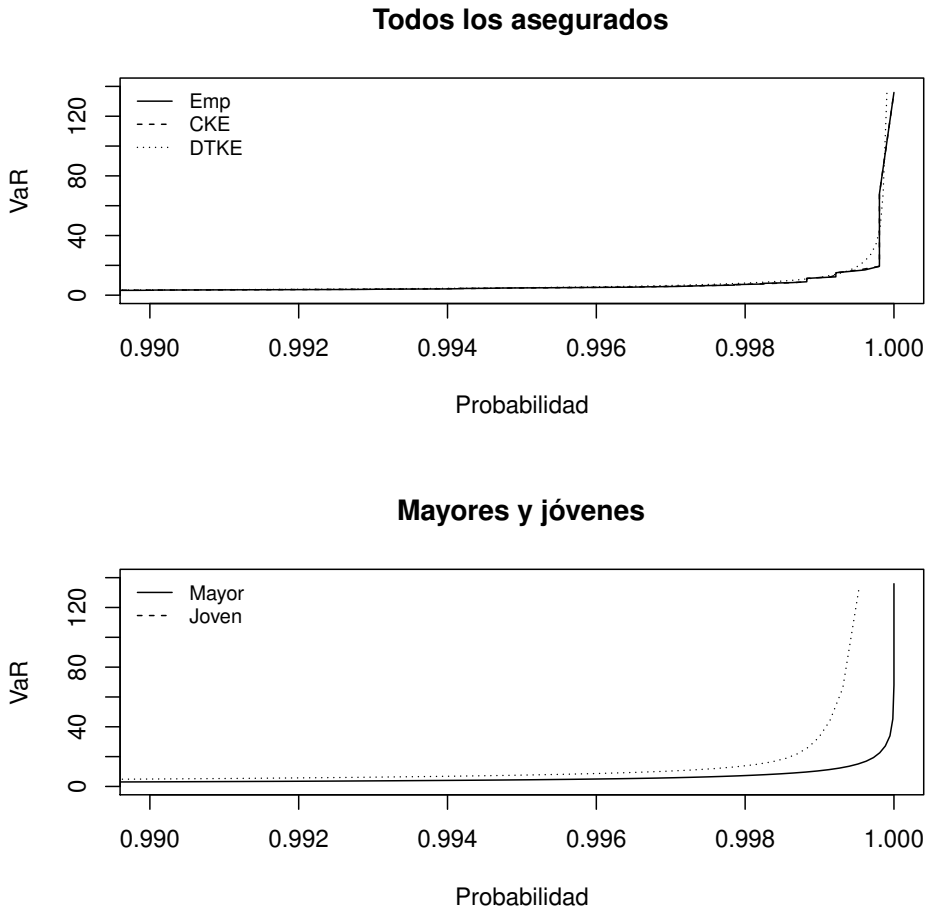


FIGURA 7.3: *VaR* estimado para niveles de confianza superiores al 99 % (eje x). Arriba: Comparación de los tres métodos para todos los asegurados. Las líneas sólidas, discontinuas y punteadas corresponden con los siguientes métodos no paramétricos: la distribución empírica, la estimación núcleo clásica y la estimación núcleo doble transformada, respectivamente. Abajo: El *VaR* estimado con la estimación núcleo doble transformada dado el nivel de confianza. La línea continua y la línea punteada corresponden con los asegurados mayores y más jóvenes, respectivamente.

7.5. Conclusiones

Al analizar la distribución de las pérdidas en una clase de riesgo dada, somos conscientes de que la asimetría hacia la derecha es frecuente. Como resultado, ciertas medidas de riesgo, incluidas la varianza y la desviación estándar, que son útiles para identificar grupos cuando la distribución es simétrica, son incapaces de discriminar cuando se presentan distribuciones que contienen un número de valores extremos poco frecuentes. Alternativamente, las medidas de riesgo centradas en la cola derecha, como los cuantiles, pueden ser útiles para cuantificar el riesgo y para comparar entre diferentes clases de riesgo.

En este capítulo, hemos propuesto un sistema para medir el riesgo a partir de datos asociados con pérdidas, que no requiere alguna hipótesis estadística sobre su distribución. También hemos demostrado que ciertas modificaciones de la estimación núcleo clásica de la cdf, tales como transformaciones, permiten obtener una estimación de la medida de riesgo por encima del máximo observado en la muestra, sin asumir una forma funcional que está estrictamente vinculada a una distribución paramétrica. Dado el pequeño número de valores que normalmente se observan en la cola de una distribución, creemos que nuestro enfoque es un método práctico para la cuantificación del riesgo.

Nuestro método permite establecer una distancia entre las clases de riesgo en términos de las diferencias en el riesgo de severidades extremas. Por lo tanto, una futura línea de investigación es proponer su utilización para valorar el riesgo en determinados segmentos de clientes donde no exista una muestra de gran tamaño a la hora de valorar la severidad de los siniestros observados.

Capítulo 8

Conclusiones

A continuación, se resumen los principales resultados de la Tesis. Adicionalmente, se comentarán algunas de las limitaciones encontradas y, finalmente, se apuntarán observaciones sobre las futuras líneas de investigación.

El Capítulo 1 de este trabajo corresponde con la Introducción. En primer lugar presentamos los antecedentes que justifican la realización de esta tesis. Para ello, comenzamos exponiendo el objetivo general de este trabajo; describimos el proceso de suscripción de una póliza desde un punto de vista económico y estadístico; explicamos la importancia que tiene el cálculo de la prima a pagar tanto para el cliente como para la compañía de seguros; y mostramos cómo bajo el enfoque propuesto es posible satisfacer la necesidad de cuantificación del riesgo en una empresa de seguros. En segundo lugar, se realiza una revisión de la literatura existente. Empezamos hablando acerca del enfoque usual de las contribuciones en seguros de no-vida; a continuación mencionamos algunas de las referencias bibliográficas más relevantes vinculadas al tema de la modelización estadística a nivel actuarial, específicamente nos centramos en temas como: los modelos lineales generalizados, el análisis de riesgos individuales y el análisis de las reclamaciones a través de la modelización de la frecuencia y la severidad; seguidamente detallamos algunas de las implicaciones más importantes en seguros desde la perspectiva del cliente, resaltando la importancia de vincular aspectos como su fidelización, satisfacción, retención - rotación y rentabilidad. Interesados en abordar la problemática acerca de la rotación de clientes, describimos la principal limitación encontrada en éste sentido e introducimos el concepto de "análisis de la dependencia entre líneas de negocio" como un tema "oportunidad", hasta el momento inexplorado y que creemos supondrá un paso adelante hacia el desarrollo del sector asegurador, donde la prioridad se encuentra en el análisis de los diferentes perfiles de riesgo; por último, planteamos como futuras líneas de investigación la modelización conjunta de los costes y la retención de clientes.

En tercer lugar en la Introducción de la tesis, presentamos la descripción detallada de las bases de datos utilizadas. Exponemos las variables que componen cada una de las bases de datos utilizadas en los análisis, describimos la segmentación realizada por tipo de cliente, detallamos los filtros y el proceso de depuración utilizado y finalmente presentamos tablas resumen con las estadísticas descriptivas en cada caso. En cuarto lugar, definimos la metodología clave utilizada para el análisis de los resultados obtenidos con los modelos de retención propuestos. Finalmente, en la última parte de la Introducción, definimos el objetivo general y los objetivos específicos que nos llevan a la realización de esta tesis. Capítulo a capítulo describimos de forma resumida los análisis realizados y damos a conocer las publicaciones vinculadas a cada uno de dichos análisis.

En el Capítulo 2 nos centramos en definir una nueva medida de riesgo que permitiese clasificar a los clientes de una compañía aseguradora según el riesgo que estos aportan a la compañía. Para cada asegurado, se tuvo en cuenta su realidad individual en términos de: (i) número de pólizas contratadas en la línea de negocio de auto, (ii) la propensión a la renovación de dichas pólizas y (iii) la severidad de las reclamaciones declaradas. La construcción de esta medida de riesgo supuso la implementación de modelos de regresión multivariantes, que permitieran, a partir del análisis conjunto de la información de los clientes respecto a las diferentes pólizas contratadas en una misma LoB, cuantificar el riesgo global de cada asegurado bajo el supuesto de independencia y dependencia entre las pólizas contratadas. Esta nueva medida, centrada en los cuantiles extremos de la distribución de los costes, es decir, el valor en riesgo, permite clasificar a los clientes en función del riesgo a nivel monetario que suponen para la compañía. Éste, además de ser un enfoque distinto al usual, donde por lo general se trabaja en base a las pérdidas esperadas ("*Expected losses*"), representa una alternativa novedosa para el cálculo del riesgo del cliente, que permitiría tener en cuenta la dependencia existente entre las pólizas que tiene contratadas en una misma compañía. Adicionalmente, la metodología utilizada en éste Capítulo es totalmente extrapolable a otras áreas donde sea necesaria la cuantificación del riesgo de clientes.

En el Capítulo 3, analizamos la deserción de los clientes para una línea de negocio. Para ello, realizamos una análisis comparativo entre el poder predictivo de tres modelos de aprendizaje de máquinas o *machine learning* versus un modelo clásico de regresión univariante. Los resultados obtenidos nos permitieron evidenciar cómo la elección del mejor modelo para la estimación de las probabilidades de renovación de una póliza de seguros depende del criterio de ajuste utilizado para la comparación entre modelos y, a su vez, de las preferencias y necesidades del negocio en un momento particular. Los criterios propuestos en esta parte representan una herramienta sencilla que facilita la

elección del *threshold* o *cut-off*, a partir del cual y, en la práctica, es posible clasificar a los clientes en función de la propensión a la renovación de sus pólizas.

A continuación, en el Capítulo 4 nos centramos en el análisis de las probabilidades de deserción bajo el supuesto de dependencia entre líneas de negocio. Aquí, la dependencia entre ellas se modeliza de dos formas distintas. Primero, imponiendo normalidad conjunta entre los errores de las variables latentes y luego asumiendo funciones de distribución marginales probit introducimos dependencia normal y no-normal entre líneas de negocio, a través, del uso de la cópula Gaussiana y t-Student respectivamente. Con el uso de las cópulas para la modelización de la dependencia pudimos detectar nuevas variables relevantes para el análisis de la deserción de los asegurados. En éste sentido, la combinación de información del cliente a nivel global mejora la capacidad predictiva de los modelos de retención y, por ende, la valoración del riesgo del cliente. Por tanto, teniendo en cuenta que podrían incorporarse más líneas de negocio al análisis y, con ello, más información, es muy recomendable que los modelos utilizados dentro de las aseguradoras puedan escalar a otro nivel de uso más general, que el de solo los informes específicos e individualizados por tipos concretos de pólizas.

En el Capítulo 5 presentamos algunas reflexiones sobre el reto que ha supuesto, para la sociedad en general y las aseguradoras en particular, la digitalización a gran escala de la información. En éste sentido, hacemos énfasis en la inminente necesidad de un cambio de paradigma que facilite la adaptación al entorno dinámico en el que vivimos. A partir de aquí, damos a conocer la visión a nivel empresarial y técnico de lo que supone la incorporación del *Analytics* en una compañía de seguros. Esbozamos las posibles limitaciones que en la práctica supondría un cambio cultural como el planteado. También, proponemos algunas ideas para la gestión eficiente del cambio. Por último, mostramos su inherente complejidad mediante un caso práctico ya testado con éxito, que reproduce lo expuesto en los tres capítulos anteriores.

En la parte II de este trabajo, presentamos dos estudios vinculados con la cuantificación del riesgo. En el Capítulo 6 se realiza una análisis detallado acerca de cómo se ve afectado el valor en riesgo estimado para diferentes carteras de activos financieros. Una de las conclusiones más relevantes es que la selección del modelo de dependencia entre pares de variables para la modelización del riesgo es fundamental. Esta aplicación es fácilmente extrapolable al sector seguros, donde por ejemplo sería útil para la modelización conjunta de impagos, costes siniestrales y deserción de clientes teniendo en cuenta diferentes líneas de negocio, como lo son auto y hogar. Por su parte, en el Capítulo 7, presentamos una metodología sencilla para la cuantificación del riesgo en

distribuciones con valores extremos. Esta metodología representa una alternativa para lo que en seguros comúnmente es conocido como *large claims*.

Finalmente, creemos que analizar el posible impacto que tendría la aplicación de los modelos presentados en este trabajo, pero ahora incluyendo por ejemplo la rentabilidad de la cartera y además añadiendo el coste que supone para la compañía la decisión de renovación de cada cliente, podría ser un área de análisis interesante, que aportaría valor al negocio y permitiría establecer propuestas relevantes cuya implementación supondría un beneficio tanto para los asegurados como para la aseguradora.

En este estudio se han explorado, modelizado y analizado las relaciones de dependencia existentes entre las pólizas de un mismo cliente individual desde la doble perspectiva de siniestralidad y deserción. Según hemos comprobado en la bibliografía existente este enfoque de análisis conjunto es un ámbito poco explorado en el mercado asegurador español. Así que se espera que esta Tesis, ayude a las compañías de seguros a reinventar la forma de cuantificar el riesgo de sus clientes, en especial, el de aquellos que tienen múltiples riesgos contratados con la compañía. También se espera que, entre otras cosas, sirva como motivación para dar apertura a una nueva era donde el análisis global de la información del asegurado se incorpore dentro de los procesos usuales de negocio, simplifique los procesos de creación, extracción y tratamiento de las bases de datos, favorezca - desde el inicio de la relación contractual - la venta cruzada de productos y permita la toma de decisiones estratégicas más asertivas respecto a la selección de mejores riesgos.

Limitaciones

Entre las principales limitaciones podemos mencionar las siguientes:

- Las dependencias entre la retención y el coste de los siniestros no se han abordado en este trabajo.
- La variable número de siniestros, que es básica en tarificación, no se ha tenido en cuenta en nuestros análisis. Por tanto, nuestro análisis está limitado en este sentido.
- En la actualidad, con la gran cantidad de información disponible gracias a la telemática (*telematics*), sería posible enriquecer los análisis realizados en éste trabajo. Sin embargo, nos vimos limitados al uso de variables clásicas, dado que no disponíamos de información a nivel de uso de los automóviles (kilómetros recorridos, zonas de conducción, velocidad media de conducción, etc).

Implementación

Los resultados obtenidos en éste estudio son implementables a nivel operativo en cualquier compañía de seguros.

Futuras líneas de investigación

Una pregunta frecuente a nivel negocio es: ante todas las opciones de modelización disponibles, ¿qué modelo escoger para el cálculo de las probabilidades de deserción?. Como respuesta podemos afirmar que es posible implementar tantos modelos como se desee. Respecto a la elección acerca de la probabilidad estimada, por ejemplo, se podría asignar a cada cliente la probabilidad de renovación más frecuente a partir de los resultados obtenidos con cada modelo. Otra opción, es utilizar el modelo que mejor ajusta cada vez que se evalúe la renovación de las pólizas de una cartera o bien utilizar la probabilidad media obtenida por distintas aproximaciones, por ejemplo, siguiendo la aproximación de los *random forests*.

Una propuesta a futuro, es establecer un proceso sistemático de valoración de clientes bajo un enfoque multiproducto, que facilite la toma de decisiones estratégicas en la compañía, al combinar la clasificación de riesgo individual del cliente vs la propensión a la cancelación de su(s) póliza(s).

Apéndice A

Estadísticas descriptivas de los datos del Capítulo 2

TABLA A.1: Estadísticas descriptivas - Variables cuantitativas

Variable	N	Media	STD	Mínimo	Máximo
client_age	18656	52.85	13.42	18.00	90.00
firstdriver_age	18656	52.83	13.21	20.00	90.00
firstdriver_agelicense	18656	29.50	12.22	1.00	70.00
veh_guilty	18656	0.87	1.17	0.00	10.00
pol_newpremium	18656	368.05	179.85	104.34	2853.28
pol_lastrenewal	18656	373.94	186.22	99.12	2968.86
mediator_currentpol	18656	2062.15	2648.94	2.00	15381.00
mediator_cancelpol	18656	5409.90	6793.16	0.00	36159.00
veh_power	18656	108.61	37.45	45.00	300.00
veh_age	18656	10.95	5.24	0.00	28.00
veh_weightpower	18656	12.29	2.67	0.00	26.00
pol_supplements	18656	2.29	2.10	0.00	27.00
pol_diffpremium	18656	-5.90	80.43	-1219.36	1881.46
pol_malus	18656	-186.61	126.39	-1992.90	475.08
pol_surcharge	18656	-110.45	142.96	-2749.73	767.55
pol_malus_last	18656	-191.29	130.58	-1943.93	554.26
pol_surcharge_last	18656	-93.34	133.25	-2247.75	724.45
pol_age	18656	5.24	4.19	1.00	44.00
exp	18656	0.80	0.23	0.33	1.00

*Hace referencia a las pólizas de auto vigentes al momento del estudio

TABLA A.2: Estadísticas descriptivas - Variables categóricas

Variable	**Categoría**	**Frecuencia (%)**
client_sex	Mujer	32.63
	Varón	67.37
pol_guarantees	O	5.11
	TRCF	15.65
	TRSF	4.85
	T	74.39
pol_other	0	81.06
	1	18.94
veh_seats	2	0.61
	4	5.31
	5	86.63
	6	0.21
	7	7.25
veh_fueltype	D	57.78
	G	41.84
	O	0.38
mediator_type	A	60.04
	C	39.96
veh_seconddriver	No	88.42
	Si	11.58
pol_waytopay	A	80.43
	S	16.75
	T	2.82

*Hace referencia a las pólizas de auto vigentes al momento del estudio

TABLA A.3: Descriptivas de las variables agregadas a nivel cliente

Variable	**N**	**Media**	**STD**	**Mínimo**	**Máximo**
npol_motor_ag	17212	1.08	0.29	1.00	3.00
nclaims_motor_ag	3264	1.19	0.47	1.00	6.00
cost_motor_ag	3264	2141.35	18399.08	-656.69	1025319.88

TABLA A.4: Resumen de las variables categóricas en función de la siniestralidad

Variable	**Categoría**	**Total (%)**	**Frecuencia (%)**	**Severidad (media)**
client_sex	Mujer	33.43	6.34	2783.59
	Varón	66.57	12.62	1818.90
pol_other	0	80.70	15.30	2189.40
	1	19.30	3.66	1940.46
npol_motor_ag	1	86.37	16.38	2139.07
	2	12.68	2.41	1990.24
	3	0.95	0.18	4366.35

Apéndice B

Estadísticas descriptivas de los datos del Capítulo 4

TABLA B.1: Estadísticas descriptivas - Variables cuantitativas Auto

Variable	N	Media	STD	Mínimo	Máximo
client_age	22106	57.37	13.21	18.00	91.00
firstdriver_age	22106	57.20	13.16	22.00	90.00
firstdriver_agelicense	22106	34.06	11.80	1.00	71.00
veh_guilty	22106	0.90	1.20	0.00	12.00
pol_newpremium	22106	388.98	192.18	111.89	2178.43
pol_lastrenewal	22106	400.22	200.03	11.83	2773.63
client_mcurrentpol	22106	1.20	0.79	0.00	21.00
mediator_currentpol	22106	1911.91	2180.87	2.00	15381.00
mediator_cancelpol	22106	5022.39	5720.97	1.00	36159.00
veh_power	22106	113.39	38.40	45.00	300.00
veh_age	22106	10.74	5.17	0.00	37.00
veh_weightpower	22106	12.02	2.63	0.00	26.00
veh_seats	22106	5.09	0.65	2.00	7.00
pol_supplements	22106	2.72	2.24	0.00	31.00
pol_diffpremium	22106	-11.24	86.63	-1473.06	1418.05
pol_malus	22106	-200.46	139.26	-1770.93	727.73
pol_surcharge	22106	-144.22	167.44	-2769.21	677.19
pol_malus_last	22106	-208.57	144.39	-1748.29	1020.62
pol_surcharge_last	22106	-124.22	157.83	-3467.21	666.01
pol_age	22106	6.02	4.60	1.00	45.00
exp	22106	0.84	0.21	0.33	1.00

*Hace referencia a las características de las pólizas de auto

TABLA B.2: Estadísticas descriptivas - Variables categóricas Auto

Variable	**Categoría**	**Frecuencia (%)**
client_sex	Mujer	26.31
	Varón	73.69
pol_guarantees	O	5.23
	TRCF	20.06
	TRSF	6.18
	T	68.53
pol_other	0	3.89
	1	96.11
veh_seats	2	0.64
	4	5.13
	5	86.08
	6	0.19
	7	7.96
veh_fueltype	D	58.15
	G	41.42
	O	0.43
mediator_type	Agente	70.48
	Corredor	29.52
veh_seconddriver	No	89.55
	Si	10.45
pol_waytopay	A	83.84
	S	13.55
	T	2.61

*Hace referencia a las características de las pólizas de auto

TABLA B.3: Estadísticas descriptivas - Variables cuantitativas Hogar

Variable	**N**	**Media**	**STD**	**Mínimo**	**Máximo**
client_age	22106	57.37	13.21	18.00	91.00
pol_capi_continent	22106	128608.81	111954.70	300.00	2669437.00
pol_capi_content	22106	36747.14	34868.16	50.00	1220303.34
home_guilty	22106	0.85	1.11	0.00	15.00
pol_newpremium	22106	253.58	181.71	31.09	5591.88
pol_lastrenewal	22106	250.12	181.78	51.59	5401.48
client_hcurrentpol	22106	1.02	0.42	0.00	25.00
mediator_currentpol	22106	1763.11	1617.76	1.00	9445.00
mediator_cancelpol	22106	4644.83	4629.76	0.00	31919.00
pol_supplements	22106	1.76	1.88	0.00	19.00
pol_diffpremium	22106	3.46	36.10	-1997.23	726.18
pol_malus	22106	-11.28	31.65	-897.10	0.00
pol_surcharge	22106	-17.26	56.89	-1479.90	2351.31
pol_malus_last	22106	-11.13	31.57	-875.54	0.00
pol_surcharge_last	22106	-18.19	57.08	-2247.04	2255.60
pol_age	22106	7.17	5.30	1.00	29.00
exp	22106	0.86	0.20	0.33	1.00

*Hace referencia a las características de las pólizas de hogar

TABLA B.4: Estadísticas descriptivas - Variables categóricas Hogar

Variable	**Categoría**	**Frecuencia (%)**
client_sex	Mujer	26.31
	Varón	73.69
home_type	A	2.61
	B	0.28
	C	61.76
	D	2.34
	G	33.01
pol_other	0	4.93
	1	95.07
mediator_type	Agente	70.45
	Corredor	29.55
pol_waytopay	A	93.43
	S	6.02
	T	0.55

*Hace referencia a las características de las pólizas de hogar

Apéndice C

Descomposición *pair-copula*

Comenzamos con la definición de la cópula C. Sea $X = (X_1, ..., X_k)$ un vector aleatorio multivariante de k variables continuas y F su función de distribución acumulada conjunta (*cumulative distribution function - cdf*), que contiene información sobre el comportamiento individual de las variables y su estructura de dependencia. La compleja estructura de dependencia no lineal que a menudo existe entre las pérdidas o ganancias, hace que sea necesario buscar enfoques que proporcionen una manera de separar la estructura de dependencia del comportamiento marginal. Las cópulas proporcionan una manera fácil de hacer esto.

Sklar (1959), pionero en la definición de cópulas, a través de su famoso teorema de Slark, demuestra la existencia de una cópula C, a través de la cual es posible establecer una relación funcional entre una función de distribución multivariante y sus funciones de distribución marginales. Entonces, dada F la cdf conjunta asociada al vector aleatorio k-dimensional X, con marginales univariantes $F_1, \ldots, F_k$, existe una cópula C tal que:

$$F(x_1, \ldots, x_k) = C(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)). \tag{C.1}$$

Si denotamos $u_j = F_j(x_j)$, $j = 1, \ldots, k$, a los valores provenientes de una Uniforme$[0, 1]$, entonces es posible definir a la cópula C como:

$$C(u_1, \ldots, u_k) = F(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_k)), \tag{C.2}$$

donde F_j^{-1}, $j = 1, \ldots, k$, es la inversa de la función de distribución asociada con la marginal j. Además, si las $F_j(x_j)$, $j = 1, \ldots, k$, son contínuas C es única.

A partir de la expressión (C.1) es posible deducir tras derivar, la función de densidad de probabilidad conjunta (*probability density function - pdf*)

$$f(x_1,\ldots,x_k) = c_{1,\ldots,k}(F_1(x_1),\ldots,F_k(x_k)) \cdot f_1(x_1)\ldots f_k(x_k), \tag{C.3}$$

para alguna cópula de densidad $c_{1,\ldots,k}$ k-variante.

Por otro lado, la pdf conjunta puede ser definida, por ejemplo, en términos de la siguiente descomposición:

$$\begin{aligned} f(x_1,\ldots,x_k) &= f_k(x_k) \cdot f_{k-1|k}(x_{k-1}|x_k) \cdot f_{k-2|k-1,k}(x_{k-2}|x_{k-1},x_k)\ldots f_{1|2,\ldots,k}(x_1|x_2,\ldots,x_k) \\ &= f_k(x_k) \cdot \prod_{j=1}^{k-1} f_{j|j+1,\ldots,k}(x_j|x_{j+1},\ldots,x_k), \end{aligned} \tag{C.4}$$

donde $f_{A|B}(A|B)$ es la función de densidad condicional de A dado B. Sabemos que la función de densidad condicional puede expresarse como:

$$f_{j|j+1,\ldots,k}(x_j|x_{j+1},\ldots,x_k) = \frac{f_{j,j+1|j+2,\ldots,k}(x_j,x_{j+1}|x_{j+2},\ldots,x_k)}{f_{j+1|j+2,\ldots,k}(x_{j+1}|x_{j+2},\ldots,x_k)}. \tag{C.5}$$

cada descomposición similar a la expuesta en (C.4) supone una estructura distinta de dependencia que define la interrelación entre variables.

Usando el teorema de Sklar y calculando la derivada para obtener la función de densidad de probabilidad, es posible re-definir la expresión (C.5) en términos de cópulas:

$$\begin{aligned} & f_{j|j+1,\ldots,k}(x_j|x_{j+1},\ldots,x_k) = \\ &= c_{j,j+1|j+2,\ldots,k}(F_{j|j+2,\ldots,k}(x_j|x_{j+2},\ldots,x_k), F_{j+1|j+2,\ldots,k}(x_{j+1}|x_{j+2},\ldots,x_k)) \\ &\cdot f_{j|j+2,\ldots,k}(x_j|x_{j+2},\ldots,x_k). \end{aligned} \tag{C.6}$$

Reemplazando recursivamente (C.6) obtenemos que:

$$\begin{aligned} & f_{j|j+1,\ldots,k}(x_j|x_{j+1},\ldots,x_k) = \\ &= c_{j,j+1|j+2,\ldots,k}(F_{j|j+2,\ldots,k}(x_j|x_{j+2},\ldots,x_k), F_{j+1|j+2,\ldots,k}(x_{j+1}|x_{j+2},\ldots,x_k)) \\ &\cdot c_{j,j+2|j+3,\ldots,k}(F_{j|j+3,\ldots,k}(x_j|x_{j+3},\ldots,x_k), F_{j+2|j+3,\ldots,k}(x_{j+2}|x_{j+3},\ldots,x_k)) \\ &\cdot \ldots \\ &\cdot c_{j,k}(F_j(x_j),F_k(x_k)) \cdot f_j(x_j), \end{aligned} \tag{C.7}$$

simplificando la notación como $c_{A,B}(F_A(x_A),F_B(x_B))) = c_{A,B}$ y $f_A(x_A) = f_A$, con $A \neq B$, entonces:

$$f_{j|j+1,\ldots,k}(x_j|x_{j+1},\ldots,x_k) = \\ = f_j \cdot \prod_{i=1}^{k-j} c_{j,j+i|j+i+1,\ldots,k},\ \forall j = 1,\ldots,k-1. \tag{C.8}$$

Sustituyendo el resultado (C.8) en (C.4) obtenemos una posible factorización de la distribución conjunta expresada en términos de cópulas bivariadas. La pdf conjunta es:

$$f(x_1,\ldots,x_k) = \prod_{j=1}^{k-1}\prod_{i=1}^{k-j} c_{j,j+i|j+i+1,\ldots,k} \prod_{l=1}^{k} f_l \tag{C.9}$$

y la cdf conjunta es:

$$F(x_1,\ldots,x_k) = \prod_{j=1}^{k-1}\prod_{i=1}^{k-j} C_{j,j+i|j+i+1,\ldots,k}. \tag{C.10}$$

En general, la expresión (C.10) representa una distribución multivariante en función de $k(k-1)/2$ parámetros de dependencia.

D-vine

Un *D-vine* es una secuencia de m árboles que denotamos como $T_1,\ldots,T_m$. Cada árbol tiene n nodos y las asociaciones entre los nodos de cada árbol se denominan "*edges*". En este caso, para cada árbol con n nodos tenemos $n-1$ *edges* y por cada *edge* surge una *pair-copula*.

El *D-vine* es un caso especial de *R-vine* en el que es necesario elegir el orden de las variables (nodos) en el árbol superior, T_1. A continuación, se presenta un ejemplo de *D-vine* para el caso de seis dimensiones

$$\begin{aligned} f(x_1,x_2,x_3,x_4,x_5,x_6) = & \underbrace{f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \cdot f_6}_{\text{nodes in } T_1} \cdot \underbrace{c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{45} \cdot c_{56}}_{\text{edges in } T_1 - \text{nodes in } T_2} \\ & \cdot \underbrace{c_{13|2} \cdot c_{24|3} \cdot c_{35|4} \cdot c_{46|5}}_{\text{edges in } T_2 - \text{nodes in } T_3} \cdot \underbrace{c_{14|23} \cdot c_{25|34} \cdot c_{36|45}}_{\text{edges in } T_3 - \text{nodes in } T_4} \\ & \cdot \underbrace{c_{15|234} \cdot c_{26|345}}_{\text{edges in } T_4 - \text{node in } T_5} \cdot \underbrace{c_{16|2345}}_{\text{node in } T_5} \cdot \end{aligned} \tag{C.11}$$

El resultado de la cdf conjunta en seis dimensiones es:

$$
\begin{aligned}
&F(x_1,x_2,x_3,x_4,x_5,x_6) = \\
&C_{12}(F_1(x_1),F_2(x_2)) \cdot C_{23}(F_2(x_2),F_3(x_3)) \cdot C_{34}(F_3(x_3),F_4(x_4)) \\
&\cdot C_{45}(F_4(x_4),F_5(x_5)) \cdot C_{56}(F_5(x_5),F_6(x_6)) \\
&\cdot C_{13|2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2)) \cdot C_{24|3}(F_{2|3}(x_2|x_3),F_{4|3}(x_4|x_3)) \\
&\cdot C_{35|4}(F_{3|4}(x_3|x_4),F_{5|4}(x_5|x_4)) \cdot C_{46|5}(F_{4|5}(x_4|x_5),F_{6|7}(x_6|x_5)) \\
&\cdot C_{14|23}(F_{1|23}(x_1|x_2,x_3),F_{4|23}(x_4|x_2,x_3)) \\
&\cdot C_{25|34}(F_{2|34}(x_2|x_3,x_4),F_{5|34}(x_5|x_3,x_4)) \\
&\cdot C_{36|45}(F_{3|45}(x_3|x_4,x_5),F_{6|45}(x_6|x_4,x_5)) \\
&\cdot C_{15|234}(F_{1|234}(x_1|x_2,x_3,x_4),F_{5|234}(x_5|x_2,x_3,x_4)) \\
&\cdot C_{26|345}(F_{2|345}(x_2|x_3,x_4,x_5),F_{6|345}(x_6|x_3,x_4,x_5)) \\
&\cdot C_{16|2345}(F_{1|2345}(x_1|x_2,x_3,x_4,x_5),F_{6|2345}(x_6|x_2,x_3,x_4,x_5)).
\end{aligned} \tag{C.12}
$$

Cada *pair-copula* $C_{A|B}$ requiere de al menos un parámetro $\theta_{A|B}$ que tiene que estimarse. Al final, tendremos tantos parametros estimados como *edges* más uno exista en el *D-vine*, en nuestro caso $6(6-1)/2 = 15$ parámetros de *pair-copula*.

Por otro lado, sabemos que las cdfs condicionales pueden calcularse como:

$$
F_{j|B} = \frac{\partial C_{j,i|B}(F_{j|B}(x_j|B),F_{i|B}(x_i|B))}{\partial F_{i|B}(x_i|B)}, \tag{C.13}
$$

donde B es el grupo de variables que condicionan diferentes a i y j. Además, sabemos que para dos variables:

$$
F_{j|i} = \frac{\partial C_{j,i}(F_j(x_j),F_i(x_i))}{\partial F_i(x_i)}. \tag{C.14}
$$

Entonces, si podemos aproximar las cdfs univariantes, podemos obtener cdfs condicionales recursivamente comenzando con (C.14) y, por lo tanto, podemos expresar la cópula y su densidad en funciones de estas derivadas parciales.

Resultados adicionales

TABLA C.1: Compañías en las carteras de activos.

	Stock 1 (R1)	**Stock 2 (R2)**	**Stock 3 (R3)**	**Stock 4 (R4)**	**Stock 5 (R5)**	**Stock 6 (R6)**
P1	Petrobras	Google	Coca-cola	General Motors	DirecTV	Moody's Corporation
P2	Petrobras	Google	Coca-cola	Procter & Gamble	Telefónica Brasil	Telecom Italia
P3	Merck	Novartis	Pfizer	Deutsche Bank	ING	Santander
P4	Merck	Novartis	Pfizer	Sanofi	GlaxoSmithKline	AstraZeneca

TABLA C.2: Modelos ARMA(P,Q)-GARCH(p,q) ajustados para los rendimientos de cada serie temporal y medidas de dispersión de cada rendimiento filtrado de las series temporales.

Activos	**ARMA**	**GARCH**	**STD**	**IQR**	**Rango**
Petrobras	(0,0)	(0,0)	0.024	0.027	0.256
Google	(0,0)	(0,0)	0.016	0.015	0.217
Coca-cola	(1,1)	(1,1)	0.010	0.011	0.096
General Motors	(0,0)	(1,1)	0.021	0.024	0.206
DirecTV	(0,0)	(1,1)	0.014	0.015	0.142
Moody's Corporation	(0,0)	(1,1)	0.021	0.019	0.224
Procter & Gamble	(0,0)	(1,1)	0.009	0.010	0.100
Telefónica Brasil	(0,0)	(0,0)	0.018	0.020	0.167
Telecom Italia	(0,0)	(1,1)	0.027	0.031	0.215
Merck	(0,0)	(1,1)	0.012	0.012	0.114
Novartis	(1,0)	(1,1)	0.011	0.012	0.107
Pfizer	(1,1)	(1,1)	0.012	0.012	0.101
Sanofi	(0,0)	(1,1)	0.016	0.018	0.140
GlaxoSmithKline	(0,0)	(1,1)	0.011	0.013	0.101
AstraZeneca	(0,0)	(1,1)	0.012	0.013	0.135
Deutsche Bank	(1,1)	(1,1)	0.029	0.030	0.288
ING	(1,1)	(1,1)	0.018	0.020	0.167
Santander	(0,0)	(1,1)	0.025	0.027	0.219

Fuente: Datos diarios obtenidos de Yahoo Finance entre Enero del 2011 y Diciembre del 2013.

TABLA C.3: Parámetros de los modelos utilizados en el estudio de simulación.

	Parámetro	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
θ_{12}	ρ	0.402	0.408	0.469	0.474
	d.f.	7.8	8.9	11.2	11.8
θ_{23}	ρ	0.431	0.433	0.514	0.515
	d.f.	12.2	13.1	7.1	6.9
θ_{34}	ρ	0.413	0.408	0.562	0.559
	d.f.	9.2	8.6	6.7	5.8
θ_{45}	ρ		0.424		0.581
	d.f.		7.0		5.8
θ_{56}	ρ		0.491		0.668
	d.f.		11.1		5.3
$\theta_{13\|2}$	ρ	0.200	0.209	0.514	0.513
	d.f.	61.1	57.7	4.2	4.2
$\theta_{24\|3}$	ρ	0.367	0.366	0.522	0.523
	d.f.	21.0	22.2	9.7	9.8
$\theta_{35\|4}$	ρ		0.362		0.234
	d.f.		10.4		8.9
$\theta_{46\|5}$	ρ		0.451		0.441
	d.f.		18.1		8.3
$\theta_{14\|23}$	ρ	0.255	0.257	0.151	0.151
	d.f.	24.9	22.7	19.0	19.1
$\theta_{25\|34}$	ρ		0.170		0.271
	d.f.		96.9		11.1
$\theta_{36\|45}$	ρ		0.259		0.181
	d.f.		37.4		62.6
$\theta_{15\|234}$	ρ		0.198		0.077
	d.f.		12.9		83.687
$\theta_{26\|345}$	ρ		0.185		0.191
	d.f.		60.6		92.7
$\theta_{16\|2345}$	ρ		0.166		0.148
	d.f.		12.5		80.1

TABLA C.4: Valores teóricos del VaR y CVaR obtenidos a partir de los modelos teóricos, en el estudio de simulación.

	Cuantil	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
VaR	99 %	3.764	5.582	3.841	5.720
	99.50 %	3.844	5.524	3.905	5.827
CVaR	99 %	0.090	0.158	0.069	0.116
	99.50 %	0.061	0.300	0.042	0.074

Apéndice D

Cuantificación del riesgo - estudio de simulación

Para analizar la precisión de los diferentes métodos, generamos 1000 muestras aleatorias *bootstrap* de los costes de los asegurados más jóvenes y mayores. Cada muestra aleatoria tiene el mismo tamaño que la muestra original, pero las observaciones se eligen con reemplazo, así que algunas pueden estar repetidas y algunas pueden estar excluidas. Estimamos el VaR_α para cada muestra *bootstrap*. En la Tabla D.1 se expone la media y el coeficiente de variación (CV). El coeficiente de variación se utiliza para comparar la precisión, dado que las estimaciones no paramétricas, excepto para la estimación empírica, tienen algún sesgo cuando el tamaño de la muestra es finito. La media y el CV del VaR_α estimado para las muestras *bootstrap*, con $\alpha = 0.95$ y $\alpha = 0.995$, se muestran tanto para los costes de las reclamaciones de los conductores más jóvenes como para los costes de los conductores mayores y de forma global para todos los conductores juntos.

La distribución empírica supone que la máxima pérdida posible es el máximo observado en la muestra. Sin embargo, como la muestra es finita y los valores extremos son escasos, estos valores extremos pueden no proporcionar una estimación precisa del VaR_α. Entonces, necesitamos "extrapolar el cuantil", es decir, necesitamos estimar el VaR_α en una zona de la distribución donde casi no tenemos información de la muestra. En la Tabla D.1 observamos que las muestras *bootstrap* son similares para todos los métodos en $\alpha = 0.95$, pero difieren cuando $\alpha = 0.995$. Además, si analizamos los coeficientes de variación observamos que, para los asegurados más jóvenes, los dos métodos núcleo son más precisos que la estimación empírica.

Dado que las medias obtenidas en las estimaciones del VaR_α para los conductores más jóvenes son mayores que las medias obtenidas para los conductores mayores, concluimos que los conductores más jóvenes tienen una distribución con una cola más pesada que la presentada por los asegurados más antiguos. Para los conductores mayores, y de

TABLA D.1: Resultados de la simulación *bootstrap* para la estimación del valor en riesgo, con el nivel de tolerancia α, en los datos de los costes de las reclamaciones vinculados a pólizas de auto

		Jóvenes		**Mayores**		**Todos**	
		Media	CV	Media	CV	Media	CV
α=0.95	Emp	1145.02	0.124	1001.57	0.040	1021.92	0.034
	CKE	1302.19	0.104	1060.24	0.051	1086.88	0.045
	DTKE	1262.58	0.105	1008.28	0.054	1049.64	0.045
α=0.995	Emp	5580.67	0.297	4077.89	0.134	4642.61	0.093
	CKE	5706.69	0.282	4134.66	0.123	4643.42	0.087
	DTKE	7794.70	0.217	4444.75	0.095	4883.85	0.080

manera similar para todos los titulares de pólizas, la estimación empírica parece ser el mejor enfoque en $\alpha = 0.95$, pero no en $\alpha = 0.995$.

Cuando $\alpha = 0.995$, la subestimación del método de distribución Empírica (Emp) es evidente en comparación con el nivel del cuantil inferior en $\alpha = 0.95$. El método DTKE tiene el menor coeficiente de variación en comparación con los otros métodos.

Bibliografía

Aas, K., Czado, C., Frigessi, A., y Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Aldrich, J. H. y Nelson, F. D. (1984). *Linear probability, logit, and probit models*, volume 45. Sage.

Alemany, R., Bolancé, C., y Guillen, M. (2013). A nonparametric approach to calculating value-at-risk. *Insurance: Mathematics and Economics*, 52(2):255–262.

Alemany, R., Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016). Combining parametric and non-parametric methods to compute value-at-risk. *Economic Computation & Economic Cybernetics Studies & Research*, 50(4):61–74.

Almer, B. (1963). Individual risk theory and risk statistics as applied to fire insurance. *ASTIN Bulletin*, 2(3):365–379.

Altman, N. y Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46(2):195–214.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Artzner, P., Delbaen, F., Eber, J.-M., y Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Ashford, J. y Sowden, R. (1970). Multi-variate probit analysis. *Biometrics*, pp. 535–546.

Au, W.-H., Chan, K. C., y Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *Evolutionary Computation, IEEE Transactions on*, 7(6):532–545.

Avanzi, B., Taylor, G., y Wong, B. (2016). Correlations between insurance lines of business: An illusion or a real phenomenon? some methodological considerations. *Astin Bulletin*, 46(2):225–263.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328.

Bahraoui, Z., Bolancé, C., y Pérez-Marín, A. M. (2014). Testing extreme value copulas to estimate the quantile. *SORT*, 38(1):89–102.

Bailey, R. A. y Simon, L. J. (1960). Two studies in automobile insurance ratemaking. *Astin Bulletin*, 1(4):192–217.

Bedford, T. y Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 30:1031–1068.

Beerli, A., Martin, J. D., y Quintana, A. (2004). A model of customer loyalty in the retail banking market. *European Journal of Marketing*, 38(1/2):253–275.

Belles-Sampera, J., Guillen, M., y Santolino, M. (2014). Beyond value-at-risk: Gluevar distortion risk measures. *Risk Analysis*, 34(1):121–134.

Bermúdez, L., Ferri, A., y Guillen, M. (2013). A correlation sensitivity analysis of non-life underwriting risk in solvency capital requirement estimation. *Astin Bulletin*, 43(1):21–37.

Bolancé, C. (2010). Optimal inverse beta (3, 3) transformation in kernel density estimation. *SORT-Statistics and Operations Research Transactions*, 34(2):223–238.

Bolancé, C., Alemany, R., y Padilla-Barreto, A. E. (2018a). Impact of D-vine structure on risk estimation. *Journal of Risk*, 20(5):1–32.

Bolancé, C., Bahraoui, Z., y Artís, M. (2014). Quantifying the risk using copulae with nonparametric marginals. *Insurance: Mathematics and Economics*, 58:46–56.

Bolancé, C., Cao, R., y Guillen, M. (2018b). Estimación máximo verosimil condicionada del modelo lineal generalizado con función de ligadura no paramétrica. En Sarabia, J. M., Prieto, F., y Guillen, M., editores, *Contributions to Risk Analysis: RISK 2018*, pp. 93–100. Cuadernos de la Fundación Mapfre, Madrid.

Bolancé, C., Guillen, M., Gustafsson, J., y Nielsen, J. P. (2012). *Quantitative operational risk models*. CRC Press, Boca Ratón (USA).

Bolancé, C., Guillen, M., y Nielsen, J. P. (2003a). Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics*, 32(1):19–36.

Bolancé, C., Guillen, M., y Nielsen, J. P. (2008a). Inverse beta transformation in kernel density estimation. *Statistics and Probability Letters*, 78(13):1757–1764.

Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016a). Predicting defection in non-life motor and home insurance. *Lectures on Modeling and Simulation: a selection from AMSE 2016*, 2:107–120.

Bolancé, C., Guillen, M., y Padilla-Barreto, A. E. (2016b). Predicting probability of customer churn in insurance. En León, R., Muñoz-Torres, M. J., y Moneva, J. M., editores, *Modeling and Simulation in Engineering, Economics and Management: International Conference, MS 2016, Teruel, Spain, July 4-5, 2016, Proceedings*, número 254, pp. 82–91. Springer International Publishing, Cham.

Bolancé, C., Guillen, M., y Pinquet, J. (2003b). Time-varying credibility for frequency risk models: estimation and tests for autoregressive specifications on the random effects. *Insurance: Mathematics and Economics*, 33(2):273–282.

Bolancé, C., Guillen, M., y Pinquet, J. (2008b). On the link between credibility and frequency premium. *Insurance: Mathematics and Economics*, 43(2):209–213.

Bolton, R. N., Kannan, P. K., y Bramlett, M. D. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, 28(1):95–108.

Boos, D. D. (1982). A test for asymmetry associated with the hodges-lehmann estimator. *Journal of the American Statistical Association*, 77(379):647–651.

Boser, B. E., Guyon, I. M., y Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. En *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, New York, NY, USA. ACM.

Boucher, J.-P., Denuit, M., y Guillen, M. (2007). Risk classification for claim counts: A comparative analysis of various zeroinflated mixed poisson and hurdle models. *North American Actuarial Journal*, 11(4):110–131.

Boucher, J.-P., Denuit, M., y Guillen, M. (2009). Number of accidents or number of claims? an approach with zero-inflated poisson models for panel data. *Journal of Risk and Insurance*, 76(4):821–846.

Bowman, A., Hall, P., y Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85(4):799–808.

Brechmann, E. C. y Czado, C. (2013). Risk management with high-dimensional vine copulas: An analysis of the euro stoxx 50. *Statistics and Risk Modeling*, 30(4):307–342.

Brechmann, E. C., Schepsmeier, U., et al. (2013). Modeling dependence with C-and D-vine copulas: The r-package cdvine. *Journal of Statistical Software*, 52(3):1–27.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., y Pérez-Marín, A. M. (2008). Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection? *Journal of Risk and Insurance*, 75(3):713–737.

Brown, R. L. y Gottlieb, L. R. (2007). *Introduction to ratemaking and loss reserving for property and casualty insurance*. Actex Publications.

Buch-Larsen, T., Nielsen, J. P., Guillen, M., y Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516.

Byrd, R. H., Lu, P., Nocedal, J., y Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Christoffersen, P., Jacobs, K., Jin, X., y Langlois, H. (2013). Dynamic dependence in corporate credit. *Dynamic Dependence in Corporate Credit, Rotman School of Management Working Paper*, (2314027).

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Cortes, C. y Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cossette, H., Gaillardetz, P., Marceau, E., y Rioux, J. (2002). On two dependent individual risk models. *Insurance: Mathematics and Economics*, 30(2):153 – 166.

Crosby, L. A. y Stephens, N. (1987). Effects of relationship marketing on satisfaction, retention, and prices in the life insurance industry. *Journal of Marketing Research*, 24(4):404–411.

Darsow, W. F., Nguyen, B., Olsen, E. T., et al. (1992). Copulas and markov processes. *Illinois Journal of Mathematics*, 36(4):600–642.

De Jong, P., Heller, G. Z., et al. (2008). *Generalized linear models for insurance data*, volume 136. Cambridge University Press, Cambridge.

de Melo Mendes, B. V., Semeraro, M. M., y Leal, R. P. C. (2010). Pair-copulas modeling in finance. *Financial Markets and Portfolio Management*, 24(2):193–213.

Denuit, M., Dhaene, J., Goovaerts, M., y Kaas, R. (2005). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley and Sons, New York.

Denuit, M., Maréchal, X., Pitrebois, S., y Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley and Sons, New York.

Dionne, G. et al. (2000). *Handbook of Insurance*. Springer, Berlín.

Dissmann, J., Brechmann, E. C., Czado, C., y Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis*, 59:52–69.

Dobbin, K. K. y Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(1):31.

Dowd, K. y Blake, D. (2006). After var: the theory, estimation, and insurance applications of quantile-based risk measures. *Journal of Risk and Insurance*, 73(2):193–229.

Dreiseitl, S. y Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5):352 – 359.

Ekinci, Y. y Duman, E. (2015). Intelligent classification-based methods in customer profitability modeling. En *Intelligent Techniques in Engineering Management*, pp. 503–527. Springer, Berlín.

Embrechts, P., Lindskog, F., y McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Fahrmeir, L. y Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science and Business Media, Berlín.

Fang, K., Jiang, Y., y Song, M. (2016). Customer profitability forecasting using big data analytics: A case study of the insurance industry. *Computers and Industrial Engineering*, 101:554 – 564.

Fang, K.-T., Kotz, S., y W. Hg, K. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, New York.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fernández, C. y Steel, M. F. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Fornell, C. (1992). A national customer satisfaction barometer: The swedish experience. *the Journal of Marketing*, pp. 6–21.

Fornell, C. y Wernerfelt, B. (1987). Defensive marketing strategy by customer complaint management: a theoretical analysis. *Journal of Marketing Research*, pp. 337–346.

Fornell, C. y Wernerfelt, B. (1988). A model for customer complaint management. *Marketing Science*, 7(3):287–298.

Frank, M. J. (1979). On the simultaneous associativity of f (x, y) and x + y - f (x, y). *Aequationes Mathematicae*, 19(1):194–226.

Frees, E. W. (2009). *Regression modeling with actuarial and financial applications*. Cambridge University Press, Cambridge.

Frees, E. W., Derrig, R. A., y Meyers, G., editores (2014). *Predictive modeling applications in actuarial science*, volume 1. Cambridge University Press, Cambridge.

Frees, E. W., Jin, X., y Lin, X. (2013). Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, 7(2):258–287.

Frees, E. W., Lee, G., y Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, 4(1):4.

Frees, E. W. y Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

Friedl, M. A. y Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61(3):399–409.

Gelman, A. y Hill, J. (2006). *Missing-data imputation*, pp. 529–544. Analytical Methods for Social Research. Cambridge University Press, Cambridge.

Genest, C., Ghoudi, K., y Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Gigliarano, C., Figini, S., y Muliere, P. (2014). Making classifier performance comparisons when ROC curves intersect. *Computational Statistics and Data Analysis*, 77:300–312.

Gourieroux, C. y Jasiak, J. (2011). *The econometrics of individual risk: credit, insurance, and marketing*. Princeton University Press, Princeton.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Guelman, L. y Guillen, M. (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, 41(2):387–396.

Guelman, L., Guillen, M., y Pérez-Marín, A. M. (2012). *Random Forests for Uplift Modeling: An Insurance Customer Retention Case*, pp. 123–133. Springer, Berlin.

Guelman, L., Guillen, M., y Pérez-Marín, A. M. (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics*, 58:68–76.

Guelman, L., Guillen, M., y Pérez-Marín, A. M. (2015a). A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems*, 72:24–32.

Guelman, L., Guillen, M., y Pérez-Marín, A. M. (2015b). Uplift random forests. *Cybernetics and Systems*, 46(3-4):230–248.

Guillen, M. (2014). Regression with categorical dependent variables. *Predictive Modeling Applications in Actuarial Science*, 1:65–86.

Guillen, M. (2016). Big data en seguros. *Índice. Revista de Estadística y Sociedad*, 67(Abril):28–30.

Guillen, M., Nielsen, J. P., y Pérez-Marín, A. M. (2008). The need to monitor customer loyalty and business risk in the european insurance industry. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 33(2):207–218.

Guillen, M., Nielsen, J. P., Scheike, T. H., y Pérez-Marín, A. M. (2012). Time-varying effects in the analysis of customer loyalty: A case study in insurance. *Expert Systems with Applications*, 39(3):3551–3558.

Guillen, M., Sarabia, J. M., y Prieto, F. (2013). Simple risk measure calculations for sums of positive random variables. *Insurance: Mathematics and Economics*, 53(1):273–280.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Günther, C.-C., Tvete, I. F., Aas, K., Sandnes, G. I., y Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1):58–71.

Haberman, S. y Renshaw, A. E. (1996). Generalized linear models and actuarial science. *The Statistician*, pp. 407–436.

Hallowell, R. (1996). The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study. *International journal of service industry management*, 7(4):27–42.

Harrell, F. E. y Davis, C. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640.

Hastie, J. (1998). *Encyclopedia of Biostatistics*, volume 4. Wiley, Chichester.

Hornik, K., Meyer, D., y Karatzoglou, A. (2006). Support vector machines in r. *Journal of Statistical Software*, 15(9):1–28.

Hothorn, T., Hornik, K., y Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hsiao, C., Kim, C., y Taylor, G. (1990). A statistical perspective on insurance rate-making. *Journal of Econometrics*, 44(1):5–24.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan.

Jahromi, A. T., Stakhovych, S., y Ewing, M. (2016). Customer churn models: A comparison of probability and data mining approaches. En *Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing*, pp. 144–148. Springer, Berlín.

Jeong, H., Gan, G., y Valdez, E. (2018). Association rules for understanding policyholder lapses. *Risks*, 6(3):69.

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pp. 120–141.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press, Boca Ratón (USA).

Jørgensen, B. y Paes De Souza, M. C. (1994). Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 1994(1):69–93.

Keaveney, S. M. (1995). Customer switching behavior in service industries: An exploratory study. *The Journal of Marketing*, pp. 71–82.

Kim, J. K., Song, H. S., Kim, T. S., y Kim, H. K. (2005). Detecting the change of customer behavior based on decision tree analysis. *Expert Systems*, 22(4):193–205.

Klugman, S. A., Panjer, H. H., y Willmot, G. E. (1998). *Loss Models: From Data to Decisions*. John Wiley and Sons, New York.

Klugman, S. A., Panjer, H. H., y Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. John Wiley and Sons, New York, 2 edition.

Kumar, D. y Garg, A. (2013). A study of data mining techniques for churn prediction. *International Journal of Science, Engineering and Computer Technology*, 3(1):1.

Kunreuther, H. C., Pauly, M. V., y McMorrow, S. (2013). *Insurance and behavioral economics: Improving decisions in the most misunderstood industry*. Cambridge University Press, Cambridge.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica*, 51(2):507–512.

Lin, Y. y Grace, M. F. (2007). Household life cycle protection: Life insurance holdings, financial vulnerability, and portfolio implications. *Journal of Risk and Insurance*, 74(1):141–173.

Liu, Y., Zhang, H. H., y Wu, Y. (2011). Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, 106(493):166–177.

Livingstone, D. J., Manallack, D. T., y Tetko, I. V. (1997). Data modelling with neural networks: advantages and limitations. *Journal of Computer-aided Molecular Design*, 11(2):135–142.

Low, R. K. Y., Alcock, J., Faff, R., y Brailsford, T. (2013). Canonical vine copulas in the context of modern portfolio management: Are they worth it? *Journal of Banking and Finance*, 37(8):3085–3099.

Low, R. K. Y., Faff, R., y Aas, K. (2016). Enhancing mean–variance portfolio selection by modeling distributional asymmetries. *Journal of Economics and Business*, 85:49–72.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., y Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10):60–68.

McCullagh, P. y Nelder, J. A. (1983). *Generalized linear models. Monographs on statistics and applied probability*. Chapman and Hall, London.

McCullagh, P. y Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC Press, London.

McNeil, A. J., Frey, R., y Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton University Press, Princeton.

McNeil, A. J., Frey, R., y Embrechts, P. (2015). *Quantitative Risk Management: Concepts, techniques and tools*. Princeton University Press, Princeton.

Meyer, D., Dimitriadou, E., Hornik, L., Weingessel, A., Leisch, F., Chang, C., et al. (2012). Package e1071: Misc functions of the department of statistics (e1071), tu wien. *R package version*, pp. 1–6.

Meyer, D. y Wien, F. T. (2001). Support vector machines. *R News*, 1(3):23–26.

Min, A. y Czado, C. (2010). Bayesian inference for multivariate copulas using pair-copula constructions. *Journal of Financial Econometrics*, 8(4):511–546.

Min, A. y Czado, C. (2014). Scomdy models based on pair-copula constructions with application to exchange rates. *Computational Statistics and Data Analysis*, 76:523–535.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., y Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *Neural Networks, IEEE Transactions on*, 11(3):690–696.

Myatt, G. J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley and Sons, New York.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer Science and Business Media, New York.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science and Business Media, New York.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., y Mason, C. H. (2006). Defection detection: Measuring and understanding the accuracy of customer churn models. *Journal of Marketing Research*, 43(2):204–211.

Newhouse, J. P., Group, R. C. I. E., Staff, I. E. G., et al. (1993). *Free for all?: lessons from the RAND health insurance experiment*. Harvard University Press, Cambridge.

Nikoloulopoulos, A. K., Joe, H., y Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics and Data Analysis*, 56:3659–3673.

Oh, D. y Patton, A. J. (2013). Time-varying systemic risk: Evidence from a dynamic copula model of cds spreads. *Economic Research Initiatives at Duke (ERID) Working Paper*.

Ophem, H. V. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15(2):228–237.

Padilla-Barreto, A. E. (2018). Joint modelling modelling for customer lapses in the insurance sector. En Sarabia, J. M., Prieto, F., y Guillen, M., editores, *Contributions to Risk Analysis: RISK 2018*, pp. 219–226. Cuadernos de la Fundación Mapfre, Madrid.

Padilla-Barreto, A. E., Bolancé, C., y Guillen, M. (2016). Cuantificación del riesgo para la tarificación en seguros de automóvil. *Anales del Instituto de Actuarios Españoles*, 22:1–24.

Padilla-Barreto, A. E., Guillen, M., y Bolancé, C. (2017). Big-data analytics en seguros. *Anales del Instituto de Actuarios Españoles*, 23:1–19.

Pinquet, J., Guillen, M., y Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin: The Journal of the IAA*, 31(2):337–348.

Pinquet, J., Guillen, M., Bolancé, C., et al. (2000). Long-range contagion in automobile insurance data: estimation and implications for experience rating. Technical report, THEMA (Théorie Economique, Modélisation et Applications), Université de Cergy-Pontoise.

Rad, H., Low, R. K. Y., y Faff, R. (2016). The profitability of pairs trading strategies: distance, cointegration and copula methods. *Quantitative Finance*, 16(10):1541–1558.

Reavis, M. W. (2012). *Insurance: Concepts and Coverage: Property, Liability, Life, Health and Risk Management*. FriesenPress, Berlín.

Reichheld, F. F. y Teal, T. (2001). *The loyalty effect: The hidden force behind growth, profits, and lasting value*. Harvard Business Press, Cambridge.

Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, pp. 116–119.

Righi, M. B. y Ceretta, P. S. (2013). Analyzing the dependence structure of various sectors in the brazilian market: A pair copula construction approach. *Economic Modelling*, 35:199–206.

Righi, M. B. y Ceretta, P. S. (2015). Forecasting value at risk and expected shortfall based on serial pair-copula constructions. *Expert Systems with Applications*, 42(17):6380–6390.

Ruppert, D. y Cline, D. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. *Annals of Statistics*, 22(1):185–210.

Russom, P. et al. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34.

Schmidt, R. y Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2):307–335.

Sheather, S. J. y Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press, Boca Ratón (USA).

Sklar, A. (1959). Fonctions de repartition á n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449–460.

Smith, K. A., Willis, R. J., y Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the Operational Research Society*, pp. 532–541.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., y Snoeck, M. (2015). Profit maximizing logistic regression modeling for customer churn prediction. En *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pp. 1–10. IEEE.

Suykens, J. y Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300.

Swanepoel, J. W. y Van Graan, F. C. (2005). A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scandinavian Journal of Statistics*, 32(4):551–562.

Swedloff, R. (2014). Risk classification's big data (r) evolution. *Connecticut Insurance Law Journal*, 21:339.

Tao, D., Tang, X., Li, X., y Wu, X. (2006). Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1088–1099.

Terrell, G. R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85(410):470–477.

Thuring, F., Nielsen, J. P., Guillen, M., y Bolancé, C. (2012). Selecting prospects for cross-selling financial products using multivariate credibility. *Expert systems with Applications*, 39(10):8809–8816.

Thuring, F., Nielsen, J. P., Guillen, M., y Bolancé, C. (2013). Segmenting and selecting cross-sale prospects using dynamic pricing. En *ICORES*, pp. 103–108.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., y Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*, 67(4):30–45.

Wand, M. P., Marron, J. S., y Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353.

Weiß, G. N. y Scheffer, M. (2015). Mixture pair-copula-constructions. *Journal of Banking and Finance*, 54:175–191.

Weiß, G. N. y Supper, H. (2013). Forecasting liquidity-adjusted intraday value-at-risk with vine copulas. *Journal of Banking and Finance*, 37(9):3334–3350.

Werner, G. y Modlin, C. (2010). *Basic Ratemaking*. Casualty Actuarial Society, 4 edition.

Williams, C. A. y Heins, R. M. (1985). *Risk management and insurance*. McGraw-Hill Companies, New York.

Winkelmann, R. (2009). Copula-based bivariate binary response models. Technical report, Working Paper, Socioeconomic Institute, University of Zurich.

Yeo, A. C., Smith, K. A., Willis, R. J., y Brooks, M. (2001). Modeling the effect of premium changes on motor insurance customer retention rates using neural networks. En *Computational Science-ICCS 2001*, pp. 390–399. Springer, New York.

Zeithaml, V. A., Berry, L. L., y Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing*, pp. 31–46.

Zellner, A. y Lee, T. H. (1965). Joint estimation of relationships involving discrete random variables. *Econometrica*, 33(2):382–394.