UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2020, Vol. 13, No. 1, pp. 67–85

Aplicación de cópulas para modelar la pérdida total en una cartera de seguros vehiculares

Application of copulas to model total loss in a vehicle insurance portfolio

Maria José Bianco[a]
mariajosebianco@economicas.uba.ar

Yennyfer Feo[b]
feceyen@gmail.com

Lucas Barreda Frank[c]
barredalucas7@gmail.com

Resumen

El objetivo del presente trabajo se centra en analizar la problemática asociada con el estudio de las dos variables que influyen en la determinación de la pérdida técnica para el negocio asegurador. En este sentido, son puestas a prueba dos metodologías para el cálculo de los siniestros totales esperados en una cartera de seguros de autos en Brasil. En primera instancia, se consideró un modelo estadístico univariado denominado tradicional, del cual se encontró que la distribución Log-Normal fue la de mejor ajuste. Como segunda metodología, fue calibrado un modelo de cópula que permite incorporar distintos tipos y grados de asociación estocástica para las variables frecuencia y severidad. Los resultados mostraron que estas presentan comportamientos extremos, con una correlación de Kendall negativa y baja (-0.24), pero que rechazan la hipótesis de independencia al 5 % de confianza. Con este marco, la cópula de Clayton rotada 270 grados, con marginales Exp-Poisson y Log-Normal para la frecuencia y severidad respectivamente, presentó las mejores estimaciones de pérdida llegando a una diferencia de 12 % con la pérdida promedio empírica de la base. Para finalizar, se detectó que asumir independencia entre severidad y frecuencia para este caso de estudio, llevaría a sobrestimaciones significativas de la pérdida esperada.

***Palabras clave*:** Cópulas – Asociación Estocástica – Pérdida Esperada.

Abstract

[a]Investigadora Centro de Investigación en Métodos Cuantitativos Aplicados a la Economía y la Gestión (CMA)-Facultad de Ciencias Económicas, Universidad de Buenos Aires
[b]Investigadora (CMA)-Facultad de Ciencias Económicas, Universidad de Buenos Aires
[c]Investigador (CMA)-Facultad de Ciencias Económicas, Universidad de Buenos Aires

The main objective of this paper is to analyze the problems associated with the study of the variables that influence the determination of the technical capacity for the insurance business. In this sense, we tests two methodologies for calculating the total expected losses in a real portfolio of car insurance. In the first step, we modeled a conventional univariate statistical model, where we found that the Log-Normal distribution was the best fit. As a second methodology, a copula model was calibrated to incorporate the stochastic association between variables frequency and severity. The results showed that both distribution have extreme behaviors, with a negative and low Kendall correlation (-0.24), but that reject the hypothesis of independence at 5 % confidence. The rotated Clayton 270 degrees copula, with the marginal distributions Exp-Poisson and Log-Normal for frequency and severity respectively, presented the best estimates of the expected loss with a difference of 12 % in relation with the empirical expected loss of the real data base. Finally, we detected that the hypothesis of independence between severity and frequency for this practical exercise leads to significant overestimates of the expected loss.

***Keywords*:** Mixed - Stochastic Association - Expected Loss.

1. Introducción

En el área actuarial, el correcto modelado y estimación de la distribución de pérdida total de una cartera de seguros constituye un eje central para el proceso de administración del riesgo. Desde un punto de vista individual, permite determinar la prima y las cargas correspondientes a cada asegurado. Desde un nivel agregado, cuantificar el riesgo de la cartera significa establecer la capacidad total y de retención de la misma, lo que llevará a su correspondiente trato con contratos de reaseguro (Shi, Feng, & Ivantsova, 2015).

En la variada literatura de este tema se encuentra presente el supuesto de independencia, lo cual puede significar un análisis restrictivo y llevar a un sesgo en la estimación de los parámetros, y por lo tanto, de la distribución predictiva (Krämer, Brechmann, Silvestrini, & Czado, 2013). Por tal motivo, Gschlößl & Czado (2007) relajan dicho supuesto y modelan el monto promedio incluyendo como covariable al número de siniestros, la cual resulta significativa. En adelante, diversos trabajos ? por ejemplo en Czado, Kastenmeier, Brechmann, & Min (2012) - desarrollaron modelos basados en Song, Li, & Yuan (2009), donde se presenta una nueva metodología para la aplicación de los GLM al análisis de regresión conjunta con datos correlacionados mediante una cópula Gaussiana.

De este modo, adentrándonos en la problemática asociada con la posibilidad de considerar tipos de dependencia más generales y flexibles para las variables cantidad de siniestros y monto promedio de pagos, el eje del trabajo se centra en aplicar y comparar dos metodologías para el cálculo de pérdida esperada a una cartera real de seguros de automóviles en Brasil.

El presente artículo se estructura de la siguiente manera. La segunda sección es un

breve resumen de la teoría de cópulas, donde se resaltan los teoremas principales y familias de cópulas que se tendrán en cuenta en el apartado de aplicación a una serie de datos. Luego, la sección tres se centra en la aplicación de dos metodologías para calcular la pérdida esperada. En la primera, con el objetivo de definir completamente el modelo de pérdida total para el caso tradicional, se estimaron los parámetros para las distribuciones Log-Normal, Weibull, Pareto y Gamma, luego, una vez seleccionada la distribución teórica de pérdida, se estimó la esperanza. Con respecto a la segunda metodología, se utilizó un modelo de cópula para el modelado conjunto del monto de los pagos y el número de siniestros. Considerando la distribución marginal del pago por siniestros, se propone el testeo para la distribución Gamma, Weibull y Log-normal. En cuanto a la distribución marginal para la frecuencia de siniestros, a causa de que la frecuencia en la base de datos estudiada no es una variable entera debido a que se encuentra ponderada por la exposición de las pólizas, se estimó un modelo Exponencial y Exp-Poisson. Posteriormente se probaron familias de cópulas conocidas: Gaussiana, Frank, Gumbel rotada y Clayton rotada. Una vez modelada la cópula, se estimó la perdida esperada. Finalmente, se tiene la cuarta sección donde se exponen los resultados principales obtenidos y recomendaciones para futuros trabajos.

2. Cópulas

Siguiendo a Nelsen (2006), podemos entender las cópulas desde dos perspectivas. En primer lugar, como funciones que unen o "copulan" una función de distribución multivariada con sus funciones de distribución marginal univariantes, o, de otro modo, como funciones de distribución conjuntas cuyas marginales unidimensionales se distribuyen uniforme. En el sentido de la primera caracterización, la cópula representa una forma paramétrica conveniente para modelar la estructura de dependencia en distribuciones conjuntas con 2 o más variables aleatorias. De la segunda, surge la posibilidad del uso de estas funciones como una fuente para la construcción de distribuciones multivariadas a partir de definir correctamente las marginales a ser insertadas dentro de la misma.

Sin embargo, para que esto sea posible, resulta necesario que las cópulas cumplan con ciertas propiedades. En particular, dada una cópula bidimensional con variables aleatorias X e Y, cuyas funciones de distribución marginales son $F_1(X)$ y $F_2(Y)$, respectivamente, y denotando $u = F_1(X)$,$v = F_2(Y)$, se debe cumplir que:

$$C(u, v) : [0; 1]x[0; 1] \to [0; 1] \tag{1}$$

$$C(u, 0) = C(0, v) = 0; C(u, 1) = u; C(1, v) = v \tag{2}$$

para todo u_1, u_2, v_1, v_2 en [0;1] tal que $u_1 \leq u_2$ y $v_1 \leq v_2$ se tiene que

$$C((u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{3}$$

La primera propiedad establece que la cópula actúa sobre el rango de dos funciones de distribución, y las siguientes 2 postulan condiciones sobre los valores asignados en el conjunto de salida de esta, que permiten que cualquier función de distribución bivariante pueda ser una cópula (condiciones de frontera y propiedad 2-creciente).

El concepto de cópula fue introducido en el contexto de espacios métricos probabilísticos por Sklar, dando respuesta concreta a uno de los problemas formulados por M. Fréchet acerca de la relación entre una función de distribución multidimensional y sus marginales de menor dimensión. En dicho trabajo, Sklar demuestra mediante el teorema que ahora lleva su apellido, que una función de distribución conjunta puede ser expresada a partir de sus distribuciones marginales y una cópula que establezca la relación funcional entre ellas, la cual es única en caso de tratarse con marginales continuas. En términos formales,

Teorema de Sklar. *Sea H una función de distribución conjunta con marginales F_1 y F_2. Existe una cópula C tal que para todo $x, y \in \mathcal{R}$:*

$$H(x, y) = C[F_1(x), F_2(y)] \tag{4}$$

Si F_1 y F_2 son continuas, entonces C es única; en cualquier otro caso C sólo está determinada de forma única sobre el conjunto $RanF_1 x RanF_2$. También, si C es una cópula dada, F_1 y F_2 funciones de distribución dadas, entonces H definida mediante (4) es una función de distribución conjunta con marginales F_1 y F_2.

En el caso de trabajar con vectores aleatorios con funciones de distribución marginales discontinuas (no aplica para los datos de este trabajo), la unicidad en el teorema de Sklar se demostraría en función a las subcópulas. Una subcópula es una función C' con las siguientes propiedades:

- $DomC' = S_1 S_2$ donde S_1 y S_2 son subconjuntos del vector identidad ? que contienen al 0 y al 1.
- C' es creciente
- Para todo $(u, v) \in S_1 X S_2$ se tiene $C'(u, 1) = u$ y $C'(1, v) = v$.

Por lema, todo subcópula se puede extender (no necesariamente de manera única) a una cópula. Por lo que del teorema de Sklar se desprende que en el caso de las marginales no sean continuas, la cópula está unívocamente determinada en el dominio de la subcópula $RangoS_1 RangoS_2$. Para profundizar en el caso de marginales discretas, ver Illanes, G (2013).

Como consecuencia inmediata de este teorema surge su célebre corolario, el cual resulta de suma utilidad en el sentido que permite, dada la distribución conjunta de dos variables aleatorias continuas y sus respectivas funciones marginales, extraer la cópula subyacente y, luego con esta, construir nuevas distribuciones conjuntas de probabilidad bivariadas a partir de insertar en la misma distintas densidades marginales.

Corolario. *Sea H una función de distribución conjunta bivariada con marginales continuas F_1 y F_2, y sea C la única cópula tal que el Teorema de Sklar se cumple. Entonces para cualquier $(u, v) \in (R^2)$*

$$C(u, v) = H[F_1^{(-1)}(u), F_2^{(-1)}(v)] \tag{5}$$

La utilización de las cópulas se ha vuelto cada vez más atractivo en el sentido que otorga una herramienta con gran flexibilidad para el modelado de distribuciones conjuntas de vectores aleatorios en una manera relativamente sencilla, ya que únicamente requiere especificar la función que une (cópula) y sus marginales. Entre las principales ventajas de su empleo, tenemos que permiten trabajar la estructura de dependencia de las variables y su comportamiento marginal de forma independiente. Además, proveen una manera natural de estudiar medidas de dependencia, dado que captura aquellas propiedades de la distribución conjunta que son invariantes bajo transformaciones estrictamente crecientes, lo cual hace conveniente expresar dichas medidas en términos de la cópula (Embrechts, Höing, & Juri, 2003). Entre las medidas más utilizadas se encuentra la correlación Tau de Kendall, la cual cuantifica relaciones no necesariamente lineales, siendo en este sentido superior a la correlación de Pearson. A su vez se utiliza directamente como función de evaluación del contraste de independencia. Para un análisis más detallado sobre la teoría de cópulas puede consultarse Schweizer & Sklar (2011) o Nelsen (2006).

2.1. Familias de cópulas

En el estudio acerca del ajuste de un modelo a un conjunto de datos en particular, los métodos multivariados deben ser analizados, por un lado, por el tipo de dependencia estocástica que poseen en su estructura y, por el otro, por el grado de asociación entre las variables involucradas. Entre los tipos de dependencia, Joe (1997) incluye: (I) singularidades en algunas curvas o superficies; (II) dependencia positiva y negativa; (III) dependencia intercambiable o flexible; (IV) decrecimiento en el tiempo del grado de dependencia.

Como se ha visto con el Teorema de Sklar y su corolario, es posible a partir de una colección de cópulas construir distribuciones bivariadas con marginales totalmente arbitrarias. Dichas colecciones son conocidas como familias de cópulas, que se diferencian principalmente por el tipo y el grado de dependencia que logran captar, y de las cuales podemos encontrar en un importante número dentro de la literatura actual. Para la construcción de estas existen diversas metodologías, como ser el método de inversión (donde es explotado el corolario para producir cópulas directamente desde las funciones de distribución conjuntas), algebraico (son estudiadas las relaciones que involucran la distribución bivariada con sus marginales) y el de funciones generadoras, entre otros. A continuación, se exponen algunas familias de cópulas que fueron utilizadas para este trabajo junto con sus características principales. Una mayor profundidad puede encontrarse en Joe (1997) y Krupskii & Joe (2015).

Cópula Gaussiana

Perteneciente a la familia de cópulas elípticas, es una de las más utilizadas en la práctica por ser fuente de distribuciones que recogen muchas de las convenientes propiedades de la normal multivariante. Construida a partir del método de inversión, se encuentra definida como sigue:

$$C_\rho(u,v) = \phi_2(\phi^{-1}(u), \phi^{-1}(v); \rho) \tag{6}$$

$$C_\rho(u,v) = \int_{\infty}^{\phi(u)} \int_{\infty}^{\phi(v)} \frac{1}{2\pi(1-\rho^2)^{0.5}} \left(\frac{q^2 - 2\rho qt + t^2}{-2(1-\rho^2)} \right) \partial q \partial t \tag{7}$$

Donde $\phi_2(.,.,;\rho)$ es la función de distribución normal bivariada con coeficiente de correlación de Pearson ρ, $\phi(.)$ es la distribución acumulada de una normal estándar univariada con $\phi^{-1}(.)$ su función inversa, y q y t son dos variables auxiliares.

Una de las principales ventajas de esta cópula es que permite capturar el rango completo de dependencia en términos de la cota superior e inferior de Fréchet y, además, codifica la asociación de igual manera que lo hace la normal bivariada con el coeficiente de correlación lineal, pero con la posibilidad de utilizar distintas marginales. Un aspecto negativo es que para $\rho < 1$ la dependencia en las colas es nula.

Familia Arquimediana

Una de las familias más estudiadas en la literatura por proveer una estructura para el modelado de distribuciones bivariadas a través de una función generadora univariante. En términos formales, *Se dice que una cópula C es Arquimediana si puede ser expresada de la forma:*

$$C(u,v) = \phi_\theta^{-1}\left[\phi_\theta(u) + \phi_\theta(v)\right] \tag{8}$$

para alguna función convexa decreciente ϕ_θ, conocida como función generadora, definida en $(0;1]$ que satisface $\phi_\theta^{-1}(1) = 0$; por convención $\phi_\theta^{-1}(t) = 0$ cuando $t \geq \phi_\theta(0)$

Las condiciones dadas para la definición son necesarias y suficientes para que (8) sea una función de distribución bivariada (Schweizer & Sklar, 2011). Estas cópulas cumplen con propiedades como simetría, asociatividad y curvas de nivel convexas, y cuentan con la ventaja de poder expresar el coeficiente de dependencia en las colas en términos de los generadores. Además, presenta una forma sencilla para expresar varios cálculos, en particular, para la Tau de Kendall que viene dada por:

$$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\acute{\phi}(t)} dt \tag{9}$$

Las cópulas arquimedianas más conocidas son las que se detallan en la tabla 1 con su respectiva definición y función generadora. La cópula Frank posee la capacidad de capturar el rango completo de dependencia de forma tal que, alcanza la cota inferior de Fréchet cuando $\theta \to -\infty$, la superior con $\theta \to +\infty$ y la de independencia cuando $\theta \to 0$. En este sentido, posee características similares a la Gaussiana (la Frank tampoco es dependiente en las colas), pero con la diferencia que la primera proporciona cantidades cerradas, lo cual la vuelve más sencilla de programar. En cuanto a las dos cópulas restantes, ambas son útiles para el modelado de dependencia en las colas, diferenciándose principalmente en que la Gumbel lo captura en la cola superior y la Clayton en la inferior.

Cópula	$C(u,v)$	$\phi_\theta(t)$
Frank	$\frac{1}{\theta}\ln\left(\frac{1+(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$
Gumbel	$\exp\left(-\left[(-\ln(u))^\theta+(-\ln(v))^\theta\right]^{\frac{1}{\theta}}\right)$	$(-\ln(t))^\theta$
Clayton	máx $\left[\left(u^{-\theta}+v^{-\theta}\right)^{\frac{1}{\theta}},0\right]$	$\frac{1}{\theta}(t^\theta-1)$

Tabla 1: Cópulas Arquimedianas - Definición y función generadora. *Fuente: Nelsen (2006).*

2.2. Cópulas en el área actuarial

Como fue mencionado en la primera sección, el modelamiento conjunto mediante la utilización de cópulas se ha vuelto muy atractivo en muchos campos de investigación donde resulta de interés la dependencia entre variables aleatorias, y para las cuales el supuesto de normalidad multivariada no se ajusta a los datos. Siendo una herramienta de gran utilidad en el sentido que define estructuras de correlación, esta teoría encuentra diversas aplicaciones en áreas como por ejemplo finanzas, donde puede ser utilizada para la estimación del riesgo de crédito y el valor a riesgo para distintos activos dentro de un mismo portafolio, la valuación de derivados y el cálculo del capital económico en empresas aseguradoras o instituciones financieras en general (Cherubini, Luciano, & Vecchiato, 2004).

En particular para el área actuarial, las cópulas han sido introducidas por Frees & Valdez (1998) y Klugman & Parsa (1999), entre otros autores, como un potencial campo de estudio en el modelado de dependencias entre pérdidas para la administración del riesgo. Entre las principales aplicaciones está el análisis de patrones de asociación en la mortalidad para grupos de individuos con seguros colectivos, la estimación del tiempo de vida para causas de decremento múltiple, la optimización de coberturas mediante ordenamiento estocástico y el modelado de distribuciones en pérdidas multivariadas, como por ejemplo siniestros y gastos. Dentro de esto último también se incluye el tema de interés para el presente trabajo, el cual se basa en definir un modelo de cópula que permita establecer una estructura de dependencia para las dos componentes de la pérdida técnica del seguro (frecuencia y severidad).

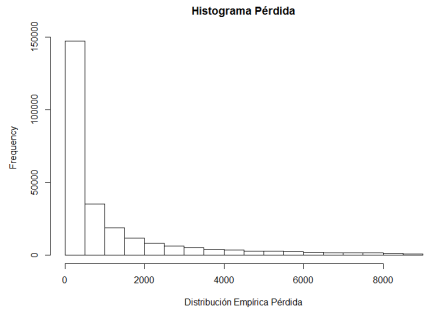
La estimación de dicha pérdida constituye una de las principales preocupaciones para los actuarios en cuanto al proceso de decisión del riesgo que toma la compañía aseguradora. De un lado permite, en la relación asegurado y aseguradora, definir la prima y cargas por parte del primero, y respecto al vínculo entre aseguradora y reaseguradora, establecer la retención y capacidad máxima viable para el negocio. En el sentido de esta problemática, Klugman, Panjer, & Willmot (2012) mencionan una serie de ventajas en cuanto al modelado por separado de la distribución para la frecuencia y la severidad, en lugar de hacerlo directamente desde la variable pérdida. Por mencionar algunas de ellas:

- Los efectos de la inflación, tanto económica como en adición de siniestros, se reflejan en las pérdidas sufridas por los asegurados y los reclamos pagados por las compañías de seguros. Tales efectos a menudo se enmascaran cuando las pólizas de seguros tienen deducibles y límites de pólizas que no dependen de la inflación y se usan los resultados agregados.
- Los modelos desarrollados para las pérdidas no cubiertas a los asegurados, los costos de siniestros a las aseguradoras y los de las cesiones a los reaseguradores pueden ser mutuamente consistentes. Esta característica es útil para una aseguradora directa al estudiar las consecuencias de transferir pérdidas a un reasegurador.
- La forma de la distribución de pérdida depende de las formas de ambas distribuciones de frecuencia e intensidad. Por ejemplo, si la distribución de severidad tiene una cola mucho más pesada que la del número de reclamaciones, la primera determinará la forma de la cola para la distribución de pérdida agregada, y será insensible a la elección del comportamiento de la frecuencia.

Por último, una de las principales razones por las cuales puede ser sesgada la estimación de la pérdida, se debe a la gran volatilidad que presenta la severidad para diferentes números de siniestros dentro de una misma cartera de riesgo. Como consecuencia de ello, resulta necesario considerar métodos que permitan una mayor flexibilidad en cuanto al modelado del comportamiento de estas variables, por lo que, en este sentido, el empleo de cópulas nos brinda una gran utilidad.

3. Aplicación a una base de datos de Brasil

La base de datos utilizada en este trabajo proviene de **Brvnhins**, ubicada en la librería CASDataset (Dutang, Charpentier, & Dutang, 2015) que puede ser consultada desde el software libre R (URL https://www.R-project.org/), y que contiene 1.965.355 registros de pólizas de AUTOSEG (un acrónimo de Sistema estadístico para automóviles en Brasil). Cada registro incluye características de riesgo, monto de la reclamación y el historial de reclamos para distintos amparos de riesgo (robo, incendio y destrucción parcial y total) sin considerar límites ni deducibles



Valor pérdida	
Número de Observaciones	252.797
Valor promedio	1.032,8
Desviación estándar	1.624,9
Sesgo	2,4
Valor mínimo	0,014
Valor máximo	8.787,5

Figura 1: **Histograma y descriptivos de la pérdida.** *Fuente: Elaboración propia.*

para el año 2011. De esta forma, la pérdida total de la cartera es definida como el agregado de los montos reclamados (ponderados por la exposición) provenientes de las diferentes coberturas, y valorados en su totalidad como un costo por parte de la aseguradora. De lo disponible, se seleccionaron las pólizas que presentaron al menos un reclamo en el tiempo expuesto de un año, lo cual da una base final de 252.797 registros.

3.1. Estimación función de pérdida

Para el análisis de la primera metodología, se trabajará con la función a estimar denominada pérdida (las cifras son en reales Br), cuyo cálculo está dado de la siguiente manera (Kellison & London, 2011):

$$S_i = \frac{\text{Total suma reclamada}}{\text{total exposición}} \tag{10}$$

La pérdida podría considerarse como el total de la suma reclamada, sin embargo para calcular la pérdida total o ultimate, es necesario ponderar los montos por el riesgo devengado, de manera tal de poder tener en consideración todos aquellos siniestros que podrán ocurrir para las pólizas que no finalizaron dentro del año en cuestión. Por esta razón, la pérdida total queda definida como la suma reclamada dividida por el total de exposición de la cartera. Dentro de la base bajo estudio, la función de distribución empírica de la pérdida presentó el comportamiento que muestra la figura 1.

Como se observa en el histograma anterior, la función de pérdida tendría las características de las distribuciones de familias asimétricas, sesgadas a izquierda, larga cola derecha y leptocúrticas. De este modo, estas características nos dan fundamento para testear en la primera fase, las funciones de distribución Log-Normal, Weibull, Gamma y Pareto.

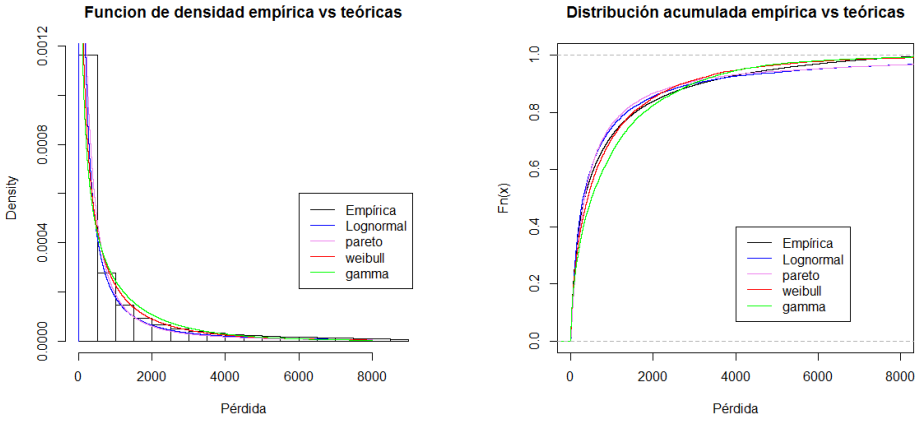


Figura 2: **Función empírica vs teóricas.** *Fuente: Elaboración propia.*

En la tabla 2, se muestra los parámetros estimados por método de máxima verosimilitud para las cuatro funciones:

Distribución	Pareto	Log-Normal	Gamma	Weibull
Parámetro de escala	1,086	5,7	0,511	0,6307
parámetro de forma	362,849	1,804	0,00049	716,762

Tabla 2: Estimación de la distribución univariada. *Fuente: Elaboración propia.*

En el gráfico 2, se comparan las funciones de densidad teóricas con respecto a la empírica mediante las gráficas de las funciones de densidad y de distribución acumulada:

Según el pp-plot (probability plot) de la figura 3 que grafica las coordenadas $(F_n(x_i), F^*(x_i))$, donde $F_n(x_i)$ es el valor que toma la función empírica en x_i y $F^*(x_i)$ es la función de distribución teórica ajustada, entre más ajustados los puntos a la recta de 45 grados mejor. De esta forma, se puede observar que la función Log-Normal es la que mejor se ajusta a la distribución empírica de los datos, y, por el contrario, la de peor desempeño entre los percentiles 30 y 90 es la función Gamma.

Empero, si bien las pruebas gráficas permiten sacar algunas conclusiones iniciales sobre las distribuciones y su ajuste a los datos, es necesario una especificación más formal mediante las pruebas de bondad de ajuste.

Las pruebas de bondad de ajuste son herramientas objetivas por medio de las cuales se puede aceptar o rechazar una distribución como un modelo adecuado a los datos observados. Entre las más utilizadas se encuentra la prueba Chi-cuadrado, en donde es comparado intervalo por intervalo, qué tan cerca está el número de datos observados en la muestra al número de datos esperados con base en la distribución.

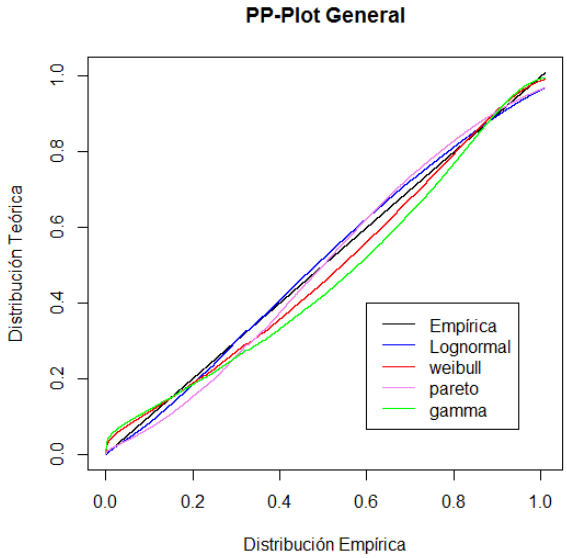


Figura 3: **PP-Plot General.** *Fuente: Elaboración propia.*

La prueba Kolmogorov- Smirnov (KS) calcula la máxima diferencia absoluta, sobre todo el rango, entre la función de distribución empírica y la función de distribución ajustada a los datos. En la prueba de Anderson-Darling (AD), en contraste con la prueba anterior, se busca la mayor diferencia entre las distribuciones empírica y teórica, siendo esta un promedio ponderado del cuadrado de dichas diferencias sobre todo el rango de la distribución. También están los criterios de información para evaluar la parsimonia de un modelo, entre las que se encuentran el criterio de Akaike (AIC) y Bayesiano (BIC). Según estas pruebas, se elige aquel modelo que tenga el menor criterio de información (IC).

La tabla 3 contiene la estimación de los parámetros para cada una de las distribuciones y su desempeño de acuerdo con las mencionadas pruebas:

Distribución	Chisq	AD	KS	AIC	BIC
Paretro	37515,1	1736,3	0,048	3909342	3909363
Log-Normal	24333,7	690,3	0,032	3897954	3897975
Weibull	31702,3	1063,2	0,039	3899356	3899377
Gamma	47427,2	2940,5	0,074	3912751	3912772

Tabla 3: Pruebas de bondad de ajuste en modelo univariado. *Fuente: Elaboración propia.*

Si se asigna una clasificación de las distribuciones con base en las pruebas de ajuste, donde se le coloca 1 a la distribución cuya prueba arrojó el menor valor, 2 a la

segunda menor y así sucesivamente, se observa en la tabla 4 que la distribución con menores valores en todas las pruebas (lo que significa mejor ajuste) fue la Log-Normal, lo que ya se había percibido en las pruebas gráficas.

Distribución	Chisq	AD	KS	AIC	BIC
Paretro	3	3	3	3	3
Log-Normal	1	1	1	1	1
Weibull	2	2	2	2	2
Gamma	4	4	4	4	4

Tabla 4: Clasificación de distribuciones. *Fuente: Elaboración propia.*

El cálculo estimado de los valores esperados se muestra en la tabla 5:

Distribución	E(X)	Dif.
Valor promedio dist. Empírica	1032,85	
Paretro	1022,54	-10,31
Log-Normal	1071,65	38,80
Weibull	988,89	-43,96
Gamma	1026,23	-6,62

Tabla 5: Clasificación de distribuciones. *Fuente: Elaboración propia.*

3.2. Modelo de cópula

En esta metodología el objetivo es estimar las funciones de distribución marginal para la frecuencia, severidad, y la cópula bivariada. Se diferencia de la anterior en que no es solo una función total la que se estima, si no que se le da importancia a la relación entre el monto del siniestro y la frecuencia de su ocurrencia.

$$\text{Frecuencia} f_i = \frac{\text{Total número reclamos}}{\text{total exposición}} \tag{11}$$

$$\text{Severidad} X_i = \frac{\text{Total suma reclamada}}{\text{total número reclamos}} \tag{12}$$

Pasar de la ecuación (10) a la (11) y (12) sólo requiere multiplicar y dividir por el número total de reclamos y hacer la agrupación correspondiente. En este caso, la pérdida total queda determinada por el producto entre la frecuencia (que es ultimate debido a que se encuentra ponderada por la exposición) y la severidad (que también es utimate porque considera que el monto de los siniestros reclamados no sufrirán modificaciones).

Al ver en la figura 4 ambas funciones, y tener estas un coeficiente de correlación muy bajo de -0.1677614, parecería no haber un comportamiento conjunto marcado. Sin embargo, la prueba de independencia Pearson's Chi-squared arroja

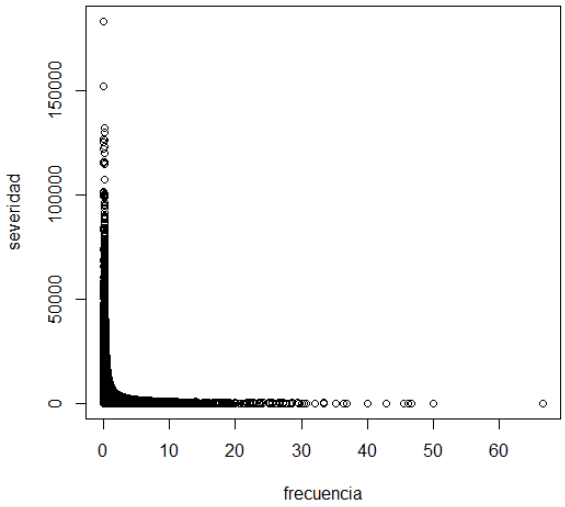


Figura 4: **Severidad vs frecuencia.** *Fuente: Elaboración propia.*

$\mathcal{X}^2 = 27599000$, $df = 252800$, $p - value < 2.2e - 16$ rechaza la hipótesis nula de independencia al 5 %, por lo que podemos pensar en el uso de cópulas.

Como se mencionó en secciones anteriores, la correlación Tau de Kendall cuantifica relaciones no necesariamente lineales, sino que se utiliza directamente como función de evaluación del contraste de independencia, y como parámetro de las familias de cópulas mencionadas. Para los datos de esta base la Tau fue de -0.2426089.

En cuanto a la parametrización de las distribuciones individuales, a causa de que la frecuencia en esta base de datos no es particularmente una variable con n entero (debido a la exposición de las pólizas), se estima en la tabla 6 una función Exponencial y Exp-Poisson para la primer marginal:

Distribución	Exponencial	Exp-Poisson
Parámetro de escala	1,7231	1,7224
Parámetro de forma		0,8043
Log-likelihood	-194317	-178657
AIC	3888636	376453
BIC	388647	376462

Tabla 6: Parametrización y pruebas de ajuste para la frecuencia. *Fuente: Elaboración propia.*

Para la severidad, se estimaron en la tabla 7 los parámetros para las distribuciones Gamma, Log-Normal y Weibull.

Distribución	Log-likelihood	AIC	BIC
Gamma	-2203725,66	4407455,32	4407476,20
Log-Normal	-2190420,98	4380845,96	4380866,84
Weibull	-2194091,10	4388186,21	4388207,09

Tabla 7: Parametrización y pruebas de bondad de ajuste para la severidad. *Fuente: Elaboración propia.*

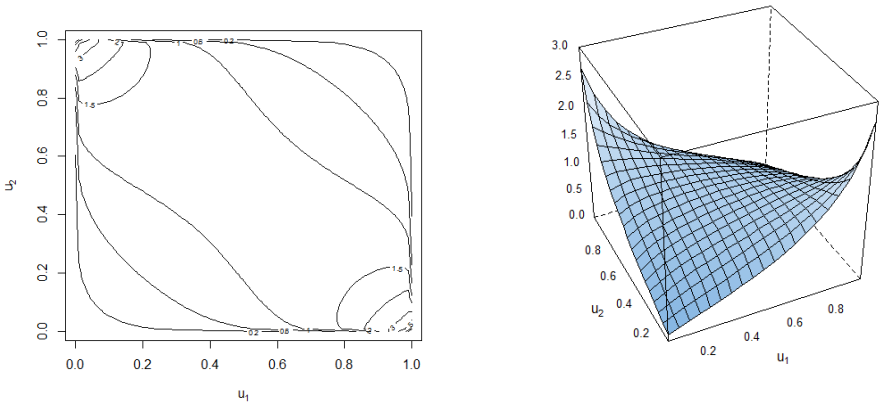


Figura 5: **Contorno y función de distribución de la cópula Gaussiana simulada.** *Fuente: Elaboración propia.*

Las pruebas dieron mejor ajuste para las distribuciones Exp-Poisson y la Log-Normal. Por otra parte, la correlación entre las funciones de densidad estimadas mantiene el orden de la Tau alrededor de -0.24.

Para la combinación de marginales seleccionadas, se testearon las cópulas Gaussiana, Clayton rotada 270 grados, Gumbel rotada 90 grados y Frank, cuyas estimaciones se presentan a continuación.

En las figuras 5, 6, 7 y 8 se observan las líneas de contorno o gráfico de líneas de densidad y la función de distribución de las cópulas simuladas, donde se pretende mostrar la estructura de dependencia entre la severidad y la frecuencia. En los cuatro casos se observa una dependencia negativa (por como los centroides se oponen en la parte superior izquierda e inferior derecha), y se evidencia la sensibilidad de dependencia en las colas. Además, puede notarse mayor concentración en los extremos de las funciones, la cual resulta en algunos casos más pronunciada que en otros, como se verá a continuación.

Como se mencionó anteriormente, la cópula Gaussiana es simétrica, lo cual se corrobora en el gráfico 5. Los parámetros estimados fueron $par = -0.4$ y $tau = -0.3$, con un criterio de *maximized loglikelihood* de $579,1$ y un AIC de $-1156,2$.

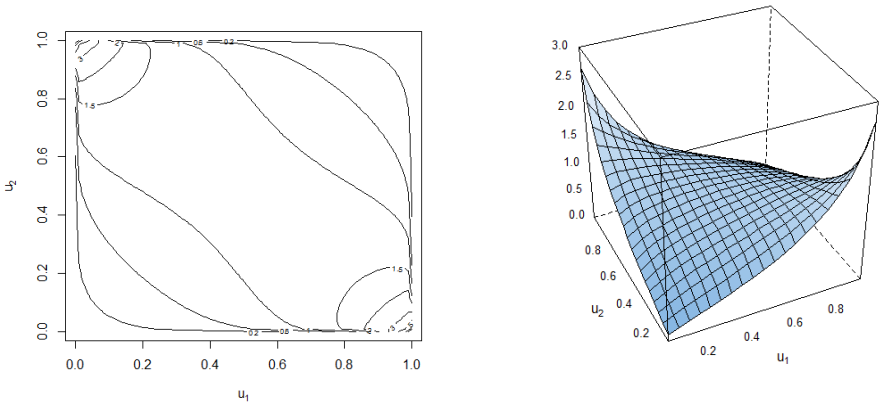


Figura 6: **Contorno y función de distribución de la cópula Clayton rotada 270^0 simulada.** *Fuente: Elaboración propia.*

La simulación de esta cópula arrojaría una pérdida esperada de $1325,88$ BR.

La cópula Clayton rotada 270^0 resalta mayor dependencia en el caso de niveles altos de severidad con poca frecuencia. Los parámetro estimados fueron $par = -0.79$ y $tau = -0.28$, con un criterio de maximized loglikelihood de $700,07$ y un AIC de $-1398,14$. La simulación de esta cópula arrojaría una pérdida esperada de 1167,318 BR.

La cópula Gumbel rotada 90^0 presentó similar comportamiento que la Clayton rotada 270^0 para valores altos de severidad con poca frecuencia, y también resalta un comportamiento leve en el costado de mayor frecuencia y poca severidad. Los parámetro estimados fueron $par = -1.39$ y $tau = -0.28$, con un criterio de maximized loglikelihood de $682,6$ y un AIC de $-1363,2$. La simulación de esta cópula arrojaría una pérdida esperada de 1286,495 BR.

La cópula Frank fue la que mostró una estructura de dependencia menos rígida en las colas, con parámetros estimados $par = -2.37$, $tau = -0.25$ y un criterio de *maximized loglikelihood* de $614,9$ y un AIC de $-1227,8$. La simulación de esta cópula arrojaría una pérdida esperada de 1971,36 BR, resultando ser la cuantía más alta comparada con las otras familias de cópulas.

En todas las simulaciones realizadas, la familia de mejor resultado y más cercana a la realidad fue la Clayton rotada 270^0 (además de ser la de mayor criterio de máxima verosimilitud y menor AIC). Adicionalmente, se probaron las mismas cópulas con otras combinaciones de marginales y la seleccionada fue la que arrojó mejores estimaciones. Por otro lado, se corrobora que para esta cópula estimada, las funciones u, v se distribuyen uniforme $[0, 1]$ como se observa en la figura 9.

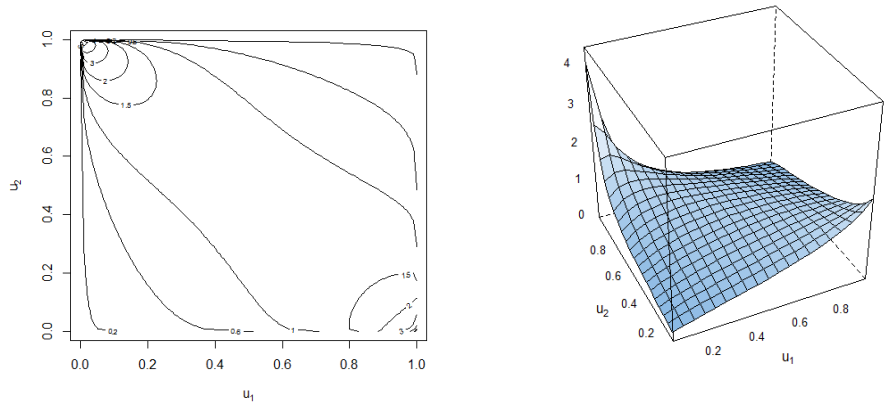


Figura 7: **Contorno y función de distribución de la cópula Gumbel rotada 90^0 simulada.** *Fuente: Elaboración propia.*

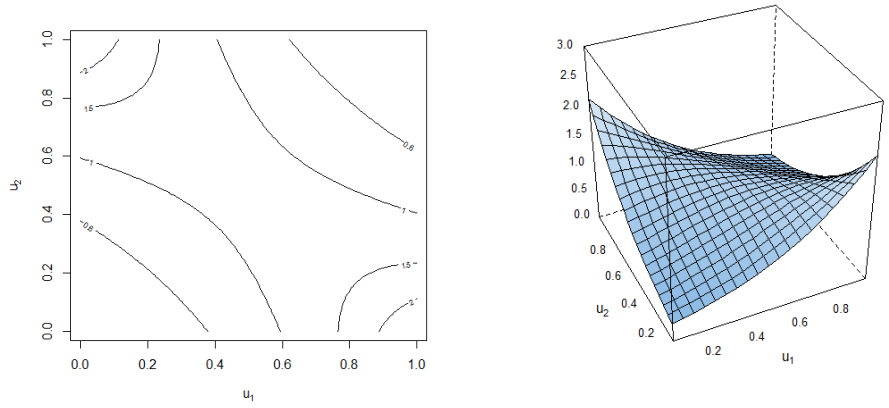


Figura 8: **Contorno y función de distribución de la cópula Frank simulada.** *Fuente: Elaboración propia.*

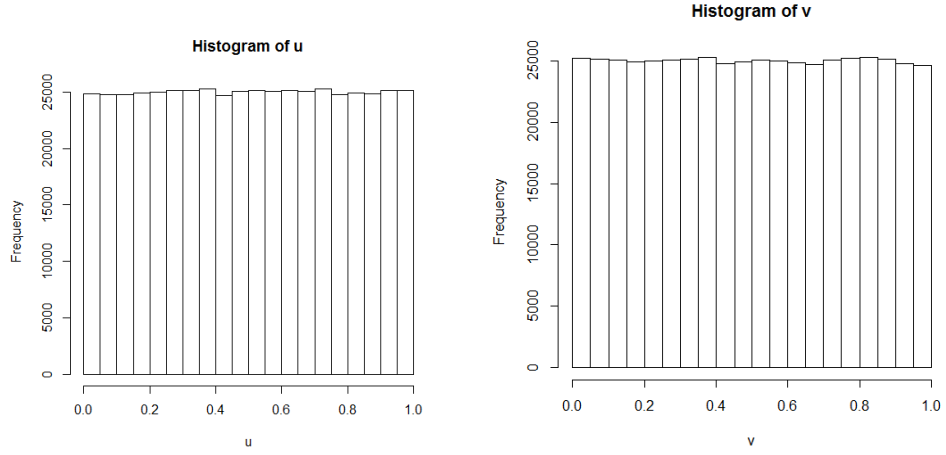


Figura 9: **Histograma para u y v de la cópula ajustada Clayton rotada 270.** *Fuente: Elaboración propia.*

Ahora bien, una vez elegidas las marginales Exp-Poisson ($shape = 1.7231832$, $rate = 0.80581$) y Log-Normal ($meanlog = 6.6036$, $sdlog = 1.90053$), la estimación para la esperanza de las distribuciones marginales arrojó:

$$\hat{E}(\text{Frecuencia}) = 0.777 \tag{13}$$

$$\hat{E}(\text{Severidad}) = 4407.519Br \tag{14}$$

Cuyo producto asumiendo independencia es:

$$\hat{E}(\text{Frecuencia})x\hat{E}(\text{Severidad}) = 3427.29Br \tag{15}$$

E incorporando información de la cópula Clayton rotada 270^0 ($par = -0.79$, $tau = -0.28$):

$$\hat{E}(\text{Pérdida}) = 1167Br \tag{16}$$

En este caso de estudio, el resultado demuestra el riesgo de asumir independencia en las funciones para calcular la pérdida, dado que bajo este supuesto se sobreestimaría la esperanza en un 194 %.

4. Ideas finales y trabajo futuro

De las fases mencionadas se encontró que la distribución con mejor ajuste para la distribución de la pérdida fue la Log-Normal tanto en las pruebas gráficas, como en las numéricas, estimando la perdida esperada con una diferencia de 4 % de la cifra empírica de la base. En el caso de la segunda parte, las variables de frecuencia y severidad tienen comportamientos muy extremos, una correlación Tau de Kendall negativa y baja (-0,24), que sin embargo rechaza la hipótesis de independencia con un 5 % de confianza. La mejor combinación de marginales estimadas fue Exp-Poisson para la función de densidad de la frecuencia y Log-Normal para la distribución de severidad, con la familia bivariada Clayton rotada 270^0. Esta familia, la cual es asimétrica y da mayor enfoque a la relación en las colas de las marginales, presentó las menores estimaciones de perdida llegando a una diferencia de 12 % con la perdida promedio empírica de la base. La familia que mayor sobrestimaciones arrojó fue la copula Frank.

Cabe resaltar que de acuerdo al estudio que se pretenda realizar, puede resultar conveniente el modelado univariado o con cópulas. En el caso que se le quiera dar un mayor énfasis a los eventos extremos y funciones de cola larga, o realizar simulaciones para percibir los comportamientos marginales, se recomienda el uso de modelos bivariados teniendo en cuenta la asociación estocástica. Por otro lado, si el objetivo se centra en una estimación global, la metodología de unificar la información en una sola distribución sin tener en cuenta el tipo de asociación entre las funciones de severidad y frecuencia es más robusta. Sin embargo, se debe tener en cuenta que asumir independencia entre dichas distribuciones llevaría a grandes sobrestimaciones de la perdida esperada.

Se propone como trabajo futuro la consideración de diferentes familias de cópulas, así como también el estudio para una estimación empírica de una cópula no paramétrica. Esta última resulta ser un instrumento útil para construir contrastes no paramétricos de independencia, y realizar análisis exploratorios de los datos (Genest, Rémillard, & Beaudoin, 2009). Otra posible línea de investigación que surge de este trabajo, se relaciona con la incorporación de medidas de riesgo, las cuales ayudarían a entender de una manera más acabada la problemática asociada con la dependencia en las colas para las distribuciones de frecuencia y severidad.

Recibido: 12/12/2019
Aceptado: 04/05/2020

Referencias

(n.d.).

Cherubini, U., L. E. & Vecchiato, W. (2004), *Copula methods in finance.*, John Wiley Sons.

Czado, C., K. R. B. E. C. & Min, A. (2012), 'A mixed copula model for insurance claims and claim sizes.', *Scandinavian Actuarial Journal* **4**, 278–305.

Dutang, C., C. A. & Dutang, M. C. (2015), ' Package CASdatasets.', *Rcram* .

Embrechts, P., H. A. & Juri, A. (2003), 'Using copulae to bound the value-at-risk for functions of dependent risks', *Finance and Stochastics* **7(2)**, 145–167.

Frees, E. W. & Valdez, E. A. (1998), 'Understanding relationships using copulas.', *North American actuarial journal* **2(1)**, 1–25.

Genest, C., R. B. & Beaudoin, D. (2009), 'Goodness-of-fit tests for copulas: A review and a power study.', *Insurance: Mathematics and economics* **44(2)**, 199–213.

Gschlößl, S. & Czado, C. (2007), 'Spatial modelling of claim frequency and claim size in non-life insurance.', *Scandinavian Actuarial Journal* **3**, 202–225.

Illanes, G. (2013), *Cópulas paramétricas y no paramétricas con aplicaciones en riesgo bancario. Tesis de maestría.*, Facultad de ingenieria Universidad de la Republica. Montevideo. Uruguay.

Joe, H. (1997), *Multivariate models and multivariate dependence concepts.*, Chapman and Hall.

Kellison, S. G. & London, R. L. (2011), *Risk Models and Their Estimation.*, Actex Publications.

Klugman, S. A., P. H. H. & Willmot, G. E. (2012), *Loss models: from data to decisions*, John Wiley and Sons.

Klugman, S. A. & Parsa, R. (1999), 'Fitting bivariate loss distributions with copulas. ', *Insurance: mathematics and economics* **24**, 139–148.

Krupskii, P. & Joe, H. (2015), ' Structured factor copula models: Theory, inference and computation. ', *Journal of Multivariate Analysis* **138**, 53–73.

Krämer, N., B. E. C. S. D. & Czado, C. (2013), 'Total loss estimation using copula-based regression models.', *Insurance: Mathematics and Economics* **53**, 829–839.

Nelsen, R. (2006), *An introduction to copulas*, 2 edn, SpringerScience Business Media, New York.

Schweizer, B. & Sklar, A. (2011), *Probabilistic metric spaces*, Courier Corporation.

Shi, P., F. X. & Ivantsova, A. (2015), 'Dependent frequency–severity modeling of insurance claims', *Insurance: Mathematics and Economics* **64**, 417–428.

Song, P. X., L. M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using Gaussian copulas.', *Biometrics* **65(1)**, 60–68.