

# Ajuste de distribución de pérdidas para seguro de accidentes automovilísticos

Valentín Chavarría<sup>1</sup>, Randall Chaves<sup>1</sup>, and & Brandon Guido<sup>1</sup>

valentin.chavarria@ucr.ac.cr, randall.chavesleon@ucr.ac.cr, brandon.guido@ucr.ac.cr,

## RESUMEN

Para el correcto funcionamiento de una aseguradora es necesario tener conocimiento adecuado sobre el comportamiento de las pérdidas. En este trabajo se realiza un ajuste de la distribución de pérdidas para un seguro que cubre vehículos automotores. Los datos usados provienen de pólizas emitidas entre el 2004 y 2005. Para la calibración se opta por un enfoque de cópulas paramétricas. En primer lugar, se ajusta de manera separada la frecuencia y severidad. Posteriormente se unen ambas distribuciones con el uso de diversas cópulas que permite la simulación y obtención de información sobre las pérdidas de la aseguradora. Con este método se ajustó una binomial negativa y log-normal para las distribuciones marginales. Estas distribuciones fueron seleccionadas a través de pruebas de hipótesis (Chi-cuadrado, Kolmogorov-Smirnoff) y del estadístico AIC. Se realizó un total de cuatro ajustes con cópulas distintas. A través de estas se obtuvo cuatro candidatos a distribución para las pérdidas de la empresa. Posteriormente se evaluó el monto del VaR y CVaR correspondiente a la cartera de seguros. De los resultados anteriores se observó que el uso de cópulas lograba replicar de manera satisfactoria el comportamiento de los datos, asimismo como de su dependencia, sin caer en un costo computacional demasiado elevado. Esto agrega peso a la idea de implementar de manera más cotidiana el uso de cópulas en este tipo de problemas y no limitarse a métodos usuales de combinación de distribuciones.

**PALABRAS CLAVE:** Vehículo, estimación, cópulas, simulación.

## INTRODUCCIÓN

La adquisición de un automóvil es, para la mayoría de las personas, una inversión de alto valor monetario. Por eso lo ideal es siempre contar con un seguro que pueda respaldar cualquier eventualidad no esperada que se le presente al propietario del vehículo. Tener acceso a un seguro automotriz tiene la capacidad de proteger ante asaltos, accidentes o inclusive la pérdida total del vehículo.

Debido a esto es importante que exista un mercado establecido de seguros y una cultura que promueva la adquisición de estos. Por otro lado, para que una compañía de seguros pueda mantenerse competitiva en el mercado de seguros necesita tener muy bien identificado sus

posibles escenarios, y trabajar al margen de estos. En esta investigación se estudia la forma para poder modelar las pérdidas por reclamación de siniestros que podría presentar la compañía.

Este tipo de problemas ha sido abordado ampliamente dentro de la literatura. Uno de los enfoques más trabajados es el del ajuste de distribuciones a la frecuencia y severidad a través de estimación paramétrica (Omari, Nyambura, & Mwangi, 2018). Por lo general se utiliza inferencia estadística y algoritmos relacionados a la verosimilitud para la selección de modelos. En este trabajo se menciona que usualmente las distribuciones negativas binomiales y geométricas pueden ser buenos ajustes, asimismo como la log-normal en el caso de la severidad.

Otro enfoque es el del uso de cópulas. Al igual que en el caso anterior se inicia con un ajuste a las distribuciones de la frecuencia y severidad (en este caso llamadas marginales), posteriormente se une ambas distribuciones con una cópula. Esto permite que se pueda asumir dependencia entre los reclamos y el tamaño de los reclamos (Czado, Kastenmeier, Brechmann, & Min, 2012). De acuerdo con los autores, esto es un factor que no debería ser ignorado en la modelación. Posterior a eso se puede utilizar mecanismos de simulación para lograr obtener datos referentes a las pérdidas de la empresa; como el valor esperado u otros de mayor interés económico como el VaR o el CVaR.

Este trabajo abarca una descripción teórica de los modelos a utilizar en la metodología. Posteriormente se detalla los resultados obtenidos y a partir de esto se elabora una sección de conclusiones que resume el comportamiento obtenido de las cópulas, asimismo como recomendaciones para futuros trabajos.

### METODOLOGÍA

En este trabajo se utiliza los datos del libro *Generalized linear models for insurance data*. Dicha base de datos cuenta con 10 variables y 67856 observaciones. Estas variables se basan en información sobre las pólizas de seguro para vehículos vendidos entre 2004 y 2005. Las variables corresponden a:

1. veh\_value: valor del vehículo en escala de 10000 dolares.
2. exposure : valor entre 0 y 1 que representa la exposición del asegurado.
3. clm : ocurrencia del reclamo (variable booleana).
4. numclaims : número de reclamos.
5. claimcst0 : monto de reclamos.
6. veh\_body : variable categórica sobre el tipo de vehículo. En la tabla 7 se puede ver el detalle de las categorías encontradas en esta variable.
7. veh\_age : variable categórica sobre la edad del vehículo.

## SEGURO DE ACCIDENTES AUTOMOVILÍSTICOS

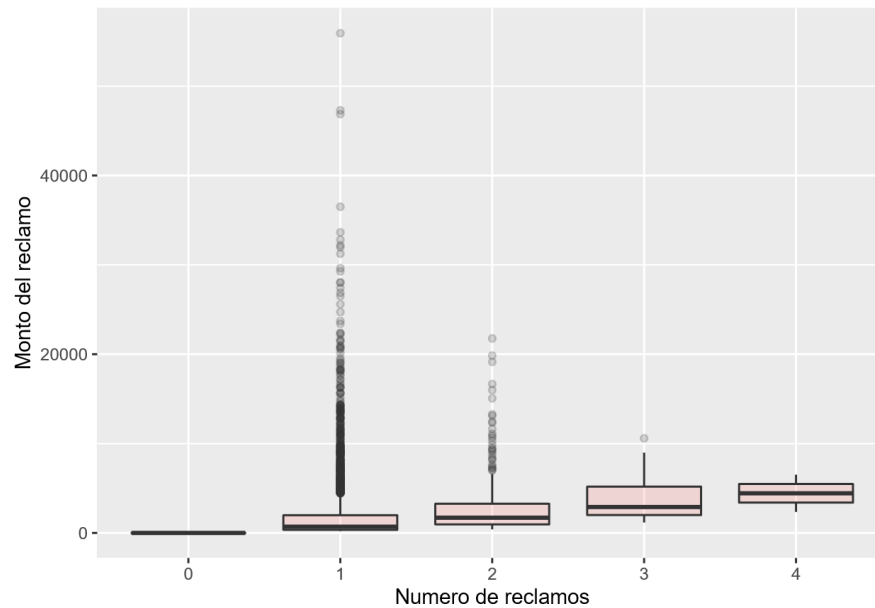
8. gender : género del conductor; M, F.
9. area: área de residencia del conductor; A, B, C, D, E, F.
10. agecat : categoría de edad del conductor; 1 (el más joven), 2, 3, 4, 5, 6.

**Tabla 1**

*Cuadro resumen de estadísticos de distribución para frecuencia y severidad*

Variable	Mínimo	Primer cuantil	Media	Mediana	Tercer cuantil	Máximo
Monto de reclamo	0	0	137.270167	0	0	55922.13
Cantidad de reclamos	0	0	0.072757	0	0	4.00

De acuerdo con la tabla obtenida, la población de estudio corresponde a los reclamos de pólizas por concepto de accidentes. De manera específica se tiene que las variables de estudio van a ser la frecuencia y severidad de los reclamos presentes en la base de datos.



**Figura 1**

*Gráfico de cajas del monto del reclamo agrupados según la cantidad de reclamos*

Asimismo, se realiza una serie de análisis estadísticos de los datos. Donde gracias a la figura 1 y la tabla 1 se detecta que las variables poseen gran acumulación en valores inferiores, donde la mayor acumulación se presenta en 0. Además, es posible observar la presencia de outliers. Todo esto puede causar problemas al momento de estimación, debido a que se genera una fuerte tendencia, se utilizan modelos que ayuden a mitigar estos inconvenientes.

En la tabla 1 se nota como los primeros tres cuantiles para las variables con los montos reclamados y la cantidad de reclamos es de 0. En el caso de los montos, se observa además una

## SEGURO DE ACCIDENTES AUTOMOVILÍSTICOS

media de 137.27, esto da indicios de que los datos tienen un comportamiento con alta dispersión. Lo anterior se confirma al observar la tabla 2, en esta se muestra algunos estadísticos de dispersión, en el que destaca el alto valor de la varianza, asimismo como el de la desviación estándar. Con el coeficiente de variación se verifica que los montos representan una mayor variabilidad que las cantidades de reclamos.

**Tabla 2**

*Estadísticos de dispersión para frecuencia y severidad*

Distribución	Desviación estándar	Coeficiente de variación
Severidad	1056.2977675	7.695028
Frecuencia	0.2782038	3.823739

**Tabla 3**

*Frecuencia según tipo de vehículo y condición de reclamo*

Tipo de vehículo	Sin reclamo	Con reclamo	Porcentaje de los reclamos
BUS	39	9	0.19
CONVT	78	3	0.06
COUPE	712	68	1.47
HBACK	17651	1264	27.34
HDTOP	1449	130	2.81
MCARA	113	14	0.30
MIBUS	674	43	0.93
PANVN	690	62	1.34
RDSTR	25	2	0.04
SEDAN	20757	1476	31.92
STNWG	15088	1173	25.37
TRUCK	1630	120	2.60
UTE	4326	260	5.62

La tabla 3 corresponde a una de contingencia. En esta se realiza el conteo de acuerdo al tipo de vehículo y la cantidad de contratos con reclamo y sin reclamo que tiene guardada la base de datos de la aseguradora. Asimismo, se agrega el porcentaje de los reclamos que está para cada uno de los 13 tipos de vehículo. En este último cálculo sale a relucir que estos son bastante diferentes entre sí, siendo los vehículos de clase *HBACK*, *SEDAN* y *STNWG* los que presentan la mayoría de reclamos dentro de la base de datos. Esto puede afectar el funcionamiento de los modelos que se utilicen para la calibración de distribuciones de pérdidas, ya que se tiene indicios de correlación entre la variable del tipo de vehículo y la cantidad de reclamos.

Con todo esto estudiado es necesario plantear el método para desarrollar la estimación paramétrica mediante la verosimilitud. En un modelo paramétrico  $f(x, \theta)$ , la función de verosimilitud conecta los datos observados y el modelo de probabilidad. De esta forma se puede realizar inferencia estadística sobre  $\theta$ . Para eso primero se resalta la ley de verosimilitud, la cual dice que sea  $X_1, \dots, X_n$  una muestra de variables aleatorias independientes idénticamente distribuidas con la misma función de densidad  $f(x, \theta)$  un espacio de parámetros  $\Theta$ . Para  $\theta \in \Theta$ , entre mayor sea el valor de  $L(\theta)$ , más concuerda con los datos observados,  $L(\theta) = \prod_{i=1}^n f(X_i | \theta)$  es la función de verosimilitud. Entonces, se debe seleccionar como estimador de  $\theta$  el valor de  $\theta \in \Theta$  que maximice la función de verosimilitud (Rossi, 2018). Es decir,

$$\hat{\theta}(X) = \operatorname{argmax} L(\theta)$$

Y como el logaritmo es una función estrictamente creciente, entonces maximizar  $L(\theta)$  es equivalente a maximizar el  $\log L(\theta)$ . La función de log verosimilitud es

$$l(\theta) = \ln L(\theta) = \ln \prod_{i=1}^n f(x_i | \theta)$$

Este método, también conocido como MLE, es un método que se utiliza muy seguido en la estimación de una amplia variedad de problemas. Normalmente, presenta mejores estimaciones comparadas con las de otros métodos, como el método de estimación por mínimos cuadrados, el método de cuantiles y el método de momentos. Esto se debe a que este método utiliza toda la información sobre los parámetros contenidos en los datos y genera un estimador flexible con mejores propiedades asintóticas (Rossi, 2018). Esto se da principalmente si la muestra es grande.

Ahora bien, aunque la teoría comentada previamente es sumamente enriquecedora para la modelación de pérdidas, es importante recalcar la necesidad de inclusión de modificaciones. Cuando se diseña un seguro, por lo general se asume que la cantidad de personas que se vean afectadas por el siniestro en cuestión no debería llegar a ser la mayoría. Si la aseguradora no delimita esto de manera adecuada y sucede el escenario donde hay una gran cantidad de reclamos, es probable que esta no sea vea capacitada para hacerle frente a las obligaciones financieras asociadas. Por esta razón, es usual que a la hora de examinar bases de datos con reclamos de seguros, se detecte que una gran parte de estas observaciones en realidad no presentan reclamos, es decir, representan costos cero.

Por ello, dentro del campo actuarial, se plantea la teoría de modificaciones a las distribuciones usuales, de manera que se pueda tomar en consideración este efecto. En el trabajo de Klugman, Panjer, and Willmot (2012), los autores exponen como a la hora de ajustar distribuciones usuales puede ocurrir que estas no permiten capturar bien el comportamiento de los datos, principalmente el lado izquierdo de la distribución y de manera particular para el caso de 0. Ante esto surge la clase de distribuciones  $(a, b, 1)$ ; en estas distribuciones se tiene que se puede eliminar el peso de probabilidad en cero (distribuciones cero-truncadas) o se puede modificar el peso que se asigna en este punto (distribuciones cero-modificadas). Ambas se van a ajustar a

escenarios particulares y es necesario manejar la teoría detrás de ambos modelos para aumentar las probabilidades de éxito en el ajuste de distribuciones relacionadas con la frecuencia.

Para verificar los ajustes se utilizaron principalmente dos pruebas, la prueba de Kolmogorov-Smirnov para la severidad y la prueba Chi-Cuadrado para la frecuencia Klugman et al. (2012). Las hipótesis de estas pruebas son,  $H_0$ , Los datos provienen de la población establecida en el modelo.

El estadístico de la prueba Kolmogorov-Smirnov es  $D = \max_{t \leq x \leq u} |F_n(x) - F^*(x)|$ , donde  $t$  es el truncamiento izquierdo y  $u$  es el punto de censura derecha. El estadístico de la prueba Chi-cuadrado es  $\chi^2 = \frac{1}{d} \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k}$ , donde  $E_k$  es el número esperado de observaciones en el intervalo y  $O_k$  es el número de observaciones en el intervalo.

Como método de unión de estas distribuciones se plantea utilizar cópulas para obtener distribuciones bivariadas de frecuencia y severidad. Kramer, Brechmann, Silvestrini, and Czado (2013) mencionan que una cópula bivariada  $C : [0,1] \times [0,1] \rightarrow [0,1]$  es una función de distribución acumulativa en  $[0,1] \times [0,1]$  con las marginales distribuidas uniformemente. En el caso bivariado, se establece que para toda función de distribución conjunta  $F_{X,Y}$  de una variable aleatoria bivariada con funciones de distribución marginales  $F_X$  y  $F_Y$ , existe una cópula bivariada  $C$  tal que  $F_{X,Y}(x, y) = C(F_X(x), F_Y(y))$ . Se enfocó la investigación en cuatro distintas familias de cópulas, la Clayton, la Gumbel, la t Student y la Frank.

Este trabajo se realizó a través del lenguaje de programación R (R Core Team, 2013) y se usó paquetes como ggplot2 (Wickham, 2016), tidyverse (Wickham y col., 2019) copula (Hofert y col., 2022) y actuar (Dutang y col., 2008).

## RESULTADOS

Dentro del análisis exploratorio de los datos se mostró que la proporción de observaciones dentro de la base en las que no hay reclamos es sumamente alta. Esta característica implicaba que los ajustes en el que se incluía la completitud de los datos no era satisfactorio. Por esta razón se decide la implementación de distribuciones de la clase  $(a, b, 1)$  para la frecuencia. En particular, las truncadas en cero. Para esto se utilizó como candidatos iniciales a la poisson, la binomial negativa y la geométrica. En el caso de la severidad se ajustó de igual manera solo las observaciones estrictamente positivas de la información.

La selección previa de distribuciones discretas para la frecuencia se realiza con base en lo observado en la literatura para trabajos con objetivos y condiciones similares. Para determinar la que obtuvo el mejor ajuste sobre los datos se utiliza el valor de AIC, asimismo como de los resultados de la prueba chi cuadrado. Los resultados de estas pruebas están en la tabla 4. Entonces, al calibrar se obtuvo que para ambos criterios el modelo favorecido corresponde al que utiliza una distribución binomial negativa truncada en cero. Resalta el hecho de que la distribución geométrica también obtiene resultados adecuados, esto se debe a que esta corresponde a un caso particular de la negativa binomial. De esta manera, se tiene que para la distribución marginal a incluir dentro

de la cópula se toma a la binomial negativa como representante del factor de frecuencia de las pérdidas de la empresa aseguradora.

**Tabla 4**

*Comparación de estadísticos para las distribuciones de frecuencia*

Distribución	AIC	Valor p
zt geométrica	2334.43	0.851
zt negativa binomial	2336.13	0.959
zt poisson	2339.60	0.028

El segundo componente a tomar en cuenta para el ajuste de la cópula corresponde a la distribución a utilizar para la severidad. En este apartado se realiza una transformación inicial de los datos de la base. Esto se hace con el objetivo de que la variable a modelar corresponda a la pérdida por reclamo promedio. De esta manera, la relación entre las pérdidas y las variables que se está modelando corresponden a un producto.

Para modelar la severidad se tuvo como candidatos a las distribuciones de Frechet, log-normal y weibull. La idea de utilizarlas surge debido a sus propiedades de colas más pesadas, lo que permitiría capturar mejor valores elevados que podrían calificar de *outliers* en otras distribuciones de uso común, por ejemplo, la normal. Dentro del análisis exploratorio se encontró que había elementos dentro de la severidad que tenían valores muy elevados respecto al grueso de los datos, por lo que era necesario tener este aspecto en consideración. Además, se tiene también la particularidad de que los datos son números estrictamente positivos, por lo que esto también debía ser tomado en cuenta. Al realizar análisis sobre los modelos obtenidos para estas distribuciones se observa en la tabla 5 que la que aporta un mejor comportamiento y verosimilitud es la log-normal. Esto es consistente con lo encontrado dentro de trabajos previos.

**Tabla 5**

*Comparación del AIC y  $D_{KS}$  para las distribuciones de severidad*

Distribución	AIC	D
Weibull	78519	0.177
Log-Normal	77153	0.105
Frechet	76149	0.932

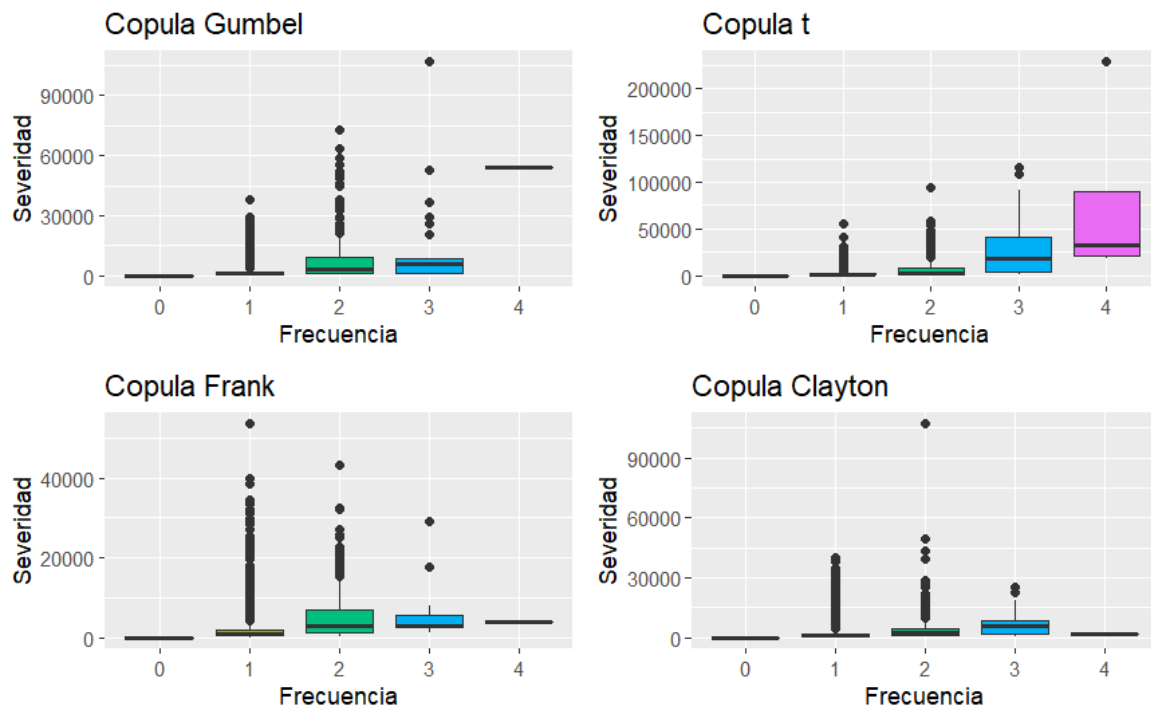
Posteriormente, con las distribuciones marginales que se obtuvieron, se procedió al ajuste de cópulas. En este paso se decide utilizar cuatro tipos distintos de cópula; gumbell, Frank, Clayton y una elíptica de tipo t de Student. La selección de estas se hizo basado en lo encontrado en literatura relacionada. Además, el hecho de incluir varios modelos permite que exista apertura a la inclusión de el criterio experto de la aseguradora en la selección de la cópula que refleje de manera

## SEGURO DE ACCIDENTES AUTOMOVILÍSTICOS

adecuada la situación particular de la empresa. Por otra parte, permitió comparar los resultados y ver la variabilidad que se obtuvo debido a la selección de uno u otro modelo de cópula.

Para el ajuste se utiliza como parámetro la correlación empírica de Kendall. De esta manera se incluye el efecto que tiene la dependencia. Este apartado es vital, pues permite desligarse del supuesto de independencia que se parte en el ajuste usual para distribuciones de frecuencia y severidad.

Finalmente, se implementó un método de simulación de variables aleatorias para incluir el efecto de cuando no hay reclamos en el seguro. De esta manera se tiene que la variable de pérdida corresponde a una composición entre la cópula para datos cuando hay un reclamo y otra variable aleatoria con valor cero que representa los escenarios donde la simulación indica que no hubo reclamo. En la figura 2 se puede observar el gráfico de cajas de las simulaciones cuando se utilizó las distintas cópulas. La cópula t Student no conservó la estructura de la base datos original, por otro lado, los datos simulados con la cópula Clayton conservó de mejor forma la estructura original.



**Figura 2**

*Gráficos de caja para las pérdidas obtenidas por simulación*

Con estas simulaciones de pérdidas es posible entonces realizar cálculos de utilidad para la empresa aseguradora. Tal es el caso del VaR y el CVaR. En la tabla 6 se observa los valores obtenidos para los estadísticos mencionados. En ellos destaca que la cópula de Clayton y de Frank son las que mantienen valores más cercanos al cálculo de VaR no paramétrico. También se resalta que todos los estimadores de VaR paramétricos son inferiores al empírico. Esto se puede deber a que las



distribuciones de severidad que se utilizaron tuvieron problemas para capturar el comportamiento que tenía la base al haber valores máximos extremadamente diferentes al resto de los datos. El caso del CVaR mantiene un comportamiento similar, además destaca el elevado monto que contiene. Este duplica el estimado del VaR para cualquiera de los métodos seleccionados, lo que una vez más se puede entender ante la presencia de valores extremos bastante considerables.

**Tabla 6**

*Comparación de estadísticos por tipo de cópula con VaR al 99%*

<b>Modelo</b>	<b>VaR</b>	<b>CVaR</b>
Gumbell	3117.71	7713.73
T student	3066.63	8354.83
Clayton	3224.53	7072.49
Frank	3211.18	7122.43
Empírico	3625.28	8299.33

## CONCLUSIONES

La función de las asegurados como ente controlador del riesgo es fundamental dentro de las sociedades contemporáneas. La existencia de seguros le permite a las personas afrontar siniestros de manera que no se vea amenazada su capacidad adquisitiva ni el acceso a calidad de vida. Con esta idea en mente, se planteó en el trabajo la tarea de calibrar modelos que permitan a una aseguradora de automóviles obtener datos necesarios para su funcionamiento.

Como una primer conclusión se obtuvo que la base de datos mantiene un comportamiento similar a lo encontrado en trabajos que abordaron problemas similares. Las distribuciones que se adaptaron mejor para la frecuencia (binomial negativa) y la severidad (log-normal) coincidieron con las recomendadas.

Además, se encontró que el uso de cópulas es una herramienta estadística sumamente útil para resolver este tipo de problemas. Por una parte se logró simular el comportamiento manifestado por la base de datos original con bastante éxito. Asimismo, se obtuvo modelos que incorporaron dentro de su lógica la dependencia existente entre la cantidad de siniestros y el monto que es cancelado por los mismos.

Finalmente, también se mostró que la influencia del tipo de cópula a utilizar no afectaba el resultado de manera considerable. De las cuatro cópulas que se utilizaron, se observó que los valores de las arquimedianas se mantuvieron cercanos entre sí. La elíptica t de Student se alejó más de los valores empíricos, lo que sugiere que en los datos convendría usar cópulas de la familia arquimediana.

Para futuros trabajos se recomienda incluir un catálogo más amplio de cópulas. Como se mencionó, hubo indicio de que el uso de cópulas arquimedianas podía acoplarse mejor con los datos. Asimismo, debería tenerse en cuenta la división de los cálculos de seguro de acuerdo al tipo

de vehículo, pues el problema de valores extremos se debió a la presencia de vehículos heterogéneos. Una tercer recomendación sería el ampliar el estudio del comportamiento de la cola derecha con distribuciones de valor extremo para excesos sobre un umbral. Una posible aplicación sería tomar el VaR como umbral y analizar el comportamiento de las excepciones con una distribución de este tipo.

### AGRADECIMIENTO

Externamos un agradecimiento al profesor Maikol Solís por la colaboración en la corrección, revisión y guía metodológica en la elaboración de este trabajo.

### REFERENCIAS

- Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, 2012(4), 278–305.
- Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: From data to decisions* (Vol. 715). John Wiley & Sons.
- Kramer, N., Brechmann, E. C., Silvestrini, D., & Czado, C. (2013). Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3), 829–839.
- Omari, C. O., Nyambura, S. G., & Mwangi, J. M. W. (2018). *Modeling the frequency and severity of auto insurance claims using statistical distributions*.
- Rossi, R. J. (2018). *Mathematical statistics: An introduction to likelihood based inference*. John Wiley & Sons.
- Hofert M., Kojadinovic I., Maechler M. & Yan J. (2022). copula: Multivariate Dependence with Copulas. R package version 1.1-1 URL <https://CRAN.R-project.org/package=copula>.
- Dutang C, Goulet V, Pigeon M (2008). “actuar: An R Package for Actuarial Science.” *Journal of Statistical Software*, 1-37. doi:10.18637/jss.v025.i07. <https://doi.org/10.18637/jss.v025.i07>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

**ANEXOS**

**Tabla 7**

*Tipos de vehículo contenidos en la base de datos*

Código	Tipo de vehículo
BUS CONVT	Convertible
COUPE HBACK	Auto hatchback
HDTOP	<i>Hardtop</i>
MCARA	Caravana
MIBUS	Minibus
PANVN	Van
RDSTR	<i>Roadster</i>
SEDAN	Automovil
STNWG	Vagón de tren
TRUCK TE	Camión