



UNIVERSIDAD
COMPLUTENSE
MADRID

**FACULTAD DE CIENCIAS ECONÓMICAS Y
EMPRESARIALES**

**MÁSTER EN CIENCIAS ACTUARIALES Y
FINANCIERAS**

TRABAJO DE FIN DE MÁSTER

TÍTULO: Modelización e identificación de siniestros de valor extremo

AUTOR: Gabriel Berjano Rosado

TUTORES: José María Lorenzo Magán & Enrique Riego Miedes

CURSO ACADÉMICO: 2023 - 2024

CONVOCATORIA: Septiembre

DECLARACIÓN DE NO PLAGIO

D./Dña. **Gabriel Berjano Rosado** con NIF _____, estudiante de Máster en la Facultad de Ciencias Económicas y Empresariales de la Universidad Complutense de Madrid en el curso **2023 - 2024**, como autor/a del trabajo de fin de máster titulado **Modelización e identificación de siniestros extremos** y presentado para la obtención del título correspondiente, cuyo/s tutor/es son: **José María Lorenzo Magán y Enrique Riego Miedes**

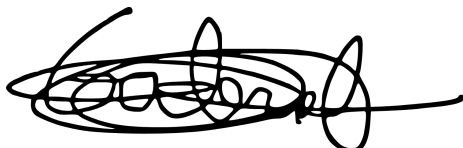
DECLARO QUE:

El trabajo de fin de máster que presento está elaborado por mí y es original. No copio, ni utilizo ideas, formulaciones, citas integrales e ilustraciones de cualquier obra, artículo, memoria, o documento (en versión impresa o electrónica), sin mencionar de forma clara y estricta su origen, tanto en el cuerpo del texto como en la bibliografía. Así mismo declaro que los datos son veraces y que no he hecho uso de información no autorizada de cualquier fuente escrita de otra persona o de cualquier otra fuente.

De igual manera, soy plenamente consciente de que el hecho de no respetar estos extremos es objeto de sanciones universitarias y/o de otro orden.

En **Madrid**, a **13** de **septiembre** de **2024**

Fdo.:



AGRADECIMIENTOS

En primer lugar, quiero agradecer tanto a mi madre como a mi padre toda la confianza que han depositado en mí todos estos años y todo el sacrificio y el trabajo incondicional que han hecho para conseguir que esté hoy aquí escribiendo esto, sin ellos no habría sido posible.

Para continuar, me gustaría dedicarles este trabajo a mis abuelos, tanto para los que aún siguen conmigo como para los que por desgracia ya no están a mi lado.

Por último, quería darles las gracias a mis tutores, José María y Enrique, por el gran trato y los consejos que he recibido de su parte a la hora de realizar este trabajo y por la sabiduría que me han transmitido a lo largo del máster.

RESUMEN

Este TFM trata la teoría de los valores extremos: qué es, cuándo aplicarla, cómo aplicarla, métodos de aplicación, identificación de valores extremos y modelización de los mismos mediante ciertas distribuciones. Para todo ello, se habla de la distribución generalizada del valor extremo (GEV) y de sus variantes Fréchet, Weibull y Gumbel y de la distribución generalizada de Pareto (GPD). Asimismo, se procede posteriormente a aplicar dicha teoría sobre un caso práctico donde se identifican los valores extremos por varios métodos, se estiman los parámetros de la distribución mediante Hill y/o máxima verosimilitud y se ajusta la distribución a alguna de las anteriores. Se emplearán métodos tanto analíticos como gráficos y se verificará por procedimientos de bondad de ajuste.

Palabras clave

Colas, distribución de Pareto generalizada, distribución generalizada del valor extremo, Fréchet, Gumbel, máximo, siniestros, umbral, valores extremos, Weibull.

Abstract

The theory of extreme values is discussed: what it is, when to apply it, how to apply it, application methods, identification of extreme values and their modelling by means of certain distributions. To this end, the generalized extreme value distribution (GEV) and its variants Fréchet, Weibull and Gumbel and the generalized Pareto distribution (GPD) are discussed. The theory is then applied to a practical case where extreme values are identified by various methods, the distribution parameters are estimated by the Hill method and/or by maximum likelihood, and the distribution is adjusted to one of the above. Both analytical and graphical methods will be used, and it will be verified by goodness-of-fit procedures.

Key words

Claims, extreme values, Fréchet, generalized extreme value distribution, generalized Pareto distribution, Gumbel, maximum, tails, threshold, Weibull.

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN.....	1
1.1 Motivación.....	2
1.2 Objetivos	2
1.3 Materiales y métodos.....	2
2. TEORÍA DE LOS VALORES EXTREMOS	3
2.1 Método de máximos de bloque (BM)	4
2.1.1 Distribución de valores extremos generalizada	5
2.1.1.1 Gumbel.....	7
2.1.1.2 Fréchet	7
2.1.1.3 Weibull	8
2.2 Método de excesos sobre un umbral (POT)	8
2.2.1 Distribución de Pareto generalizada	9
2.3 Estimación de los parámetros de la distribución de la cola	10
2.3.1 Máxima verosimilitud	10
2.3.2 Estimador de Hill	11
3. ANÁLISIS PRÁCTICO	12
3.1 Presentación de la base de datos.....	12
3.2 Identificación del umbral	17
3.2.1 Método de Tukey	17
3.2.2 Función empírica de exceso medio (ME-PLOT)	18
3.2.3 Gráfico de Hill	19
3.2.4 Estimación visual / subjetiva	20
3.3 Estudio de la cola de la distribución	20
3.3.1 Parámetros de la cola.....	20
3.3.1.1 Máxima verosimilitud.....	20

3.3.1.2 Estimador de Hill	20
3.3.2 Distribución generalizada de valores extremos	20
3.3.3 Distribución de Pareto generalizada	21
3.4 Prueba de Kolmogorov-Smirnov	23
4. RESULTADOS Y CONCLUSIONES.....	24
5. LIMITACIONES Y RECOMENDACIONES	25
6. BIBLIOGRAFÍA.....	26

ANEXOS

Anexo I: Base de datos en Excel (anexo digital)

Anexo II: Base de datos modificada sin coste 0 en Excel (anexo digital)

Anexo II: Script en R (anexo digital)

ÍNDICE FIGURAS

Figura 1. Distribución Gumbel (Pérez, 2004).	7
Figura 2. Distribución Fréchet (Pérez, 2004).	8
Figura 3. Distribución Weibull (Pérez, 2004).	8
Figura 4. Pareto Generalizada comparada con la cola de una distribución normal (Victor A. Rico. Teoría de valores extremos, n. d.).	10
Figura 5. Valor del vehículo.	12
Figura 6. Grado de exposición.	13
Figura 7. Indicación de siniestro.	13
Figura 8. Número de siniestros.	14
Figura 9. Cuantía de los siniestros.	14
Figura 10. Tipología del vehículo.	15
Figura 11. Antigüedad del vehículo.	15
Figura 12. Género del propietario.	16
Figura 13. Vehículos por zonas.	16
Figura 14. Edad del propietario.	17
Figura 15. Función empírica de exceso medio.	18
Figura 16. Gráfico de Hill.	19
Figura 17. Función de densidad de la distribución frente a la densidad teórica.	21
Figura 18. Gráfico cuantil-cuantil.	21
Figura 19. Excesos del umbral.	22
Figura 20. Distribución de los excesos.	22

ÍNDICE DE TABLAS

Tabla 1. Dominios de atracción (Pérez, 2004).	6
---	---

1. INTRODUCCIÓN

Para el sector asegurador es de vital importancia preocuparse por la solvencia, para así poder continuar ejerciendo cobertura de los riesgos a la sociedad. Por ello, se requiere cierto volumen de primas para poder cubrir la siniestralidad. Esta siniestralidad es incierta y por tanto se va a identificar con una variable aleatoria.

Sin embargo, esta variable aleatoria no puede ser explicada por una única función de distribución de probabilidad. Esto es debido a las colas de la distribución. Para solucionar esto se implementa la Teoría del Valor extremo que es "Una disciplina estadística que desarrolla un conjunto de modelos y métodos tanto paramétricos como no paramétricos con el objeto de describir, cuantificar y modelizar los casos raros los cuales se distribuyen, no bajo la 'ley de los grandes números', sino bajo la ley de los pequeños números" (Pérez, 2004)

Cuando se analiza clásicamente cualquier tipo de datos, la tendencia es ignorar los valores extremos, pero en algunos casos es de vital importancia no hacerlo porque, aunque sean hechos poco probables, tienen un alto impacto.

La teoría del valor extremo postula que el valor más grande de un conjunto de datos tiende a una distribución asintótica que solo depende de la cola de la distribución. En otras palabras, es el estudio de las colas de las distribuciones.

El modelo para el que se desarrolla la teoría de valores extremos está enfocado a describir el comportamiento estadístico de

$$M_n = \max \{X_1, \dots, X_n\}$$

donde X_1, \dots, X_n es una secuencia de variables aleatorias independientes con distribución común F y M_n representa el máximo del proceso sobre n unidades de tiempos de observación.

Primeramente, se asentarán de forma breve las bases de la teoría que rodea al fenómeno de los valores extremos y después se finalizará con un caso práctico y se establecerán unas conclusiones sobre el proceso.

1.1 Motivación

El motivo por que elegí este trabajo es que siempre me ha llamado la atención la estadística y la matemática, sobre todo centradas en el ámbito de los seguros, motivo por el cual ingresé en este máster. Una vez lo estaba cursando descubrí que me apasionaba el estudio de los siniestros y lo decidí enfocar en el estudio de los siniestros extremos, para profundizar más sobre ello y porque pienso que tiene bastante más relevancia de la que se le atribuye.

1.2 Objetivos

En este trabajo, nos vamos a centrar en modelizar siniestros extremos de una cartera de seguros de automóvil. Ajustaremos un modelo para la cola.

Los objetivos de este trabajo son: primeramente, obtener el umbral por el que vamos a identificar los siniestros extremos, y para ello definiremos un valor límite que nos indicará el punto de separación, lo obtendremos por diferentes métodos. En segundo lugar, una vez presentemos las distribuciones que necesitamos para ajustar nuestros datos, se procederá a justar la cola a ellas para identificar cuál ajusta mejor. El modelo se ajustará con los parámetros que proporcione el método de máxima verosimilitud. Por último se hará una prueba de bondad de ajuste para verificar dicho ajuste.

1.3 Materiales y métodos

Los materiales utilizados para elaborar este trabajo han sido Excel y R y los métodos utilizados han sido la bondad de ajuste, la prueba de Kolmogorov-Smirnov, los contrastes de hipótesis, método de Hill, máxima verosimilitud, métodos visuales subjetivos, el método de Tukey y exceso medio.

2. TEORÍA DE LOS VALORES EXTREMOS

Engloba las herramientas estadísticas empleadas para modelar y pronosticar las distribuciones que surgen cuando se estudian eventos extremos.

Es la rama de la estadística que centra su estudio en las colas de la distribución. Nos solemos fijar en la parte central de los datos que es la más probable y descartamos los extremos para que no interfieran en el resultado, pero hay momentos donde esta información es muy útil.

Para su aplicación, se puede dividir la muestra por bloques de cuantía fija donde se va a escoger el máximo valor de cada bloque. Estos valores veremos que ajustarán a lo que denominaremos distribución de valores extremos generalizada.

También se puede fijar un umbral, un umbral es, en la teoría de valores extremos, un valor específico utilizado para identificar los eventos extremos de la muestra de datos, los eventos que lo exceden se consideran eventos extremos. El umbral está sujeto al problema de la varianza y del sesgo, ya que cuanto mayor sea el umbral menor es el número de observaciones y aumenta la varianza del ajuste. Además, si el umbral es demasiado bajo, se incrementa el sesgo al modelizar observaciones que no pertenecen a la cola (Pérez, 2004).

Por tanto, hay dos modelos: ambos modelos difieren en el modo en que cada uno clasifica las observaciones que se consideran eventos extremos.

Para ambos métodos se disponen de dos teoremas fundamentales que permiten caracterizar las distribuciones asintóticas correspondientes:

Para máximos de bloques (Block máxima, BM) tenemos la distribución de valores extremos generalizada (GEV) y para excesos del umbral (Peaks over the threshold, POT) tenemos la distribución de Pareto

generalizada (GDP). Son las dos distribuciones que usaremos para modelizar.

En BM se dividen los datos en bloques de igual tamaño y se seleccionan los datos máximos para cada bloque y en POT se consideran como datos extremos los que exceden cierto umbral.

El reto es identificar el tamaño de los bloques y el valor del umbral, debido a que, si creas pocos bloques, al elegirse el máximo estarías perdiendo valores y si se hace lo contrario sobrecargarías de valores potencialmente no necesario. De forma análoga, si el umbral es un valor muy alto estás perdiendo valores que podrían ser extremos y si es muy bajo estás introduciendo un peso a los valores extremos que no debería de incluirse.

Tras esta breve introducción, se procede a explicar con más profundidad lo anterior.

2.1 Método de máximos de bloque (BM)

En el criterio de máximo de bloques los datos se seleccionan por bloques y el valor máximo de cada uno de estos bloques es tomado como un valor extremo. Ahora bien, bajo este criterio una de las grandes limitaciones es que solamente se obtiene un valor extremo y esto no puede ser enteramente preciso y/o adecuado.

El método máximo de bloques tiene como principal deficiencia, la insuficiencia en el uso de los datos, es decir, se desperdician observaciones que pueden ser consideradas extremas ya que solo se toma un solo punto dentro de un bloque. Por esta razón, en la práctica se ha sustituido en gran medida por métodos basados en excedencias sobre un umbral, en donde se usan todos los datos que son extremos en el sentido de que sobrepasan un determinado nivel alto pre-determinado.

Dada una sucesión de variables aleatorias e idénticamente distribuidas Y_1, \dots, Y_m con función de distribución común F (distribución base) y un tamaño de bloque $k \in \mathbb{N}$, el máximo de bloque será:

$$X_i = \max_{(i-1)k < j < ik} Y_j$$

Con $i = 1, 2, \dots, n$ y $n = \frac{m}{k}$ (número de máximos)

Por tanto, la función de distribución del máximo de bloque, X_i será:

$$P(X_i \leq x) = F(x)^k$$

Por lo que la función de distribución del máximo de bloques es la potencia k -ésima de la función de distribución máxima. Por tanto, la función de distribución del máximo de bloques depende de F que suele ser desconocida en aplicaciones reales. Por ello, consideramos la distribución asintótica.

2.1.1 Distribución de valores extremos generalizada

Lo mejor de esta distribución es que como no sabemos a qué distribución se converge, generaliza diferentes tipos de distribuciones extremas, sin necesidad de conocer la distribución específica a las que convergen los datos. Esta característica ofrece un uso más factible de cálculo y estimadores, permitiendo su aplicación en una gran variedad de contextos con desconocimiento de la forma exacta de la distribución de los valores extremos (Gilli & Këllezi, 2006).

Teorema de Fisher-Tippet-Gnedenko: La distribución asintótica de los máximos de bloque de variables aleatorias e idénticamente distribuidas se puede aproximar mediante la distribución de valores extremos generalizada:

$$GEV(x; \xi, \sigma, \mu) = e^{-\left(1 + \xi \left(\frac{x - \mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}} \quad (1)$$

Donde:

$$\xi, \mu \in \mathbb{R}$$

$$\sigma > 0$$

$$1 + \xi \left(\frac{x - \mu}{\sigma} \right) > 0$$

σ es la escala.

μ es la localización.

ξ es el índice de la cola de la distribución, la forma, cuanto mayor sea más larga es la cola, más pesada es, esto indica que hay una mayor probabilidad de observar valores extremos alejados del umbral. Es lo que hace que la GEV sea Gumbel (colas medias) cuando $\xi = 0$, Fréchet (colas pesadas) cuando $\xi > 0$ o Weibull (colas suaves) cuando $\xi < 0$. (Mario M. Pizarro, 2021)

Dominio de atracción:

Tabla 1. Dominios de atracción (Pérez, 2004).

Distribución inicial $F(x)$	Distribución límite para los máximos $G(x)$
Exponencial Gamma Normal Log-Normal	Gumbell (Tipo I)
Pareto Cauchy Burr Log-Gamma	Fréchet (Tipo II)
Uniforme Beta	Weibull (Tipo III)

Las tres distribuciones se pueden unir en un único modelo, la GEV, haciendo $\xi = 1/\alpha$.

Las siguientes funciones son distribuciones de valores extremos, las límite.

La distribución límite $G(x)$ siempre pertenece a alguna de las 3, sea cual sea $F(x)$.

2.1.1.1 Gumbel

(La primera fórmula contiene posición y escala):

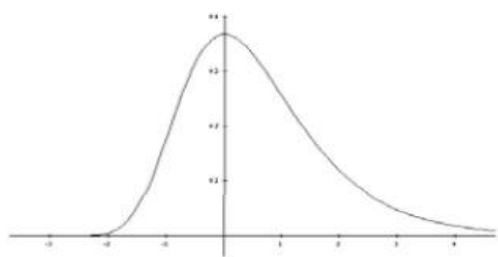
$$G(x; \mu, \sigma) = e^{-e^{-\frac{x-\mu}{\sigma}}} \quad (2)$$

$$\mu \in \mathbb{R}$$

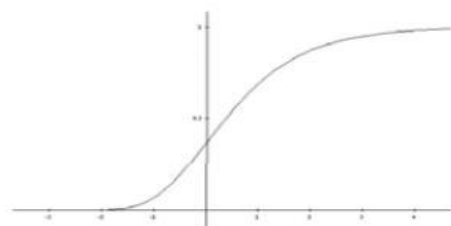
$$\sigma > 0$$

$$\text{Dom}(x) \in \mathbb{R}$$

$$G_0(x) = e^{-e^{-x}}, \quad x \in \mathbb{R} \quad (3)$$



Función de densidad de Gumbel

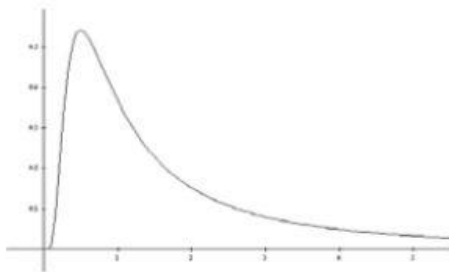


Función de distribución de Gumbel

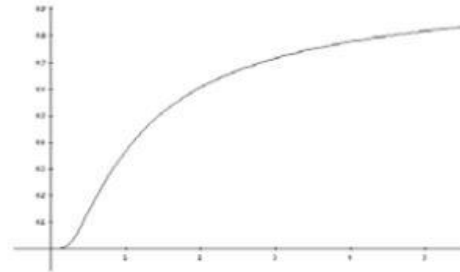
Figura 1. Distribución Gumbel (Pérez, 2004).

2.1.1.2 Fréchet

$$G_1(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ e^{-x^{-\alpha}} & \text{si } x > 0 \end{cases} \quad \alpha > 0 \quad (4)$$



Función de densidad de Fréchet

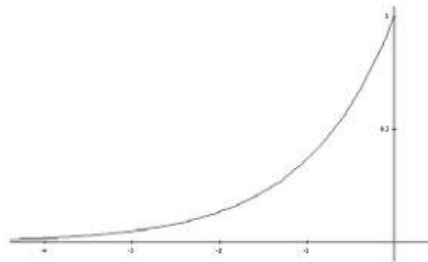


Función de distribución de Fréchet

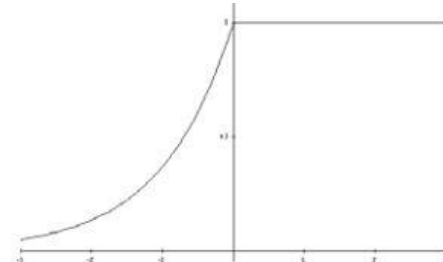
Figura 2. Distribución Fréchet (Pérez, 2004).

2.1.1.3 Weibull

$$G_2(x) = \begin{cases} 1 & \text{si } x > 0 \\ e^{-(-x)^{-\alpha}} & \text{si } x \leq 0 \end{cases} \quad \alpha < 0 \quad (5)$$



Función de densidad de Weibull



Función de distribución de Weibull

Figura 3. Distribución Weibull (Pérez, 2004).

2.2 Método de excesos sobre un umbral (POT)

El método POT es útil para grandes observaciones que exceden un umbral alto u . Este método es más útil que el máximo por bloques en aplicaciones prácticas, debido al uso más eficiente de los datos en

valores extremos. Desde el punto de vista de un inversor, se está interesado en las pérdidas que exceden un umbral u . Dado un conjunto de datos X_1, \dots, X_n de una función de distribución desconocida denotada como F , existe un número aleatorio N_u de pérdidas que excederá el umbral u .

Estos datos se denotan como $\tilde{x}_1, \dots, \tilde{x}_n$.

La selección del umbral u , conlleva una compensación entre sesgo y varianza en la estimación. Valores muy bajos del umbral generan sesgo en la estimación, mientras que valores muy altos del umbral generan alta varianza en la estimación.

2.2.1 Distribución de Pareto generalizada

El problema después de la elección del umbral es ver qué función ajusta a los excesos. Esto se hace condicionando la distribución a que supere el umbral, que denotaremos por u :

Teorema de Pickands, Balkema y de Haan:

Sea X una variable aleatoria con función de distribución F , x_0 el punto final derecho y un umbral $u < x_0$, F^u es la función de los excesos de X sobre u . Esto implica que toda distribución tiene un umbral a partir del cual sus excesos se distribuyen como una distribución de Pareto generalizada (Superintendencia Financiera de Colombia, 2020).

La distribución de los excesos converge en un umbral infinito a una distribución generalizada de Pareto:

$$\lim_{u \rightarrow \infty} P(X - u \leq y | X > u) = G(y; \xi, \sigma) \quad (6)$$

(Pérez, 2004)

$$G_{\xi,\sigma}(y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma}y\right)^{-\frac{1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - e^{-\frac{y}{\sigma}} & \text{si } \xi = 0 \end{cases} \quad (7)$$

(Gilli & K llezi, 2006)

σ es la escala

ξ es el  ndice de la cola de la distribuci n

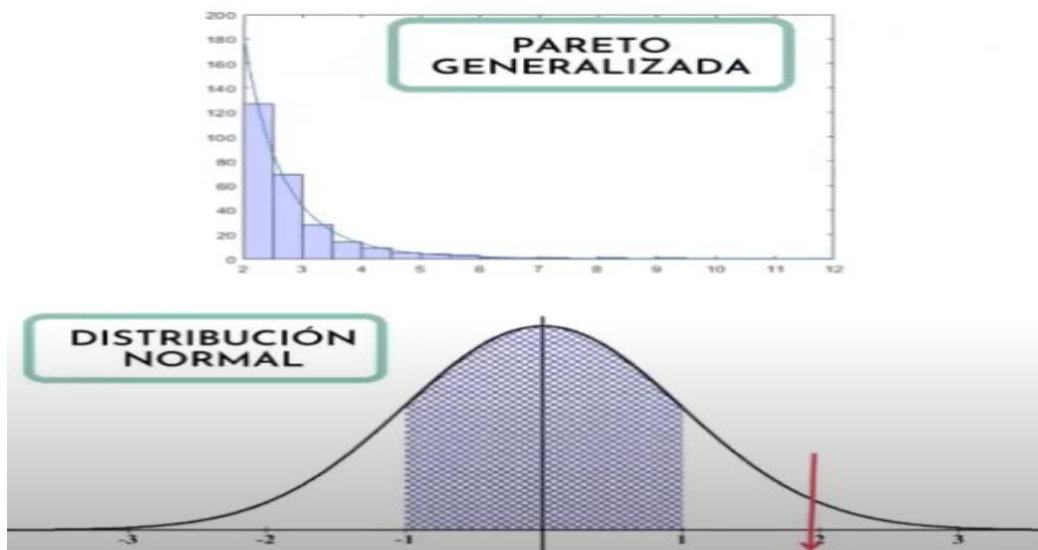


Figura 4. Pareto Generalizada comparada con la cola de una distribuci n normal (Victor A. Rico. Teor a de valores extremos, n. d.).

Se puede apreciar que la cola de la distribuci n es la Pareto.

2.3 Estimaci n de los par metros de la distribuci n de la cola

2.3.1 M xima verosimilitud

Para $\gamma \neq 0$ y $x = (x_1, x_2, \dots, x_n)$, la funci n logar tmica de verosimilitud viene dada, para $\theta = \{\lambda, \delta, \gamma\}$, por:

$$\log(L(\theta; x)) = -n \log(\delta) - (1 + \gamma) \sum_{i=1}^n z_i - \sum_{i=1}^n e^{-z_i} \text{ donde } z_i = \frac{1}{\gamma} \log \left[1 + \gamma \frac{x_i - \lambda}{\delta} \right]$$

Para $\theta = \{\lambda, \delta\}$ y $\gamma=0$, se obtiene:

$$\log(L(\theta; \mathbf{x})) = -n \log(\delta) - \sum_{i=1}^n e^{-\frac{x_i - \lambda}{\delta}} - \sum_{i=1}^n \left(\frac{x_i - \lambda}{\delta} \right) \quad (8)$$

(Abalasei, 2017)

Tras esto solo quedaría usar algún método por el que maximizar estas funciones, como por ejemplo paquetes o funciones en R o mediante cálculo numérico.

2.3.2 Estimador de Hill

Partiendo del método de máxima verosimilitud y en contexto donde se está estudiando la cola de una distribución, es preferible usar los datos que exceden el umbral. Para estos casos, el estimador de Hill para el parámetro $1/\xi$ se define como:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log X_{(n-i+1)} - \log X_{(n-k)} \quad (9)$$

(Mora Valencia, 2010)

3. ANÁLISIS PRÁCTICO

Las figuras y las conclusiones de las salidas del código de R de más abajo son de elaboración propia todas ellas. Se adjunta en los anexos digitales, así como un análisis descriptivo general de la variable y de los extremos de la misma.

Si en algún momento los cálculos difieren para alguna parte en concreto es porque se ha hecho uso de la base de datos alternativa, modificada excluyendo el coste 0.

3.1 Presentación de la base de datos

La base de datos se llama dataCar y es una base de datos basada en las pólizas de seguro de automóvil a un año entre 2004 y 2005.

Existen diversas formas de obtener esta base de datos, como por ejemplo el paquete de R "insuranceData" o la Escuela de Negocios de Wisconsin, aunque el origen de la base de datos es (De Jong P., Heller G.Z. (2008), Generalized linear models for insurance data, Cambridge University Press).

La base de datos se compone de 67856 observaciones y 10 variables:

- Veh_value: Valor del vehículo en decenas de miles de dólares.

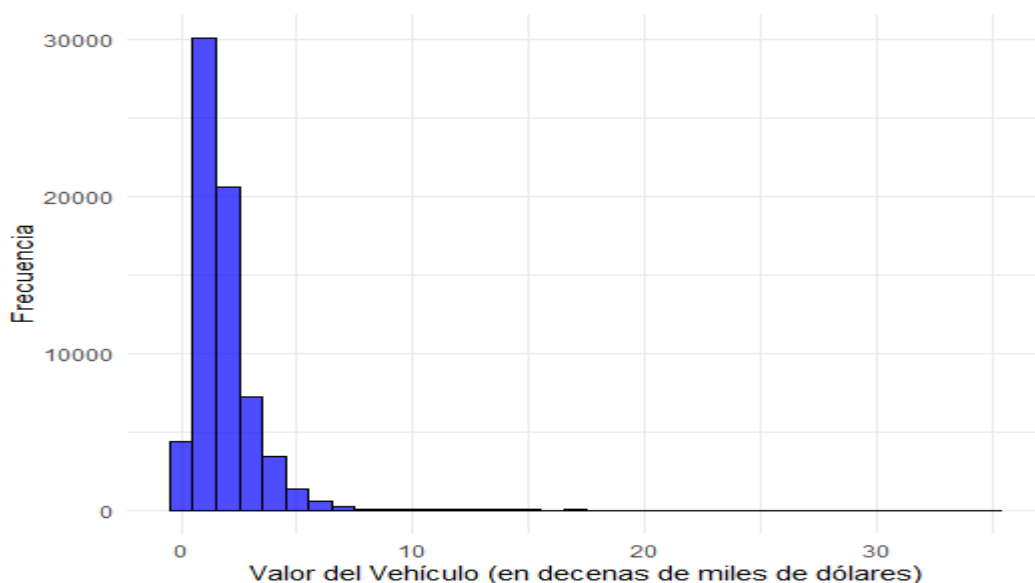


Figura 5. Valor del vehículo.

- Exposure: Grado de exposición del seguro, varía entre 0 y 1 y es la proporción del tiempo en un año que el vehículo está asegurado.

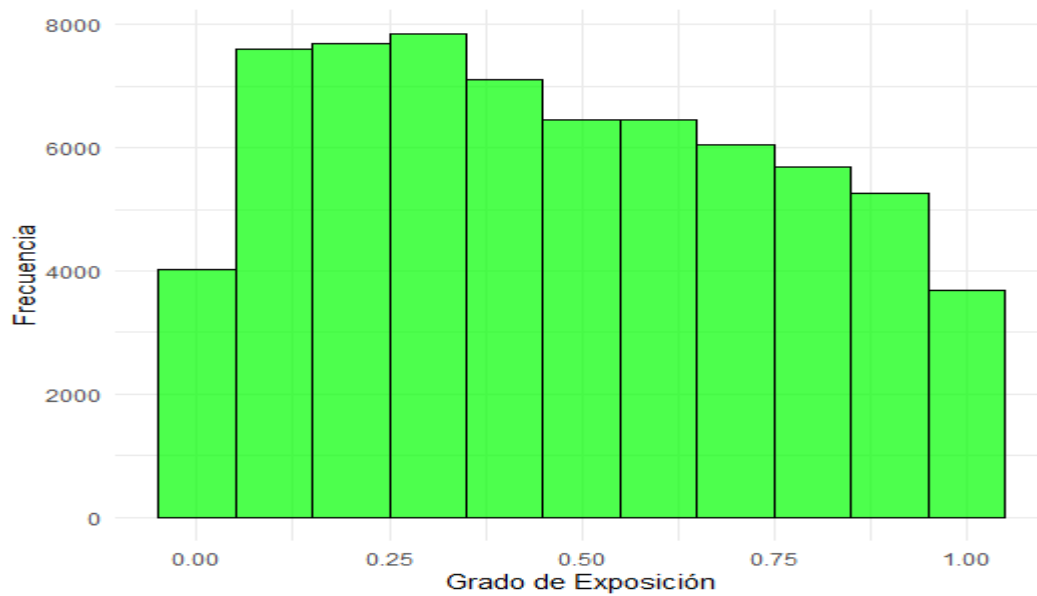


Figura 6. Grado de exposición.

- Clm: Variable dicotómica que indica si ha habido (1) o no (0) un siniestro.

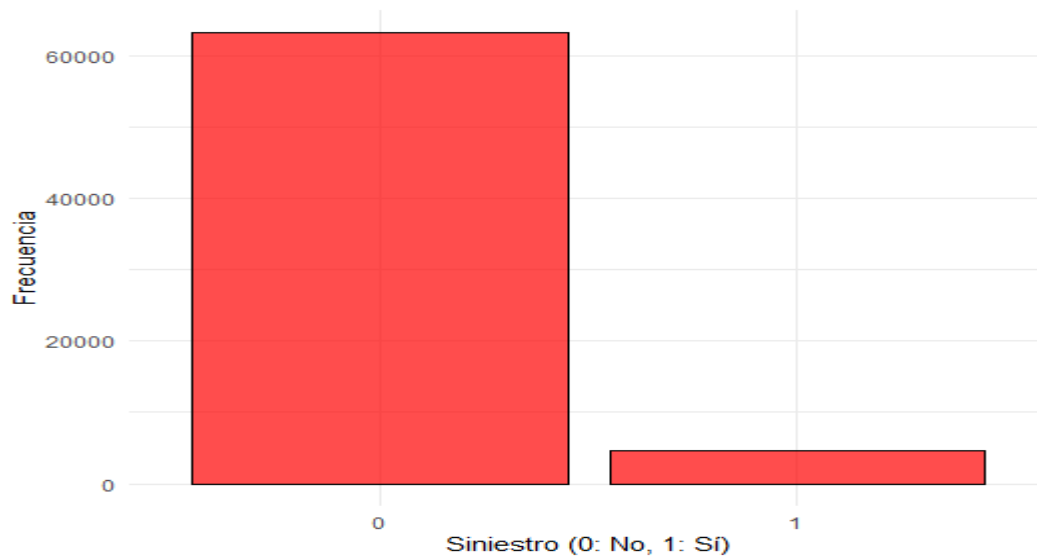


Figura 7. Indicación de siniestro.

- Numclaims: Número de siniestros.

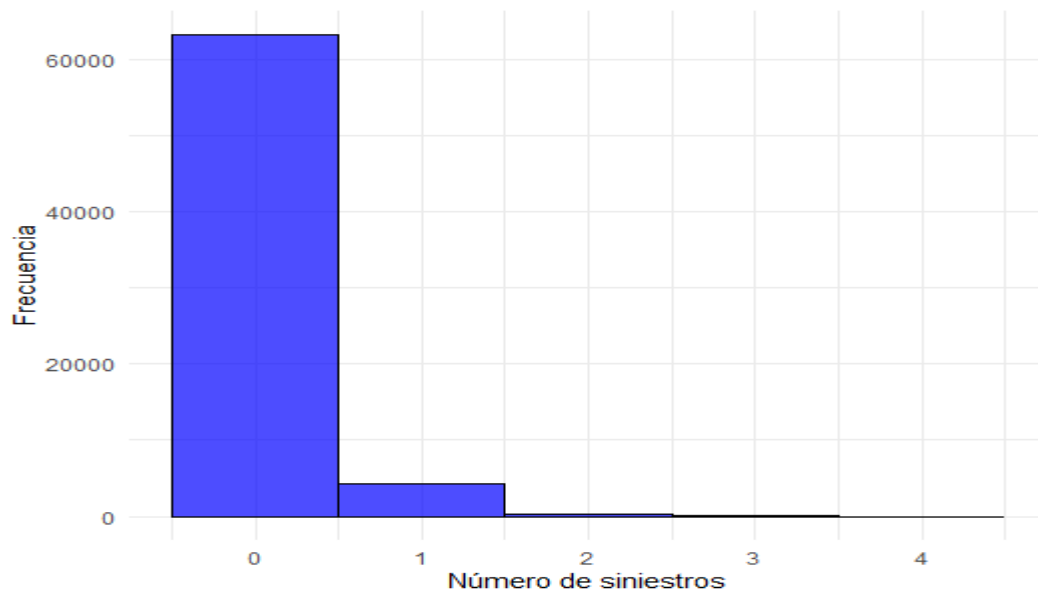


Figura 8. Número de siniestros.

- Claimcst0: Cuantía del siniestro. Es la variable a analizar en nuestro estudio. Se pueden diferenciar en ella 3 tramos: el tramo de siniestro frecuentes de hasta una cuantía de unos 10000, el tramo de siniestros intermedios de hasta unos 20000 y a partir de ahí los siniestros extremos.

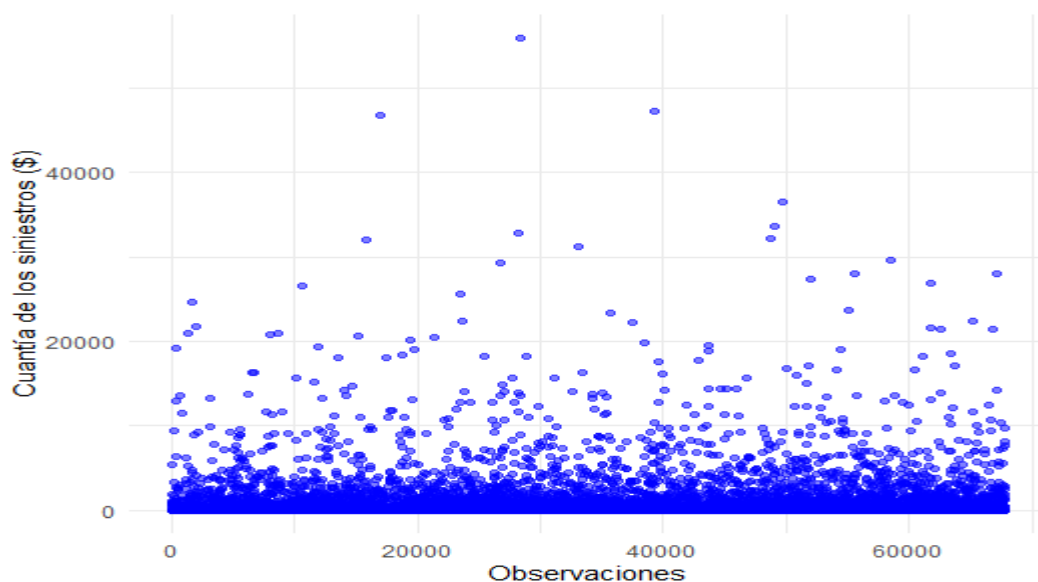


Figura 9. Cuantía de los siniestros.

- Veh_body: Tipología del vehículo codificada como BUS, CONV, COUPE, HBACK, HDTOP, MCARA, MIBUS, PANVN, RDSTR, SEDAN, STNWD, TRUCK y UTE.

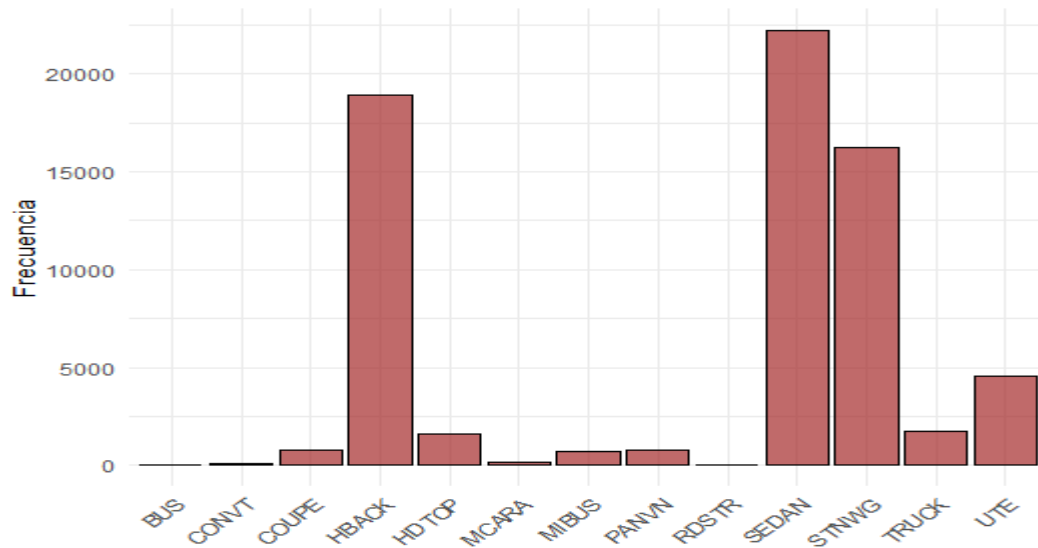


Figura 10. Tipología del vehículo.

- Veh_age: Edad del vehículo, dividida en las categorías 1, 2, 3 y 4, siendo 1 la de menor tiempo.

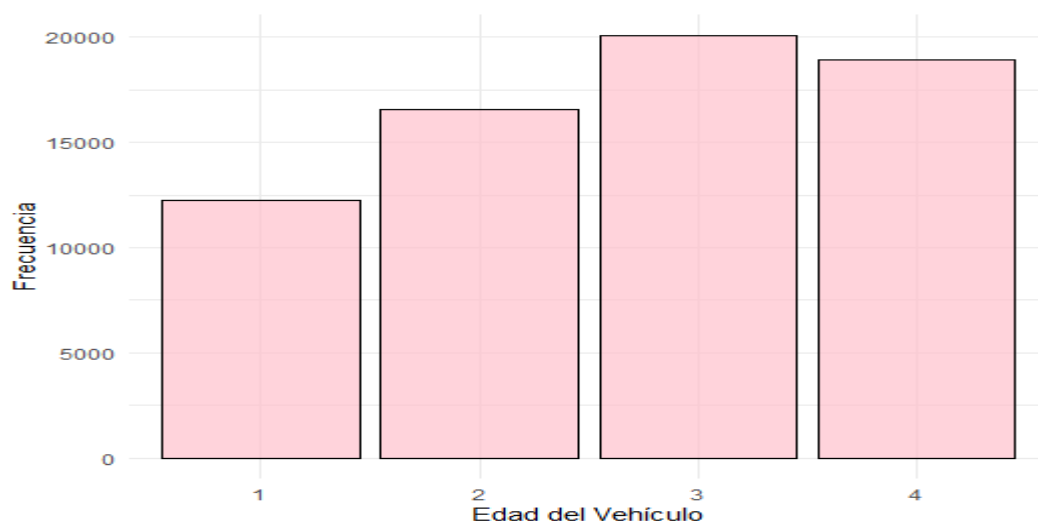


Figura 11. Antigüedad del vehículo.

- Gender: Género del propietario del vehículo. Es un factor con 2 niveles, M si es masculino y F si es femenino.

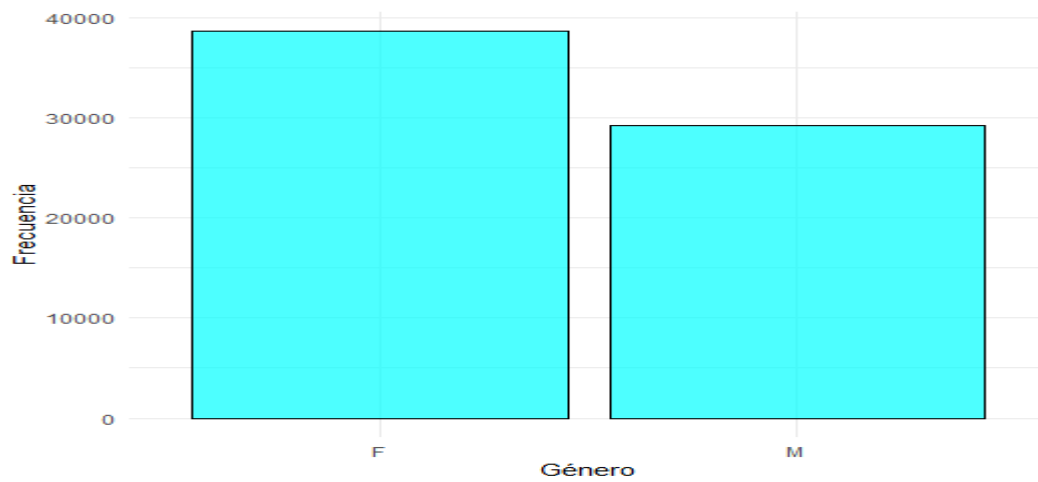


Figura 12. Género del propietario.

- Area: Zona geográfica donde se encuentra el vehículo. Tiene 6 categorías: A, B, C, D, E y F.

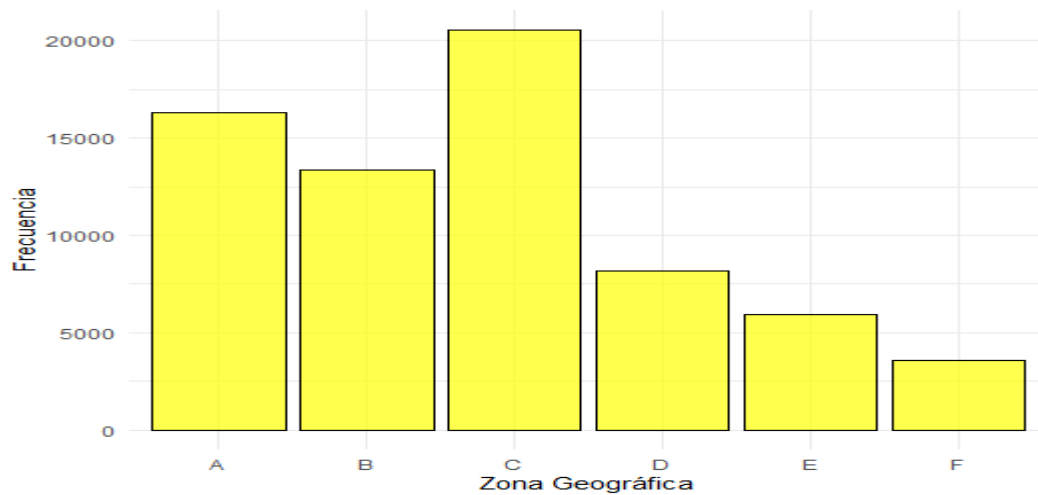


Figura 13. Vehículos por zonas.

- Agecat: Edad del propietario del vehículo, que se divide en categorías del 1 al 6 en orden ascendente de edad.

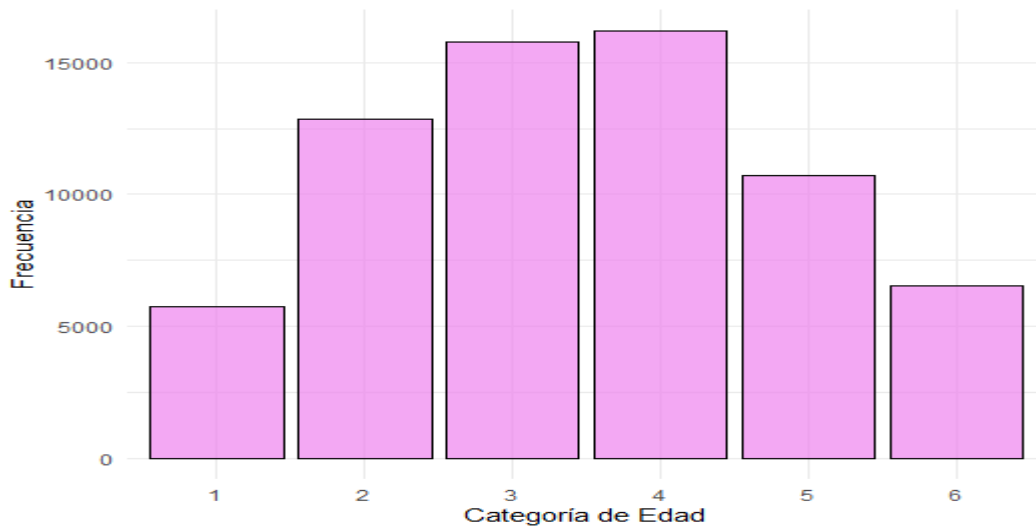


Figura 14. Edad del propietario.

3.2 Identificación del umbral

3.2.1 Método de Tukey

Sirve para identificar valores atípicos. Utiliza los cuartiles y el rango intercuantílico (IQR).

Lo primero es ordenar los datos de mayor a menor y calcular el primer cuartil, la mediana y el segundo cuartil.

Posteriormente calculamos el rango intercuartílico:

$$IQR = Q_3 - Q_1 \quad (10)$$

Finalmente determinamos los límites para los valores atípicos:

$$\begin{cases} \text{Límite inferior} = Q_1 - 1.5 * IQR \\ \text{Límite superior} = Q_3 + 1.5 * IQR \end{cases} \quad (11)$$

Siguiendo estas indicaciones:

Q1 es 353,77, Q3 es 2091,43 e IQR es 1737,65. Y, por tanto, los límites son -2252,71 y 4697,91. Ignoramos el lado inferior por la tipología de nuestro estudio y concluimos que los valores superiores a 4697,91 se consideran atípicos o extremos. Tiene sentido dado que usa los

percentiles 25 y 75, no son casos tan extremos como el 1% de los mayores siniestros que hemos usado anteriormente.

3.2.2 Función empírica de exceso medio (ME-PLOT)

Esta gráfica representa la esperanza de los valores que han superado el umbral y por tanto es una función de distribución condicionada

Se estima con la siguiente expresión dada una muestra ordenada de forma descendente:

$$E_{k,n} = \hat{e}_n\{X_k\} = \frac{\sum_{i=1}^k X_i}{k} - X_{k+1} \quad k = 1, \dots, n-1 \quad (12)$$

No usamos la suma de los excesos, sino los datos de la muestra ($u=X_{k+1}$) y por tanto "la función de exceso medio empírica es la media aritmética de los k mayores valores muestrales (Pérez, 2004)".

Para representar la gráfica se toma la función empírica $E_{k,n}$ como variable dependiente y $u=X_{k+1}$ como variable independiente.

Finalmente, nos fijamos en el valor a partir del cual la línea asciende.

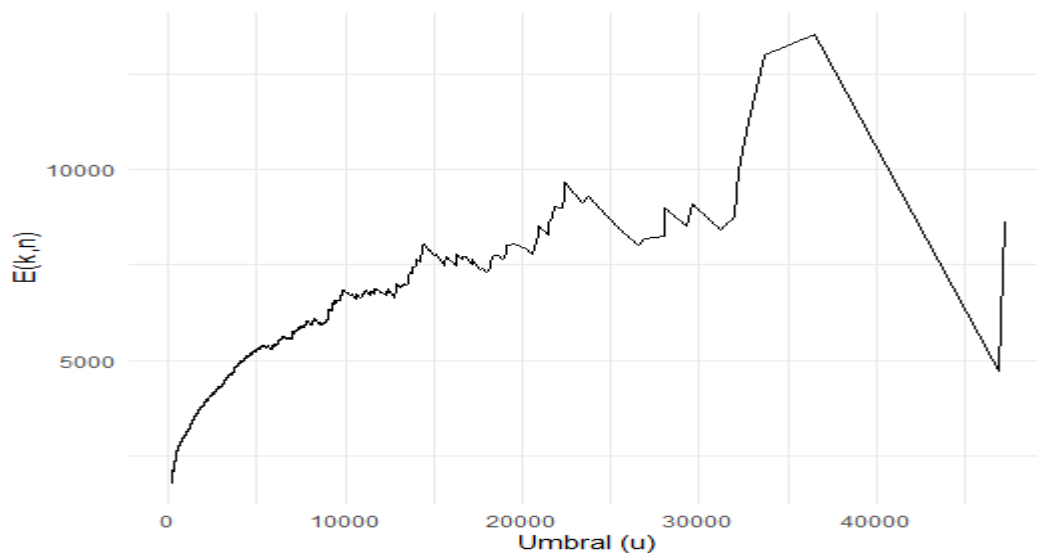


Figura 15. Función empírica de exceso medio.

Como podemos observar, el cambio más brusco de tendencia se produce cuando el umbral es 32000 aproximadamente. Podría ser otra alternativa de umbral, todavía más extrema.

3.2.3 Gráfico de Hill

El método de Hill se utiliza para el umbral que requiere la distribución de Pareto generalizada, únicamente cuando $\xi > 0$. Este método consiste en estimar el parámetro ξ ordenando las observaciones de la forma $Y_1 \geq \dots \geq Y_{(n)}$ y calcular el estimador de Hill (Gomez, 2022):

$$\xi_{i,n}^{Hill} = \frac{1}{i} \sum_{j=1}^i \ln \frac{Y_{(j)}}{Y_{(i)}} \quad 2 \leq i \leq n \quad (13)$$

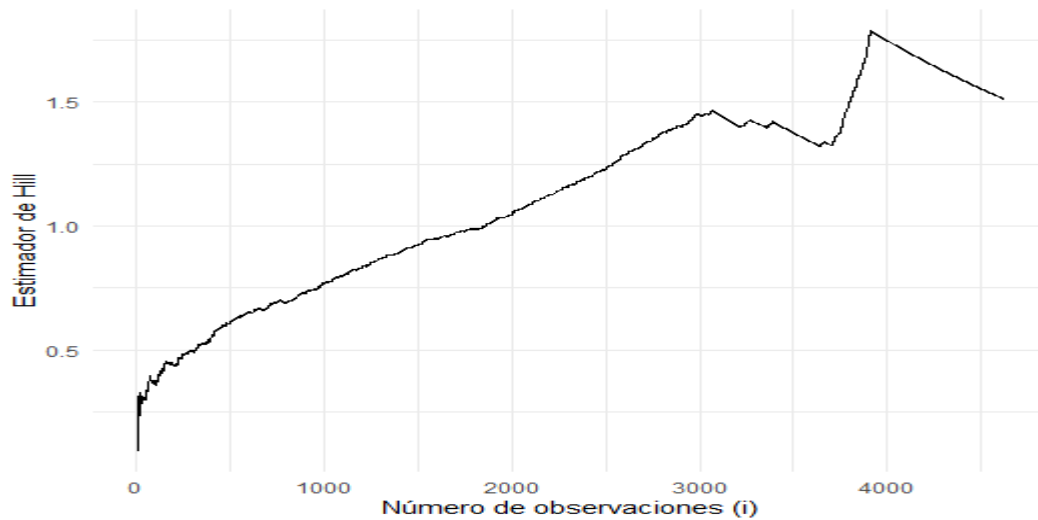


Figura 16. Gráfico de Hill.

Podemos comprobar como en el gráfico de Hill tiene una tendencia similar al anterior método. Lo distintivo de este método frente al anterior es que podemos ir viendo cuál va a ser el estimador de Hill para la forma de la cola y que podemos ver la cantidad de siniestros que se consideran extremos, en este caso hasta el salto brusco, unos

3700 siniestros considerados extremos, concretamente los 3700 de mayor cuantía.

3.2.4 Estimación visual / subjetiva

Este método se basa en fijar un percentil como umbral que suele ser 5% o 1% o simplemente graficar los datos y a simple vista establecer el umbral. En este caso se fijó el percentil 99 y el resultado del umbral fue de 17937,13. A simple se puede apreciar que una elección aparentemente buena sería optar por fijarlo en unos 20000, así que tiene sentido.

3.3 Estudio de la cola de la distribución

Empleamos un umbral del percentil 99 para el estudio.

3.3.1 Parámetros de la cola

Vamos a calcular por los dos métodos previamente estudiados los tres parámetros que forman la distribución.

3.3.1.1 Máxima verosimilitud

Mediante el paquete “extremes” en R podemos aplicar máxima verosimilitud con la función “fevd”. Los resultados son: localización (μ) es 5324,91; la escala (σ) es 1983,39 y la forma (ξ) es 0,7128.

3.3.1.2 Estimador de Hill

Respecto al estimador de Hill, aplicando la fórmula de la teoría, nos aporta un valor de 0,6657 al parámetro de forma, lo cual está bastante cerca de la otra estimación, puesto que difiere en menos de cinco centésimas.

3.3.2 Distribución generalizada de valores extremos

Dado que el parámetro de forma es positivo, estamos frente a una distribución Fréchet, razón por la cual se ha podido aplicar el estimador de Hill. Esto indica que la cola es pesada.

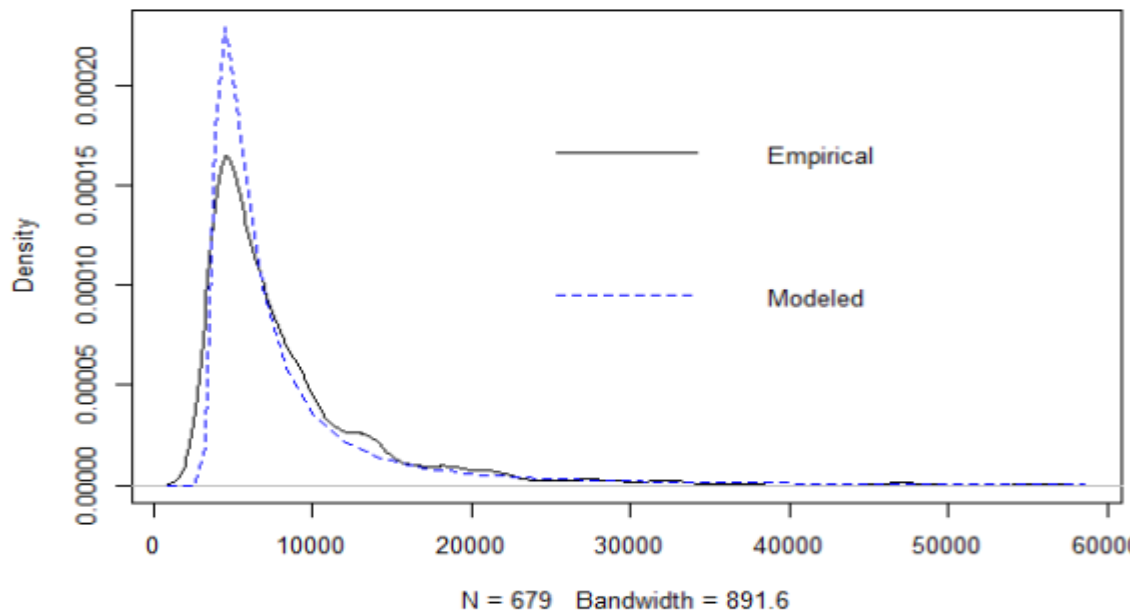


Figura 17. Función de densidad de la distribución frente a la densidad teórica.

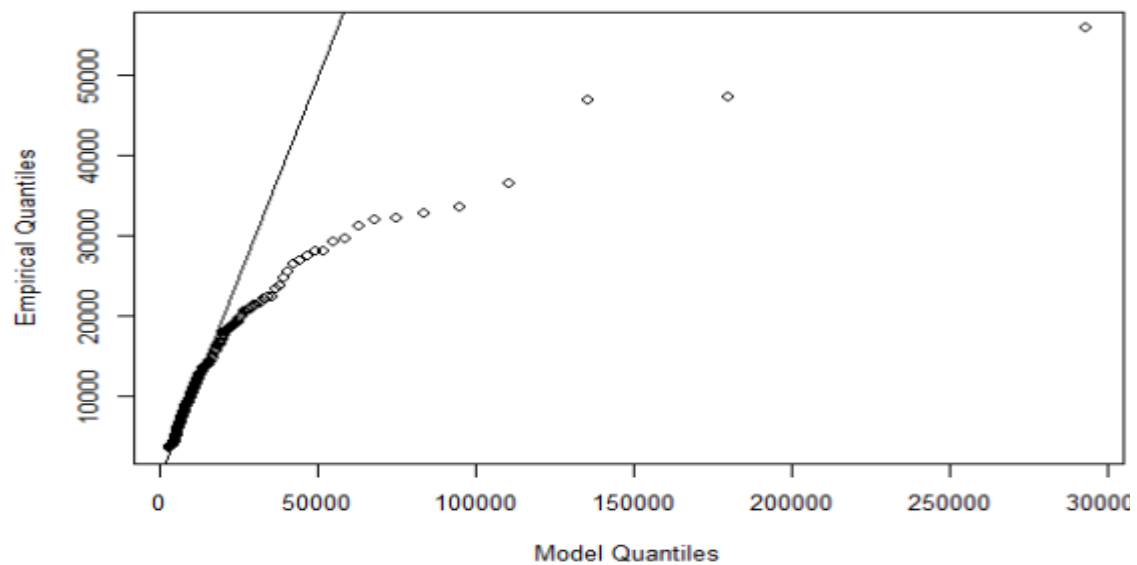


Figura 18. Gráfico cuantil-cuantil.

Se puede apreciar como a partir del 20000 como habíamos dicho la variable deja de seguir la distribución inicial, indicando que la cola se distribuye de otra manera.

3.3.3 Distribución de Pareto generalizada

Lo primero es presentar gráficamente los excesos, para posteriormente modelizarlos:

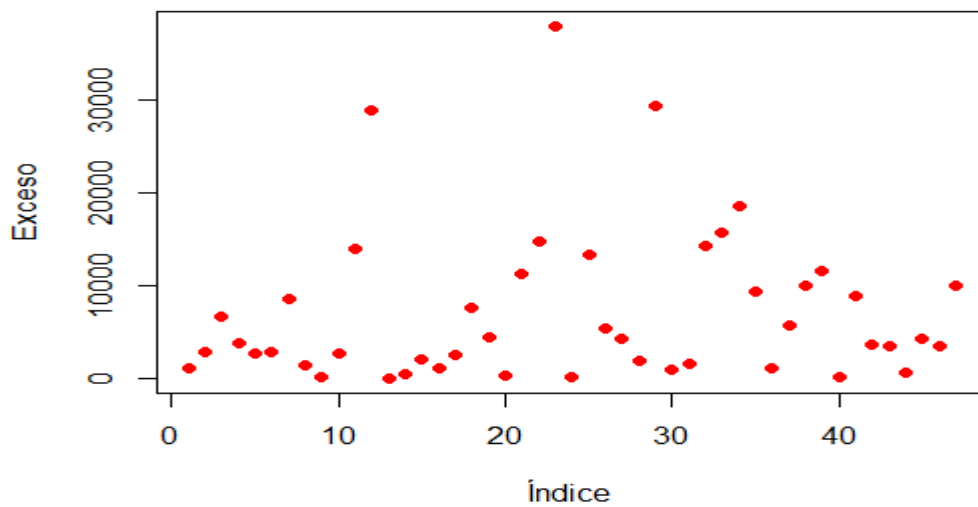


Figura 19. Excesos del umbral

Usando R, realizamos el ajuste de los excesos a la distribución de Pareto generalizada:

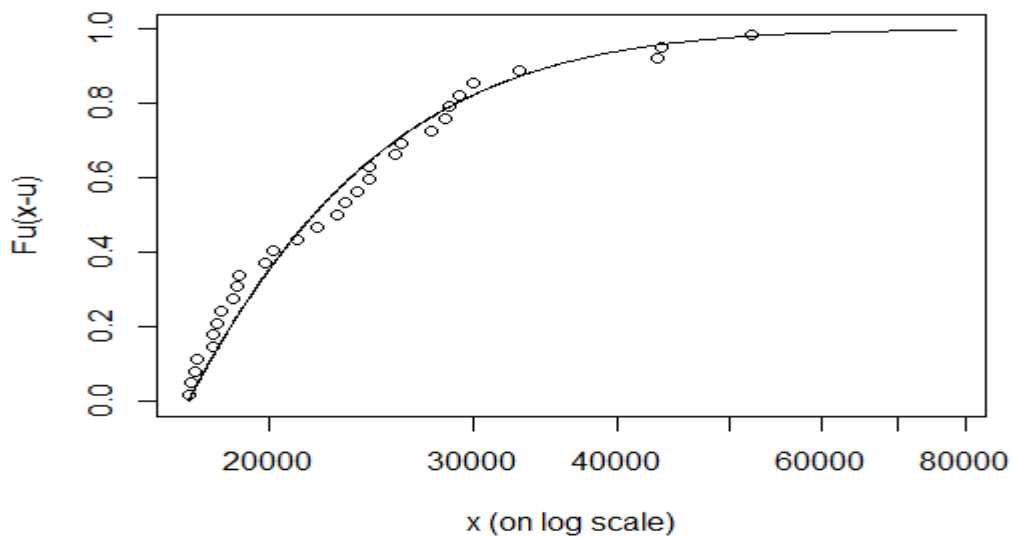


Figura 20. Distribución de los excesos.

Como podemos apreciar, la distribución empírica casa muy bien con la distribución teórica, por lo tanto, podemos decir que los excesos se distribuyen como una distribución de Pareto generalizada.

3.4 Prueba de Kolmogorov-Smirnov

Esta prueba se realiza para decidir si la muestra proviene de una distribución específica. Se define por el siguiente contraste de hipótesis:

H_0 : Los datos siguen la distribución.

H_1 : Los datos no siguen la distribución.

Para rechazar o aceptar la hipótesis nula, el estadístico de la prueba empleado es:

$$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right) \quad (14)$$

Donde F es la distribución acumulativa teórica.

La hipótesis nula es rechazada siempre y cuando si D es superior al valor crítico obtenido en la tabla de Kolmogorov-Smirnov o el p-valor sea inferior al nivel requerido (Superintendencia Financiera de Colombia, 2020).

En nuestro estudio se comprobaría que la cola de la distribución es una Fréchet:

H_0 : Los datos siguen una distribución Fréchet.

H_1 : Los datos no siguen una distribución Fréchet.

4. RESULTADOS Y CONCLUSIONES

Hemos empleado el método del umbral para la identificación de los siniestros extremos y hemos ajustado esos valores resultantes a una distribución generalizada de valores extremos. Posteriormente, hemos identificado los parámetros de la distribución y hemos llegado a la conclusión de que, por el valor del parámetro de forma, es una distribución de tipo Fréchet.

Por último, seleccionamos los excesos de los siniestros extremos respecto del umbral y comparamos su distribución con la teórica de una distribución de Pareto generalizada, siendo exitosa la comparación.

Por tanto, se ha comprobado que la cola de la distribución siempre se distribuye como una distribución de valores extremos generalizada, independientemente de cuál de los tres tipos sea, y que los excesos se convergen a una distribución de Pareto generalizada.

5. LIMITACIONES Y RECOMENDACIONES

Las limitaciones encontradas a lo largo del trabajo han sido varias:

La primera es que la base de datos es una base de datos genérica de uso habitual. Esto es debido a que me fue imposible obtener otra base de datos.

La segunda es la elección del umbral, ya que, según el método que se elija, el umbral va a ser diferente y los resultados pueden variar.

La tercera limitación ha sido la extensión del trabajo. No se ha podido analizar el método de máximos de bloques ni realizar el contraste KS.

Tampoco se ha podido incidir demasiado en ciertos puntos y, por ello, continuaré con las recomendaciones:

Si se quisiese retomar, profundizar o replicar el estudio realizado, recomendaría elaborar una base de datos y complementar la parte teórica actual, ya que es realmente compleja y va mucho más allá de lo anteriormente descrito en este trabajo. En cuanto a la parte práctica, realizaría el método de máximo de bloques y lo intentaría ajustar a una distribución de valores extremos generalizada. Para finalizar lo contrastaría con KS.

6. BIBLIOGRAFÍA

Abalasei, R. F. (2017). *VALORES EXTREMOS Teoría y aplicaciones 2016-2017*.

Actuariales, C. (2024). *Matemática de los Seguros No Vida : Modelos , Medición del Riesgo y Solvencia*.

Alves, I. F., & Neves, C. (2016). Extreme Value Theory: An Introductory Overview. *Extreme Events in Finance: A Handbook of Extreme Value Theory and Its Applications*, 53–95. <https://doi.org/10.1002/9781118650318.ch4>

Bahraoui, Z., Bolancé, C., & Pérez-Marín, A. M. (2014). Testing extreme value copulas to estimate the quantile. *Sort*, 38(1), 89–102. <https://doi.org/10.2139/ssrn.2361163>

Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J., De Waal, D., & Ferro, C. (2005). Statistics of extremes: Theory and applications. *Statistics of Extremes: Theory and Applications*, 1–490. <https://doi.org/10.1002/0470012382>

Carvalho, J. V. F., & Oliveira, L. H. A. (2024). We are Living on the Edge: Managing Extreme-Severity Claims Using Extreme Value Theory. *Brazilian Business Review*, 21(3). <https://doi.org/10.15728/bbr.2022.1245.en>

Chen, H., & Volpe, R. P. (2010). Reproduced with permission of the copyright owner . Further reproduction prohibited without. *Journal of Allergy and Clinical Immunology*, 130(2), 556. <http://dx.doi.org/10.1016/j.jaci.2012.05.050>

Christophe Dutang (2023). CRAN Task View: Extreme Value Analysis. Version 2023-11-04. URL <https://CRAN.R-project.org/view=ExtremeValue>.

Christophe Dutang and Arthur Charpentier (2024). *CASdatasets: Insurance datasets*, R package version 1.2-0, DOI [10.57745/P0KHAG](https://doi.org/10.57745/P0KHAG).

Diawara, D., & Kane, L. (2010). *Applying of the extreme value theory for determining extreme claims in the automobile insurance sector: case of a china car insurance*. 1–31.

Embrechts, P. (n.d.). *Extremal Events*.

Eric Gilleland, Richard W. Katz (2016). extRemes 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software*, 72(8), 1-39. doi:10.18637/jss.v072.i08

Escola, M. (2022). *Beatriz Marques de Sousa Sinistros Graves – Quantos e quanto*.

Escuela de negocios de Wisconsin. Datadescriptions.pdf, Base de datos.<https://instruction.bus.wisc.edu/jffrees/jffreesbooks/Regression%20Modeling/BookWebDec2010/data.html>

Frees, E. W. (n.d.). *Instructors' Manual for Regression Modeling with Actuarial and Financial Applications*.

Ghaddab, S., Kacem, M., Peretti, C. De, & Belkacem, L. (2023). Extreme severity modeling using a GLM-GPD combination: application to an excess of loss reinsurance treaty. *Empirical Economics*, 65(3), 1105–1127. <https://doi.org/10.1007/s00181-023-02371-4>

Gilleland, E. (2022). *Package 'in2extRemes'*. 1–4.

Gilleland, E. and Katz, R. W., 2016: in2extremes: Into the R Package extremes - Extreme Value Analysis for Weather and Climate Applications. NCAR Technical Note, NCAR/TN-523+STR, 102 pp., DOI: 10.5065/D65T3HP2.

Gilleland, E., Ribatet, M., & Stephenson, A. G. (2013). A software

- review for extreme value analysis. *Extremes*, 16(1), 103–119.
<https://doi.org/10.1007/s10687-012-0155-0>
- Gilli, M., & Këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(2–3), 207–228. <https://doi.org/10.1007/s10614-006-9025-7>
- Gomez, H. E. (2022). *Máster Interuniversitario en Estadística e Investigación Operativa*.
- Gumbel, E. J. (2012). Statistics of Extremes. In *Statistics of extremes*. (Online-ausg). Dover Publications.
- Karlsson, M., Wang, Y., & Ziebarth, N. R. (2024). Getting the right tail right: Modeling tails of health expenditure. *Journal of Health Economics*, 97(May), 102912.
<https://doi.org/10.1016/j.jhealeco.2024.102912>
- Katz, R. (2014). *Statistics of Weather and Climate Extremes*. iii, 22–25. <http://www.isse.ucar.edu/extremevalues/extreme.html>
- Laudagé, C., Desmettre, S., & Wenzel, J. (2019). Insurance: Mathematics and Economics Severity modeling of extreme insurance claims for tariffication. *Insurance: Mathematics and Economics*, 88, 77–92.
<https://doi.org/10.1016/j.insmatheco.2019.06.002>
- Mario M. Pizarro (2021). Teoría de Valores Extremos: Distribución de Máximos de bloque.
- Mcneil, A. J., & Mathematik, D. (1997). *ESTIMATING THE TAILS OF LOSS SEVERITY DISTRIBUTIONS*. 27(1).
- Mora Valencia, A. (2010). Estimadores del índice de cola y el valor en riesgo.
- Olivar Danetzy, Lara Naryaly (2020). Teoría del Valor Extremo.
- Package, T., & Datasets, T. I. (2024). *Package 'CASdatasets'*.

- Percz-fructuoso, M. J., & Perez, A. G. (2010). *y finanzas Analyzing solvency with extreme value theory : an application to the Spanish motor liability insurance market*. 20, 35–48.
- Pérez, A. G. (2004). La teoria del valor extremo: una aplicación al sector asegurador. *Hisopano - Italiano de Matemática*, 28802, 27–53.
- Salvadori, G., Michele, C. De, Kottegoda, N. T., & Rosso, R. (2007). Univariate Extreme Value theory. *Extremes in Nature*, 1–112. https://doi.org/10.1007/1-4020-4415-1_1
- SELVAKUMAR, V., SATPATHI, D. K., KUMAR, P. T. V. P., & HARAGOPAL, V. V. (2022). Modeling of Motor Insurance Extreme Claims through Appropriate Statistical Distributions. *Studies of Applied Economics*, 40(S1), 1–13. <https://doi.org/10.25115/eea.v40is1.5685>
- Started, G., Courses, S., Assessment, C. I., & Guide, Q. S. (n.d.). *Extreme value analysis in R with extRemes and in2extRemes What is EVA*. 6–11.
- Stephenson, A. (2024). *Functions for Extreme Value Distributions - Package "evd."* <https://cran.r-project.org/web/packages/evd/evd.pdf>
- Superintendencia de Colombia. Investigaci, D. E., Subdirecci, D., & Investigaci, D. E. (2020). *Documento técnico identificación de siniestros extremos*. 1–20.
- The, S., Statistics, A., & March, N. (2018). *EXTREME VALUE MODELLING OF WATER-RELATED INSURANCE CLAIMS* Author (s): Christian Rohrbeck , Emma F . Eastoe , Arnoldo Frigessi and Jonathan A . Tawn Published by: Institute of Mathematical Statistics Stable URL : [https://www.jstor.org/stable/10.2307/265.12\(1\), 246–282](https://www.jstor.org/stable/10.2307/265.12(1), 246–282).

Víctor A. Rico (2021). Teoría de Valores extremos: Distribución Pareto Generalizada en R.

Vilar Zanón, J. L., & Zhou, N. (n.d.). *Application of Extreme Value Theory to hail risk assessment and management: a study in Spanish wine grapes insurance* .

Wang, Y., Haff, I. H., & Huseby, A. (2020). Insurance : Mathematics and Economics Modelling extreme claims via composite models and threshold selection methods. *Insurance: Mathematics and Economics*, 91, 257–268.
<https://doi.org/10.1016/j.insmatheco.2020.02.009>

Wolny--Dominiak, A., & Maintainer, M. T. (2022). *Package "insuranceData" Type Package Title A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance*.
<http://instruction.bus.wisc.edu/jfrees/jfreesbooks/RegressionModeling/BookWebDec2010/>