Costes esperados por daños corporales en seguro de automóviles e influencia en reservas

Grupo 3
Víctor Alonso Lara
David López Avakian
Sergio Obando Henao
Víctor Manuel Pérez
Miquel Trullols Salat

19 de Diciembre de 2025

Índice

1. Objetivo

Este estudio tiene como objetivo aplicar los conocimientos en modelos estadísticos y cuantificación de riesgos al análisis de un caso práctico en el ámbito del seguro de automóviles. En particular, se centra en el estudio de los costes esperados por daños corporales y su influencia en el cálculo de reservas técnicas. Se trata de un tema de gran relevancia, especialmente en el mercado español, donde la compensación por daños corporales representa más del 60 % del coste total en el seguro de responsabilidad civil del automóvil [Santolino, 2011].

Para su desarrollo, el trabajo se estructura en dos fases:

- Revisión de la literatura existente, con especial atención a los modelos utilizados en la estimación de costes y reservas.
- Modelización del número y la cuantía de los siniestros, mediante técnicas estadísticas aplicadas a datos reales.

Antes de abordar el análisis, se presentarán dos conceptos fundamentales que permiten contextualizar el problema y establecer las bases teóricas del estudio:

- En España, el **seguro obligatorio de automóviles** está regulado por el Real Decreto Legislativo 8/2004. La ley establece que todo propietario de un vehículo a motor con estacionamiento habitual en España debe contratar y mantener un seguro que cubra la responsabilidad civil por los daños causados a personas o bienes durante la circulación del vehículo [Estado, 2004].
- Los **daños corporales** son las lesiones físicas o psíquicas sufridas por una persona en un accidente de circulación, incluyendo lesiones temporales, secuelas permanentes y fallecimiento. Su valoración e indemnización se regulan en la Ley 35/2015 mediante el Baremo oficial [Jefatura Estado, 2015].

2. Estado de la cuestión: Revisión de la literatura existente

2.1. Número de siniestros

La frecuencia de siniestros es un elemento fundamental en la estimación de los costes esperados por daños corporales en el seguro del automóvil, ya que una predicción más precisa del número de reclamaciones permite ajustar con mayor exactitud las reservas técnicas y reducir la incertidumbre asociada a los siniestros futuros. En los últimos años, distintos estudios han buscado mejorar la modelización de esta variable mediante enfoques capaces de reflejar la complejidad y las dependencias que caracterizan los datos de siniestralidad.

Los siniestros con daños corporales constituyen un subconjunto específico dentro del conjunto total de siniestros en el seguro de automóviles, dado que no todos los siniestros implican necesariamente lesiones físicas.

En este contexto, Álvarez Jareño y Muñiz Rodríguez analizan la idoneidad de las distribuciones clásicas para modelizar el número de siniestros en carteras de seguros de responsabilidad civil de automóviles. A partir del estudio de 15 carteras, los autores identifican diversas anomalías muestrales recurrentes que cuestionan la validez de la distribución de Poisson como modelo base. Entre estas anomalías destacan el contagio, la sobre-dispersión (varianza superior a la media), el inflado de ceros (frecuencia excesiva de asegurados sin siniestros), el desinflado de unos (subestimación de asegurados con un único siniestro) y la presencia de

colas más pesadas (subestimación de conductores con múltiples siniestros). Estas irregularidades evidencian que el supuesto de independencia entre eventos y la igualdad entre media y varianza del modelo de Poisson no se cumplen en la práctica [Álvarez et al., 2010].

Para abordar estas limitaciones, los autores proponen la reparametrización de distribuciones alternativas que ofrecen un mejor ajuste a los datos observados: la distribución binomial negativa, la distribución Polya-Aeppli, la distribución Poisson Inversa Gaussiana y la distribución Poisson Pascal Generalizada. Estas distribuciones permiten capturar la presencia de colas pesadas, proporcionando así una base más robusta para la modelización de la frecuencia siniestral.

Otros autores como Pechon, Trufin y Denuit (2018) analizan la frecuencia de siniestros en el seguro obligatorio de responsabilidad civil automotriz tomando al hogar como unidad de riesgo. A diferencia de los modelos que tratan cada póliza de manera independiente, los autores incorporan efectos aleatorios correlacionados a través de mezclas Poisson–LogNormal y Poisson–Gamma, con el propósito de capturar la dependencia entre los miembros de un mismo hogar y la heterogeneidad no observada. Los resultados muestran una correlación significativa entre las siniestralidades de los cónyuges, cercana al 40 %, lo que confirma la existencia de una propensión común al riesgo. Este enfoque permite afinar las estimaciones de frecuencia y, en consecuencia, mejorar la valoración de los daños corporales y la suficiencia de las provisiones técnicas [Pechon et al., 2018].

Otro enfoque a tener en cuenta es el de [Tzougas and di Cerchiara, 2023]. En su artículo, han desarrollado una clase de modelos de regresión Poisson bivariados mixtos con dispersión variable, orientados a modelizar de forma conjunta la frecuencia de reclamaciones por daños corporales y la frecuencia de reclamaciones por daños materiales en el seguro de responsabilidad civil de automóviles. Estos modelos incorporan distribuciones de mezcla para capturar la variabilidad no explicada y permiten analizar simultáneamente las dos variables correlacionadas. Además, son capaces de reflejar tanto la sobre dispersión como la correlación positiva entre ambas frecuencias, lo que representa un avance significativo en la modelización multivariada del riesgo.

La literatura revisada evidencia que comprender la frecuencia de siniestros no solo mejora la precisión en la estimación de los daños corporales, sino que también constituye un componente esencial en la gestión del riesgo y en la sostenibilidad del sistema asegurador.

2.2. Cuantía del siniestro

Los siniestros con daños corporales en el seguro de automóviles se caracterizan por una alta variabilidad en sus costes. En España, durante 2005, la mayoría de estos siniestros costaron menos de 1.500 euros, pero un 0,5 % superaron los 300.000 euros, y algunos casos graves, como lesiones tetrapléjicas, pueden superar el millón de euros [Santolino and Ayuso, 2007].

Como primera aproximación, se realiza un análisis descriptivo de los factores que determinan el coste por daños corporales. Por ejemplo, Marter y Weisberg (1991) clasifican los siniestros de tráfico en cuatro categorías según el tipo de lesión sufrida por la víctima —esguince, fractura, contusión y herido grave— y, para cada una, comparan elementos como el coste médico total, el coste sin hospitalización, el proveedor de asistencia, la frecuencia de visitas y el período de curación [Santolino Prieto, 2011]. Santolino también subraya la relevancia del análisis descriptivo como punto de partida en su estudio sobre indemnizaciones por daños corporales en seguros de auto fijadas judicialmente en Cataluña y Aragón durante el período 2001-2003 [Santolino, 2011].

A partir del análisis descriptivo, la literatura propone modelos para estimar el coste de los siniestros. Weisberg y Derrig [Santolino, 2011] plantean el uso del modelo Tobit, adecuado para datos censurados, donde la indemnización no puede ser inferior a cero ni superar los límites legales o de póliza. La variable dependiente es la indemnización (continua y censurada), mientras que las explicativas incluyen factores dicotómicos —contratación de abogado, lesión grave, fractura, indicios de exageración— y cuantitativos —porcentaje de culpa, coste médico total, semanas de incapacidad—. Este enfoque permite estimar el impacto de factores médicos y legales sobre el logaritmo de la indemnización esperada, ajustando por censura. Los resultados muestran que la contratación de un abogado, la clasificación de la lesión como grave y la presencia de fracturas incrementan la indemnización, mientras que los indicios de exageración la reducen.

El modelo logit ordenado es una herramienta estadística adecuada para analizar variables categóricas jerárquicas, especialmente cuando las categorías tienen un orden natural. En el contexto de siniestros, este modelo permite clasificar la severidad de los eventos en distintos niveles como leve, moderado y grave, y evaluar cómo diferentes factores influyen en la probabilidad de que un siniestro pertenezca a una categoría de mayor severidad. El objetivo principal de este modelo es identificar los factores que incrementan la probabilidad de que un siniestro se clasifique en niveles superiores de pérdida, lo que resulta fundamental para la gestión del riesgo. Para abordar limitaciones del modelo logit ordenado clásico y capturar mejor la complejidad de los datos, se pueden considerar varias extensiones: ordenado mixto, ordenado heterocedástico y multinomiales [Santolino, 2011].

Santolino propone un modelo econométrico log-lineal para explicar el logaritmo de la indemnización total, incorporando variables como edad, tipo de lesión, tipo de vehículo y sexo del lesionado. Un hallazgo relevante es que ni el tipo de vehículo ni la edad del conductor resultan significativos al 10 % de nivel de confianza. Además, se observa que las mujeres reciben indemnizaciones mayores que los hombres y que, cuando el perito necesita más de una visita a la víctima, la cuantía indemnizatoria tiende a incrementarse [Santolino, 2011].

La predicción del coste de indemnización es un aspecto crítico para las compañías aseguradoras, ya que determina la capacidad de la entidad para cumplir con sus obligaciones futuras. Las aseguradoras deben disponer de reservas suficientes que garanticen la estabilidad financiera y la solvencia del ramo. Este desafío se intensifica en los siniestros corporales cuya indemnización se reclama por vía judicial, dado que la resolución suele demorarse durante meses o incluso años. En consecuencia, estos expedientes permanecen abiertos en la contabilidad de la compañía, lo que obliga a realizar provisiones adecuadas para cubrir el coste esperado. [Santolino and Ayuso, 2007].

Por último, se recomienda a las aseguradoras prestar especial atención a los siniestros que superan el percentil 90-95 %, ya que representan casos atípicos con costes significativamente elevados. Estos expedientes requieren una evaluación más exhaustiva para verificar la consistencia de los gastos médicos reclamados y detectar posibles exageraciones o prácticas fraudulentas. Un análisis detallado en esta franja no solo contribuye a reducir el riesgo de sobre indemnización, sino que también permite optimizar la asignación de reservas [Weisberg and Derrig,].

3. Análisis metodológico escogido y resultados

El objetivo inicial es comprender la estructura del conjunto de datos y familiarizarnos con las variables disponibles, ya que esto constituye la base para cualquier análisis exploratorio. En el Cuadro 1 se presenta una descripción resumida de las variables correspondientes a una cartera de una aseguradora en Francia, que incluyen características del vehículo, del conductor, de la póliza y los montos asociados a los reclamos, entre otros.

Cuadro 1: Diccionario de Variables

Variable	Tipo[(1)]	Descripción
IDpol	2	Número de póliza.
ClaimNb	2	Número de siniestros.
Exposure	2	Tiempo de vigencia y exposición al riesgo, en años.
Power	1	Potencia del coche (en orden ascendente, de d a o).
CarAge	2	Antigüedad del vehículo, en años.
DriverAge	2	Edad del conductor, en años.
Brand	1	Marca del vehículo.
Gas	1	Tipo de combustible: Diesel o Regular.
Density	2	Número de habitantes por km² en la ciudad del conductor.
Region	1	Región de la póliza en Francia.
ClaimAmount	2	Costo total del reclamo.
InjuryAmount	2	Costo de compensación por lesiones corporales.
PropertyAmount	2	Costo por daños materiales.

[(1)] 1 = variable categórica, 2 = variable numérica.

3.1. Análisis descriptivo univariado y bivariado de la base de datos

3.1.1. Análisis descriptivo univariado

Se realiza un análisis descriptivo de las variables numéricas considerando medidas de tendencia central, dispersión y forma (Cuadro 9). Adicionalmente, para las variables Exposure, CarAge, DriverAge y Density se aplica una técnica de segmentación mediante el método de k-means (Figura 1). Concluido este análisis, se procede al estudio de las variables categóricas, examinando su distribución de frecuencias y representaciones gráficas para identificar patrones relevantes. Los aspectos más significativos se detallan a continuación:

- **Número de siniestros (ClaimNb):** Presenta valores muy bajos, con media de 0,04 y mediana cero, debido a que el 96,1 % de las pólizas no registraron siniestros. La distribución es altamente sesgada a la derecha (asimetría 5,78) y leptocúrtica (curtosis 38,79), concentrada en cero pero con algunos valores extremos que generan gran variabilidad relativa, reflejada en una desviación estándar de 0,22 frente a una media muy pequeña.

- **Tiempo de vigencia y exposición al riesgo (Exposure):** Caracterizada por una duración promedio de medio año (media 0,56; mediana 0,54) con baja variabilidad (desviación estándar 0,37), es decir, pólizas con duraciones similares. La distribución es prácticamente simétrica (asimetría -0,05) y platicúrtica (curtosis -1,57), más plana que la normal, con menor concentración en torno a la media y colas ligeras.

- **Antigüedad del vehículo (CarAge):** Tiene una media de 7,5 años y una alta dispersión (desviación estándar 5,76), lo que refleja grandes diferencias entre autos nuevos y antiguos, con casos extremos de hasta 100 años. La distribución está sesgada a la derecha (asimetría 1,21), predominando vehículos relativamente nuevos, y es leptocúrtica

(curtosis 8,3), concentrada cerca de la media pero con mayor probabilidad de valores extremos.

- **Edad del conductor (DriverAge):** Una media de 45,3 años y una mediana cercana (44), lo que indica una distribución equilibrada. Presenta un rango amplio (18 a 99 años) y una dispersión moderada (desviación estándar 14,33), reflejando diversidad sin valores extremos desproporcionados. La distribución muestra un ligero sesgo hacia edades mayores (asimetría 0,46) y una forma cercana a la normal (curtosis -0,3), sin colas pronunciadas.

- **Densidad poblacional por km2 en la ciudad del conductor (Density):** Presenta una media muy superior a la mediana (1987.33 vs 287), lo que indica una distribución fuertemente sesgada hacia valores altos. La gran variabilidad (desviación estándar 4.779,6) y la asimetría positiva (4,13) reflejan diferencias significativas entre zonas de baja y alta densidad. La curtosis (17,72) confirma una distribución leptocúrtica, con concentración en valores bajos y presencia de colas largas hacia la derecha.

- **Costo total del reclamo (ClaimAmount):** muestra una distribución extremadamente desbalanceada y con valores atípicos. El promedio es de 832,57 euros, pero la mediana es cero, lo que indica que más del 50 % de las pólizas no tienen reclamos. El mínimo también es cero, mientras que el máximo alcanza los 20.368,3 miles de euros, evidenciando la presencia de valores muy altos. La desviación estándar, de 41.847 euros, refleja una gran dispersión, y tanto la asimetría (375,9) como la curtosis (166.362,6) confirman que la distribución está fuertemente sesgada hacia la derecha, con alta concentración en valores bajos y numerosos outliers. Dado que el 96,1 % de las pólizas no registraron reclamos, se elaboraron histogramas (Figura 2) considerando únicamente aquellas con siniestro, para evitar que la concentración en cero oculte el comportamiento del resto:

  - El primer histograma, que corresponde al 99 % de las pólizas con siniestro, muestra que la mayoría de los reclamos son de bajo coste: el 96,9 % no supera los 50 mil euros, solo el 2,4 % está entre 50 mil y 100 mil euros y menos del 1 % excede los 100 mil euros. El coste promedio es de 13.560 euros, lo que confirma que los siniestros de alto importe son excepcionales y que existe una fuerte concentración en valores bajos.
  - El histograma del 1 % de pólizas más costosas muestra que, aunque pertenecen a la cola de la distribución, el 98,8 % de los reclamos extremos se concentra entre 161,5 mil y 5.000 miles de euros, con un promedio de 577,7 miles de euros. Los valores más altos, entre 10.000 y 25.000 miles de euros, apenas representan el 0,6 % cada uno, confirmando que los siniestros de importe máximo son casos excepcionales con gran impacto potencial.

- **Costo de compensación por daños corporales (InjuryAmount):** Indica una fuerte concentración en valores nulos y una cola muy prolongada con pocos reclamos de gran magnitud. La media es 615,9 euros, mientras que la mediana y el mínimo son cero, y el máximo alcanza 19.973 miles de euros, evidenciando casos excepcionales. La dispersión es elevada (desviación estándar 40.867 euros) y la distribución está fuertemente sesgada a la derecha (asimetría 375,25) y leptocúrtica (curtosis 165.244,5), con alta concentración en valores bajos y presencia de outliers. Histogramas con pólizas con siniestros confirman concentración en valores bajos y una cola con pocos casos de gran coste (Figura 2).

- **Costo por daños materiales (PropertyAmount):** La distribución muestra una fuerte concentración en valores nulos y casos positivos menos extremos que en *InjuryAmount*. El promedio es 216,7 euros, la mediana y el mínimo son cero, y el máximo llega a 575,5

miles de euros, lo que indica reclamos significativos pero no desproporcionados. La dispersión es alta (desviación estándar: 1.554 euros) y la asimetría (129,54) y curtosis (45.682,36) confirman un sesgo marcado hacia la derecha. Para evitar que los ceros oculten el patrón real, se elaboraron dos histogramas: el primero concentra el 99 % de los casos por debajo de 12.000 euros, destacando los rangos de 8.000-10.000 euros (26,3 %), 4.000-6.0000 euros (23,7 %) y 2.000-4.000 euros (22,3 %). El segundo muestra el 1 % restante, casi todo entre 10.400 y 100 mil euros, con algunos reclamos aislados de mayor coste. Aunque existen siniestros elevados, la mayoría se sitúa en valores moderados. Ambos gráficos se incluyen en la (Figura 2).

Tras el análisis detallado de las variables numéricas, se continúa con el estudio de las variables categóricas, examinando su distribución de frecuencias y representación gráfica (Figura 03) para identificar patrones relevantes:

- **Potencia del coche (Power):** Representa la potencia del vehículo desde la menor (d) hasta la mayor (o), cuenta con 12 categorías distintas. La categoría más frecuente es f, con 95.902 registros, lo que equivale al 23,2 % del total. Las categorías d, e, f y g concentran la mayor parte de los datos, con porcentajes de 16,5 % , 18,6 % , 23,2 % y 22,1 % respectivamente, sumando en conjunto aproximadamente 80,4 % del total, mientras que las demás categorías, representan solo el 13,2 % , lo que indica que la mayoría de los vehículos se sitúan en rangos de potencia media.

- **Marca del vehículo (Brand):** Cuenta con 7 categorías distintas, la más frecuente es “Renault, Nissan or Citroen”, con 218.591 registros, lo que equivale aproximadamente al 52,8 % del total, mostrando una clara predominancia de estas marcas en el conjunto de datos. Le siguen Japanese (except Nissan) or Korean con 79.228 registros (19,1 %), Opel, General Motors or Ford con 37.477 (9,0 %) y Volkswagen, Audi, Skoda or Seat con 32707 (7,9 %), mientras que el resto de marcas representan solo el 11,1 %. Esta distribución indica que más de la mitad de los vehículos pertenecen a marcas del grupo Renault-Nissan-Citroën, evidenciando una concentración significativa en este segmento.

- **Tipo de combustible (Gas):** Presenta únicamente 2 categorías, regular y diesel. La categoría más frecuente es regular, con 207.610 registros, lo que representa aproximadamente el 50,2 % del total, mientras que Diesel concentra el 49,8 % restante. Esta distribución muestra un equilibrio casi perfecto entre ambos tipos de combustible.

- **Región de la póliza (Region):** Cuenta con 10 categorías diferentes, la más frecuente es Centre, con 160.814 registros, lo que representa aproximadamente el 38,8 % del total, seguida por Île-de-France con 16,9 %, Bretagne con 10,2 %, Pays-de-la-Loire con 9,4 %, Aquitaine con 7,6 % y Nord-Pas-de-Calais con 6,6 %, mientras que el resto de regiones suman el 10,5 %. Esta distribución evidencia una fuerte concentración en la región Centre, que por sí sola agrupa más de un tercio de los registros, mientras que las demás regiones presentan proporciones significativamente menores.

3.1.2. Análisis descriptivo bivariado

El análisis de correlación entre las variables numéricas muestra que la mayoría de las relaciones son débiles o prácticamente nulas, lo que indica baja dependencia lineal entre ellas (Pearson). Destacan correlaciones muy altas, como era de esperarse, entre *ClaimAmount* e *InjuryAmount* ($\approx 99\,\%$), así como entre *ClaimAmount* y *PropertyAmount* ($\approx 64\,\%$). También se observa una correlación considerable entre *ClaimNb* y *PropertyAmount* ($\approx\ 66\,\%$), indicando que un mayor número de reclamaciones tiende a relacionarse con mayores daños materiales. El resto de las variables, como *CarAge*, *DriverAge*, *Density* y *Exposure*, presentan correlaciones muy bajas con las variables de coste, lo que sugiere que no influyen significativamente en

el valor de los siniestros. En general, el patrón confirma que las variables monetarias están fuertemente interrelacionadas, mientras que las características del vehículo y del conductor tienen poca relación lineal con los montos reclamados. En el Anexo se incluye la matriz de correlaciones (Cuadro 10), así como un mapa de calor (Figura 4).

Se analiza la relación entre variables categóricas y dos indicadores clave: el número de reclamos (*ClaimNb*) y el coste de daños corporales (*InjuryAmount*).

Los vehículos de menor potencia concentran más siniestros, alcanzando hasta 3 o 4 reclamos por póliza, mientras que los de mayor potencia no superan los dos, lo que indica que la potencia no incrementa la frecuencia de siniestros (Figura 5). En cuanto al coste de lesiones, la mayoría de los valores se sitúan cerca de cero, reflejando montos bajos en la mayoría de los casos. Sin embargo, se observan outliers en categorías como f y n, con cifras cercanas a 20.000 miles de euros. En general, no se aprecia una relación clara entre la potencia y el importe de las lesiones corporales (Figura 6).

En cuanto a la marca del vehículo, se observa que la frecuencia de siniestros se mantiene en niveles bajos para la mayoría de las categorías, con valores que oscilan entre 1 y 2 reclamos por póliza. Sin embargo, existen casos puntuales que alcanzan hasta 4 reclamos, especialmente en marcas japonesas (excepto Nissan) o coreanas, así como en Opel, General Motors o Ford (Figura 7). En cuanto al coste por daños corporales, la mayoría de los valores se sitúan cerca de cero, reflejando importes reducidos en la mayoría de los casos. No obstante, destacan dos casos extremos que corresponden a pólizas de vehículos de marca Renault, Nissan o Citroën, con cifras que alcanzan aproximadamente los 20.000 miles de euros (Figura 8).

Respecto al tipo de combustible, se observa que la frecuencia de siniestros se mantiene en niveles similares tanto para vehículos diésel como para los de gasolina regular, con valores que oscilan entre 1 y 2 reclamos por póliza. No obstante, se presentan casos puntuales que alcanzan hasta 3 o 4 reclamos en ambas categorías, lo que indica que el tipo de combustible no incrementa la frecuencia de siniestros (Figura 9). En relación con el coste de lesiones corporales, la mayoría de los valores se sitúan cerca de cero, reflejando importes bajos en la mayoría de los casos. Sin embargo, se identifican outliers significativos en vehículos de gasolina regular (Figura 10).

La frecuencia de siniestros se mantiene en niveles bajos para la mayoría de las regiones, con valores que oscilan entre 1 y 2 reclamos por póliza, aunque se observan casos puntuales que alcanzan hasta 3 o 4 reclamos en Centre, Ile-de-France y Nord-Pas-de-Calais, mientras que Basse-Normandie, Haute-Normandie y Limousin destacan por presentar un menor número de siniestros (Figura 11). En lo referente al coste de lesiones corporales, la mayoría de los valores se sitúan cerca de cero, reflejando importes reducidos en la mayoría de los casos; sin embargo, se identifican outliers significativos en la región Centre, donde se registran los costes más extremos (Figura 12).

3.2. Modelización seleccionada y objetivos a alcanzar

3.2.1. Número de siniestros

A continuación, desarrollaremos los distintos modelos estudiados a lo largo de la asignatura *Modelización estadística*, indicando en cada caso las razones por las que no resultan aplicables cuando corresponda. Aunque la base de datos incluye información sobre el coste de los siniestros, estas variables no se han incorporado en los modelos de frecuencia porque

el coste depende directamente de la ocurrencia del siniestro. Incluirlas generaría un sesgo y una falsa sensación de precisión, además de romper la separación habitual entre frecuencia y severidad que se utiliza en la tarificación actuarial.

3.2.1.1 Clúster jerárquico divisivo

Para el análisis de la variable *ClaimNb*, caracterizada por una alta concentración de ceros y sobredispersión, se ha optado por un algoritmo de Árboles de Inferencia Condicional. Para garantizar la robustez del modelo y evitar el sobreajuste en una cartera de gran volumen, se definió un objeto de control conservador:

- **Nivel de confianza:** 95,00 % ($\alpha = 0{,}05$).
- **Tamaño de Nodos:** Se requiere un mínimo de 3.000 observaciones para intentar una división (min_split) y 1.500 para un nodo terminal (min_bucket).
- **Profundidad:** Limitada a 3 niveles (max_depth) para priorizar la interpretabilidad de los drivers principales.

El modelo (Figura 13) segmenta la cartera priorizando la variable *Exposure*. A continuación, se resume la jerarquía de decisión y los nodos resultantes:

- **Rama de baja *Exposure*** ($Exposure \leq 0{,}29$)
  Este segmento se subdivide nuevamente según el tiempo de vigencia y exposición al riesgo, en años:
  - **Muy Baja ($\leq 0{,}08$):** La variable discriminante es la antigüedad del vehículo *(CarAge)*.
    - **Nodo 4:** Autos nuevos (≤ 5 años). Riesgo mínimo ($err = 0{,}6\,\%$).
    - **Nodo 5:** Autos antiguos (> 5 años). Riesgo bajo ($err = 1{,}2\,\%$).
  - **Media-Baja ($> 0{,}08$):** La variable discriminante es la marca del vehículo (*Brand*).
    - **Nodo 7:** Marcas Occidentales (Fiat, Renault, VW, etc.). Riesgo medio ($err = 2{,}7\,\%$).
    - **Nodo 8:** Marcas Asiáticas. Riesgo menor ($err = 1{,}6\,\%$).

- **Rama de alta *Exposure*** ($\texttt{Exposure} > 0{,}29$)
  En contratos de mayor duración, la *Brand* es el primer divisor, seguido de factores geográficos o de exposición:
  - **Marcas Occidentales:** Se discriminan por zona geográfica (*Region*).
    - **Nodo 11:** Regiones de alta siniestralidad (e.g., Ile-de-France). Riesgo máximo ($err = 6{,}1\,\%$).
    - **Nodo 12:** Regiones de baja siniestralidad (e.g., Bretagne). Nodo mayoritario ($err = 4{,}9\,\%$).
  - **Marcas Asiáticas:** Se discriminan nuevamente por *Exposure*.
    - **Nodo 14:** Exp. $\leq 0{,}65$. Riesgo moderado ($err = 3{,}0\,\%$).
    - **Nodo 15:** Exp. $> 0{,}65$. Riesgo alto ($err = 5{,}2\,\%$).

De manera adicional, se aplicó la función *sctest* (Structural Change Test) en R, que permite comparar estadísticamente las variables candidatas para determinar cuál es la más adecuada al momento de dividir un nodo. Este procedimiento contrasta la hipótesis nula de independencia entre cada predictor y la variable de respuesta (*ClaimNb*), utilizando los p-valores como

criterio para confirmar que las divisiones realizadas por el árbol de decisión son sólidas y estadísticamente justificadas.

Si bien el valor más frecuente de *ClaimNb* es cero (la mayoría de contratos no registran siniestros), el árbol consigue detectar y segmentar el riesgo oculto, mostrando que la probabilidad de ocurrencia de un siniestro varía entre 0.6 % y 6.1 % según el perfil del asegurado.

3.2.1.2 Clúster no jerárquico - k means

Es un método de agrupamiento que busca dividir un conjunto de datos en *k* grupos (clústers) basados en similitud. Es un algoritmo iterativo que elige *k* centros (centroides) iniciales y asigna cada observación al centro más cercano. Luego recalcula cada centro como la media de los puntos en su grupo y repite este proceso hasta que las asignaciones no cambian. El objetivo es que los grupos sean lo más parecidos posibles entre sí y diferentes de los otros grupos.

Para aplicar el algoritmo k-means, se seleccionaron las variables *Exposure*, *CarAge*, *DriverAge*, *Density* y *Power* (esta última convertida previamente a formato numérico). Todas fueron normalizadas mediante estandarización ($(x - media)/desv.estandar$).

Para determinar el número óptimo de clústers (k), utilizamos el método del codo, que analiza la relación entre el número de clústers y la suma total de cuadrados dentro de los grupos (WSS). El punto donde la curva deja de disminuir de forma pronunciada y comienza a aplanarse se denomina *codo*. En nuestro caso, para las variables analizadas, el número óptimo de clústers es 4 (Figura 14). A continuación, se presentan los principales comentarios que resumen los resultados detallados en el (Cuadro 11):

- **Exposure:** Clúster 4 tiene la mayor vigencia (1 año en promedio, rango de 0,8 a 2 años), representa el 35 %. Clúster 1 la más corta (0,1 años en promedio, rango hasta 0,3 años), aporta el 31 %. Ambos suman 66 % del total.
- **CarAge:** Clúster 1 concentra vehículos nuevos (2,7 años en promedio, rango de 0 hasta 6 años) con 49 %. Clúster 2, antigüedad intermedia (9,3 años en promedio, rango 7–13 años), aporta 35 %. Ambos simbolizan el 84 %
- **DriverAge:** Clúster 2 reúne edades medias (40,9 años en promedio, rango de 35–47) con 32 %. Clúster 3, conductores mayores (54 años en promedio, rango 48–62 años), representa 29 %.
- **Density:** Clúster 1 con zonas poco pobladas (398,9 km2, rango hasta 2.317), constituyendo el 80 % del total.
- **Power:** Clúster 2 concentra potencia media-baja (3,7 en promedio, rango de 3–5) con el 50 %. Clúster 1, la más baja (1,5 en promedio, rango hasta 3), aporta 35 %.

3.2.1.3 Análisis PCA

Se aplicó el método de Análisis de Componentes Principales (PCA) sobre la matriz de correlaciones para reducir la dimensionalidad y explorar patrones entre cinco variables: *Exposure*, *CarAge*, *DriverAge*, *Density* y *Power*. Este procedimiento permite sintetizar la información en componentes no correlacionados que explican la mayor parte de la varianza original.

La matriz de correlaciones (Figura 15) muestra relaciones en general débiles entre las variables, con valores máximos en torno a 0.19 (*Exposure–DriverAge*). Destacan correlaciones positivas leves entre *Exposure* y *CarAge* (0.14) y negativas moderadas entre *Exposure* y *Density* (-0.11). Las demás asociaciones son prácticamente nulas.

Los resultados del PCA indican que los dos primeros componentes explican el 48.36 % de la varianza total (26.13 % y 22.23 %, respectivamente), mientras que el tercer componente eleva la proporción acumulada al 68.28 %. Esto sugiere que una representación bidimensional es adecuada para visualización, aunque la inclusión de un tercer componente permitiría capturar mayor estructura.

El biplot (Figura 16) muestra que el Componente 1 diferencia principalmente *Exposure* y *DriverAge* frente a *Density*, reflejando las correlaciones observadas. El Componente 2 captura variabilidad adicional asociada a la edad del conductor, mientras que *Power* presenta baja calidad de representación en este plano, indicando que su aporte se concentra en componentes posteriores. En conjunto, el PCA ofrece una síntesis útil para análisis exploratorio y visualización.

3.2.1.4 Modelo de regresión lineal clásico

No resulta adecuado cuando la variable dependiente es discreta, ya que este modelo parte del supuesto de que dicha variable es continua y sigue una distribución normal, condición que no se cumple en este caso.

3.2.1.5 Modelo Poisson

Adecuado para datos de conteo (números enteros no negativos), que permite identificar qué variables explicativas influyen en la frecuencia de ocurrencia de un evento. Para la construcción del modelo se adoptó un enfoque incremental:

- Se partió de un modelo inicial que únicamente incluía una única variable explicativa, `offset(log(Exposure))`. No todos los vehículos están asegurados el mismo tiempo (*Exposure*). Un coche asegurado 12 meses tiene más oportunidades de registrar siniestros que otro asegurado solo 6 meses. Si no se ajusta esta diferencia, el modelo podría interpretar erróneamente que el coche de mayor exposición es más riesgoso. El término `offset(log(Exposure))` corrige este problema incorporando la exposición directamente en el predictor, de modo que el modelo estime tasas de siniestros por unidad de exposición. Así, si la exposición se duplica, la expectativa de siniestros también se duplica, sin necesidad de estimar un coeficiente adicional.
- Posteriormente, se incorporaron variables adicionales una a una, relacionadas con el conductor, el vehículo y el entorno.
- Tras cada incorporación, se aplicó un test ANOVA y se revisó el Akaike Information Criterion (AIC). El ANOVA se utilizó para evaluar si la nueva variable aportaba una mejora estadísticamente significativa respecto al modelo anterior, justificando así su inclusión. Por su parte, el AIC permitió comparar la calidad global del ajuste, penalizando la complejidad del modelo para evitar sobreajuste.

El modelo Poisson estimado tiene la siguiente especificación:

$$\begin{aligned}\log\left(\mathbb{E}[\text{ClaimNb}]\right) = \log(\text{Exposure}) + \beta_0 + \beta_1\,\text{DriverAge} \\ + \beta_2\,\text{Density} + \beta_3\,\text{CarAge} + \beta_4\,\text{RegionCluster} \\ + \beta_5\,\text{Gas} + \beta_6\,\text{PowerNumCluster} + \beta_7\,\text{BrandCluster}\end{aligned}$$

En un modelo Poisson con enlace logarítmico, el efecto de cada variable sobre la frecuencia esperada de siniestros se interpreta a través de la razón de incidencia (IRR, Incidence Rate Ratio), calculada como $IRR = e^{\beta}$, donde β es el coeficiente estimado para cada variable.Un valor de IRR mayor que 1 indica que la variable incrementa el número esperado de siniestros,

mientras que un valor menor que 1 indica que lo reduce. En el (Cuadro 12) se presentan los estadísticos del modelo, a partir de los cuales se derivan los principales comentarios e interpretaciones sobre el impacto de cada variable en la frecuencia de siniestros:

- Por cada año adicional en la edad del conductor (*DriverAge*), la frecuencia esperada de siniestros disminuye $\approx$ 1,09 %.
- Por cada incremento de 1.000 habitantes/km2 en la ciudad del conductor (*Density*), el número esperado de siniestros aumenta $\approx$ 1,8 %.
- Por cada año adicional en la antigüedad del vehículo (*CarAge*), la frecuencia esperada disminuye $\approx$ 1,40 %.
- Las regiones del clúster *b* (*Basse-Normandie, Bretagne, Centre, Haute-Normandie y Pays-de-la-Loire*) presentan una frecuencia esperada $\approx$ 18,42 % menor respecto al clúster *a* (*Aquitaine, Ile-de-France, Limousin, Nord-Pas-de-Calais y Poitou-Charentes*).
- Los vehículos con combustible *Regular* tiene una una frecuencia esperada $\approx$ 10,74 % menor que los de combustible *Diesel*.
- Los vehículos de marca *Japonesa (excepto Nissan) o Coreana* presentan una frecuencia esperada 20,31 % menor en comparación con el resto de marcas.
- Los vehículos cuya potencia se encuentra en el rango *i–k* presentan una frecuencia esperada 13,87 % mayor respecto a los de potencia *d–e*.

El modelo Poisson ajustado muestra un buen comportamiento para explicar la frecuencia de siniestros. La media (0,043) y la varianza (0,050) son similares, y el cociente entre la desviación y los grados de libertad (0,276) se sitúa muy por debajo de 1. Estos indicadores descartan la presencia de sobredispersión e incluso sugieren una ligera subdispersión. Por ello, no es necesario recurrir a extensiones como el cuasi-Poisson o la binomial negativa, que se emplean habitualmente para corregir sobredispersión.

3.2.1.6 Modelos lineales generalizados (GLM)

Existen diversos modelos que constituyen casos particulares. En este trabajo nos centraremos en aquellos que se aplican a variables dependientes de elección discreta.

3.2.1.6.1 Elección binaria - logit

Describe situaciones en las que la variable dependiente del modelo econométrico es discreta y toma únicamente dos valores posibles. En nuestro caso, utilizaremos variables ficticias binarias que adoptan los valores 0 y 1, donde 1 indica que el individuo ha tenido al menos un siniestro y 0 indica que no ha tenido siniestros. En estos modelos, se supone una relación no lineal entre las variables, dónde el término de error sigue una distribución logística.

La especificación del modelo es la siguiente:

$$\begin{aligned}\operatorname{logit}\big(\Pr(\text{ClaimNB} \mid X)\big) = \beta_0 &+ \beta_1\,\text{Exposure} + \beta_2\,\text{CarAge} + \beta_3\,\text{DriverAge} + \\ &+ \beta_4\,\text{Density} + \beta_5\,\text{GasRegular} + \beta_6\,\text{RegionCluster}_b + \\ &+ \beta_7\,\text{BrandCluster}_b + \beta_8\,\text{PowerNumbCluster}_{f\text{-}h} + \\ &+ \beta_9\,\text{PowerNumbCluster}_{i\text{-}k} + \beta_{10}\,\text{PowerNumbCluster}_{l\text{-}o}\end{aligned}$$

Los principales hallazgos del modelo, cuyos resultados se presentan en el (Cuadro 13), son los siguientes:

- Tiempo de exposición es altamente significativo. A mayor periodo de exposición, mayor es la probabilidad de reclamo. Por cada unidad adicional, los odds (razón entre la probabilidad de que ocurra el evento y la de que no ocurra) se multiplican por 3,46, lo que indica que el evento es mucho más probable.
- Por cada habitante adicional por km², los odds aumentan en torno a 0,001738 %, lo que implica un incremento muy leve en la probabilidad.
- Antigüedad del vehículo: Tiene un efecto inverso. Por cada año adicional, los odds se multiplican por 0,9882, disminuyendo en 1,18 %, lo que significa que el evento es menos probable.
- Edad del conductor: También reduce la probabilidad de reclamo. Por cada año adicional, los odds se multiplican por 0,9921, disminuyendo en 0,79 %.
- Los vehículos del clúster b (marcas japonesas, excepto Nissan, o coreanas) tienen odds de reclamo un 31,7 % menores que los del clúster base a.
- Cambiar a un clúster de potencia superior aumenta los odds entre 11 % y 15 %, mientras que el clúster más cercano al de referencia no es significativo y presenta un odds ratio menor.
- En términos regionales, pertenecer al clúster b reduce los odds en 8,81 % respecto al clúster a.
- Los vehículos con gas regular tienen odds un 9,83 % menores que los diésel.

A continuación, se construyó la matriz de confusión con los datos de la muestra. La primera decisión fue definir el umbral (threshold) adecuado para clasificar las predicciones. Dado que la variable objetivo presenta una distribución altamente desbalanceada (96 % de los casos son 0 y solo 4 % son 1), se utilizó el análisis ROC para seleccionar el umbral óptimo.

El test ROC generó la curva que se muestra en la (Figura 17), donde el punto de corte óptimo se ubica en 0,63, valor que se ha adoptado como umbral (threshold) para la clasificación. En esta matriz, la celda (0,0) corresponde a los verdaderos negativos, la celda (1,0) a los falsos negativos (errores de tipo II), la celda (0,1) a los falsos positivos (errores de tipo I) y la celda (1,1) a los verdaderos positivos. En el (Cuadro 14) se muestran los resultados.

Al observar la matriz, notamos que todas las predicciones corresponden al valor 0. Esto indica que el modelo logit no es adecuado para esta base de datos. La razón principal es el fuerte desbalance en la distribución de la variable objetivo: la mayoría de los registros toman el valor 0. Esta desproporción provoca que el modelo logit no pueda estimar correctamente, ya que tiende a predecir siempre la clase mayoritaria.

3.2.1.6.2 Elección multinomial

No lo utilizamos porque este tipo de modelo está pensado para variables cualitativas con categorías sin orden, como elegir entre colores o marcas. En nuestro caso, el número de siniestros (0, 1, 2, . . .) tiene un orden natural y representa cantidades, no categorías independientes. Si tratáramos cada número como una categoría separada, el modelo ignoraría esa relación.

3.2.1.6.3 Elección multinomial ordenada

Se utiliza cuando la variable dependiente tiene categorías con orden, pero no es estrictamente numérica. Por ejemplo, medir el nivel de satisfacción: bajo, medio y alto. Aquí hay un orden, pero no sabemos si la distancia entre *bajo* y *medio* es igual a la de *medio* y *alto*. En cambio, el número de siniestros es una variable de conteo, donde las diferencias sí son cuantitativas y significativas.

3.2.1.6.4 Elección multinomial anidada

Está pensado para decisiones jerárquicas entre alternativas cualitativas, no para cantidades. Usarlo para número de siniestros sería conceptualmente incorrecto porque no hay un proceso de elección jerárquico.

3.2.2. Costo del siniestro

En las siguientes secciones presentaremos los diferentes modelos abordados durante la asignatura *Cuantificación de riesgos*. Para ello, hemos trabajado únicamente con las pólizas que registran algún coste (pólizas con siniestros). Además, las cifras originales se han dividido por 1.000 con el fin de facilitar el análisis, ya que desde el punto de vista estadístico es preferible trabajar con números más pequeños: esto reduce problemas de escala sin alterar la interpretación de los resultados.

3.2.2.1 Modelización no paramétrica

En primer lugar, se ha representado la distribución empírica de la variable *InjuryAmount* (considerando únicamente los valores estrictamente positivos), tal como se muestra en la (Figura 18). Asimismo, en el Cuadro 2 se muestra el Valor en Riesgo (VaR) de dicha distribución empírica para distintos niveles de confianza:

Cuadro 2: Mod. no paramétrica

Nivel de confianza	VaR
$99,0\,\%$	157,0
$99,5\,\%$	349,8
$99,9\,\%$	1.284,4

Nota: cifras expresadas en miles de euros.

Tal como se muestra en el Cuadro 2, el 99,9 % de los costes (considerando únicamente aquellos estrictamente positivos) no supera los 1.284,4 miles de euros, que corresponde al valor máximo observado en dicho percentil.

A continuación, se obtiene la estimación de la función de densidad mediante el método de núcleo, utilizando tanto el núcleo gaussiano como el núcleo Epanechnikov, y considerando distintas opciones para el cálculo del parámetro de suavizado (bandwidth). Cabe destacar que los valores del parámetro de suavizado no varían entre ambos núcleos. Se han evaluado tres configuraciones:

- Opción por defecto: proporciona un parámetro de suavizado de 814,2.
- Método SJ, que arroja un valor de 3,379.
- Valor fijo, definido subjetivamente, estableciendo el parámetro en 3.

La (Figura 19) muestra las funciones de densidad obtenidas para cada configuración. Se observa que, a medida que aumenta el parámetro de suavizado, la función de densidad se vuelve más plana, reflejando una mayor suavización de la distribución.

Asimismo, se ha obtenido la estimación de la función de distribución mediante el método de núcleos para diferentes valores del parámetro de suavizado (bandwidth) (Figura 20): 86 (valor por defecto), 2, 50 y 100. A continuación, el Cuadro 3 presenta los valores del VaR (miles de euros) correspondientes a cada una de estas estimaciones:

Cuadro 3: Modelización no paramétrica

Nivel de confianza	86	2	50	100
$99,0\,\%$	157.9	158.0	158.0	157.9
$99,5\,\%$	350.4	350.6	350.5	350.4
$99,9\,\%$	1.290,1	1.290,0	1.290,0	1.290,1

Nota: cifras monetarias expresadas en miles de euros.

Al comparar con el Cuadro 2, correspondiente a la distribución empírica, se observa que todos los valores del VaR, independientemente del parámetro de suavizado (h), son superiores. Esto se debe a que el proceso de suavización extiende la función de distribución más allá del valor máximo observado en los datos originales.

Además, el Cuadro 3 muestra que, a medida que el nivel de confianza se aproxima a 1 (por ejemplo, 0,999), un mayor parámetro de suavizado implica que la función de distribución tarda más en alcanzar el valor 1, lo que se traduce en un VaR más elevado. Por ejemplo, para un parámetro de suavizado igual a 100, el 99,9 % de los costes no supera los 1.290,1 miles de euros, mientras que con un parámetro de suavizado igual a 2, el 99,9 % de los costes no supera los 1.290,0 miles de euros.

3.2.2.2 Modelización paramétrica

En la modelización paramétrica se estimaron primero distintas distribuciones clásicas, como Weibull, Lognormal, Log-logística y Gamma, sobre los costes positivos reescalados, evaluando su capacidad para reproducir la asimetría y la cola que define el riesgo extremo. El ajuste por máxima verosimilitud permitió estimar los parámetros de las distribuciones y comparar su calidad relativa mediante el criterio de información de Akaike (AIC). Con los modelos seleccionados se calcularon los valores de referencia del Valor en Riesgo (VaR) en los niveles de confianza del 95 %, 99 % y 99,5 %.

También se estimaron modelos en la escala logarítmica de los costes con el objetivo de mejorar la representación de la cola y capturar de manera más precisa los siniestros severos. Esta transformación permite estabilizar la variable y facilita el ajuste de distribuciones flexibles, entre ellas la Gaussiana (equivalente a la log-normal en la escala original), la GH, la Hiperbólica, la t-Student, la NIG y la VG. Para cada familia se consideraron versiones simétricas y asimétricas, analizando sus parámetros de forma y escala con el fin de evaluar la capacidad de reproducir tanto la asimetría como el comportamiento en la cola. El AIC calculado sobre log(costes) no es directamente comparable con el de las distribuciones clásicas; para corregirlo se suma $2 * \sum log(costes)$. Así, los valores de AIC quedan en la misma escala que los modelos ajustados directamente sobre los costes, permitiendo una comparación homogénea.

En todos los casos, los modelos asimétricos ofrecieron un mejor ajuste, destacando especialmente las distribuciones VG y GH asimétricas como las más parsimoniosas y con estimaciones de riesgo más coherentes en las colas extremas. Los VaR calculados y reportados en el análisis corresponden igualmente a estas versiones asimétricas. En el Cuadro 4 se presentan los resultados obtenidos:

Cuadro 4: Modelización paramétrica

Distribución	$VaR_{99{,}0\,\%}$	$VaR_{99{,}5\,\%}$	$VaR_{99,\,\%}$	**AIC**
Weibull	49	114	151	94.128
Gamma	70	131	159	99.511
Log-Normal	80	365	637	94.767
Log-Logística	94	723	1.712	94.499
GH (asim.)	34	131	239	91.922
Hyp (asim.)	34	128	224	92.027
t-Student (asim.)	43	127	192	93.183
NIG (asim.)	40	128	209	92.839
VG (asim.)	34	132	240	91.919

Nota: cifras monetarias expresadas en miles de euros.

La comparación entre los modelos clásicos y las distribuciones de la familia Generalized Hyperbolic (GH) revela diferencias claras en la representación del comportamiento extremo de los costes positivos. Las funciones tradicionales muestran limitaciones: aunque algunas ofrecen un ajuste relativamente mejor, sus valores de AIC siguen siendo elevados y evidencian dificultades para capturar la concentración de probabilidad en las colas. Esta rigidez se refleja también en el cálculo del Valor en Riesgo (VaR), donde tienden a inflar los cuantiles en percentiles altos, generando estimaciones poco realistas en escenarios severos.

En contraste, las distribuciones GH ajustadas en escala logarítmica logran mejoras sustanciales tanto en los valores de AIC como en la estimación del VaR. Su capacidad para incorporar colas más pesadas y una estructura flexible permite reflejar con mayor fidelidad la variabilidad en los extremos, evitando la inflación de cuantiles y ofreciendo resultados coherentes con la evidencia empírica. Dentro de este grupo, GH y VG destacan por alcanzar los AIC más bajos, mientras que la Hiperbólica se consolida como una alternativa robusta al producir valores de VaR consistentes y estables en escenarios de riesgo severo.

En suma, las distribuciones de la familia GH superan las limitaciones de los modelos clásicos y se perfilan como herramientas idóneas para la gestión del riesgo extremo, combinando ajuste global y estimaciones realistas del VaR de la variable *InjuryAmount*.

3.2.2.3 Teoría de valores extremos

El análisis de la variable *InjuryAmount*, basado en la Teoría de Valores Extremos (EVT), confirma la existencia de una distribución con cola pesada (heavy-tailed distribution). Para verificar el comportamiento de la cola y validar la elección del umbral, se han empleado las siguientes herramientas visuales:

- **Hill Plot:** El estimador de Hill se utiliza en la Teoría de Valores Extremos para calcular el índice de cola (ξ) de distribuciones con colas pesadas, a partir de los valores más extremos de la muestra. Este estimador permite evaluar si la cola sigue un comportamiento tipo Pareto y es clave para seleccionar un umbral adecuado. En la (Figura 21), el eje horizontal representa el orden de los estadísticos y el vertical las estimaciones de ξ. La línea negra muestra cómo varía el estimador según el número de datos extremos considerados, mientras que la línea discontinua horizontal indica una zona de estabilidad alrededor de $\xi \approx 0.9$. La línea vertical azul señala el umbral elegido (u=21).
- **Mean Excess Plot (Gráfico de Exceso Medio):** La (Figura 22) representa el exceso medio en función del umbral. Se observa una tendencia creciente y aproximadamente lineal con pendiente positiva a medida que aumenta el umbral, lo que indica que los

valores extremos se incrementan de forma sistemática. En el marco de la Teoría de Valores Extremos (EVT), esta pendiente positiva es característica de distribuciones con cola pesada, confirmando que el parámetro de forma (ξ) es estrictamente positivo. Esto descarta distribuciones de cola ligera, como la exponencial, y respalda la idoneidad del modelo Pareto para la cola.

- **CV-Plot (Gráfico de Coeficiente de Variación):** En el contexto de la Teoría de Valores Extremos, el valor del CV es un indicador del tipo de cola:
  - Si el CV es mayor que 1, la distribución presenta cola pesada, característica de distribuciones tipo Pareto.
  - Si el CV es igual a 1, la cola sigue un comportamiento exponencial, típico de distribuciones con cola ligera.
  - Si el CV es menor que 1, la cola es ligera, asociada a distribuciones con soporte finito.

  En la (Figura 23), la línea azul se mantiene claramente por encima de 1 en la región relevante, lo que confirma la presencia de una cola pesada y respalda la idoneidad del modelo GPD (Generalized Pareto Distribution) para los excesos.

Basándose en la estabilidad observada en los gráficos de diagnóstico, se estableció un umbral de $u \approx 21$. Este valor permite aislar las 1.616 observaciones más extremas de la muestra, que corresponden aproximadamente al $10\,\%$ superior de la distribución de pérdidas ($P(X > u) \approx 0{,}10$).

Se ajustó la Distribución Pareto Generalizada (GPD) a los excesos sobre el umbral mediante el método de Máxima Verosimilitud (MLE). Los parámetros estimados reflejan la severidad de la cola: parámetro de forma ($\tilde{\zeta}$) igual a $0{,}8831$ (error estándar: $0{,}0456$) y parámetro de escala (ψ) igual a $19{,}2994$ (error estándar: $0{,}9092$).

El valor positivo y significativamente distinto de cero de $\tilde{\zeta} \approx 0{,}88$, considerando su bajo error estándar, confirma la presencia de una cola pesada.

Con estos parámetros, se calculó el capital necesario para cubrir el riesgo al nivel de confianza del $99{,}5\,\%$ ($q = 0{,}995$), comparando los resultados obtenidos mediante el enfoque Pareto simple (Hill) frente al enfoque POT-GPD (Peaks Over Threshold + Generalized Pareto Distribution). El primero asume que toda la cola sigue una distribución Pareto, utilizando el estimador de Hill para aproximar el índice de cola, mientras que el enfoque POT-GPD se centra únicamente en los excesos sobre un umbral y los modela con la Distribución Pareto Generalizada, tal como recomienda la Teoría de Valores Extremos. Esta última aproximación es más robusta, ya que no impone que toda la distribución sea Pareto, sino que describe con mayor precisión el comportamiento extremo.

Cuadro 5: Teoría de valores extremos

Metodología	**Parámetros estimados**	$VaR_{99{,}5\,\%}$
Pareto simple	$\hat{\xi}^{Hill} \approx 1{,}13$	284,3
POT-GPD	$\tilde{\zeta} \approx 0{,}88$, $\psi \approx 19{,}30$	285,8

Nota: cifras monetarias expresadas en miles de euros.

Como se muestra en el Cuadro 5, ambos modelos convergen hacia un riesgo extremo similar, situando el VaR en torno a 285 mil euros. En particular, el enfoque POT-GPD estima

un $VaR_{99{,}5\,\%} \approx 285{,}8$ miles de euros, lo que implica que, con una probabilidad del $0{,}5\,\%$, las pérdidas podrían superar este monto. Este valor, muy superior a la media, evidencia que el riesgo se concentra en eventos de baja frecuencia y alto impacto.

3.2.2.4 Distribuciones compuestas

Una distribución compuesta es un modelo que combina dos distribuciones diferentes para representar adecuadamente datos con dos zonas de comportamiento:

- El cuerpo (valores pequeños o moderados), donde la frecuencia es alta y se ajusta bien con distribuciones como Weibull o Lognormal.
- La cola (valores extremos), donde ocurren pérdidas muy grandes y se necesita una distribución de cola pesada como Pareto.

El punto que separa ambas zonas es el umbral θ, que en este caso se ha determinado mediante el Hill Plot del apartado anterior (observación 1616). A partir de este umbral, se ajustan los parámetros de cada parte y se asignan pesos r y $1 - r$ para reflejar la proporción de datos en cada zona, optimizando por máxima verosimilitud.

El (Cuadro 15) muestra que los modelos simples (Weibull y Lognormal) no capturan bien ambos extremos: Weibull subestima el riesgo extremo ($VaR_{99{,}5\,\%}$: 151 miles de euros) y Lognormal lo sobreestima ($VaR_{99{,}5\,\%}$: 637 miles de euros). Las distribuciones compuestas reducen este sesgo: Weibull–Pareto es el mejor modelo al presentar el mayor loglik y el menor AIC (Akaike Information Criterio), con un $VaR_{99{,}5\,\%}$ más realista (309 miles de euros), lo que lo hace preferible para gestión de riesgo. Cuanto mayor sea el logaritmo de la función de verosimilitud (menos negativo), mejor explica el modelo los datos, mientras que menor sea el AIC, mejor explica el modelo los datos porque este criterio penaliza la complejidad y premia el buen ajuste.

Asimismo, cuando se graficó la distribución Lognormal–Pareto (Figura 24), se observó un salto en el umbral que une ambas distribuciones, por lo que se optó por incluir la restricción de continuidad en la función de densidad, determinando que el valor de r debería ser de 0,83. Sin embargo, al incluir esta restricción, la precisión del modelo se vio perjudicada, ya que el loglik disminuyó y el AIC aumentó, lo que indica que imponer continuidad no siempre mejora el ajuste global, aunque puede ser útil para evitar discontinuidades visuales en la densidad.

3.2.2.5 Distribuciones multivariadas

Se procederá a modelizar conjuntamente los costes por daños materiales y los costes por daños corporales, aplicando previamente una transformación logarítmica a ambas variables. Sobre estos datos transformados se evaluará qué distribución multivariante ofrece el mejor ajuste. Una vez seleccionada la distribución óptima, se calcularán las métricas de riesgo VaR y TVaR para distintos niveles de confianza: 99 %, 99,5 % y 99,9 %.

En primer lugar, se aplicaron pruebas de normalidad a las transformaciones logarítmicas de los costes por separado, utilizando el test de Jarque-Bera. En ambos casos se obtuvieron p-valores inferiores a $2{,}2e-16$, por lo que se rechaza la hipótesis nula con una confianza superior al 99 %, concluyendo que los log-costes no siguen una distribución normal. Si las marginales no son normales, la distribución conjunta tampoco puede ser una normal multivariante. Para confirmarlo, se aplicó el test de Mardia sobre las dos variables transformadas en logaritmos, obteniendo p-valores de 0 tanto para asimetría como para curtosis. Por tanto, se rechaza la hipótesis nula de normalidad multivariante con una confianza superior al 99 %. Ante este resultado, se procedió a ajustar distribuciones más flexibles: la hiperbólica generalizada, la t-Student, la hiperbólica, la normal inversa gaussiana y la varianza-gamma, cada una en sus variantes simétrica y asimétrica.

Cuadro 6: Distribuciones multivariadas

Distribución	AIC escala original (sim.)	AIC escala original (asim.)
t-Student	622.497	613.658
Hiperbólica	623.676	610.371
NIG	622.630	611.958
VG	623.992	606.617
GHip	622.439	606.563

La distribución que ofrece el mejor ajuste a los datos es la hiperbólica generalizada multivariante asimétrica. En este caso, los parámetros estimados para la distribución son los siguientes:

$$\lambda = 0{,}6063, \quad \bar{\alpha} = 0{,}0193$$

$$\mu = \begin{bmatrix} 7,7859 \\ 9,1153 \end{bmatrix}, \quad \sigma = \begin{bmatrix} 4{,}6606 & -0{,}3728 \\ -0{,}3728 & 0{,}1868 \end{bmatrix}, \quad \gamma = \begin{bmatrix} -0{,}1719 \\ -0{,}7390 \end{bmatrix}$$

En la (Figura 25) se presenta un mapa de calor de los datos. Se observa que, aunque la distribución hiperbólica generalizada multivariante asimétrica ofrece un mejor ajuste que la distribución gaussiana, su desempeño en los extremos sigue siendo insuficiente.

Con estos parámetros se procedió a simular el VaR y el TVaR para niveles de confianza del 99 %, 99,5 % y 99,9 %, obteniendo los siguientes resultados:

Cuadro 7: Distribuciones multivariadas

Nivel de confianza	VaR	TVaR
$99,0\,\%$	765	2.620
$99,5\,\%$	2.487	5.238
$99,9\,\%$	37.978	26.152.044

Nota: cifras monetarias expresadas en miles de euros.

3.2.2.6 Cópulas

Las variables *InjuryAmount* y *PropertyAmount* presentan dependencias negativas por rangos. Por ello, no resulta adecuado ajustar una cópula explícita distinta. Además, debido a la distribución de los datos, la única opción razonable es aplicar la cópula gaussiana. A continuación, se presentan los resultados obtenidos utilizando la cópula gaussiana en dos escenarios: con marginales log-normales y con marginales log-logísticas.

Cuadro 8: Cópulas

Marginales	AIC ajustado	$VaR_{99,0\,\%}$	$VaR_{99,5\,\%}$	$VaR_{99,9\,\%}$
Log-normales	62.834	375	643	2.013
Log-logísticas	62.651	738	1.712	12.487

Nota: cifras monetarias expresadas en miles de euros.

Tal como se muestra en el Cuadro 8, la cópula gaussiana con marginales log-logísticas ofrece el mejor ajuste a los datos, ya que presenta el menor valor de AIC. En este escenario, para un nivel de confianza del 99.9 %, la pérdida máxima esperada asciende a 12.487 miles de euros.

4. Informe ejecutivo

Por desarrollar.

5. Anexos

5.1. Análisis descriptivo univariado y bivariado de las bases de datos

5.1.1. Análisis descriptivo univariado

Cuadro 9: Análisis descriptivo univariado - variables numéricas

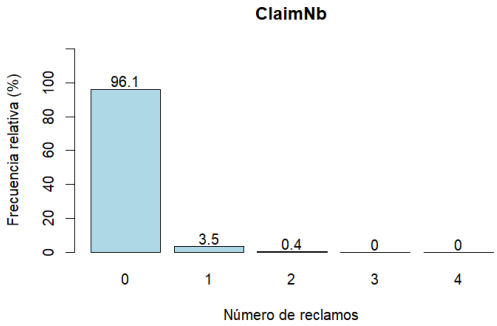
Concepto	Claim Nb	Exposure	CarAge	DriverAge	Density
Media	0,04	0,56	7,53	45,32	1.987,33
Mediana	0	0,54	7	44	287
Mínimo	0	0	0	18	2
Máximo	4	1,99	100	99	27000
Desv. Estándar	0,22	0,37	5,76	14,33	4.779,6
Asimetría	5,78	-0,05	1,21	0,46	4,13
Curtosis	38,79	-1,57	8,3	-0,3	17,72

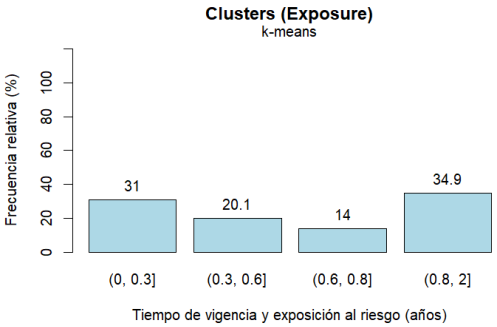
Cuadro 9: Análisis descriptivo univariado - variables numéricas

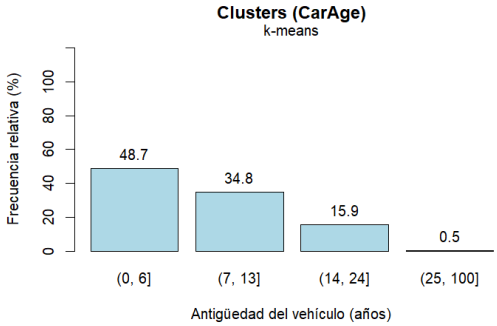
Concepto	Claim Amount	Injury Amount	Property Amount
Media	832,57	615,9	216,67
Mediana	0	0	0
Mínimo	0	0	0
Máximo	20.368.330	19.792.821	575.508,8
Desv. Estándar	41.847	40.867,4	1.554,24
Asimetría	375,9	375,25	129,54
Curtosis	166.362,6	165.244,5	45.682,36

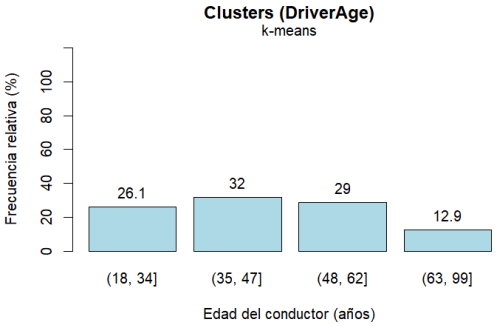
(Volver)

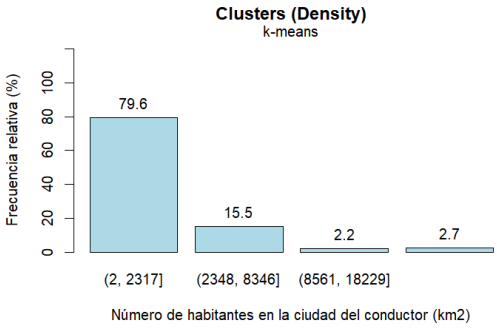
Figura 1: Análisis descriptivo univariado





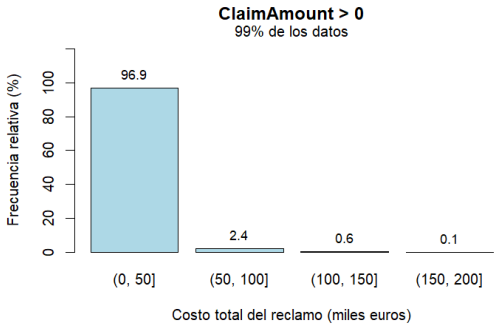


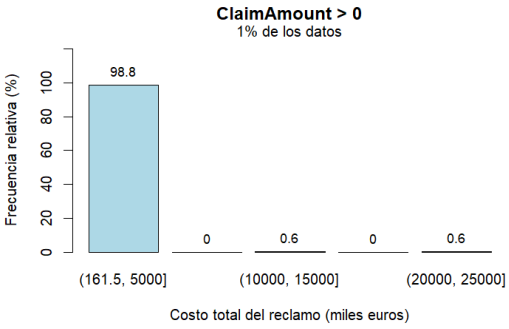


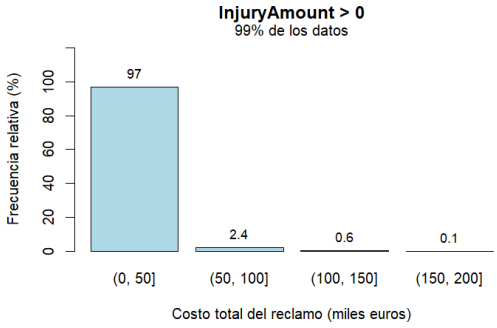


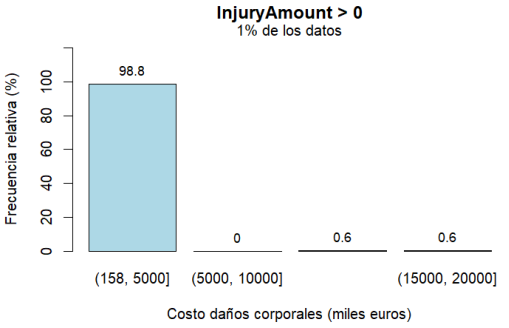
(Volver)

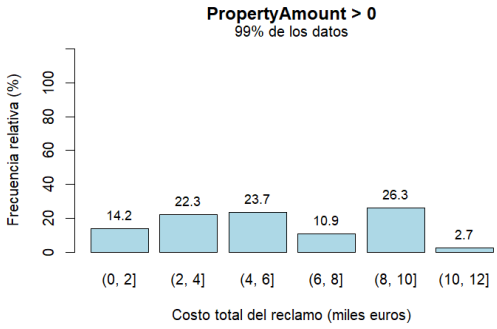
Figura 2: Análisis descriptivo univariado (Variables numéricas) - Histogramas

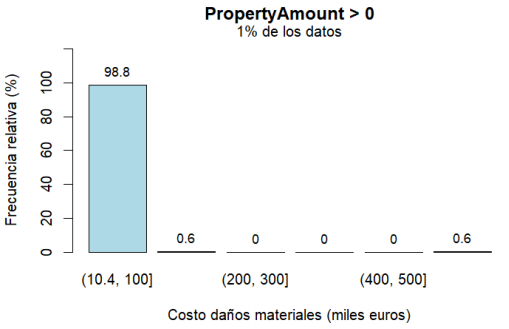






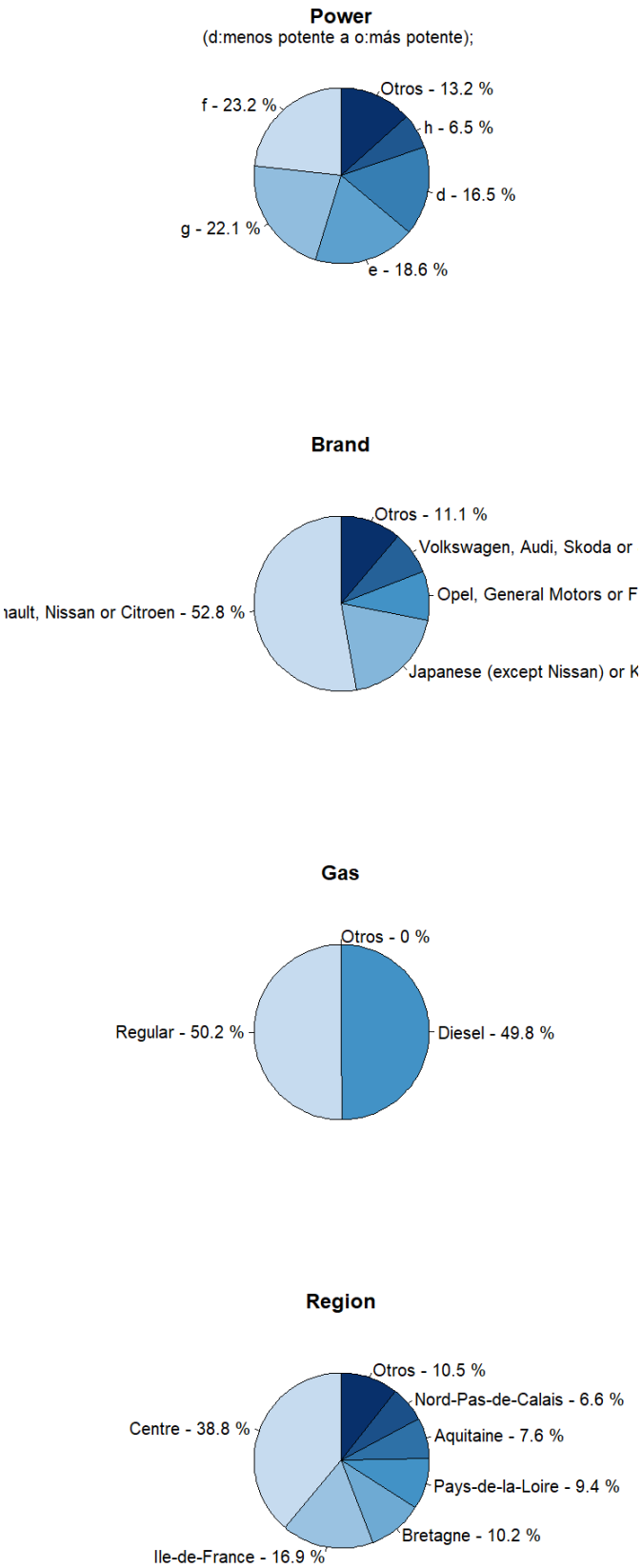






(Volver)

Figura 3: Análisis descriptivo univariado (Variables categóricas) - Gráfico de pastel



(Volver)

5.1.2. Análisis descriptivo bivariado

Cuadro 10: Matriz de correlación de Pearson

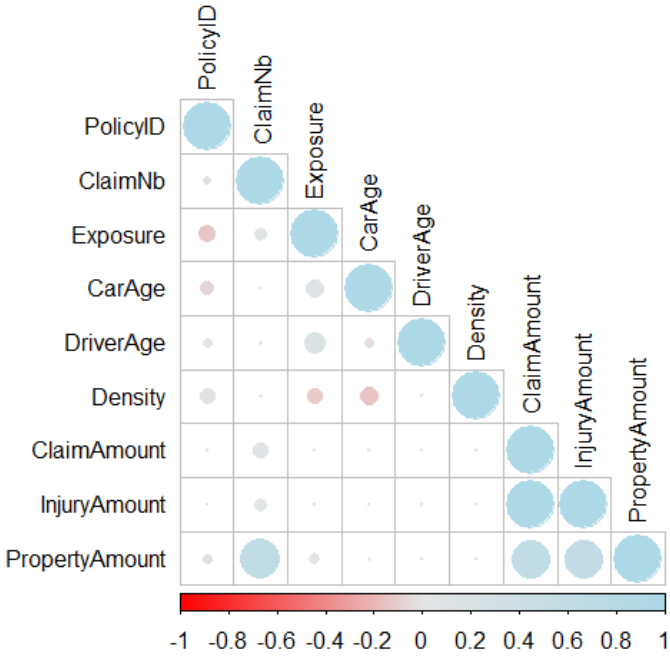
Variable	PolicyID	ClaimNb	Exposure	CarAge	DriverAge	Density
PolicyID	1	-	-	-	-	-
ClaimNb	-0,0327	1	-	-	-	-
Exposure	-0.1324	0,0761	1	-	-	-
CarAge	-0.0789	0,0025	0,1399	1	-	-
DriverAge	0,0487	-0,0075	0,1943	-0,0465	1	-
Density	0,1022	0,0089	-0,1121	-0,1423	-0,0016	1
ClaimAmount	-0,0054	0,0977	0,0022	0,0016	-0,0045	-0,0013
InjuryAmount	-0,0041	0,0749	0,0002	0,0014	-0,0045	-0,0014
PropertyAmount	-0,0389	0,6620	0,0559	0,0062	-0,0038	0,0012

Cuadro 10: Matriz de correlación de Pearson

Concepto	Claim Amount	Injury Amount	Property Amount
PolicyID	-	-	-
ClaimNb	-	-	-
Exposure	-	-	-
CarAge	-	-	-
DriverAge	-	-	-
Density	-	-	-
ClaimAmount	1	-	-
InjuryAmount	0,9996	1	-
PropertyAmount	0,6417	0,6190	1

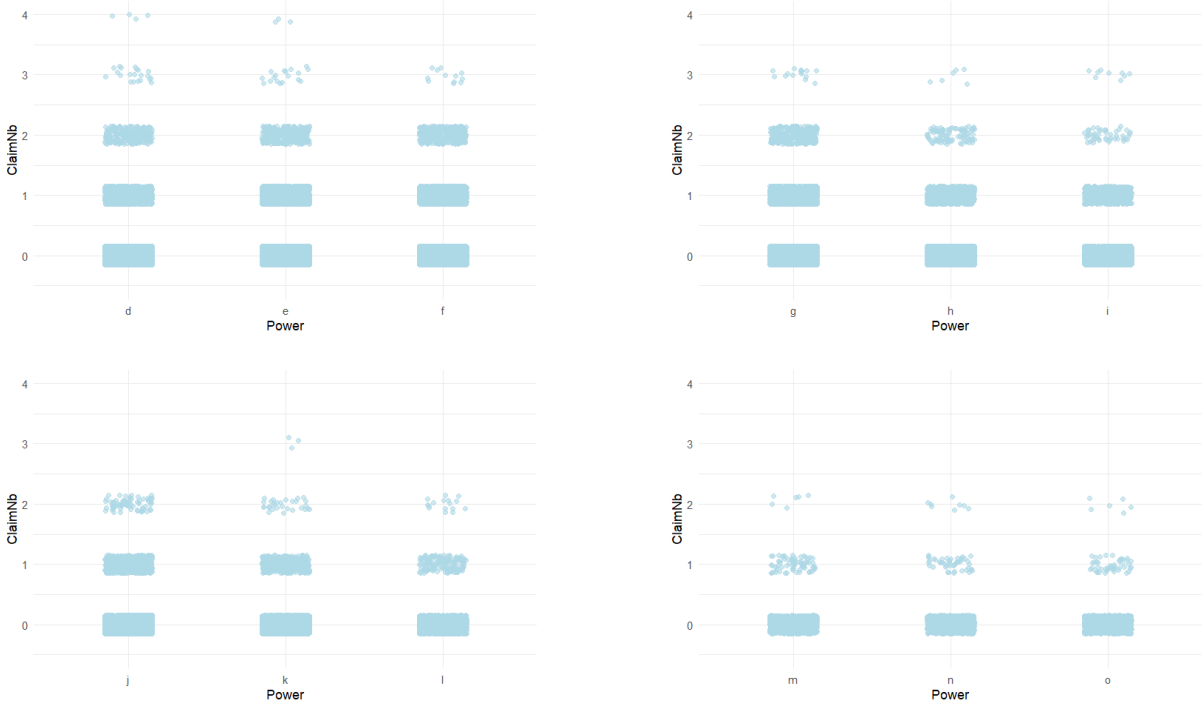
(Volver)

Figura 4: Matriz de correlación de Pearson - Mapa de calor



(Volver)

Figura 5: Jitter plot - Power vs ClaimNb

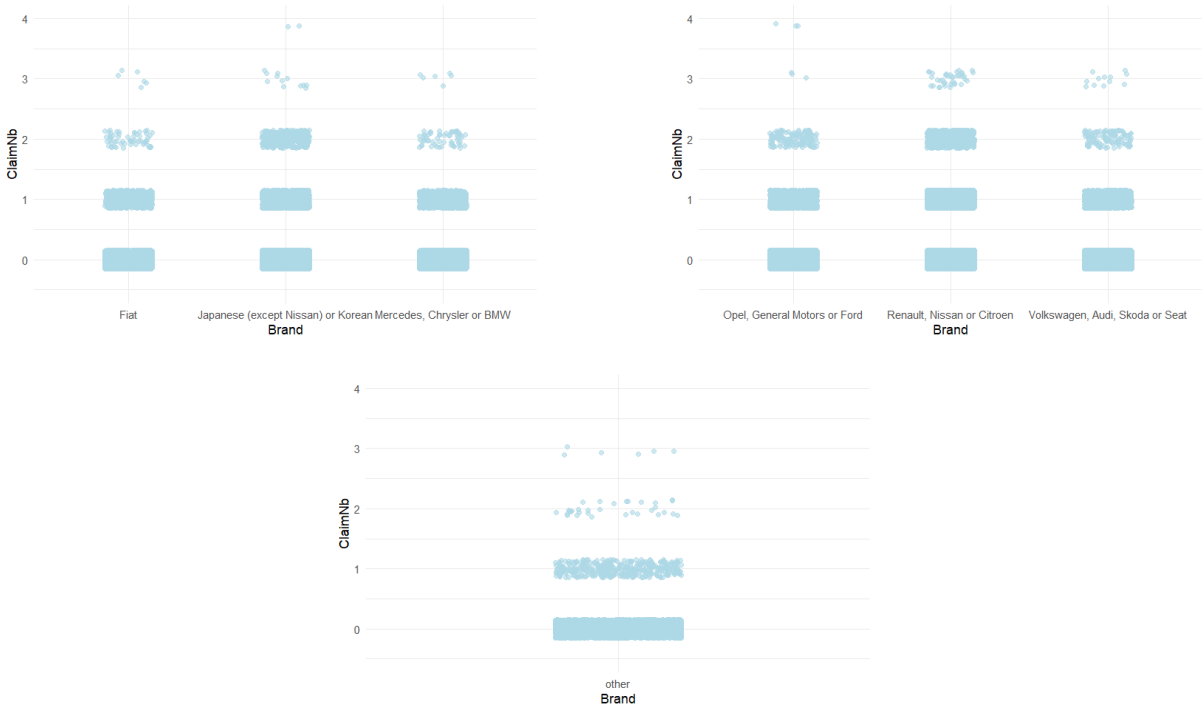


(Volver)

Figura 6: Jitter Plot - Power vs InjuryAmount

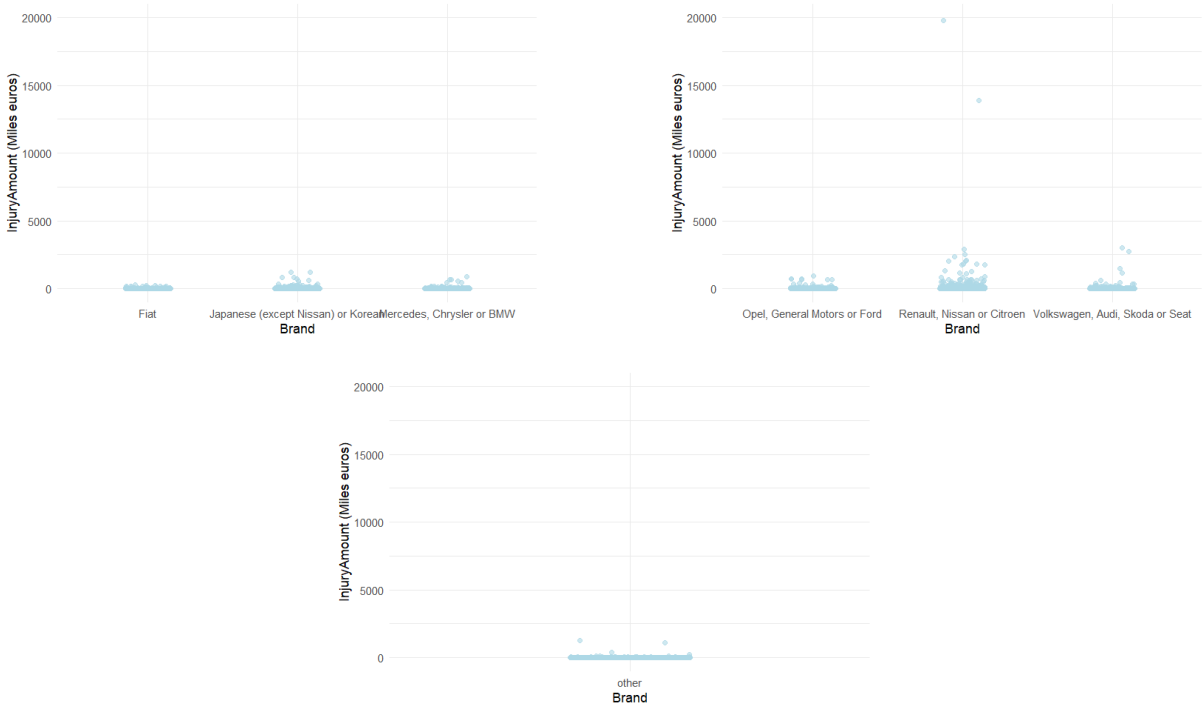
(Volver)

Figura 7: Jitter plot - Brand vs ClaimNb



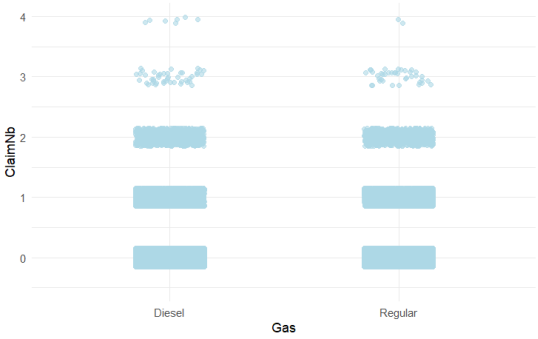
(Volver)

Figura 8: Jitter Plot - Brand vs InjuryAmount



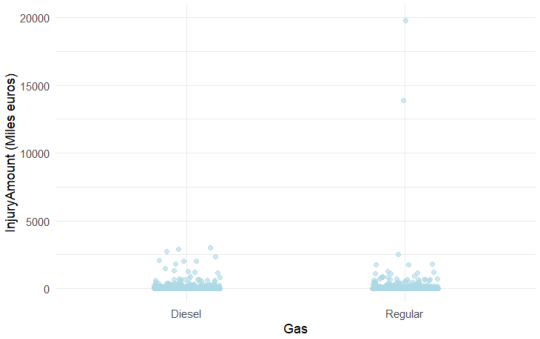
(Volver)

Figura 9: Jitter plot - Gas vs ClaimNb



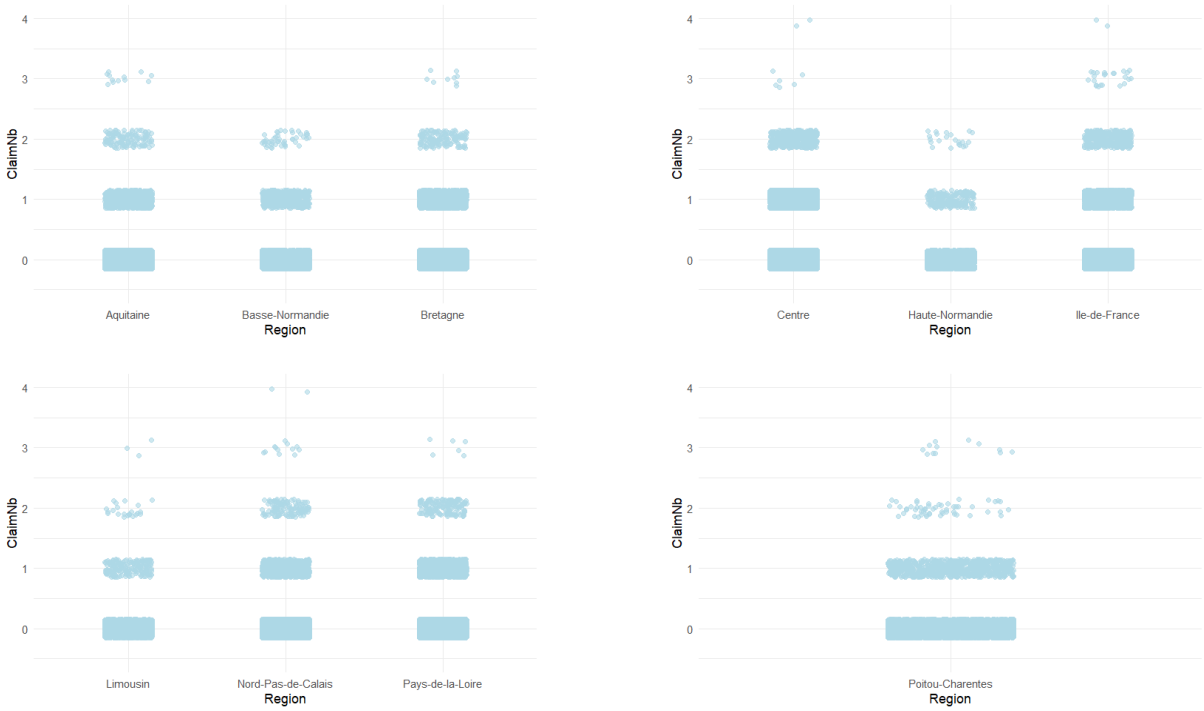
(Volver)

Figura 10: Jitter Plot - Gas vs InjuryAmount



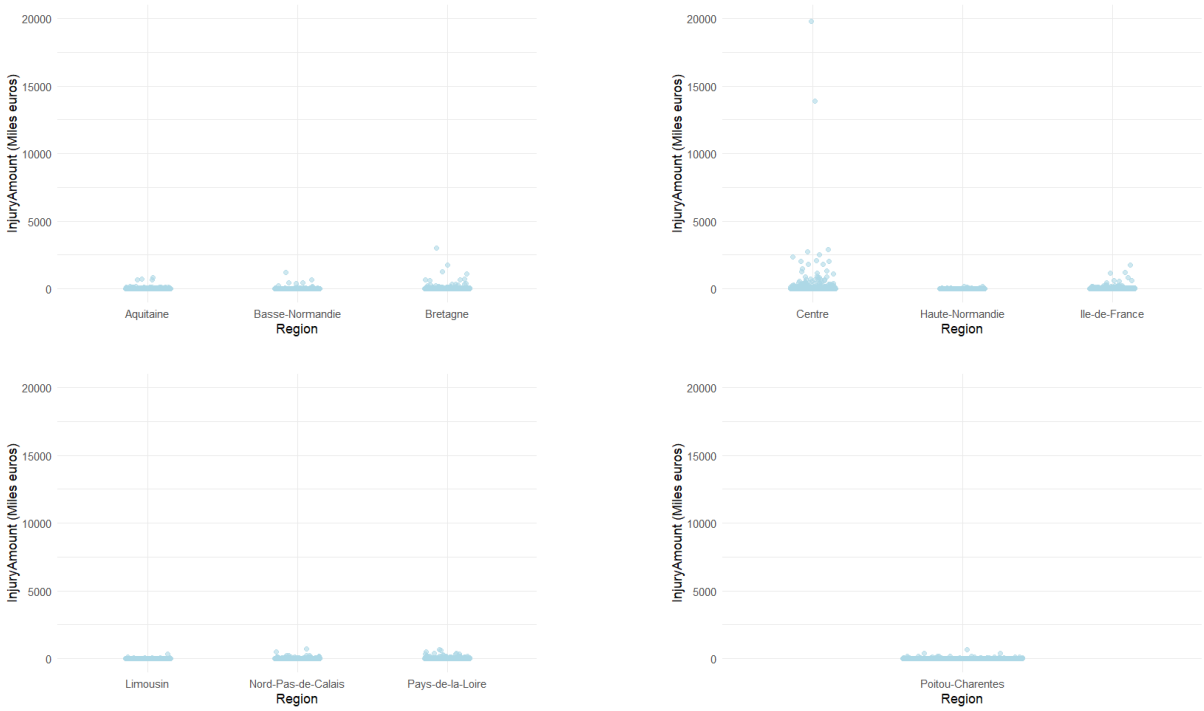
(Volver)

Figura 11: Jitter plot - Region vs ClaimNb



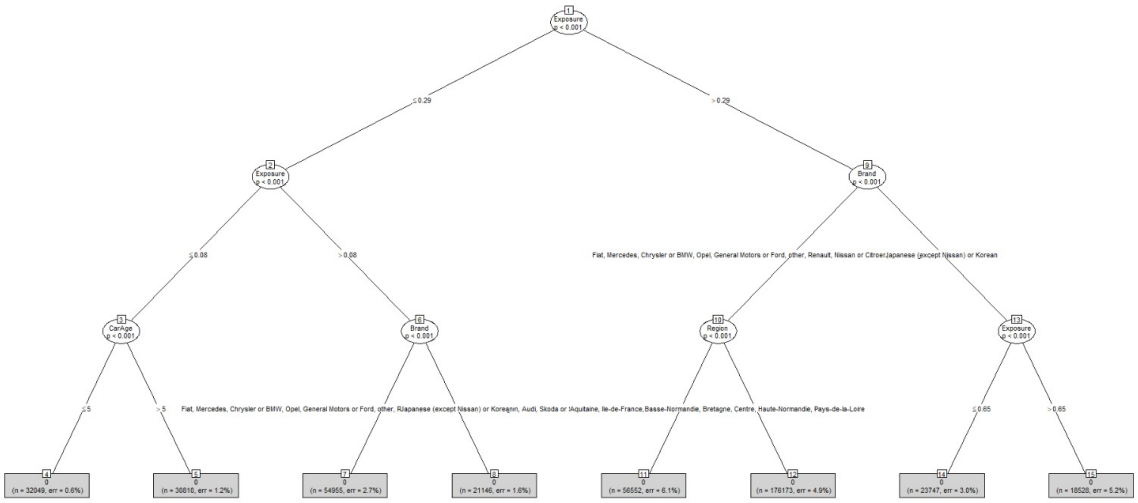
(Volver)

Figura 12: Jitter Plot - Region vs InjuryAmount



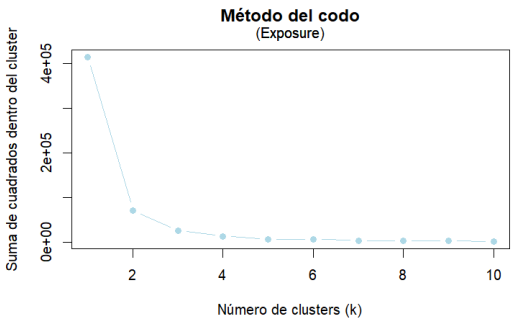
(Volver)

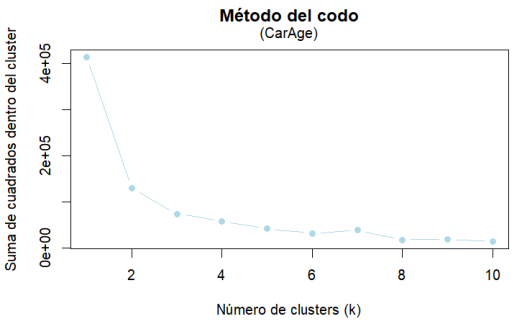
Figura 13: Clúster jerárquico aglomerativo

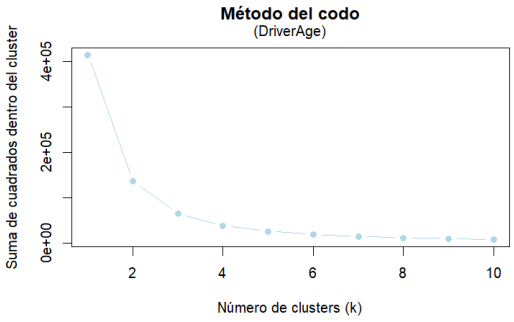


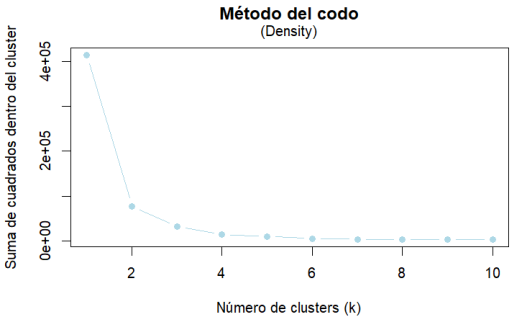
(Volver)

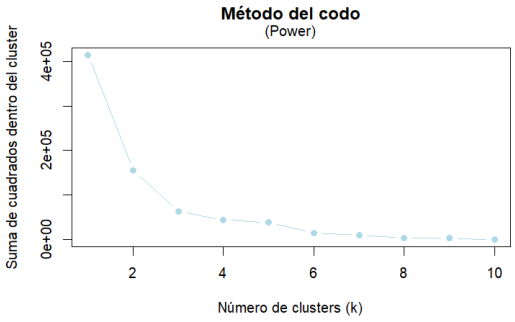
Figura 14: Clúster no jerárquico - k means











(Volver)

5.2. Modelización seleccionada y objetivos a alcanzar

5.2.1. Número de siniestros

5.2.1.1 Clúster jerárquico divisivo

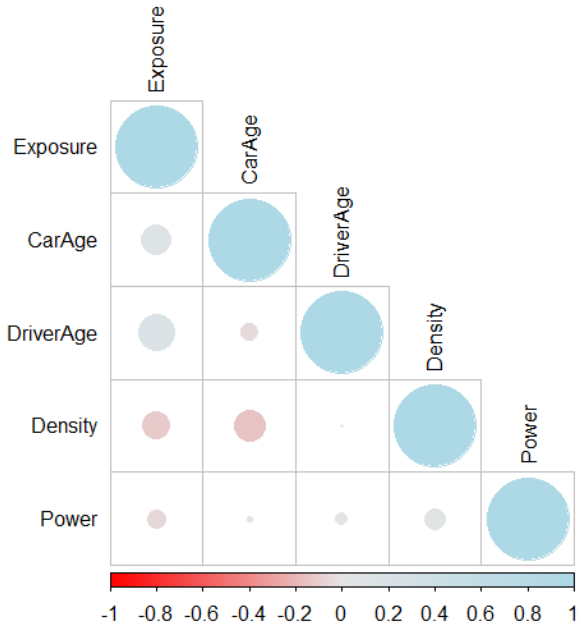
5.2.1.2 Clúster no jerárquico - k means

Cuadro 11: Clúster no jerárquico - k means

Concepto	Clúster	Exposure	CarAge	DriverAge	Density	Power
Centroide estand.	1	-1,2	-0,8	-1,2	-0,3	-0,9
	2	-0,4	0,4	-0,3	0,5	0,1
	3	0,4	1,5	0,6	2,3	1,7
	4	1,2	4,2	1,8	5,2	3,3
Centroide desestand.	1	0,1	2,7	28,5	398,9	1,5
	2	0,4	9,9	40,9	4.257,8	3,7
	3	0,7	16,2	53,9	12.810,2	6,8
	4	1,0	32,0	71,1	26.692,8	10,0
Mínimo	1	0,0	0	18	2	1
	2	0,3	7	35	2.348	3
	3	0,6	14	48	8.561	6
	4	0,8	25	63	20.000	9
Máximo	1	0,3	6	34	2.317	3
	2	0,6	13	47	8.346	5
	3	0,8	24	62	18.229	8
	4	2,0	100	99	27.000	12
Tamaño	1	128.390	201.485	108.167	329.523	145.339
	2	83.237	144.258	132.408	64.073	214.001
	3	58.021	65.981	120.016	9.015	45.274
	4	144.312	2.236	53.369	11.349	9.346

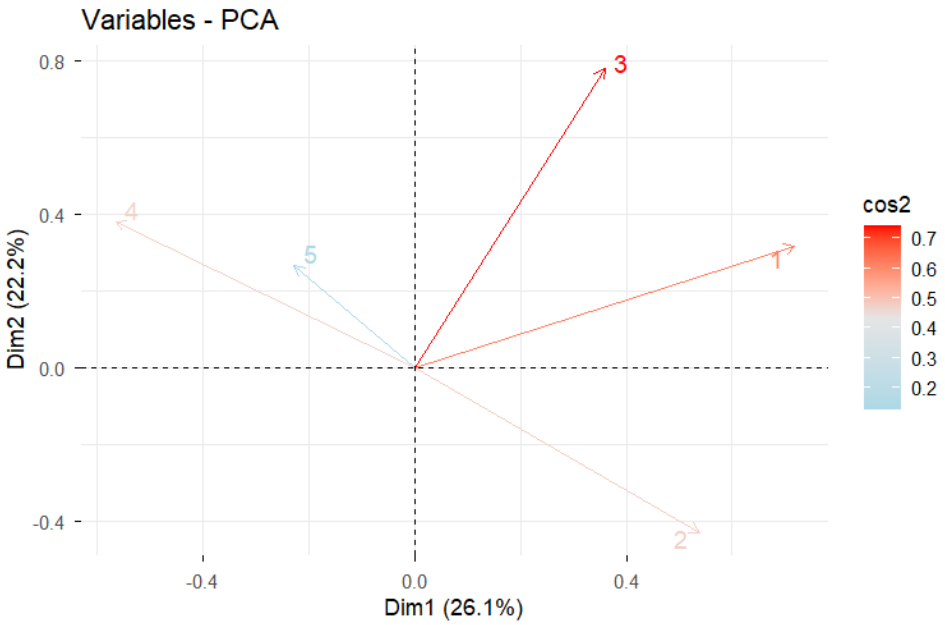
(Volver)

Figura 15: Análisis de Componentes principales (PCA)



(Volver)

Figura 16: Análisis de Componentes principales (PCA)



(Volver)

5.2.1.3 Análisis PCA

5.2.1.4 Modelo Poisson

Cuadro 12: Modelo Poisson

Coeficientes	Estimado	p-value	e^{β}	IRR (%)
Intercepto	-1.782e+00	$< 2e-16$	0.1682	-83.1755
DriverAge	-1.095e-02	$< 2e-16$	0.9891	-1.0890
Density	1.808e-05	$< 2e-16$	1.0000	0.0018
CarAge	-1.414e-02	$< 2e-16$	0.9860	-1.4045
RegionClusterb	-2.036e-01	$< 2e-16$	0.8158	-18.4208
GasRegular	-1.136e-01	5.05e-13	0.8926	-10.7398
BrandClusterb	-2.270e-01	$< 2e-16$	0.7969	-20.3102
PowerNumClusterf-h	4.235e-03	0.8003	1.0042	0.4244
PowerNumClusteri-k	1.299e-01	6.00e-07	1.1387	13.8664
PowerNumClusterl-o	1.009e-01	0.0563	1.1062	10.6173

(Volver)

5.2.1.5 Modelos lineales generalizados (GLM)

5.2.1.5.1 Elección binaria - logit

Cuadro 13: Elección binaria - logit

Concepto	Coeficiente	p-value	Odds-ratio
Constante	-3.435e+00	$< 2e-16$	0.0322
Tiempo de exposición	1.241e+00	$< 2e-16$	3.4604
Edad del vehículo	-1.182e-02	1.46e-13	0.9882
Edad del conductor	-7.929e-03	$< 2e-16$	0.9921
Densidad poblacional	1.738e-05	$< 2e-16$	1.00001738
Gas diesel	-	-	1
Gas regular	-1.035e-01	9.75e-10	0.9017
RegionClusterA	-	-	1
RegionClusterB	-9.216e-02	2.11e-06	0.9119
BrandClusterA	-	-	1
BrandClusterB	-3.811e-01	$< 2e-16$	0.6830
PowerNumbClusterd-f	-	-	1
PowerNumbClusterf-h	2.051e-02	0.2554	1.0207
PowerNumbClusteri-k	1.454e-01	1.98e-07	1.1565
PowerNumbClusterl-o	1.071e-01	0.0596	1.1129

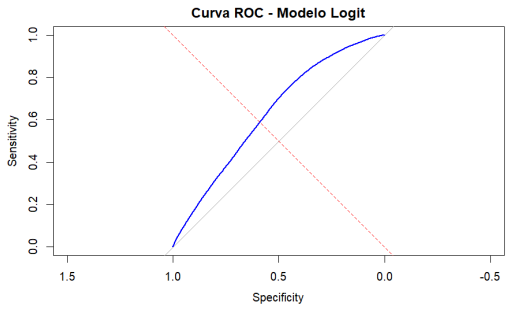
(Volver)

Cuadro 14: Elección binaria - logit

Real / Predicción	0	1
0	397.779	16.181
1	0	0

(Volver)

Figura 17: Elección binaria - logit

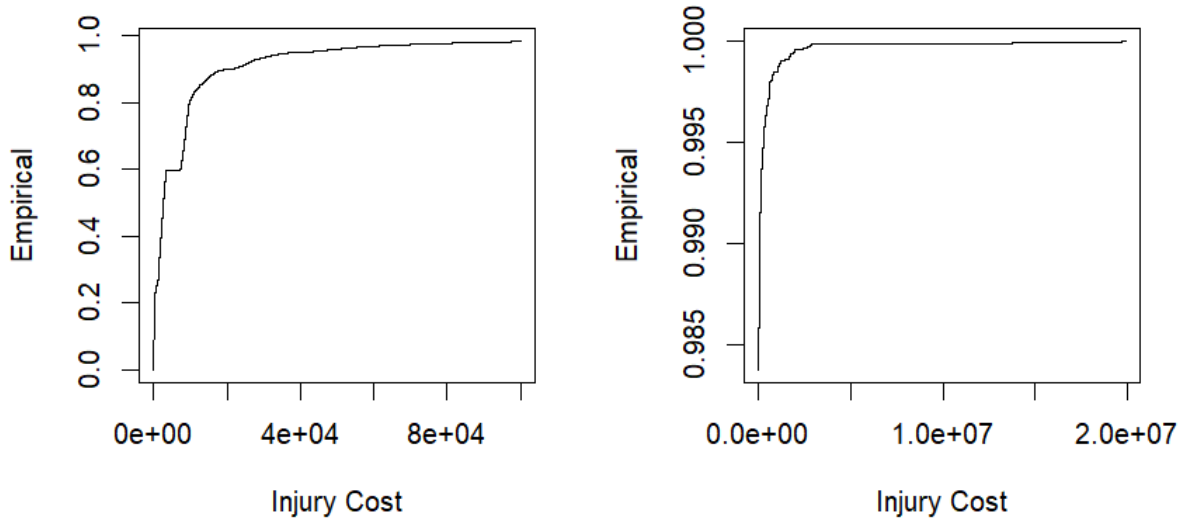


(Volver)

5.2.2. Costo del siniestro

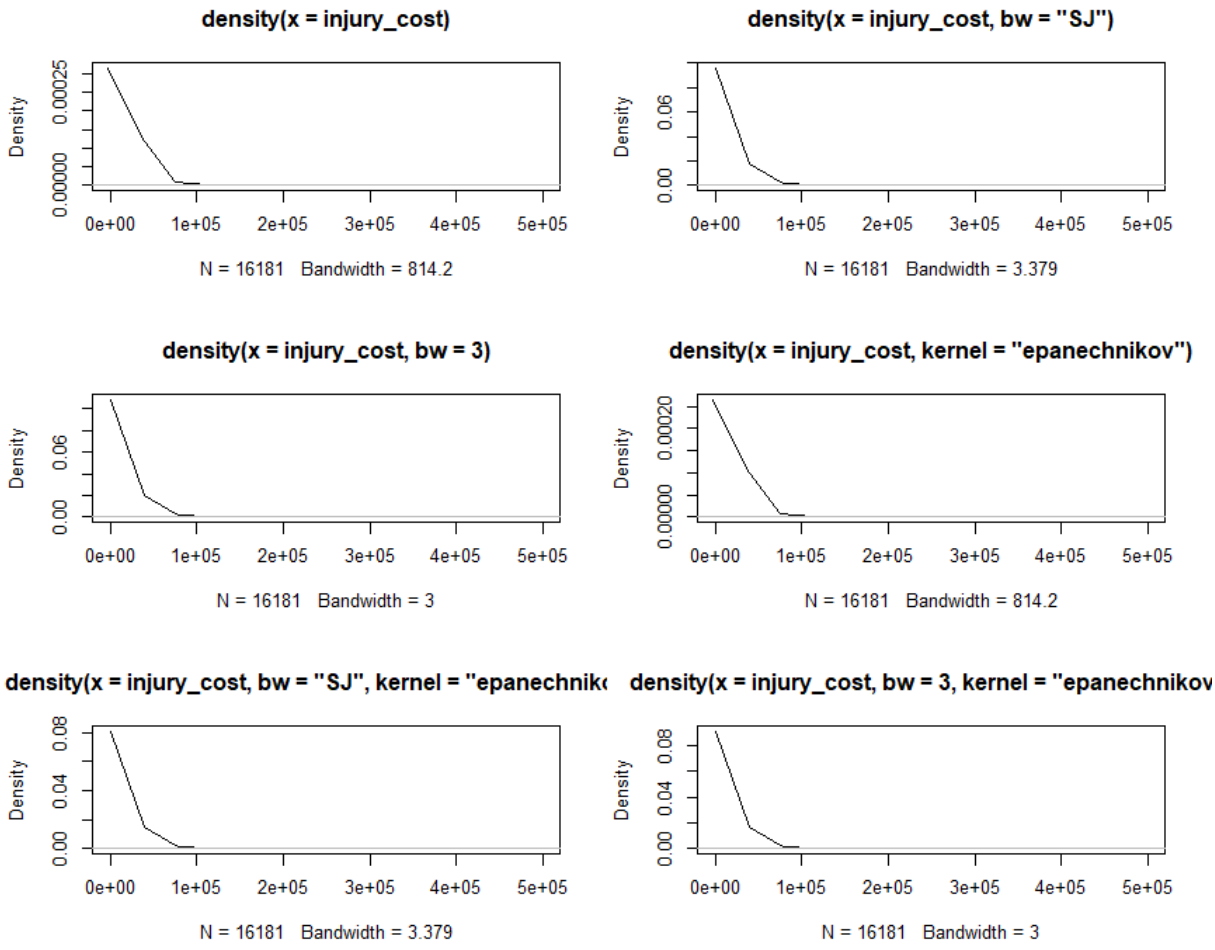
5.2.2.1 Modelización no paramétrica

Figura 18: Mod. no paramétrica



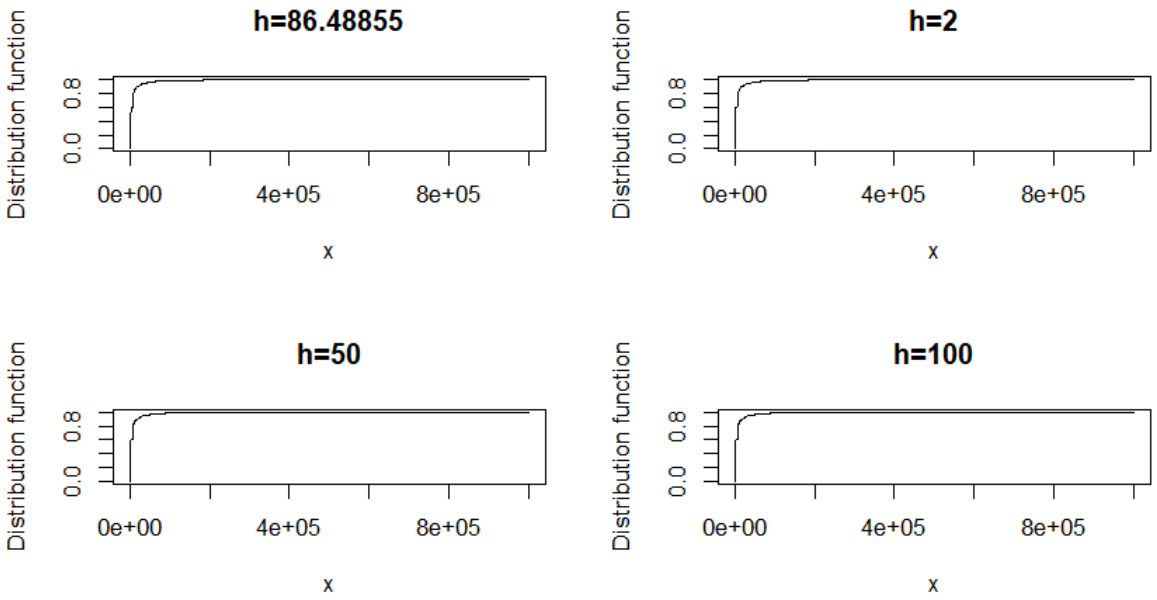
(Volver)

Figura 19: Modelización no paramétrica



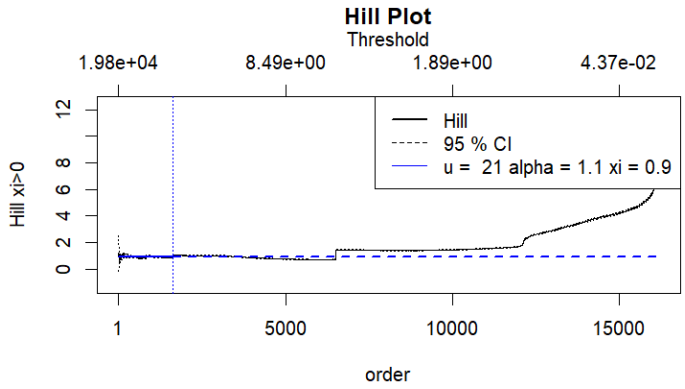
(Volver)

Figura 20: Modelización no paramétrica



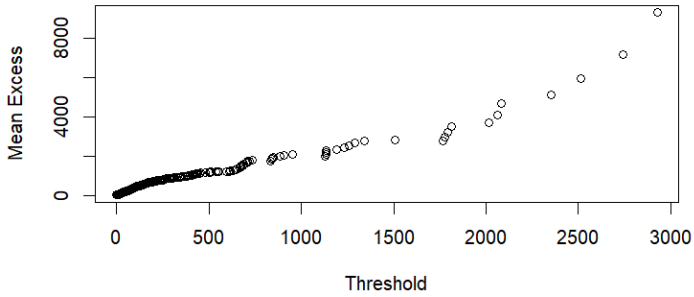
(Volver)

Figura 21: Teoría de valores extremos



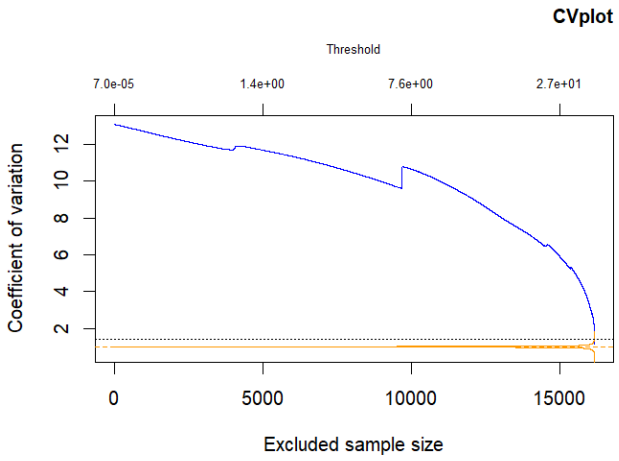
(Volver)

Figura 22: Teoría de valores extremos



(Volver)

Figura 23: Teoría de valores extremos



(Volver)

5.2.2.2 Modelización paramétrica

5.2.2.3 Teoría de valores extremos

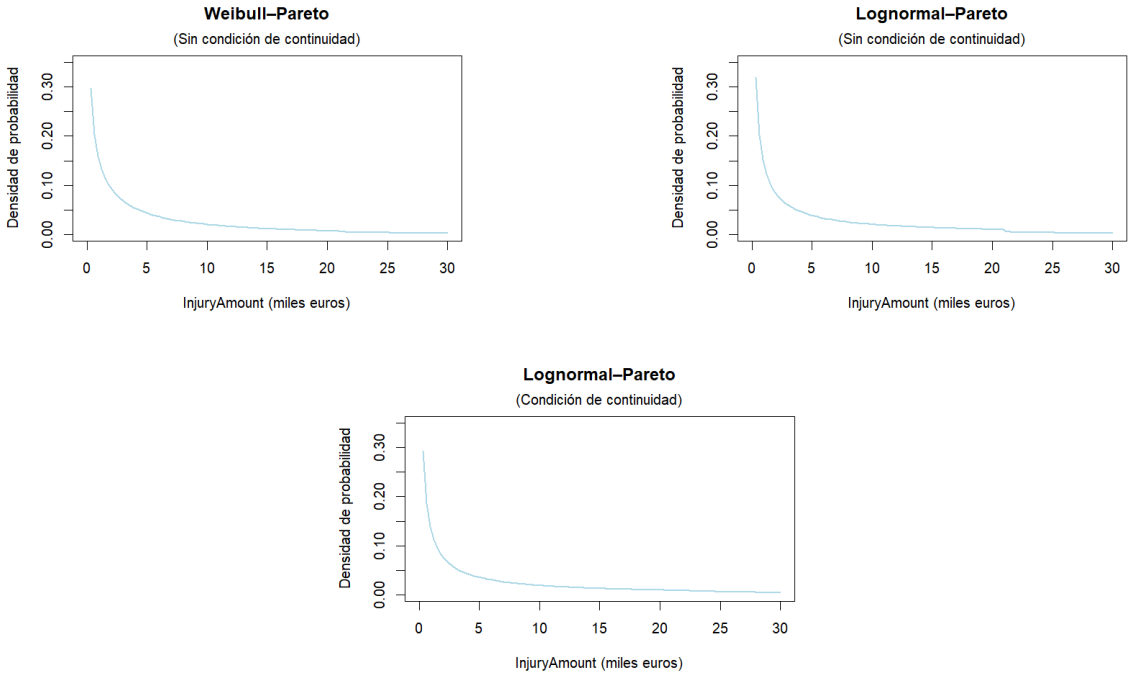
5.2.2.4 Distribuciones compuestas

Cuadro 15: Distribuciones compuestas

Concepto	Wb	LN	Wb-Pareto	LN-Pareto	LN-Pareto (C)
scale	5,71	-	5,53	-	-
shape	0,51	-	0,59	-	-
meanlog	-	0,71	-	3,02	3,02
sdlog	-	2,23	-	3,37	3,37
alpha	-	-	1,11	1,11	1,11
theta	-	-	21,09	21,09	21,09
r	-	-	0,90	0,90	0,83
loglik	-47.062	-47.382	-46.356	-46.450	-48.809
AIC	94.219	94.767	92.720	92.906	97.621
$VaR_{99,5\,\%}$ (miles euros)	151	637	309	309	509

(Volver)

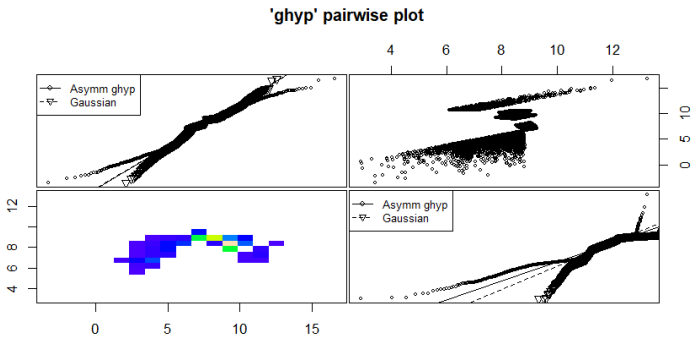
Figura 24: Distribuciones compuestas



(Volver)

5.2.2.5 Distribuciones multivariadas

Figura 25: Distribuciones multivariadas



(Volver)

Referencias

[Estado, 2004] Estado, J. (2004). Boe-a-2004-18911 real decreto legislativo 8/2004, por el que se aprueba el texto refundido de la ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor.

[Jefatura Estado, 2015] Jefatura Estado (2015). Ley 35/2015, de 22 de septiembre, de reforma del sistema para la valoración de los daños y perjuicios causados a las personas en accidentes de circulación.

[Pechon et al., 2018] Pechon, F., Trufin, J., and Denuit, M. (2018). Multivariate modelling of household claim frequencies in motor third party liability insurance. *ASTIN Bulletin: The Journal of the IAA*, 48(3):969–993.

[Santolino and Ayuso, 2007] Santolino, M. and Ayuso Gutiérrez, M. (2007). Una revisión metodológica de la valoración actuarial de los siniestros con daños corporales en el seguro del automóvil. *Anales del Instituto de Actuarios Españoles*, (13):143–172.

[Santolino, 2011] Santolino, M. (2011). *Métodos econométricos para la valoración cualitativa y cuantitativa del daño corporal en el seguro del automóvil*. PhD thesis, Universitat de Barcelona. Book Title: Métodos econométricos para la valoración cualitativa y cuantitativa del daño corporal en el seguro del automóvil ISBN: 9788469368862.

[Tzougas and di Cerchiara, 2023] Tzougas, G. and di Cerchiara, A. P. (2023). Bivariate Mixed Poisson Regression Models with Varying Dispersion. *North American Actuarial Journal*, 27(2):211–241.

[Weisberg and Derrig,] Weisberg, H. I. and Derrig, R. A. (1998). Quantitative methods for detecting fraudulent automobile bodily injury claims.

[Álvarez et al., 2010] Álvarez Jareño, J. A., Muñiz Rodríguez, P., Álvarez Jareño, J. A., and Muñiz Rodríguez, P. (2010). Reparametrización de las principales distribuciones de probabilidad en el estudio del número de siniestros debido: determinación del índice de dispersión. *Anales del Instituto de Actuarios Españoles*, 16:1–24.