Informe ejecutivo: 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

10 de Enero de 2026

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.

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 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.

Los siniestros con daños corporales constituyen un subconjunto específico dentro del conjunto total de siniestros en el seguro de automóviles.

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 sobredispersión (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 Pólya-Aeppli, la distribución Poisson Inversa Gaussiana y la distribución Poisson Pascal Generalizada.

Otros autores como Pechon, Trufin y Denuit 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 [Pechon et al., 2018].

Otro enfoque a tener en cuenta es el de [Tzougas and di Cerchiara, 2023]. En su artículo, desarrollan 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.

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, pueden superar el millón de euros [Santolino and Ayuso, 2007].

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, 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, 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— y cuantitativos —porcentaje de culpa, coste médico total—. Los resultados muestran que la contratación de un abogado, la clasificación de la lesión como grave y la presencia de fracturas incrementan la indemnización.

En el contexto de siniestros, el modelo logit ordenado 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. Para abordar limitaciones del modelo logit ordenado clásico, 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 [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. 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, lo que obliga a realizar provisiones adecuadas para cubrir el coste esperado [Santolino, 2011].

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. Un análisis detallado en esta franja no solo contribuye a reducir el riesgo de sobreindemnización, sino que también permite optimizar la asignación de reservas [Weisberg and Derrig, 1998].

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

El análisis realizado sobre la cartera de seguros muestra que la frecuencia de siniestros es muy baja: el 96 % de las pólizas no registraron ningún reclamo, lo que se traduce en una media prácticamente nula. Esto indica que la mayoría de los asegurados no generan coste, aunque existe un pequeño grupo con múltiples siniestros que concentra el riesgo y puede tener un impacto significativo en la compañía.

En cuanto a la duración de las pólizas, se observa una gran homogeneidad, con una vigencia promedio cercana a medio año y poca variabilidad entre contratos. Esto sugiere que la exposición al riesgo es bastante uniforme en términos temporales.

Respecto a las características del vehículo y del conductor, el perfil predominante corresponde a conductores adultos, con una edad media de 45 años, y vehículos relativamente nuevos, con una antigüedad promedio de 7,5 años. Aunque existen casos extremos, como vehículos muy antiguos, estos son excepcionales y no alteran la tendencia general.

El análisis del entorno geográfico revela una fuerte disparidad en la densidad poblacional: la mayoría de los asegurados reside en zonas poco pobladas, pero también hay pólizas en grandes ciudades, lo que incrementa la exposición al riesgo en entornos urbanos.

En relación con los costes de los siniestros, se confirma una alta concentración en valores nulos, ya que más de la mitad de las pólizas no presentan reclamaciones. Sin embargo, cuando

ocurren, la mayoría de los reclamos son de bajo importe, aunque existen casos excepcionales que superan varios millones de euros. El coste promedio del total del reclamo es de 832,57 euros y el 96.9— % de las pólizas que reportaron algún siniestro no superan los 50 mil euros.

Estos siniestros graves, aunque poco frecuentes, representan un riesgo económico considerable y justifican la necesidad de contar con reservas suficientes para cubrir eventos extremos.

Por último, el análisis de las variables categóricas muestra que la cartera está concentrada en vehículos de potencia media y en marcas como Renault, Nissan o Citroën, que representan más de la mitad del total. El tipo de combustible se distribuye de manera equilibrada entre diésel y gasolina, mientras que la región Centre concentra una proporción significativa de pólizas. Esta concentración en determinados segmentos puede influir en la exposición agregada al riesgo.

En síntesis, el análisis confirma que el riesgo está altamente concentrado en pocos casos de gran severidad, mientras que la mayoría de los asegurados no generan coste. Este patrón refuerza la importancia de una adecuada gestión del riesgo extremo y de provisiones suficientes para garantizar la estabilidad financiera.

3.1.2. Análisis descriptivo bivariado

El coste total del siniestro está dominado por las lesiones corporales. Prácticamente, cuando hay lesiones, el importe global del reclamo se dispara y explica casi todo el coste. Los daños materiales también contribuyen, pero con un peso claramente menor. Esto significa que nuestra exposición económica se juega en el ámbito corporal, no en el material: es ahí donde se concentran los eventos que comprometen reservas y resultados.

La relación entre número de siniestros y coste no es lineal ni estable. Aunque más siniestros tienden a asociarse con más daños materiales, no vemos un patrón consistente que los conecte con mayores lesiones corporales. En otras palabras, la severidad manda: un único siniestro con daños corporales relevantes puede tener más impacto que varios siniestros materiales de bajo importe.

Las variables del perfil (edad del conductor, antigüedad del vehículo, potencia, marca, tipo de combustible) no muestran una relación clara y útil con el coste de las lesiones. Hay outliers en diversas categorías, pero no justifican estrategias de selección de riesgo o tarificación basadas en estas características. Esto es importante: no hay palancas obvias en el perfil del cliente o del coche que reduzcan el coste por lesiones.

El entorno (regiones y densidad poblacional) sí muestra matices: en zonas de alta exposición (e.g., grandes áreas urbanas) se observa algo más de frecuencia, pero sin evidencia de que ello se traduzca en lesiones corporales más costosas de forma sistemática. El mensaje para Gerencia es que frecuencia y severidad se comportan de forma distinta y deben gestionarse con herramientas diferentes.

Por último, los casos extremos aparecen dispersos entre categorías y regiones. No están concentrados en un segmento específico, por lo que la gestión del riesgo extremo debe ser transversal, con foco en procedimientos de reclamación y control de costes médicos/legales, más que en la segmentación del cliente o del vehículo.

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*. 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

El análisis se aplicó sobre la variable número de siniestros (ClaimNb), caracterizada por una alta concentración de pólizas sin reclamos y algunos casos con múltiples siniestros. El objetivo fue segmentar la cartera para identificar perfiles con distinta probabilidad de siniestro mediante un árbol jerárquico que organiza las divisiones en función de las variables más influyentes.

La primera partición se realizó sobre la exposición (duración efectiva del contrato), utilizando como umbral el valor de 0,29 años. Este corte divide la cartera en dos grandes grupos: pólizas con baja exposición (igual o inferior a 0,29 años) y pólizas con alta exposición (superior a 0,29 años). A partir de esta división principal, el modelo profundiza en cada rama con criterios adicionales.

En el grupo de baja exposición, el siguiente factor relevante es la antigüedad del vehículo, que separa autos nuevos de los más antiguos, y posteriormente la marca del vehículo, distinguiendo entre marcas occidentales y asiáticas.

En el grupo de alta exposición, la primera variable discriminante es la marca, seguida por la región geográfica, que permite diferenciar zonas con comportamientos distintos, y finalmente un nuevo corte en exposición en torno a 0,65 años para refinar el análisis.

Esta estructura jerárquica confirma que la duración del contrato es el factor más determinante, mientras que la marca y la región aportan información adicional en contratos de mayor vigencia, y la antigüedad del vehículo resulta útil en contratos cortos.

3.2.1.2 Clúster no jerárquico - k means

Se aplicó k-means de forma independiente sobre cada variable clave—exposición del contrato (Exposure), antigüedad del vehículo (CarAge), edad del conductor (DriverAge), densidad poblacional (Density) y potencia del vehículo (Power)—tras estandarizar los datos. El número óptimo de clústeres se determinó mediante el método del codo, que indicó que cuatro grupos por variable ofrecían el mejor equilibrio entre simplicidad e información.

En exposición, los grupos más relevantes son el de larga duración, con una media cercana a 1 año y casos hasta 2 años, y el de muy corta, con una media de 0,1 años y un máximo de 0,3. Ambos concentran la mayor parte de la cartera. En antigüedad del vehículo, predominan los clústeres de vehículos nuevos (0–6 años, media 2,7) y intermedios (7–13 años, media 9,9), que juntos representan la mayoría del parque. En edad del conductor, los segmentos más frecuentes son adultos (35–47 años, media 40,9) y conductores mayores (48–62 años, media 53,9). En densidad poblacional, el clúster de baja densidad (media 399 hab/km², máximo 2.317) concentra la mayor parte de las pólizas, mientras que los entornos metropolitanos (hasta 27.000 hab/km²) son minoritarios. Por último, en potencia, los niveles bajos y medio-bajos (categorías 1–5) agrupan la gran mayoría de los vehículos, mientras que los niveles altos son

residuales.

Esta segmentación independiente por variable permite describir la cartera en términos claros: contratos mayoritariamente de corta o larga duración, vehículos nuevos o intermedios, conductores adultos, entornos poco densos y potencias bajas o medias.

3.2.1.3 Análisis PCA

El PCA se aplicó para sintetizar la información de cinco variables cuantitativas —exposición (Exposure), antigüedad del vehículo (CarAge), edad del conductor (DriverAge), densidad poblacional (Density) y potencia (Power)— y comprobar si existía redundancia entre ellas.

El resultado es claro: no hay una única dimensión que explique el conjunto. Con dos componentes se recoge aproximadamente el 48 % de la variabilidad y, aun ampliando a tres componentes, se llega solo a 68 %. En términos prácticos, cada variable aporta información distinta y conviene mantenerlas todas en el análisis. Entre ellas, exposición y densidad son las que más estructura aportan cuando se observan de forma conjunta; la edad del conductor añade información adicional, y la potencia tiene un peso menor en la explicación global.

En resumen, el PCA confirma que no procede simplificar eliminando factores y que el seguimiento debe concentrarse especialmente en exposición y densidad, sin perder de vista el resto.

3.2.1.4 Modelo Poisson

El modelo Poisson se empleó para explicar la frecuencia de siniestros (ClaimNb) como variable de conteo con alta presencia de ceros, ajustando por la exposición mediante un offset en log(Exposure) para trabajar en términos de tasa.

La construcción siguió un proceso incremental: partimos de un modelo básico con la exposición y fuimos añadiendo variables una a una, evaluando en cada paso con ANOVA si la nueva inclusión mejoraba significativamente el modelo anterior. Con este criterio, el mejor modelo fue el que incluye todas las variables explicativas disponibles, sin incorporar variables de coste (para mantener la separación actuarial entre frecuencia y severidad).

Para mejorar la interpretabilidad, trabajamos con clústeres en la construcción del modelo: regiones, marcas y potencia en tramos derivados de la segmentación previa.

Los principales resultados son:

- Edad del conductor y antigüedad del vehículo muestran efectos de reducción: conforme aumentan, la frecuencia esperada disminuye ligeramente.
- Densidad poblacional aporta una señal positiva (más densidad, mayor frecuencia).

Desde el punto de vista técnico, el modelo ajusta bien y se descarta sobredispersión; lo que hace innecesarias extensiones como cuasi-Poisson o binomial negativa.

3.2.1.5 Modelos lineales generalizados (GLM)

3.2.1.5.1 Elección binaria - logit

El modelo Logit se aplicó para analizar la probabilidad de que una póliza registre al menos un siniestro frente a no tener ninguno, transformando la variable objetivo en binaria (0 = sin siniestros, 1 = con siniestros). Este enfoque permite estimar cómo influyen las características del contrato, del vehículo y del asegurado en la ocurrencia del evento, considerando relaciones

no lineales entre variables.

La construcción del modelo siguió el mismo criterio incremental que en Poisson: partimos de un modelo básico y fuimos incorporando variables una a una, evaluando con el AIC si la inclusión mejoraba el ajuste. También en este caso se trabajó con clústeres para variables categóricas (marca, región y potencia). El modelo final incluyó todas las variables explicativas relevantes, sin incorporar costes, para mantener la separación entre frecuencia y severidad.

Los resultados confirman que la exposición es el factor más influyente: por cada unidad adicional, la razón de probabilidades (odds) se multiplica por 3,46, lo que indica un incremento muy significativo en la probabilidad de siniestro. La edad del conductor y la antigüedad del vehículo reducen ligeramente la probabilidad a medida que aumentan, mientras que la densidad poblacional aporta un efecto positivo, aunque pequeño. Entre las variables categóricas, se observan diferencias entre clústeres de marca y potencia, y los vehículos con gasolina regular presentan odds menores que los diésel.

A pesar de estos hallazgos, el modelo Logit mostró limitaciones importantes: la variable objetivo está fuertemente desbalanceada (96 % de las pólizas sin siniestros), lo que provocó que el modelo tendiera a predecir siempre la clase mayoritaria. Incluso tras ajustar el umbral con análisis ROC, la matriz de confusión evidenció que no se lograba una clasificación efectiva. El modelo no resulta adecuado como herramienta predictiva en esta base de datos debido al desbalance extremo.

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 la modelización no paramétrica trabajamos únicamente con pólizas que presentan costes positivos para lesiones corporales (InjuryAmount) y daños materiales (PropertyAmount), con el fin de evitar la distorsión que provoca la concentración de ceros. El objetivo fue estimar el Valor en Riesgo (VaR) al 99,5 % sin imponer una forma funcional, comparando la distribución empírica con la estimación por núcleo (kernel).

Sobre la distribución empírica, el VaR 99,5 % de InjuryAmount asciende a 349,8 miles de euros, mientras que el de PropertyAmount es de 10,5 miles de euros. El VaR 99,5 % del total (ClaimAmount) se sitúa en 354,3 miles de euros. Si se suman los VaR marginales de lesiones y materiales, el resultado (359,3 miles de euros) queda por encima del VaR del total, lo que muestra que la agregación simple de riesgos marginales puede ofrecer una visión más conservadora que el comportamiento conjunto del total en este nivel de confianza.

Con estimación por núcleo (bandwidth por defecto, h $\approx$ 86,49), el VaR 99,5 % de InjuryAmount aumenta ligeramente hasta 350,4 miles de euros, y el de PropertyAmount permanece en torno a 10,5 miles de euros; el ClaimAmount se mantiene en 354,3 miles de euros. La suma de VaR marginales bajo kernel alcanza 361,0 miles de euros, de nuevo por encima del VaR del total. Este patrón confirma dos ideas clave: primero, el riesgo extremo está dominado por las lesiones corporales, ya que el aporte de materiales al cuantil alto es reducido; segundo, la

suavización por kernel tiende a extender la cola y, por tanto, a elevar levemente el VaR frente a la empírica en lesiones, mientras que en materiales el efecto es prácticamente neutro. En consecuencia, al 99,5 % conviene reportar tanto el VaR del total como la suma de VaR marginales y señalar explícitamente que el método kernel produce valores ligeramente superiores a la empírica en lesiones, reflejando una asignación de mayor peso a escenarios extremos coherente con la cola pesada observada en los datos.

3.2.2.2 Modelización paramétrica

Para estimar el coste por daños corporales (InjuryAmount) bajo supuestos paramétricos se contrastaron, primero, distribuciones clásicas ajustadas directamente sobre los costes positivos (Weibull, Gamma, Log-normal y Log-logística) y, después, familias más flexibles ajustadas sobre la escala logarítmica —en sus versiones asimétricas— como Generalized Hyperbolic (GH), Hiperbólica, t-Student, NIG y Variance-Gamma (VG). La comparación se realizó mediante AIC homogéneo (corrigiendo el de los modelos en log-costes) y VaR en distintos niveles de confianza.

Los resultados muestran un patrón claro: las familias clásicas presentan sesgos en cola en sentidos opuestos —Weibull subestima los cuantiles altos, mientras que la Log-logística los sobreestima y la Log-normal se sitúa en un punto intermedio con colas todavía elevadas.

En cambio, las familias flexibles en log-costes mejoran el ajuste global: GH asimétrica alcanza AIC 91.922 con VaR 99,5 % $\approx$ 239 mil euros, y VG asimétrica AIC 91.919 con VaR 99,5 % $\approx$ 240 mil euros.

El mejor ajuste estadístico (AIC) lo aportan GH y VG asimétricas en log-costes, pero su VaR 99,5 % se sitúa sustancialmente por debajo de los valores empíricos/no paramétricos reportados ($\approx$ 350 mil euros), lo que indica que los modelos paramétricos tienden a ser menos conservadores frente a la cola pesada observada en los datos reales.

Como contraste adicional, para las distribuciones GH y VG el VaR 99,5 % del total (ClaimAmount) supera la suma de los VaR marginales (Injury + Property), lo que evidencia que el comportamiento conjunto puede generar agregaciones más severas que las estimaciones individuales, un matiz relevante para la gestión del riesgo agregado.

En definitiva, se recomienda priorizar familias flexibles en log-costes (GH y VG asimétricas) por calidad de ajuste (AIC); no obstante, conviene interpretarlas con cautela si el objetivo es capturar de forma conservadora la cola pesada de InjuryAmount.

3.2.2.3 Teoría de valores extremos

El diagnóstico de cola pesada se apoyó en tres pruebas complementarias sobre InjuryAmount. El Hill Plot mostró una zona amplia de estabilidad, esta estabilidad es precisamente el indicador de que estamos ante una cola con comportamiento del tipo Pareto. El Mean Excess Plot presentó una tendencia creciente y aproximadamente lineal, rasgo característico de distribuciones de cola pesada. El CV-Plot mantuvo el coeficiente de variación por encima de 1 en la región relevante, lo que descarta tanto colas exponenciales (CV $\approx$ 1) como colas ligeras (CV $<$ 1) y refuerza la conclusión de cola pesada.

La coherencia entre estas tres evidencias valida que la severidad observada se concentra en un subconjunto de eventos de baja frecuencia y alto impacto, para los cuales los enfoques estándar tienden a infraestimar el riesgo.

Sobre esa base, el modelo POT-GPD registra un VaR 99,5 % que se sitúa en torno a 286

mil euros, muy próximo al estimado por el enfoque Pareto simple (284 mil euros), lo que aporta robustez al resultado y confirma que la cuantificación de la cola mediante EVT es más conservadora que los modelos paramétricos en log-costes.

En resumen, esto significa que POT-GPD captura mejor el comportamiento extremo observado, ofreciendo un punto de referencia prudente para dimensionar reservas en percentiles altos; su uso como complemento a los modelos paramétricos resulta aconsejable cuando la prioridad es mitigar el riesgo de subestimación en escenarios críticos.

3.2.2.4 Distribuciones compuestas

El análisis parte de una limitación estructural: las distribuciones simples no logran representar simultáneamente el comportamiento del cuerpo y la cola de la distribución de costes. Weibull, por ejemplo, ajusta bien la zona central pero subestima los cuantiles extremos, mientras que Lognormal tiende a inflarlos, generando estimaciones poco realistas en escenarios severos.

Para abordar esta complejidad, se aplicaron modelos compuestos, que combinan dos distribuciones: una para el cuerpo (Weibull o Lognormal) y otra para la cola (Pareto), separadas por un umbral determinado mediante diagnóstico EVT (Hill Plot). Este enfoque permite que cada parte del modelo se especialice en la zona donde ofrece mejor ajuste, evitando el sesgo sistemático que se observa en los modelos simples.

Los resultados confirman que la combinación Weibull–Pareto es la más robusta, alcanzando el mayor loglik y el menor AIC entre las alternativas evaluadas. Además, su estimación del VaR 99,5 % $\approx$ 309 mil euros se sitúa en un rango intermedio, reduciendo la subestimación de Weibull y la sobreestimación de Lognormal.

Un aspecto relevante es que imponer continuidad en la función de densidad (para evitar saltos en el umbral) no siempre mejora el ajuste global: aunque suaviza la transición visual entre cuerpo y cola, incrementa el AIC y reduce la verosimilitud, lo que indica que la prioridad debe ser la precisión estadística antes que la estética del modelo.

3.2.2.5 Distribuciones multivariadas

El objetivo de esta sección fue modelizar conjuntamente los costes por daños corporales (InjuryAmount) y materiales (PropertyAmount), reconociendo que ambos presentan dependencia y colas pesadas. Las pruebas de normalidad (Jarque-Bera y Mardia) confirmaron que ni las marginales ni la distribución conjunta siguen un patrón gaussiano, lo que nos empuja a recurrir a familias más flexibles.

Se evaluaron distribuciones multivariadas avanzadas —Generalized Hyperbolic (GH), Hiperbólica, t-Student, Normal Inversa Gaussiana (NIG) y Variance-Gamma (VG)— en versiones simétricas y asimétricas. El criterio AIC fue el eje de comparación. Los resultados muestran que la GH asimétrica ofrece el mejor ajuste global (AIC más bajo). Sin embargo, incluso con este modelo, el comportamiento en los extremos sigue siendo insuficiente: el VaR 99,5 % calculado, 2.487 miles de euros, es muy sensible y puede alcanzar valores desproporcionados, reflejando la dificultad de representar simultáneamente colas pesadas en dos dimensiones.

3.2.2.6 Cópulas

Las variables InjuryAmount y PropertyAmount presentan dependencias negativas por rangos. Por ello, no resulta adecuado ajustar una cópula explícita. Además, debido a la distribución de los datos, la única opción razonable es aplicar la cópula gaussiana.

Se trabajó la cópula gaussiana en dos escenarios: con marginales log-normales y con marginales log-logísticas. 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 99.5 %, la pérdida máxima esperada asciende a 1.712 miles de euros, frente a los 646 miles de euros de las marginales con log-normales.

En términos de riesgo, el VaR conjunto al 99,5 % bajo esta cópula alcanza valores muy elevados, muy superiores a los VaR marginales, lo que evidencia que la agregación bajo dependencia puede generar escenarios extremos significativamente más severos que la suma de riesgos individuales.

3.2.2.7 Comparativo

Los resultados para InjuryAmount convergen en la misma conclusión: el riesgo extremo de la línea está dominado por la cola de lesiones corporales y su cuantificación es sensible al enfoque de modelización empleado.

Los métodos no paramétricos sitúan el VaR 99,5 % alrededor de 350 miles de euros (empírico 349,8; kernel 350.4), lo que ofrece una referencia fiel y conservadora de la realidad muestral, coherente con una cola pesada. El kernel incrementa levemente el cuantil respecto a la empírica, como es esperable cuando el suavizado asigna algo más de masa a los extremos.

Frente a ello, las familias paramétricas flexibles en log-costes (GH y VG asimétricas) mejoran los criterios de ajuste global (AIC), pero producen VaR sustancialmente menores (239-240 miles de euros). Esto indica que estos modelos, aun siendo útiles para describir el cuerpo de la distribución y para tareas analíticas, no son suficientes por sí solos para fijar reservas o capital en niveles regulatorios, porque tienden a infraestimar la severidad en el extremo.

La EVT (POT-GPD y Pareto) estima un VaR de 284-286 miles de euros, sirviendo como una referencia conservadora y coherente con la evidencia de cola pesada identificada en el análisis. Por su parte, el modelo compuesto Weibull–Pareto ofrece un VaR intermedio (309 miles de euros).

4. Conclusiones

El análisis de la cartera de automóviles confirma un perfil de riesgo inequívoco y altamente relevante para la toma de decisiones: la frecuencia es muy baja y la intensidad está extraordinariamente concentrada en un número reducido de siniestros con lesiones corporales, que son los que determinan la volatilidad de resultados y las necesidades de reserva. En la muestra trabajada, aproximadamente el 96 % de las pólizas no registran siniestros, mientras que un subconjunto minoritario concentra importes elevados que explican el coste agregado y exponen a la entidad a pérdidas extremas; este patrón exige prestar particular atención en la cola de la distribución, donde realmente se juega el riesgo económico, sin descuidar la gobernanza de la frecuencia.

En el análisis de frecuencia, utilizando un modelo GLM Poisson estable y sin indicios de sobredispersión, se identificaron dos patrones relevantes. Por un lado, la edad del conductor y la antigüedad del vehículo presentan un efecto atenuador: a medida que aumentan, la frecuencia esperada de siniestros disminuye ligeramente. Por otro lado, la densidad poblacional muestra una relación positiva con la siniestralidad, lo que refleja un mayor riesgo en entornos urbanos con alta concentración de tráfico.

En el análisis de segmentación, se aplicaron dos enfoques complementarios: clúster jerárquico divisivo y k-means no jerárquico, además de un análisis PCA para evaluar redundancias. El clúster jerárquico, construido sobre la variable número de siniestros, mostró que la exposición temporal es el criterio más discriminante: el primer corte separa pólizas de corta duración (menor igual a 0.29 años) de aquellas con mayor vigencia (mayor a 0.29 años). Dentro de cada grupo, se identificaron subdivisiones relevantes: en contratos cortos, la antigüedad del vehículo y la marca aportan diferenciación; en contratos largos, la marca y la región geográfica son los factores que refinan la segmentación. Este patrón muestra que la duración del contrato es el eje estructural del riesgo.

Por su parte, el análisis k-means se aplicó sobre variables clave (exposición, antigüedad del vehículo, edad del conductor, densidad poblacional y potencia), determinando cuatro grupos óptimos por variable. Los resultados describen una cartera concentrada en dos extremos: pólizas de muy corta duración (0,1 años) y pólizas de larga duración (1 año), vehículos mayoritariamente nuevos o intermedios, conductores adultos (35–47 años) y entornos de baja densidad, con potencias bajas o medias. Esta segmentación no revela perfiles de riesgo extremo concentrados en un clúster específico, sino que confirma que el riesgo está disperso y depende más de la exposición acumulada que de características del vehículo o del asegurado.

El PCA mostró que no existe una dimensión única que explique la variabilidad global: incluso con tres componentes principales, solo se captura el 68 % de la información. Esto significa que cada variable aporta señales distintas y que simplificar eliminando factores no es recomendable.

Por último, el modelo logit binario para predecir la ocurrencia de al menos un siniestro se descartó debido al fuerte desbalance de clases (96 % de pólizas sin siniestros), que limita su capacidad predictiva y podría inducir decisiones basadas en una falsa sensación de precisión.

En el bloque de intensidad, se muestra que InjuryAmount domina el coste total cuando hay lesiones corporales, mientras que PropertyAmount (daños materiales) resulta marginal en los cuantiles altos; esto reafirma que la exposición económica decisiva de la entidad está en el ámbito corporal, no en el material.

La evidencia no paramétrica—comparando la distribución empírica con la estimación por núcleo—muestra cuantiles extremos elevados y confirma que el suavizado por kernel extiende ligeramente la cola, elevando el VaR frente a la empírica.

En paralelo, las familias paramétricas clásicas (Weibull, Gamma, Lognormal, Log-logística) presentan sesgos contrapuestos en la cola: Weibull tiende a subestimar los extremos, mientras que Log-logística los sobreestima; la Lognormal queda en un punto intermedio, aún con colas elevadas. Aun así, las familias más flexibles (Generalized Hyperbolic, Variance-Gamma, NIG, t asimétrica) mejoran los criterios de ajuste global (AIC), pero no resuelven por completo el reto del extremo: sus VaR al 99,5 % resultan sistemáticamente inferiores a los obtenidos con referencias empíricas o con Teoría de Valores Extremos (EVT). La EVT—diagnosticada mediante Hill Plot, Mean Excess Plot y CV-Plot—corrobora la presencia de cola pesada en lesiones corporales y, mediante ajustes POT-GPD y Pareto, entrega cuantiles altos más prudentes que las familias paramétricas en log-costes. Esta capa metodológica ofrece un piso conservador que alinea la cuantificación con el comportamiento real del extremo y que, por tanto, resulta especialmente útil para la planificación de reservas.

Como solución intermedia y de alto valor práctico, los modelos compuestos (por ejemplo, Weibull–Pareto para cuerpo y cola, separados por un umbral identificado con EVT) presentan el buen ajuste central con una representación fiel del extremo, alcanzando mejor verosimilitud y AIC que alternativas simples y situando el VaR en un rango intermedio razonable que reduce tanto la subestimación de familias ligeras como la sobreestimación de las más pesadas.

El análisis de distribuciones multivariadas confirma que los log-costes por lesiones corporales y daños materiales no siguen un patrón normal conjunto, lo que dificulta su modelización con enfoques clásicos. Las pruebas de normalidad multivariante y los ajustes realizados muestran que las familias flexibles, como la Generalized Hyperbolic (GH) en versión asimétrica, ofrecen el mejor ajuste global según el criterio AIC. Sin embargo, incluso con estos modelos, la representación de los extremos sigue siendo insuficiente: el VaR conjunto al 99.5 % puede alcanzar valores muy elevados y sensibles a la especificación, reflejando la complejidad de capturar simultáneamente colas pesadas en dos dimensiones.

La agregación del riesgo entre lesiones y materiales añade una dimensión crítica para decisiones de capital: la dependencia entre ambas variables puede amplificar significativamente el VaR conjunto, situándolo muy por encima de los VaR marginales e incluso de su suma. En el ejercicio, la cópula gaussiana con marginales log-logísticas ofreció el mejor compromiso de ajuste y se utilizó para estimar el VaR conjunto al 99.5 %, mostrando que la agregación bajo dependencia puede generar escenarios más severos que los sugeridos por un análisis aislado de cada componente.

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

[Álvarez et al., 2010] Álvarez, 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.