CAPÍTULO IV

Modelos predictivos del riesgo y aplicaciones a los seguros

Montserrat Guillen*
María Láinez
Ana M. Pérez-Marín
Eduardo Sánchez

El análisis del riesgo estudia los sucesos extraordinarios, qué los causa y cómo mitigar su probabilidad de ocurrencia. En seguros, básicamente importa la frecuencia y la severidad de los siniestros. Usando medidas como los cuantiles alejados de la mediana, la modelización predictiva permite detectar factores que afectan al riesgo. Tras la presentación de la regresión cuantílica como modelo básico y sus generalizaciones, se realiza una recopilación de casos de estudio en el ámbito asegurador, en situaciones de datos masivos y en particular en el análisis de datos telemáticos en seguros del automóvil.

Palabras clave: regresión cuantílica, extremos, seguros de automóvil, datos telemáticos, accidentes.

* Montserrat Guillen y Ana M. Pérez-Marín agradecen el apoyo de la Fundación BBVA en los proyectos de investigación en Big Data, del Ministerio de Ciencia e Innovación (proyecto número PID2019-105986GB-C21) y del programa ICREA Academia. Los autores agradecen las aportaciones de MAPFRE España y MAPFRE SA.

1. INTRODUCCIÓN

El análisis del riesgo tiene por objeto estudiar fenómenos extraordinarios, grandes accidentes y catástrofes que ocasionan cuantiosas pérdidas. Como campo científico, al igual que las ciencias actuariales, siempre ha quedado en tierra de nadie por su multidisciplinariedad; en economía, matemáticas, ingeniería, bioestadística y demografía. Sin embargo, en las últimas décadas, el impacto que el desarrollo tecnológico ha ejercido en el ámbito de los seguros y la gestión de riesgos ha impuesto al estadístico, o al científico de datos, como un perfil imprescindible en esta materia.

Resulta paradójico que el análisis de riesgos se acabe vinculando al big data cuando al hablar de riesgos solemos pensar en fenómenos catastróficos para los que no hay prácticamente ni información, ni antecedentes. Los grandes siniestros son infrecuentes, y se dan en circunstancias incontroladas que los hacen altamente peligrosos, donde además es difícil identificar una regularidad estadística, y con ello diseñar medidas para evitar su ocurrencia.

A pesar de la dificultad de recopilar información para el análisis de riesgos, existen métodos adaptados de por sí a la predicción de fenómenos de muy baja frecuencia. Por ejemplo, para el seguro de automóviles donde la probabilidad de que un conductor sufra un accidente durante un año no suele superar el 10 % en la mayoría de países desarrollados, o para el seguro de vida, donde se trabaja con probabilidades incluso mucho más reducidas, la metodología cuantitativa está bien establecida. Las bases de datos con miles de asegurados ya eran habituales en la segunda mitad del siglo XX, y las entidades aseguradoras han venido utilizando con total normalidad las técnicas de modelización que permiten analizar lo que en inglés se denominan los *rare events.*

Las nuevas tecnologías, los sensores y la internet de las cosas no han hecho más que ampliar las posibilidades del análisis de riesgos, para convertirlo en un auténtico paraíso de la información. Además, y añadido a los logros obtenidos, se ha generado una nueva demanda unánime en la sociedad: la necesidad de fomentar la prevención. La generación de datos para evitar catástrofes es un tema para el que los ciudadanos reclaman establecer niveles de vigilancia, incluso más exhaustivos de lo que sería considerado una invasión a su propia privacidad. Hay numerosos ejemplos de ello, desde la creación de avisos meteorológicos que informan a la población de la inminencia de fenómenos adversos tales como tormentas, huracanes o tsunamis, hasta la prevención de riesgos alimentarios, y lógicamente las pandemias. En consecuencia, el análisis de riesgos se ha transformado completamente en los últimos años, proporcionando respuesta no solo a aspectos meramente predictivos sino también preventivos. En definitiva, la disponibilidad de bases de datos de gran volumen ha multiplicado las posibilidades del análisis de riesgos y ha dado lugar a una nueva era en la elaboración de modelos predictivos, que sirven para anticipar patrones fuera de lo normal, y de modelos *prescriptivos,* cuya utilidad es crear sistemas de protección.

En este capítulo, se muestra cómo se puede implementar el análisis predictivo de los cuantiles para la modelización del riesgo, dejando atrás los modelos de regresión tradicionales focalizados en el análisis de la media. Seguidamente, se ofrecen algunas generalizaciones. También se incluyen varios ejemplos con resultados relativos al uso de datos telemétricos en el seguro de automóviles y, concretamente, en pólizas de pago por uso. Para terminar, se realiza una revisión de otras aplicaciones del big data en los seguros, y se concluye presentando algunas líneas emergentes en este ámbito.

2. PREDECIR FRECUENCIA Y SEVERIDAD ESPERADAS

El seguro es un mecanismo autoprotector en el que un colectivo solidario, formado por los asegurados, se hace cargo de compensar a sus miembros cuando alguno de ellos sufre un siniestro. El vínculo queda establecido a través de una entidad aseguradora, suscribiendo una póliza y realizando el pago de su correspondiente prima. En la mayoría de productos aseguradores, el principal escollo es anticipar cuál es la probabilidad de ocurrencia de un accidente, y si este se produce, cuál se prevé que sea su magnitud[1].

El planteamiento del problema como probabilidad de ocurrencia y seguidamente del coste económico del siniestro, es muy parecido al problema de impago de un crédito, donde por una parte, se modeliza la probabilidad de impago y, posteriormente, la cuantía esperada de la pérdida si el impago ya se ha producido. Sin embargo, el coste máximo total, o el parcial que queda por pagar, en una operación crediticia es una cuantía acotada. En la mayoría de siniestros en el sector asegurador, los límites no son tan claros *a priori,* ya que si bien los bancos, al conceder créditos con cuantía establecida, tienen un intervalo de oscilación de las pérdidas dentro de un margen, la inmensa mayoría de los productos aseguradores, incluso a pesar de que existan cláusulas sobre máximos de responsabilidad pactados contractualmente en las pólizas, pueden llegar a tener rangos de variación entre el coste del siniestro máximo y del siniestro medio que pueden calificarse de gigantescos o, a efectos prácticos, desconocidos. Con la excepción de la mayoría de los seguros de vida, en los que la indemnización ya queda fijada en la póliza y el siniestro solamente se produce una vez, el resto de seguros admiten que pueda producirse más de un accidente durante la vigencia de la póliza y, además como ya se ha mencionado, antes de que ocurran dichos accidentes existe una elevada incertidumbe sobre su potencial severidad. Los productos que tienen: severidad desconocida y posibilidad de reiteración de siniestro en un mismo periodo de cobertura son los más habituales, de hecho se conocen como los *seguros generales.* Tal es el caso por ejemplo, en un seguro de salud, donde se puede requerir asistencia en varias ocasiones, acudiendo a uno o más especialistas e incurriendo en gastos médicos difícilmente previsibles.

[1] Aquí no trataremos las consecuencias de los accidentes desde perspectivas ajenas a la compensación económica. La prevención de accidentes y gestión de riesgos pueden ir mucho más allá y considerar daños irreparables, como la pérdida de vidas humanas.

El cálculo de primas más básico consiste en multiplicar el número esperado de siniestros por su coste medio. Obtenida esta cantidad, que suele referirse a un periodo anual, se aplican ajustes de seguridad y recargos para gastos de administración y de adquisición, y se determina el precio final que pagará el tomador del seguro, lo que en términos técnicos se conoce como su *prima de transferencia del riesgo.* La suma de las primas de los asegurados de un mismo colectivo garantiza fondos suficientes para hacer frente de forma mancomunada y solidaria a todos los siniestros del colectivo. Debido a la gran responsabilidad asumida por las entidades aseguradoras en su compromiso de resarcir de las pérdidas a sus asegurados, el sector en su totalidad queda sujeto a una regulación férrea, a un nivel incluso más exigente que el aplicado en otras áreas de actividad del sector financiero. Todo ello implica un control de la solvencia de las entidades y, sobre todo, una garantía de corrección de los cálculos actuariales necesarios para proveer las primas. De ahí el papel fundamental del análisis de riesgos basado en los datos.

Como el principio del cálculo de precios se fundamenta en un modelo predictivo orientado a modelizar el valor esperado de una variable de conteo (la frecuencia de siniestralidad) y de una variable positiva (el coste o severidad), que generalmente se asume no acotada y como ya se ha comentado antes, es asimétrica a la derecha, los modelos lineales generalizados, árboles de clasificación, redes neuronales y *random forests,* entre otros, son los métodos de *machine learning* que se vienen utilizando con total normalidad en los departamentos actuariales de las entidades aseguradoras y que vinculan la siniestralidad a factores o características del objeto asegurado y de quien lo asegura. De ese modo, dichos modelos estadísticos predictivos sirven como base para establecer una prima suficiente, y distinta para cada tipología de cliente y cada contrato.

Uno de los grandes debates en el sector de los seguros actualmente surge a raíz del impulso que el big data ha ejercido en la personalización de las primas. La creciente disponibilidad de información permite que el número esperado y la cuantía esperada de los siniestros pueda ajustarse a un elevado número de características de riesgo particulares de quienes suscriben las pólizas, un número de factores a tener en cuenta que es muy superior al conjunto que se utilizaba décadas atrás. De ese modo, se ha visto incrementada la capacidad de diseñar sistemas de tarificación muy granulares que tienen en cuenta cada vez más información individual. La capacidad predictiva de los modelos y su adaptación a entornos con datos masivos choca entonces con el principio de mutualización. Y es en este punto donde emerge la inquietud de saber cuáles son los límites de la personalización de los precios, ya que si se pudiera llegar a predecir exactamente quién va a sufrir un accidente, y quién no, se acabaría estableciendo un precio para el primer grupo que sería igual al valor total de los accidentes que van a experimentar y un precio igual a cero para el segundo grupo, por lo que el propio concepto de la solidaridad en el seguro desaparecería. No existe un consenso sobre los límites de la ultra-segmentación de las primas, pero sí medidas que permiten detener un proceso

de individualización que conduzca a niveles de desigualdad excesiva de prima entre el colectivo de los asegurados[2].

Sin embargo, el big data abre una nueva perspectiva en el uso de los datos en los seguros y esa no es otra que la *prevención,* es decir, la predicción del riesgo anticipando la ocurrencia del siniestro y relegando el mero cálculo del precio a un segundo plano. A ello, ha contribuido muy notablemente la disponibilidad de información prácticamente en tiempo real.

2.1. Notación general y con datos telemáticos

Introducimos aquí la notación que va a utilizarse en el resto del capítulo. Se supone periodicidad anual en el contrato de seguro. Sea n el número de asegurados, sean N_i y S_{ij} respectivamente, el número de siniestros del asegurado i-èsimo, y la cuantía del j-ésimo siniestro del asegurado i, $i = 1,...,n, j = 1,...,N_i$ definida esta última únicamente si $N_i > 0$. Sean $X_{1i},...,X_{ki}$ las k características observables que suelen determinarse a partir de la formalización del contrato. Por ejemplo, la edad del asegurado, su antigüedad en la compañía y las características del objeto asegurado, como su superfície y localización en el caso de una vivienda, o marca, modelo, potencia y zona de conducción en un vehículo. Denominaremos $Z^*_{1i}(t),...,Z^*_{mi}(t)$, al conjunto de m variables telemáticas que están asociadas al objeto asegurado, es decir, que pueden medirse una vez ya está vigente el contrato e informan en tiempo real sobre su uso durante un periodo T y que se actualizan en intervalos de tiempo t. Denotaremos por $Z_{1i},...,Z_{si}$ a las s características anuales que resumen las observaciones telemáticas para cada asegurado i. Por ejemplo, en el caso del automóvil, con las tecnologías actuales puede medirse el total de kilómetros recorridos, número de trayectos realizados, la velocidad media de cada trayecto, frenazos, aceleraciones u otras medidas sin necesidad de tener localización exacta del vehículo. Cómo utilizar esta información telemática es uno de los objetivos de los modelos predictivos del riesgo en un entorno de datos masivos.

La información telemática permite conocer con detalle la exposición al riesgo, es decir, el intervalo de tiempo en el que realmente el asegurado puede tener un accidente que corresponde al momento en el que se encuentra conduciendo[3]. Hay casos en los que la exposición al riesgo es permanente, por ejemplo en los seguros de salud, pero en el seguro del automóvil a más kilómetros recorridos, mayor es la exposición y por lo tanto

---

[2] Hay factores que actualmente no pueden utilizarse para la determinación de precios. Por ejemplo, en la Unión Europea, como en un número creciente de países en el mundo, el principio de no discriminación impide que el sexo del asegurado pueda utilizarse como elemento diferencial en las tarifas, aunque sí puede servir internamente para analizar el riesgo que asume una entidad de seguros.

[3] Hay que tener en cuenta que los automóviles pueden sufrir percances aunque no están funcionando, por ejemplo estando aparcados pueden recibir un golpe de un tercero. Los siniestros de robo son un claro ejemplo también de exposición al riesgo con el vehículo parado. En el ejemplo concreto de recibir un golpe de un tercero, el siniestro estaría cubierto por la responsabilidad civil del culpable del golpe.

a igualdad de condiciones, quienes recorren más kilómetros tienen una probabilidad de sufrir accidentes sensiblemente superior a quienes recorren menos. El total de kilómetros recorridos en un año es una de las principales características telemáticas disponibles, pongamos Z_{1i}, que suele denotarse por D_i, y que puede utilizarse en el seguro de automóviles como una aproximación de la exposición al riesgo y, además, como elemento esencial en el pago por kilómetro.

2.2. Modelizar la frecuencia

El modelo de Poisson es el modelo básico para predecir el número esperado de siniestros y puede especificarse como:

$$E(N_i \mid X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) = exp\,(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \gamma_1 Z_{1i} + ... + \gamma_s Z_{si}), \qquad [1]$$

siendo $\theta = (\beta_0,\beta_1,...,\beta_k,\gamma_1,...,\gamma_s)$ el vector de parámetros a estimar. Se suele usar la notación matricial $X_i^{a'}\theta$ para el predictor lineal, siendo $X_i^{a'} = \left(X_i^{'}, Z_i^{'}\right)$ el conjunto de regresores distinguiendo entre los que provienen del contrato $X_i^{'}$ y los que provienen de la telemetría $Z_i^{'}$. Se supone que N_i sigue una distribución de Poisson de parámetro $exp\left(X_i^{a'}\theta\right)$.

Cuando se utiliza una variable de exposición al riesgo, también denominada *offset*, D_i, para el i-ésimo individuo, el modelo se expresa como:

$$\begin{aligned} &lE(N_i \mid D_i, X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) \\ &= D_i exp\left(\beta_0^o + \beta_1^o X_{1i} + ... + \beta_k^o X_{ki} + \gamma_1^o Z_{1i} + ... + \gamma_s^o Z_{si}\right) \\ &= exp\left(\log\left(D_i\right) + \beta_0^o + \beta_1^o X_{1i} + ... + \beta_k^o X_{ki} + \gamma_1^o Z_{1i} + ... + \gamma_s^o Z_{si}\right). \end{aligned} \qquad [2]$$

siendo $\theta^o = \left(\beta_0^o,\beta_1^o,...,\beta_k^o,\gamma_1^o,...,\gamma_s^o\right)$ el vector de parámetros a estimar.

Los parámetros se estiman por máxima verosimilitud y, como parte de los modelos lineales generalizados, se utilizan el conjunto de herramientas de inferencia de esta familia de modelos. De todos modos, como en muchos casos suele haber sobredispersión en los datos o un exceso de ceros, lo que se aconseja es usar algunas extensiones del modelo básico de Poisson como el modelo binomial negativo, que aquí omitimos.

2.3. Modelizar la cuantía

Para modelizar la cuantía de los siniestros se puede especificar un modelo para el coste, siendo cero si no ha habido ningún siniestro[4]. Para modelizar la cuantía, se puede utilizar un modelo Gamma donde la variable es estrictamente positiva. Así el modelo puede especificarse como:

[4] Si el asegurado ha sufrido más de un siniestro, se puede modelizar la media de los costes de los siniestros que ha sufrido cada asegurado.

$$E(S_i) = exp\ (\alpha_0 + \alpha_1 X_{1i} + ... + \alpha_l X_{ki} + \alpha_{l+1} Z_{1i} + ... + \alpha_{l+s} Z_{si}) \quad [3]$$

siendo S_i el coste medio de los siniestros del asegurado i, en caso de que este haya tenido algún siniestro, y 0 en caso contrario, $\alpha_0,..., \alpha_{l+s}$ el vector de parámetros a estimar en el modelo para las cuantías en las que se han incluido k características no-telemáticas del i-ésimo individuo y otras s de carácter telemático. Para simplificar, se han supuesto los mismo factores predictivos en los dos modelos: frecuencia y cuantía, pero el número y tipo de factores podría cambiar. El método de estimación es máxima verosimilitud. En el caso de la severidad, es difícil obtener resultados que permitan distinguir el impacto de los factores predictivos sobre la media, por lo que en la práctica no es extraño tomar solamente una constante y trabajar directamente con el importe medio de los siniestros.

3. PREDECIR CUANTILES

Dado que la frecuencia de siniestralidad es generalmente muy baja, la predicción del riesgo puede centrarse en algunos indicadores telemáticos, variables aleatorias continuas, que se sabe que están positivamente asociados a una mayor siniestralidad, por ejemplo, los excesos de velocidad o la conducción nocturna, entre otros, mediante la regresión cuantílica.

La regresión cuantílica es un modelo que especifica una relación entre los cuantiles de la variable respuesta R_i y un conjunto de covariables para el i-ésimo individuo. Si consideramos una especificación en la que distinguimos entre variables telemáticas y aquellas que no lo son, el modelo se expresa como:

$$Q_\tau(R_i \mid X_{1i},...,X_{ki},Z_{1i},...,Z_{si}) = \beta_0^\tau + \beta_1^\tau X_{1i} + ... + \beta_k^\tau X_{ki} + \gamma_1^\tau Z_{1i} + ... + \gamma_s^\tau Z_{si}, \quad [4]$$

siendo τ el nivel del cuantil a estimar (por ejemplo, 90 %, 95 %, 99 %), $Q_\tau(R_i \mid X_i^a) = X_i^a \theta^\tau$ el cuantil condicional a ajustar para la variable R_i, y $\theta^\tau = \left(\beta_0^\tau, \beta_1^\tau, ..., \beta_k^\tau, \gamma_1^\tau, ..., \gamma_s^\tau\right)$ el vector de parámetros a estimar. Se puede introducir una transformación en el predictor lineal de forma análoga a como se realiza en la media. Sin embargo, el procedimiento de estimación del modelo de regresión cuantílica generalizado es más complejo y no está disponible actualmente en los paquetes de estimación habituales.

4. OTROS MODELOS PREDICTIVOS DEL RIESGO

Recientemente, se han planteado nuevos modelos predictivos del riesgo, más allá de la regresión cuantílica, que amplían los modelos existentes en dos direcciones: a) para especificaciones no lineales del modelo de riesgo, como se acaba de mencionar en el párrafo anterior, y b) para otras medidas de riesgo como la esperanza condicionada de la cola. En este último caso, se establece un modelo conjunto en el que se estima simultáneamente un modelo para el cuantil y un modelo para la esperanza de los valores más

allá del cuantil. En el ámbito actuarial y de las ingenierías se supone que R_i es una variable aleatoria no negativa y, por lo tanto, el riesgo se encuentra en la parte derecha de la distribución, ya que se modelizan pérdidas y estas se suponen positivas. Sin embargo, en el ámbito financiero, cuando la variable respuesta son los rendimientos, que pueden ser positivos o negativos, entonces el riesgo se encuentra en los valores negativos, por lo que el interés reside en la cola izquierda de la distribución. Este cambio de signo, entre ambos entornos, ha propiciado un cierto distanciamiento entre las investigaciones y el uso de notaciones diferentes, que suelen confundir.

Los modelos conjuntos de medidas de riesgo surgen de la literatura financiera y, por lo tanto, en lugar de cuantiles se habla de valor en riesgo y en lugar de esperanza condicionada de la cola *(tail conditional expectation),* se utiliza el término *expected shortfall regression.* Para simplificar la notación, diremos que $Q_\tau(R_i \mid X_i^a)$ denota el cuantil condicional y que $E(R_i \mid R_i \geq Q_\tau(R_i \mid X_i^a), X_i^a) = CTE_\tau(R_i \mid X_i^a)$, siendo el modelo conjunto de regresión cuantílica y de regresión de esperanza condicional de la cola:

$$Q_\tau(R_i \mid X_i^a) = \beta_0^\tau + \beta_1^\tau X_{1i} + \ldots + \beta_k^\tau X_{ki} + \gamma_1^\tau Z_{1i} + \ldots + \gamma_s^\tau Z_{si}, \quad [5]$$

$$CTE_\tau(R_i \mid X_i^a) = \beta_0^{c\tau} + \beta_1^{c\tau} X_{1i} + \ldots + \beta_k^{c\tau} X_{ki} + \gamma_1^{c\tau} Z_{1i} + \ldots + \gamma_s^{c\tau} Z_{si}. \quad [6]$$

La estimación de los parámetros puede realizarse con un estimador de momentos o bien optimizando la función de pérdida correspondiente al modelo conjunto para ambas medidas de riesgo.

Los modelos anteriores permiten identificar factores que elevan el riesgo de ciertas variables respuesta, y, en particular, tienen interés las que están asociadas a una mayor accidentabilidad, como son el total de kilómetros recorridos y los excesos de velocidad en el caso del seguro de automóviles. En la siguiente sección se ilustran algunos casos de uso de los modelos anteriores, viendo el impacto de la inclusión de la información telemática.

5. EL SEGURO DE AUTOMÓVIL: PAGO POR KILÓMETRO

Una de las aplicaciones más reciente de la telemática es la utilización de sensores en los vehículos para monitorizar la conducción. No escapa a nadie que parte de este campo tiene como último objetivo lograr el transporte autónomo, en el que no sea necesaria la intervención de un humano al volante. Aunque se han logrado ciertos avances, y sobre todo en elementos de ayuda a la conducción como el control de velocidad, la distancia al vehículo precedente, el aparcamiento automático o el sensor de elementos alrededor del vehículo, no se vislumbra cuándo se podrá eliminar completamente el conductor. Sin embargo, la telemetría sí es una realidad y se está integrando de tal forma en los automóviles que el seguro no es ajeno a la disponibilidad de información que ello permite utilizar.

El pago por kilómetro es una de las formas de aseguramiento que más ha dado que hablar en los últimos años y se conoce bajo el acrónimo en ingles *PAYD, pay as you drive*[5]. Actualmente es posible comprar este tipo de producto en muchos países del mundo, incluida la mayoría de países de la Unión Europea. Dado que se aplica al seguro del automóvil, que es un seguro obligatorio, todo vehículo tiene una cuota mínima que corresponde a una prima básica, cuyo importe se incrementa de forma proporcional a la distancia recorrida por el vehículo.

Esta sección muestra un caso real de una cartera de vehículos en España que disponían de un dispositivo telemático de registro de datos de conducción instalado en los automóviles. Mostraremos tres aproximaciones al análisis del riesgo. En primer lugar, el tradicional, en el que la frecuencia de siniestralidad se explica por factores clásicos como la edad o la potencia del vehículo. Se muestra cómo la inclusión de indicadores telemáticos permite mejorar las predicciones incluso en una situación básica en la que no se utiliza la exposición al riesgo y el modelo para la frecuencia es el modelo de Poisson más simple. Al incluir la distancia total recorrida como *offset* en el modelo se aprecia la importancia del efecto de exposición al riesgo y se puede interpretar el resultado obtenido como una forma sencilla de obtener un coste por kilómetro adaptado a las características de riesgo. En un segundo paso, se presenta un modelo de regresión cuantílica y un modelo de regresión para la esperanza condicionada de la cola que permite detectar factores asociados a elevados valores de la exposición al riesgo. Este modelo permite detectar qué características influyen en exposiciones extremas al riesgo, percentil 90 %, lo que permite identificar segmentos de mayor peligrosidad que el resto. En un tercer caso, se muestra el comportamiento en curvas de referencia que comparan los kilómetros recorridos por encima de los límites de velocidad permitidos y los kilómetros recorridos para algunos asegurados que han sufrido un siniestro. Dicho análisis permite detectar pautas concretas que podrían asociarse a un incremento del riesgo.

5.1. Número de siniestros por kilómetro recorrido

En este apartado se muestra una aplicación empírica basada en una muestra de 11.937 asegurados que tienen una póliza PAYD en vigor durante todo el año 2018. Las variables de que se dispone se muestran en la tabla 1. Entre ellas se encuentran las variables clásicas utilizadas en tarificación (edad, antigüedad del carnet, antigüedad y potencia del vehículo y número de siniestros) así como variables telemáticas (kilometraje total recorrido durante el año, porcentaje de kilómetros recorridos en horario nocturno, por vías urbanas y por encima de los límites de velocidad).

En la tabla 2 se muestran los correspondientes estadísticos descriptivos. La edad media de los asegurados es 31,26 años, con 11,09 años de antigüedad media de carnet. La

[5] El PAYD también es conocido como *usage-based insurance (UBI)* cuando incorpora indicadores de uso del vehículo además de la distancia recorrida.

TABLA 1.

DEFINICIÓN DE LAS VARIABLES EN LOS DATOS SOBRE SEGUROS DE AUTOMÓVILES, 2018

Variable	*Descripción*
edad	Edad del asegurado a fecha 01.01.2018
carnet	Antigüedad del carnet de conducir (en años) a fecha 01.01.2018
antigveh	Antigüedad del vehículo (en años) a fecha 01.01.2018
potencia	Potencia del vehículo asegurado (en cc)
km_totales_mil	Distancia total (en miles de kilómetros) recorridos durante todo 2018
noctur	Porcentaje de kilómetros conducidos durante 2018 por la noche
velocidad	Porcentaje de kilómetros conducidos durante 2018 por encima del límite de velocidad
urban	Porcentaje de kilómetros conducidos durante 2018 por vías urbanas
nsin	Número de siniestros por responsabilidad civil con culpa durante 2018

Fuente: Elaboración propia.

antigüedad media del vehículo es de 10,85 años, y la potencia es de 98,33 en promedio. A lo largo de 2018 los asegurados de la muestra han conducido en promedio 8.661 kilómetros. La Figura 1 muestra el histograma del kilometraje total, que mide la exposición al riesgo, y que como puede apreciarse tiene una marcada asimetría positiva. Se tiene además que un 16.04 % de los kilómetros totales recorridos durante 2018 se han circulado por la noche, un 26.57 % en vías urbanas y un 0.31 % por encima del límite de velocidad permitida. En promedio han tenido 0.053 siniestros de responsabilidad civil con culpa durante 2018.

TABLA 2.

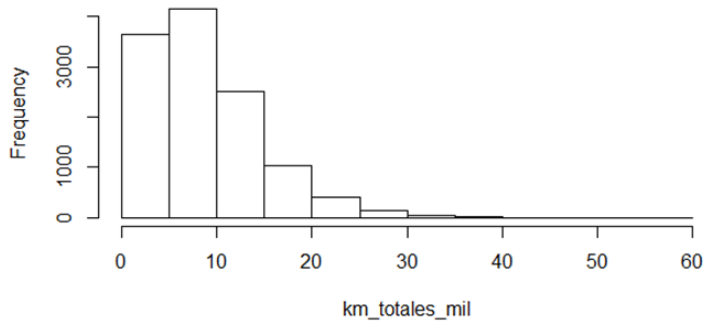
ESTADÍSTICOS DESCRIPTIVOS DE LAS VARIABLES EN LOS DATOS SOBRE SEGUROS DE AUTOMÓVILES, 2018

Variable	*Media*	*Desv. est.*	*Mínimo*	*Máximo*
edad	31,256	4,809	17,000	82,000
carnet	11,095	4,531	0,000	53,000
antigveh	10,859	4,751	0,085	46,118
potencia	98,334	27,554	23,000	418,000
km_totales_mil	8,662	5,983	0,001	56,639
noctur	16,037	16,862	0,000	100,000
velocidad	0,310	0,390	0,000	1,940
urban	26,572	16,497	0,000	100,000
nsin	0.053	0.235	0,000	2,000

Fuente: Elaboración propia.

FIGURA 1

HISTOGRAMA DEL KILOMETRAJE TOTAL DE 2018 EN MILES



Fuente: Elaboración propia.

En la tabla 3 se muestran los resultados de la estimación de los tres modelos de Poisson propuestos para el número de siniestros. El primero de ellos (modelo 1) incluye como variables explicativas únicamente las variables clásicas utilizadas en tarificación. Se ha excluido la edad, dado que está muy correlacionada con la antigüedad del carnet y esta última tiene un efecto más claro a la hora de explicar el número de siniestros. El segundo de ellos (modelo 2) incluye además de las variables clásicas, las variables telemáticas. Finalmente, el modelo 3 incluye las mismas variables explicativas que el modelo 2, pero además añade como offset el logaritmo del kilometraje anual en miles.

Respecto al modelo 1, se observa que la única variable con efecto significativo a la hora de explicar la siniestralidad es la antigüedad del carnet. En concreto, a más antigüedad, el número esperado de siniestros disminuye. Respecto al modelo 2, se observa (además del efecto significativo de la antigüedad del carnet) que los porcentajes de conducción nocturna y por vía urbana tienen un efecto significativo a la hora de explicar la siniestralidad. En concreto, conducir por la noche o por vía urbana se asocia a un mayor número de siniestros. Al añadir el logaritmo del kilometraje total (en miles) como offset, se producen algunos cambios en el efecto del resto de variables explicativas. La antigüedad del carnet deja de tener efecto significativo, en cambio sí lo tiene la antigüedad del vehículo, de manera que cuanto más antiguo es el vehículo mayor es el número esperado de siniestros. Por otra parte, respecto a las variables telemáticas, únicamente la conducción urbana tiene un efecto significativo a la hora de explicar la siniestralidad, de manera que, a más conducción urbana, mayor es el número de siniestros. Por lo que respecta a la bondad de ajuste de los modelos, se concluye que la inclusión de las variables telemáticas mejora el ajuste, dada la reducción que se produce en el valor del

AIC (criterio de información de Akaike) al pasar del modelo 1 al 2. Se observa además que la inclusión del logaritmo del kilometraje total como offset del modelo mejora el ajuste (el AIC vuelve a reducirse al pasar el modelo 2 al 3).

TABLA 3.

RESULTADOS DE LA ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN DE POISSON PARA SINIESTROS CON CULPA SOBRE SEGUROS DE AUTOMÓVILES, MUESTRA DE 2018 (N=11,937)

Parámetro	*Modelo 1*	*Modelo 2*	*Modelo 3 (con* offset*)*
constante	-2,895***	-3,205***	-6,071***
sexo	0,110	0,079	0,024
carnet	-0,026**	-0,023*	-0,019
antigveh	-0,001	-0,003	0,024**
potencia	0,0020	0,002	0,002
noctur	–	0,007**	0,004
velocidad	–	0,071	-0,036
urban	–	0.006*	0,028***
ln(km_totales_mil)	–	–	1,000
AIC	5.019,715	5.011,130	4.996,762

Notas: *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05.
Fuente: Elaboración propia.

De los anteriores modelos deducimos que la inclusión de la información telemática mejora el ajuste del modelo.

5.2. Modelos de predicción de kilómetros e indicadores mediante cuantiles

En este apartado, ajustamos un modelo gamma para predecir la media del kilometraje anual en función de una serie de variables explicativas. Así mismo, ajustamos también un modelo de regresión cuantílica para los diferentes percentiles del kilometraje anual, en particular los percentiles 5 %, 25 %, 50 %, 75 % y 95 %. Los resultados se muestran en la tabla 5.

Por lo que respecta al modelo gamma que estima la media del kilometraje anual, todas las variables explicativas tienen efecto significativo. En concreto, los hombres realizan más kilómetros que las mujeres, por lo que están más expuestos al riesgo de sufrir un accidente. La potencia del vehículo, la conducción nocturna y por encima de los límites de velocidad se asocian también a un mayor kilometraje anual. Por otro lado, a medida que aumenta la conducción por vía urbana, así como la antigüedad del carnet de conducir y del vehículo, se reduce el kilometraje anual.

TABLA 4.

RESULTADOS DE LA ESTIMACIÓN DE LOS MODELOS DE REGRESIÓN GAMMA Y CUANTÍLICA A NIVELES 5 %, 25 %, 50 %, 75 % Y 95 % PARA EL TOTAL DE KILÓMETROS RECORRIDOS, MUESTRA DE 2018 (N=11,937)

		Percentil				
	Media	*5 %*	*25 %*	*50 %*	*75 %*	*95 %*
constante	9,717***	3205,329***	9380,727***	12977,480***	16678,573***	22657,529***
sexo	0,040***	56,727	169,423	233,063*	591,666***	1265,352***
carnet	-0,003*	-15,701*	-28,747*	-21,388	-19,840	-16,449
antigveh	-0,030***	-87,374***	-210,693***	-233,622***	-235,848***	-209,408***
potencia	0,001***	-0,104	1,555	3,977*	7,456**	15,803*
urban	-0,022***	-26,975***	-91,363***	-131,313***	-171,819***	-231,402***
noctur	0,005***	-5,213***	23,124***	39,614***	54,663***	66,698***
velocidad	0,094***	884,493***	971,182***	841,673***	733,010***	911,299**
AIC	234.724	238.928	236.640	236.947	240.428	250.092

Notas: *** p-value < 0.001, ** p-value < 0.01, * p-value < 0.05.

Fuente: Elaboración propia.

Respecto a los modelos de regresión cuantílica, se observa que el sexo únicamente resulta significativo para la estimación de percentiles elevados, y su efecto siempre es el mismo: ser hombre se asocia a valores más elevados del kilometraje o exposición al riesgo. Por otro lado, la antigüedad del carnet tiene efecto significativo y negativo para percentiles bajos, en concreto, a más antigüedad del carnet menores son los percentiles 5 % y 25 % del kilometraje total. La antigüedad del vehículo tiene efecto significativo y negativo para cualquier percentil, por tanto, se asocia a una reducción del kilometraje. La potencia del vehículo solo resulta tener efecto significativo para la estimación de percentiles elevados de la exposición al riesgo, en concreto a mayor potencia mayores son los percentiles 50 %, 75 % y 95 % del kilometraje. La conducción por vía urbana siempre se asocia con menores valores de los percentiles del kilometraje total, mientras que la conducción nocturna y por encima de los límites de velocidad se asocian con un mayor nivel de exposición al riesgo (excepto en el caso de la conducción nocturna para el percentil 5 %, que tiene efecto contrario).

El análisis realizado permite detectar qué características influyen en exposiciones extremas al riesgo (percentil 95 %). En concreto, se identifica a los hombres con vehículos nuevos y potentes, que circulan poco por vías urbanas, pero además con mayor proporción en horario nocturno y por encima del límite de velocidad, como el segmento más expuesto al riesgo y por lo tanto el de mayor peligrosidad.

5.3. Curvas de referencia, tiempo/kilómetros

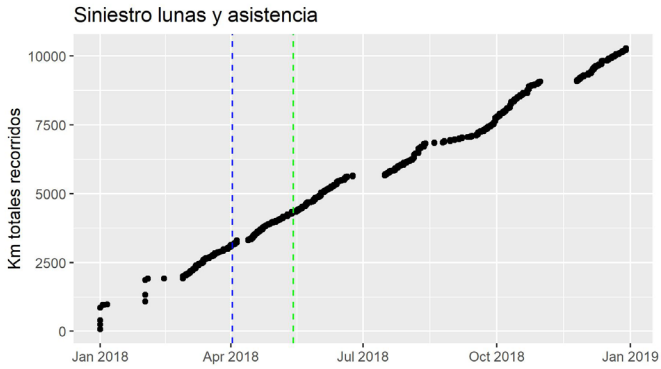
En este apartado se muestran algunas curvas de referencia para conductores de la muestra. Dichos diagramas permiten realizar un *tracking* de los conductores. En la figura 2

se muestra el total de kilómetros acumulados a lo largo del año 2018 y en la figura 3, se ve la acumulación de kilómetros recorridos con excesos de velocidad, cuyo patrón cambia notablemente en la segunda mitad del año.

El análisis longitudinal de los datos telemáticos permite establecer indicadores de riesgo sobre cambios inesperados de comportamientos y anomalías, pudiéndose utilizar como herramientas de prevención ante actitudes al volante que se asocian a mayor accidentabilidad.

FIGURA 2

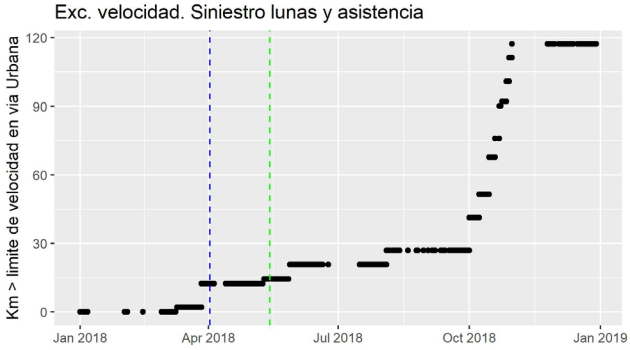
CONDUCTOR CON DOS SINIESTROS, TOTAL DE KILÓMETROS RECORRIDOS ACUMULADOS A LO LARGO DE 2018. LOS PERIODOS DE NO UTILIZACIÓN DEL VEHÍCULO SE ENCUENTRAN EN BLANCO



Fuente: Elaboración propia.

FIGURA 3

CONDUCTOR CON DOS SINIESTROS, TOTAL DE KILÓMETROS RECORRIDOS ACUMULADOS CON EXCESO DE VELOCIDAD EN VÍA URBANA A LO LARGO DE 2018. LOS PERIODOS DE NO UTILIZACIÓN DEL VEHÍCULO SE ENCUENTRAN EN BLANCO



Fuente: Elaboración propia.

6. OTRAS APLICACIONES DEL BIG DATA EN LOS SEGUROS

Para finalizar, realizamos un breve resumen de aportaciones en el ámbito del big data y los seguros, iniciando nuestro recorrido en un trabajo de Bologa, Bologa y Florea (2013) donde el análisis se centra en la detección del fraude. Efectivamente, la detección de comportamientos sospechosos ha sido una de las mayores preocupaciones de los aseguradores desde el inicio del siglo XXI, por cuanto supone un incremento de los costes. La efectividad de los métodos para identificar relaciones en redes dió un enorme impulso a la disciplina porque se demostró capaz de detectar núcleos de generación de fraude. Por ejemplo, talleres que de acuerdo con los asegurados incrementaban los costes de reparación. En el trabajo mencionado, se daban casos relacionados con el uso y recobro de asistencia sanitaria que nunca se utilizó, los autores concluyeron que la mejor estrategia era la combinación de la investigación de siniestros tradicional y los modelos predictivos.

Algunas contribuciones como Zhang (2017) no mejoran especialmente la capacidad predictiva de los modelos pero sí la ganancia en tiempo con la paralelización de algoritmos en el tipo de bases de datos de siniestros en el que se tienen varios millones de pólizas. En otros casos se propone la combinación de métodos como Lin *et al.* (2017) para el análisis de la siniestralidad o para la proyección de los beneficios futuros de un cliente de una entidad de seguros (Fang, Jiang Song, 2016).

Porrini (2017) plantea los grandes retos de los datos masivos que manejan las aseguradores, desde la perspectiva del marco regulatorio europeo. En su trabajo expone los principales elementos de preocupación: privacidad de los asegurados (los datos pueden contener información sensible), discriminación (los factores de tarificación pueden favorecer a determinados tipos de ciudadanos y perjudicar a otros) e impedimentos a la competencia (un mal uso de los datos puede dar ventajas a una empresa frente a otra). Arumugam y Bhargavi (2019) abordan el uso de datos masivos en seguros con un sistema de cálculo de primas a nivel teórico, pero a diferencia de Zhang *et al.* (2018) que sí establecen una correlación entre un indicador de riesgo y la ocurrencia de accidentes, los primeros no llegan a mostrar la implementación.

La pregunta más inquietante la lanzan Barry y Charpentier (2020) al cuestionar si el big data va a cambiar el sector asegurador. Hasta el momento, parece que los viejos modelos persisten y tener más datos significa solamente un incremento de los indicadores de riesgo más que un cambio de enfoque que revolucione el seguro. Meyers y Hoyweghen (2020) relatan con detalle un reciente experimento realizado en Bélgica para hallar evidencias de la relación entre la mejora de la conducción y la disminución de la siniestralidad mediante el uso de un gran volumen información sobre los asegurados. El estudio relata algunas dificultades como la mala calidad de recogida de datos a través del Smartphone, la comunicación con los asegurados (qué indicadores habría que darles) y finalmente, cómo aconsejarles, cuándo y cómo. Dichos autores alientan a

superar las dificultades de tarificar en base al comportamiento, es decir a los patrones de conducción observados en los asegurados.

7. CONCLUSIONES

Nos encontramos en una nueva era tanto por lo que respecta al fundamento de los seguros como a la reorientación de su finalidad.

La utilización de elementos telemáticos permite un mayor conocimiento del riesgo y los modelos de predicción de riesgos son buena muestra de la capacidad que tal información proporciona en la mejora de la modelización.

Los modelos de predicción del riesgo tienen una orientación diferente a los tradicionales y permiten detectar factores asociados a indicadores de mayor peligrosidad. En este sentido, el empleo de la regresión cuantílica frente a la metodología de modelos lineales generalizados presenta algunas ventajas por cuanto que permite conocer las causas del comportamiento de la variable respuesta en escenarios extremos, como los de mayor siniestralidad. Esto puede ayudar a la adopción de medidas preventivas que reduzcan la siniestralidad.

Quedan por resolver numerosos aspectos de la modelización del riesgo, como la elección del nivel de tolerancia o la propia medida de riesgo, dado que es bien sabido que dicha elección es crucial para la posterior interpretación de resultados.

Para finalizar, debemos añadir que en el entorno de los datos masivos, preocupan tres aspectos: los algoritmos de aprendizaje y su aceleración, la estabilidad estructural de los resultados, es decir, cada cuanto tiempo es válido un modelo que se alimenta constantemente de datos y, finalmente, la depuración de la información o el uso de indicadores sintéticos por cuanto deben ser capaces de recoger la esencia de aquello que están midiendo.

Referencias

Arumugam, S. y Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data,* 6(1), pp. 86.

Barry, L. y Charpentier, A. (2020). Personalization as a promise: Can Big Data change the practice of insurance?. *Big Data & Society,* 7(1). https://doi.org/10.1177/2053951720935143

Bologa, A. R., Bologa, R. y Florea, A. (2013). Big data and specific analysis methods for insurance fraud detection. *Database Systems Journal,* 4(4), pp. 30-39.

Fang, K., Jiang, Y. y Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. *Computers & Industrial Engineering,* 101, pp. 554-564.

Lin, W., Wu, Z., Lin, L., Wen, A. y Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee Access,* 5, pp. 16568-16575.

Meyers, G. y Hoyweghen, I. V. (2020). 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data & Society,* 7(1). https://doi.org/10.1177/2053951720914650

Porrini, D. (2017). Regulating Big Data effects in the European insurance market. *Insurance Markets and Companies,* 8, pp. 6-15.

Zhang, Y. (2017). Bayesian analysis of big data in insurance predictive modeling using distributed computing. *ASTIN Bulletin,* 47(3), pp. 943-961.

Zhang, H., Xu, L., Cheng, X., Chen, W. y Zhao, X. (2017). Big data research on driving behavior model and auto insurance pricing factors based on UBI. En: *International Conference On Signal And Information Processing, Networking And Computers* (pp. 404-411). Singapore: Springer.