Rect@ Vol 10 Diciembre 2009. Pp 267 - 284

Recibido

10/10/2009

Revisado

2/11/2009

Aceptado

11/11/2009

Diseño de perfiles de asegurados en el ramo del automóvil

Segovia-González, M.M. mmseggon@upo.es
Guerrero. F.M. fguecas@upo.es
Herranz, P..: pherpei@upo.es
Departamento de Economía, Métodos Cuantitativos e Historia Económica.
Universidad Pablo de Olavide.
Edificio Blanco White num. 3.
Ctra. de Utrera km. 1. 41013 Sevilla

RESUMEN

En los últimos años las compañías de seguros cada vez son más competitivas. Por ello es muy importante que tengan un buen sistema de clasificación de los individuos según su comportamiento frente a la siniestralidad. Realizaremos un trabajo empírico con los datos de una compañía aseguradora española. Se dará una estimación de la función del riesgo de ocurrencia de un siniestro en función de la edad del conductor, teniendo en cuenta otras variables relevantes. Además, obtendremos los tramos de edades que mejor o peor se comportan frente al riesgo. Para ello utilizaremos la técnica del análisis de componentes principales funcional y una estimación de la función de correlación. Con los resultados obtenidos por dicha técnica funcional se llevara a cabo un análisis cluster para detectar grupos homogéneos en el comportamiento de los siniestros. Esto permitiría a las compañías aseguradoras establecer una tarificación para cada uno de los grupos.

.

Palabras claves: R*amo del seguro, siniestros, análisis multivariante*

ABSTRACT

In recent years, insurance companies are increasingly competitive. It is therefore very important that they have a good system of classification of individuals according to their behavior in accident ratio. We will work with empirical data from a Spanish insurance company. It will give an estimate of the role of the risk of occurrence of an incident based on driver age, taking into account other relevant variables. In addition, we get the age ranges that better or worse they behave toward risk. To do this we use the technique of functional principal component analysis and an estimate of the

correlation function. With the results obtained by this functional technique we will conduct a cluster analysis to identify homogeneous groups in the behavior of claims. This would allow insurance companies to establish a tariff for each of the groups.

Keywords: insurance, claim, multivariate analysis

1.- Introducción

En el ramo del seguro de automóvil operan un gran número de aseguradoras que ofrecen productos muy similares. Existen trabajos empíricos con datos de automovilistas franceses (Chiappori and Salanié, 2000) y de Quebec (Dionne *et al.*, 2001) que muestran que no hay signos de selección adversa en estos mercados si se tiene una segmentación del riesgo lo más cercana posible a la realidad. Por otra parte, hay trabajos que ponen de relieve cómo las características socioeconómicas, las características del vehículo y de la póliza contratada también influyen en este proceso de decisión. Además, se muestra que cuando el asegurado está satisfecho con el trato recibido y con el servicio que se le ha dado, posee una mayor predisposición a renovar su póliza con su actual aseguradora (Pujol and Bolancé, 2004). Las compañías aseguradoras cada vez estén más interesadas en establecer métodos eficientes para el control de la evolución de los accidentes y sus costes asociados (Ayuso y Guillén, 1999).

Las tarifas del ramo de automóviles están estructuradas en función de una serie de variables. En este sentido destacamos la aplicación de métodos multivariantes y cálculo actuarial en la selección de riesgos en la tarificación (Boj *et al.*, 2004 y Guillén *et al.*, 2005). Asimismo, existen estudios que describen los comportamientos de fraude teniendo en cuenta las características de los asegurados y de las pólizas (Artis *et al.*, 1999, Ayuso *et al.*, 1999, Artis *et al.*, 2002).

Nos interesa conocer qué características son a priori más influyentes a la hora de tener una mayor o menor siniestralidad. Para ello, podemos ver las recomendaciones de la Unión Española de Entidades Aseguradoras y Reaseguradoras (UNESPA), así como diversos estudios (Melgar y Guerrero, 2005; Boj *et al.*, 2004 y Guillén *et al.* 2005), que nos indican una serie de características que son importantes. Dichas características son: la gama del vehículo, la zona de circulación, el uso del vehículo y las circunstancias personales tanto del conductor como del asegurado, la edad, el sexo, los años de posesión del carné de conducir, etc. Entre otras, es muy importante conocer las circunstancias personales del asegurado: sexo, edad, años de carnet, estado civil... Todas estas características son tenidas en cuenta en las diferentes modalidades de pólizas del automóvil. En cuanto a la variable edad, se suele distinguir entre jóvenes y no jóvenes dependiendo de los hábitos de vida. Sin embargo, no se especifican diferentes tarifas por tramos de edad a lo largo de la vida del asegurado. Esto se podría tener en cuenta a través del Análisis en Componentes Principales Funcional (ACPF). De esta forma, además de considerar las características que influyen en este fenómeno, la edad del conductor se utiliza de manera más precisa (Segovia-Gonzalez, *et al.*, 2009). En este trabajo pretendemos agrupar a aquellos individuos que se comportan de forma homogénea frente a la siniestralidad, haciendo uso de los resultados del ACPF y del análisis cluster. Para el desarrollo de esta idea hemos estructurado el trabajo de la siguiente forma.

En la sección 2 describimos la muestra con la que vamos a trabajar, estableciendo una serie de perfiles. Asimismo mostraremos las estimaciones de las funciones de riesgo con las que trabajaremos y posteriormente en la sección 3 los resultados del ACPF. Haciendo uso de la interpretación de las componentes obtenidas y de las

puntuaciones de cada una de ellas obtenidas en un trabajo anterior, se lleva a cabo en la sección 4 un análisis cluster. Se darán una serie de grupos de asegurados que se comportan de forma similar en cuanto al riesgo de ocurrencia de un siniestro. Para ello haremos uso de los resultados obtenidos por el ACPF y del análisis cluster. Esta información será de gran interés para las compañías de seguros a la hora de tarificar. Por último, presentamos las conclusiones más relevantes y la bibliografía utilizada.

2.- Descripción de la muestra

La información cedida por la entidad aseguradora se corresponde al período de 2001 a 2003. Se dispone de 271.800 asegurados y de 88.337 siniestros. Nos restringiremos a los individuos asegurados en dicho período que en el caso de haber tenido algún siniestro la culpa sea directa o compartida, conduzcan un vehículo de tipo turismo y cuyo uso sea particular. No consideramos a los taxis, camiones, motos, etc. puesto que estos vehículos tienen un riesgo de siniestro diferente. Hacemos todo esto para que el estudio no englobe a poblaciones muy heterogéneas respecto al comportamiento de la siniestralidad. Una vez aplicadas las restricciones anteriores y depurados los datos nos encontramos con 175.191 asegurados y 30.483 siniestros que serán el objeto de nuestro estudio.

2.1.- Estratificación de la muestra

Consideramos una serie de características que son importantes para el estudio de la siniestralidad. Dichas variables son: el tipo de turismo, el sexo del conductor habitual y la zona geográfica. En cuanto a las variables zona geográfica y tipo de turismo agrupamos las zonas centro y norte, así como los turismos de gama media y alta, con objeto de obtener un tamaño muestral suficiente para poder realizar inferencias. Teniendo en cuenta esta estratificación y después de realizar una exhaustiva depuración de los datos obtenemos una serie de perfiles que se muestran en la Tabla 1.

En definitiva, tenemos 12 perfiles distintos y nuestro principal propósito es ver cómo se comportan los siniestros en las distintas edades por las que va pasando el conductor.

Por tanto, estudiaremos el *riesgo de ocurrencia de un siniestro* en cada uno de los perfiles y en cada una de las edades. Dicho riesgo lo definimos como el cociente entre el número de siniestros ocurridos en un perfil determinado a una edad determinada y el número total de asegurados que tienen dicho perfil y dicha edad. Los periodos de edades que tenemos van de los 18 a los 88 años; no obstante, al disponer de muy pocos asegurados menores de 25 y mayores de 71 años, hemos decidido estudiar de forma agrupada ambos grupos de individuos (que se han denotado con el valor 25^- y 71^+). En estos tramos tendremos únicamente el comportamiento global. Luego, en nuestro estudio, la variable principal será el riesgo de ocurrencia de un siniestro del perfil i-ésimo en la edad t, con $i = 1, ..., 12$ y $t = 25^-, 25, 26, \ldots, 70, 71, 71^+$. En estudios posteriores, pretendemos disponer de la información de un número mayor de asegu-

Perfiles	**Sexo**	**Gama del turismo**	**Zona geográfica**
1	Mujer	Media-Alta	Sur
2	Mujer	Media-Alta	Centro-Norte
3	Mujer	Media-Alta	Mediterránea
4	Mujer	Baja	Sur
5	Mujer	Baja	Centro-Norte
6	Mujer	Baja	Mediterránea
7	Hombre	Media-Alta	Sur
8	Hombre	Media-Alta	Centro-Norte
9	Hombre	Media-Alta	Medierránea
10	Hombre	Baja	Sur
11	Hombre	Baja	Centro-Norte
12	Hombre	Baja	Mediterránea

Tabla 1: Perfiles considerados en el estudio de la siniestralidad.

rados, a pesar de ser esto muy complicado de conseguir, ya que la mayoría de las compañías aseguradoras se muestran reticentes a dar información de su cartera de clientes. Para obtener las estimaciones de las funciones con las que vamos a trabajar, aplicaremos el criterio de mínimos cuadrados. Es decir, lo que haremos será minimizar para el perfil i-ésimo el valor de:

$$MC = \sum_{j=t_0}^{t_1} \left[y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_{ik}(t_j) \right]^2$$

y tenemos, por tanto, que:

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_{ik}(t), \text{ para todo } i = 1, \ldots, 12.$$

El procedimiento empleado ha sido el método de las funciones base, en este caso se utilizarán funciones base del tipo B-splines (Segovia-Gonzalez *et al.*, 2009). Mostramos los datos originales y las curvas que aproximan a algunos de los perfiles con los que estamos realizando el estudio (Figura 1). Dichas curvas tienen las características que hemos descrito anteriormente (Tabla 1). Asimismo, presentamos el error cometido en cada una de ellas aplicando el criterio de mínimos cuadrados (MC).

3.- Cálculo de las componentes principales funcionales regularizadas

Una vez que tenemos las estimaciones de las 12 curvas nos disponemos a llevar a cabo el ACPF regularizado. Para ello se utiliza el algoritmo desarrollado por

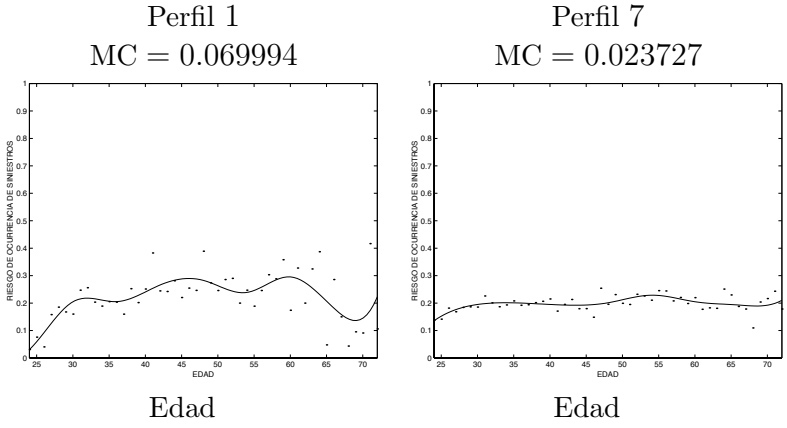

Figura 1: Riesgo de ocurrencia de un siniestro.

Ramsay (2003). Tanto el marco teórico utilizado como las ventajas de utilizar la técnica funcional en vez del análisis de componentes principales clásico se puede ver en Segovia-Gonzalez *et al.* (2009). En ambos análisis, por la forma en que se plantea el problema de optimización, se sabe que la primera componente principal será la que acumule la mayor variabilidad del proceso original; seguidamente, estará la segunda componente principal, la tercera y así sucesivamente.

Lo que obtendremos son estimaciones, teniendo que la h-ésima componente principal vendrá dada por:

$$\hat{\xi_h} = \int_{t_0}^{t_1} \hat{X}(t)\hat{f}_h(t)dt, \;\; h = 1, 2, 3, 4 \dots ,$$

donde $\hat{X}(t)$ y $\hat{f}_h$ son estimaciones de las curvas del proceso original y de la autofunción h-ésima, respectivamente. Además, para cada perfil podremos tener la puntuación en cada una de las componentes. Se tiene, para la componente h-ésima, que las puntuaciones en cada uno de los individuos considerados serán $\hat{\xi}_h = (\hat{\xi}_{1h}, \hat{\xi}_{2h}, \dots , \hat{\xi}_{12h})$ con:

$$\hat{\xi}_{ih} = \int_{t_0}^{t_1} \hat{x_i}(t)\hat{f}_h(t)dt, \;\; \text{para todo } i = 1, \dots , 12,$$

siendo $\hat{f}_h(t)$ y $\hat{x_i}(t)$ las estimaciones de la autofunción h-ésima y de la función correspondiente al perfil i-ésimo, respectivamente. A continuación, mostraremos información acerca de las primeras 4 autofunciones (Figura 2) y de las puntuaciones correspondientes en cada una de las 4 primeras componentes principales para los distintos perfiles (Tabla 2).

La elección del número óptimo de componentes principales se trata de forma más detallada en la sección siguiente.

Perfiles	**1ª Componente**	**2ª Componente**	**3ª Componente**	**4ª Componente**
1	0.454173	−0.123593	0.027439	0.008693
2	0.222051	0.169771	−0.127397	−0.032286
3	0.003438	−0.122665	−0.146860	0.038278
4	0.131653	−0.011524	0.101664	−0.018921
5	−0.004001	−0.109032	−0.013333	−0.098486
6	−0.231211	0.072026	−0.025004	−0.080370
7	0.167503	0.107711	0.098156	0.026214
8	0.006047	0.073371	0.022335	0.042549
9	−0.094135	0.016874	−0.053302	0.089978
10	−0.110139	−0.001048	0.041606	−0.015850
11	−0.256171	−0.043447	0.041342	0.026575
12	−0.289207	−0.028445	0.033354	0.013626

Tabla 2: Puntuaciones de los distintos perfiles en cada una de las componentes.

3.1.- Elección del número de componentes principales

Cuando realizamos un ACPF, una de las cuestiones que nos surgen es la elección del número óptimo de componentes. Se pretende poder explicar el proceso en estudio con un número de componentes no muy elevado. En la práctica, rara vez los investigadores utilizan un único criterio para determinar cuántas componentes extraer. Un ajuste demasiado alto o demasiado bajo hará que la estructura de los datos no se plasme de forma clara. Los criterios que utilizaremos serán el criterio de contraste de caída y el criterio de porcentaje de la varianza (Hair, 2000). Para aplicar el criterio de contraste de caída representaremos el gráfico de sedimentación (véase Figura 3).

En el eje de ordenadas de la Figura 3 hemos representado el valor de los autovalores obtenidos al aplicar el ACPF con regularización. Podemos observar que, a partir de la cuarta componente, la pendiente disminuye bastante. Luego, una elección lógica sería quedarnos con las cuatro primeras componentes principales.

Asimismo, con la información de los autovalores podremos, en cada periodo, conocer qué porcentaje de variabilidad del proceso viene explicado por cada una de las componentes. En nuestro caso, la primera componente explica el 70.1919%; la segunda, el 12.8608%; la tercera, el 8.9408% y la cuarta, el 4.0766%. Con las tres y cuatro primeras componentes explicamos el 91.9935% y el 96.0701% de la variabilidad total, respectivamente. Por tanto, otra forma de elegir con cuántas componentes nos quedamos sería elegir el conjunto de ellas que expliquen un número razonable de porcentaje de variabilidad del proceso original. Tomar tres o cuatro componentes podría ser razonable, pues tendríamos explicado el 92% o el 96%, respectivamente.

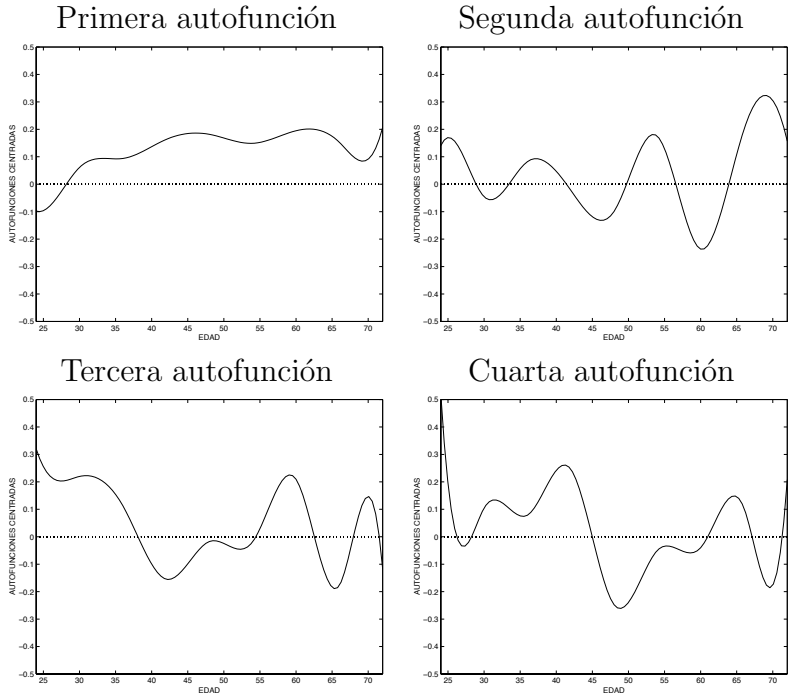

Figura 2: Cuatro primeras autofunciones centradas.

3.2.- Bondad del ajuste

A continuación, estimaremos el error que cometemos al reconstruir el proceso original por medio de las componentes principales obtenidas. El error cuadrático medio viene dado por:

$$ECM^{(m)}(t) = \frac{1}{12}\sum_{i=1}^{12}[\hat{x_i}(t) - \hat{x_i}^{(m)}(t)]^2,$$

donde la estimación del proceso reconstruido, utilizando m componentes, se calcula como:

$$\hat{x_i}^{(m)}(t) = \sum_{h=1}^{m} \hat{f}_h(t)\hat{\xi}_{ih},$$

siendo $\hat{f}_h(t)$ la autofunción h-ésima estimada y $\hat{\xi}_{ih}$ el valor que la observación i-ésima toma en la componente h-ésima. En la Figura 4, se representan las funciones del error cuadrático medio según hayamos utilizado en la reconstrucción una, dos, tres o cuatro componentes principales. Como los errores cometidos son muy pequeños, resulta difícil observar en las gráficas las mejoras obtenidas al utilizar un mayor número de componentes. Por ello, mostraremos los valores que toman dichas funciones en una serie de periodos. Lo presentamos en las Tablas 4 y 5 del anexo 1, para las edades $25^-, 25, \ldots, 71, 71^+$, aunque lo podríamos presentar para cualquier edad $t \in [t_0, t_1]$. Se observa que el error disminuye conforme se incrementa el número de componentes utilizadas.

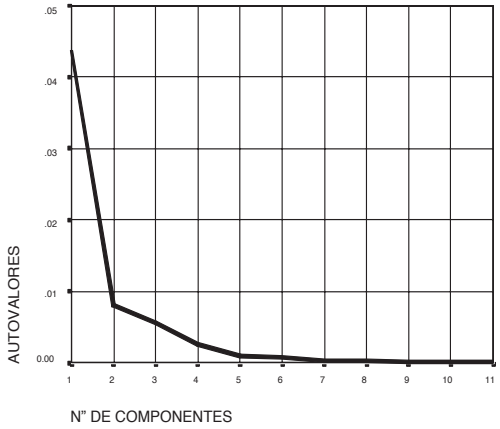

Figura 3: Gráfico de sedimentación.

3.1.- Interpretación de las componentes principales funcionales

La estimación de la función correlación entre el proceso y la componente h-ésima, $(h = 1, 2, 3, 4)$ vendrá dada por:

$$\hat{r}_h(t) = r(\hat{X}(t), \hat{\xi}_h) = \frac{\sqrt{\hat{\rho}_h}}{\hat{\sigma}(t)} \hat{f}_h(t),$$

siendo $\hat{X}(t)$ la estimación de las curvas obtenidas para los distintos perfiles, $\hat{\rho}_h$ y $\hat{f}_h(t)$ el autovalor y la autofunción estimada para la componente h-ésima, respectivamente, y $\hat{\sigma}(t)$ denota la estimación de la desviación típica del proceso original. De forma general, consideramos que existe una fuerte correlación si ésta es superior o igual a 0.7, en valor absoluto (Hair, 2000). En la Figura 5 representamos la función de correlación existente entre el proceso estimado y las primeras cuatro componentes.

Se puede observar que la primera componente está muy correlacionada positivamente con el tramo que va de los 31 a los 67 años, ambos inclusive, y con los mayores de 71 años. La información que obtenemos para los mayores de setenta y un año no es muy fiable, pues en nuestro estudio los tuvimos que agrupar. Por tanto, no tendremos en cuenta la significación estadística detectada para la primera componente en dicho tramo de edad. La segunda y tercera componente están muy correlacionadas positivamente con el tramo de edad que va de los 68 a los 70 años y de los 27 a los 30 años, respectivamente. Y por último, la cuarta componente no está fuertemente correlacionada con ningún tramo de edad. Por ello, el estudio lo realizaremos considerando las tres primeras componentes principales, ya que así explicamos el 91.99% de la variabilidad total del proceso.

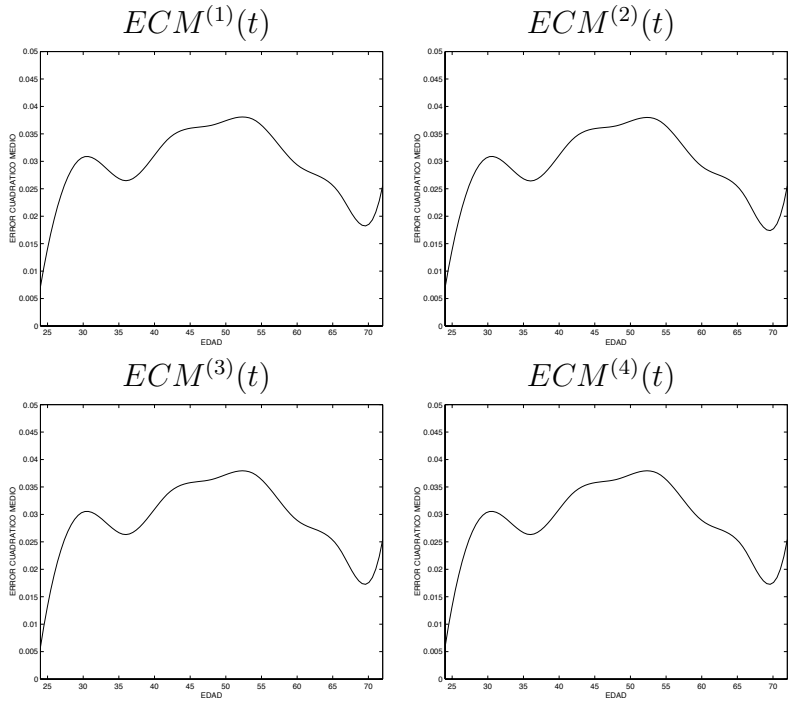

Figura 4: Funciones del error cuadrático medio.

4.- Interpretación de los resultados

4.1.- Comportamiento de los distintos perfiles

Utilizando la información obtenida por la función de correlación y la puntuación de cada perfil en cada una de las componentes principales, podremos ver en cada uno de los tramos de edad detectados el comportamiento de cada uno de los perfiles.

En la Figura 6 se muestra una representación de dichos valores para el caso de las tres primeras componentes.

Con respecto a la primera de ellas podremos conocer cuándo en el tramo de edad de los 31 a los 67 años el riesgo de ocurrencia de un siniestro es superior o inferior en los distintos perfiles considerados. No obstante, si la correlación obtenida en el apartado anterior hubiera sido negativa, los perfiles que tomaran puntuaciones más altas se comportarían de forma que el riesgo de ocurrencia de un siniestro en ese tramo fuese inferior con respecto a los perfiles que en esa variable tomen valores más pequeños.

Según las características de cada uno de estos perfiles, se puede observar que los hombres y los individuos con coche de gama baja, en general, toman una puntuación más pequeña en esta componente. En concreto, las dos puntuaciones más pequeñas se corresponden con hombres que conducen coches de gama baja y su lugar de residencia es la zona mediterránea o la zona centro-norte. Por el contrario, los dos valores más altos en dicha variable se corresponde con mujeres que conducen un

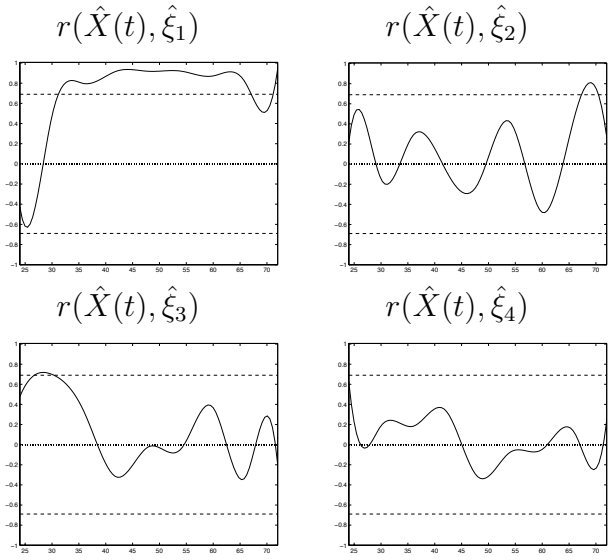

Figura 5: Funciones de correlación.

coche de gama media-alta y residen en la zona sur o en la zona centro-norte.

Si nos centramos en la segunda componente principal, obtuvimos que está muy correlacionda positivamente con el tramo de edad que va de los 68 a los 70 años, ambos inclusive. Utilizando las puntuaciones de cada perfil en esta componente podremos establecer un orden para cada uno de los perfiles en función del comportamiento de los siniestros en este tramo de edad. Los individuos que mejor se comportan son las mujeres con coche de gama media-alta y residentes en el sur y el mediterráneo

Si pasamos a estudiar la tercera componente principal, obtuvimos que dicha componente está directamente correlacionada con el tramo de edad que va de los 27 a los 30 años, ambos inclusive. En función de las puntuaciones que toman en la tercera componente, de menor a mayor, los perfiles aparecen en el siguiente orden $P_3, P_2, P_9, P_6, P_5, P_8, P_1, P_{12}, P_{11}, P_{10}, P_7$ y P_4. En dicho tramo de edad, la zona con más riesgo de siniestros es la zona sur, siendo los hombres que conducen coches de gama baja los más propensos a que tengan algún siniestro en dicho rango de edad.

Hasta ahora, hemos evaluado los perfiles uno a uno según la información de las distintas componentes. Al tener doce perfiles distintos y haber detectado 3 tramos de edades distintos, podemos establecer 36 supuestos distintos en la tarificación. El establecer tantos grupos distintos puede suponer a la compañías un gran problema. Por un lado, la gestión de una gama más amplia de tarifas se hace más compleja, y por otra parte, al obtener una información más particionada de las pólizas por la variedad de perfiles, el estudio comparativo con los resultados de años anteriores es difícilmente abordable. Estos motivos hicieron que la aseguradora que nos suministró los datos nos solicitara la reducción del número de perfiles de asegurados.

Por ello nos planteamos a continuación utilizar de forma conjunta toda la información obtenida con las tres componentes principales.

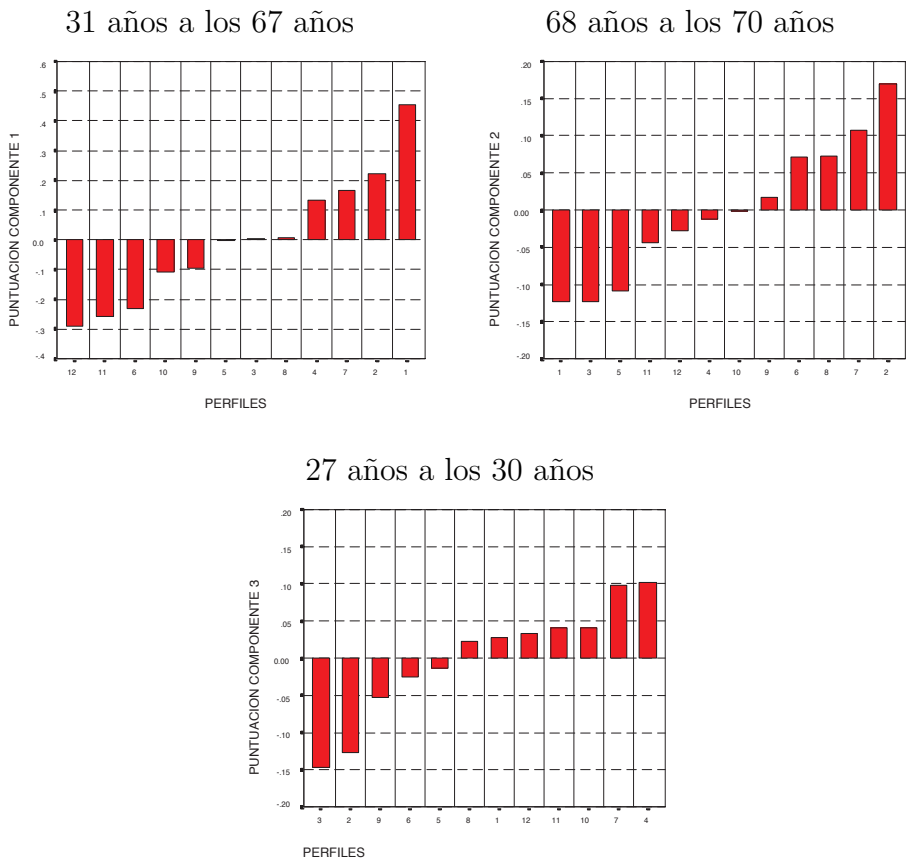

Figura 6: Puntuaciones de las tres primeras componentes en los distintos perfiles.

4.2.- Agrupación homogénea de los perfiles

El Análisis Cluster (AC) es una técnica multivariante cuyo objetivo principal es la comparación de los objetos basándose en las variables especificadas por el investigador y no en la estimación de ellas. Sokal y Sneath (1963) son dos de los autores que más han influido en el desarrollo de esta técnica. En nuestro estudio, pretendemos agrupar a aquellos individuos que se comportan de forma homogénea o similar según las puntuaciones que toman en las tres primeras componentes principales que hemos estimado en la sección anterior. Nos restringimos a las tres primeras, pues con ellas tendremos explicado el 92% de la variabilidad del proceso original. Además, como se ha visto con anterioridad la cuarta componente no está fuertemente correlacionada con ningún tramo de edad, con lo cual podremos prescindir de ella en este análisis.

Vamos a aplicar un método jerárquico aglomerativo denominado *método del promedio entre grupos*. Dicho método utiliza información de todas las distancias entre pares de individuos, y no solamente de los más alejados o de los más próximos (Johnson, 2000; Hair, 2000). En la Figura 7 mostramos el dendograma obtenido al llevar a cabo el análisis de los distintos individuos utilizando las tres primeras componentes del caso funcional.

Mostramos en la Figura 8 una representación gráfica de las puntuaciones que

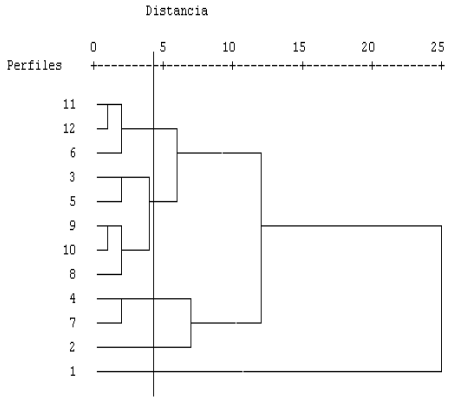

Figura 7: Dendograma.

toman en las tres primeras componetes los distintos perfiles.

El primer grupo (G1) es el formado por el perfil P_1, es decir, las mujeres del sur cuya gama de coche es media-alta. En el segundo grupo (G2) encontramos al perfil P_2, que se corresponde con las mujeres con coche de gama media-alta de la zona centro-norte. En el tercer grupo (G3) localizamos a los perfiles P_3, P_5, P_8, P_9 y P_{10}. Se agrupan, por tanto, las mujeres que conducen un coche de gama media-alta y pertenecen a la zona mediterránea, las que tienen un coche de gama baja y residen en la zona centro-norte, así como los hombres de la zona centro-norte y mediterránea que conducen un coche de gama media-alta y los del sur que tienen un coche de gama baja. En el cuarto grupo (G4) se encuentran los perfiles P_4 y P_7, es decir, las mujeres con coche de gama baja y los hombres con coche de gama media-alta, todos ellos del sur. En el quinto grupo (G5) se tienen los perfiles P_6, P_{11} y P_{12}, con lo que se agrupa a los hombres y mujeres del mediterráneo con coche de gama baja y los hombres del centro-norte también con coche de gama baja.

Debido a que los grupos primero, segundo y cuarto están formados por muy pocos elementos, en concreto por uno en los dos primeros grupos y por dos individuos en el cuarto, vamos a obviar estos grupos al realizar los contrastes.

Estudiamos si existen diferencias significativas en cuanto a las puntuaciones medias de la primera componente en los grupos tres y cinco. Aplicamos el test de la t de Student con la corrección de Welch. Como el p-valor obtenido es menor a 0.05, concluimos que existen diferencias en cuanto a las puntuaciones medias de la primera componente en ambos grupos. El intervalo de confianza al 95% de la diferencia de puntuaciones medias entre ambos grupos viene dado como $I_{\mu_3-\mu_5} = (0.1440, 0.2942)$, por lo que la puntuación media en el grupo quinto es inferior a la del grupo tres. Esto quiere decir que el grupo quinto descrito anteriormente verifica que en el tramo de edad que va desde los 31 a los 67 años tienen un menor riesgo de tener un siniestro que los individuos del grupo tres. En cuanto a las puntuaciones obtenidas en la

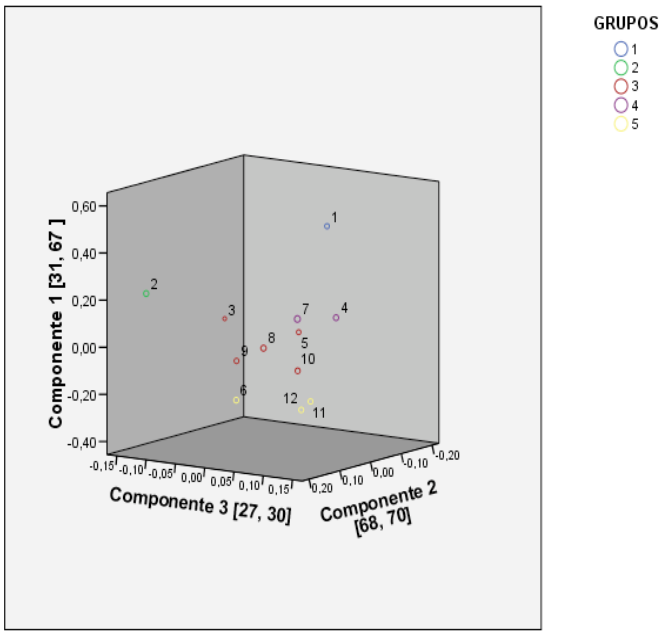

Figura 8: Grupos del análisis cluster.

segunda componente principal tenemos que tras verificar que se cumplen todas las hipótesis de aplicación del test de la t de Student concluimos que no existen diferencias significativas, ya que el p-valor obtenido es 0.634. Es decir, en ambos grupos no se puede afirmar que sus medias se comporten de forma distinta en el tramo de edad de los 68 a los 70 años. De forma análoga, cuando utilizamos las puntuaciones obtenidas en la tercera componente obtenemos que no existen diferencias significativas, ya que se obtiene un p-valor igual a 0.362.

Hemos creado la Tabla 3 utilizando los resultados del análisis cluster. Los grupos vendrán ordenados según el riesgo de siniestro en los distintos tramos de edad.

27 años a los 30 años	31 años a los 67 años	68 años a los 70 años
G2	G5	G1
G5 y G3	G3	G5 y G3
G1	G4	G4
G4	G2	G2
	G1	

Tabla 3: Comportamiento de los distintos grupos.

Se plantea la posibilidad de que la compañía establezca 13 grupos de individuos que tienen un comportamiento distinto según el riesgo de ocurrencia de un siniestro. Asimismo, se muestran ordenados de menor a mayor riesgo para cada uno de los rangos de edad detectados con el ACPF.

5.- Conclusión

Este trabajo se centra en el seguro del automóvil y tiene como punto de partida los resultados del análisis realizado a los datos cedidos por una compañía aseguradora en otros trabajos anteriores. Con la información obtenida al aplicar el ACPF y la técnica multivariante del análisis por conglomerados, hemos ordenado a diferentes grupos de conductores según el riesgo de ocurrencia de un siniestro en distintos tramos de edad. Estos tramos de edad los hemos detectado a su vez al hacer uso de la función de correlación entre el proceso estudiado y cada una de las componentes principales funcionales. De esta forma hemos pasado de tener 36 grupos distintos de tarificación (12 perfiles y 3 tramos de edades) a tener 13 grupos distintos. El reducir los grupos de conductores que se comportan de forma distinta es de suma importancia para las compañías. Sobretodo para aquellas que trabajan con un volumen pequeño o mediano de clientes. Ya que, un incremento amplio del número de tarifas en este tipo de compañías puede complicar su gestión e impedir el análisis comparativo de sus resultados respecto a los de años anteriores. En este trabajo lo que damos es un orden entre los grupos según se comporten mejor o peor en cuanto a la siniestralidad teniendo en cuenta múltiples variables y trabajando con las funciones de riesgo a lo largo de la edad del conductor, en vez de con vectores p dimensionales. Una línea de continuidad para trabajos futuros será la cuantificación de dichos riesgos.

6.- Referencias

- Artis M., Ayuso M., Guillén M. (1999). Modeling different types of automobile insurance fraud behaviour in Spanish market. *Mathematics and Economics*, **824**, 1-2, pp. 67-81.

- Artis M., Ayuso M., Guillén, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, **69**, 3, pp. 325 − 340.

- Ayuso M., Guillén M. (1999). Modelos de detección de fraude en el seguro del automóvil. *Cuadernos actuariales*, **8**, pp. 135 − 149.

- Ayuso M., Guillén M., Artis M. (1999). Técnicas cuantitativas para la detección del fraude en el seguro del automóvil. *Anales del instituto de actuarios españoles*, **5**, pp. 51 − 83.

- Boj E., Claramunt M.M., Fortiana J. (2004). Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación. *Cuadernos de la Fundación Mapfre estudios*. Instituto de Ciencias del Seguro. Madrid.

- Chiappori P.A., Salanié B. (2000). Testing for Asymmetric Information in Insurance Markets. *Journal of political Economics*, **108**, 1, pp. 56 − 78.

- Dionne G.C., Gourièroux, Vanasse C. (2001). Testing for Evidence od Adverse Selection in the Automobile Insurance Market: A Comment. *Journal os political Economy*, **109**, 2, pp. $444 - 453$.

- Guillén M., Ayuso M., Bermúdez L., Morillo I. (2005). El Seguro de automóviles: estado actual y perspectiva de la técnica actuarial. *Fundación Mapfre estudios*. Instituto de Ciencias del Seguro. Madrid.

- Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (2000). *Análisis Multivariante*. Prentice Hall.

- Johnson, D.E. (2000). *Métodos multivariados aplicados al análisis de datos*. International Thomson Editores, S.A.

- Melgar M.C., Guerrero, F.M. (2005). Los siniestros en el seguro del automóvil: un análisis econométrico aplicado. *Estudios de Economía Aplicada (Asepelt)*, pp. $355 - 375$.

- Pujol M., Bolancé C. (2004). *La matriz valor fidelidad en el análisis de los asegurados en el ramo del automóvil*. Fundación Mapfre estudios. Instituto de Ciencias del Seguro. Madrid.

- Ramsay, J.O. (2003). *R and S-PLUS Functions for functional data analysis*. McGill University.

- Sokal R.R., Sneath P.H.A. (1963). *Principles of numerical taxonomy*. W.H. Freeman and Co.

- Segovia-Gonzalez M.M., Guerrero F.M., Herranz, P. (2009). Explaining functional principal component analysis to actuarial science with an example on vehicle insurance. *Insurance: Mathematics and Economics*, **45**, 2, pp. $278 - 285$.

- UNESPA. $http://www.unespa.es/Nociones\_seguro/7.cfm$.

Anexo

Edad	$\mathbf{ECM^{(1)}}$	$\mathbf{ECM^{(2)}}$	$\mathbf{ECM^{(3)}}$	$\mathbf{ECM^{(4)}}$
25^-	0,0071420	0,0069320	0,0063500	0,0058100
25	0,0141210	0,0139300	0,0135680	0,0133470
26	0,0198240	0,0196680	0,0194180	0,0193810
27	0,0242950	0,0241820	0,0240090	0,0239600
28	0,0275790	0,0275110	0,0273300	0,0272780
29	0,0297200	0,0296910	0,0294410	0,0294210
30	0,0307630	0,0307610	0,0304640	0,0304410
31	0,0308060	0,0307970	0,0305120	0,0304640
32	0,0301060	0,0301000	0,0298440	0,0297850
33	0,0289960	0,0289890	0,0287770	0,0287190
34	0,0278070	0,0277830	0,0276280	0,0275770
35	0,0268580	0,0268180	0,0267160	0,0266740
36	0,0264710	0,0264200	0,0263580	0,0263210
37	0,0268790	0,0268260	0,0267820	0,0267420
38	0,0279430	0,0278940	0,0278510	0,0278010
39	0,0294330	0,0293930	0,0293390	0,0292760
40	0,0311210	0,0310890	0,0310190	0,0309400
41	0,0327770	0,0327500	0,0326630	0,0325690
42	0,0341690	0,0341420	0,0340450	0,0339390
43	0,0351280	0,0350910	0,0349950	0,0348820
44	0,0357090	0,0356560	0,0355720	0,0354560
45	0,0360280	0,0359570	0,0358900	0,0357750
46	0,0362000	0,0361110	0,0360650	0,0359540
47	0,0363410	0,0362380	0,0362130	0,0361090

Tabla 4: Error cuadrático medio en diferentes edades.

Edad	$\mathbf{ECM^{(1)}}$	$\mathbf{ECM^{(2)}}$	$\mathbf{ECM^{(3)}}$	$\mathbf{ECM^{(4)}}$
48	0,0365650	0,0364550	0,0364480	0,0363520
49	0,0369480	0,0368460	0,0368410	0,0367590
50	0,0374060	0,0373200	0,0373080	0,0372410
51	0,0378170	0,0377400	0,0377280	0,0376710
52	0,0380580	0,0379800	0,0379740	0,0379210
53	0,0380060	0,0379190	0,0379100	0,0378620
54	0,0375370	0,0374340	0,0374040	0,0373660
55	0,0365770	0,0364420	0,0363820	0,0363560
56	0,0352400	0,0350630	0,0349710	0,0349540
57	0,0336880	0,0334680	0,0333430	0,0333350
58	0,0320830	0,0318270	0,0316740	0,0316720
59	0,0305860	0,0303110	0,0301400	0,0301380
60	0,0293590	0,0290910	0,0289110	0,0289100
61	0,0285070	0,0282730	0,0280990	0,0280960
62	0,0278960	0,0277090	0,0275540	0,0275420
63	0,0273360	0,0271880	0,0270580	0,0270370
64	0,0266370	0,0264990	0,0263940	0,0263630
65	0,0256090	0,0254280	0,0253450	0,0253070
66	0,0240590	0,0237650	0,0236960	0,0236520
67	0,0219490	0,0214650	0,0213960	0,0213530
68	0,0198370	0,0191490	0,0190750	0,0190340
69	0,0184350	0,0176090	0,0175280	0,0174870
70	0,0184520	0,0176320	0,0175500	0,0175070
71	0,0205980	0,0200080	0,0199390	0,0198860
71^{+}	0,0255830	0,0255260	0,0254880	0,0254170

Tabla 5: Error cuadrático medio en diferentes edades.