

# **Análisis del comprador de seguros de vida en el periodo 2008-2012: búsqueda y análisis de segmentos de interés a través de técnicas de data mining**

**Primer Borrador de 4 de noviembre de 2013**

**Lorenzo Escot  
José Andrés Fernandez Cornejo  
Javier Portela**



**UNIVERSIDAD  
COMPLUTENSE  
MADRID**

## **Análisis del comprador de seguros de vida en el periodo 2008-2012: búsqueda y análisis de segmentos de interés a través de técnicas de data mining**

1)	Introducción .....	3
2)	Variables utilizadas y sus categorías .....	4
3)	Estudio comparativo de la evolución temporal de las pólizas en los años de crisis, atendiendo a sus características de prima, duración y capital. Relación con segmentos poblacionales de edad y profesión. ....	5
4)	Estudio comparativo de la evolución temporal de las pólizas en los años de crisis, atendiendo a sus características de edad y profesión. ....	16
4)	Estudio comparativo de la evolución temporal de las pólizas en los años de crisis, atendiendo a sus características de edad y profesión. ....	17
5)	Conclusiones y recomendaciones .....	21
	ANEXO .....	24

## 1) Introducción

La evolución del número de pólizas de vida en general en el periodo de crisis (2008-2012) sigue una pauta ascendente que alcanza un máximo en 2010 y es descendente a partir de entonces. Sin embargo, una curva agregada de totales puede esconder subsegmentos de la población para los cuales la evolución puede ser radicalmente diferente, incluso contradictoria (ascendente en lugar de descendente) con la curva general.

Con la información disponible se pretende en este informe estudiar cómo la forma de la curva de evolución en el periodo de crisis del número de pólizas de vida depende de ciertas características como son el tipo de póliza en cuanto a la terna prima, duración y capital, y características socioeconómicas como la edad y la profesión del asegurado.

Se realizarán dos análisis en cierto modo simétricos.

En el primero se construye una categorización de pólizas en cuanto a sus características de prima, duración y capital, y seguidamente se agrupan esas categorías mediante un análisis cluster en función de su evolución en el periodo 2008-2012. Se extrae información a continuación sobre las características de segmentos de edad y profesión de los asegurados, relacionadas con los clusters extraídos.

En el segundo análisis se contruye una categorización de los asegurados en cuanto a su edad y profesión y se agrupan esas categorías mediante un análisis cluster en función de su evolución en el periodo 2008-2012. Se extrae información a continuación sobre las características de las polizas suscritas (en función de prima, capital, duración) en los clusters extraídos.

En las conclusiones se comenta el interés y los límites de los resultados obtenidos y se discute la posibilidad de utilizar otro tipo de información complementaria.

## 2) Variables utilizadas y sus categorías

Tras un análisis previo de las diferentes variables incluidas en los ficheros suministrados por ICEA y de la capacidad de cada una de ellas para discriminar entre diferentes tipos de segmentos de comprador de seguros de vida durante el periodo 2008-2012, hemos centrado nuestro análisis en las siguientes variables

### Edad

Edad<=24	edad2=1
25=<Edad<=34	edad2=2
35=<Edad<=44	edad2=3
45=<Edad<=54	edad2=4
55=<Edad<=64	edad2=5
65=<Edad<=34	edad2=6

### Profesión

0	Desconocida
1	Autónomos y Comerciantes
2	Empresarios
3	Profesionales Liberales
4	Agricultores y Ganaderos
5	Directivos y Mandos Intermedios
6	Empleados con Titulación
7	Empleados sin Titulación
8	Amas de Casa
9	Jubilados y Pensionistas
10	Otra

### Prima, duración y capital

En todo el documento, las variables prima, duración y capital pasan a ser prima2, dura2 y cap2, categorizadas como sigue:

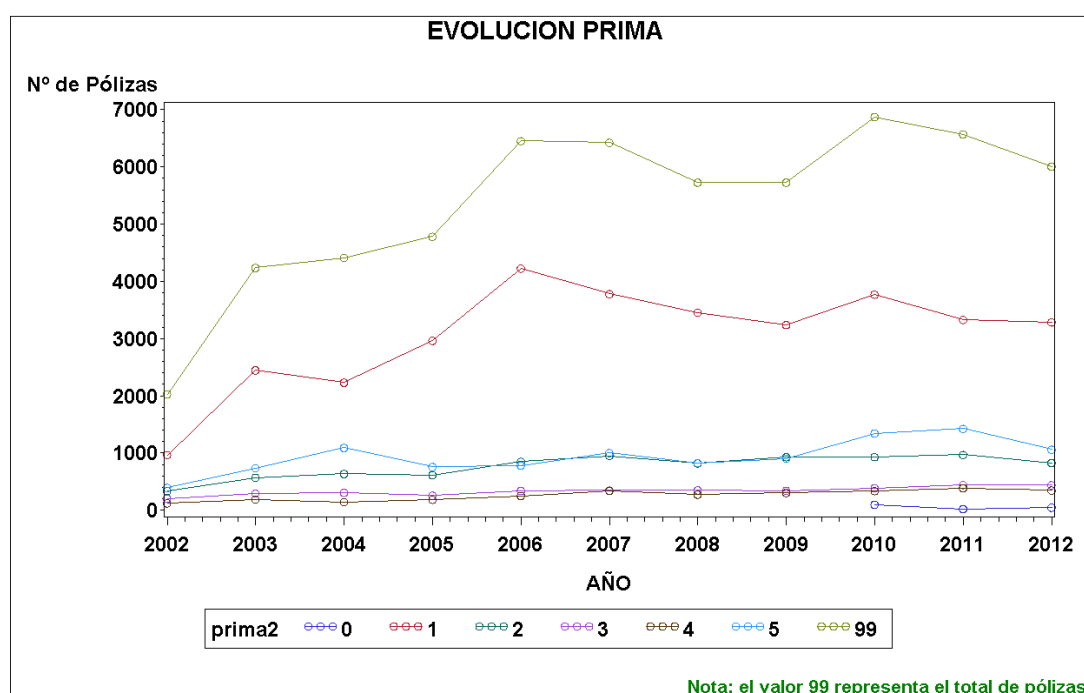
prima=0	prima2=0
0<prima<=300.5	prima2=1
300.51=<prima<=601	prima2=2
601.01=<prima<=901.5	prima2=3
901.51=<prima<=1202	prima2=4
1202.01=<prima	prima2=5
capital=0	cap2=0
0=<capital<=6000	cap2=1
6011=<capital<=60100	cap2=2
60101 =<capital	cap2=3
duracion=0	dura2=0
0<duracion<=4	dura2=1
4<duracion<=20	dura2=2
20<duracion	dura2=3

### 3) Estudio comparativo de la evolución temporal de las pólizas en los años de crisis, atendiendo a sus características de prima, duración y capital. Relación con segmentos poblacionales de edad y profesión.

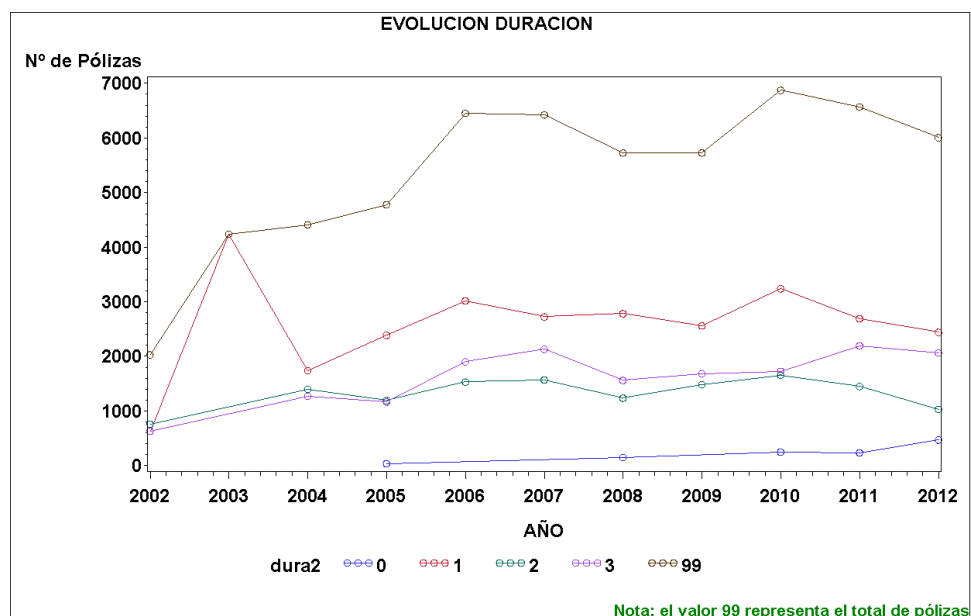
En primer lugar es necesario un análisis descriptivo para evaluar las técnicas más apropiadas para buscar resultados interesantes a través de técnicas de data mining.

En las gráficas del número de pólizas, se observa de 2008 a 2012, en el total de pólizas, una parte estable y ascendente hasta 2010, y un descenso hasta 2012 hasta niveles de 2009 (la curva que aparece en la parte superior, es la que representa el total de pólizas).

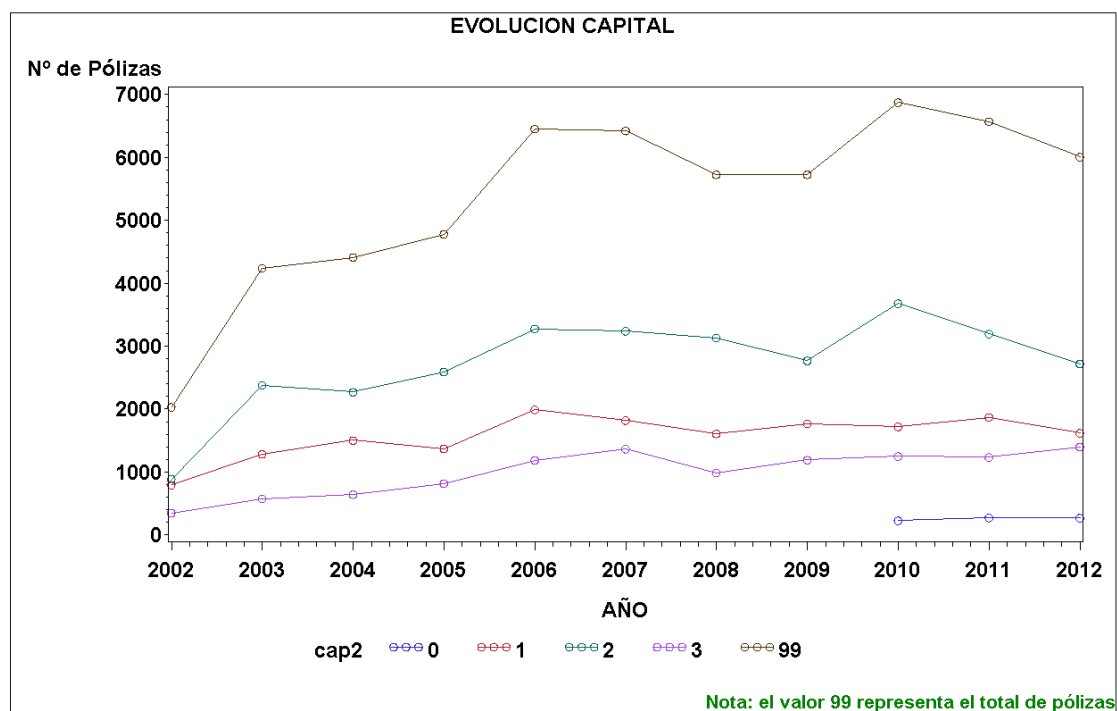
Segregando por tipo de prima, se observa que las primas pequeñas (prima2=1) son las más numerosas y marcan la pauta general de la evolución de primas. Otros tipos de prima parecen comportarse de manera estable. En la base de datos se ve que hay un cierto número de primas desconocidas de 2010 a 2012 (prima2=0).



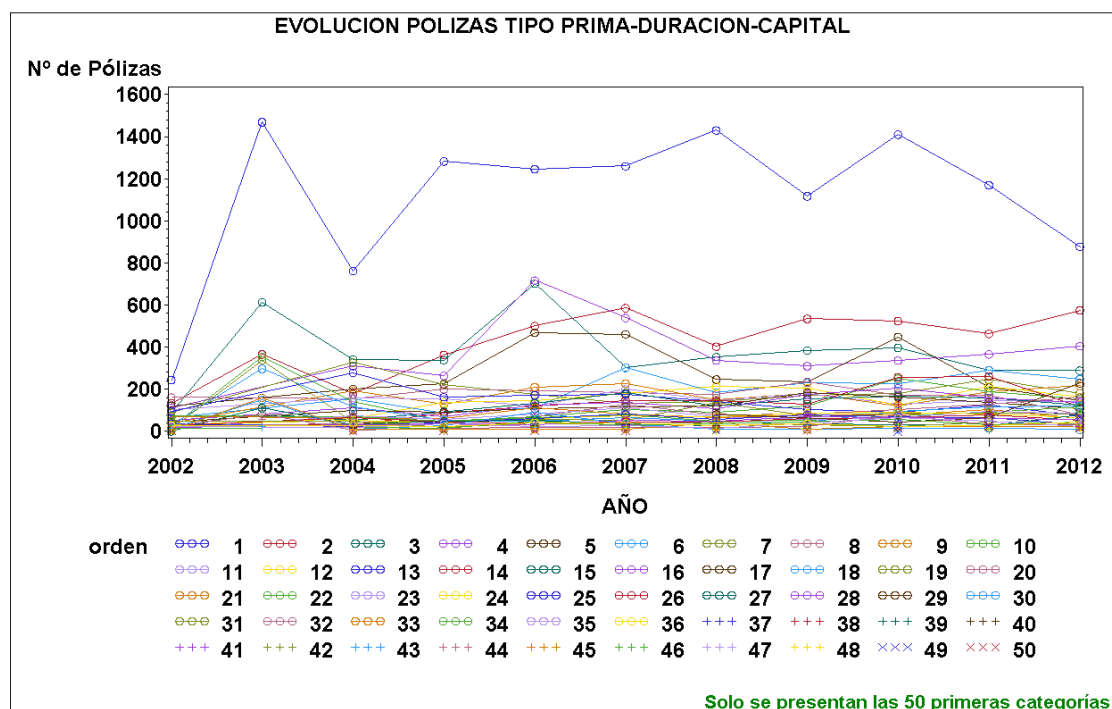
Las diferentes duraciones del contrato siguen pautas diferentes: las pólizas de duración baja y media siguen la pauta general, pero las de duración alta toman una dirección ligeramente ascendente para disminuir, muy poco, de 2011 a 2012.



En cuanto al capital, las pólizas de capital medio marcan la pauta general mientras que las de capital bajo y alto marcan pautas muy similares.



Sin embargo es necesario observar la evolución de las combinaciones de prima con duración y capital , pues pueden arrojar más luz sobre la tipología de las pólizas y su evolución en los años de crisis.



Se observa que la mayoría de las pólizas pertenecen a la combinación marcada en la figura como orden=1. Se muestran aquí las características de los primeros 10 tipos de pólizas:

prima2	dura2	cap2	COUNT	orden
1	1	2	12272	1
1	1	3	4631	2
1	1	1	4106	3
1	3	2	3700	4
1	2	2	2767	5
5	3	1	1901	6
5	2	2	1865	7
1	2	1	1719	8
1	3	3	1661	9
5	1	2	1546	10

La póliza mayoritaria tipo 1, que marca la tónica general de la evolución en crisis, es la que conjuga prima y duración baja, capital asegurado medio. La póliza tipo dos (prima, duración bajas pero capital alto) sigue una evolución contraria a la anterior, ascendente, así como la póliza tipo 4.

Al existir tantas combinaciones (tantos tipos de pólizas), surge la necesidad de intentar agrupar aquellas que han seguido una pauta de evolución general similar en años de crisis. Esto se realiza a través del análisis cluster, que construirá grupos de tipos de pólizas basándose en la evolución temporal en el periodo. Previamente se procede a la creación de cinco variables base para el análisis cluster, que corresponden respectivamente al total de pólizas en cada uno de los años 2008-2012.

Del análisis cluster realizado (se omiten los detalles), se obtienen cuatro clusters de tipos de pólizas claramente diferenciados en cuanto a su evolución en 2008-2012

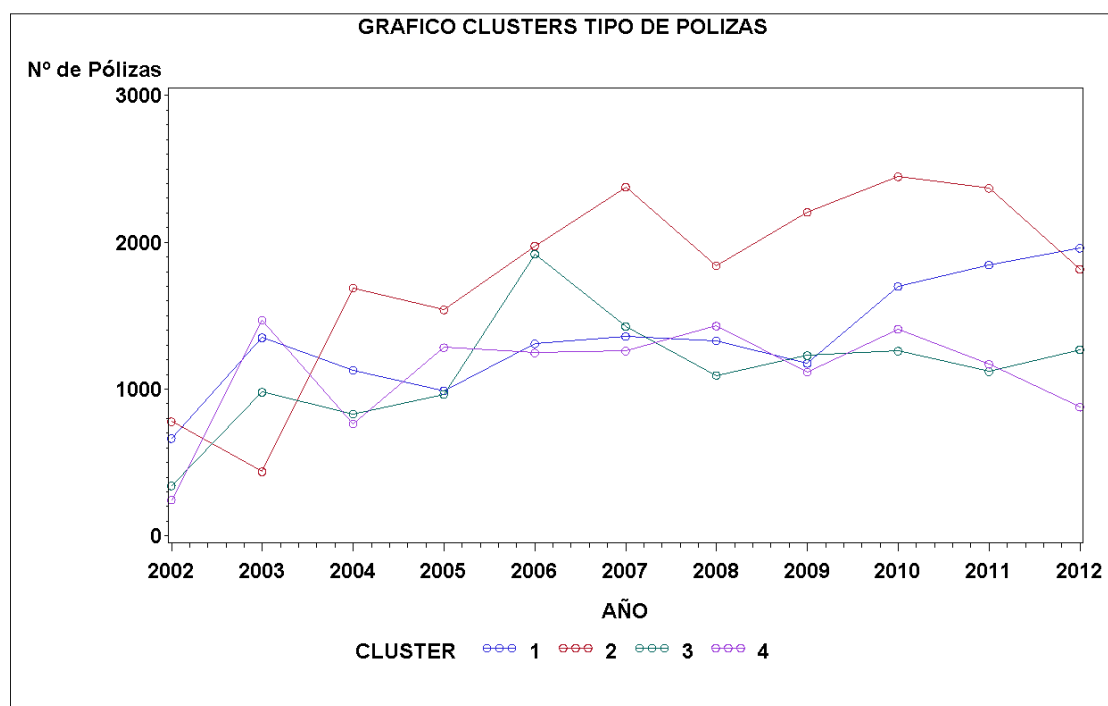
En la tabla presentada, el total 14815 , por ejemplo, se refiere a las pólizas que pertenecen al cluster 1 (en todo el periodo 2002-2012, aunque la información utilizada para crear los clusters sea solamente la de la crisis 2008-2012).

#### Cluster pólizas tipo

CLUSTER	COUNT	PERCENT
1	14815	25.1097
2	19477	33.0113
3	12437	21.0793
4	12272	20.7996

La tabla completa de las características de cada cluster está en el anexo, pero cabe destacar que el cluster cuatro está compuesto solamente por las pólizas prima muy baja, capital medio y duración baja, que sirve una curva fuertemente descendente a partir de 2010. El cluster 3 lo componen pólizas de prima muy baja, capital y duración bajos, y pólizas de prima muy baja , capital medio y duración alta, y de prima muy baja, capital alto, duración baja. Este cluster sigue una curva estable o ascendente a partir de 2011.

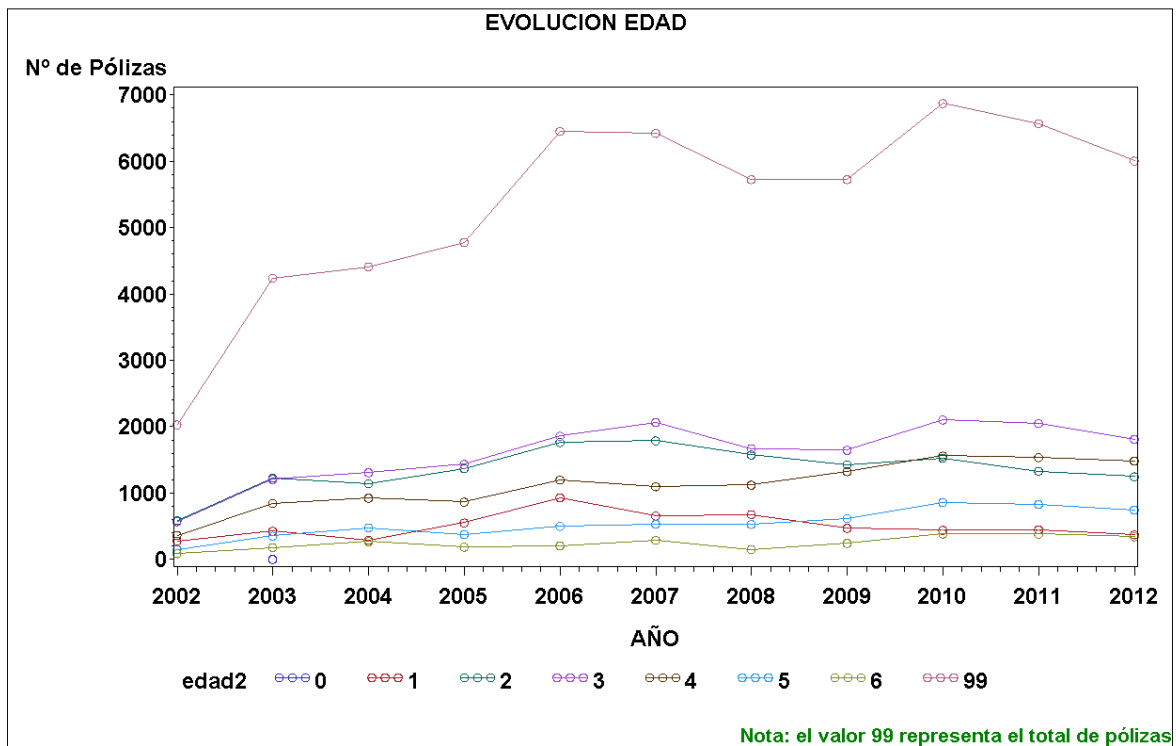
El cluster 2 marca una evolución ascendente de 2008 a 2010 y descendente después y lo componen primas de todo tipo salvo las intermedias, con capitales mayormente medio-altos. El cluster 1 lo componen vel resto de pólizas, y su evolución es contraria a la general, siguiendo una curva ascendente.



Cabe preguntarse ahora cómo los diferentes perfiles de asegurados se reparten entre estos clusters de pólizas. La información de la que se dispone sobre los asegurados es muy limitada, y solamente las variables edad y profesión dan información relevante (se ha probado a incluir otras como estado civil, localidad, etc. sin buenos resultados).



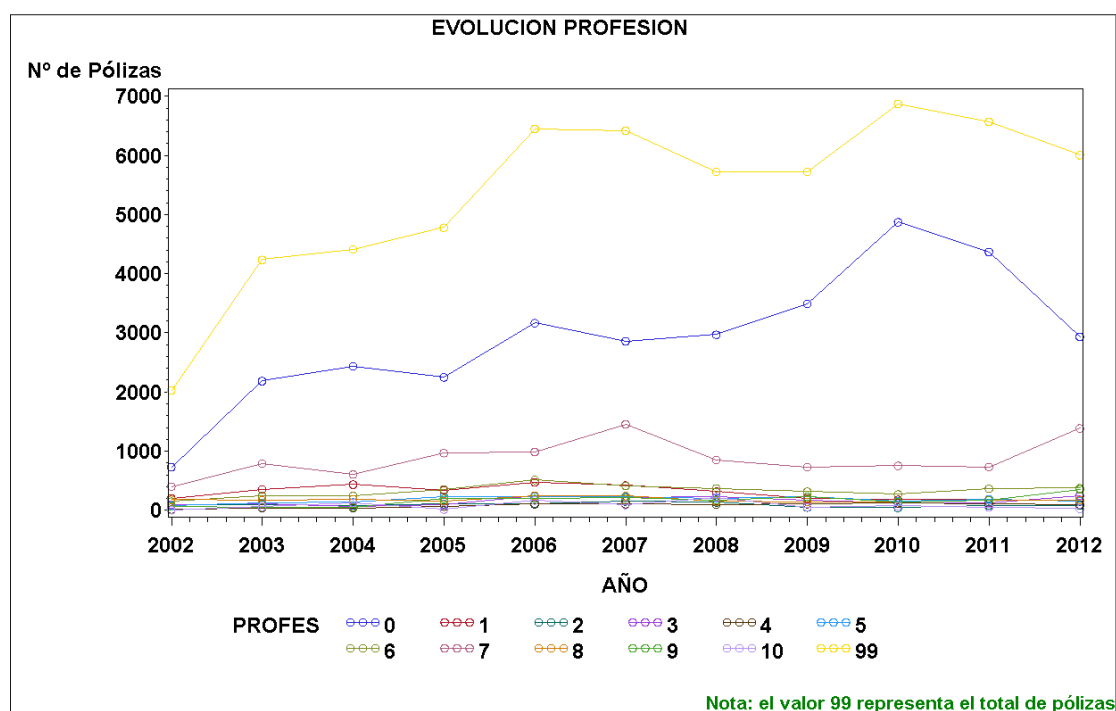
En primer lugar se observa la evolución por separado de los distintos grupos de edad y profesión. Por separado, da la impresión de que los grupos de edad se comportan de manera similar.



En cuanto a las profesiones, recordemos los códigos:

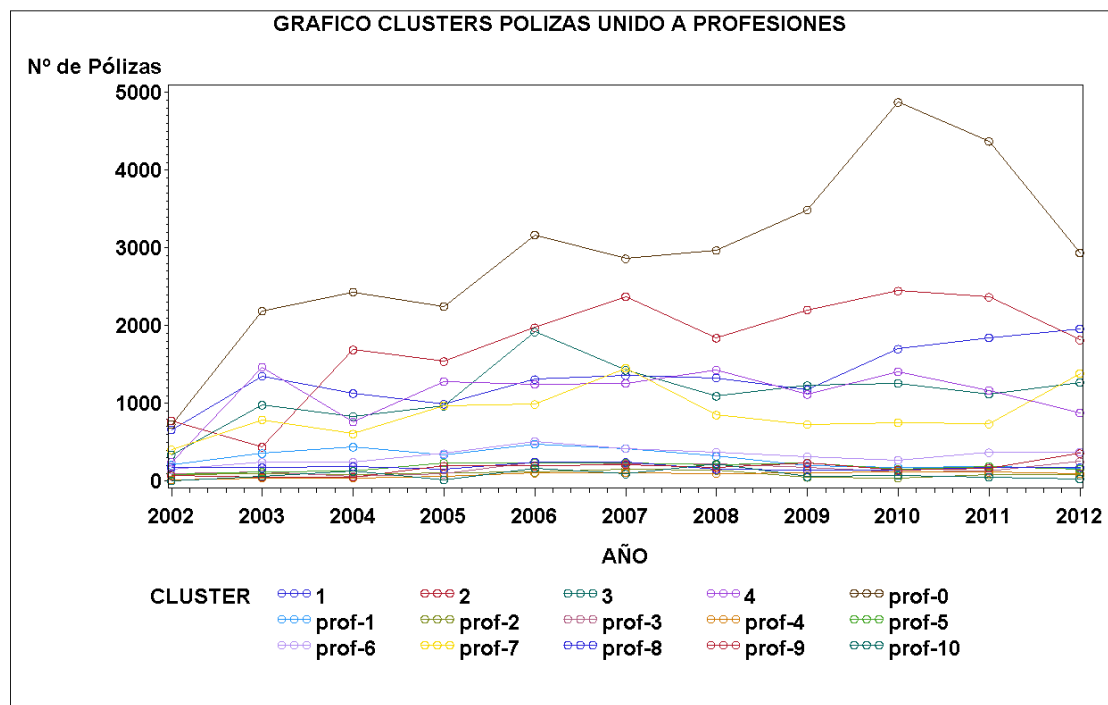
- 0 Desconocida
- 1 Autónomos y Comerciantes
- 2 Empresarios
- 3 Profesionales Liberales
- 4 Agricultores y Ganaderos
- 5 Directivos y Mandos Intermedios
- 6 Empleados con Titulación
- 7 Empleados sin Titulación
- 8 Amas de Casa
- 9 Jubilados y Pensionistas
- 10 Otra

Los empleados sin titulación (profes=7) siguen una curva contraria a la evolución general, mientras que las profesiones codificadas como 0 o desconocidas siguen la pauta general. En esta categoría entrarían tanto los parados, como aquellos que no quieren identificar su profesión, incluidos por ejemplo los trabajadores en el sector de la “economía sumergida” : empleadas del hogar, fontanería, pintura, etc.



La edad por sí sola no ofrece diferencias, pero es interesante proyectar las curvas de profesiones con los clusters de pólizas observados anteriormente.

A priori en el gráfico se puede observar que la profesión 7 (empleados sin titulación) sigue una evolución que podría estar entre los clusters 1 y 3. Otras profesiones como la 0, estarían repartidas entre los clusters 2 y 4. En todo caso, es menos preciso observar estas evoluciones por separado que relacionarlas estadísticamente. Para ello se utilizará un árbol de decisión en el que se segmentarán los grupos de edad y profesión para ver su distribución en términos de los distintos clusters de pólizas.



Para obtener el resultado de los árboles de decisión se ha optado por realizar árboles diferentes por grupos de edad (se ha intentado incluir edad como variable explicativa pero los resultados eran menos claros). Los resultados generales están en el anexo. Se presentará aquí una explicación de los más destacados.

En primer lugar hay que recordar la distribución porcentual de tipos de pólizas en los cuatro clusters:

#### Cluster polizas tipo

CLUSTER	COUNT	PERCENT
1	14815	25.1097
2	19477	33.0113
3	12437	21.0793
4	12272	20.7996

Los resultados del árbol darán para cada grupo, diferentes porcentajes en cada uno de los clusters de pólizas, y es bueno tener en cuenta la distribución general de todas las pólizas analizadas en los cuatro clusters (25%,33%,21%,20%). Hay que señalar que aunque los clusters están creados en función del comportamiento en crisis, se ha tenido en cuenta todo el periodo de 2002 a 2012 para la construcción de segmentos con el árbol, para aprovechar más toda la información .

Se ha realizado en primer lugar un árbol para todas las edades consideradas en conjunto. Se observa que las profesiones 8,9,10 (amas de casa, jubilados y pensionistas y otros) tienen una distribución porcentual entre los clusters diferente a las pólizas en general:

Este grupo de profesiones ubica un 40% de sus pólizas en el cluster 2 (frente a solo un 33% en general en todas las profesiones). Este cluster tiene una evolución ascendente de 2008 a 2010 y descendente después y primas de todo tipo salvo las intermedias, con capitales mayormente medio-altos; se pueden ver las características del cluster en el apartado de la definición de los clusters.

Estas profesiones tienen respecto al total, un porcentaje menor de pólizas tipo 3 y 4 (en torno al 12% para los jubilados y pensionistas, y al 16% para amas de casa y otros, frente a cerca de 20% en todas las profesiones), y ligeramente mayor a la tónica general de pólizas de tipo 1 (cerca del 27%).

El resto de profesiones sigue una proporción de pólizas similar a la generalidad.

Si se desagrega por edades y se realiza el árbol se obtienen también algunos resultados interesantes.

#### Para los jóvenes (edad2=1, menores de 25 años),

En las profesiones 5 y 8, Amas de Casa y Directivos o Mandos intermedios, casi un 50% de las pólizas pertenecen al cluster 2.

En profesiones liberales y pensionistas, tiene más preponderancia el cluster 1 (33%) que en la tónica general (25%), y menos el 3 y 4. Recordemos que el cluster 1 lo comportan gran variedad de pólizas, y sigue evolución ascendente durante la crisis.

#### Para adultos (edad2=2,3,4, 25 -54 años),

En este rango de edades cabe destacar que las profesiones 2 y 5 (empresarios, Directivos o Mandos intermedios) tienen una distribución diferente al resto de la tónica general, con mayor inclinación hacia el cluster 1 (31% frente al 25%) y menor hacia los 2,3,4 de la generalidad de las profesiones.

Las profesiones 3 y 9 (liberales y jubilados-pensionistas) siguen una distribución parecida al caso anterior pero menos marcada (29% en el cluster 1).

#### Para adultos (edad2=5, 54-65 años),

Aquí no se destacan grandes diferencias entre las profesiones. Este grupo de edad en general, está centrado en los clusters 1 y 2.

#### Para mayores (edad2=6, más de 65 años),

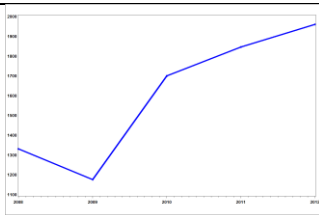
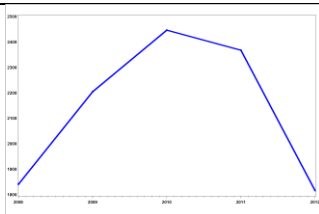
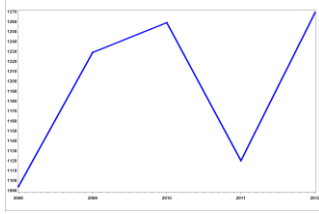
En este grupo sí existe diferencia entre las profesiones: jubilados, pensionistas, amas de casa, autónomos y comerciantes y agricultores siguen esta distribución en los cuatro clusters (gran preponderancia del cluster 2):

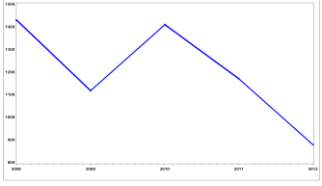
1	:	21.7%
2	:	72.8%
3	:	4.9%
4	:	0.7%

El resto de profesiones tampoco tiene gran representación en los clusters 3 y 4, pero en el 1 tiene cerca de un 36% frente al 21.7% de las profesiones mencionadas más arriba, y en el cluster 2 un 57%.

Como resumen, se puede señalar que hay determinados grupos que hacen ascender la representación de cada cluster y otros que la hacen descender, en términos relativos respecto a la proporción general. Recordemos que cuando una porción de determinado colectivo pertenece a un cluster determinado, esto significa por un lado, que contrata el tipo de pólizas correspondientes a ese cluster, y que por otro lado sigue la evolución temporal en tiempo de crisis correspondiente a ese cluster.

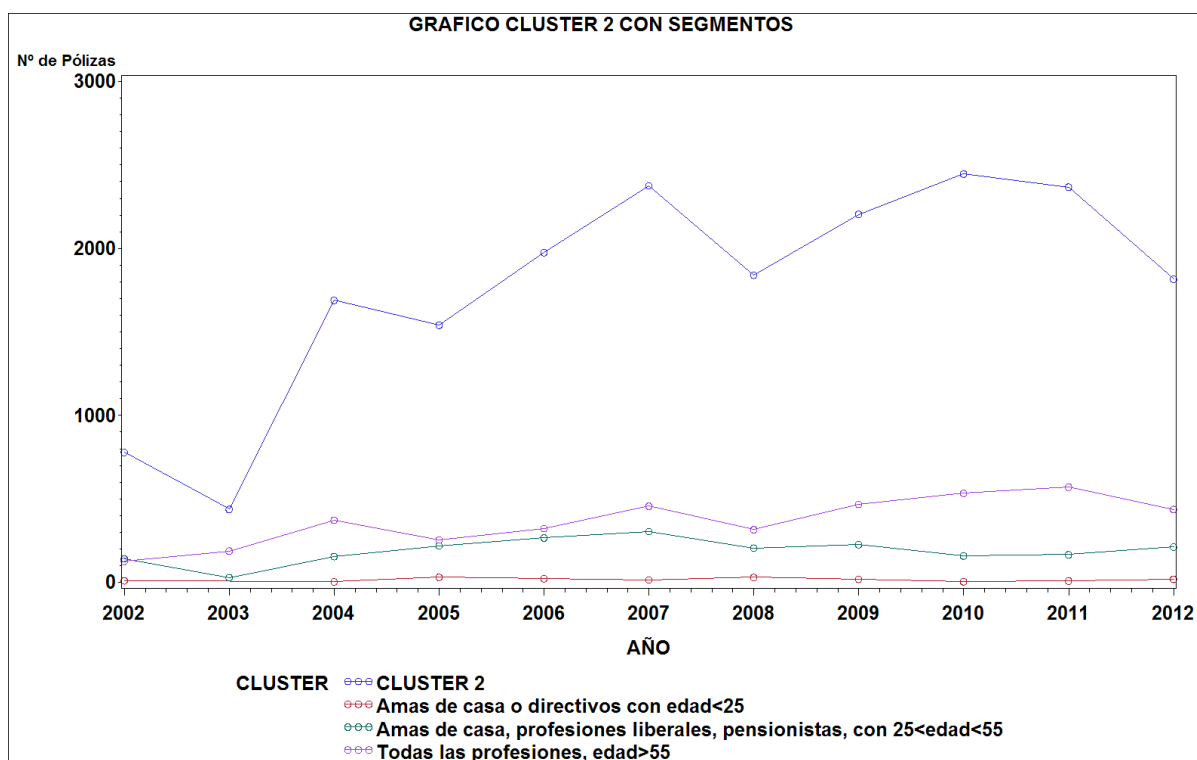
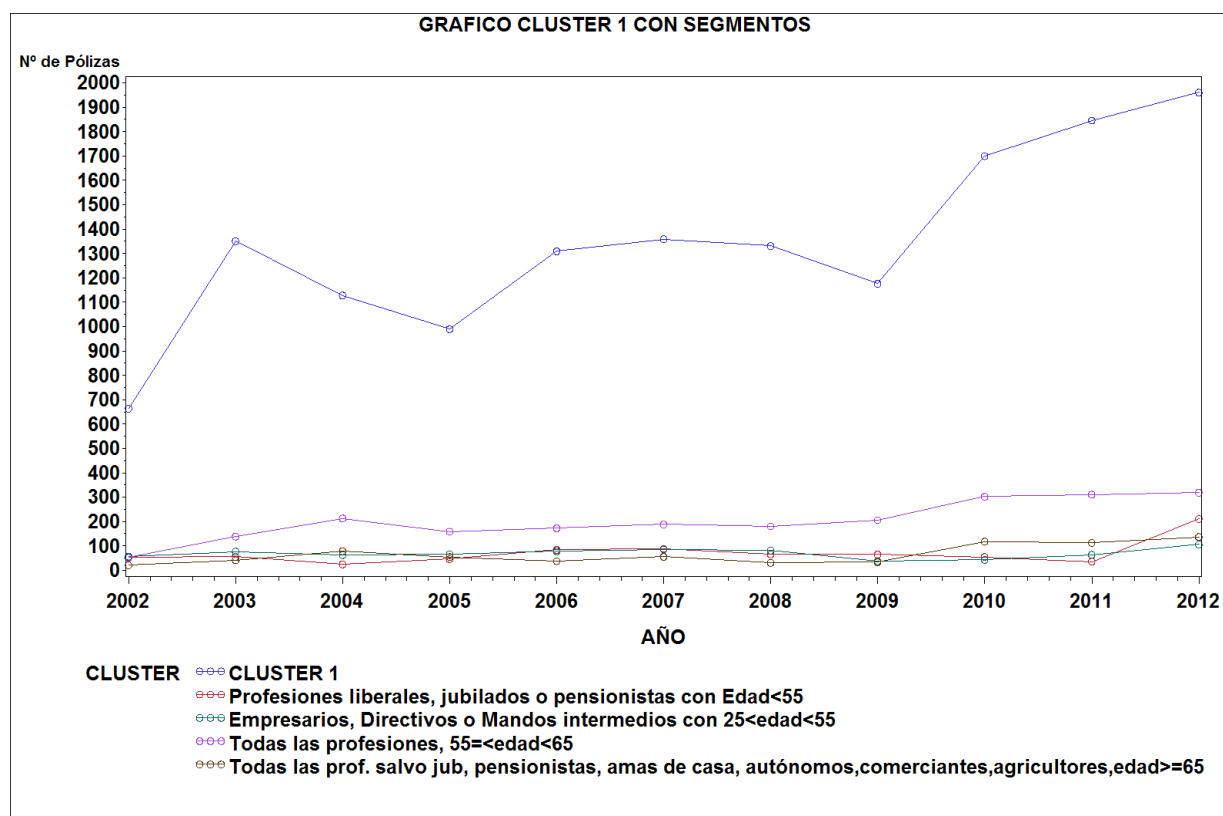
Es necesario recordar que no está representada toda la población en todos estos segmentos señalados (aunque sí la gran mayoría), sino que son aquellos con un comportamiento más claro y homogéneo en cuanto a las pautas estudiadas.

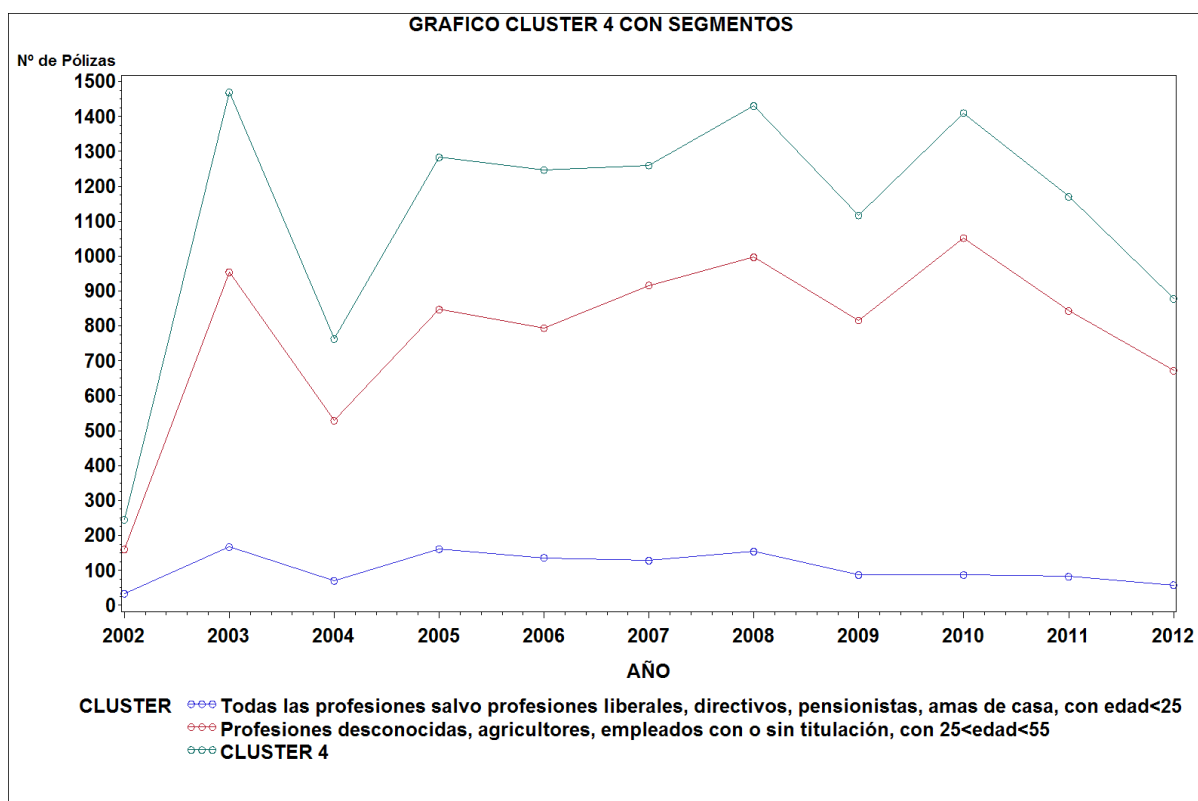
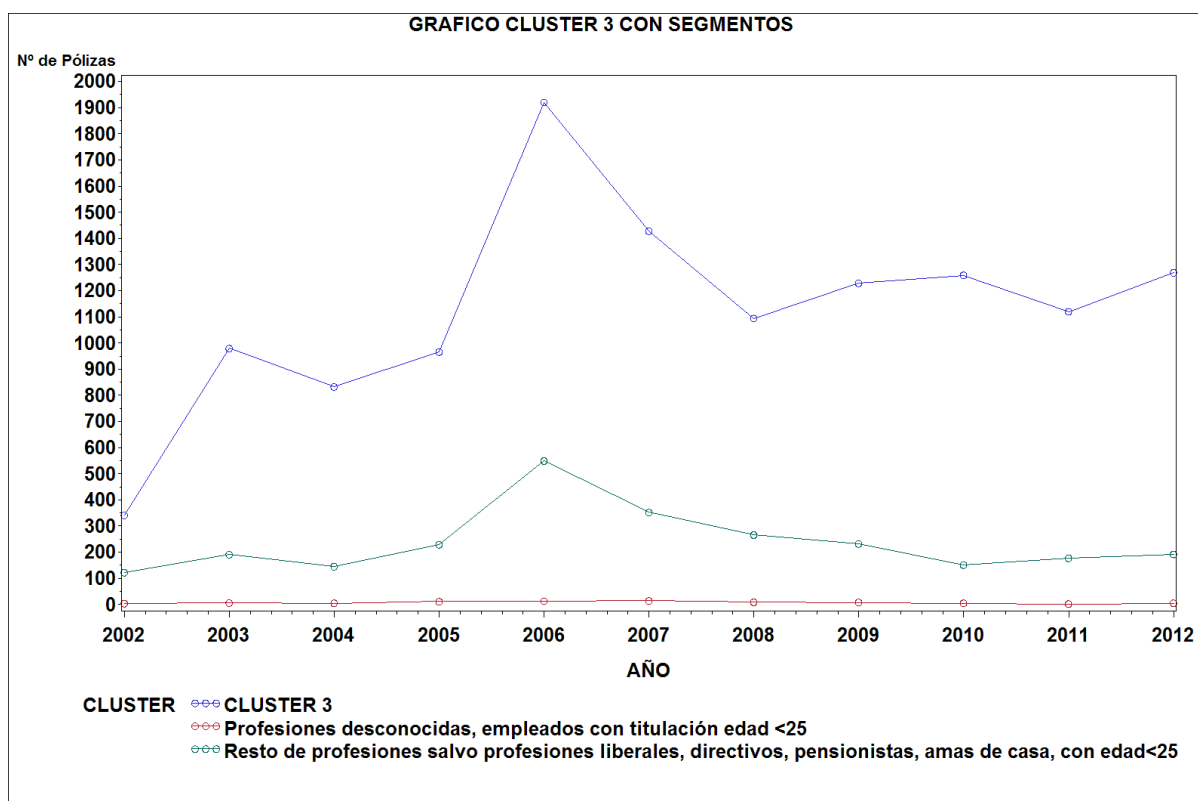
Cluster	Evolución 2008-2012	Tipo de pólizas	Grupos con claramente mayor representación del cluster que la proporción general
1		Pólizas no representadas en los clusters 2,3,4 (25% del total)	Profesiones liberales, jubilados o pensionistas con Edad<55  Autónomos o comerciantes, Directivos o Mandos intermedios con 25<edad<55  Todas las profesiones, 55=<edad<65  Todas las profesiones <b>salvo</b> jubilados, pensionistas, amas de casa, autónomos y comerciantes y agricultores, con edad>=65
2		primas de todo tipo salvo las intermedias, capitales mayormente medio-altos (33% del total)	Amas de casa o directivos con edad<25  Amas de casa, profesiones liberales, pensionistas, con 25<edad<55  Todas las profesiones, 55<edad
3		prima muy baja, capital y duración bajos  prima muy baja, capital medio y duración alta  prima muy baja, capital alto, duración baja (21% del total)	Profesiones desconocidas, empleados con titulación edad <25.  Resto de profesiones <b>salvo</b> profesiones liberales, directivos, pensionistas, amas de casa, con edad<25

4		<p>prima muy baja, capital medio y duración baja</p> <p>(20% del total)</p>	<p>Todas las profesiones <b>salvo</b> profesiones liberales, directivos, pensionistas, amas de casa, con edad&lt;25</p> <p>Profesiones desconocidas, agricultores, empleados con o sin titulación, con 25&lt;edad&lt;55</p>
---	---	---	---

Una vez observados esos segmentos poblacionales de interés, es necesario corroborar gráficamente cómo su evolución corresponde adecuadamente a cada cluster. Salvo algún segmento, su evolución en crisis (y antes de la crisis) corresponde bien a la evolución del cluster, con lo cual los segmentos poblacionales están bien definidos en cuanto a su evolución, género de pólizas y características de edad y profesión.

Hay que tener en cuenta que en los gráficos siguientes solo se proyectan ciertos colectivos de interés pertenecientes a cada cluster, en realidad cada cluster (tipos de pólizas) abarca prácticamente toda la población pero con menor o mayor participación de cada colectivo. Además de cada colectivo solamente se proyectan aquellas pólizas que corresponden al cluster correspondiente (en prima, duración, capital).



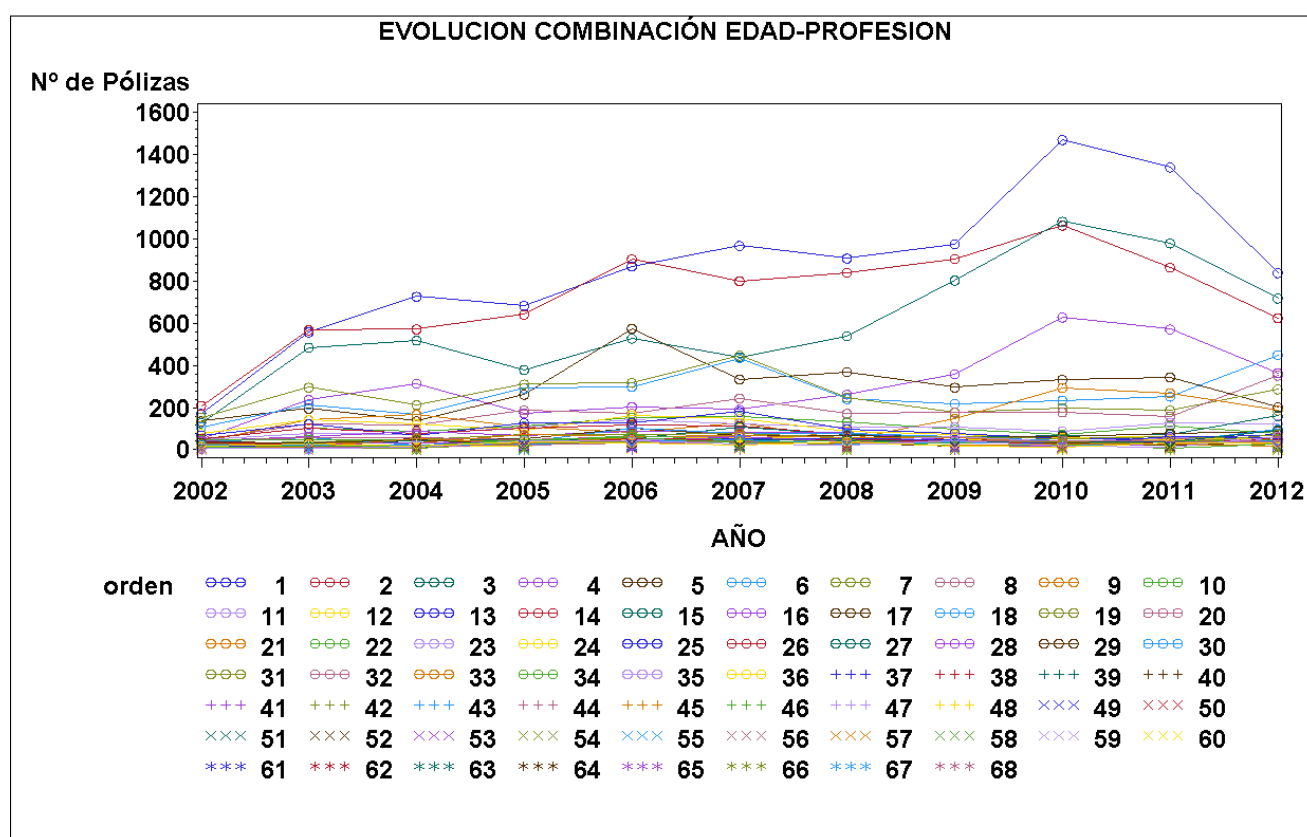




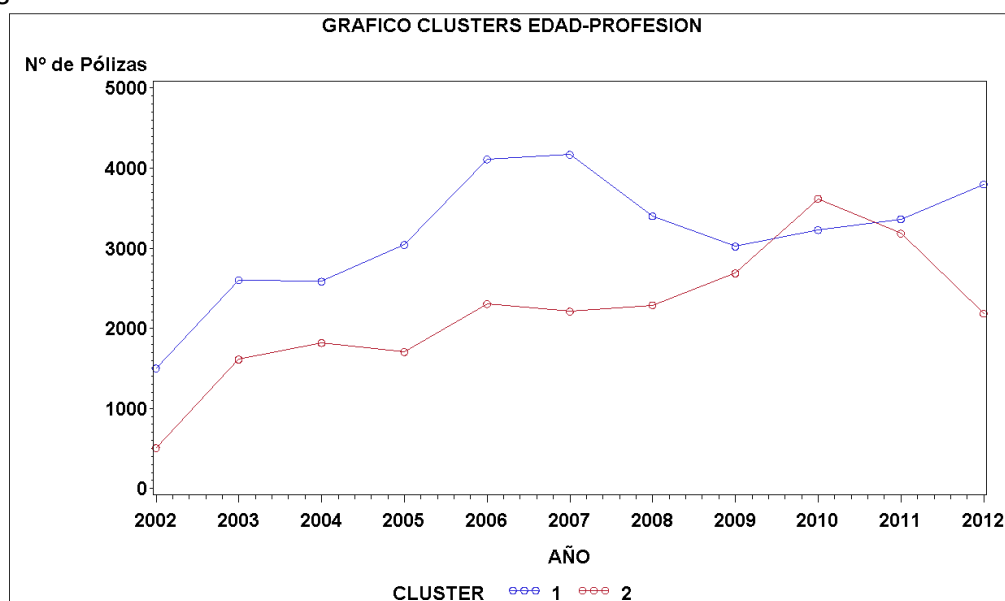
#### 4) Estudio comparativo de la evolución temporal de las pólizas en los años de crisis, atendiendo a sus características de edad y profesión.

En este apartado se realizará un análisis en cierto modo simétrico respecto al anterior: se crearán clusters que engloben los segmentos de edad o profesión en cuanto a un comportamiento similar en la evolución de pólizas contratadas en tiempos de crisis. Se espera observar estructuras similares a las observadas en el apartado anterior, y corroborarlas.

Como se ve en el gráfico, hay ciertas combinaciones de edad y profesión que siguen la evolución general en la crisis, y otras que van en sentido ascendente.



Realizando el análisis cluster para obtener grupos de comportamiento homogéneo, se llega a solamente dos clusters



El cluster 2 está compuesto exclusivamente por pólizas contratadas por asegurados con profesión desconocida (lo que puede significar, parado, economía sumergida, etc.) con edades comprendidas entre 25 y 55 años. Este cluster tiene una evolución ascendente hasta 2010 y descendente después. El cluster 1 engloba al resto de individuos, y su evolución global toma una curva ascendente.

CLUSTER	COUNT	PERCENT
1	34820	59.0760
2	24121	40.9240

Es conveniente recordar que aunque diferentes tramos de edad estén englobados en el cluster 1, entre ellos es posible que haya ciertos comportamientos de evolución posiblemente no ascendentes, lo que ocurre es que su evolución es más parecida al cluster 1 que al 2 (en los términos de la distancia utilizada por el algoritmo cluster). Sin embargo, el gráfico en el que se ve la evolución de todas las combinaciones edad-profesión permite pensar que hay cierta homogeneidad interna en cada uno de los clusters.

De manera similar al apartado anterior, se realiza un árbol de decisión, donde se determinan grupos de pólizas atendiendo a sus categorías de prima, capital y duración, para ver a qué cluster de población pertenecen en mayor o menor porcentaje.

En los resultados, cabe destacar lo siguiente:

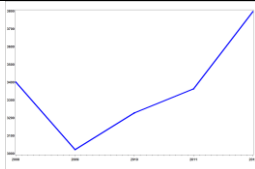
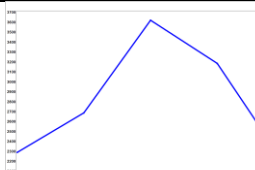
Las pólizas de duración 0, 2 y 3 (desconocida, y de más de 4 años)  
y las pólizas de duración 1 (<4 años) , capital bajo y prima alta (>1200 euros)  
y las pólizas de duración 1 (<4 años), capital desconocido y medio alto, y prima alta

tienen mayor presencia respecto a la tónica general en el cluster 1 (evolución ascendente), y menor presencia en el cluster 2 (evolución descendente).

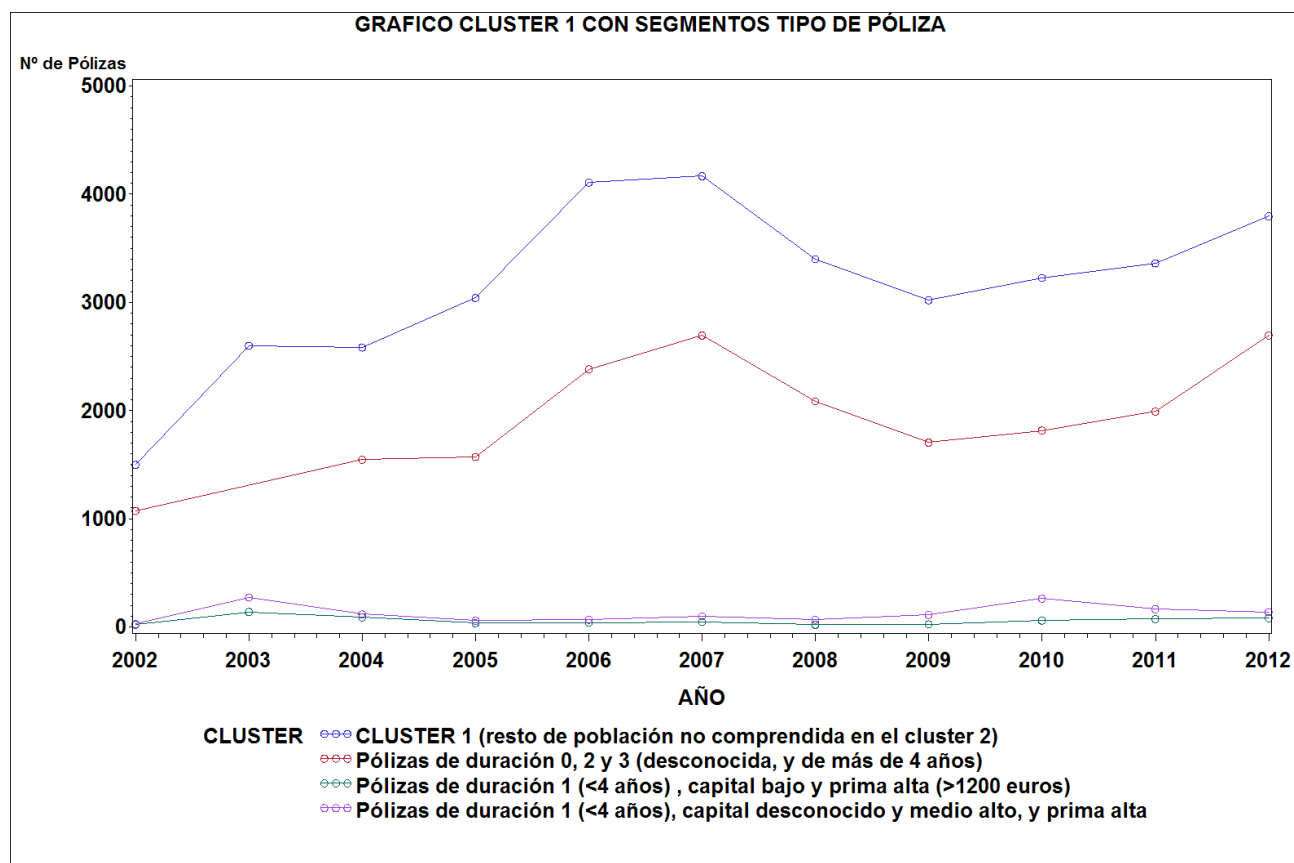
Las pólizas de duración 1 (<4 años), y prima menor que 1200 euros,

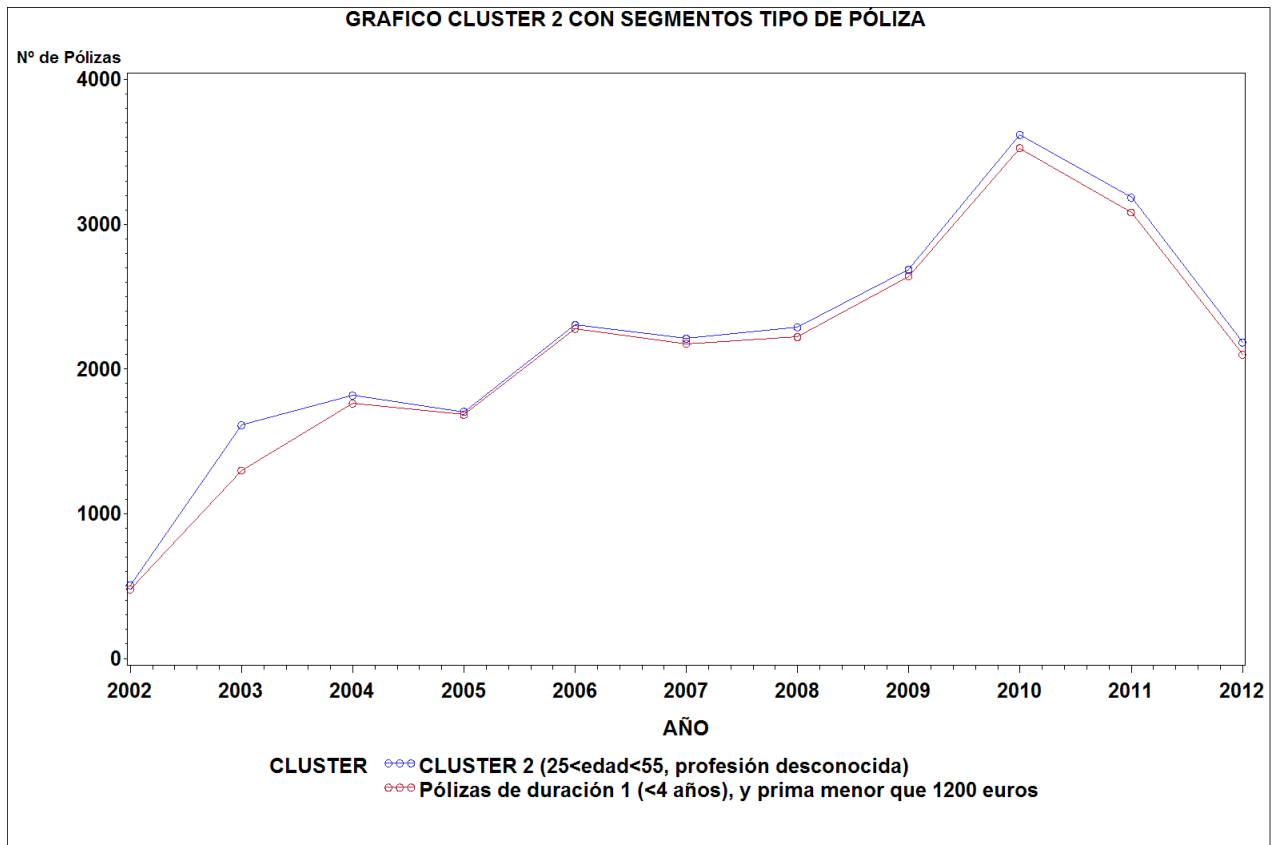
tienen mayor presencia respecto a la tónica general en el cluster 2 (evolución descendente), y menor presencia en el cluster 1 (evolución ascendente).

Se puede construir una tabla resumen así:

Cluster	Evolución 2008-2012	edad-profesión	Tipos de Pólizas con claramente mayor representación del cluster que la proporción general
1		Población no comprendida en el cluster 2.	<p>Pólizas de duración 0, 2 y 3 (desconocida, y de más de 4 años)</p> <p>Pólizas de duración 1 (&lt;4 años) , capital bajo y prima alta (&gt;1200 euros)</p> <p>Pólizas de duración 1 (&lt;4 años), capital desconocido y medio alto, y prima alta</p>
2		Profesión desconocida , edad entre 25 y 55 años	Pólizas de duración 1 (<4 años), y prima menor que 1200 euros

Se realiza ahora un estudio gráfico de estos segmentos definidos de tipos de pólizas relativamente a los clusters estudiados. Se observa cómo esos segmentos o tipos de pólizas detectados corresponden bien a la misma evolución del cluster, en general. Se recuerda que no está representados todos los tipos de pólizas sino sólo aquellos con un comportamiento más claro y homogéneo en cuanto a las pautas estudiadas.





## 5) Conclusiones y recomendaciones

La evolución del número de pólizas en general en el periodo de crisis sigue una pauta ascendente que alcanza un máximo en 2010 y es descendente a partir de entonces. Sin embargo, se observa que la evolución del total de pólizas en este periodo puede llegar a ser muy diferente según el tipo de póliza suscrita, y el segmento poblacional de edad y profesión. Además, se observa que el comportamiento en crisis y antes de ella no es constante en general, con lo cual análisis estadísticos que agreguen datos en dos periodos “antes de la crisis” y “durante toda la crisis” no estarían bien planteados, hay que tener en cuenta la serie en su totalidad.

Buscando pautas generales que permitan describir las características comunes de las pólizas y sus asegurados según su comportamiento durante la crisis, se han encontrado de manera indirecta grupos relacionados de edad y profesión con la evolución temporal y el tipo de pólizas mayoritarias. Se ha abordado el problema de manera simétrica, actuando sobre clusters de pólizas tipo, y sobre clusters de asegurados. La razón es que combinar a priori ambos tipos de información para obtener una segmentación que los relacione (via cluster o árboles) no da resultados claros al corresponder a distintos conceptos.

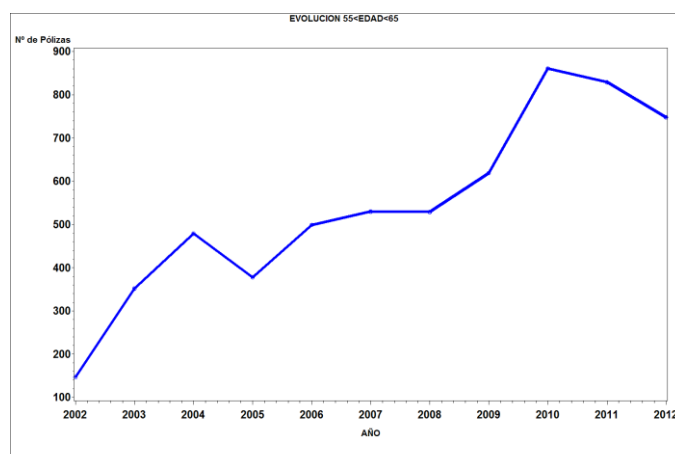
La inclusión de más información sería interesante. La localidad no parece ayudar mucho (se incluye un gráfico en el anexo donde se observa que el año 2010 tiene demasiados registros con localidad desconocida, lo que no permite utilizar esta variable con garantías). En general es sabido que la información del comportamiento de consumo tiene más fuerza predictiva y estadística para orientar el comportamiento a la hora de suscribir un seguro, que las simples variables socioeconómicas. Así, parámetros como que el asegurado tenga otros tipos de seguros o servicios relacionados permitiría construir más segmentos significativos, y más sólidos que los simplemente basados en la edad y profesión.

Además, la no existencia de un código individual identificador de cada póliza impide conocer, por ejemplo, su continuidad en el tiempo, patrones de cambio durante la crisis de un tipo de póliza a otra, o bien establecer modelos para la cancelación de los contratos, evaluar el efecto de la crisis sobre la cancelación o derivación, etc. En concreto, no se sabe en los datos analizados qué pólizas son “repetidas” (continuación de la póliza del año anterior) o nuevamente contratadas, lo que puede complicar su interpretación al duplicarse de algún modo la información.

En cuanto a los resultados obtenidos, se han detectado segmentos poblacionales con un comportamiento de evolución muy similar, y también en cuanto a las primas contratadas. Hay que tener en cuenta la dificultad de la interpretación de los resultados: se trata de detectar comportamientos generalmente similares para profundizar a posteriori sobre esta información, pero a nivel individual el carácter predictivo de esta información es muy bajo.

Como ejemplo nos centramos en el estudio de cluster de pólizas tipo. Concretamente en el comportamiento del colectivo de todas las profesiones, edades entre 55 y 65 ( $\text{edad}=5$ ). Este colectivo tiene la particularidad de tener un alto porcentaje, 37%, en el cluster 1 (con sus tipos específicos de pólizas). Este 37% del colectivo de  $55 < \text{edad} < 65$  sigue una evolución ascendente en el tiempo de crisis. Sin embargo, también en este mismo colectivo hay un 40% que suscriben las pólizas tipo del cluster 2, con una evolución ascendente de 2008-2010 y descendente en 2010-2012. Igualmente, este colectivo tiene porcentajes cercanos al 10% en las pólizas tipo del cluster 3 y 4, con sus propias evoluciones.

Si se proyecta la curva global correspondiente al colectivo  $55 < \text{edad} < 65$ , no se tendrá en realidad una noción clara de la evolución temporal de las pólizas contratadas, al coexistir en el mismo colectivo perfiles con evolución contradictoria. Por ejemplo, la evolución general del colectivo  $55 < \text{edad} < 65$  es similar a la tónica general en el periodo de crisis:



Observando esta curva es difícil imaginar que hay subsegmentos de ese mismo colectivo que siguen simultáneamente curvas de evolución en crisis como las de cada uno de los clusters 1,2,3,4, pero así es, y su agregación da lugar a la curva arriba expuesta.

Para mayor confusión, este colectivo concreto está englobado, en el análisis de cluster por edad-profesión, en el colectivo que tiene una curva global ascendente en crisis. Pero al estar agregado a otros colectivos, la importancia de su curva se diluye (aquí es necesario señalar que los cluster edad profesión se basan en combinaciones edad-profesión y ninguna de ellas consiste en “todas las profesiones,  $55 < \text{edad} < 65$ ”).

Por lo tanto todas las conclusiones que se extraen deben de tomarse en términos porcentuales, y poniendo especial cuidado en las agregaciones y desagregaciones antes de llegar a conclusiones generales o específicas.

Para cada colectivo de interés se puede incidir más en ellos observando el detalle estadístico con la ayuda de los datos. Además, los segmentos detectados pueden verse como un primer paso para obtener información más valiosa, utilizando conocimientos cualitativos y/o añadiendo nueva información en términos de variables estadísticas.

Por último cabe añadir que el modo en que se ha realizado el estudio admite más variantes y proyecciones. Aunque se han construido los clusters teniendo en cuenta solamente la evolución en tiempo de crisis, igualmente podría hacerse un análisis en el periodo pre-crisis y comparar resultados, pero de

momento no hemos considerado relevante ese análisis. Igualmente la misma manera de proceder puede aplicarse a determinados segmentos de la población (pólizas solamente de Barcelona, por ejemplo) o bien con otras variables de interés. Es decir, básicamente se puede cambiar el periodo de análisis, la población objetivo o las variables para formar los clusters o para construir segmentos relativos a los clusters. Otras variaciones incluyen tener en cuenta las variables prima y capital en su forma continua original; en general se obtiene más información de esa manera, pero en este caso hemos preferido la simplicidad de su versión categorizada, en lugar de una pequeña ganancia en precisión.

## ANEXO

### 1) Composición de los clusters obtenidos con prima, capital y duración

Cluster polizas tipo

CLUSTER	COUNT	PERCENT
1	14815	25.1097
2	19477	33.0113
3	12437	21.0793
4	12272	20.7996

CLUSTER	prima2	cap2	dura2	COUNT	PERCENT
1	0	0	1	56	0.0949
1	0	1	3	34	0.0576
1	1	0	0	289	0.4898
1	1	0	1	46	0.0780
1	1	0	3	102	0.1729
1	1	1	0	67	0.1136
1	1	2	0	421	0.7135
1	1	3	0	120	0.2034
1	1	3	2	312	0.5288
1	2	0	3	51	0.0864
1	2	1	1	391	0.6627
1	2	1	3	778	1.3186
1	2	2	1	1016	1.7220
1	2	3	2	221	0.3746
1	2	3	3	707	1.1983
1	3	1	1	479	0.8119
1	3	1	2	358	0.6068
1	3	1	3	714	1.2101
1	3	2	1	329	0.5576
1	3	2	2	651	1.1034
1	3	2	3	526	0.8915
1	3	3	1	302	0.5119
1	3	3	2	111	0.1881
1	3	3	3	201	0.3407
1	4	0	3	35	0.0593
1	4	1	1	262	0.4441
1	4	1	2	544	0.9220
1	4	1	3	849	1.4390
1	4	2	1	141	0.2390
1	4	2	2	470	0.7966
1	4	2	3	197	0.3339
1	4	3	1	156	0.2644
1	4	3	2	93	0.1576
1	4	3	3	122	0.2068
1	5	0	3	118	0.2000
1	5	1	1	1039	1.7610
1	5	1	2	1422	2.4101
1	5	2	0	89	0.1508
1	5	3	1	348	0.5898
1	5	3	2	319	0.5407
1	5	3	3	329	0.5576
2	1	1	2	1719	2.9135
2	1	1	3	1477	2.5033
2	1	2	2	2767	4.6898
2	1	3	3	1661	2.8152
2	2	1	2	1170	1.9830
2	2	2	2	1276	2.1627



2	2	2	3	1480	2.5084
2	2	3	1	1294	2.1932
2	5	1	3	1901	3.2220
2	5	2	1	1546	2.6203
2	5	2	2	1865	3.1610
2	5	2	3	1321	2.2389
3	1	1	1	4106	6.9592
3	1	2	3	3700	6.2711
3	1	3	1	4631	7.8490
4	1	2	1	12272	20.7996

## 2) Resultados del árbol de decisión con cluster de prima, capital, duración como variable dependiente y profesión como independiente, para todas las edades

### PROFESIONES EN GENERAL TODAS LAS EDADES

IF PROFES IS ONE OF: 0 1 4 6 7

THEN

N : 49621  
 1 : 24.2%  
 2 : 32.1%  
 3 : 21.9%  
 4 : 21.8%

IF PROFES IS ONE OF: 2 3 5

THEN

N : 4691  
 1 : 31.7%  
 2 : 33.2%  
 3 : 18.7%  
 4 : 16.4%

IF PROFES IS ONE OF: 8 10

THEN

N : 2857  
 1 : 26.4%  
 2 : 40.8%  
 3 : 16.3%  
 4 : 16.5%

IF PROFES EQUALS 9

THEN

N : 1832  
 1 : 29.7%  
 2 : 45.4%  
 3 : 13.6%  
 4 : 11.3%

## 3) Resultados del árbol de decisión con cluster de prima, capital, duración como variable dependiente y profesión como independiente, realizando el árbol para cada grupo de edad.

### EDAD 1

IF PROFES IS ONE OF: 0 6

THEN

N : 3493  
 1 : 15.5%  
 2 : 31.4%  
 3 : 30.9%  
 4 : 22.2%

IF PROFES IS ONE OF: 1 2 4 7  
THEN

N	:	1402
1	:	19.4%
2	:	31.1%
3	:	23.3%
4	:	26.2%

IF PROFES IS ONE OF: 3 9  
THEN

N	:	224
1	:	33.0%
2	:	30.4%
3	:	19.6%
4	:	17.0%

IF PROFES IS ONE OF: 5 8 10  
THEN

N	:	410
1	:	22.4%
2	:	48.8%
3	:	17.3%
4	:	11.5%

EDAD2=2,3,4

IF PROFES IS ONE OF: 0 4 6 7  
THEN

N	:	35452
1	:	22.6%
2	:	29.7%
3	:	23.5%
4	:	24.2%

IF PROFES EQUALS 1  
THEN

N	:	2580
1	:	27.4%
2	:	30.7%
3	:	20.5%
4	:	21.3%

IF PROFES IS ONE OF: 8 10  
THEN

N	:	1935
1	:	25.7%
2	:	35.7%
3	:	17.4%
4	:	21.2%

IF PROFES IS ONE OF: 2 5  
THEN

N	:	2407
1	:	31.6%
2	:	30.7%
3	:	19.7%
4	:	18.0%

IF PROFES IS ONE OF: 3 9

THEN

N	:	2469
1	:	28.9%
2	:	35.4%
3	:	19.3%
4	:	16.4%

Edad2=5

IF PROFES IS ONE OF: 0 1 2 3 5 7 9 10

THEN

N	:	5222
1	:	38.2%
2	:	40.2%
3	:	10.7%
4	:	10.8%

IF PROFES IS ONE OF: 4 6 8

THEN

N	:	701
1	:	35.1%
2	:	39.5%
3	:	16.5%
4	:	8.8%

Edad2=6

IF PROFES IS ONE OF: 8 9 1 4

THEN

N	:	762
1	:	21.7%
2	:	72.8%
3	:	4.9%
4	:	0.7%

IF PROFES IS ONE OF: 0 3

THEN

N	:	1614
1	:	37.7%
2	:	57.7%
3	:	2.7%
4	:	1.9%

IF PROFES IS ONE OF: 5 6 7 2 10

THEN

N	:	329
1	:	35.3%
2	:	56.2%
3	:	7.6%
4	:	0.9%

#### 4) Composición de los clusters obtenidos con edad y profesión.

CLUSTER	COUNT	PERCENT
1	34820	59.0760
2	24121	40.9240

CLUSTER	edad2	PROFES	COUNT	PERCENT
1	1	0	3188	5.4088
1	1	1	261	0.4428
1	1	3	98	0.1663
1	1	5	165	0.2799
1	1	6	310	0.5259
1	1	7	1053	1.7865
1	1	9	126	0.2138
1	1	10	197	0.3342
1	2	1	799	1.3556
1	2	2	192	0.3257
1	2	3	400	0.6786
1	2	4	209	0.3546
1	2	5	504	0.8551
1	2	6	1160	1.9681
1	2	7	2841	4.8201
1	2	8	346	0.5870
1	2	9	289	0.4903
1	2	10	238	0.4038
1	3	1	1088	1.8459
1	3	2	354	0.6006
1	3	3	648	1.0994
1	3	4	327	0.5548
1	3	5	624	1.0587
1	3	6	1131	1.9189
1	3	7	2916	4.9473
1	3	8	529	0.8975
1	3	9	397	0.6736
1	3	10	196	0.3325
1	4	1	703	1.1927
1	4	2	284	0.4818
1	4	3	420	0.7126
1	4	4	228	0.3868
1	4	5	469	0.7957
1	4	6	690	1.1707
1	4	7	1936	3.2846
1	4	8	482	0.8178
1	4	9	339	0.5752
1	4	10	155	0.2630
1	5	0	3362	5.7040
1	5	1	322	0.5463
1	5	2	138	0.2341
1	5	3	148	0.2511
1	5	4	85	0.1442
1	5	5	197	0.3342
1	5	6	283	0.4801
1	5	7	728	1.2351
1	5	8	343	0.5819
1	5	9	279	0.4734
1	5	10	86	0.1459
1	6	0	1611	2.7332
1	6	1	112	0.1900
1	6	7	200	0.3393
1	6	8	219	0.3716
1	6	9	415	0.7041
2	2	0	8001	13.5746
2	3	0	9513	16.1399
2	4	0	6607	11.2095

### 5) Resultados del árbol de decisión con los clusters edad-profesión como variable dependiente y prima, duración y capital como independientes.

dura2 IS ONE OF: 0 2 3

THEN

N : 30585  
1 : 64.0%  
2 : 36.0%

IF cap2 EQUALS 1  
AND prima2 EQUALS 5  
AND dura2 EQUALS 1

THEN

N : 1032  
1 : 62.3%  
2 : 37.7%

IF cap2 IS ONE OF: 0 2 3  
AND prima2 EQUALS 5  
AND dura2 EQUALS 1

THEN

N : 1879  
1 : 74.5%  
2 : 25.5%

IF prima2 EQUALS 2  
AND dura2 EQUALS 1

THEN

N : 2694  
1 : 57.7%  
2 : 42.3%

IF prima2 IS ONE OF: 0 1 3 4  
AND dura2 EQUALS 1

THEN

N : 22751  
1 : 51.2%  
2 : 48.8%

### 6) Gráfico de la evolución de pólizas por localidad (se observa el excesivo número de datos desconocidos en 2010)

