UNIVERSIDADE DA CORUÑA

Facultad de Economía y Empresa

Trabajo de Fin de Grado

Análisis Cluster. Un ejemplo aplicado a la identificación de perfiles de consumidores.

Jessica Martínez Bouza

Tutor: Prof. Dr. Xosé Manuel Martínez Filgueira

Grado en Administración y Dirección de Empresas

Año 2014

Trabajo de Fin de Grado presentado en la Facultad de Economía y Empresa de la Universidad de A Coruña para la obtención del Grado en Administración y Dirección de Empresas

RESUMEN

La Estadística se puede interpretar como un conjunto de herramientas destinadas a obtener información a partir de datos. En el presente trabajo se pretende mostrar, a través de una aplicación, su utilidad en áreas de interés para la empresa.

El tema elegido es el estudio de perfiles de consumo, en este caso aplicado al consumo de seguros, usando Análisis Cluster.

El enfoque incluye una parte teórica en la que se describen las técnicas estadísticas multivariantes, en general, y con más detalle el análisis cluster. Posteriormente voy a aplicar este tipo de análisis a un caso real, con un ejemplo donde se pretende conocer el perfil de los consumidores españoles frente a decisiones sobre el mercado de seguros privados.

Para realizar este estudio se tienen en cuenta factores económicos, sociológicos, y demográficos, utilizando la información procedente de la Encuesta de Presupuestos Familiares realizada por el INE en el año 2012. Todo esto permite demostrar que el análisis cluster es un instrumento de utilidad en la empresa ya que reúne una gran capacidad de procesamiento y análisis de la información y conlleva a la obtención de resultados muy interesantes.

Palabras clave: análisis multivariante, análisis cluster, perfil de consumidores, seguros.

Número de palabras: 9.332

ABSTRACT

Statistics can be interpreted as a set of tools designed to elicit information from data. In the present work aims to show, through an application, its usefulness in areas of interest to the company.

The theme chosen is the study of consumption profiles, in this case applied to the consumption of insurance, using Cluster Analysis.

The approach includes a theoretical part in describing the multivariate statistical techniques, in general, and with more detail the cluster analysis. Later i am going to apply this type of analysis to a real case, with an example where it is intended to know the profile of Spanish consumers compared to decisions on the private insurance market.

To perform this study takes into account factors economic, sociological, and demographic, using information from the Household Budget Survey carried out by the INE in the year 2012. All this allows you to prove that the cluster analysis is a useful instrument in the company as it brings together a great capacity for processing and analysis of information and leads to obtaining results very interesting.

Key words: multivariate analysis, cluster analysis, consumer profile, insurances.

Number of words: 9.332

ÍNDICE

INTRODUCCIÓN 5
1. MÉTODOS ESTADÍSTICOS MULTIVARIANTES Y ANÁLISIS CLUSTER 7
1.1. MÉTODOS ESTADÍSTICOS MULTIVARIANTES 7
1.2. ANÁLISIS CLUSTER 11
2. APLICACIÓN A UN CASO REAL: ANÁLISIS DEL PERFIL DE COMPRADORES DE SEGUROS 19
2.1. INTRODUCCIÓN 19
2.2. CONCEPTOS 20
2.2.1. Actividad aseguradora: empresa de seguros 20
2.2.2. Consumidor de seguros y comportamiento del consumidor 21
2.3. DATOS Y VARIABLES EMPLEADOS 22
2.3.1. Variables utilizadas 23
2.3.2. Análisis cluster 24
2.3.3. Explicación de los perfiles 26
2.4. RESULTADOS OBTENIDOS 27
2.4.1. Grupos a considerar 27
2.4.2. Interpretación de los grupos 27
2.4.3. Perfiles de los grupos en función de las variables explicativas 29
2.4.4. Síntesis 31
CONCLUSIONES 32
BIBLIOGRAFÍA 34
ANEXOS 37
ANEXO A. VARIABLES EXPLICATIVAS 37
ANEXO B. DENDOGRAMAS DE LAS COMUNIDADES AUTÓNOMAS 40
ANEXO C. CÁLCULOS CLÚSTERES ESPAÑA 43
ANEXO D. METODOLOGÍA Y TABLAS RESUMEN VARIABLES EXPLICATIVAS 47
ANEXO E. FRECUENCIAS RELATIVAS, CONTRASTE CHI CUADRADO Y DIAGRAMAS DE ASOCIACIÓN DE LAS VARIABLES EXPLICATIVAS 54

INTRODUCCIÓN

En el mundo de la empresa es importante analizar la información para lograr una adecuada planeación y control apoyándose en los estudios de pronósticos, presupuestos, etc. Esto da lugar a que la correcta aplicación de la metodología en la investigación empresarial precise disponer de herramientas auxiliares para recoger, organizar, presentar, analizar e interpretar datos.

La Estadística ofrece una gran variedad de herramientas analíticas en la toma de decisiones bajo incertidumbre (Mendenhall y Reinmuth, 1978). Según Peña (2001), esta ciencia es el resultado de la unión de dos disciplinas que evolucionan independientemente hasta coincidir en el s. XIX: la primera es el cálculo de probabilidades y la segunda es la "estadística" que estudia la descripción de datos. La integración de ambas líneas de pensamiento da lugar a una ciencia que estudia cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos.

Muchos autores definieron la Estadística como un instrumento útil para relacionar la teoría y la práctica (Sarabia y Pascual, 2005). El propio Fisher definió la Estadística como "la Matemática de los datos observacionales". Otra definición es la aportada por Kendall y Stuart (1979), quienes afirman que "la Estadística es la rama de los métodos científicos que trata los datos obtenidos contando o midiendo las propiedades de poblaciones de fenómenos naturales".

En el presente trabajo se pretende mostrar, a través de un ejemplo, la importancia del análisis de la información para la empresa y del interés de la Estadística como técnica para obtener esa información a partir de datos. El método estadístico que voy a utilizar es el análisis cluster, el cual forma parte del análisis multivariante y me va a permitir obtener información acerca de los perfiles de consumidores en un caso concreto. Con relación a esto, Hair, Anderson, Tatham y Black (1999) afirman:

> Actualmente los directivos no pueden fiarse de las antiguas aproximaciones donde se consideraban consumidores homogéneos y caracterizados por un reducido número de variables demográficas. En su lugar, deben desarrollar estrategias para atraer a numerosos segmentos

> de clientes con diversas características demográficas y psicográficas en un mercado con múltiples restricciones (legales, económicas, competitivas, tecnológicas, etc). Sólo a través de las técnicas de análisis multivariante se pueden examinar adecuadamente las relaciones múltiples de este tipo para llegar a una comprensión de la toma de decisiones más completa y realista. (p.2)

Teniendo en cuenta lo expuesto hasta ahora, con este trabajo pretendo mostrar el interés que tienen las técnicas estadísticas en el trabajo relacionado con nuestro perfil profesional. Para ello se van a describir este tipo de técnicas en general, dando más detalle al análisis cluster, técnica que se utilizará posteriormente en mi ejemplo de determinación del perfil de consumidores de seguros, basado en el artículo “El perfil del consumidor en el sector asegurador español” de Albarrán y De Pablos (2001).

La estructura de este trabajo está divida en dos grandes apartados: en el primero se aborda la decisión teórica de los métodos multivariantes, dando una visión general y otra particular del análisis cluster; esta descripción está basada en gran medida en el libro escrito por Hair et al. (1999). En el segundo apartado se realiza la aplicación práctica del análisis cluster, comenzando con la descripción del artículo usado como referencia y de los conceptos relacionados con el mundo de los seguros; a continuación se describe la metodología empleada, tanto en el propio análisis cluster como en la interpretación de sus resultados y concluye con la descripción de los propios resultados del análisis.

Finalmente, se expondrán unas conclusiones sobre todo lo expuesto que sinteticen todo el trabajo realizado.

1. MÉTODOS ESTADÍSTICOS MULTIVARIANTES Y ANÁLISIS CLUSTER

1.1. MÉTODOS ESTADÍSTICOS MULTIVARIANTES:

El análisis multivariante comprende una serie de métodos estadísticos para realizar el tratamiento conjunto de datos relativos a diversas variables. Para ser considerado verdaderamente multivariante, todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido (Hair et al., 1999). Su razón de ser radica en un mejor entendimiento del fenómeno objeto de estudio obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir. Esta capacidad de procesar y analizar información convierte a estos métodos en un instrumento de utilidad en todos aquellos lugares en los que se necesita analizar información como por ejemplo en el mundo de la empresa.

Algunos de los métodos del análisis multivariante son métodos puramente descriptivos, que realizan un estudio de los datos muestrales exclusivamente. En otros métodos, sin embargo, se trata de realizar inferencias acerca de parámetros poblacionales. Una clasificación de las técnicas multivariantes es la que tiene en cuenta la naturaleza de las variables dependientes e independientes que se utilizan en cada método, así como el objetivo que se persigue. Las variables dependientes son aquellas que constituyen el núcleo central de la investigación, mientras que las variables independientes son variables que se utilizan para explicar el comportamiento de las variables dependientes (Uriel, 1995).

Entre las técnicas más conocidas, Uriel (1995) destaca las siguientes:

- Análisis multivariante de la varianza: es la generalización del análisis de la varianza al caso de que se trate de determinar la influencia de uno o más factores sobre más de una variable dependiente.

- Regresión multivariante: es una generalización del modelo de regresión lineal en la cual se cuantifica la influencia que ejercen variables de distinto tipo sobre una variable dependiente de carácter continuo.

- Análisis de correlación canónica: este análisis es también una generalización del modelo de regresión lineal y establece la interdependencia entre dos conjuntos de

variables, ambos con variables métricas. Para hacerlo busca combinaciones lineales de variables explicadas y de variables explicativas y busca relacionarlas de forma que se maximice la correlación entre ellas.

- Análisis discriminante: se aplica para caracterizar mediante un conjunto de variables independientes las diferencias existentes entre distintos grupos y también para clasificar observaciones cuando se desconoce el grupo al que pertenecen. En este caso se considera que la variable dependiente es de naturaleza categórica, y que cada una de sus categorías indica los elementos de un grupo diferente.

- Modelos logit binomial y multinomial: estos métodos tienen, en su objetivo, una gran similitud con los métodos de análisis discriminante pero, en un planteamiento, es una extensión de la regresión lineal a variables dependientes categóricas. Las variables independientes pueden ser continuas, discretas o categóricas.

- Análisis de componentes principales: es una técnica de reducción de datos que trata de transformar un conjunto de variables en otro conjunto de menor número de variables, con la particularidad de que las nuevas variables están incorrelacionadas entre sí. Cuando se aplica esta técnica no se formula ningún modelo teórico.

- Análisis factorial: su finalidad es también la de reducción de datos. Sin embargo, a diferencia del análisis de componentes principales, se formula un modelo teórico en el cual se explica el comportamiento de variables observables mediante factores comunes y factores únicos no observables que se obtienen en el proceso de análisis de datos.

- Análisis de correspondencias: es un método de reducción de datos aplicable a variables categóricas.

- Escalas multidimensionales: son un conjunto de técnicas que utilizan las proximidades entre objetos para realizar una representación espacial de los mismos. Es decir, el objetivo de este método es la agrupación.

- Análisis cluster: es una técnica de agrupación. Su objetivo es la partición de un conjunto de objetos o individuos en grupos tales que los objetos pertenecientes a un mismo grupo son muy similares entre sí pero muy diferentes a los objetos pertenecientes a otros grupos.

Para los métodos multivariantes, Kendall (1980) establece una clasificación en la que pone el acento en si las técnicas se basan en relaciones de dependencia entre las variables establecidas a priori, o bien, si se basan en relaciones de interdependencia no presupuestas a priori:

- Las técnicas basadas en relaciones de dependencia establecen a priori una distinción entre una o más variables dependientes o endógenas a explicar, y otras variables que utilizaremos para explicar las primeras, llamadas independientes, exógenas o predictivas. Entre estas técnicas destacan la regresión múltiple, con una variable dependiente cuantitativa; el análisis discriminante, con una única variable dependiente cualitativa; el análisis multivariante de la varianza, con varias variables dependientes cuantitativas; o el análisis canónico, con varias variables dependientes cualitativas.

- Las técnicas basadas en relaciones de interdependencia no establecen ninguna distinción a priori entre variables y su objetivo principal es organizar los datos de forma que sean más manejables y comprensibles. Entre ellas podemos destacar el análisis factorial, el análisis cluster o el escalamiento multidimensional.

Los diferentes métodos que constituyen el análisis de dependencia pueden ser a su vez divididos en dos tipos según el número de variables dependientes y el tipo de escalas de medida empleadas para las variables.

Según Hair et al. (1999), el análisis de dependencia puede incluso ser clasificado en función del tipo de escala de la variable con variables métricas (numéricas/cuantitativas) o no métricas (cualitativas/categóricas). En lo que sigue se van a indicar algunos ejemplos de aplicación de métodos multivariantes teniendo en cuenta las características métricas de las variables utilizadas. Si el análisis implica una única variable dependiente que es métrica, la técnica apropiada es el análisis de regresión múltiple. Por otro lado, si la única variable dependiente es no métrica (categórica), se podría aplicar el análisis discriminante múltiple o el modelo logit. En contraste, cuando el problema del investigador implica varias variables dependientes, hay otras técnicas estadísticas apropiadas. En el caso de trabajar con variables dependientes métricas ocurre lo siguiente: si las independientes son no métricas se puede aplicar el análisis de la varianza y si las independientes son métricas se aplica la correlación canónica. Si varias variables dependientes son no métricas, entonces pueden transformarse a través de una variable ficticia de código (0-1) y puede

utilizarse también el análisis de correlación canónica. Finalmente, con variables dependientes e independientes no métricas, tiene sentido usar el análisis de correspondencias.

Existe una estrecha relación entre los diversos procedimientos de dependencia, que pueden ser vistos como una familia de técnicas. El análisis canónico puede considerarse como el modelo general en el cual se basan otras muchas técnicas multivariantes, dado que sitúa la mínima restricción respecto al tipo y número de variables tanto de valor teórico dependiente como independiente (Hair et al., 1999).

Con respeto a los métodos de análisis de interdependencia, todos ellos tienen en común que las variables son analizadas simultáneamente en un esfuerzo por encontrar una estructura subyacente para el conjunto total de variables o sujetos. Hair et al. (1999) señalan lo siguiente:

- Si se está analizando la estructura de las variables, entonces el análisis factorial es la técnica apropiada.

- Si los casos o los encuestados se van a agrupar para representar una estructura, entonces seleccionaremos el análisis cluster.

- Si el interés está en la estructura de objetos, deberían aplicarse las técnicas de análisis multidimensional.

Además también señalan que, al igual que ocurre con el análisis de dependencia, deberían considerarse las propiedades de las técnicas de medición. Generalmente, el análisis factorial y el análisis cluster se consideran análisis de interdependencia métricos. Sin embargo, los datos no métricos pueden ser transformados a través de una variable ficticia codificada para usarlos con análisis factorial y análisis cluster. Se han desarrollado tanto las aproximaciones métricas como las no métricas al análisis multidimensional. Si se van a analizar las interdependencias entre objetos medidos por datos no métricos, el análisis de correspondencias es la técnica apropiada.

1.2. ANÁLISIS CLUSTER:

Para mostrar el interés de las técnicas estadísticas multivariantes se va a realizar en este trabajo un ejemplo de cómo se aplica una de ellas. El método escogido es el análisis cluster, aplicado con el objetivo de buscar los perfiles de consumidores de un grupo de productos. Por esta razón, se van a describir con más detalle las características de dicha técnica.

Como ya hemos visto, el análisis cluster es un método multivariante que pertenece al análisis de interdependencia. El principal propósito del análisis cluster es agrupar objetos (es decir, encuestados, productos u otras entidades) basándose en las características que poseen; se trata de clasificar a los individuos y a los objetos en conglomerados, de tal forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado. Los conglomerados de objetos resultantes deberían mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados). Por tanto, si la clasificación es acertada, los objetos dentro de los conglomerados estarán muy próximos cuando se representen gráficamente, y los diferentes grupos estarán muy alejados. (Hair et al., 1999).

Pérez (2004), utiliza otros términos asignados al concepto de análisis cluster: análisis de conglomerados, análisis tipológico y clasificación automática. Todos ellos pueden funcionar como sinónimos. Para Sokal y Sneath (1963), dos de los autores que más han influido en el desarrollo del análisis cluster, la clasificación es uno de los procesos fundamentales de la ciencia, ya que los fenómenos deben ser ordenados para que podamos entenderlos. Tanto el análisis cluster como el análisis discriminante sirven para clasificar individuos en categorías; la principal diferencia entre ambos radica en que en el análisis discriminante se conoce a priori el grupo de pertenencia, mientras que el análisis cluster sirve para ir formando grupos homogéneos de conglomerados.

Según Hair et al. (1999), la intención fundamental del análisis cluster es clasificar individuos, lo que también se denomina hacer una taxonomía, la cual consiste en ver que grupos se producen automáticamente en función de las relaciones entre individuos. Pero, al hacer esta clasificación también aparecen otras utilidades; una de ellas es la de simplificación o reducción de los datos ya que en muestras grandes, al agrupar a los individuos más similares, nos va a permitir tener una visión

más clara de los elementos que forman la muestra, algo complejo de ver cuando el tamaño de la muestra es muy grande. El análisis cluster también nos proporciona una manera de identificar relaciones entre individuos y, al hacer el agrupamiento, es más fácil encontrar dichas relaciones.

El análisis cluster es muy fácil de aplicar, dado que en su planteamiento clásico no necesita supuestos para la distribución de las variables, dando lugar a un análisis meramente descriptivo. Sin embargo, esta misma simplicidad tiene sus inconvenientes. Uno de ellos es que no permite hacer inferencia estadística porque no se parte de unos supuestos previos. Además, este análisis siempre proporciona una solución, independientemente de que en la realidad esa agrupación exista o no, por lo que podríamos obtener unos grupos creados que no son más que una consecuencia de un artificio matemático. Otro inconveniente es que la solución obtenida depende en gran medida del método y de otros factores que debe decidir el investigador, lo que lo hace un poco inestable y hace aconsejable que se comprueben varios agrupamientos para ver hasta que punto se introducen diferencias (De la Fuente, 2011).

A pesar de no establecerse hipótesis previas para la aplicación del análisis cluster, es necesario hacer unos análisis previos para detectar problemas que puedan afectar a los resultados. Hair et al. (1999) consideran importante seguir una serie de pasos:

- Lo primero a tener en cuenta es la selección de las variables: las variables con las que se trabaje deben ser apropiadas para el problema analizado dado que la elección de una variable no relevante podría dirigir la dirección de la agrupación y dar lugar a resultados que se aparten del objetivo de análisis. Por lo tanto, el investigador debe darse cuenta de incluir sólo aquellas variables que caracterizan los objetos que se están agrupando y que se refieran específicamente a los objetivos del análisis cluster; esta técnica no tiene un medio para diferenciar las variables relevantes de las irrelevantes.

- La multicolinealidad es otro aspecto importante debido a que actúa reforzando el efecto de las variables relacionadas; por esta razón, es conveniente comprobar que no tenemos variables muy relacionadas. En caso de detectar la multicolinealidad, se puede actuar de varias maneras: eliminando variables redundantes, reduciendo el número de variables mediante análisis factorial o componentes principales, o utilizando una medida que compense los efectos de la correlación (distancia de Mahalanobis).

- La existencia de datos anómalos también puede interferir en la agrupación, por lo que es interesante detectarlos a priori o, en tal caso, tenerlos en cuenta a la hora de analizar los resultados. Dichos datos distorsionan la verdadera estructura y hacen que los conglomerados deducidos no sean representativos de la verdadera estructura de la población. Una manera de tratar los anómalos pasa por realizar el análisis con ellos y sin ellos y decidir a posteriori, una vez sabemos cuáles son sus efectos sobre dicho análisis. El análisis cluster ayuda a identificar los datos anómalos dado que, si trabajamos con muchas variables, es difícil detectar anómalos globales a simple vista; sin embargo, aparecen al realizar las agrupaciones porque les resulta difícil encajar en los grupos que se van formando.

- Por último, también debemos tener en cuenta la tipificación puesto que la dimensión de los datos puede afectar al análisis; una variable con valores de la orden de miles tendrá más influencia en el grupo que otra de la orden de las unidades (si se homogenizan se igualaría su importancia). Sin embargo, el proceso de tipificación puede ocultar la estructura subyacente de los grupos, lo que implica que no debemos hacer la tipificación de manera automática, sino observando cada situación.

Existen diferentes aspectos que debemos recordar a la hora de aplicar el análisis cluster, ya que según nos decidamos por hacerlo de una forma o de otra podemos tener diferencias en los resultados. Los aspectos a considerar según Hair et al. (1999) son:

- ¿Cómo medir las semejanzas entre los objetos? ¿Cómo relacionar observaciones con los grupos y los grupos entre sí? Es decir, qué medida utilizar para ver qué objetos están cercanos y poder agruparlos. A medida que vamos agrupando individuos, el algoritmo deberá trabajar con los grupos y con los individuos o sólo con los grupos. Existen tres métodos para medir la similitud: medidas de correlación, medidas de distancia y medidas de asociación.

- ¿Cuántos grupos obtener? Hay que decidir cuántos grupos se formarán finalmente. El número de grupos a formarse dependerá del método de agrupación elegido (jerárquico o no jerárquico) como veremos posteriormente. Esto puede hacerse analizando los resultados o decidiéndolo a priori en función de la información que tengamos. No existe un procedimiento objetivo o estándar, por lo que los investigadores han desarrollado varios criterios y han llegado a diferentes conclusiones sobre posibles reglas de parada:

- Examinar alguna medida de similitud o distancia entre los conglomerados a cada paso sucesivo, donde la solución cluster se define cuando la medida de similitud excede a un valor especificado o cuando los valores sucesivos entre los pasos dan un salto súbito.

- Aplicar alguna regla estadística o adaptar un test estadístico, tal como las correlaciones "point-biseral/tau" o el ratio de verosimilitud.

- Complementar el juicio estrictamente empírico con cualquier conceptualización de las relaciones teóricas que pueda sugerir un número natural de conglomerados.

El análisis cluster se empieza midiendo la semejanza (o diferencia) entre las observaciones. De hecho, los métodos de análisis cluster se consideran que trabajan más con la matriz de distancia que con las observaciones; pero, las distancias que se utilizan dependen del tipo de datos que se manejen. Pérez (2004) diferencia los siguientes tipos de distancias:

- Para datos numéricos, es decir, de escala de intervalo, se suele utilizar la distancia euclídea (distancia ordinaria medida con una regla entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras) o la distancia de Mahalanobis (distancia al cuadrado entre los centroides de dos poblaciones). Un caso particular es el de los coeficientes de correlación de Pearson y Sperman, ya que no son una distancia propiamente dicha, sino que actúan como una medida de la relación entre los valores de los individuos.

- Para variables binarias, variables que sólo toman dos valores (que se pueden denominar como 0 y 1 como ausencia y presencia de una característica) existe una infinidad de medidas; para construirlas, se considera una tabla en la que denominan a, b, c y d las diferentes combinaciones de presente y ausente de una variable en los dos individuos a los que queremos medir la cercanía. De entre estas medidas, Lévy y Varela (2003) consideran que la concordancia simple es la medida más elemental que se puede aplicar, ya que mide cuantas veces coinciden los valores de las variables en dos observaciones.

Lévy y Varela (2003) contemplan además otras medidas:

- Para variables nominales (pueden poseer más de dos valores), ordinales (los diferente valores de la variable se encuentran ordenados en una secuencia) o de razón (cuentan con un cero absoluto por debajo del cual no existen medidas de esa variable) resulta más práctico adaptarlas a los casos anteriores que diseñar medidas específicas para ellas.

- Si se combinan diferentes tipos de variables en el mismo estudio, se pueden tratar de varias maneras: en primer lugar podemos homogeneizar dichas variables, haciendo las transformaciones necesarias para que todas las variables sean del mismo tipo. Otra posibilidad es hacer análisis diferentes, agrupando las variables del mismo tipo o también se puede optar por transformar datos cuantitativos en cualitativos. Una alternativa sería aplicar el coeficiente de similitud de Gower, que es una medida diseñada precisamente para este tipo de situaciones.

El siguiente paso es decidir el método o tipo de agrupamiento a utilizar. Según Lévy y Varela (2003) dicha decisión depende del tipo de escala de medida de nuestras variables y del propósito particular de nuestro estudio. Hay dos enfoques: métodos jerárquicos y métodos no jerárquicos.

- El procedimiento jerárquico consiste en hacer agrupaciones individuales en cada paso, ya sea comenzando con todos los individuos separados y agrupando los más cercanos hasta que todos estén agrupados en uno (algoritmo de aglomeración) o al revés, todos los individuos en un único grupo que se va separando en grupos más pequeños, hasta que todos los individuos estén de nuevo separados (algoritmos divisivos). Peña (2002) sugiere que los algoritmos de aglomeración requieren menos tiempo de cálculo y son los más utilizados. El esquema de agrupamiento de los clústeres jerárquicos se observa bien con el dendograma[1], que es un gráfico en forma de árbol en el que se representan las agrupaciones de individuos, los cuales aparecen separados en un extremo, saliendo una línea de cada uno de ellos. Cada agrupación se indica uniendo las líneas de los individuos agrupados, y esa fusión aparece a una altura que indica la distancia que existe entre ellos. Continuamos sucesivamente hasta que queda una única línea, lo cual indicará que se realizaron todas las agrupaciones.

[1] Se pueden ver ejemplos de dendogramas en el anexo B

- En el procedimiento no jerárquico hay que decidir previamente el número de grupos que se quiere formar, y el algoritmo de agrupamiento irá asignando los individuos a cada uno de los grupos que se pretende formar. Hair et al. (1999) señalan que el esquema más utilizado es el de las k-medias. Su aplicación empieza definiendo las semillas del conglomerado, es decir, unos puntos tantos como grupos se decidiesen formar, alrededor de los cuales se irán creando los conglomerados. Cuando esté realizada la agrupación para todos los individuos, se calcula el centroide para cada uno de ellos y se compara con las semillas. Si la diferencia es inferior a una cuota entonces se para y los clústeres formados serán los grupos que se obtengan en esa iteración; si, por lo contrario, la diferencia es mayor que la cuota, los centroides pasan a ser las nuevas semillas y se repite el proceso.

Una vez tenemos elegido el tipo de agrupamiento a utilizar, se van empezando a organizar los grupos, por lo que debemos medir la semejanza o cercanía entre puntos y grupos o entre grupos. Ahora bien, hay diversas maneras de medir la cercanía entre grupos ya que depende de si estamos trabajando con el método jerárquico o con el no jerárquico. Según Hair et al. (1999), si trabajamos con el método jerárquico, los algoritmos más habituales son los siguientes:

- El encadenamiento simple (método single): este procedimiento se basa en la distancia mínima. Encuentra a los dos objetos separados por la distancia más corta y las coloca en el primer conglomerado. Posteriormente se encuentra la distancia más corta, y o bien un tercer objeto se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros.

- El encadenamiento completo: esta técnica se basa en la distancia máxima entre individuos de cada conglomerado, la cual representa la esfera más reducida que puede incluir todos los objetos en ambos conglomerados.

- El encadenamiento medio (método average): el criterio de aglomeración de este encadenamiento es la distancia media de todos los individuos de un conglomerado con todos los individuos de otro. Estas técnicas no dependen de los valores extremos al igual que en el encadenamiento simple o completo, sino que la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos.

- El método de Ward: en este método la distancia entre dos conglomerados es la suma de los cuadrados entre dos conglomerados sumados para todas las variables. En cada paso del procedimiento de aglomeración, se minimiza la suma de los cuadrados dentro del conglomerado para todas las particiones (el conjunto completo de conglomerados disjuntos o separados) obtenida mediante la combinación de dos conglomerados con aproximadamente el mismo número de observaciones.

- El método del centroide: según esta técnica, la distancia entre los dos conglomerados es la distancia entre sus centroides. Los centroides de los grupos son los valores medios de las observaciones de las variables en el valor teórico del conglomerado. En este método, cada vez que se agrupa a los individuos, se calcula un nuevo centroide. Se aplica generalmente sólo con variables continuas (Peña, 2002).

Para Bisquerra (1989) ninguno de los métodos anteriores proporciona una solución óptima a todos los problemas debido a algunas indeterminaciones que constituyen los fundamentos del análisis cluster (recordemos que es posible llegar a distintos resultados según el método elegido). El buen criterio del investigador, el conocimiento del problema y la experiencia profesional sugerirán el método más adecuado.

Por otra parte, los procedimientos no jerarquizados utilizan, según Hair et al. (1999), una de las siguientes aproximaciones para asignar las observaciones individuales de uno de los conglomerados:

- Umbral secuencial: este método empieza seleccionando una semilla de conglomerado e incluye todos los objetos que caen dentro de una distancia previamente especificada. Cuando todos los objetos dentro de la distancia están incluidos, se selecciona una segunda semilla y se incluyen todos los objetos dentro de la distancia previamente especificada. A continuación, se selecciona una tercera semilla y el proceso continúa como se ha descrito.

- Umbral paralelo: este umbral selecciona varias semillas de conglomerado simultáneamente al principio y asigna objetos dentro de la distancia umbral hasta la semilla más cercana. A medida que el proceso avanza, se puede ajustar las distancias umbral para incluir más o menos objetos en los conglomerados.

- Optimización: esta práctica permite la reubicación de los objetos. Si, en el curso de la asignación de los objetos, un objeto se acerca más a otro conglomerado que no es el que tiene asignado en este momento, entonces un procedimiento de optimización cambia el objeto al conglomerado más parecido (cercano).

Según Hair et al., (1999), la regla de parada en cuanto a la formación de grupos depende del método que se haya elegido:

- En las agrupaciones jerárquicas se usan algunos contrastes que indican cuando la diferencia es significativa como para parar. Sin embargo, también existen métodos que se basan en informaciones descriptivas y, dado que es muy frecuente que el planteamiento en el análisis cluster sea descriptivo, serán los que comentemos. La idea para decidir la parada se basa en la distancia que hay entre cada agrupamiento; en el momento en el que se observa que ésta es suficientemente grande, o que se produce un salto importante en su incremento, entonces tendremos un motivo para pensar que los grupos son suficientemente diferentes como para que no sea razonable seguir haciendo el agrupamiento. Estas diferencias entre las distancias de agrupamiento se pueden apreciar en el dendograma o analizando directamente sus valores. En la práctica no siempre está claro que exista ese punto donde las diferencias aconsejan parar, o incluso podemos dudar entre varios; por esta razón, también se tiene en cuenta la interpretación de los resultados, es decir, que el resultado obtenido tenga un sentido y de una interpretación razonable.

- En los métodos no jerárquicos la regla de parada es una decisión previa. El algoritmo empieza con los grupos ya decididos. Esta decisión está basada en la información de la que se dispone para realizar el análisis o también se puede realizar una agrupación previa mediante un cluster jerárquico, que proporciona tanto el número de grupos como las semillas de los conglomerados, las cuales se convierten en los centroides del cluster previo.

En el análisis cluster es muy importante la interpretación de los datos, ya que siempre se obtiene un resultado numérico pero éste no tiene porqué tener sentido, por lo que interesa averiguar su significado antes de poder aceptarlo como una agrupación correcta. Para ello, podemos hacer una representación gráfica de las variables utilizadas o diferentes análisis con variables explicativas como un análisis de la varianza o un análisis discriminante (Hair et al., 1999).

La última parte del análisis es la validación de los resultados. Se trata de comprobar la estabilidad de los resultados. Con respecto a esto, en Cuadras (2014) se indica una crítica al análisis cluster, que es el excesivo repertorio de distancias y métodos de clasificación, lo cual puede producir clasificaciones dispares para los mismos datos.

2. APLICACIÓN A UN CASO REAL: ANÁLISIS DEL PERFIL DE COMPRADORES DE SEGUROS

2.1. INTRODUCCIÓN:

En este apartado voy a aplicar el análisis cluster a una situación real. Para ello me voy a basar en el artículo “El perfil del consumidor en el sector asegurador español” elaborado por Albarrán y De Pablos (2001) en la Revista de Investigación Operacional. En este artículo se busca conocer el comportamiento de los hogares españoles frente a decisiones sobre el mercado de seguros privados tales como la suscripción o no de una determinada póliza, el gasto efectuado en la misma, número de distintas pólizas contratadas, etc.; para ello se tienen en cuenta factores de tipo económico, sociológico, demográfico y psicológico, lo cual ayuda a definir un perfil del tipo de consumidor español de seguros. Los seguros que se estudian en este artículo son los siguientes: vida, salud, multirriesgo hogar, automóvil y otros (seguro de viaje, responsabilidad civil del cazador…) y las comunidades autónomas analizadas son: Andalucía, Castilla y León, País Vasco, Cataluña, Navarra, Madrid, Canarias y Baleares.

Para llevar a cabo este estudio se utiliza la información procedente de la Encuesta Básica de Presupuestos Familiares (EBPF) del año 1990 realizada por el INE en España.

La metodología utilizada es un análisis cluster con un estudio descriptivo previo. Las variables explicativas de los perfiles se separan según sean características propias del sustentador principal (actividad profesional, categoría socio-profesional, condición socioeconómica, edad, nivel de estudios y sexo), características económicas del hogar (posesión o no de un automóvil, régimen de tenencia de la vivienda, superficie total, superficie útil, número de perceptores de ingresos, nivel de ingresos y fuente principal de ingresos) y características sociales y demográficas del hogar

(provincia, comunidad autónoma, tamaño del municipio, estrato, número de miembros y tipo de hogar).

En cuanto a las variables dependientes, se ha tenido en cuenta el gasto en cada uno de los tipos de seguro estudiados y el gasto total, el número de pólizas contratadas de cada tipo de seguro por el hogar y, por último, variables dicotómicas que expresan la tenencia o no de alguna modalidad de los cinco ramos analizados.

En mi caso el objetivo es similar, es decir, conocer el perfil de los hogares españoles frente a decisiones sobre el mercado de seguros (seguros de vivienda, sanidad, transporte, civil y enterramiento). Se emplean variables semejantes pero en algunos casos las he modificado debido a la diferente información contenida en los datos empleados que proceden de la Encuesta de Presupuestos Familiares del año 2012.

2.2. CONCEPTOS:

2.2.1. Actividad aseguradora: empresa de seguros:

Se considera entidad, compañía o empresa de seguros (en sentido estricto) a la empresa o sociedad dedicada a la práctica del seguro. Algunas de sus características esenciales, según Castelo y Guardiola (1992) son:

- Exclusividad de actuación: la entidad está dedicada a la práctica de operaciones de seguro y actividades relacionadas.

- Sometimiento a normas de vigilancia oficial: el carácter social y público de la actividad aseguradora y el establecimiento por la empresa aseguradora de las condiciones técnicas (tarifas) económicas (primas) y documentales (pólizas) que regirán las relaciones contractuales entre ella y el asegurado o los Organismos oficiales, es lo que justifica que se institucionalice una especial vigilancia técnica, económica y financiera.

- Operaciones en masa: las entidades de seguros tratan de conseguir el mayor número posible de asegurados para compensar y diversificar riesgos.

- Exigencias legales: las exigencias están materializadas en capitales mínimos iniciales y otras garantías financieras.

La actividad aseguradora forma parte de la actividad económica y mercantil de los países. Posee unas características específicas, peculiares y complejas relacionadas con la prestación de servicios que surgen debido a la existencia de riesgos económicos que afectan a los individuos, a las empresas y a la sociedad en general (Albarrán, 1999a).

El seguro constituye la forma más perfecta y técnicamente eficaz para la cobertura de riesgos ya que transforma los individuales en colectivos y el asegurador los transfiere a una organización estructurada con la técnica y operativa adecuadas para garantizar su compensación, en caso de ocurrir el evento (Eurostat, 1988).

En resumen, una entidad aseguradora es una sociedad dedicada a la práctica del seguro de forma exclusiva, sometida fuertemente a la normativa legal, y con una actividad propia de servicios que tiene invertido su proceso productivo (primero se cobra para posteriormente ofrecer el servicio en el caso de que acontezca el riesgo).

2.2.2. Consumidor de seguros y comportamiento del consumidor:

La condición fundamental para perfilar el concepto de consumidor es la de ser destinatario final de un producto, actividad o servicio. A su vez, la directiva 93/13/CEE, confirmando esta idea y matizándola más, considera consumidor a “toda persona física que actúe con un propósito ajeno a su actividad profesional”.

Según Albarrán (1999b), el consumidor de seguros es quien contrata los servicios de cobertura de riesgos y, eventualmente, las prestaciones indemnizatorias que proporcionan las empresas aseguradoras. Pueden ser, tanto el tomador del seguro (suscriptor de la póliza) como el asegurado, e incluso, se puede extender al tercer beneficiario (caso frecuente en los seguros personales) o al tercer perjudicado (persona no participante en el contrato, a la cual se le causa un daño, cuyo riesgo es objeto de cobertura de un seguro de responsabilidad civil y, por tanto, tiene derecho a que se resarzan las consecuencias de dicho daño).

Es comprador de seguros, según Castelo y Guardiola (1992), en sentido estricto, la persona que en sí misma o en sus bienes o intereses económicos está expuesta al riesgo; también añaden que, en sentido amplio, asegurado es quien suscribe la póliza con la entidad comprometiéndose al pago de las primas estipuladas y teniendo derecho al cobro de las indemnizaciones que se produzcan como consecuencia de un siniestro. En la práctica, el término asegurado engloba los

conceptos de tomador, asegurado, beneficiario y tercer perjudicado. Todos son consumidores, en sentido material, al utilizar o disfrutar del servicio de cobertura de riesgos (Sánchez, 1981).

Embid, Martin y Zorrilla (1998) entienden por comportamiento de compra del consumidor aquel por el que, mediante un proceso racional o irracional, selecciona, compra, usa y dispone de productos o servicios para satisfacer sus necesidades y deseos.

En este trabajo se analiza el comportamiento individual como consumidor final. Los factores que influyen en la demanda de los seguros privados y de los aspectos determinantes en los hábitos de compra y modificación del comportamiento de los consumidores de seguros son, en gran medida, aspectos demográficos, económicos o socioculturales.

2.3. DATOS Y VARIABLES EMPLEADOS:

Debido a la dificultad de obtener información sobre el gasto en seguros, la fuente estadística de la que he extraído los datos utilizados en el análisis empírico ha sido la Encuesta de Presupuestos Familiares (EPF) del año 2012, cuyos datos están recogidos en el Instituto Nacional de Estadística (INE). Cabe mencionar las limitaciones a las que se enfrenta todo el análisis de estas características como pueden ser la falta de información y la posibilidad de error.

La Encuesta de Presupuestos Familiares es una de las encuestas más antiguas de las que realiza el Instituto Nacional de Estadística, con el objetivo de obtener información sobre la naturaleza y destino de los gastos de consumo, así como sobre diversas características relativas a las condiciones de vida de los hogares. Los gastos de consumo que se registran en dicha encuesta se refieren tanto al flujo monetario que destina el hogar al pago de determinados bienes y servicios de consumo final, como al valor de determinados consumos no monetarios efectuados por los hogares.

Dicha encuesta sustituye a la Encuesta Continua de Presupuestos Familiares (ECPF) que estuvo en vigor desde el año 1997 hasta 2005. Obviamente, la EPF ha evolucionado en diversos aspectos como por ejemplo en el tipo de población considerada, en el tamaño de la muestra (hasta 24.000 hogares), en el nivel de desagregación del gasto, en el sistema de recogida o diseño de cuestionarios, e

incluso ha adoptado distintas formas en lo que a su periodicidad se refiere (de trimestral a anual).

Debido a las grandes posibilidades que ofrece la información proveniente de las Encuestas de Presupuestos Familiares (EPF) para su utilización por parte de una gran diversidad de usuarios, estas encuestas son consideradas tradicionalmente como encuestas multiobjetivo. Entre estos objetivos se consideran los siguientes:

- Poner a disposición de los investigadores y del sistema de indicadores sociales en general, datos estadísticos sobre distintos campos de preocupación social (equipamiento, vivienda, nutrición, sanidad, enseñanza, turismo).

- Obtención de un conjunto de variables: distribución de hogares o de personas según determinadas variables de clasificación (sexo, edad, nivel de estudios, etc.) o tamaño medio de los hogares.

Se trata de la única encuesta pública a escala nacional que dispone de información sobre el gasto en seguros de forma desagregada en dos aspectos: respecto al asegurado (real o potencial), tomando al hogar como la unidad básica de cómputo y también respecto al grado de agrupación por ramos y modalidades, aunque, en este aspecto, la clasificación y diferenciación de los mismos no sea muy rigurosa.

2.3.1. Variables utilizadas:

Las variables explicativas que voy a utilizar en este trabajo resumen las características demográficas del hogar y los factores económicos y socioculturales en función de los cuales se modeliza el comportamiento de los clientes de seguros. Dichas variables son las siguientes:

- Características propias del sustentador principal: situación profesional, situación socioeconómica, condición socioeconómica, edad, nivel de estudios y sexo.

- Características económicas del hogar: régimen de tenencia de la vivienda, superficie útil, número de personas que ingresan, nivel de ingresos y fuente principal de ingresos.

- Características sociales y demográficas del hogar: comunidad autónoma, tamaño del municipio, número de miembros de la familia y número de miembros ocupados.

Las variables dependientes consideradas en mi análisis están relacionadas con el consumo de seguros privados. Como ya dijimos anteriormente, voy a estudiar los gastos en seguros de vivienda, sanidad, transporte, civil y enterramiento. Conviene tener muy claro las características de cada uno de estos seguros:

- Seguros ligados a la vivienda: pagos por seguros efectuados tanto por los propietarios como por los inquilinos ocupantes de la vivienda, incluyendo incendio, robo, daños por el agua…**Se excluyen** los seguros pagados habitualmente por los propietarios para cubrir los riesgos que se puedan presentar en el edificio.

- Seguros ligados a la sanidad (seguros de enfermedad y accidente): cuotas pagadas a los seguros médicos no obligatorios, satisfechos directamente a entidades particulares de asistencia sanitaria. También se incluye aquí el seguro escolar.

- Seguros ligados al transporte: seguros relacionados con el transporte personal sobre el vehículo y sus ocupantes. Incluye también los seguros de viaje y equipaje.

- Seguros de responsabilidad civil: seguro de responsabilidad civil por los daños causados a terceros o a sus bienes. **Se excluyen** los que resultan de la utilización de un vehículo personal.

- Seguros de enterramiento.

2.3.2. Análisis cluster:

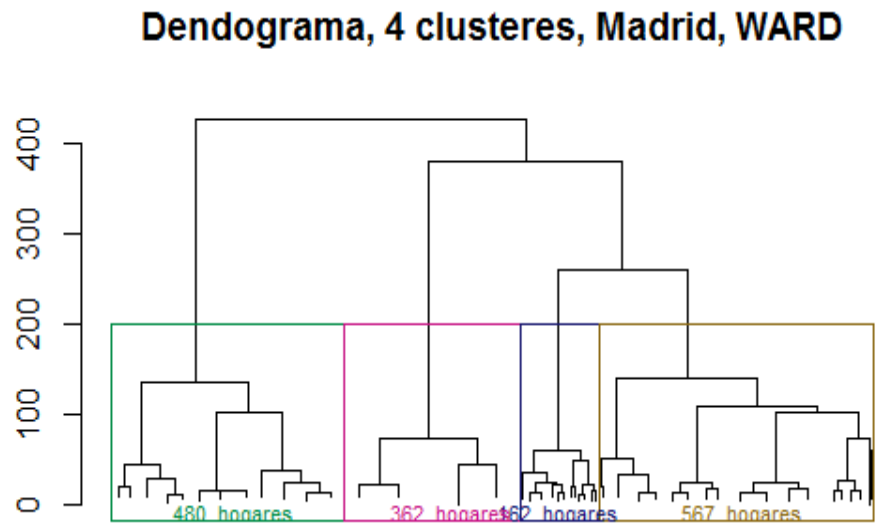
El análisis cluster ha sido la técnica estadística que he elegido para conocer el perfil de los compradores de los seguros privados en España. Este análisis nos permitirá conocer qué características de los individuos se relacionan con las pautas de consumo de seguros.

Una decisión previa es escoger una distancia y un algoritmo de agrupamiento entre las diferentes posibilidades. Como distancia he escogido la euclídea puesto que es la más habitual. Para escoger el algoritmo he realizado un ensayo previo con

menos datos, en el que comparaba los siguientes métodos: encadenamiento simple (single), encadenamiento medio (average) y Ward. Como submuestras seleccioné tres comunidades autónomas (Castilla y León, Galicia y Madrid).

La elección ha sido el método de Ward puesto que produce grupos más homogéneos, mientras que los demás forman grupos afectados por valores atípicos y, por lo tanto, muy heterogéneos y de poca utilidad. Se puede observar esta diferencia en los dos dendogramas[2] siguientes con datos de la comunidad de Madrid; uno está hecho con el método de Ward y otro con el método average. Se ve claramente que con el método de Ward se obtienen grupos homogéneos en tamaño:

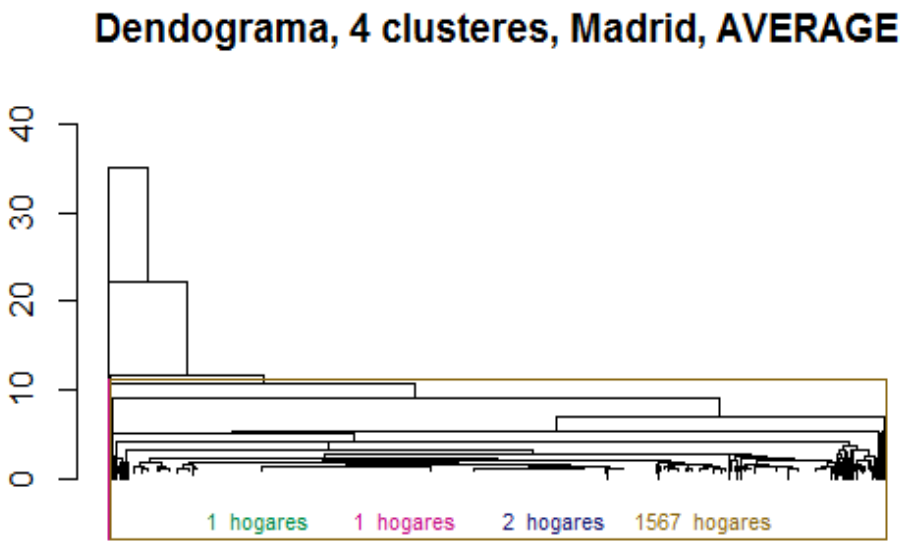
- Grupo 1: 480 hogares
- Grupo 2: 362 hogares
- Grupo 3: 162 hogares
- Grupo 4: 567 hogares



Por lo contrario, con el método average se consiguen grupos que se alejan mucho de la homogeneidad:

- Grupo 1: 1 hogar
- Grupo 2: 1 hogar
- Grupo 3: 2 hogares
- Grupo 4: 1567 hogares

[2] Se pueden observar el resto de los dendogramas en el anexo B



Esta homogeneidad es habitual en este método. Por ejemplo, Martín y De Paz (2007) aseguran que Ward es el método de la varianza mínima debido a que busca separar conglomerados cuya unión conlleve el menor incremento de la varianza. Además, según Pardo y Cabarcas (2001, p.76), el método de Ward utiliza la distancia entre grupos que cumple con el objetivo de buscar clases que tengan menos inercia intra-clases, como criterio de homogeneidad estadística.

2.3.3. Explicación de los perfiles:

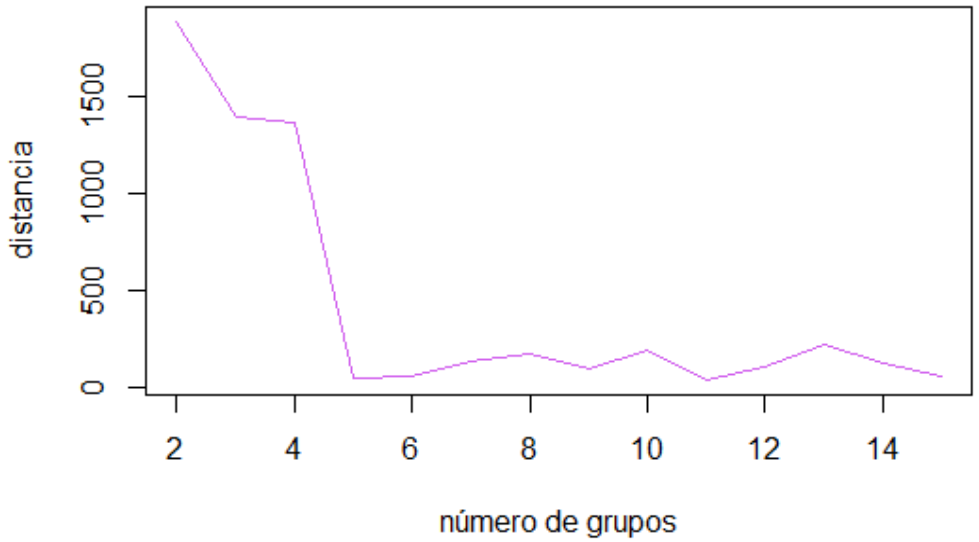
Dado que las variables usadas para explicar los perfiles de los grupos son cualitativas (discretas o han sido agrupadas en intervalos) se puede aplicar en ellas técnicas de análisis de variables cualitativas.

Inicialmente se realiza un contraste de independencia de la χ^2 y, para aquellos casos en los que se obtenga asociación significativa (nivel de significación α=5%), se va a usar como medida de esta asociación la V de Cramer. Posteriormente, para aquellas variables en las que se mida una asociación relevante (valores de V en torno a 0,2) se buscará en qué categorías es más apreciable dicha asociación mediante diagramas de asociación.

2.4. RESULTADOS OBTENIDOS

2.4.1. Grupos a considerar:

La primera decisión que se toma es el número de grupos apropiado. Me he basado en la evolución de distancias de agrupamiento, que se pueden ver representadas en el siguiente gráfico.



Se ve claramente que los grupos de España tienen que ser 4, ya que es el momento en el cual se produce el salto más grande.

2.4.2. Interpretación de los grupos:

Se ha separado la interpretación de los grupos, que realizo con respecto a sus características de consumo de seguros, de la interpretación de sus perfiles, en la que se trata de dar una interpretación de las diferencias anteriores usando las diferentes variables explicativas.

- **Grupo 1 (3522 hogares):** destaca por tener el mayor porcentaje de compradores de seguros en vivienda (95,60%) y en transporte (99,50%), los cuales son los seguros más frecuentes. Además, si nos fijamos en la proporción de seguros por grupos, vemos que los restantes seguros están infrarrepresentados en dicho grupo (-3,57% en enfermedad, -2,85% en civil y -11,49% en enterramiento). Respecto al gasto medio, vemos que este grupo es el que tiene un mayor gasto en el seguro de transporte (1.095€) y es el segundo grupo que tiene un mayor gasto en el de vivienda (303,90€) superándole solamente el grupo 2 (382,80€) del cual hablaremos a

continuación. También coincide dicho razonamiento cuando nos fijamos únicamente en la gente que gasta en seguros. Asimismo, al calcular el primer cuartil (25% de los s en compradores) del grupo 1 vemos como es el que mayor gasto tiene en vivienda (181,162€) y en transporte (673,70€), mientras que al calcular la mediana (50% de los compradores) percibimos que sigue siendo el grupo con mayor gasto en transporte (857,50€) pero en vivienda vuelve a superarlo el grupo 2, con una diferencia de sólo 7,6€, ya que el grupo 1 tiene un gasto de 250€ y en el grupo 2 el gasto es de 257,6€. El cálculo del tercer cuartil (75% de los compradores) nos reafirma la explicación de la mediana dado que el grupo 1 es el grupo con mayor gasto en transporte (1.200€) pero es el segundo grupo con mayor gasto en vivienda (379,10€) ya que lo supera el grupo 2 (390€).

- **Grupo 2 (2789 hogares):** se caracteriza por ser el que mayor gasto tiene en todos los seguros excepto en transporte y en enterramiento, en los cuales es el segundo grupo que más gasta (382,80€ en vivienda, 1.235,61€ en enfermedad, 674,60€ en transporte, 25,16€ en civil y 86,50€ en enterramiento). Como ya dijimos anteriormente, el mayor gasto en transporte lo tiene el grupo 1 (1.095€) y, como veremos más adelante, el mayor gasto en enterramiento lo tiene el grupo 4 (301,05€). El mismo razonamiento nos sirve si nos fijamos solamente en los compradores de seguros. Cabe destacar el seguro de enfermedad en este grupo, ya que es el que recibe un mayor gasto con mucha diferencia respecto del que recibe en los otros grupos y también es el seguro en el que más se gasta dentro del propio grupo; además, en este grupo dicho seguro tiene el 89,10% de compradores, un porcentaje enorme y más aún si lo comparamos con los porcentajes del resto de grupos. Fijándonos en la proporción de seguros por grupos, vemos como el seguro de enfermedad en el grupo 2 es del 48,60%, por lo que podemos decir que este grupo se caracteriza por recoger a los compradores de seguros de enfermedad. Con el seguro civil ocurre algo parecido ya que este grupo es el que más gasto le dedica a dicho seguro también con bastante diferencia respecto del resto de grupos; además, el porcentaje de compradores de este seguro en este grupo es un 11,70% que, a pesar de ser un porcentaje bajo, es el más alto si lo comparamos con el porcentaje que dedican a dicho seguro el resto de grupos. En cuanto a la proporción de seguros, vemos como en este grupo el seguro civil tiene un 53,94%, lo que nos lleva a concluir que este grupo se caracteriza por recoger a los compradores del seguro civil, al igual que sucede con el seguro de enfermedad.

- **Grupo 3 (7214 hogares):** es el grupo que tiene los porcentajes más bajos de compradores en todos los seguros (56,10% en vivienda, 5% en enfermedad, 71,30% en transporte, 0,4% en civil y 4,1% en enterramiento). Del mismo modo, es el grupo que menos gasto medio le dedica a todos los seguros (125,30€ en vivienda, 2,22€ en enfermedad, 282,10€ en transporte, 0,11€ en civil y 0,39€ en enterramiento). Además, todos los seguros están infrarrepresentados en este grupo, ya que la proporción de seguros es la siguiente: -8,04% en vivienda, -24,11% en enfermedad, -3,67% en transporte, -27,76% en civil y -30,16% en enterramiento. Al calcular el primer cuartil para este grupo, nos encontramos con que para el 25% de los compradores el gasto es totalmente inexistente en todos los seguros (0€). La mediana nos indica que el 50% de los compradores gasta 100€ en vivienda y 300€ en transporte que, como ya vimos anteriormente son los dos seguros en los que más gasta este grupo 3 ya que son los dos seguros más habituales en los hogares. El tercer cuartil nos informa de que el 75% de los compradores gasta 222,20€ en vivienda y 421,80€ en transporte, lo cual nos vuelve a confirmar que el grupo 3 dedica sus gastos principalmente a los seguros de vivienda y transporte.

- **Grupo 4 (8263 hogares):** este grupo es muy peculiar, ya que destaca por su enorme porcentaje de compradores en el seguro de enterramiento (100%), es decir, todos los hogares de este grupo tienen contratado un seguro de enterramiento. También tiene un alto porcentaje en el seguro de vivienda (76,10%) y en el seguro de transporte (76,90%). A pesar de estos porcentajes, si nos fijamos en la proporción de seguros por grupo con respecto del total, vemos como excepto vivienda (1,03%) y enterramiento (44,76%) todos los seguros están infrarrepresentados en el grupo 4 (-21,02% en enfermedad, -1,57% en transporte y -23,42% en civil). En cuanto al gasto medio, los seguros en los que más se gasta en este grupo son: transporte (439,50€), seguido de enterramiento (301,05€) y, por último, vivienda (186,10€). El gasto en el resto de seguros es insignificante.

2.4.3. Perfiles de los grupos en función de las variables explicativas:

Para elaborar estos perfiles me voy a basar en el análisis de las diferentes variables cualitativas ya citadas. Para todas ellas el contraste de independencia ha salido significativo[3], sin embargo la asociación medida con la V de Cramer[4] no es alta, por lo que he tomado como referencia aquellas variables en las que esta medida es

[3] No hay independencia porque (p-value= $2e^{-16}$) < (α=5%). El p-value está en el anexo E
[4] La tabla de la V de Cramer se observa en el anexo D

relevante. En estos casos se han analizado sus residuos tipificados mediante diagramas de asociación[5], para así observar qué categorías están sobrerrepresentadas o infrarrepresentadas en cada uno de los grupos.

- **Grupo 1:** este grupo es el que tiene más categorías sobrerrepresentadas respecto al nivel de ocupados (*entre 2 y 4 miembros ocupados/hogar*). Respecto a la situación profesional del sustentador principal, la categoría sobrerrepresentada es *ocupado* y respecto a la situación socioeconómica, la categoría sobrerrepresentada es *trabajadores no manuales de la industria y los servicios.* El tener varios miembros del hogar ocupados, conlleva a tener un nivel de ingresos aceptable (el sustentador principal tiene unos ingresos medios de *2500€ a 5000€ mensuales*). Además, el sustentador principal tiene un buen nivel de estudios, ya que las categorías sobrerrepresentadas son *educación secundaria de segundo ciclo* y *educación superior*. El régimen de tenencia sobrerrepresentado en los hogares de este grupo es *propiedad con préstamo o hipoteca en curso*. Las comunidades autónomas[6] que aparecen sobrerrepresentadas en este grupo son *Comunidad Foral de Navarra* y *País Vasco.*

- **Grupo 2:** este grupo tiene *2 miembros ocupados/hogar.* La situación profesional del sustentador principal sobrerrepresentada en este grupo es *ocupado* y la situación socioeconómica sobrerrepresentada es *trabajadores no manuales de la industria y los servicios,* al igual que sucedía en el grupo 1. Este grupo es el que tiene un mayor nivel de ingresos (el sustentador principal tiene unos ingresos medios de *3000€ a 7000€ mensuales*) y, además, es el grupo en el que el sustentador principal tiene el mejor nivel de estudios, siendo *educación superior* la categoría sobrerrepresentada. El régimen de tenencia sobrerrepresentado en este grupo es *propiedad con préstamo o hipoteca en curso*, al igual que en el grupo 1. Los hogares que pertenecen a este grupo están sobrerrepresentados en *Cataluña*.

- **Grupo 3:** en este grupo aparecen las siguientes categorías sobrerrepresentadas: el sustentador principal de los hogares está *desocupado*, pero hay *1 miembro ocupado/hogar*. Además, es el grupo con menor nivel de ingresos (el

[5] Los diagramas de asociación se encuentran en el anexo E

[6] Hay que indicar que la V de Cramer de esta variable es 0,18, por lo que su importancia en la definición de los grupos es baja, pero he decidido incluirla puesto que también se realiza en el artículo que he utilizado como referencia y además permite una cierta orientación geográfica de los perfiles.

sustentador principal ingresa *menos de 1000€ al mes*). Con respecto al nivel de estudios, en el grupo 3 no hay ninguna categoría sobrerrepresentada ni infrarrepresentada, lo cual quiere decir que para ésta variable dicho grupo tiene una composición muy parecida a la de la población global de España. El régimen de tenencia sobrerrepresentado en este grupo es el *alquiler*. Por comunidades autónomas, este grupo está sobrerrepresentado en *Aragón*, en *La Rioja* y en *Melilla.*

- **Grupo 4:** en este grupo están sobrerrepresentadas las categorías *inactivo* y *jubilado* como situación profesional y situación socioeconómica del sustentador principal, respectivamente. Ningún miembro del hogar está ocupado. El nivel de ingresos del sustentador principal es más bien bajo (*entre 500€ y 1500€ al mes*). Cabe destacar también que estamos ante el grupo en el que el sustentador principal tiene el peor nivel de estudios dado que las categorías sobrerrepresentadas son *sin estudios o con estudios de primer grado* y *educación secundaria de primer ciclo*. El régimen de tenencia sobrerrepresentado en este grupo es *propiedad sin préstamo o hipoteca en curso*. Los hogares que pertenecen a este grupo están sobrerrepresentados en *Andalucía* y *Extremadura*.

2.4.4. Síntesis:

Se puede sintetizar todo lo anterior diciendo que hay dos grupos (el grupo 1 y el grupo 2) con gasto alto en seguros lo cual coincide con los hogares en mejor situación económica, mientras que los otros dos grupos (el grupo 3 y el grupo 4) presentan menos gasto y peor situación económica.

Cabe destacar las siguientes puntualizaciones:

- El grupo 2 es el que dedica un mayor consumo en los seguros de sanidad y es también el grupo que mayor nivel de ingresos tiene.

- Los grupos en los que aparecen sobrerrepresentados menores ingresos (el grupo 3 y el grupo 4) son los que tienen un mayor número de hogares.

- En el grupo 4 hay una sobrerrepresentación del seguro de enterramiento y de jubilados como situación del sustentador principal.

CONCLUSIONES

Como propuse en los objetivos de este trabajo, se ha comprobado la utilidad de la Estadística en el mundo empresarial, ya que una empresa no sólo necesita datos para poder llevar a cabo sus funciones, sino que es necesario poder tratarlos estadísticamente para conseguir resultados con los que decidir y sacar conclusiones.

Se ha comprobado también la utilidad del análisis cluster para elaborar perfiles de consumidores en un sector relevante como es el de los seguros privados, ya que mediante su utilización se han segmentado los consumidores de seguros en cuatro grandes grupos.

Se han analizado los perfiles de cada uno de esos cuatro grupos, relacionándolos con las características propias del sustentador principal y con las características económicas, sociales y demográficas de los hogares que los conforman.

Para cada uno de los grupos se han obtenido unas características que ayudan a definirlos, las cuales se pueden resumir de la siguiente forma:

- **Grupo 1:** es el grupo con mayor porcentaje de compradores en los seguros de vivienda y transporte. Es el que más gasto le dedica al seguro de transporte y es el segundo que más gasta en el seguro de vivienda. Además, es el grupo que tiene el mayor número de ocupados y un nivel aceptable de ingresos.

- **Grupo 2:** es el que tiene un mayor gasto en todos los seguros excepto en transporte y en enterramiento. Destacan en este grupo el seguro de enfermedad y el seguro civil por tener ambos el mayor porcentaje de compradores con respecto al resto de grupos. Hay 2 miembros ocupados/hogar y es el grupo con mayor nivel de ingresos y con mejor nivel de estudios.

- **Grupo 3:** es el grupo con menor porcentaje de compradores en todos los seguros y es el que menos gasto dedica a todos los seguros. Sólo hay 1 miembro ocupado/hogar y el sustentador principal está desocupado (parado). Es el grupo con menor nivel de ingresos.

- **Grupo 4:** destaca por su enorme porcentaje de compradores en el seguro de enterramiento y también tiene porcentajes altos en los seguros de vivienda y transporte. Ningún miembro del hogar está ocupado y el sustentador principal está inactivo (jubilado). Este grupo tiene un nivel bajo de ingresos y tiene el peor nivel de estudios.

BIBLIOGRAFÍA

Albarrán, I. (1999a). *La actividad aseguradora: importancia, revisión e integración de conceptos fundamentales.* Tesis Doctoral, Universidad Complutense de Madrid. Recuperada de: < http://eprints.ucm.es/6723/1/0022.pdf>

Albarrán, I. (1999b). *Mercado de seguros. Oferta y demanda.* Tesis Doctoral, Universidad Complutense de Madrid. Recuperada de: <http://eprints.ucm.es/6722/1/0021.pdf>

Albarrán y De Pablos (2001). El perfil del consumidor en el sector asegurador español. *Revista investigación operacional*, 22(2), 144-153. Recuperado de: <http://rev-inv-ope.univ-paris1.fr/files/22201/IO-22201-7.pdf>

Bisquerra, R. (1989). *Introducción conceptual al Análisis Multivariable.* Barcelona: PPU (Promociones y Publicaciones Universitarias S.A.)

Castelo, J. y Guardiola, A. (1992). *Diccionario MAPFRE de seguros* (3 ed.). Madrid: Mapfre

Cuadras, C. (2014). *Nuevos métodos de análisis multivariante.* Barcelona: CMC Editions. Recuperado de: <http://www.ub.edu/stat/personal/cuadras/metodos.pdf>

Cramer, H. (1968). *Métodos matemáticos de estadística* (4 ed.). Madrid: Aguilar

De la Fuente, S. (2011). Análisis de conglomerados. Recuperado de: <http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/CONGLOMERADOS/conglomerados.pdf>

Directiva 93/13/CEE. Glosario de términos de Protección de los consumidores: Davarra y Davarra (Asesores Jurídicos). Recuperado de: <http://www.davara.com/documentos/glosario/consumer.html>

Embid, P., Martin, M. y Zorrilla, V. (1998). *Marketing financiero.* Madrid: Mc-Graw-Hill

EUROSTAT (1988): Sistema Europeo de Cuentas Integradas (SEC), INE, Luxemburgo

Fletcher, J. (2013). *Introduction to Researh Methods*: Toronto: University of Toronto, Department of Political Science, 2013. "Lab Manual: Exercise 3A, Crosstabulation with Nominal Variables". Recuperado de: <http://groups.chass.utoronto.ca/pol242/Labs/LM-3A/LM-3A_content.htm>

Hair, J., Anderson, R., Tatham, R. y Black. W. (1999). *Análisis Multivariante* (5 ed.). Madrid: Prentice Hall Iberia

Instituto Nacional de Estadística (2012). Encuesta de Presupuestos Familiares (Base 2006). Recuperado de: <http://www.ine.es/prodyser/micro_epf2006.htm>

Instituto Nacional de Estadística (2012). Metodología de la Encuesta de Presupuestos Familiares. Recuperado de: <http://www.ine.es/metodologia/t25/t2530p458.pdf>

Kendall, M. (1980). *Multivariate Analysis* (2 ed.). Londres: Charles Griffin & Company LTD

Lévy, J-P., y Varela, J. (2003). *Análisis Multivariable para las Ciencias Sociales.* Madrid: Pearson educación S.A.

Martín, Q. y De Paz, Y. (2007). *Tratamiento estadístico de datos con SPSS.* Madrid: Paraninfo

Mendenhall, W. y Reinmuth, J. (1978). *Estadística para administración y economía*. España: Grupo Editorial Iberoamérica

Pardo, C. y Cabarcas, G. (2001). Métodos estadísticos multivariados en investigación social. *Simposio de Estadística,* 73-91. Recuperado de: <http://www.eio.uva.es/~valentin/ad3d/anadat/teoria_web/simposio_estadistica/MetEstMulInvSocialParte4_CLUSTER.pdf>

Peña, D. (2001). *Fundamentos de estadística*. Madrid: Alianza Editorial

Peña, D. (2002). *Análisis de datos multivariantes.* Madrid: McGraw-Hill

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos.* Madrid: Pearson educación S.A.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Ruiz-Maya, L., Martín, F.J., Montero, J.Mª. Y Uriz, P. (1995). *Análisis estadístico de encuestas: datos cualitativos.* Madrid: Editorial AC

Sánchez, F. (1981). Las Mutualidades y el movimiento de defensa del consumidor. *Revista Española de Seguros*, 26. Recuperado de: <http://eprints.ucm.es/6722/1/0021.pdf>

Sarabia, J. M., y Pascual, M. (2005). *Curso básico de Estadística para Economía y Administración de Empresas.* Santander: Servicio de Publicaciones de la Universidad de Cantabria

Sokal, R., y Sneath, P. (1963). *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman

Uriel, E. (1995). *Análisis de datos. Series temporales y Análisis multivariante*. Madrid: Editorial AC

Zeileis, A., Meyer, D. y Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd. *Journal of Statistical Software*, 17(3), 1–48. Recuperado de: <http://www.jstatsoft.org/v17/i03/paper>

ANEXOS

❖ ANEXO A. VARIABLES EXPLICATIVAS

ACTIVIDA: situación profesional del Sustentador Principal

I (inactivo)

D (desocupado)

O (ocupado)

Variable elaborada a partir de SITUREDSP: situación en la actividad reducida del sustentador principal y de OCUSP: ¿estaba el sustentador principal ocupado en la semana anterior a la entrevista?

CATPROFE: situación socioeconómica del Sustentador Principal (clasificación reducida)

1 Trabajadores de la industria y los servicios

2 Trabajadores no manuales de la industria y los servicios

3 Autónomos de la industria, servicios y agricultores

4 Parados

5 Jubilados y otros

CONDSOEC: condición socioeconómica del Sustentador Principal

1 Directores y gerentes

2 Técnicos y profesionales

3 Empleados de tipo administrativo y trabajadores de servicios y de comercio

4 Artesanos y trabajadores cualificados de otros sectores, operadores y montadores

5 Trabajadores en ocupaciones elementales

6 Nunca trabajo o No consta (incluye fuerzas armadas)

EDAD: edad del Sustentador Principal

"menos de 31" "entre 31 e 45" "entre 46 e 66" "más de 66"

NIVESTUD: nivel de estudios del Sustentador Principal

1 Sin estudios o con estudios de primer grado

2 Educación secundaria, primer ciclo

3 Educación secundaria, segundo ciclo

4 Educación superior

SEXO: sexo del Sustentador Principal

1 Hombre

6 Mujer

REGCICIE: régimen de tenencia de la vivienda

1 Propiedad sin préstamo o hipoteca en curso

2 Propiedad con préstamo o hipoteca en curso

3 Alquiler

4 Alquiler reducido (renta antigua)

5 Cesión semigratuita

6 Cesión gratuita

SUPUTIL: superficie útil de la vivienda

"menos de 75" "entre 75 e 100" "entre 100 e 125" "más de 125"

CCAA: comunidad autónoma de residencia

1 Andalucía

2 Aragón

3 Principado de Asturias

4 Illes Balears

5 Canarias

6 Cantabria

7 Castilla y León

8 Castilla – La Mancha

9 Cataluña

10 Comunidad Valenciana

11 Extremadura

12 Galicia

13 Comunidad de Madrid

14 Región de Murcia

15 Comunidad Foral de Navarra

16 País Vasco

17 La Rioja

18 Ceuta

19 Melilla

TAMAMUN: tamaño del municipio de residencia

1 Municipio de 100.000 habitantes o más

2 Municipio con 50.000 o más y menos 100.000 habitantes

3 Municipio con 20.000 o más y menos de 50.000 habitantes

4 Municipio con 10.000 o más y menos de 20.000 habitantes

5 Municipio con menos de 10.000 habitantes

NUMOCU: número de ocupados en el hogar

“**0**” “**1**” “**2**” “**3**” “**4**” “**5**”

FUENPRINRED: fuente principal de ingresos

1 Trabajo por cuenta propia y rentas de la propiedad y del capital

2 Trabajo por cuenta ajena

3 Pensiones, subsidios y otras prestaciones e ingresos regulares

-9 No consta

INTERIN: nivel de ingresos del Sustentador Principal

1 Menos de 500 €

2 De 500 a menos de 1000 €

3 De 1000 a menos de 1500 €

4 De 1500 a menos de 2000 €

5 De 2000 a menos de 2500 €

6 De 2500 a menos de 3000 €

7 De 3000 a menos de 5000 €

8 De 5000 a menos de 7000 €

9 De 7000 a menos de 9000 €

10 9000 o más €

NUMPERI: número de personas que ingresan

"**0**" "**1**" "**2**" "**3**" "**4**" "**5**" "**6**" "**-9**"

NUMEMIEMB: número de miembros de la familia

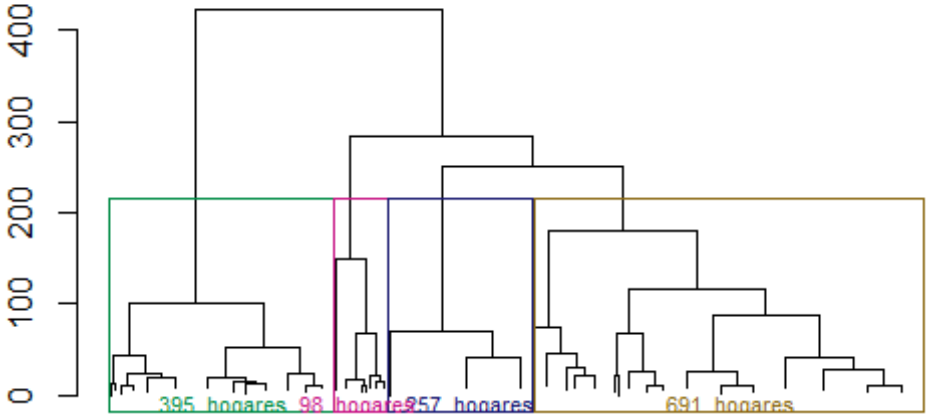
"**1**" "**2**" "**3**" "**4**" "**5**"

5 representa "5 o más"

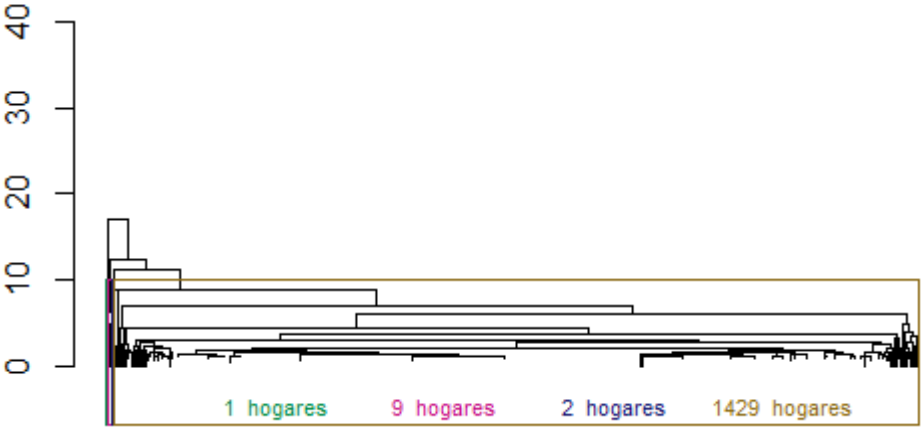
FUENTE: elaboración propia a partir del INE (2012)

❖ ANEXO B. DENDOGRAMAS DE LAS COMUNIDADES AUTÓNOMAS

Dendograma, 4 clusteres, Castilla y León, WARD

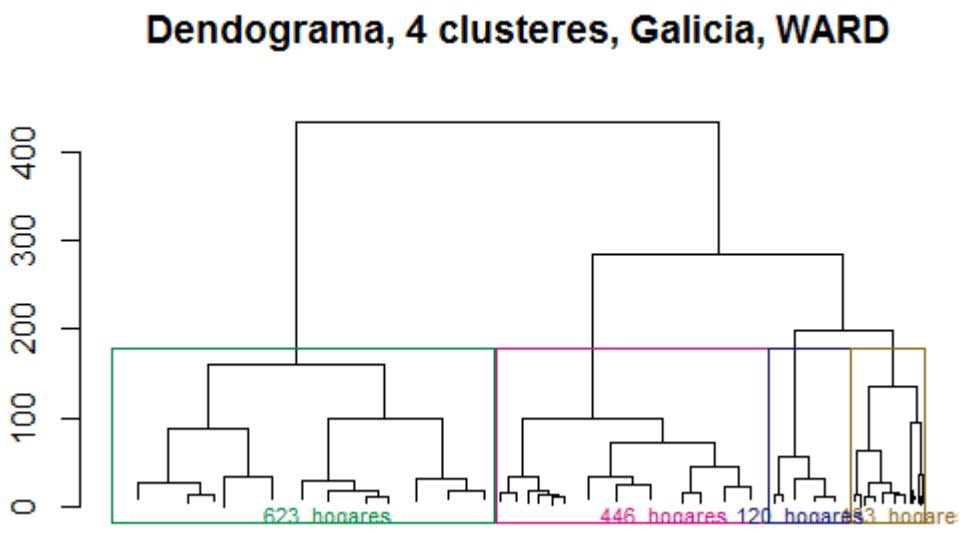


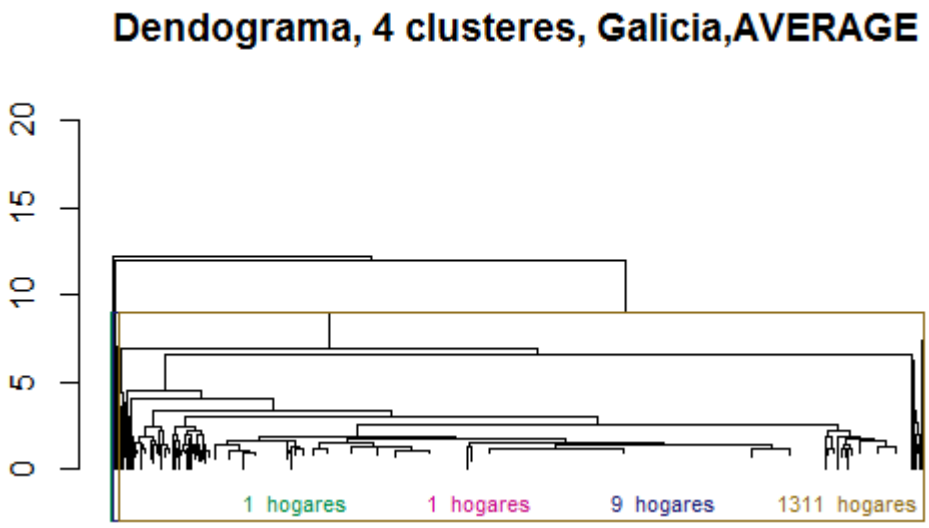
Dendograma, 4 clusteres, Castilla y León, AVERAGE

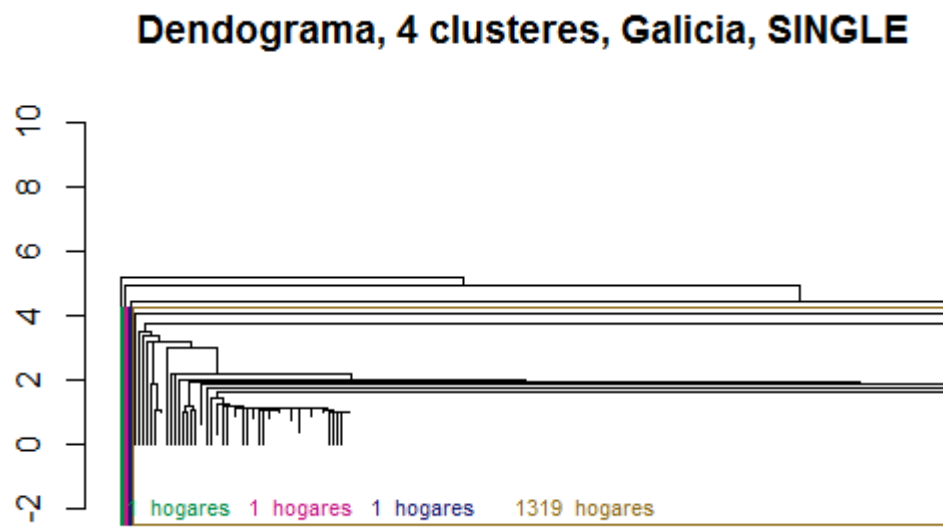


Dendograma, 4 clusteres, Castilla y León, SINGLE

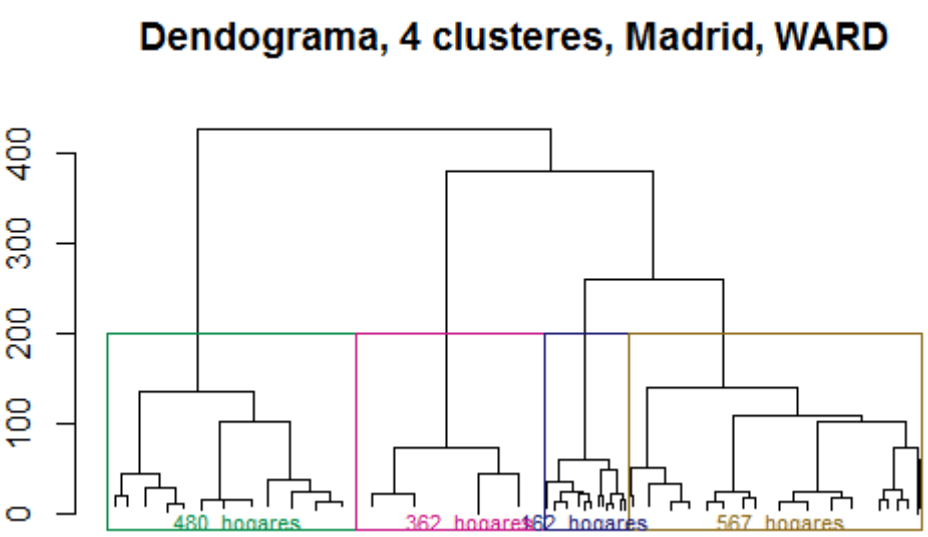
FUENTE: elaboración propia a partir del INE (2012)

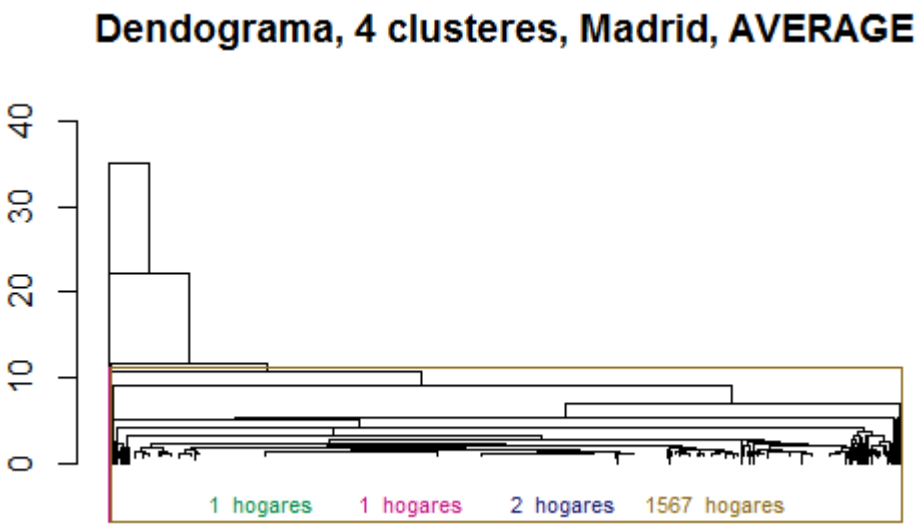


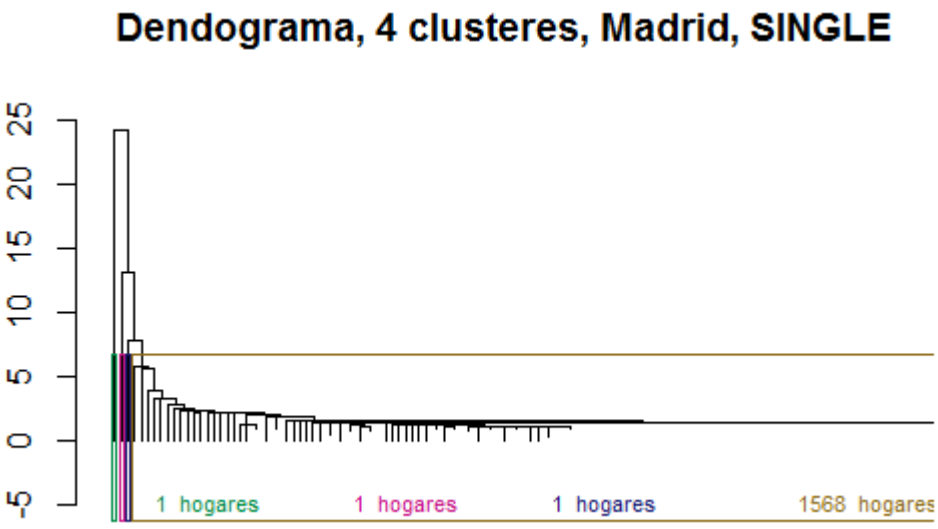




FUENTE: elaboración propia a partir del INE (2012)

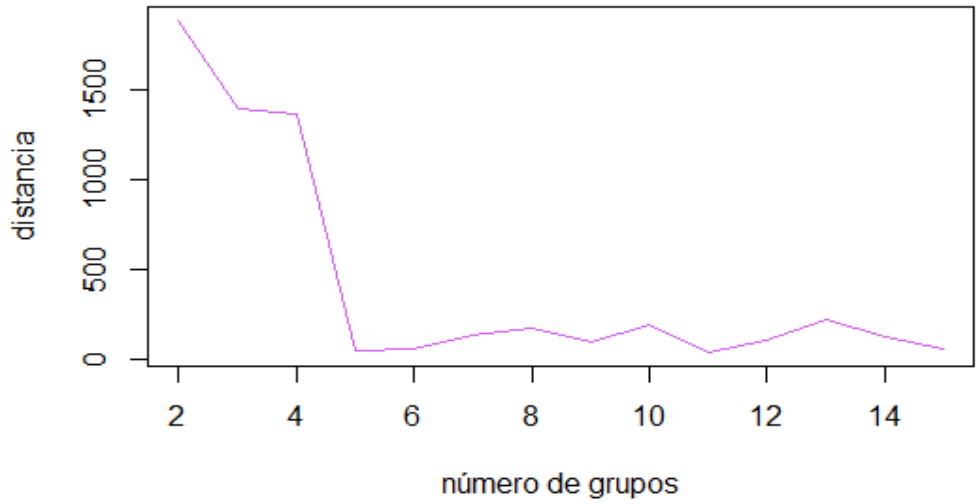






FUENTE: elaboración propia a partir del INE (2012)

❖ ANEXO C. CÁLCULOS CLÚSTERES ESPAÑA



FUENTE: elaboración propia a partir del INE (2012)

Porcentajes de compradores de seguros por grupos (%):

GRUPOS	VIVIENDA	SANIDAD	TRANSPORTE	CIVIL	ENTERRAMIENTO
1	95,60	14,50	99,50	1,80	13,40
2	87,20	89,10	88,50	11,70	34,50
3	56,10	5,00	71,30	0,40	4,10
4	76,10	8,30	76,90	0,90	100

Porcentajes de seguros por grupos (diferencias respecto a lo esperado) (%):

GRUPOS	VIVIENDA	SANIDAD	TRANSPORTE	CIVIL	ENTERRAMIENTO
1	4,65	-3,57	3,84	-2,85	-11,49
2	2,26	48,60	1,31	53,94	-3,20
3	-8,04	-24,11	-3,67	-27,76	-30,16
4	1,03	-21,02	-1,57	-23,42	44,76

Gasto medio por grupos (€/año):

GRUPOS	VIVIENDA	SANIDAD	TRANSPORTE	CIVIL	ENTERRAMIENTO
1	303,90	26,97	1095	0,29	12,57
2	382,80	1235,61	674,6	25,16	86,50
3	125,30	2,22	282,1	0,01	0,39
4	186,10	8,35	439,5	0,07	301,05

Gasto medio por grupos de los compradores de seguros (€/año):

GRUPOS	VIVIENDA	SANIDAD	TRANSPORTE	CIVIL	ENTERRAMIENTO
1	317,80	185,91	1100,30	15,73	93,97
2	438,80	1386,76	762,70	215,94	251,03
3	223,30	43,89	395,40	3,08	9,62
4	244,5	100,66	571,50	8,57	301,05

Primer cuartil (€/año):

GRUPOS	VIVIENDA 25%	SANIDAD 25%	TRANSPORTE 25%	CIVIL 25%	ENTERRAMIENTO 25%
1	181,16	0,00	673,70	0,00	0,00
2	170,00	503,30	326,00	0,00	0,00
3	0,00	0,00	0,00	0,00	0,00
4	3,35	0,00	23,30	0,00	183,70

Mediana (€/año):

GRUPOS	VIVIENDA 50%	SANIDAD 50%	TRANSPORTE 50%	CIVIL 50%	ENTERRAMIENTO 50%
1	250,00	0,00	857,50	0,00	0,00
2	257,60	1020,00	558,00	0,00	0,00
3	100,00	0,00	300,00	0,00	0,00
4	190,00	0,00	361,80	0,00	267,80

Tercer cuartil (€/año):

GRUPOS	VIVIENDA 75%	SANIDAD 75%	TRANSPORTE 75%	CIVIL 75%	ENTERRAMIENTO 75%
1	379,10	0,00	1200,00	0,00	0,00
2	390,00	1800,00	900,00	0,00	141,40
3	222,20	0,00	421,80	0,00	0,00
4	260,00	0,00	646,30	0,00	375,80

FUENTE: elaboración propia a partir del INE (2012)

Diagrama de caja del seguro de vivienda:

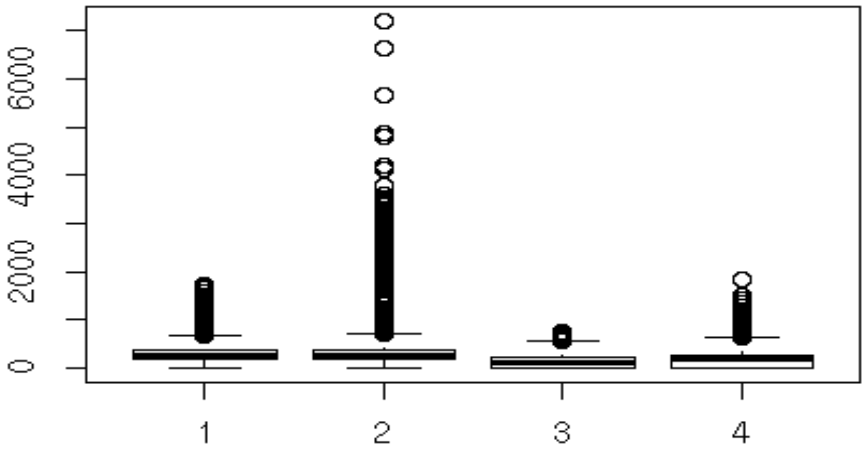


Diagrama de caja del seguro de sanidad:

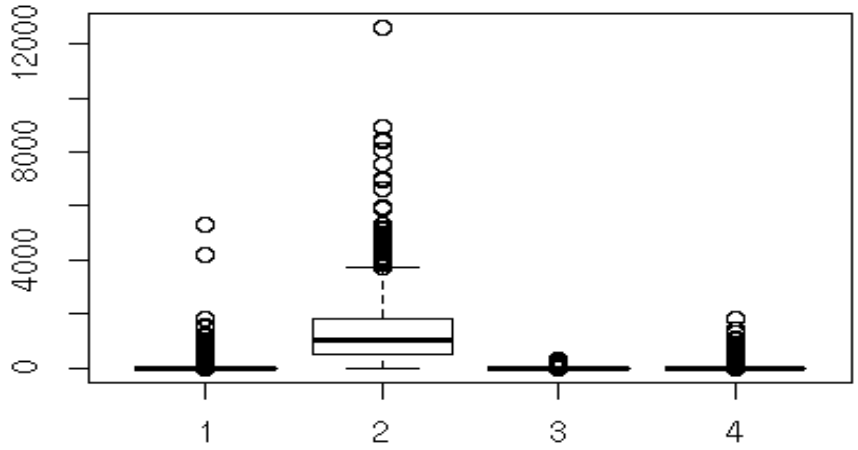


Diagrama de caja del seguro de transporte:

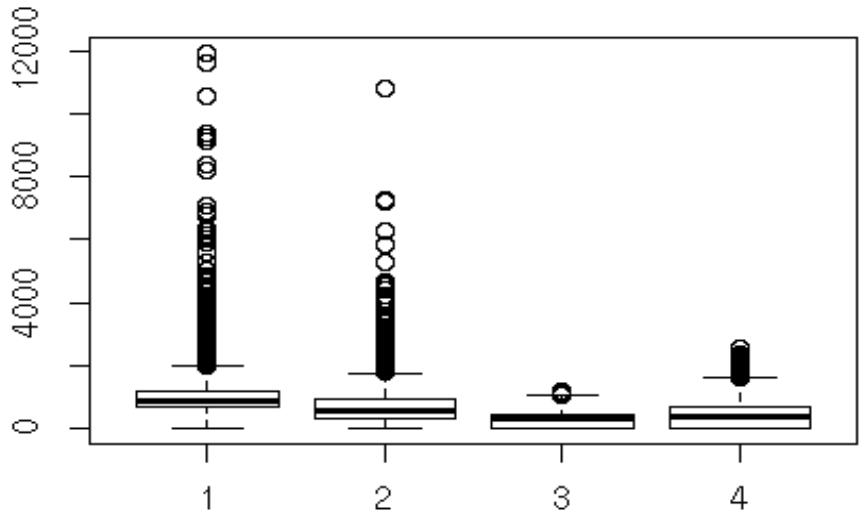


Diagrama de caja del seguro civil:

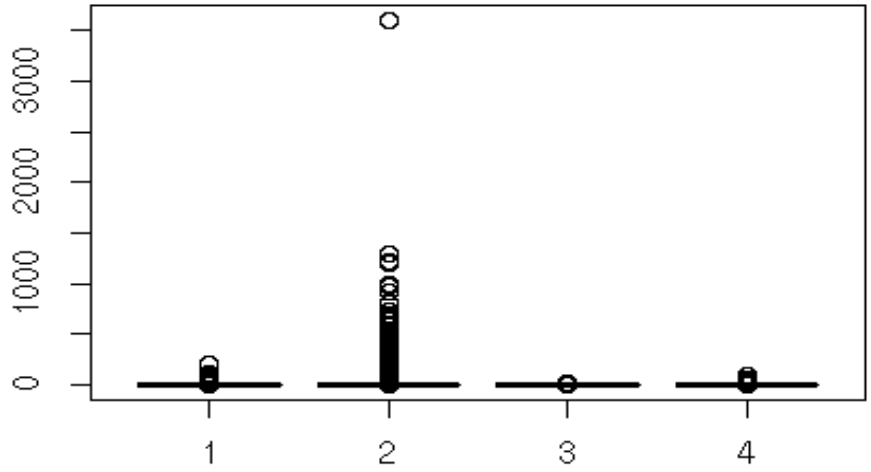
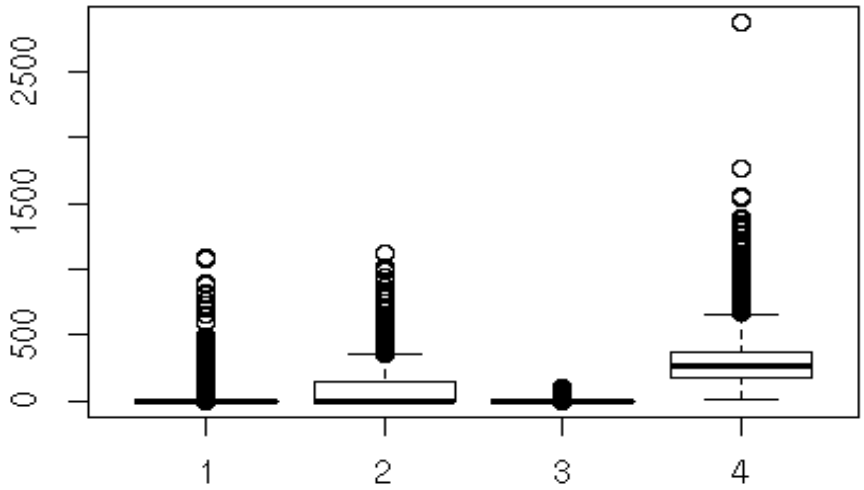


Diagrama de caja del seguro de enterramiento:



FUENTE: elaboración propia a partir del INE (2012)

❖ ANEXO D. METODOLOGÍA Y TABLAS RESUMEN DE LAS VARIABLES EXPLICATIVAS

Para analizar los perfiles de cada grupo he utilizado unos cálculos que permiten conocer la asociación existente entre las variables:

- **V de Cramer:** Cramer (1968) ha descrito este coeficiente como una medida de asociación entre variables medidas en escala nominal que utiliza el estadístico χ^2 (V y χ^2 coinciden en el caso sencillo de tablas 2 x 2). El estadístico V está dentro del intervalo [0,1]. Un valor V = 0 significa asociación inexistente y V = 1 implica una asociación perfecta.

- **Contraste chi cuadrado (χ^2) de independencia:** según Ruiz-Maya, Martín, Montero y Uriz (1995), este contraste se usa para contrastar la independencia entre dos atributos. Para construir el contraste se utiliza la fórmula del estadístico χ^2 de Pearson. Este estadístico mide la diferencia entre la frecuencia conjunta observada (n_{ij}) y la frecuencia conjunta teórica que existiría si los atributos fuesen independientes ($\hat{E}_{ij}$). Obviamente, si existe independencia su valor estará cercano a cero. Por lo contrario, si toma un valor distante de cero, se rechaza la hipótesis nula de independencia de atributos. En cuanto a los grados de libertad $(r-1)\text{x}(c-1)$, "r" y "c" son los números de las categorías de los atributos.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_{(r-1)\text{x}(c-1)} \qquad \text{siendo} \quad \hat{E}_{ij} = \frac{n_{i.}\cdot n_{.j}}{n}$$

- **Diagramas de asociación:** estos diagramas son una representación gráfica de los residuos tipificados corregidos. Visualizan las desviaciones normales entre las frecuencias reales y las esperadas, bajo cierta hipótesis de independencia. Cada celda está representada por un rectángulo cuya altura es proporcional al residual y cuya base es proporcional a la raíz cuadrada de las cuentas esperadas, de modo que el área de dicho rectángulo es proporcional a la diferencia entre frecuencias observadas y esperadas (Zeileis, Meyer y Hornik, 2006).

Tabla interpretación nivel de asociación (V Cramer):

NIVEL DE ASOCIACIÓN	DESCRIPCIÓN DE LA RELACIÓN	COMENTARIOS
0,00	No hay relación	Conociendo la variable independiente no ayuda en la predicción de la variable dependiente
0,00 a 0,15	Muy débil	Generalmente no aceptable
0,15 a 0,20	Débil	Mínimamente aceptable
0,20 a 0,25	Moderada	Aceptable
0,25 a ,30	Moderadamente fuerte	Deseable
0,30 a0,35	Fuerte	Muy deseable
0,35 a 0,40	Muy fuerte	Muy deseable
0,40 a 0,50	Preocupantemente fuerte	O es una relación muy buena o las dos variables están midiendo el mismo concepto
0,50 a 0,99	Redundante	Las dos variables están midiendo el mismo concepto
1,00	Relación perfecta	Si conocemos la variable independiente, podemos predecir perfectamente la variable dependiente

FUENTE: elaboración propia a partir de Fletcher (2013) de Toronto University

VARIABLES EXPLICATIVAS	V de CRAMER
Situación profesional del sustentador principal (ACTIVIDA)	**0,21**
Situación socioeconómica del sustentador principal (CATPROFE)	**0,20**
Condición socioeconómica (CONDSOEC)	0,18
Edad del sustentador principal (EDAD)	0,17
Nivel de estudios del sustentador principal (NIVESTUD)	**0,20**
Sexo del sustentador principal (SEXO)	0,09
Régimen de tenencia de la vivienda (REGCICIE)	**0,19**
Superficie útil de la vivienda (SUPUTIL)	0,10
Comunidad autónoma de residencia (CCAA)	**0,18**
Tamaño del municipio (TAMAMUN)	0.07
Número de ocupados en el hogar (NUMOCU)	**0,19**
Fuente principal de ingresos (FUENPRINRED)	0,17
Nivel de ingresos del sustentador principal (INTERIN)	**0,25**
Número de personas que ingresan (NUMPERI)	0,14
Número de miembros de la familia (NUMEMIEMB)	0,12

Basándome en esta tabla he construido las siguientes, teniendo en cuenta aquellas variables cuya V de Cramer está entre **0,20** y **0,25**, es decir, que tienen una relación moderada (aceptable) según Fletcher et al. (2013). Hay alguna excepción: también incluyo las variables cuya V de Cramer es de **0,19** y una variable cuya V de Cramer es **0,18** ya que quiero ver la relevancia que tienen.

	V de CRAMER
Situación profesional del sustentador principal (ACTIVIDA)	0,21
Situación socioeconómica del sustentador principal (CATPROFE)	0,20
Nivel de estudios del sustentador principal (NIVESTUD)	0,20
Régimen de tenencia de la vivienda (REGCICIE)	0,19
Comunidad autónoma de residencia (CCAA)	0,18
Número de ocupados en el hogar (NUMOCU)	0,19
Nivel de ingresos del sustentador principal (INTERIN)	0,25

A continuación se muestran dichas tablas. “Sobrerrepresentado” e “Infrarrepresentado” hacen referencia a las categorías de cada variable que están sobrerrepresentadas o infrarrepresentadas. Para saberlo, he hecho otros cálculos, los cuales están en el siguiente anexo.

Situación profesional del sustentador principal (ACTIVIDA):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- **Ocupado**	- Desocupado - **Inactivo**
2	- **Ocupado**	- **Desocupado** - Inactivo
3	- **Desocupado**	- **Inactivo**
4	- **Inactivo**	- **Ocupado**

Situación socioeconómica del sustentador principal (CATPROFE):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- Trabajadores de la industria y los servicios - **Trabajadores no manuales de la industria y los servicios** - Autónomos de la industria, servicios y agricultores	- Parados - **Jubilados y otros**
2	- **Trabajadores no manuales de la industria y los servicios** - Autónomos de la industria, servicios y agricultores	- Trabajadores de la industria y los servicios - **Parados** - Jubilados y otros
3	- **Parados**	- Trabajadores no manuales de la industria y los servicios - **Jubilados y otros**
4	- **Jubilados y otros**	- **Trabajadores no manuales de la industria y los servicios** - Autónomos de la industria, servicios y agricultores

Nivel de estudios del sustentador principal (NIVESTUD):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- Educación secundaria, segundo ciclo - **Educación superior**	- **Sin estudios o con estudios de 1er grado** - Educación secundaria, primer ciclo
2	- **Educación superior**	- **Sin estudios o con estudios de 1er grado** - Educación secundaria, primer ciclo
3	-	-
4	- **Sin estudios o con estudios de 1er grado** - Educación secundaria, primer ciclo	- Educación secundaria, segundo ciclo - **Educación superior**

Régimen de tenencia de la vivienda (REGCICIE):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- **Propiedad con préstamo o hipoteca en curso**	- Propiedad sin préstamo o hipoteca en curso - **Alquiler** - Alquiler reducido (renta antigua) - Cesión gratuita
2	- **Propiedad con préstamo o hipoteca en curso**	- **Alquiler**
3	- **Alquiler** - Cesión gratuita	- **Propiedad sin préstamo o hipoteca en curso** - Propiedad con préstamo o hipoteca en curso
4	- **Propiedad sin préstamo o hipoteca en curso** - Alquiler reducido (renta antigua)	- Propiedad con préstamo o hipoteca en curso - **Alquiler**

Comunidad autónoma de residencia (CCAA):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- Cantabria - Comunidad de Madrid - **Comunidad Foral de Navarra** - **País Vasco**	- **Andalucía** - Canarias - Extremadura
2	- Illes Balears - **Cataluña** - Comunidad de Madrid	- Principado de Asturias - Castilla y León - Castilla-La Mancha - **Extremadura** - **Comunidad Foral Navarra**
3	- **Aragón** - Illes Balears - Castilla y León - Comunidad Foral de Navarra - **La Rioja** - **Melilla**	- **Andalucía** - Comunidad de Madrid
4	- **Andalucía** - Principado de Asturias - Canarias - Comunitat Valenciana - **Extremadura** - Región de Murcia	- **Aragón** - **Illes Balears** - **Cataluña** - Comunidad de Madrid - **Comunidad Foral Navarra** - País Vasco - La Rioja - Melilla

Número de ocupados en el hogar (NUMOCU):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	**2 3 4**	**0**
2	**2**	**0**
3	0 **1**	**2 3**
4	**0**	1 **2**

Nivel de ingresos del sustentador principal (INTERIN):

GRUPOS	SOBRERREPRESENTADO	INFRARREPRESENTADO
1	- De 2000 a menos de 2500 € - **De 2500 a menos de 3000 €** - **De 3000 a menos de 5000 €** - De 5000 a menos de 7000 €	- Menos de 500 € - **De 500 a menos de 1000 €** - De 1000 a menos de 1500 €
2	- De 2500 a menos de 3000 € - **De 3000 a menos de 5000 €** - **De 5000 a menos de 7000 €** - De 7000 a menos de 9000 € - 9000 o más €	- Menos de 500 € - **De 500 a menos de 1000 €** - De 1000 a menos de 1500 €
3	- **Menos de 500 €** - **De 500 a menos de 1000 €** - De 1000 a menos de 1500 €	- De 2500 a menos de 3000 € - **De 3000 a menos de 5000 €** - De 5000 a menos de 7000 €
4	- **De 500 a menos de 1000 €** - **De 1000 a menos de 1500 €** - De 1500 a menos de 2000 €	- De 2500 a menos de 3000 € - **De 3000 a menos de 5000 €** - De 5000 a menos de 7000 €

FUENTE: elaboración propia a partir del INE (2012)

❖ ANEXO E. FRECUENCIAS RELATIVAS, CONTRASTE CHI CUADRADO Y DIAGRAMAS DE ASOCIACIÓN DE LAS VARIABLES EXPLICATIVAS

Situación profesional del sustentador principal (ACTIVIDA)

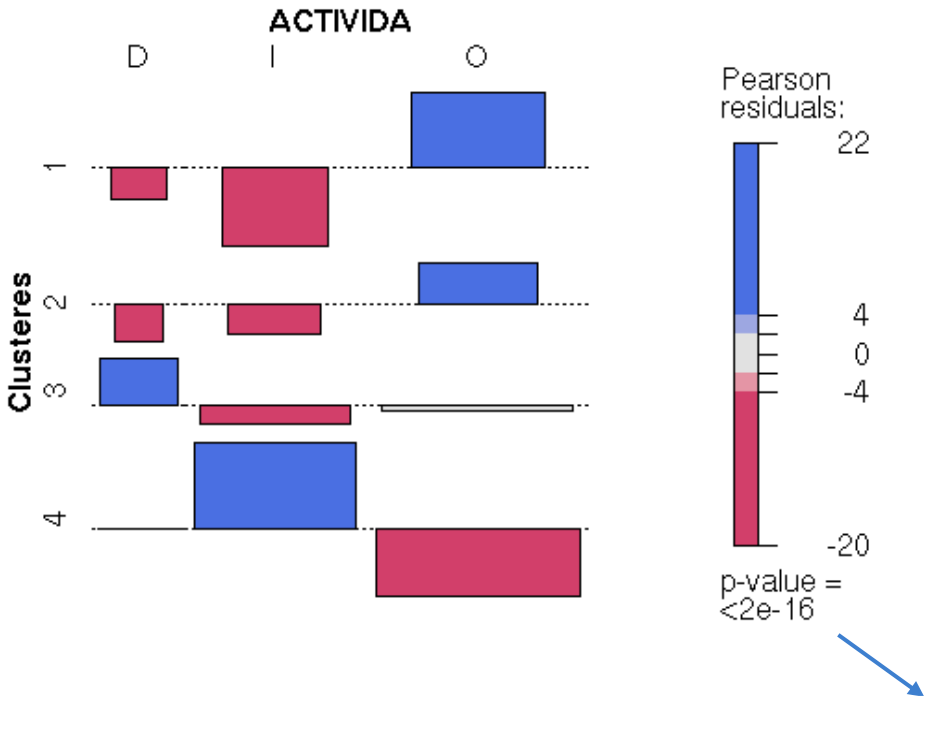
- **Frecuencias relativas:**

D	I	O
0,09	0,35	0,56

- **Frecuencias relativas por grupos:**

GRUPOS	D	I	O
Grupo 1	0,05	0,15	0,80
Grupo 2	0,04	0,26	0,70
Grupo 3	0,14	0,31	0,55
Grupo 4	0,09	0,49	0,42

- **Diagrama de asociación:**



Chi cuadrado

Situación socioeconómica del sustentador principal (CATPROFE)

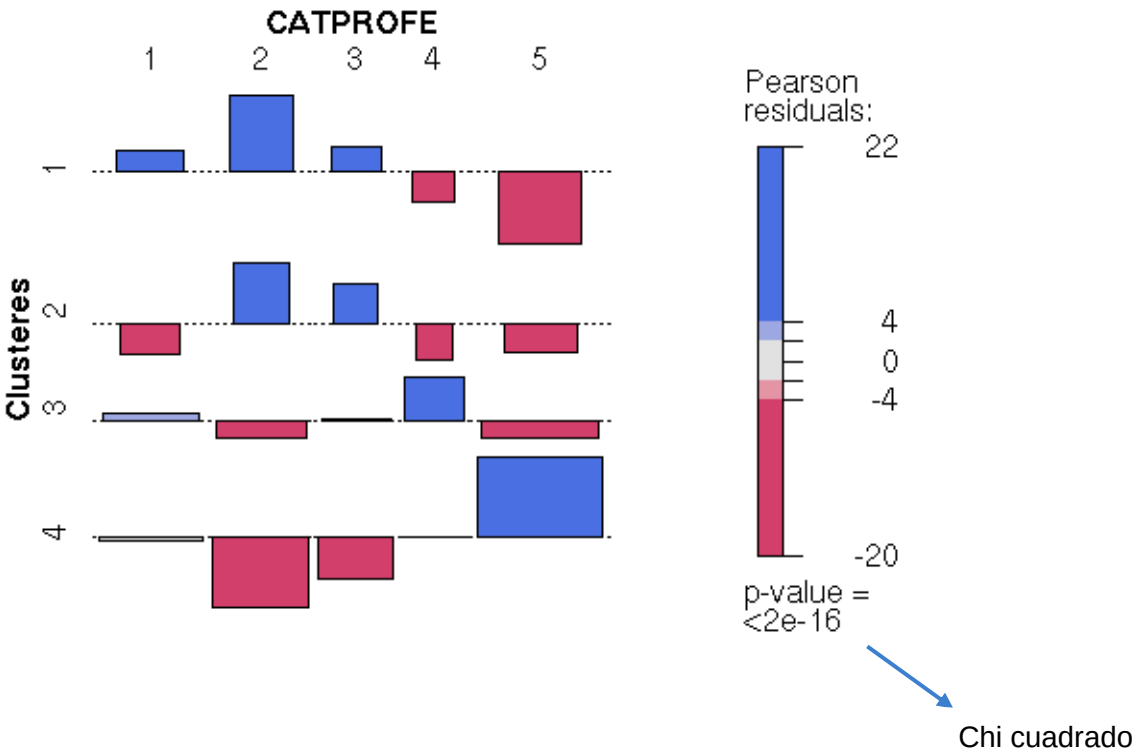
- **Frecuencias relativas:**

1	2	3	4	5
0,23	0,20	0,12	0,09	0,35

- **Frecuencias relativas por grupos:**

GRUPOS	1	2	3	4	5
Grupo 1	0,28	0,36	0,16	0,05	0,15
Grupo 2	0,16	0,34	0,20	0,04	0,26
Grupo 3	0,24	0,18	0,13	0,14	0,31
Grupo 4	0,23	0,11	0,08	0,09	0,49

- **Diagrama de asociación:**



Nivel de estudios del sustentador principal (NIVESTUD)

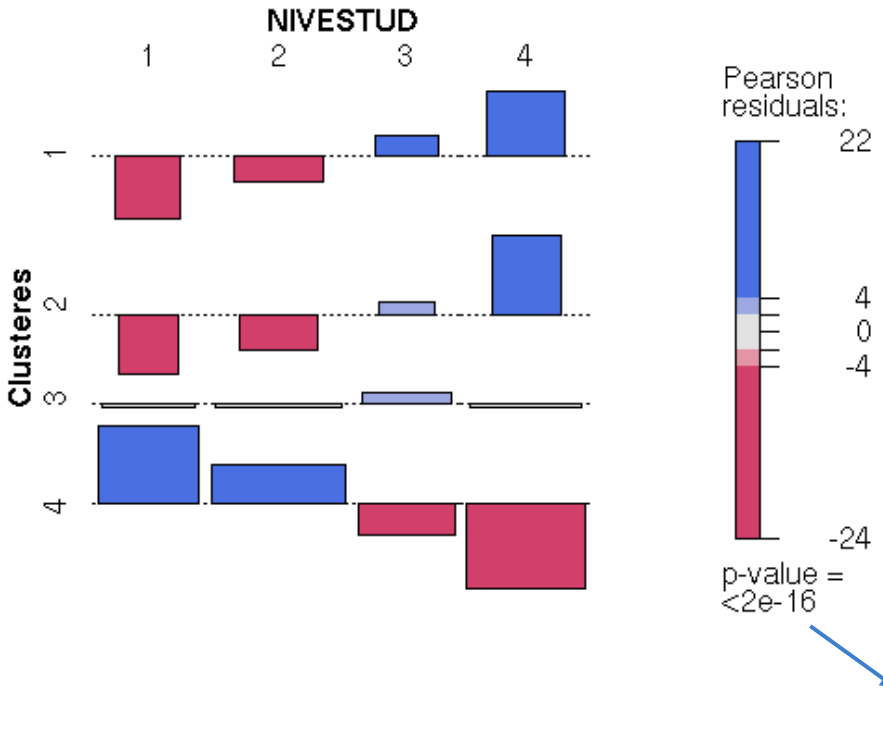
- **Frecuencias relativas:**

1	2	3	4
0,19	0,35	0,18	0,28

- **Frecuencias relativas por grupos:**

GRUPOS	1	2	3	4
Grupo 1	0,07	0,28	0,22	0,44
Grupo 2	0,05	0,24	0,21	0,50
Grupo 3	0,19	0,35	0,19	0,27
Grupo 4	0,30	0,43	0,14	0,14

- **Diagrama de asociación:**



Chi cuadrado

Régimen de tenencia de la vivienda (REGCICIE)

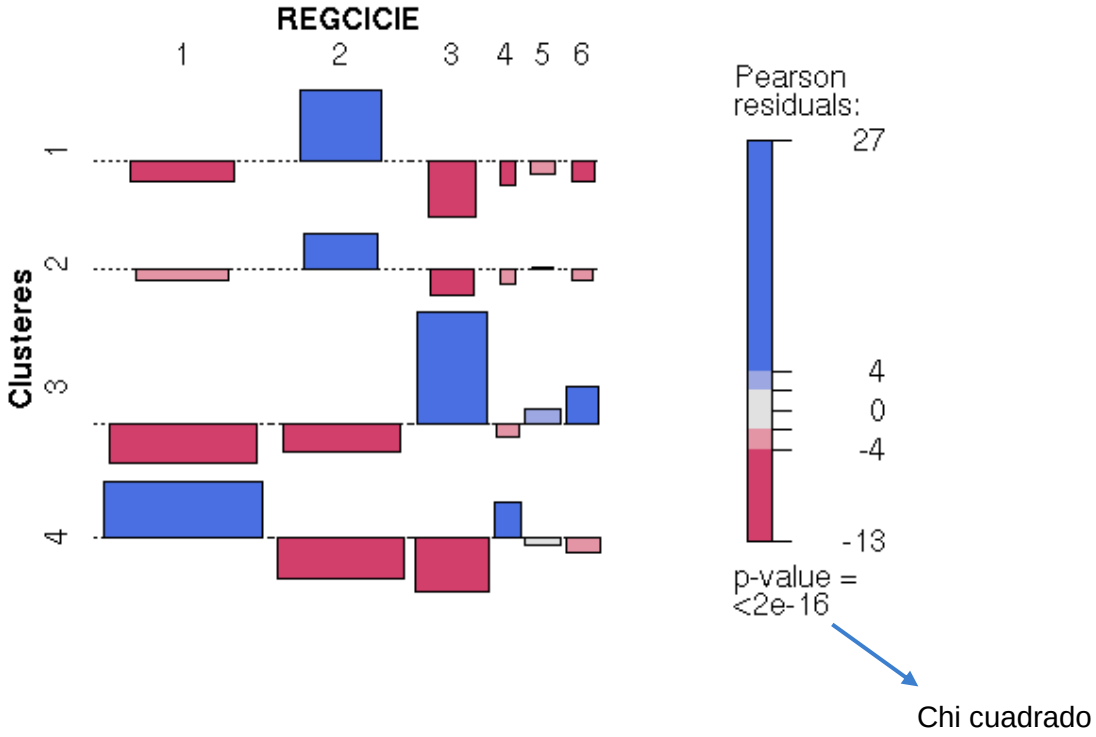
- **Frecuencias relativas:**

1	2	3	4	5	6
0,51	0,32	0,11	0,01	0,03	0,02

- **Frecuencias relativas por grupos:**

GRUPOS	1	2	3	4	5	6
Grupo 1	0,45	0,48	0,04	0,00	0,02	0,01
Grupo 2	0,47	0,41	0,07	0,00	0,03	0,02
Grupo 3	0,43	0,27	0,22	0,01	0,04	0,04
Grupo 4	0,62	0,25	0,06	0,02	0,03	0,02

- **Diagrama de asociación:**



Chi cuadrado

Comunidad autónoma de residencia (CCAA)

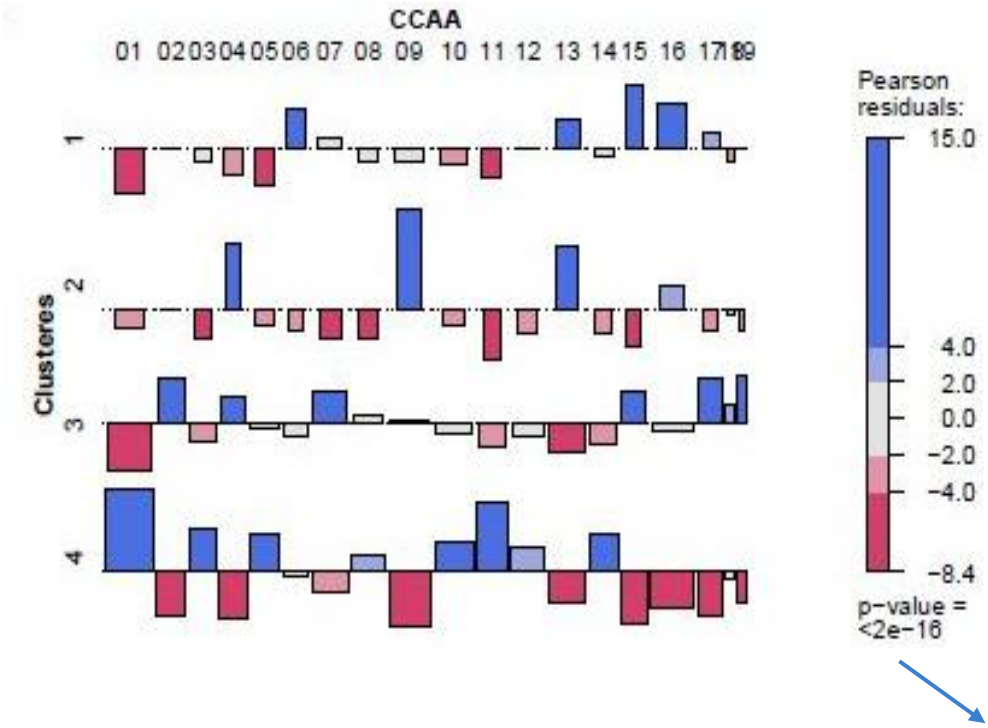
- **Frecuencias relativas:**

1	2	3	4	5	6
0,11	0,04	0,04	0,04	0,05	0,04
7	**8**	**9**	**10**	**11**	**12**
0,07	0,06	0,09	0,08	0,05	0,06
13	**14**	**15**	**16**	**17**	**18**
0,07	0,04	0,04	0,10	0,03	0,01
19					
0,01					

- **Frecuencias relativas por grupos:**

GRUPOS	1	2	3	4	5	6
Grupo 1	0,07	0,05	0,03	0,03	0,03	0,06
Grupo 2	0,10	0,05	0,02	0,08	0,04	0,02
Grupo 3	0,08	0,06	0,03	0,05	0,05	0,03
Grupo 4	0,16	0,03	0,05	0,02	0,06	0,03
GRUPOS	**7**	**8**	**9**	**10**	**11**	**12**
Grupo 1	0,07	0,05	0,08	0,07	0,03	0,06
Grupo 2	0,04	0,04	0,18	0,07	0,01	0,04
Grupo 3	0,08	0,06	0,09	0,07	0,04	0,06
Grupo 4	0,06	0,06	0,06	0,09	0,07	0,07
GRUPOS	**13**	**14**	**15**	**16**	**17**	**18**
Grupo 1	0,09	0,04	0,07	0,13	0,04	0,00
Grupo 2	0,12	0,03	0,02	0,12	0,02	0,00
Grupo 3	0,06	0,03	0,05	0,09	0,05	0,01
Grupo 4	0,06	0,05	0,02	0,08	0,02	0,00
GRUPOS	**19**					
Grupo 1	0,01					
Grupo 2	0,00					
Grupo 3	0,01					
Grupo 4	0,00					

- **Diagrama de asociación:**



Chi cuadrado

Número de ocupados en el hogar (NUMOCU)

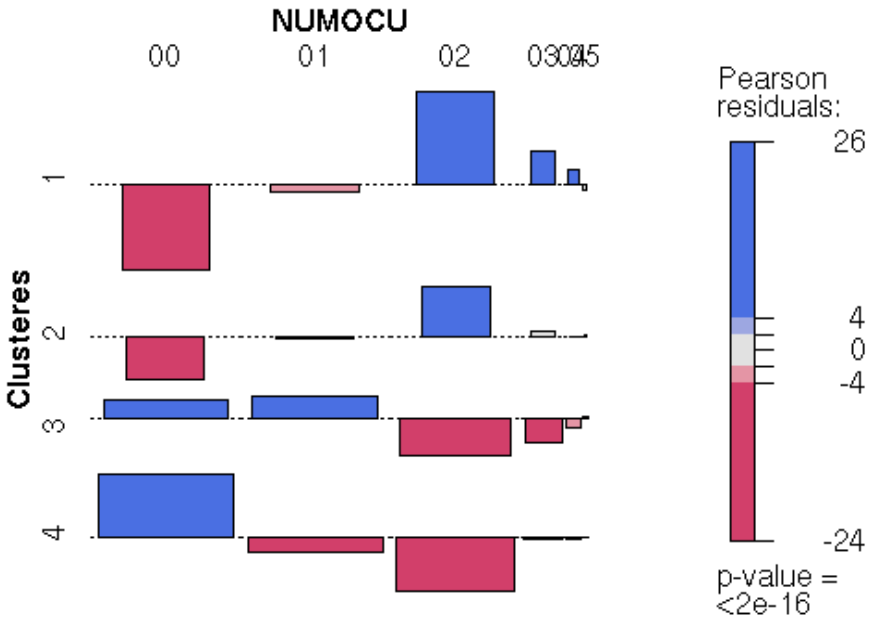
- **Frecuencias relativas:**

0	1	2	3	4	5
0,35	0,35	0,27	0,03	0,00	0,00

- **Frecuencias relativas por grupos:**

GRUPOS	0	1	2	3	4	5
Grupo 1	0,11	0,33	0,49	0,06	0,01	0,00
Grupo 2	0,21	0,34	0,40	0,03	0,00	0,00
Grupo 3	0,38	0,39	0,20	0,02	0,00	0,00
Grupo 4	0,46	0,32	0,18	0,03	0,00	0,00

- **Diagrama de asociación:**



Chi cuadrado

Nivel de ingresos del sustentador principal (INTERIN)

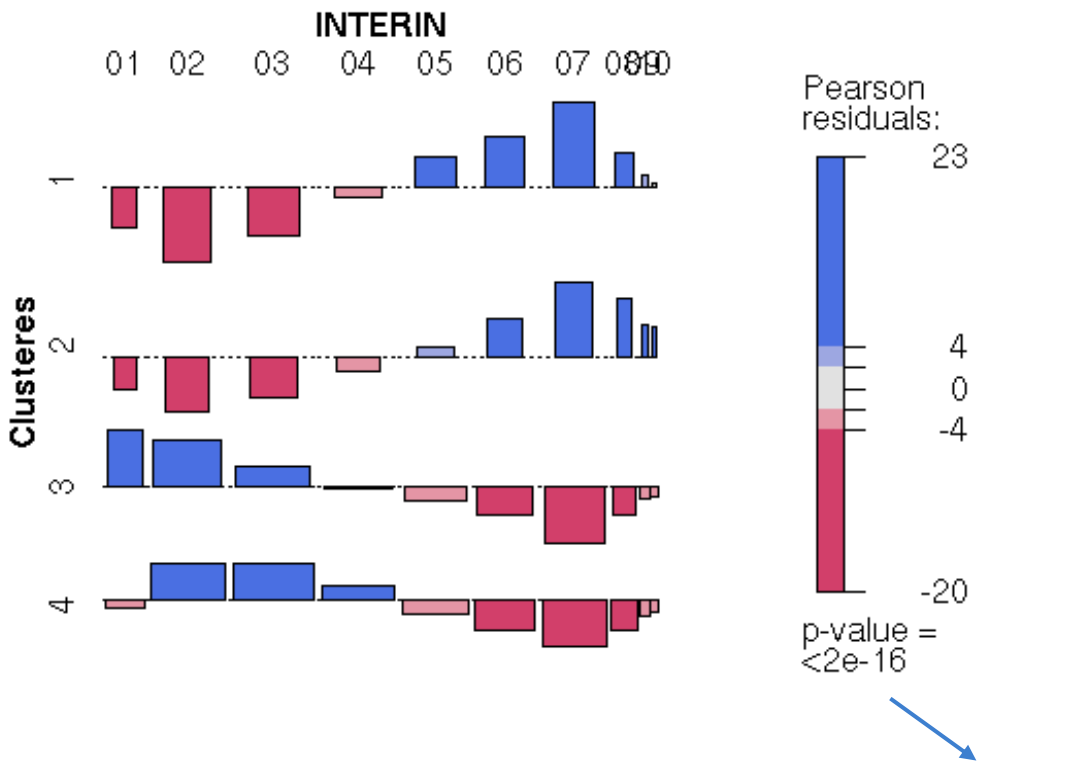
- **Frecuencias relativas:**

1	2	3	4	5	6
0,05	0,17	0,20	0,17	0,14	0,11
7	**8**	**9**	**10**		
0,13	0,02	0,00	0,00		

- **Frecuencias relativas por grupos:**

GRUPOS	1	2	3	4	5	6
Grupo 1	0,01	0,04	0,11	0,15	0,19	0,19
Grupo 2	0,01	0,06	0,11	0,14	0,15	0,18
Grupo 3	0,08	0,23	0,23	0,17	0,12	0,08
Grupo 4	0,04	0,22	0,25	0,19	0,12	0,08
GRUPOS	**7**	**8**	**9**	**10**		
Grupo 1	0,27	0,04	0,01	0,00		
Grupo 2	0,26	0,06	0,01	0,01		
Grupo 3	0,07	0,01	0,00	0,00		
Grupo 4	0,08	0,01	0,00	0,00		

- **Diagrama de asociación:**



Chi cuadrado