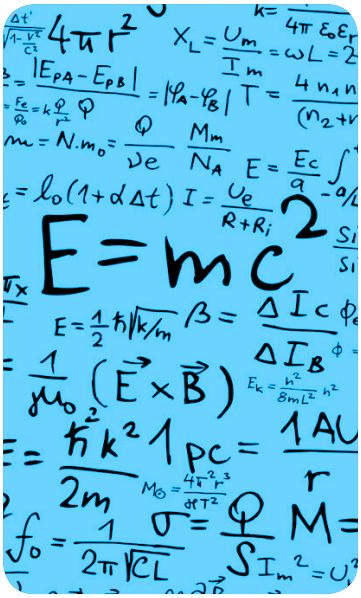
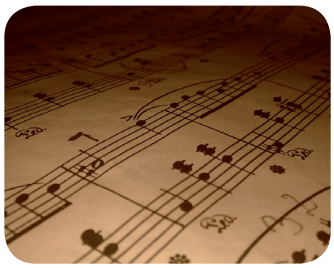




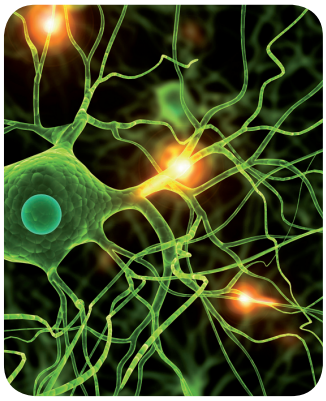
MÁSTERES de la UAM

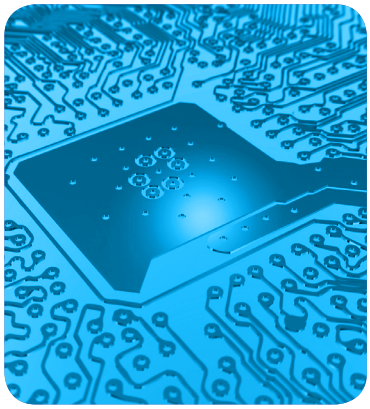
Facultad de Psicología /11-12

Máster en Metodología de las Ciencias del Comportamiento y de la Salud

UNIVERSIDAD AUTONOMA DE MADRID

excelencia Campus Internacional UAM CSIC+





UAM EDICIONES

Análisis y valoración de técnicas de clasificación de hechos delictivos y autores imputados, y correspondencia entre soluciones para el Perfilado Criminológico. Aplicación a los incendios forestales es España

Santiago Fernandez Dapena

Resumen

Objetivos: Más de la mitad de los incendios forestales en España son intencionados, pero tan sólo se llega a identificar a sus autores en un 2% de los casos. A partir de la base de datos sobre incendios forestales recopilada por las fuerzas y cuerpos de seguridad del estado se ha diseñado una metodología de análisis basada en el análisis de conglomerados, que permite la elaboración de tipologías de hechos y tipologías de incendiarios, estableciendo posteriormente mediante el análisis de correspondencias las relaciones vinculantes entre ambas tipologías. Se pretende elaborar un modelo que permita predecir las características psicosociales de un autor desconocido a partir de las evidencias halladas en la escena del incendio.

Método: Se han analizado los datos de una muestra de 300 incendios, correspondientes a los incendios con imputado detenido durante los años 2009 y 2010. El modelo se ha validado con los incendios del 2011. Se han utilizado como técnicas fundamentales los análisis de conglomerados bietápico, de k-medias y jerárquico, y el análisis de correspondencias. Se han realizado simulaciones bootstrap para comprobar la robustez de las reglas estadísticas de decisión y se propone un estadístico para valorar la concordancia entre soluciones.

Resultados: Tras la depuración previa de los datos, la tipificación de los incendios ha dado lugar a 7 grupos, en función de la información recopilada en el lugar del siniestro, y 11 perfiles psicosociales de sujetos incendiarios, en función de las características de los imputados. Se ha detectado que no siempre existe una agrupación analíticamente óptima y que, cuando existe, no tiene por qué ofrecer la mejor solución práctica al problema teórico planteado, siendo necesario recurrir a los presupuestos teóricos y a los objetivos del estudio para encontrar la solución más útil. Se ha detectado un efecto importante del orden de los casos en la creación de las tipologías y se propone una metodología de análisis para minimizar este efecto.

Conclusiones: Si bien ha sido posible identificar los problemas de análisis más importantes y se ha podido establecer un procedimiento de análisis robusto para este tipo de problemas de investigación, los resultados obtenidos deben considerarse provisionales, debido al reducido tamaño de la muestra y su incidencia sobre los grados de libertad de los modelos estimables.

Palabras clave: análisis de conglomerados, perfiles criminales, análisis de correspondencia, clasificación de incendios intencionados, técnicas de clasificación, concordancia de soluciones, influencia del orden de los casos en la solución.

Análisis y Valoración de Técnicas de Clasificación de Hechos Delictivos y Autores Imputados, y Correspondencia entre Soluciones para el Perfilado Criminológico. Aplicación a los Incendios Forestales en España.

El incendio forestal supone uno de los principales problemas de tipo ecológico y social y es la cuestión medioambiental que más preocupa a la población en España (González, Sotoca, Martínez y Martín, 2010). Tal y como definen estos autores, un incendio forestal es *"el fuego que se extiende sin control en terreno forestal y que afecta a combustibles vegetales que no estaban destinados a arder"*. En el año 2010 el número de incendios forestales ascendió a 11.722 con una superficie afectada de 54.770 hectáreas. Según los datos del Ministerio de Medio Ambiente los siniestros intencionados supusieron el 60,36% del total, seguido de los causados por negligencias o accidentes con un 23,74%. Sólo en un 1,7% de los incendios intencionados se llegó a la detención del autor (MARM, 2010), lo que representa un grave obstáculo en la lucha contra incendios. La Fiscalía de Sala de Medio Ambiente y Urbanismo de la Fiscalía General del Estado ordenó la realización de estudios conducentes a la elaboración del perfil del incendiario forestal en España (González et al., 2010).

El perfilado de delincuentes

El concepto de perfilado de delincuentes se puede definir como la acción de inferir las características de personalidad, conductuales y socio-demográficas de un delincuente basándose en el análisis de las evidencias obtenidas en la escena de un delito (Kocsis, Middledorp y Karpin, 2008; Snook, Cullen, Bennell, Taylor y Gendreau, 2008; Hicks y Sales, 2006; Hakkanen, Puolakka y Santtila, 2004; Alison, Bennell, Mokros y Ormerod, 2002; Kocsis y Cooksey, 2002; Canter, 2000). Aunque ésta es una tarea inherente a los cuerpos y fuerzas de seguridad, son muchos los países (como EEUU, Inglaterra, Australia, Finlandia o Canadá), que recurren habitualmente a profesionales externos para la elaboración de perfiles de delincuentes, especialmente en crímenes con un alto grado de violencia, crueldad o delitos en serie (Alison, Goodwill, Almond, Van de Heuvel y Winterl, 2010; Hicks y Sales, 2006; Kocsis, 2006; Canter, 2000).

Entre las principales corrientes teóricas existentes en la actualidad destacan, la Investigación para el Análisis Criminal (*Criminal Investigative Analysis,* CIA) desarrollada por la Unidad de Análisis del Comportamiento de la Oficina Federal de Investigación (*Behavioral Science Unit of the Federal Bureau of Investigation*, FBI); el Análisis de Evidencias del

Comportamiento (*Behavioral Evidence Analisis,* BEA) desarrollado por Brent Turvey; el Perfilado de Acción Criminal (*Crime Action Profiling*, CAP) de Richard Kocsis y la Psicología Investigadora (*Investigative Psichology*, IP) de David Canter. Una descripción detallada se puede encontrar en Hicks y Sales (2006) y en Kocsis (2006).

Estas corrientes teóricas se agrupan en tres aproximaciones al problema del perfilado criminal, la aproximación criminal-investigadora que basa la eficacia del perfilado en la capacidad, conocimiento y experiencia de los investigadores policiales como la CIA; la aproximación clínica-facultativa que basa la eficacia del perfilado en los conocimientos y experiencia del psicólogo colaborador como la BEA; y la aproximación estadística, basada en el análisis multivariante de la información obtenida en la escena del crimen, como la CAP y la IP (Alison et al., 2010; Snook et al., 2008; Alison, Goodwill y West, 2004). Esta última es la más extendida en la actualidad. (Canter, 2004, 2011; Alison et al., 2010; Doley, 2003; Kocsis y Cooksey, 2002; Alison et al., 2002).

En cuanto al delito específico del incendio se han realizado numerosos esfuerzos para obtener modelos que permitan clasificar a los incendiarios en distintas tipologías, así como proporcionar evidencia que sustente un determinado marco teórico (Muller, 2008; Kocsis, 1998, 2002, 2004; Häkkänen, et al., 2004; Doley, 2003; Santtila, Häkkänen y Fritzon, 2003; Fritzon y Canter, 2001; Canter y Fritzon, 1998; Kocsis, Irwin y Hayes, 1998).

En la actualidad y ante la carencia de evidencia empírica se está poniendo en duda la validez científica de las teorías sobre las que se sustenta el perfilado de delincuentes (Devery, 2010; Snook et al., 2008; Dowden, Bennell y Bloomfield, 2007; Woodhams y Toye, 2007; Eastwood, Cullen, Kavanagh y Snook, 2006).

La primera aproximación al problema del perfilado psicosocial del incendiario forestal en España la encontramos en González et al., (2010), en el que siguiendo la metodología empleada en la literatura científica internacional realizan una aproximación al problema mediante el escalamiento multidimensional (Multidimensional Scaling, MDS). Sin embargo, el MDS sufre de una excesiva sensibilidad a los casos atípicos (es un análisis de inercia y no un análisis de masa), la solución obtenida puede resultar difícilmente interpretable y generalizable y las fronteras establecidas en los gráficos bidimensionales deben decidirse de manera subjetiva (Jaworska y Chupetlovska-Anastasova, 2009). Dadas las limitaciones de las estrategias de análisis empleadas hasta la fecha proponemos una nueva aproximación para la agrupación de hechos y autores, basada en técnicas de análisis de conglomerados más recientes, como el análisis de conglomerados bietápico y la creación de reglas de producción.

El análisis de conglomerados es un conjunto de técnicas multivariantes dirigidas a la agrupación de objetos en grupos denominados conglomerados, en función de las características que comparten. La técnica pretende encontrar una estructura natural en los datos que permita el agrupamiento de los objetos, de manera que los miembros de un conglomerado sean más similares entre sí (varianza intra-conglomerado) que entre conglomerados diferentes (varianza inter-conglomerados). Por tanto, el análisis se basa en la valoración de las distancias entre los objetos y los conglomerados, siendo de fundamental importancia la elección de la medida de distancia adecuada. El análisis de conglomerados tiene una serie de limitaciones: a) es una técnica descriptiva, no inferencial; b) siempre va a crear conglomerados independientemente de la existencia o no de una estructura natural en los datos; y c) la solución obtenida puede no ser generalizable pues depende de las variables, de la medida de distancia y del método de aglomeración empleados en la conglomeración (Fielding, 2007; Hair, 1995).

Los métodos tradicionales de análisis de conglomerados y sus variaciones funcionan relativamente bien cuando todas las variables son continuas, no así cuando las variables son categóricas o cuando hay mezcla de ambos tipos (Chiu, Fang, Chen, Wang y Jeris, 2001; Guha, Rastogi y Shim, 1999). Un algoritmo ampliamente empleado para la agrupación de objetos utilizando únicamente variables medidas en escala nominal o en combinación con variables continuas, es el análisis de conglomerados en dos fases, bietápico o Twostep Cluster Analysis.

El análisis de conglomerados bietápico implementado en el SPSS emplea el algoritmo BIRCH, diseñado por Zhang, Ramakrishnan y Livny, (1996) como un algoritmo eficiente para la conglomeración de bases de datos muy grandes con variables cuantitativas (IBM SPSS, 2010a; Bacher, Wenzig y Vogler, 2004; Chiu et al., 2001). El Twostep cluster se realiza en dos fases, en una primera etapa procede a una única lectura de los datos y crea una primera conglomeración secuencial denominada pre-cluster. El número de conglomerados del pre-cluster depende de la capacidad de cálculo y memoria de computación asignada. En la segunda fase y mediante un procedimiento jerárquico realiza la conglomeración de los pre-clusters para obtener los *k* mejores conglomerados. Posteriormente con estos conglomerados se calculan los *k* centroides de los conglomerados definitivos que se obtienen al reasignar cada caso a un conglomerado en función de su proximidad a los centroides (Bacher et al., 2004; Zhang et al., 1996).

Chiu et al., (2001) añadieron al algoritmo inicial diseñado por Zhang et al., (1996) la capacidad de incluir variables categóricas en el análisis y la capacidad de selección automáti-

ca del número de conglomerados. La medida de distancia propuesta está relacionada con el decremento en el logaritmo de la verosimilitud del modelo de *k* conglomerados con respecto al modelo de *k*-1 conglomerados (tras la fusión de los dos más próximos).

Sea $\boldsymbol{x}= \{x_i: i= 1, 2,\ldots, N_k\}$ el conjunto de N_k casos *p* dimensionales, (*p* variables categóricas), pertenecientes al conglomerado *k* (*k*=1, 2,…, *K*), sea $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$ la función de densidad de probabilidad de $\boldsymbol{x}$ en el conglomerado *k* y $\boldsymbol{\theta}_k$ el vector de parámetros del modelo, el logaritmo de la función de verosimilitud de la clasificación toma la forma:

$$l = \sum_{k=1}^{K} \sum_{i \in I} \log\big(p(x_i|\theta_k)\big) = \sum_{k=1}^{K} \xi_k \quad (1)$$

donde ξ_k es la aportación de cada conglomerado al logaritmo de la función de verosimilitud. Cuando se trabaja únicamente con variables categóricas y al emplear el método de estimación por máxima verosimilitud se asumen los supuestos de independencia y distribución multinomial de los casos en las categorías de cada variable, por lo que se define $\xi_k = -n_k \cdot \sum_{p=1}^{P} E_{kp}$ donde E_{kp} es la entropía de la variable *p* en el conglomerado *k* definida como $E_{kp} = -\sum_{c=1}^{C_j} \hat{\pi}_{kpc} \cdot \log\big(\hat{\pi}_{kpc}\big)$, siendo $\hat{\pi}_{kpc}$ la probabilidad de que un sujeto del conglomerado *k* tome la categoría *c* en la variable *p*. Observamos que ξ_k es una medida de dispersión total intracluster o entropía intracluster, y por tanto *l* es una medida de entropía global intracluster. Si partimos de la solución de *k* conglomerados, la distancia entre los conglomerados *i* y *s* que se unen formando un nuevo conglomerado $\langle i, s\rangle$ está definida como:

$$\text{d(i,s)} = l_k - l_{k-1} = \sum_{k=1}^{K} \xi_k - \sum_{k=1}^{K-1} \xi_k = \xi_i + \xi_s - \xi_{\langle i,s\rangle} \quad (2)$$

donde $\xi_v = n_v \cdot \sum_{p=1}^{P} \sum_{c=1}^{C_j} \hat{\pi}_{vpc} \cdot \log\big(\hat{\pi}_{vpc}\big)$ la función distancia así calculada es no negativa y simétrica, por lo que observamos que una fusión de conglomerados irá siempre acompañada de un incremento en la entropía.

El análisis de conglomerados no proporciona una solución única y la solución obtenida está afectada por una serie de decisiones previas que deben ser tomadas por los investigadores y que influirán en la solución final, tanto en la elección de las variables finales como en la decisión del número de conglomerados finales.

Elección de variables para la conglomeración.

Numerosos autores hacen referencia a la llamada "The curse of dimensionality" (Bellman, 1961), que hace referencia a que cuando hay demasiadas variables se reduce el rendimiento en la conglomeración debido a que un exceso de variables irrelevantes puede provo-

car que casos de distintos grupos sean cada vez más parecidos al homogeneizarse las distancias entre ellos (Müller, Günnemann, Assent, y Seidl, 2009; Moise, Sander y Ester, 2008; Beyer, Goldstein, Ramakrishnan, y Shaft, 1999; Fielding, 2007; Chizi y Maimon, 2005). Por otro lado, un número excesivo de variables y de categorías por variable puede provocar que el número de parámetros a estimar sea superior al número de sujetos existente en la muestra dando lugar a modelos infraidentificados (con un número negativo de grados de libertad). Una reducción del número de variables a aquellas que son realmente importantes en la conglomeración influirá positivamente en la estabilidad. La medida de importancia relativa de las variables ha sido definida en la versión 19 de SPSS como:

$$IV_i = \frac{-\log_{10}(sig_i)}{\max_{j\in\Omega}(-\log_{10}(sig_j))} \quad (3)$$

donde Ω representa el conjunto de variables incluidas, sig=Prob(χ^2_d>X^2) siendo $X^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(N_{ij}-\widehat{N_{ij}})^2}{\widehat{N_{ij}}}$ donde N_{ij} es el número de casos observados en la categoría *i* de la variable y el cluster *j* de la conglomeración, y $\widehat{N_{ij}}$ el número de casos esperado en el cruce categoría *i* conglomerado *j*. Este algoritmo escala la importancia de manera que, en cada análisis, se asigna el valor 1 a la variable más importante y es el valor de comparación. El valor IV (importancia de la variable) así calculado representa la importancia relativa de cada variable frente a la importancia de la variable más importante en una ejecución dada del algoritmo de conglomeración (IBM SPSS, 2010a).

Número de conglomerados finales.

El procedimiento bietápico ofrece dos medidas de desajuste global en las que basar los criterios de decisión automática para determinar el número de conglomerados finales, el Criterio de Información Bayesiano (BIC) y el Criterio de Información de Akaike (AIC) (Bacher et al., 2004; Chiu et al., 2001). El uso del AIC no asume la existencia de un modelo real generador de los datos. El AIC estima la perdida de información experimentada cuando empleamos un modelo determinado para reproducir la realidad y penaliza por el número de parámetros independientes *p* del modelo, se calcula como AIC = -2*l* + 2p, siendo *l* el logaritmo de la verosimilitud. El modelo con menor AIC será el modelo con menor perdida de información, es decir, el modelo que mejor se ajusta a los datos empíricos (Burnham y Anderson, 2004). El empleo del BIC asume la existencia de un modelo real que es el buscado y penaliza por el numero de parámetros y por el tamaño de la muestra, se calcula como BIC= -2*l* + p·Ln(n). El

modelo con menor BIC será por tanto el modelo más próximo al modelo real. El empleo del BIC reproduce modelos demasiado parsimoniosos para muestras pequeñas o moderadas (Burnham y Anderson, 2004).

La versión 19 de SPSS estima el número final de conglomerados en dos etapas, en la primera etapa calcula el BIC (o el AIC) de cada uno de los modelos solicitados (por defecto desde la solución con k=15 conglomerados hasta la solución con k=1), calculando posteriormente el cambio en BIC entre cada pareja de modelos consecutivos como la diferencia en BIC entre el modelo de k conglomerados y el de k+1, dBIC(k)=BIC(k)-BIC(k+1). A continuación efectúa un cambio de escala calculando la razón de cambio en BIC como el cociente entre dBIC(k) y dBIC(1),

$$R(k) = \frac{dBIC(k)}{dBIC(1)} \tag{4}$$

donde dBIC(1) es el cambio en la pérdida de ajuste al fundir los dos últimos conglomerados (k+1=2) en un único conglomerado global (k=1).

La estimación inicial del número de conglomerados viene establecida por todas las soluciones desde k=2 conglomerados hasta el número K de conglomerados tal que la ratio R(k) sea todavía mayor que 0,04. En la segunda etapa, y para todas las soluciones que cumplen la condición anterior, se calcula el cambio en la distancia inter-conglomerados en cada paso como la razón entre las distancias de cada pareja de fusiones consecutivas,

$$Rd(k) = \frac{d_{min}(k)}{d_{min}(k+1)} \tag{5}$$

donde $d_{min}(k)$ es la mínima distancia inter-conglomerados para los k conglomerados de la etapa. Finalmente se comparan las dos mayores razones de distancia de entre las k soluciones de la estimación inicial como:

$$R = \frac{Rd(kv)}{Rd(kt)} \tag{6}$$

seleccionando el modelo del numerador si la razón de razones de medida de distancia entre estos dos modelos es mayor de 1,15 o el de mayor número de conglomerados si ésta es menor (IBM SPSS, 2010a). Independientemente de que estos cálculos de los que informa el SPSS en su documentación no son los que refleja en la tabla de salida de datos, estos valores límite, (0,04 y 1,15), empleados como puntos de corte en las etapas de selección automática de conglomerados, están establecidos por simulaciones realizadas por el equipo de SPSS. No se recomienda su empleo con bases de datos con variables de tipo mixto o categóricas, o en presencia de conglomerados con solapamiento (Bacher et al., 2004).

Influencia del orden de los casos.

Como sucede con el algoritmo de conglomerado de k medias, el algoritmo BIRCH se ve afectado por el orden de los casos (IBM SPSS, 2010a, 2010b), aunque esta influencia es menor que con otros algoritmos (Zhang et al., 1996), no ha sido cuantificada hasta ahora. Para minimizar este efecto SPSS propone la reordenación aleatoria previa de los datos, o el diseño de varios modelos con distintas reordenaciones y valorar la mejor solución de las obtenidas. La decisión de qué variables incluir en el modelo y del número final de conglomerados es determinante para la solución final, por lo que propondremos un análisis de sensibilidad al orden para analizar su influencia en la selección de las variables y en el número final de conglomerados.

El objetivo principal de este estudio es el diseño de una metodología basada en el análisis de conglomerados que sea efectiva cuando enfrentemos bases de datos de delitos y delincuentes y que permita extraer la información relevante para elaborar un modelo predictivo. Para ello realizaremos una aproximación metodológica que permita, analizando las variables del hecho delictivo y del autor del delito establecer una vinculación entre ellas. De esta forma se pretende explorar la existencia de una estructura natural en los hechos y en los autores que nos permita aproximar las características sociodemográficas de un autor en función de los aspectos del hecho que comete, con el objetivo final de orientar a las fuerzas de seguridad en la dirección más probable en la búsqueda de un delincuente a partir de las evidencias objetivas obtenidas en la escena del delito. Otro reto adicional planteado por la investigación es la representación adecuada de las soluciones, de manera que estas sean fáciles de interpretar a la vez que informen tanto de las características de los conglomerados (la descripción de los perfiles) como de las diferencias existentes entre los conglomerados.

Método

Muestra

La muestra analizada es de tipo incidental y consta de 300 incendios forestales ocurridos en España entre los años 2009 y 2010 en los que el autor ha sido detenido e imputado. Esta muestra debe considerarse una muestra clínica, puesto que no ha habido selección ni asignación aleatorias. La muestra no es representativa de la población de incendios forestales en España, sino de los que son policialmente resueltos. Todas las variables están completas y no existen valores perdidos. La opción de respuesta “no se sabe” o “desconocido” está considerada como una categoría más de análisis. Todas las variables son categóricas y se describen

en el Apéndice I. Los datos han sido tratados de forma anónima, no existe información en el archivo de datos que permita la identificación de ninguno de los autores y se solicitó el consentimiento informado por escrito a cada autor antes de la recogida de los datos personales.

Instrumentos

Los datos han sido recogidos mediante el Cuestionario para la Investigación del Perfil del Incendiario Forestal V2009b, elaborado ad hoc por la Unidad Técnica de Policía Judicial (UTPJ) de la Guardia Civil. La primera versión de este cuestionario fue realizado por un grupo de expertos formado por psicólogos de la UTPJ en colaboración con investigadores especializados en incendios forestales, partiendo del modelo de Viegas y Soeiro (2007). La versión actual, V2009b, es una modificación de la versión inicial a partir del estudio de la muestra de incendios de los años 2007, 2008 y parte del 2009 (González et al., 2010). El cuestionario consta de 68 preguntas cerradas de elección múltiple (entre 3 y 10 categorías de respuesta).

La información de los cuestionarios se ha codificado y volcado en un archivo de datos en formato SPSS para su depuración y análisis posterior. El archivo de datos está compuesto por 3 bloques de información. El primer bloque está formado por 5 variables de control e identificación de los cuestionarios. El segundo bloque comprende 20 variables sobre evidencias recogidas en la escena del incendio, a cumplimentar por los responsables de la investigación policial. El tercer bloque lo forman 43 variables sobre el autor, a cumplimentar por quien participe en la toma de declaración policial.

Método de análisis

Análisis descriptivo y depuración de datos. Se llevó a cabo una descripción univariante de todas las preguntas del cuestionario para detectar respuestas inadecuadas, errores de grabación y valores en blanco. Las respuestas en blanco fueron imputadas en la categoría No sabe/No contesta. Esta categoría se considera una categoría de respuesta más puesto que en el proceso de investigación puede faltar información, que es desconocida para el investigador, y sin embargo constituir un patrón de respuesta persistente.

En lo que respecta a las variables del autor, se adoptó como criterio descartar aquellos casos en los que más del 60% de las respuestas (incluido el sexo) correspondían a la categorías NS/NC, por considerarse que la información carecía de la calidad suficiente.

Análisis de agrupación. La estrategia general de análisis consistió en el análisis independiente de las variables del hecho (espacio de los hechos) y de las variables del autor (espacio de los autores) para, posteriormente, fundir ambos espacios en un análisis conjunto capaz de detectar las relaciones existentes entre ambos espacios (si es que las hubiera).

La técnica de agrupación elegida para la aglomeración de los casos fue el Análisis de Conglomerados en dos Fases. Se eligió esta técnica por su capacidad para gestionar variables categóricas (escala nominal), si bien se utilizaron otras técnicas concurrentes en la fase de validación.

Previamente a la elección de la solución final se valoraron dos aspectos importantes que pueden influir o incluso determinar la calidad de la solución: el número de variables utilizadas en la conglomeración y el número de conglomerados óptimo para formar la solución final. Se analizó la influencia del orden de los casos en la conglomeración, en la selección de variables y en la selección del número final de conglomerados.

Análisis de sensibilidad al orden de los casos en la conglomeración. Hemos definido un índice de concordancia (IC) entre conglomeraciones con igual número de clusters, como el porcentaje de casos que en dos conglomeraciones distintas han sido asignados al mismo conglomerado relativo. Para calcular el IC partimos de dos conglomeraciones realizadas bajo los mismos parámetros y en las que la única diferencia es la reordenación de los casos de forma aleatoria. Si la hipótesis de robustez al orden es cierta, un porcentaje muy elevado de los sujetos asignados a cada conglomerado $1, 2 \ldots K$ en la conglomeración de anclaje C_1, se corresponderán en la conglomeración C_2 de comprobación a los conglomerados 1', 2'…K' que pueden haber recibido un número de identificación del conglomerado distinto y estar ordenados de manera diferente. Dentro de los miembros n_k de cada conglomerado de la primera solución se calcula la moda para el número asignado al conglomerado de pertenencia en la conglomeración 2 y realizamos una comparación lógica asignando un 1 a todos los sujetos que perteneciendo al conglomerado *k* en la conglomeración 1 pertenecen a la categoría modal en la conglomeración 2, y 0 a los que no cumplen esta condición de coincidencia.

Tabla 1. Ejemplo de cálculo del Índice de Concordancia.

Conglomeración	Conglomerado A					Conglomerado B										Conglomerado C				
C_1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3
C_2	3	3	3	3	2	1	1	1	1	3	1	1	1	3	1	2	2	2	2	1
MODA C_2			3							1								2		
Concordancia	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0
IC	80%																			

Para una solución de *K* conglomerados, con un tamaño para cada conglomerado de n_k = (n_1,…, n_K), el índice de concordancia global se calcula como la siguiente suma lógica (Verdadero=1, Falso=0) para todos los miembros de un conglomerado en la solución inicial de anclaje:

$$IC = \sum_{j=1}^{K}\sum_{i=1}^{n_k}\frac{\left[X'_{ij} = Moda\left(X'_{ij}\right)\right]}{n_k} \quad j = 1,\ldots,K; i = 1,\ldots,n_k$$

donde X'_{ik} es el valor de la categoría del conglomerado de pertenencia en la solución de comprobación para cada caso *i* del conglomerado *k* en la solución de anclaje. Si ordenamos los conglomerados por el valor de la categoría de pertenencia de la solución de anclaje, el estadístico se simplifica como:

$$IC = \sum_{j=1}^{K}\sum_{i=1}^{n_k}\frac{\left[j = Moda\left(X'_{ik}\right)\right]}{n_k} \quad j = 1,\cdots,K; i = 1,\cdots,n_k$$

El cálculo del IC se lleva a cabo mediante sintaxis de SPSS diseñada al efecto. Se han realizado 20 conglomeraciones distintas y comparado cada una de ellas con las 19 restantes obteniendo un total de 190 comparaciones, tanto para el espacio del hecho como del autor.

Elección de variables para la conglomeración. Una vez valorado el efecto del orden de los casos en la conglomeración, diseñamos un procedimiento que realiza 100 reordenaciones aleatorias de los casos, y para cada una de estas reordenaciones se realiza un análisis bietápico con todas las variables del hecho (o autor, según corresponda). Para valorar la contribución de cada variable en la solución obtenida calcularemos el valor de la importancia según la fórmula (3). La IV varía dependiendo del número de conglomerados de la solución solicitada, por lo que este procedimiento se repite para todas las soluciones de 2 a 15 conglomerados. Como estadístico de resumen para comparar la importancia relativa de las variables se calcula el porcentaje de veces en que cada variable presenta una importancia relativa menor a 0,4 en cada solución.

Número de conglomerados finales. Emplearemos el BIC como criterio de selección por incluir una corrección por tamaño muestral que consideramos necesaria. La fórmula de cálculo del BIC del análisis bietápico es BIC= $-2\cdot\sum_{k=1}^{K}\xi_k + r_k \cdot \mathrm{Ln}(n)$ siendo r_k el número de parámetros independientes calculado como $r_k = K\cdot\sum_{p=1}^{P}\left(C_p - 1\right)$ donde C_p es el número de categorías *C* de la variable categórica *P*. Calcularemos el número de parámetros r_k y su influencia en la identificación de los modelos, debido a que un número excesivo de parámetros acompañado de un tamaño muestral pequeño puede provocar la falta de grados de libertad suficientes en la estimación.

Hemos diseñado un procedimiento que realiza 100 reordenaciones aleatorias de los casos, y para cada una de estas reordenaciones realiza un análisis bietápico para todas las soluciones de 2 a 15 conglomerados con las variables definitivas del hecho (o autor, según corresponda). Mediante la orden /PRINT IC y el sistema de gestión de resultados (SGR), capturamos para su análisis y representación gráfica los valores de los estadísticos: BIC, cambio en BIC, razón del cambio en BIC y razón de medidas de distancia para cada una de las soluciones de *k* conglomerados.

Con el objeto de obtener evidencias adicionales que den soporte a la decisión sobre el número de conglomerados, hemos empleado otras técnicas de agrupación. Hemos dicotomizado todas las variables finales seleccionadas para el análisis y realizado un análisis de conglomerados jerárquico por el método de Ward y con la distancia euclídea al cuadrado.

Solución final. Los análisis realizados se presentaron como evidencias a un panel compuesto por seis expertos en psicología, metodología, criminología e investigación policial, tras el que se tomó la decisión final. Se incluyeron en el modelo final todas aquellas variables que fueron consideradas relevantes por al menos dos jueces. En cuanto al número de conglomerados se analizaron varias soluciones, eligiendo la mas interpretable.

Una vez sometida toda la información al panel de expertos, decididas las variables finales y el número de conglomerados de hecho y de autor, para obtener el modelo definitivo se realizó una nueva reordenación aleatoria de los datos y se llevó a cabo la conglomeración definitiva de los hechos y de los autores (por separado).

Relación entre el espacio de los hechos y el de los sujetos. Se valoró el solapamiento y la relación existente entre los conglomerados de hechos y los conglomerados de autores mediante el análisis de correspondencias simple utilizando las variables categóricas que contienen los conglomerados de pertenencia de hecho y de autor. Se valoró el número de dimensiones necesarias para explicar las relaciones existentes a partir del porcentaje de inercia explicado por el número de dimensiones elegido.

Todos los análisis se han llevado a cabo mediante el procedimiento Análisis de Conglomerados Bietápico y el resto de rutinas implementadas en el paquete estadístico IBM-SPSS V19. Este procedimiento tiene una serie de opciones que hemos configurado de la siguiente forma: no realizar tratamiento de ruido, umbral del cambio en distancia inicial 0, capacidad máxima de memoria empleada 1024, máximo número de ramas 16 y profundidad máxima del árbol 16.

Validación. Para validar el modelo desarrollado con la muestra de 2009 y 2010, se recogieron, depuraron e incorporaron al archivo de datos los cuestionarios de los sujetos imputados durante el año 2011 por algún incendio forestal. Hemos utilizado como modelo base el análisis de conglomerados bietápico y la solución de 7 (hecho) y 11 (autor) conglomerados.

Para poder utilizar el modelo a nivel predictivo, se filtraron los casos anteriores al 2011 y se guardó el modelo de puntuación (scoring) mediante la sub-sentencia OUTFILE MODEL. El conglomerado de pertenencia se guardó también como una variable adicional.

Una vez guardado el modelo de puntuación, se utilizó para predecir el grupo de pertenencia de los nuevos casos. Para los casos ya analizados, la predicción coincidirá con el conglomerado asignado y almacenado anteriormente. La puntuación se lleva a cabo con la sentencia MODEL HANDLE, que puede construirse utilizando el "Asistente de puntuación" ubicado en el menú Utilidades. Su ejecución creará las variables pronóstico que contendrán la nueva asignación incluyendo los casos de la muestra de validación.

Una vez obtenidas las variables pronosticadas hemos comparado las soluciones de la muestra de análisis y de la muestra de validación y hemos puesto a prueba la hipótesis de homogeneidad de las distribuciones. Este análisis se ha realizado por separado para los espacios del hecho y del autor.

Posteriormente hemos obtenido una nueva solución de 7 conglomerados del hecho y 11 del autor empleando la muestra completa, y hemos calculado el índice de concordancia entre los modelos de 2010 y 2011. Finalmente y tras dicotomizar las variables, hemos realizado un análisis de conglomerados con el procedimiento de *k*-medias comparando la solución obtenida en ambos espacios con la solución obtenida con el análisis bietápico.

Resultados

Descripción de las variables

Tras el análisis univariante de las variables se eliminaron un total de 33 casos por falta de información en las respuestas, con lo que la muestra quedó reducida a 267 casos. El 60% de los incendios pertenece a la campaña de 2009 y el 40% a la campaña de 2010. El 68% de los cuestionarios fueron cumplimentados por la Guardia Civil, el 25% por policías autonómicas y el 5% por agentes forestales. Por comunidades autónomas, el 48% de los incendios ocurrieron en Galicia, el 15% en Andalucía, el 10% en Castilla León, repartiéndose el resto por otras comunidades autónomas con desigual incidencia.

En cuanto a los autores de los incendios, la muestra la componen 217 autores, el 95% son hombres y el 5% mujeres. Hay 17 autores reincidentes (más de un incendio), de ellos hay 10 con 4 incendios o más, destacando un sujeto con 12, otro con 15 y otro con 19 incendios imputados.

Para el espacio del hecho, la variable provincia no se empleará en el análisis al no representar evidencias del incendio. En cuanto a las variables de autor que inicialmente son 43, eliminamos del análisis las 4 variables correspondientes a los motivos por no ser objetivas y haberse comprobado que los sujetos mentían. La variable "Diagnóstico principal" por tener un 84,6 % de respuestas en la categoría No se sabe. Y las variables "Distancia del incendio al domicilio" y "Distancia del incendio al trabajo" por ser redundantes con las de localización de incendio. Las variables tenidas en cuenta en el análisis de elección de variables son las 19 variables del hecho y las 36 del autor de la Tablas 1 y 2 del Apéndice I. En la Tabla 2 podemos ver un ejemplo de las variables empleadas.

Tabla 2. Ejemplo de la descripción de las variables.

Bloque	Variable	Categoría	Recuento	% de N
Hecho	Uso principal de la zona afectada	Aprovechamiento forestal	99	37,1%
		Aprovechamiento ganadero	30	11,2%
		Aprovechamiento agrícola	91	34,1%
		Aprovechamiento cinegético	12	4,5%
		Uso recreativo / turismo rural	9	3,4%
		Interfase forestal-urbana	21	7,9%
		No se sabe	5	1,9%
	Obedece a un patrón anterior	Sí	121	45,3%
		No	144	53,9%
		No se sabe	2	,7%
Autor	Sector laboral	Agrícola	53	19,9%
		Forestal	17	6,4%
		Pesca	18	6,7%
		Industria	17	6,4%
		Administración	5	1,9%
		Comercio-hostelería	4	1,5%
		Construcción	46	17,2%
		Otros servicios	42	15,7%
		Variados	17	6,4%
		No se sabe	48	18,0%

Supuestos

El empleo de máxima verosimilitud asume los supuestos de independencia de las variables en el modelo y de distribución multinomial para las variables categóricas. En el espacio del hecho se rechaza la hipótesis de independencia en el 80% de las comparaciones y en el 96% en el espacio del autor.

Identificación del modelo

Aplicando la fórmula $r_k = K \cdot \sum_{p=1}^{P}(C_p - 1)$ para el cálculo de los parámetros a estimar, observamos que en el espacio del hecho partimos inicialmente de 19 variables con un total de 81 categorías, dando lugar a la necesidad de estimar 62**K* parámetros, por lo que con una muestra de 267 casos estarían infraidentificados todos los modelos a partir del modelo de 5 conglomerados. En el espacio del autor la situación es aún peor, pues partimos inicialmente de 36 variables con un total de 158 categorías, dando lugar a la necesidad de estimar 122**K* parámetros, por lo que los modelos de más de tres conglomerados estarían infraidentificados.

Análisis de sensibilidad al orden de los casos en la conglomeración

La variable IC, en la que se almacenan los cálculos de concordancia en la conglomeración de las 190 comparaciones realizadas, alcanza una media del 71,04% de concordancia (DT=6,22) para las variables del hecho y una media del 77,80% (DT=5,74) para las de autor. Hemos estudiado su distribución mediante las pruebas de Kolmogorov-Smirnov (KS) y Shapiro-Wilk (SW). El resultado para el IC del hecho es KS= 0,068 (p=0,034) y SW=0,993 (p=0,446). Para el IC del autor obtuvimos un KS= 0,043 (p=0,200) y SW=0,991 (p=0,321). Si bien el estadístico de Kolmogorov rechaza la hipótesis de normalidad en el IC del hecho (y siendo este estadístico sensible al tamaño muestral), analizando el estadístico de Sapiro-Wilk y los gráficos de normalidad, tomamos la decisión de mantener la hipótesis de normalidad de las dos distribuciones, para un valor de α=0.05.

La normalidad de las dos variables IC evidencia la ausencia de sesgo en la medida del porcentaje de concordancia. La distribución de los valores permite apreciar la existencia de un efecto del orden de los casos sobre el resultado final de la conglomeración, efecto que es ligeramente menor en el caso de la agrupación de los autores.

El espacio del hecho

Elección de variables para la conglomeración del hecho.

En las Figuras 1 y 2 observamos el efecto del orden de los casos en la importancia de cada variable en la conglomeración sobre la solución de 5 conglomerados. El gráfico de cajas de la Figura 2, resume la variabilidad de la importancia de las variables calculada en las reordenaciones aleatorias. Se observa que las variables más estables frente al efecto del orden son las que menos importancia tienen en las distintas conglomeraciones y también las más impor-

tantes. Las variables con importancia intermedia son más sensibles al efecto del orden de los casos.

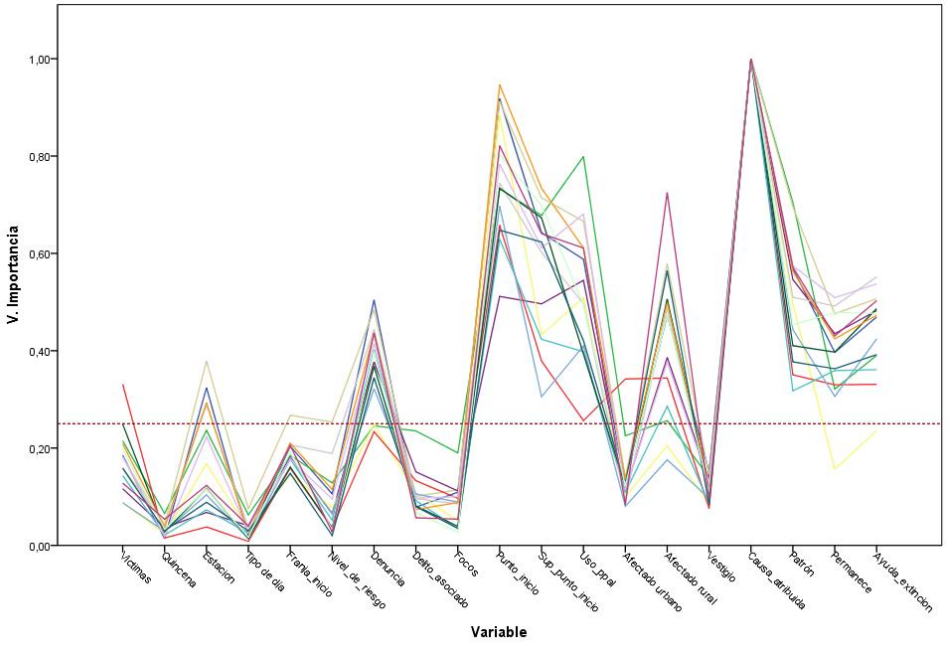


Figura 1 Efecto del orden de los casos en la importancia relativa de las variables (15 reordenaciones).

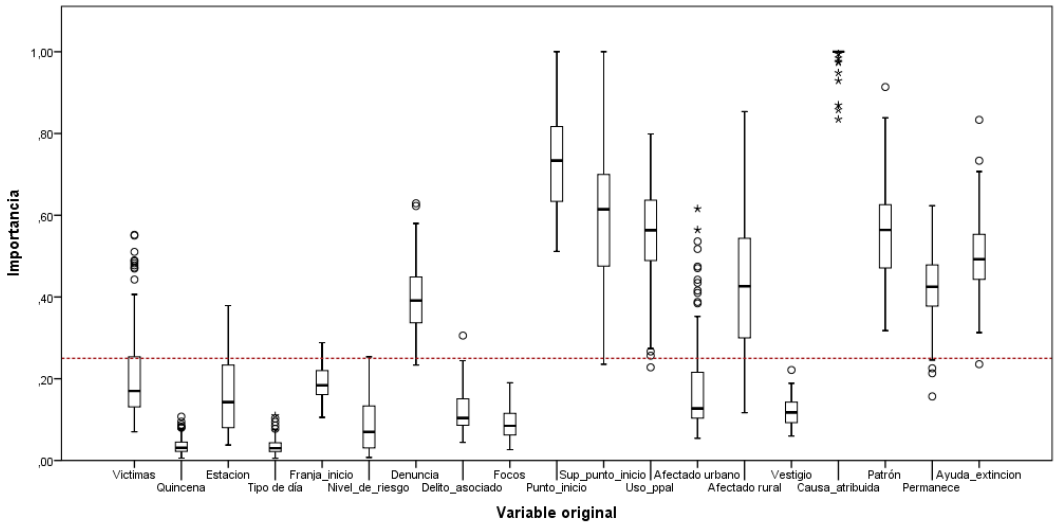


Figura 2. Efecto del orden de los casos en la importancia relativa de las variables (100 reordenaciones).

La decisión tomada por el panel de expertos fue la de establecer un punto de corte para la importancia relativa en el valor 0,25, lo que lleva a mantener las nueve variables de la Tabla 3 para el cálculo de la solución final del espacio del hecho. El índice de acuerdo (Kappa) de la solución final con la decisión de cada juez tuvo un acuerdo mínimo de 0,79 ($p<0,005$) y máximo de 1 ($p<0,005$).

Tabla 3. Variables definitivas del hecho.

Etiqueta	Nombre
Persona que denuncia	Denuncia
Punto de inicio R	R_Punto_inicio
Tipo de superficie cerca del punto de inicio	sup_punto_inicio
Uso principal de la zona afectada	uso_ppal
Afectado rural	afectado_rural
Causa atribuida	Causa_2010
Obedece a un patrón anterior	patrón
Permanece en el lugar del hecho	permanece
Ayuda en la extinción	ayuda_extincion

Número de conglomerados finales del hecho. Hemos calculado los resultados del BIC y de la Razón de Medidas de Distancia (RMD) en 100 repeticiones del análisis bietápico, desde 2 hasta 15 conglomerados, calculados tras la reordenación aleatoria de los casos con las 9 variables del hecho, (por claridad sólo se representan 15 repeticiones). Los valores del BIC decrecen hasta llegar a un mínimo que coincide con el número de conglomerados analíticamente óptimo y posteriormente vuelve a crecer según aumentamos el número de conglomerados. Aplicando el BIC como criterio de selección de modelos y según la gráfica de la Figura 3, el modelo inicial seleccionado sería el de cinco conglomerados.

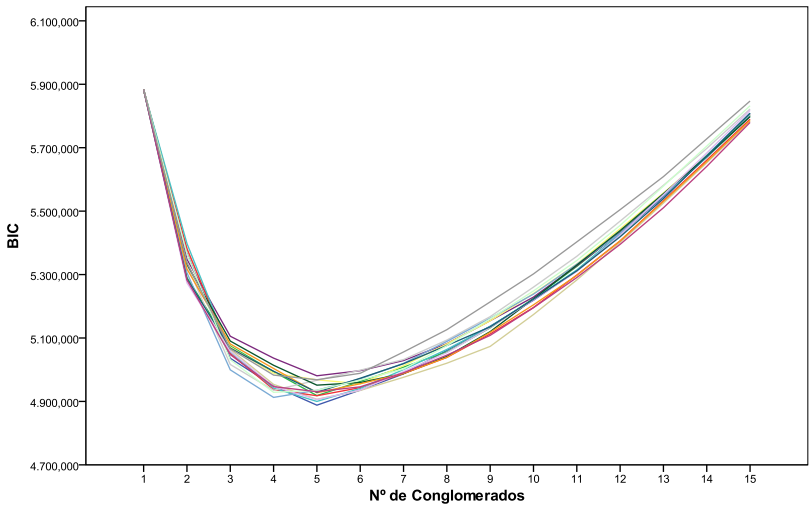


Figura 3. Efecto del orden de los casos en los valores del BIC.

La primera etapa del método automático de selección del número de conglomerados del SPSS no emplea el BIC como criterio de selección, sino la razón de cambio en BIC, R(k), calculada según la ecuación (4) (Figura 4). Seleccionando inicialmente el rango de soluciones desde la de 2 hasta la de *k* tal que R(*k*)>0,04. Aplicando este método en las 100 reordenaciones aleatorias, el rango de soluciones inicial para el espacio del hecho sería desde la de 2 hasta la de 4 conglomerados en 23 ocasiones y desde la de 2 hasta la de 5 conglomerados en 77 ocasiones.

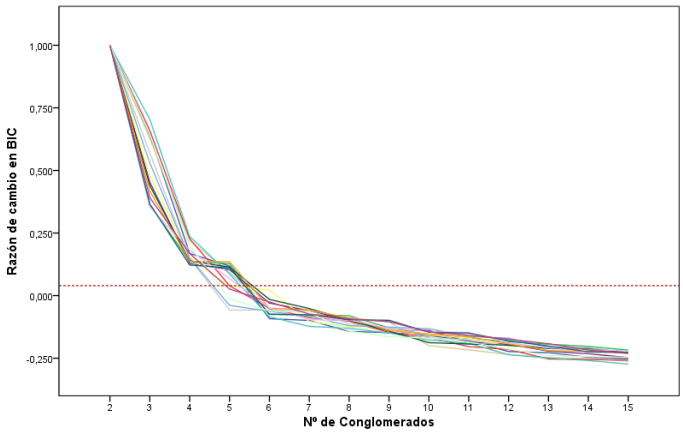


Figura 4. Efecto del orden de los casos en la razón de cambio en BIC, R(k).

En las Figuras 5 y 6 se aprecia la variabilidad producida por el orden de los casos en la razón de medidas de distancia. Al aplicar a las soluciones iniciales el criterio de la razón de medidas de distancia calculado según la ecuación (6), R=Rd(v)/Rd(t) > 1,15, el SPSS selecciona automáticamente la solución de 2 conglomerados en 28 ocasiones, la de 3 en 37, la de 4 en 9 y la solución de 5 en 26 ocasiones.

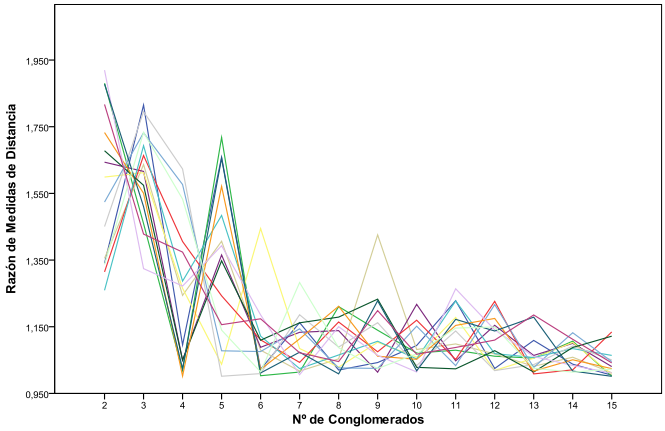


Figura 5. Efecto del orden de los casos en la razón de medidas de distancia.

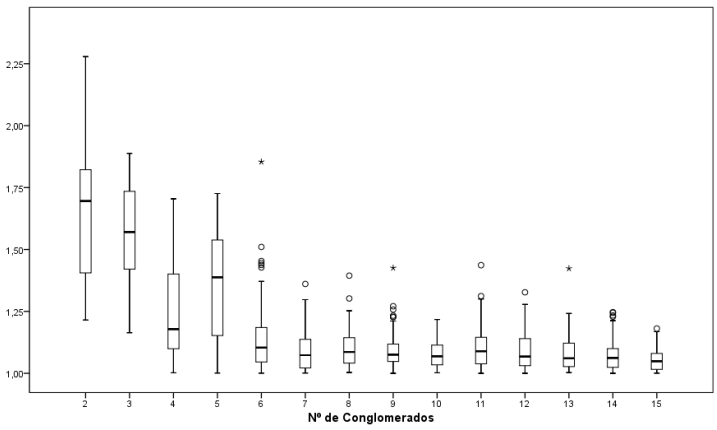


Figura 6. Efecto del orden de los casos en la variabilidad de la RMD

En la Figura 7 hemos representado la media de la razón de medidas de distancia en las 100 reordenaciones efectuadas, donde se aprecia que los mayores saltos en distancia ocurren en las soluciones de 2, 3 y 5 conglomerados, seguida de la de 4, la de 6 y la de 7, y a partir de la de 7 no se aprecian diferencias significativas en la razón de medidas de distancia hasta la de 13.

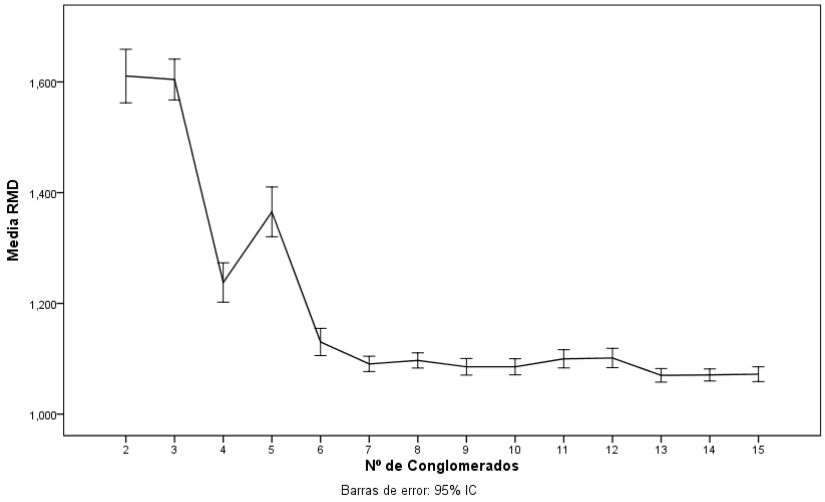


Figura 7. Media de la Razón de Medidas de Distancia

El resultado de la conglomeración jerárquica realizada con las variables dicotomizadas, con el método de Ward y la distancia euclídea al cuadrado lo observamos en la Figura 8. Las soluciones recomendadas serían la de 2, 3, 5, 7, 9 y 13 conglomerados dependiendo del grado de discriminación requerido.

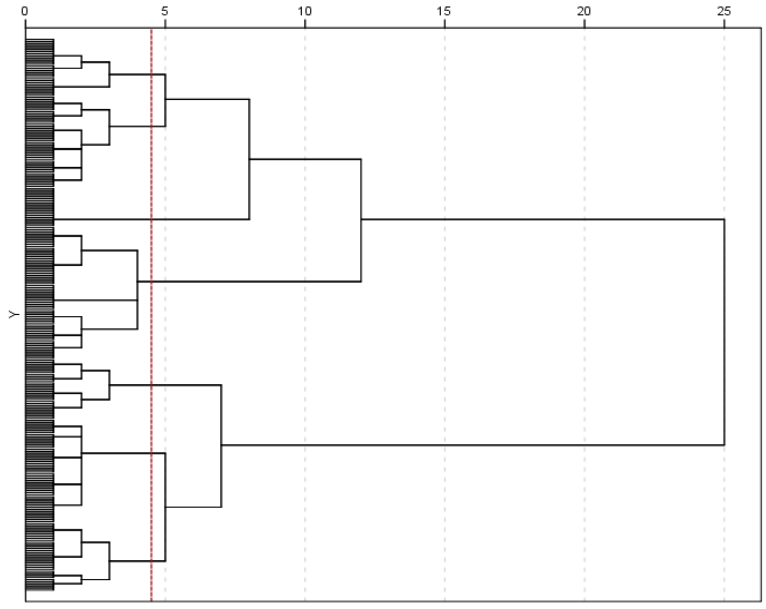


Figura 8. Dendrograma. La línea roja marcaria la solución de 7 conglomerados.

Debemos determinar el número de conglomerados que van a conformar el espacio del hecho. Si bien las soluciones de 2, 3 y 5 conglomerados son las que más veces selecciona en automático el procedimiento bietápico, ninguna de las soluciones presenta evidencias claras que nos permitan decantarnos por una u otra. Si nos decantamos por la solución que más se repite, la de tres conglomerados, siendo la más estable no tendría mucha capacidad descriptiva. Un salto en distancia en la solución de *k* conglomerados representa pérdida de homogeneidad y aumento de la entropía al unir dos conglomerados y pasar de *k+1* a *k*, si observamos la Figura 7 vemos que al pasar de una solución a otra a partir de la de 7 conglomerados hacia las de 8, 9, etc. tampoco se aprecia ningún salto en la distancia entre las soluciones que represente un incremento evidente en la entropía.

Finalmente hemos analizado la relevancia teórica de varias soluciones buscando la que nos permita una interpretación sencilla de los centroides y que tenga suficiente capacidad descriptiva en cada uno de los dos espacios, seleccionando la solución de 7 conglomerados para el espacio del hecho.

Agrupación por variables del hecho. El Espacio del hecho. Se han empleado las 9 variables de la selección definitiva reflejadas en la Tabla 3. En la Figura 9 observamos los tamaños de la solución de 7 conglomerados que da lugar a agrupaciones de un tamaño relativo entre el 7,1% (19 casos) y 24,7% (66 casos). La razón de tamaños entre el conglomerado mayor y el menor es de 1:3,47.

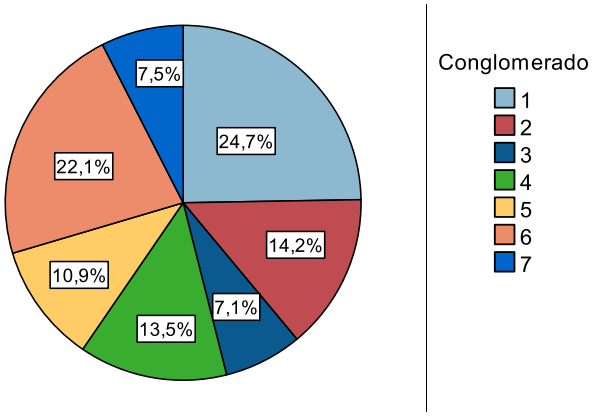


Figura 9. Solución de 7 conglomerados del espacio del hecho.

Valorando la participación de cada una de las variables en la solución obtenida, se observa que las variables que más contribuyen a la solución son causa atribuida, punto de inicio

y uso principal de la zona afectada. La importancia relativa de los predictores utilizados en el modelo, escalados de 0 (menor importancia) a 1 (mayor importancia) es la siguiente:

- La variable con mayor capacidad discriminativa es características del punto de inicio, que recibe la máxima puntuación (1,0).
- Le siguen la causa atribuida (0,99) y el uso principal de la zona afectada (0,86).
- Con una importancia escalada entre 0,60 y 0,80 sólo se encuentra el tipo de superficie cerca del punto de inicio (0,69).
- Con una importancia escalada entre 0,40 y 0,60, si obedece a un patrón anterior (0,59), afectado rural (0,50), si permanece en el lugar del hecho (0,46), si ayuda en las labores de extinción (0,45) y la persona que denuncia (0,41).

Para la descripción de los conglomerados se han empleado las variables utilizadas en la conglomeración. A continuación se resumen las características de cada conglomerado, y se les atribuye un nombre descriptivo. Los conglomerados se encuentran ordenados por su tamaño y no por su número de creación.

1. El conglomerado número 1 cuenta con 66 hechos (24,7%), está constituido mayoritariamente por incendios imprudentes (98%). Principalmente se inician en cultivos (27%), caminos y sendas (18%) o cualquier otro tipo de superficie, próximas a matorrales (54%), afectan por tanto a zonas forestales (50%), sin cultivar (20%) o agrícolas (18%) y son zonas de aprovechamiento agrícola (58%) o interfase forestal-urbana (15%). Habitualmente denuncia el propio autor (41%) o los vecinos (29%). El autor permanece en la zona (100%) y ayuda en las labores de extinción (100%). El incendio no obedece a un patrón anterior (100%). La etiqueta propuesta para este conglomerado es IMPRUDENTE AGRÍCOLA PRESENTE.

2. El conglomerado número 6 cuenta con 59 hechos (22,1%), está constituido mayoritariamente por incendios sin sentido (88%), principalmente se iniciaron en pistas forestales (61%), o carreteras y viales (32%) y se propagaron por masas forestales (97%), la zona afectada fue forestal (98%), de aprovechamiento forestal (100%). Habitualmente denuncian los vecinos (44%), agentes de la autoridad (27%) o los medios de vigilancia y extinción (19%). El autor permanece en la zona (54%) o no (46%) pero no ayuda en las labores de extinción (71%) y el incendio obedece a un patrón anterior (95%). La etiqueta propuesta para este conglomerado es SIN SENTIDO FORESTAL.

3. El conglomerado número 2 cuenta con 38 hechos (14,2%), está constituido mayoritariamente por imprudencias (68%) o instrumentales (26%), principalmente se iniciaron en cultivos (37%), sendas o caminos (29%) y se propagaron por matorrales (47%), por terreno agrícola (21%) o dehesa (18%), la zona afectada fue principalmente agrícola (39%) o forestal (24%), de aprovechamiento agrícola (63%) o ganadero (21%). Habitualmente denuncian los medios de vigilancia y extinción (45%) o los vecinos (29%). El autor no permanece en la zona (68%) ni ayuda en las labores de extinción (79%) y puede o no obedecer a un patrón anterior (50%). La etiqueta propuesta para este conglomerado es IMPRUDENTE AGRICOLA QUE ESCAPA.

4. El conglomerado número 4 engloba a 36 hechos (13,5%), está constituido principalmente por imprudencias (69%) o instrumentales (30%), se inician el interior de masas vegetales alejadas (50%) y se propagaron por masa forestal (83%), la zona afectada fue forestal (100%) de aprovechamiento forestal (56%) o ganadero (19%). Suelen denunciarlo los vecinos (44%), el propio autor (17%) o testigos (14%) y no obedecer a un patrón (92%). La etiqueta propuesta para este conglomerado es IMPRUDENTE RECREATIVO FORESTAL.

5. El conglomerado número 5 engloba a 29 hechos (10,9%), está constituido principalmente por incendios sin sentido (59%) o instrumentales (38%), se inician en carreteras o viales (41%) o caminos y sendas (31%) o pistas forestales (17%) y se propagaron por masas forestales (52%), dehesa (24%) o matorral (14%), la zona afectada fue principalmente masa forestal (93%) de aprovechamiento forestal (34%) o cinegético (24%) o interfase forestal-urbana (17%). Suelen denunciarlo los vecinos (52%) o testigos (24%) y pueden o no obedecer a un patrón (48%), el autor permanece en la escena (93%) pero no ayuda en la extinción (65%). La etiqueta propuesta es SIN SENTIDO O CINEGETICO.

6. El conglomerado número 7 engloba a 20 hechos (7,5%), está constituido principalmente por incendios sin sentido (100%), se inician en pistas forestales (100%) y se propagaron por matorral (100%), la zona afectada fue principalmente masa forestal (100%) de aprovechamiento agrícola (95%). Suelen denunciarlo los medios de vigilancia y extinción (95%) y obedecen a un patrón (100%), el autor no permanece en la escena (100%) ni ayuda en la extinción (95%). La etiqueta propuesta para este conglomerado es SIN SENTIDO AGRICOLA.

7. El conglomerado número 3 engloba a 19 hechos (7,1%), está constituido principalmente por incendios instrumentales (63%) o sin sentido (37%), se inician en carreteras o viales (74%) y se propagaron por matorral (47%), dehesa (21%) o pastizal (21%), la zona afectada fue sin cultivar (63%) o agrícola (21%), de aprovechamiento ganadero (42%) o agrícola (37%). Suelen denunciarlo los vecinos (63%) o testigos (26%) y pueden o no obedecer a un patrón (53%), el autor no permanece en la escena (100%) ni ayuda en la extinción (100%). La etiqueta propuesta para este conglomerado es INSTRUMENTAL GANADERO.

El Espacio del Autor

Elección de variables para la conglomeración del autor.

En las Figuras 10 y 11 observamos el efecto del orden de los casos sobre la importancia de cada variable en la conglomeración para la solución de 6 conglomerados. Se observa de nuevo que las variables más estables frente a la reordenación son las que están en los extremos de la importancia. En el espacio del autor se aprecia una menor variabilidad en la importancia de las variables que en el espacio del hecho.

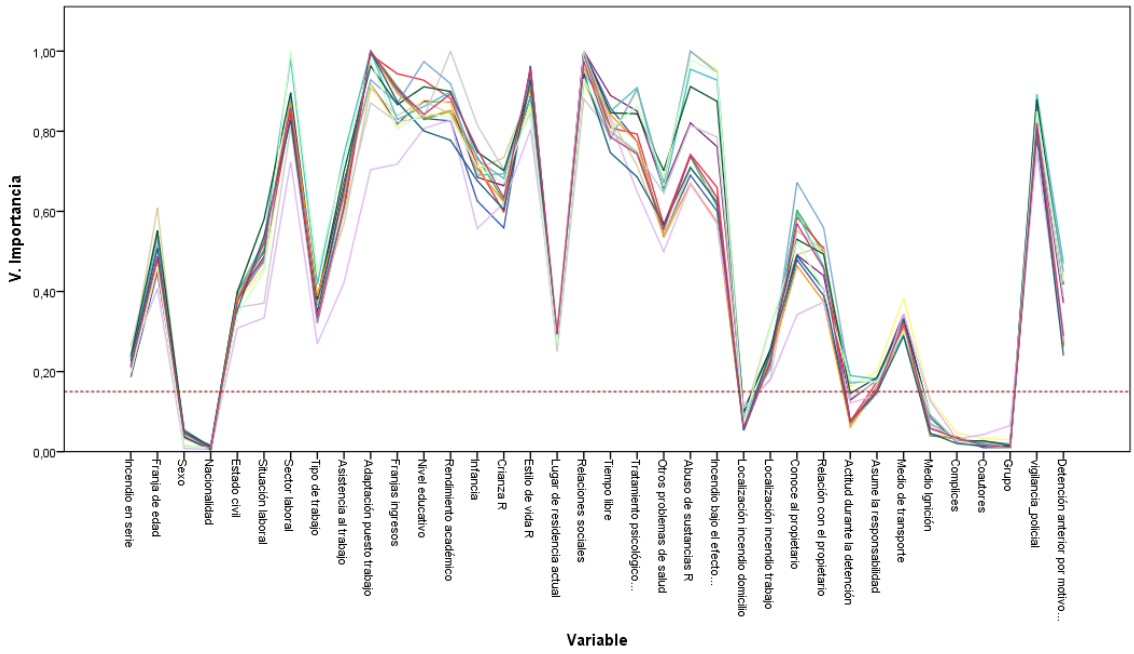


Figura 10. Efecto del orden de los casos en la importancia relativa de las variables.

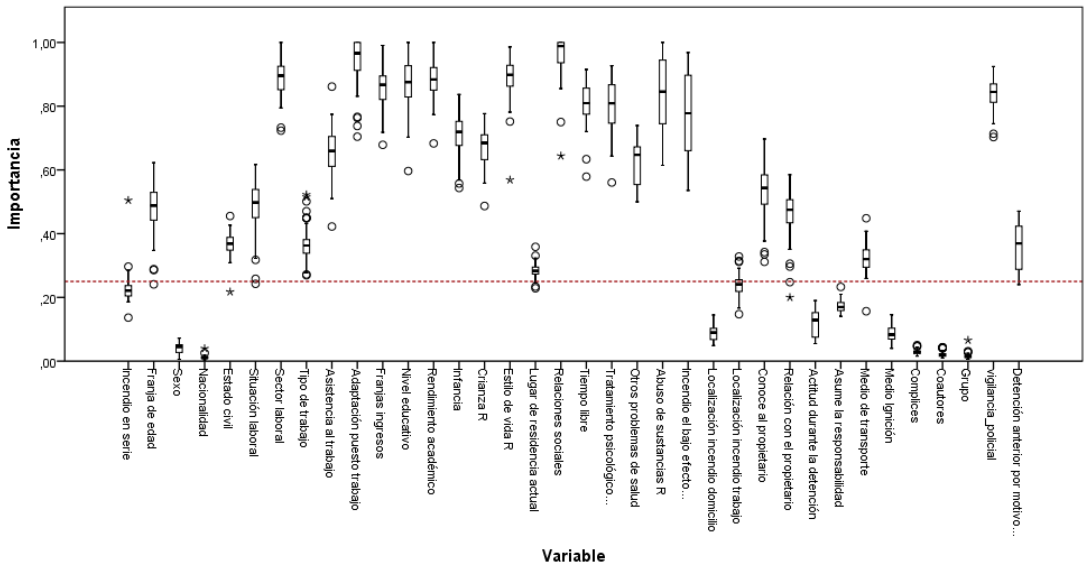


Figura 11. Efecto del orden de los casos en la importancia relativa las variables.

La decisión tomada por el panel de expertos fue la de establecer un punto de corte para la importancia relativa en el valor 0,30 y mantener las 25 variables de la Tabla 4 para el cálculo de la solución final del espacio del autor. El índice de acuerdo (Kappa) mínimo fue de 0,66 ($p<0,005$) y máximo de 1 ($p<0,005$).

Tabla 4. Variables definitivas del autor.

Etiqueta	Nombre
Franjas de edad	R_Franja_edad
Estado civil	e_civil
Situación laboral	situacion_laboral
Sector laboral	sector_laboral
Tipo de trabajo	tipo_trabajo
Asistencia al trabajo	asistencia_trabajo
Adaptación al puesto de trabajo	adaptación_trabajo
Franjas ingresos	ingresos
Nivel educativo	educacion
Rendimiento académico	rendimiento
Infancia	infancia
Crianza R	R_Crianza
Estilo de vida R	R_Estilo_vida
Relaciones sociales	relaciones_sociales
Tiempo libre	tiempo_libre
Tratamiento psicológico/psiquiátrico	psic_tratam
Otros problemas de salud	salud_otro
Abuso de sustancias R	R_abuso_sust
Incendio bajo el efecto de sustancias	sustancias_hecho
Conoce al propietario	conoce_prop
Relación con el propietario	rela_prop
Medio de transporte	R_Transporte
Vigilancia policial	vigilancia_policial
Detención anterior por motivo distinto al incendio	antecedentes
Incendio en serie	incendio_serie

Número de conglomerados finales del autor. De nuevo hemos calculado los resultados del BIC y de la RMD en 100 repeticiones del análisis bietápico desde 2 hasta 15 conglomerados, con reordenación de los casos y con las 25 variables del autor. Aplicando el BIC como criterio de selección de modelos y según la grafica de la Figura 12, el modelo inicial seleccionado sería el de 5 conglomerados. Aplicando la razón de cambio en BIC, R(*k*), como criterio de selección inicial en 100 reordenaciones aleatorias (hemos representado 15 en la Figura 13), el rango inicial de soluciones para el espacio del autor según el SPSS sería las de 2 a 5 conglomerados en 99 ocasiones y de 2 a 6 en 1 ocasión.

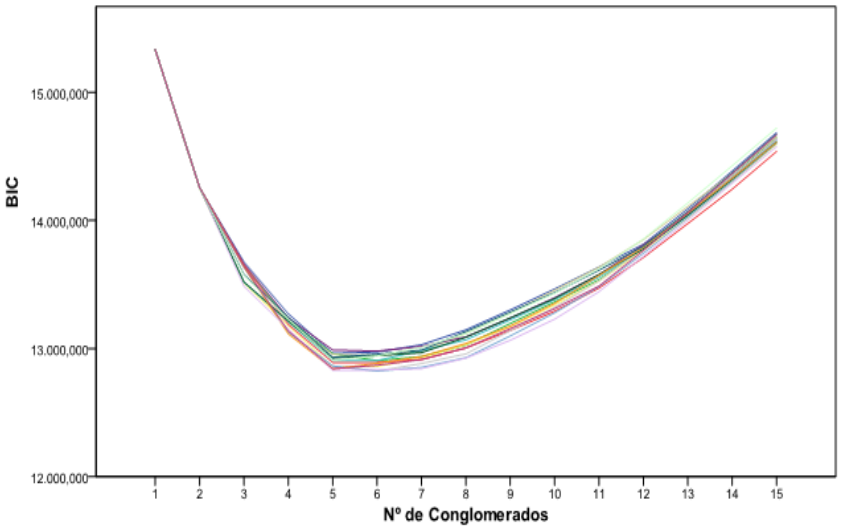


Figura 12. Efecto del orden de los casos en los valores del BIC de 1 a 15 Conglomerados.

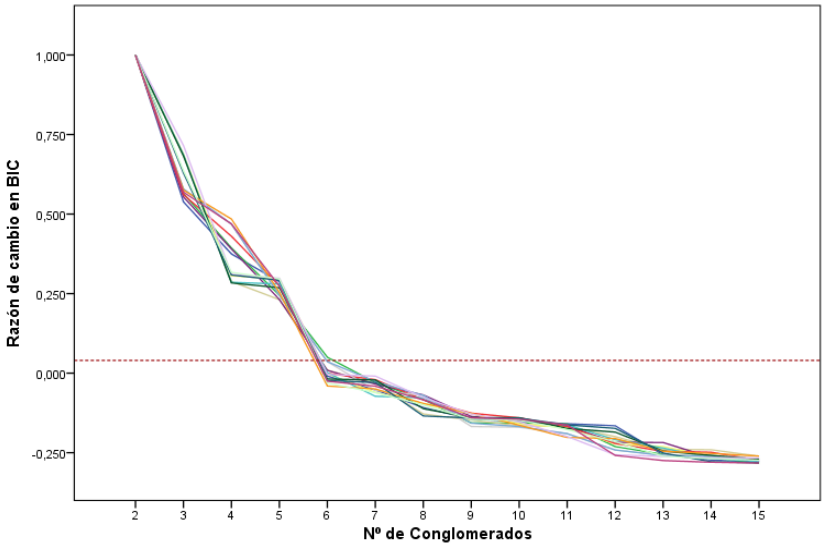


Figura 13. Efecto del orden de los casos en la razón de cambio en BIC, R(k).

En las Figuras 14 y 15 observamos el efecto del orden de los casos en 15 ordenaciones aleatorias de las 100 realizadas. Al aplicar al rango de soluciones iniciales el criterio de la razón de medidas de distancia calculado según la ecuación (6), R=Rd(v)/Rd(t) > 1,15, el SPSS selecciona automáticamente la solución de 5 conglomerados en las 100 ocasiones.

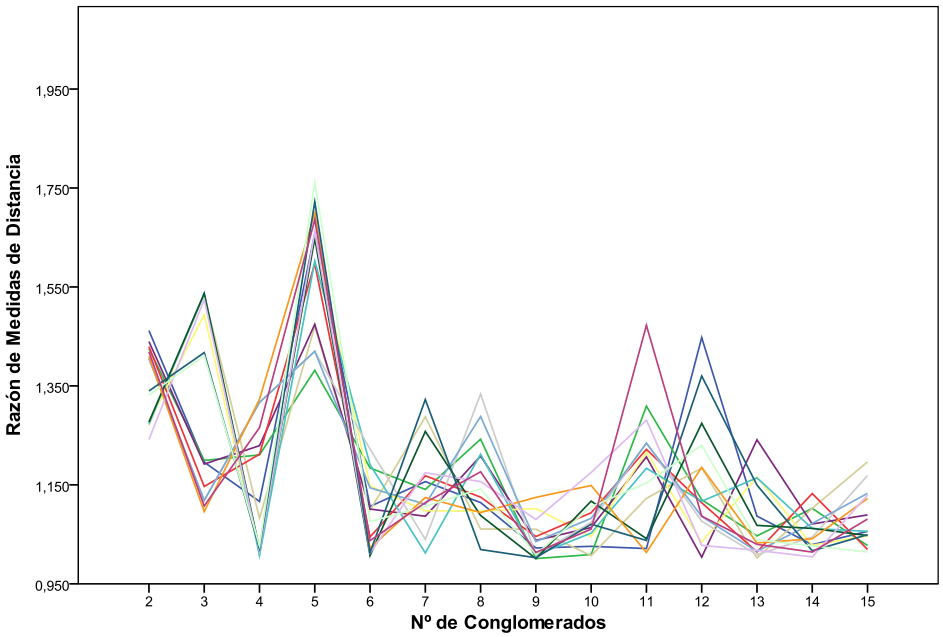


Figura 14. Efecto del orden de los casos en la Razón de Medidas de Distancia.

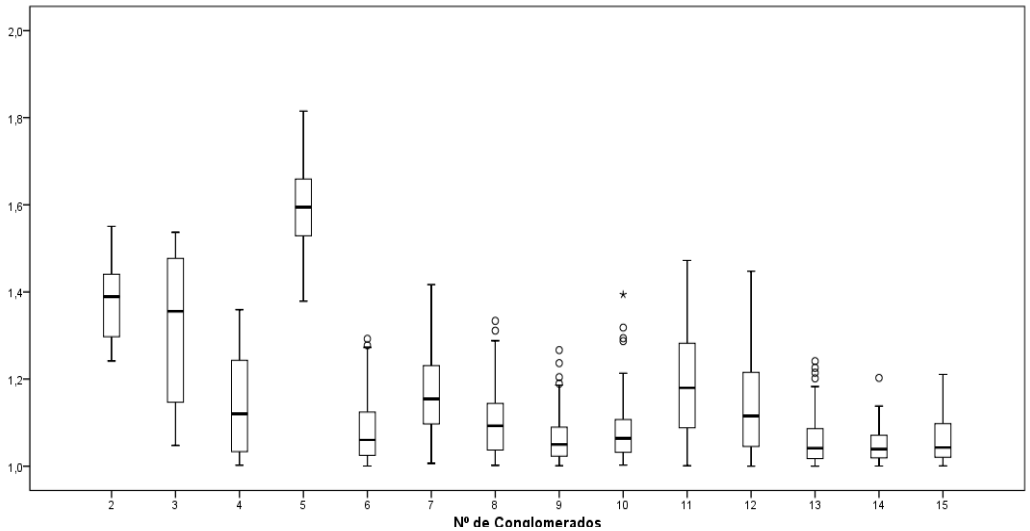


Figura 15. Efecto del orden de los casos en la variabilidad de la RMD

En la Figura 16 hemos representado la media de la razón de medidas de distancia en las 100 reordenaciones efectuadas, donde se aprecia el mayor salto en distancia en la solución

de 5 conglomerados seguida de las soluciones de 2 y 3, también se aprecian saltos significativos en la solución de 7 y 11 conglomerados.

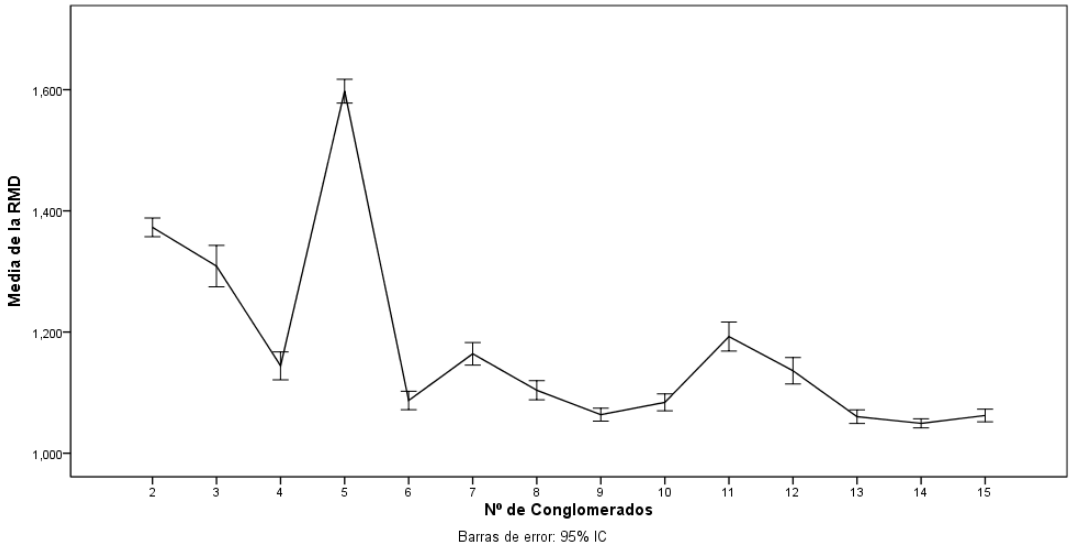


Figura 16. Media de la Razón de Medidas de Distancia

El dendrograma de clasificación de la conglomeración jerárquica realizada con las variables dicotómicas con el método de Ward y la distancia euclídea al cuadrado representado en la Figura 17, muestra que las soluciones más evidentes serían la de 2, 3, 4, 6, 7, 8 y 11 conglomerados.

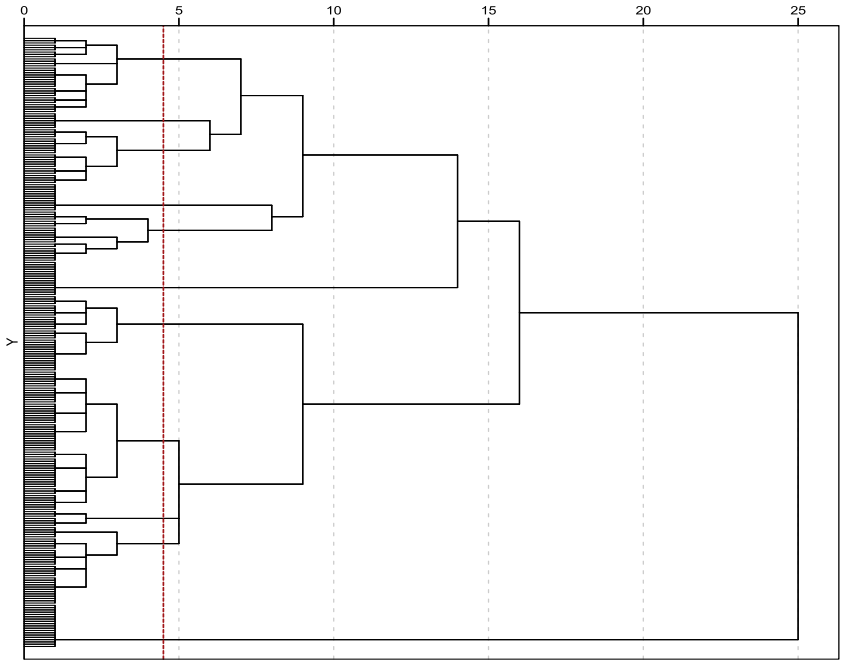


Figura 17. Dendrograma. La línea roja marcaria la solución de 11 conglomerados

La solución de 5 conglomerados ofrecida por el procedimiento bietápico parece bastante estable, pero tras el análisis de sus centroides observamos que uno de los conglomerados está compuestos por 19 casos que son la misma persona, es decir uno de los incendiarios reincidentes en la muestra, y otro de los conglomerados está formado por 15 casos que son otro de los incendiarios reincidentes, por lo que la clasificación del espacio de autor se vería reducida a tres conglomerados únicamente. Como ocurre en el espacio del hecho no tendrían suficiente capacidad descriptiva. Analizando la Figura 16 se observan saltos en distancia en las soluciones de 7 y 11 conglomerados que representan los mayores incrementos en la entropía para soluciones previas a la de 5 conglomerados. Finalmente hemos analizado la relevancia teórica de estas soluciones buscando la que nos permita una interpretación sencilla de los centroides y que tenga suficiente capacidad descriptiva en cada uno de los dos espacios, seleccionando la solución de 11 conglomerados para el espacio del autor.

Agrupación por variables del autor. El Espacio del autor. Se han utilizado las 25 variables de clasificación de la Tabla 4. En la Figura 18 observamos los tamaños relativos de los conglomerados en la solución elegida. El conglomerado más numeroso recoge el 20,2% de los sujetos (54 casos), frente al menos numeroso que agrupa el 3% de los incendios (8 casos). El tamaño relativo del mayor respecto al menor es de 1:6,75.

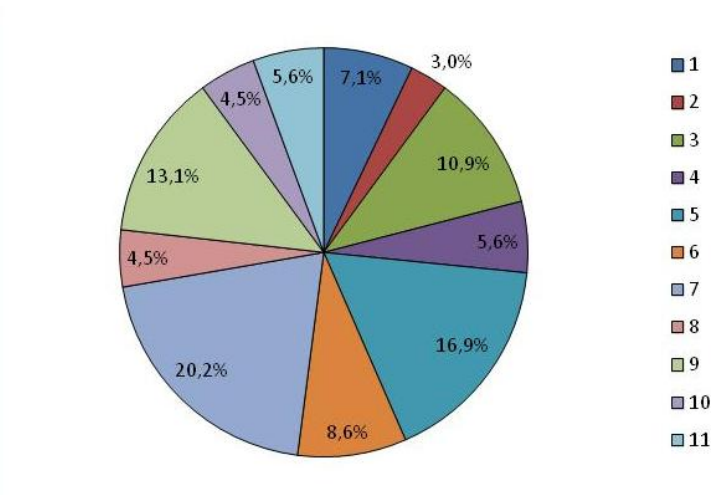


Figura 18. Solución de 11 conglomerados del espacio de autor.

En el caso de la agrupación de autores son 9 las variables que muestran una importancia relativa como predictor mayor que 0,60:

- Las variables con mayor capacidad descriptiva son la adaptación al puesto de trabajo que recibe la máxima puntuación (1,0), seguida de las variables correspondientes a la asistencia al trabajo (0,87), el sector laboral (0,82) y a las relaciones sociales (0,80).
- Con una importancia escalada entre 0,60 y 0,80 están el rendimiento académico (0,70), el incendio bajo el efecto de sustancias (0,68), el tiempo libre (0,67), el estilo de vida (0,65), la vigilancia policial (0,62) y la franja de ingresos (0,62).
- Con importancias relativas entre 0,40 y 0,60 se encuentran las variables tratamiento psicológico (0,58), el nivel educativo (0,56), situación laboral (0,53), franja de edad (0,51), el abuso de sustancias (0,48), infancia (0,46), el medio de transporte (0,45) y otros problemas de salud (0,43).
- Con importancias relativas entre 0,20 y 0,40 se encuentran las variables tipo de trabajo (0,39), la crianza (0,39), conoce al propietario (0,39), la relación con el propietario (0,28), detención anterior (0,25) e incendio en serie (0,21).
- Con importancia relativa por debajo de 0,2 únicamente se encuentra la variable estado civil (0,19).

A continuación se describen los conglomerados obtenidos. La ordenación de los conglomerados obedece a su tamaño y no a su número de orden en la solución.

1. El conglomerado número 7 está constituido por 54 sujetos (20,2%). Su perfil corresponde mayoritariamente a un sujeto que está normalmente adaptado al trabajo (98%) y nunca falta (87%), habitualmente trabaja en el sector agrícola (26%), en la industria (20%) o en la construcción (18%), en un trabajo de tipo manual (74%), está empleado (54%) o es autónomo (20%) y tiene un sueldo de entre 600 y 1200€ al mes (76%), tiene estudios primarios (48%) o elementales (44%) y aprobaba sin dificultad (41%) o con dificultad (37%). Tiene muchos amigos (85%) y en su tiempo libre prefiere estar con gente (93%), vive con su pareja (46%) o con sus padres (41%). No está bajo tratamiento psiquiátrico (93%), ni abusa de sustancias (93%), no cometió el incendio bajo el efecto de alguna sustancia (93%), ni tiene otros problemas de salud (96%), tuvo una infancia normal (94%), una crianza normal (98%), no ha sido detenido anteriormente por otros motivos (87%) y no se encontraba bajo vigilancia policial (74%). Está casado (57%) o soltero (35%) y no tenía relación con el propietario (54%) o eran vecinos (13%) y no era un incendiario en serie (87%). La etiqueta propuesta para este conglomerado es **JOVEN TRABAJADOR.**

2. El conglomerado número 5 está constituido por 45 sujetos (16,9%). Su perfil corresponden mayoritariamente a un sujeto que está normalmente adaptado al trabajo (53%) o no se sabe (44%) y nunca falta (42%) o no se sabe (42%), habitualmente trabaja en el sector agrícola (36%) o no se sabe (27%), en un trabajo de tipo manual (71%), es pensionista o jubilado (71%) y cobra menos de 600€ (47%) o entre 600 y 1200€ al mes (42%). Tiene estudios elementales (64%) o es analfabeto (29%) y no estuvo escolarizado (36%). Tiene pocos amigos (49%) o muchos (44%) y en su tiempo libre prefiere estar con gente (58%) o solo (42%), vive con su pareja (44%) o solo (31%). No está bajo tratamiento psiquiátrico (91%), ni abusa de sustancias (93%), ni cometió el incendio bajo el efecto de alguna sustancia (98%), ni tiene otros problemas de salud (78%), tuvo una infancia normal (100%), una crianza normal (84%), no ha sido detenido anteriormente por otros motivos (91%) y no se encontraba bajo vigilancia policial (84%). Está soltero (51%), separado o viudo (29%) y tiene más de 60 años (58%) o de entre 46 y 60 años (27%). El mismo es el propietario (47%) o lo conocía mucho (31%), siendo en este caso un vecino (29%), se desplazó a pie (71%) y no era un incendiario en serie (89%). La etiqueta propuesta para este conglomerado es **JUBILADO PROPIETARIO.**

3. El conglomerado número 9 está constituido por 35 sujetos (13,1%). Su perfil corresponde mayoritariamente a un sujeto que está normalmente adaptado al trabajo (91%) y nunca falta (94%), habitualmente trabaja en el sector agrícola (31%) o en otros servicios (31%), en un trabajo de tipo manual (51%) o cualificado (37%), es pensionista o jubilado (43%) o autónomo (29%) y cobra entre 600 y 1200€ al mes (43%) o más de 1200€ (28,6%), tiene estudios primarios (48%) o elementales (40%) y aprobaba sin dificultad (49%) o sacaba buenas notas (17%). Tiene muchos amigos (89%) y en su tiempo libre prefiere estar con gente (97%), vive con su pareja (94%). No está bajo tratamiento psiquiátrico (91%), ni abusa de sustancias (80%), ni cometió el incendio bajo el efecto de alguna sustancia (91%), ni tiene otros problemas de salud (86%), tuvo una infancia normal (91%), una crianza normal (89%), no ha sido detenido anteriormente por otros motivos (100%) y no se encontraba bajo vigilancia policial (97%). Está soltero (77%), tiene más de 60 años (57%) o de entre 34 y 46 años (26%) y el mismo es el propietario (49%) o lo conocía mucho (49%), siendo entonces familiar (14%) o amigo (11%). Se desplazo en turismo (69%) o a pie (20%) y no era un incendiario en serie (100%).La etiqueta propuesta para este conglomerado es **SOLTERO PROPIETARIO**.

4. El conglomerado número 3 está constituido por 29 sujetos (10,9%). Su perfil corresponde mayoritariamente a un sujeto del que no se sabe su adaptación al trabajo (52%) o esta es normal (24%) y no se sabe si falta al trabajo (59%) o nunca falta (28%), no se sabe en qué sector trabaja (52%), y no se sabe el tipo de trabajo (59%), es pensionista o jubilado (41%) o desempleado (34%) y cobra menos de 600€ al mes (62%) o no tiene ingresos (21%), tiene estudios elementales (59%) o es analfabeto (38%) y aprobaba con dificultad (31%) o suspendía (17%). No tiene amigos (62%) o tiene pocos (31%) y en su tiempo libre prefiere estar solo (86%), vive solo (52%) o con sus padres (28%). Está bajo tratamiento psiquiátrico (59%) o no (41%), abusa de sustancias (69%), pero no cometió el incendio bajo el efecto de alguna sustancia (52%) aunque pudo hacerlo (45%), no tiene otros problemas de salud (59%) o si (38%), tuvo una infancia normal (45%) aunque puede tener historia de problemas en la familia (38%) y una crianza normal (48%) o difícil (38%), si ha sido detenido anteriormente por otros motivos (52%) o no (49%) y no se encontraba bajo vigilancia policial (72%). Está casado (79%), es menor de 60 años (96%), no conoce de nada al propietario (55%) o lo conocía mucho (34%), siendo entonces vecinos (21%), se desplazó a pie (90%) y no era un incendiario en serie (52%) aunque podía serlo (48%). La etiqueta propuesta para este conglomerado es **ASOCIAL CONSUMIDOR SIN INGRESOS.**

5. El conglomerado número 6 está constituido por 23 sujetos (8,6%).Su perfil corresponde mayoritariamente a un sujeto que está normalmente adaptado al trabajo (91%) y falta poco al trabajo (65%) o nunca falta (26%), tiene trabajos variados (30%) o trabaja en la construcción (26%), pero son trabajos de tipo manual (100%), está empleado principalmente (61%) o desempleado (30%) y cobra entre 600 y 1200€ al mes (87%), tiene estudios elementales (74%) y aprobaba con dificultad (83%). Tiene muchos amigos (56%) o tiene pocos (39%) y en su tiempo libre prefiere estar con gente (83%), vive con su pareja (43%) o con sus padres (43%). Está bajo tratamiento psiquiátrico (61%) o no (39%), abusa de sustancias (100%), y cometió el incendio bajo el efecto de alguna sustancia (91%), no tiene otros problemas de salud (61%) o si (39%), tuvo una infancia normal (91%) y una crianza normal (78%), si ha sido detenido anteriormente por otros motivos (87%) y no se encontraba bajo vigilancia policial (78%). Está casado (48%) o soltero (43%) es menor de 34 años (65%), conoce poco o nada al propietario (91%), y no tenía relación con el (56%) o eran vecinos (39%), se desplazo a pie (52%) u otros medios (30%) y no

era un incendiario en serie (56%) pero puede serlo (39%). La etiqueta propuesta para este conglomerado es **JOVEN CONSUMIDOR CON ANTECEDENTES.**

6. El conglomerado número 1 está constituido por 19 sujetos (7,1%). Su perfil corresponde a un único sujeto del que se desconoce su adaptación al puesto de trabajo (100%) y no se sabe si falta al trabajo (100%), trabaja en la construcción (100%), y es un trabajo de tipo manual (100%), está desempleado (100%) y cobra entre 600 y 1200€ al mes (100%), es analfabeto (100%) y no estuvo escolarizado (100%). No se sabe si tiene amigos (100%) ni que hace en su tiempo libre (100%), vive con otros (100%). No se sabe si está bajo tratamiento psiquiátrico (100%), abusa de sustancias (100%), pero no se sabe si cometió el incendio bajo el efecto de alguna sustancia (100%), no se sabe si tiene otros problemas de salud (100%). No se sabe cómo fueron su infancia ni su crianza (100%), no ha sido detenido anteriormente por otros motivos (100%) y estaba siendo investigado como presunto autor (100%). Está casado (100%), tiene entre 34 y 46 años (100%), conoce mucho al propietario (100%) siendo este un vecino (100%), se desplazo a pie (100%) y era incendiario en serie (100%). La etiqueta propuesta para este conglomerado es **REINCIDENTE MÚLTIPLE 1**.

7. El conglomerado número 4 está constituido por 15 sujetos (5,6%). Su perfil corresponden a un único sujeto con una mala adaptación al puesto de trabajo (100%) pero que falta poco al trabajo (100%), trabaja en la pesca (100%), y es un trabajo de tipo cualificado (100%), está desempleado (100%) y no tiene ingresos (100%), tiene estudios elementales (100%) y aprobaba sin dificultad (100%). Tiene pocos amigos (100%) y en su tiempo libre prefiere estar solo (100%), vive con su pareja (100%). No está bajo tratamiento psiquiátrico (100%), no abusa de sustancias (100%) ni cometió el incendio bajo el efecto de alguna sustancia (100%), no tiene otros problemas de salud (100%). Tuvo una infancia y crianza normales (100%), si ha sido detenido anteriormente por otros motivos (100%) y estaba sometido a vigilancia policial (100%). Está soltero (100%), tiene entre 46 y 60 años (100%), conoce mucho al propietario (100%) siendo este un vecino (100%), se desplazo en un turismo (100%) y no era incendiario en serie (100%). La etiqueta propuesta para este conglomerado es **REINCIDENTE MÚLTIPLE 2**.

8. El conglomerado número 11 está constituido por 15 sujetos (5,6%). Su perfil corresponde a un sujeto del que se desconoce su adaptación al puesto de trabajo (100%) y no se sabe si falta al trabajo (100%), ni en qué sector trabaja (100%), ni de qué tipo es su trabajo

(100%), está desempleado (60%) o no se sabe (20%) y cobra entre 600 y 1200€ al mes (40%) o no tiene ingresos (40%), tiene estudios primarios (53%) o elementales (33%) y aprobaba sin dificultad (40%) o suspendía (33%). Tiene muchos amigos (93%) y en su tiempo libre prefiere estar con gente (87%), vive con su pareja (67%) o con sus padres (27%). No está bajo tratamiento psiquiátrico (100%), no abusa de sustancias (100%), ni cometió el incendio bajo el efecto de alguna sustancia (100%), no tiene otros problemas de salud (93%). Tuvo una infancia y crianza normales (100%), no ha sido detenido anteriormente por otros motivos (87%) y no estaba controlado ni vigilado por la policía (67%) o estaba siendo investigado como presunto autor (27%). Está casado (67%) o soltero (33%), es menor de 34 años (53%) o de entre 46 y 60 años (27%), es el mismo propietario (67%) o un familiar (13%) y se desplazó en un turismo (80%) y no era incendiario en serie (67%). La etiqueta propuesta para este conglomerado es **PARADO SOCIABLE PROPIETARIO.**

9. El conglomerado número 8 está constituido por 12 sujetos (4,5%). Su perfil corresponde a un único sujeto con una adaptación normal al puesto de trabajo (100%) y que nunca falta al trabajo (100%), trabaja en otros servicios (100%), y es un trabajo de tipo manual (100%), está empleado (100%) y cobra entre 600 y 1200€ al mes (100%), tiene estudios primarios (100%) y aprobaba con dificultad (100%). Tiene pocos amigos (100%) y en su tiempo libre prefiere estar solo (100%), vive con sus padres (100%). No está bajo tratamiento psiquiátrico (100%), no abusa de sustancias (100%) ni cometió el incendio bajo el efecto de alguna sustancia (100%), ni tiene otros problemas de salud (100%). Tuvo una infancia y crianza normales (100%), no ha sido detenido anteriormente por otros motivos (100%) ni estaba sometido a vigilancia policial (100%). Está casado (100%), es menor de 34 años (100%), no conoce de nada al propietario (100%) ni hay ninguna relación (100%) se desplazó en un todo terreno (100%) y si era incendiario en serie (100%). La etiqueta propuesta para este conglomerado es **REINCIDENTE MÚLTIPLE 3.**

10. El conglomerado número 10 está constituido por 12 sujetos (4,5%). Su perfil corresponde a un sujeto con una adaptación normal al puesto de trabajo (58%) o no se sabe (42%) y nunca falta al puesto de trabajo (67%) o no se sabe (33%), trabaja en el sector agrícola (25) o no se sabe en qué sector trabaja (25%) o en el sector forestal (17%), no se sabe el tipo de trabajo (42%) o este es de tipo manual (33%) o cualificado (25%), es autónomo (42%) o empleado (25%) o pensionista jubilado (17%) o desempleado (17%) y no se sabe lo que cobra (75%), no se sabe que estudios tiene (50%) o son elementales (25%) y se

desconoce su rendimiento académico (100%). Tiene muchos amigos (42%) o se desconoce (42%) y en su tiempo libre prefiere estar con gente (33%) o se desconoce (58%), vive con su pareja (33%) o se desconoce (33%) o con sus padres (17%). Se desconoce si está bajo tratamiento psiquiátrico (58%) o si lo está (25%), no abusa de sustancias (50%) o se desconoce (50%), y no cometió el incendio bajo el efecto de alguna sustancia (75%), no tiene otros problemas de salud (58%) o si (25%). Se desconoce cómo fue su infancia (92%) o su crianza (75%) o esta fue difícil (17%), no ha sido detenido anteriormente por otros motivos (100%) y no estaba controlado ni vigilado por la policía (92%). Está casado (42%) o soltero (42%), es menor de 46 años (75%) o mayor de 60 años (17%), conoce mucho al propietario (42%) o es el mismo propietario (25%) o no lo conoce de nada (25%), se desplazó en un turismo (50%) o un todo terreno (25%) o a pie (17%) y no era incendiario en serie (92%). La etiqueta propuesta para este conglomerado es **INFORMACION INCOMPLETA.**

11. El conglomerado número 2 está constituido por 8 sujetos (3%). Su perfil corresponden a un único sujeto con una regular adaptación al puesto de trabajo (100%) y que falta regularmente al trabajo (100%), trabaja en el sector agrícola (87%), y es un trabajo de tipo manual (100%), está desempleado (100%) y no tiene ingresos (100%), es analfabeto (87%) y suspendía habitualmente (100%). Tiene pocos amigos (87%) y en su tiempo libre prefiere estar con gente (87%), vive solo (100%). No está bajo tratamiento psiquiátrico (100%), abusa de sustancias (100%), cometió el incendio bajo el efecto de alguna sustancia (100%) y tiene otros problemas de salud (87%). Tuvo una infancia normal (100%) y crianza normales (87%), no ha sido detenido anteriormente por otros motivos (87%) pero estaba siendo investigado como presunto autor (100%). Está casado (100%), tiene entre 34 y 46 años (100%), no conoce de nada al propietario (87%) pero eran vecinos (100%), se desplazó a pie (100%) y no era incendiario en serie (100%). La etiqueta propuesta para este conglomerado **REINCIDENTE MÚLTIPLE 4.**

Asociación entre los espacios del hecho y de autor.

La distribución de frecuencias conjunta reflejada en la Tabla 5 permite afirmar que existe relación entre ambos métodos de clasificación, con un chi-cuadrado significativo (χ^2=517,5; gl=60; p<0,005). Analizando la Tabla 6, observamos que la solución de 2 dimensiones permite explicar el 80% de la inercia disponible en los datos y se ha seleccionado ese número de dimensiones para crear la solución.

Tabla 5. Tabla de contingencias de los conglomerados del hecho (H) frente a los del sujeto (S).

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Total
H1	0	0	4	0	12	1	20	0	20	5	4	66
H2	0	0	3	0	15	1	5	0	10	2	2	38
H3	0	0	5	0	6	6	1	0	0	1	0	19
H4	0	0	1	0	4	0	18	0	5	3	5	36
H5	0	0	11	0	4	5	4	0	0	1	4	29
H6	0	8	5	15	4	10	6	11	0	0	0	59
H7	19	0	0	0	0	0	0	1	0	0	0	20
Total	19	8	29	15	45	23	54	12	35	12	15	267

Tabla 6. Estadísticos de resumen de las dimensiones obtenibles.

					Proporción de inercia		Confianza para el Valor propio	
Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Explicada	Acumulada	Desviación típica	Correlación 2
1	,976	,952			,491	,491	,021	,229
2	,776	,602			,311	,802	,035	
3	,485	,236			,122	,924		
4	,332	,110			,057	,980		
5	,152	,023			,012	,992		
6	,121	,015			,008	1,000		
Total		1,938	517,549	,000[a]	1,000	1,000		

a. 60 grados de libertad

La solución se resume a continuación:

- La solución muestra la proximidad del sujeto reincidente múltiple 1 (S1) con los hechos sin sentido beneficio agrícola (H7) en el extremo izquierdo de la dimensión 1.
- En el extremo superior de la dimensión 2 se ubican los hechos correspondientes a los sujetos reincidente múltiple 4 (S2), reincidente múltiple 3 (S8) y reincidente múltiple 2 (S4) que se encuentran próximos a los hechos de sin sentido forestal (H6).
- Los sujetos del tipo joven consumidor de sustancias con antecedentes (S6) se encuentran a media distancia entre los hechos sin sentido forestal (H6) e instrumental ganadero (H3).
- Los sujetos del tipo asocial consumidor (S3) se encuentran muy próximos a los hechos instrumental ganadero (H3) y sin sentido beneficio cinegético (H5).
- Los eventos correspondientes a los conglomerados de sujetos, jubilado propietario (S5) y joven trabajador (S7) se encuentran próximos a los hechos de imprudente agrícola que escapa (H2).
- Los sujetos parado social propietario (S11) se encuentra en la frontera de imprudente agrícola que escapa (H2), imprudente agrícola presente (H1) e imprudente recreativo forestal (H4).

- Los casos de sujetos del tipo soltero propietario (S9) y aquellos con la información incompleta (S10) se ubican junto a los hechos imprudente agrícola presente (H1) e imprudente recreativo forestal (H4).

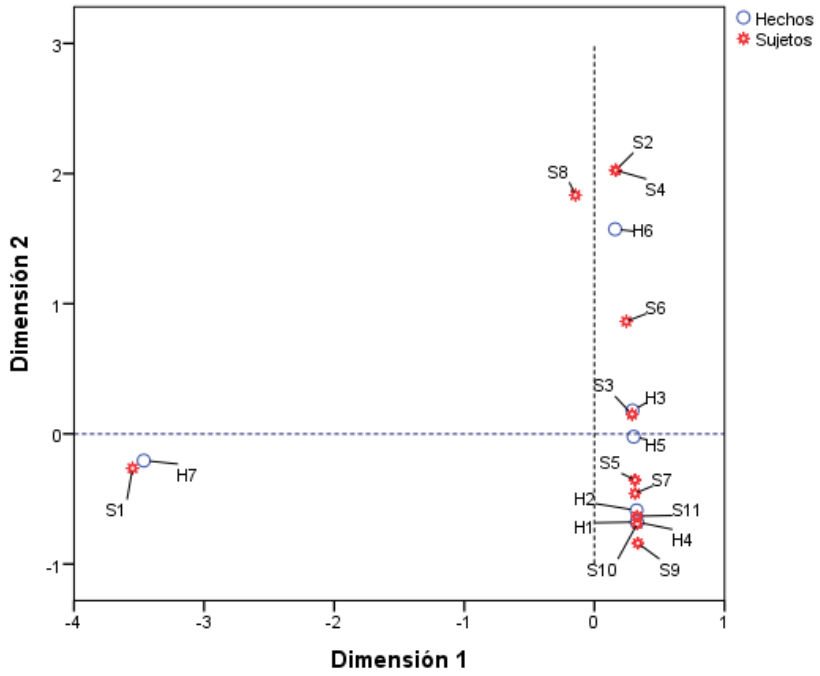


Figura 19. Solución de 2 dimensiones del análisis de correspondencias de hechos y sujetos.

Validación del modelo

En la Tabla 7 podemos observar la distribución por campañas de las muestras de estimación o histórica (2009 y 2010) y de validación (2011).

Tabla 7. Distribución de hechos registrados por año.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Campaña 2009	160	42,3	42,3	42,3
	Campaña 2010	107	28,3	28,3	70,6
	Campaña 2011	111	29,4	29,4	100,0
	Total	378	100,0	100,0	

Validación del Espacio del hecho. Podemos observar en la Figura 20 y en la Tabla 8 que la distribución de tamaños de los conglomerados no es la misma en las dos muestras. La prueba de homogeneidad nos confirma que la distribución de hechos del 2011 clasificados según el modelo de 2010 es significativamente diferente a la de los hechos de años anteriores (χ^2=14,8; gl=6; p=0,022).

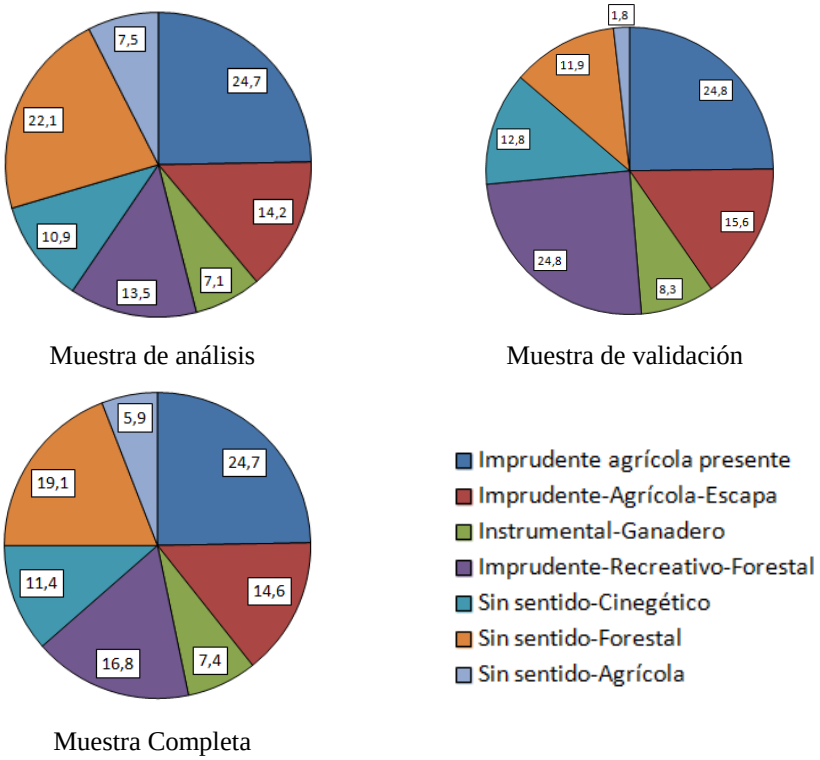


Figura 20. Tamaño de los conglomerados del hecho modelo 2010

Tabla 8. Tabla de contingencia Pronóstico7C * Año<2011 (FILTER)

			Año<2011 (FILTER)	
			Validación	Histórico
Pronóstico7C	Imprudente agrícola presente	Frec.	27_a	66_a
		% Col	24,8%	24,7%
		Residuos z	,0	,0
	Imprudente-Agrícola-Escapa	Frec.	17_a	38_a
		% Col	15,6%	14,2%
		Residuos z	,3	-,3
	Instrumental-Ganadero	Frec.	9_a	19_a
		% Col	8,3%	7,1%
		Residuos z	,4	-,4
	Imprudente-Recreativo-Forestal	Frec.	27_a	36_b
		% Col	24,8%	13,5%
		Residuos z	2,7	-2,7
	Sin sentido-Cinegético	Frec.	14_a	29_a
		% Col	12,8%	10,9%
		Residuos z	,5	-,5
	Sin sentido-Forestal	Frec.	13_a	59_b
		% Col	11,9%	22,1%
		Residuos z	-2,3	2,3
	Sin sentido-Agrícola	Frec.	2_a	20_b
		% Col	1,8%	7,5%
		Residuos z	-2,1	2,1
Total		Frec.	109	267
		% Col	100,0%	100,0%

Cada letra de subíndice indica un subconjunto de Año<2011 (FILTER) categorías cuyas proporciones de columna no difieren significativamente entre sí en el nivel ,05.

La diferencia de patrón entre las dos muestras es debida a 3 categorías. En primer lugar, en la muestra de validación sólo hay dos casos pertenecientes al conglomerado *Sin sentido-Agrícola* (z=-2,1). En segundo lugar el número de hechos pertenecientes a la clase *Imprudente-Recreativo-Forestal* ha aumentado del 13,5% al 24,3% en la muestra de validación frente a la de análisis (z=2,7). En tercer lugar, el número de hechos correspondientes a la clase *Sin sentido-forestal*, se ha reducido a casi la mitad respecto a la muestra de estimación (z=-2,3).

Los resultados sugieren que el patrón de hechos ha cambiado respecto a años anteriores, con un incremento de los hechos del tipo *Imprudente-Recreativo-Forestal* y una disminución de los del tipo *Sin sentido-Agrícola* y *Sin sentido-forestal*. No obstante si analizamos la composición de estos dos patrones observamos que en el caso del tipo *Sin sentido-Agrícola* está formado por un solo incendiario múltiple y el *Sin sentido-forestal* está formado por 6 incendiarios múltiples. Al no haber actuado estos incendiarios múltiples en el año 2011, son patrones difícilmente reproducibles en la muestra de validación. Comparando las soluciones sin estos incendiarios observamos que no hay diferencias significativas entre las distribuciones de los hechos en la muestra de análisis y de validación, (χ^2=8,17; gl=6; p=0,226).

Con el objeto de proporcionar un índice global de concordancia, hemos calculado el IC del hecho entre la solución de la muestra total con el modelo de 2010 y con el modelo de 2011 obteniendo un valor de concordancia entre las soluciones del 71,4%.

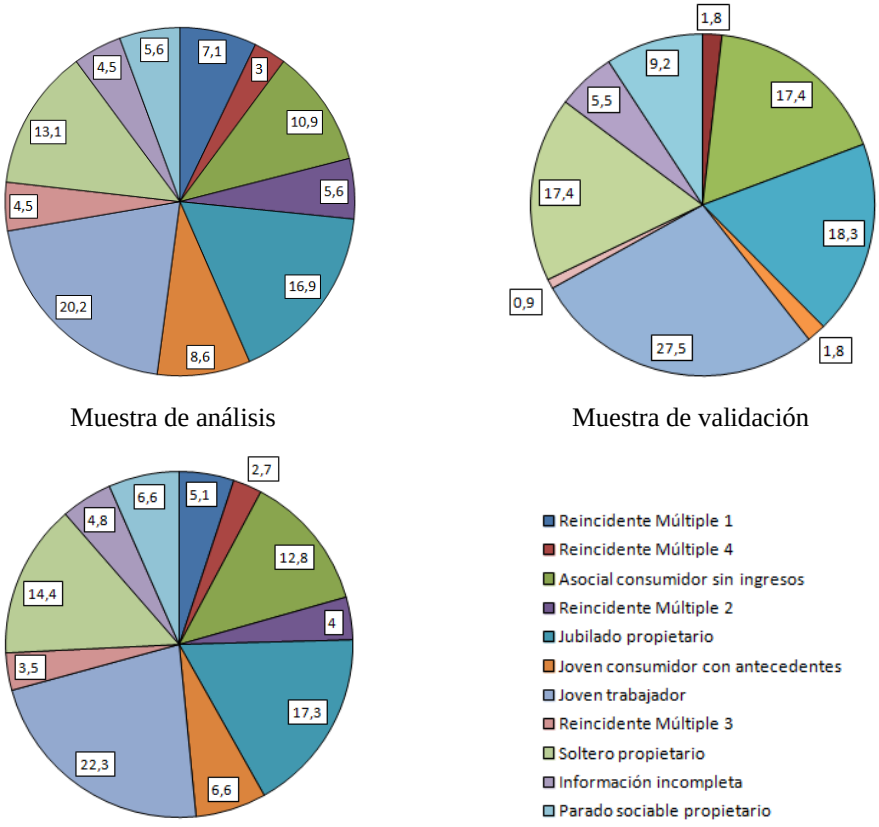
En la comparación de las soluciones del modelo de 2010 realizada mediante el análisis bietápico y el *k*-medias hemos obtenido un índice de concordancia entre las conglomeraciones del 76%.

Validación del Espacio del autor. Podemos observar en la Figura 21 y en la Tabla 9 que la distribución de tamaños de los conglomerados no es la misma en las dos muestras. La prueba de homogeneidad nos confirma que la distribución de autores del 2011 clasificados según el modelo de 2010 es significativamente diferente a la de los autores de años anteriores (χ2=29,7; gl=10; p=0,001).

La diferencia de patrón entre las dos muestras es debida principalmente a 5 categorías. En la muestra de validación no hay ningún caso correspondiente a los tipos *Reincidente Múltiple 1* (z=-2,9) *y 2* (z=-2,5), y hay un solo caso del tipo *Reincidente Múltiple 3* (z=-1,7). También se aprecia un aumento del tipo *Asocial consumidor sin ingresos* del 10,9% al 17,4%

(z=1,7), y una disminución del 8,6% al 1,8% del patrón *Joven consumidor con antecedentes* (z=-2,4).

Estos resultados sugieren que el patrón de autores ha cambiado respecto a años anteriores, con un decremento del tipo *Joven consumidor con antecedentes* y de los reincidentes múltiples que no han vuelto a actuar, por lo que no se registran casos de este tipo, aunque existan dos sujetos que reproduzcan el patrón del *Reincidente Múltiple 4*. Si analizamos los sujetos que conforman a los perfiles Reincidentes Múltiples, podemos observar que no están presentes en la muestra de validación, con lo que el χ^2 se ha visto perjudicado por la presencia de patrones no reproducibles. Realizando la comparación de las dos distribuciones de patrones sin tener en cuenta a los reincidentes múltiples observamos que no existen diferencias significativas entre las distribuciones de los autores en la muestra de análisis y de validación, (χ2=9,9; gl=7; p=0,195).



Muestra de análisis

Muestra de validación

Muestra Completa

Figura 21. Tamaño de los conglomerados del autor. Modelo 2010.

Tabla 9. Tabla de contingencia Pronóstico11C * Año<2011 (FILTER)

			Año<2011 (FILTER)	
			Validación	Histórico
Pronóstico11C		Frec.	0_a	19_b
	Reincidente Múltiple 1	% Col	0,0%	7,1%
		Residuos z	-2,9	2,9
		Frec.	2_a	8_a
	Reincidente Múltiple 4	% Col	1,8%	3,0%
		Residuos z	-,6	,6
		Frec.	19_a	29_a
	Asocial consumidor sin ingresos	% Col	17,4%	10,9%
		Residuos z	1,7	-1,7
		Frec.	0_a	15_b
	Reincidente Múltiple 2	% Col	0,0%	5,6%
		Residuos z	-2,5	2,5
		Frec.	20_a	45_a
	Jubilado propietario	% Col	18,3%	16,9%
		Residuos z	,3	-,3
		Frec.	2_a	23_b
	Joven consumidor con antece-dentes	% Col	1,8%	8,6%
		Residuos z	-2,4	2,4
		Frec.	30_a	54_a
	Joven trabajador	% Col	27,5%	20,2%
		Residuos z	1,5	-1,5
		Frec.	1_a	12_a
	Reincidente Múltiple 3	% Col	0,9%	4,5%
		Residuos z	-1,7	1,7
		Frec.	19_a	35_a
	Soltero propietario	% Col	17,4%	13,1%
		Residuos z	1,1	-1,1
		Frec.	6_a	12_a
	Información incompleta	% Col	5,5%	4,5%
		Residuos z	,4	-,4
		Frec.	10_a	15_a
	Parado sociable propietario	% Col	9,2%	5,6%
		Residuos z	1,3	-1,3
Total		Frec.	109	267
		% Col	100,0%	100,0%

Cada letra de subíndice indica un subconjunto de Año<2011 (FILTER) categorías cuyas proporciones de columna no difieren significativamente entre sí en el nivel ,05.

Con el objeto de proporcionar un índice global de concordancia, hemos calculado el IC del autor entre la solución de la muestra total con el modelo de 2010 y con el modelo de 2011 obteniendo un valor de concordancia entre las soluciones del 79,6%.

En la comparación de las soluciones del modelo de 2010 realizada mediante el análisis bietápico y el *k*-medias hemos obtenido un índice de concordancia entre las conglomeraciones del 79%.

Representación gráfica de los perfiles

La representación grafica propuesta para una adecuada lectura de las distribuciones de los perfiles en cada variable es un grafico radial como los de las Figura 22 y Figura 23 y para los perfiles de los conglomerados un gráfico de perfiles como los de las Figura 24 y Figura 25.

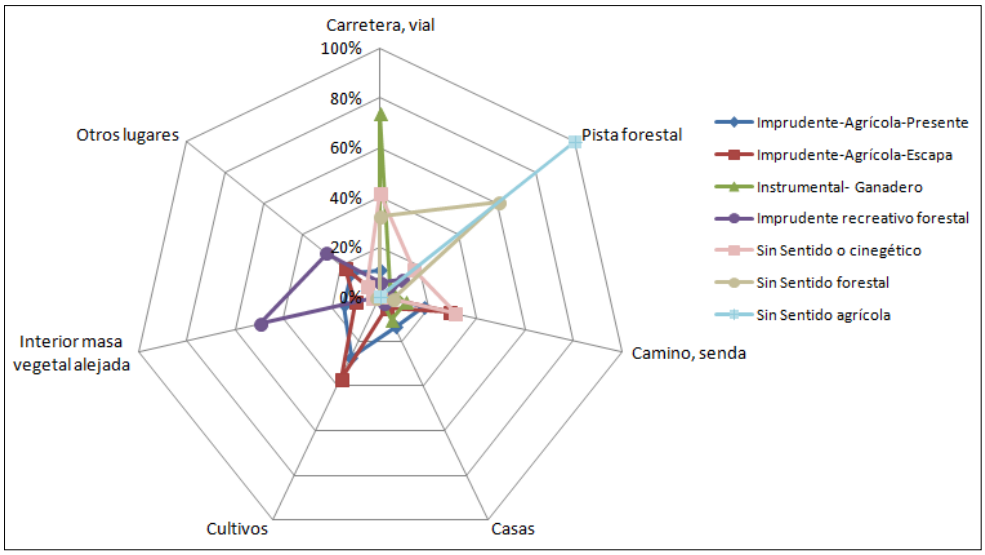


Figura 22. Gráfico radial de la variable del hecho "Superficie en el punto de inicio".

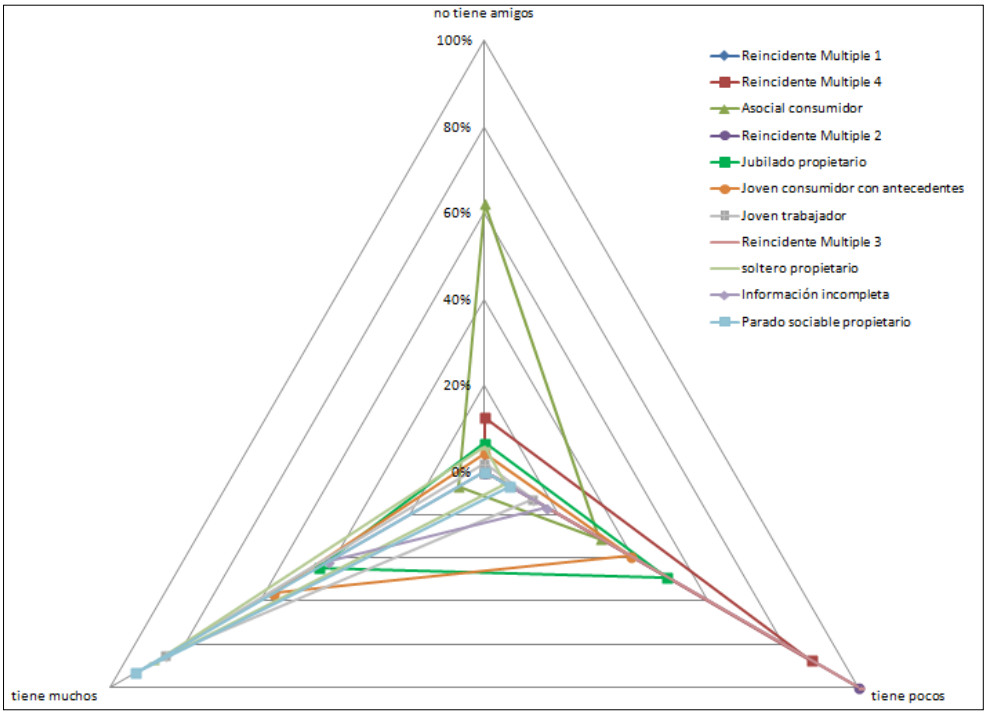


Figura 23. Gráfico radial de la variable de autor "Relaciones sociales".

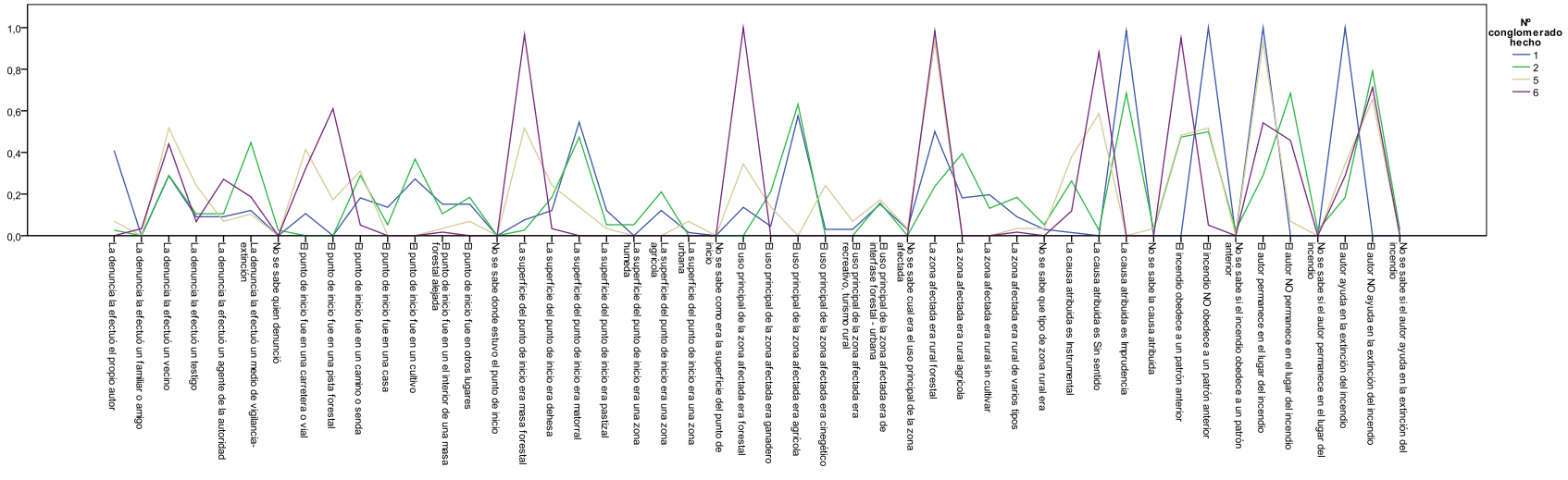


Figura 24. Gráfico de los centros de los perfiles de los conglomerados del hecho 1, 2, 5 y 6.

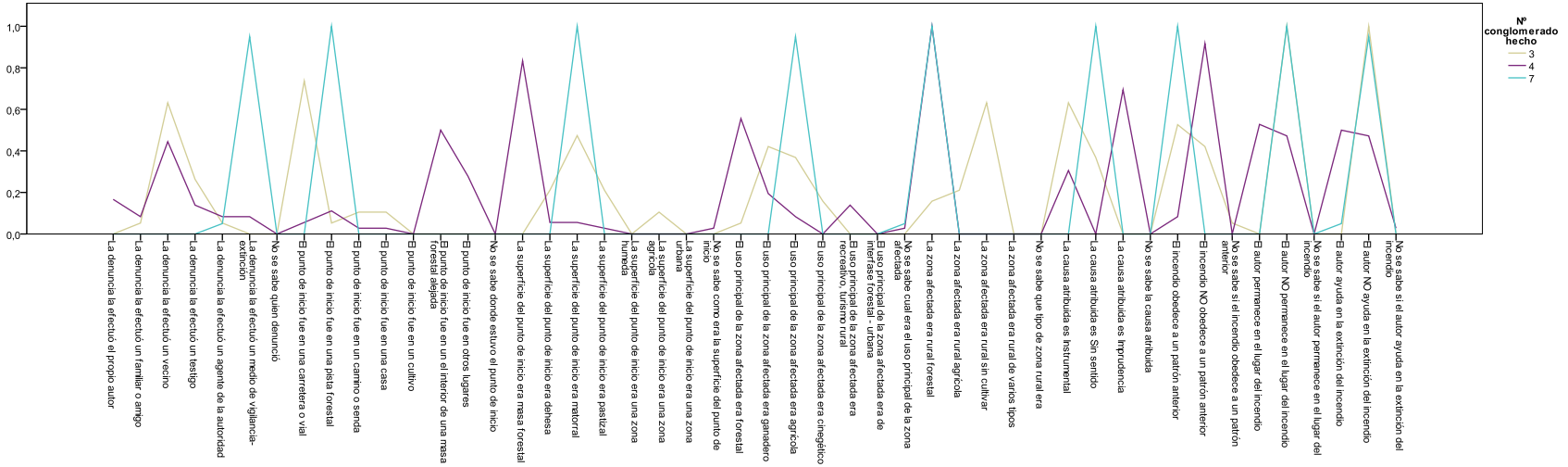


Figura 25. Gráfico de los centros de los perfiles de los conglomerados del hecho 3, 4 y 7.

Discusión y Conclusiones

El objetivo principal de este estudio ha sido el diseño de una metodología de análisis que permita la agrupación de objetos de dos bloques de información relacionados (en nuestro caso, hechos y autores de esos hechos) en clases para, posteriormente, analizar las relaciones existentes entre las agrupaciones. En último término, la clasificación realizada de manera independiente en ambos bloques debe permitir que, conociendo la pertenencia de un hecho a una de las clases del bloque de los hechos, podamos predecir la clase más probable del bloque de autores y, con ello, las características potenciales de un sujeto como autor del hecho. La técnica de análisis estadístico identificada con mayor potencial ha sido el análisis de conglomerado bietápico. El estudio llevado a cabo permite valorar, desde el punto de vista metodológico, la influencia de las decisiones analíticas previas que se deben tomar para optimizar las agrupaciones. Estas decisiones, han demostrado ser críticas en la solución final, y son la elección de las variables a emplear en las agrupaciones y el número de agrupaciones considerado adecuado para conformar la solución final.

Tras el análisis univariante de las variables se realizó un proceso de depuración de datos observándose la presencia de casos con más del 60% de los datos con valores perdidos. Esta situación provocó la pérdida de aproximadamente el 10% de la muestra, pasando de 300 casos a los 267 finales. Esta tarea de depuración es imprescindible para poder acometer cualquier análisis, si bien lo ideal sería establecer un proceso de control previo en el proceso de recogida de los datos.

Los resultados demuestran que el empleo de los procedimiento automáticos que establecen los programas estadísticos al uso no constituyen una estrategia óptima de actuación cuando las muestras empleadas no son perfectas, como suele ocurrir con las muestras reales, por lo que se plantea la necesidad del estudio y análisis en profundidad tanto de los pasos intermedios de los procedimientos como de los criterios de las decisiones automáticas que emplean los programas informáticos. Este estudio previo es un requisito indispensable para obtener la mejor solución a nuestro problema tras la aplicación de la técnica a los datos.

Los análisis de sensibilidad al orden que hemos realizado demuestran que la influencia del orden de los casos afecta a la solución final, a través de su influencia en tres aspectos: la selección de las variables definitivas, la selección del número de conglomerados y la solución final en sí misma.

La influencia del orden de los casos en la selección de las variables de clasificación es mayor en el espacio del hecho (Figura 1 y Figura 2) que en el espacio del autor (Figura 10 y Figura 11). Si bien en los dos espacios se aprecia que las variables que menos se ven influenciadas por el orden son las que se encuentran en los extremos de la importancia de las variables (IV), es decir, las que menos influyen y las que más influyen en la solución elegida.

Se observa también que la variabilidad de la influencia del orden de los casos es mayor en el espacio del hecho que en el del autor, lo que sugiere que existe una estabilidad mayor en la muestra de autores que en la de los hechos. Esta estabilidad puede ser aportada por la presencia de casos correspondientes a autores reincidentes existentes en la muestra, que proporcionan grupos de casos en los que todas las variables toman el mismo vector de valores.

Por tanto, podemos afirmar que existe un efecto del orden de los casos en la selección de las variables, si bien este efecto disminuye con la estabilidad de los datos en la muestra. Esta estabilidad de los datos la proporciona la existencia de grupos de casos con valores similares en las variables analizadas, es decir, con la existencia de grupos naturales en los datos, y que, cuanto mayor es el numero de variables incluido se hace más difícil la presencia de grupos nítidamente definidos, aumentando el solapamiento y confusión entre clusters, lo que se ha denominado "The curse of dimensionality" (Bellman, 1961). Este efecto de confusión es debido a que un exceso de variables irrelevantes reduce el rendimiento en la conglomeración al provocar que las distancias entre casos sean cada vez mas homogéneas y, por tanto, los casos y los grupos sean cada vez más parecidos (Müller, Günnemann, Assent, y Seidl, 2009; Moise, Sander y Ester, 2008; Beyer, Goldstein, Ramakrishnan, y Shaft, 1999; Fielding, 2007; Chizi y Maimon, 2005).

El análisis de la importancia de las variables nos permite eliminar del estudio las variables que menos influencia tienen en la agrupación. Esta reducción de las variables necesarias tiene dos objetivos, por un lado minimizar el efecto de confusión producido por el exceso de variables, y por otro lado minimizar el número de parámetros a estimar por el modelo, lo que aumentaría los grados de libertad, la estabilidad y la generalizabilidad del modelo.

Hemos recurrido a la valoración criterial externa de jueces para determinar que variables suprimir conjugando el valor de la IV calculada y su importancia teórica (especialmente para la investigación policial), de esta forma hemos eliminado todas aquellas variables que no alcanzaban un valor de 0,25 en la importancia IV y que fueron consideradas prescindibles por los jueces. En el espacio del hecho se han eliminado un total de 10 variables y 11 en el espa-

cio de autor, reduciendo de esta forma el número de parámetros a estimar en 174 en el espacio del hecho y 373 en el del autor.

En cuanto al efecto del orden en la decisión del número de conglomerados que tendrá la solución final, de nuevo se aprecia que el efecto es mayor en el espacio del hecho que en el espacio del autor. En las Figuras 5, 6 y 14, se observa la variabilidad en la razón de medidas de distancia, criterio empleado por el procedimiento bietápico para la selección automática del número de conglomerados. Los resultados del análisis de sensibilidad muestran que en el caso del espacio del hecho el SPSS no es capaz de decantarse por una solución estable en todas las repeticiones analizadas. Las soluciones de 2, 3 y 5 conglomerados tienen la misma posibilidad de ser elegidas. En el espacio del hecho sí se observa que la solución automática realizada por el procedimiento es la misma en las cien repeticiones, lo que sugiere que el efecto del orden en la solución final del número de conglomerados disminuye, de nuevo, con la estabilidad de los datos en la muestra. A pesar de la solución estable ofrecida por el procedimiento para el espacio de autor, se ha comprobado que dicha solución está sesgada por la presencia de los incendiarios reincidentes, que proporcionan estabilidad a la solución al representar agrupaciones naturales pero proporcionan una solución muy poco descriptiva para el resto de sujetos. Por ello, para decidir la solución final de número de conglomerados tanto en el espacio del hecho como del autor, es necesario recurrir a la comparación de distintas soluciones teóricas y seleccionar la que permita agrupaciones estables, que sean suficientemente descriptivas y que sean fácilmente interpretables.

Habiendo decidido para cada espacio, el número de variables finales necesarias para el modelo, decidido el número óptimo de conglomerados finales y estimada la solución final, el análisis de sensibilidad al orden realizado mediante el cálculo del IC en los dos espacios permite afirmar de nuevo la existencia de un efecto del orden en la asignación de los casos a los conglomerados. Se vuelve a constatar que el efecto es mayor en el espacio del hecho que en el espacio del autor, con una diferencia de medias de 6,76 puntos, ($t=-11,05$; $p<0,005$). A pesar de que el efecto del orden es menor en el espacio del autor, consideramos este efecto relevante pues, empleando las mismas variables y número de conglomerados en 100 soluciones reordenadas, por término medio, el 29,96% de los hechos y el 22,2% de los sujetos no se asignan al mismo conglomerado en las distintas soluciones de agrupación.

Los estadísticos empleados por el SPSS para la selección definitiva del número de conglomerados son el BIC y la razón de medidas de distancia según la fórmula (5). El BIC es una medida de desajuste global que nos permite comparar modelos y está penalizado por el

número de parámetros y el tamaño muestral. Según este estadístico, las soluciones a elegir hubiesen sido la de 5 conglomerados para el hecho (Figura 8) y la de 5, 6 o 7 conglomerados para el autor (Figura 14). Esta situación se debe a la mayor penalización que reciben las soluciones de más conglomerados al tener que estimar más parámetros. Sin embargo del análisis de la formula: $l = \sum_{k=1}^{K} \sum_{i \in I} \log\left(p(x_i|\theta_k)\right) = \sum_{k=1}^{K} -n_k \cdot \sum_{p=1}^{P} \sum_{c=1}^{C_j} \hat{\pi}_{kpc} \cdot \log\left(\hat{\pi}_{kpc}\right)$, donde $\hat{\pi}_{kpc}$ es la probabilidad de que un sujeto del conglomerado *k* tome la categoría *c* en la variable *p*, se desprende que, una vez fijado el número de variables en el análisis para cada *k* conglomerados, y asumiendo que $\hat{\pi}_{kpc}$ se mantiene constante en distintas muestras al tener la misma distribución, la fórmula del BIC quedaría simplificada a: BIC= -2·*n*·Cte + cte·Ln(*n*). Pudiéndose apreciar que la penalización por *n* de los modelos más complejos es menor según aumenta *n*. Por lo que es de esperar que según aumente el tamaño muestral las soluciones de más número de conglomerados no se vean tan perjudicadas, entrando a formar parte de las posibles soluciones finales.

Por otro lado, el segundo criterio empleado por el SPSS en la selección del numero de conglomerados finales (la razón de medidas de distancia), analiza el incremento relativo en la entropía según se funden los conglomerados más homogéneos en cada paso. Al igual que ocurre con cualquier procedimiento jerárquico, las solución elegida depende del salto relativo en distancia, es decir aquella en la que al unir dos conglomerados el salto es mayor, con lo que llevar a cabo ese paso de fusión no sería aconsejable. Pero, como observamos en la Figura 12 para el espacio del hecho, la fusión sucesiva de conglomerados desde la solución de 15 conglomerados no presenta ningún salto en distancia que nos permita tomar una decisión, hasta llegar a la solución de 5 conglomerados. De hecho, la solución propuesta por el procedimiento automático del análisis de conglomerados bietápico del SPSS no es estable. Dependiendo del orden de los casos en el archivo de datos selecciona las soluciones de 2, 3, 4 y 5 conglomerados el 28, 37, 9 y 26% de las veces, respectivamente. Sin embargo estas soluciones son muy poco descriptivas, agrupar a los incendios solamente en "imprudentes", "sin sentido" o "instrumentales", da lugar a un sistema de clasificación que no tendría mucha capacidad predictiva para clasificar los casos futuros. Habiendo rechazado estas soluciones por falta de capacidad descriptiva, no disponemos de ningún estadístico que nos permita tomar una decisión analíticamente.

A pesar de que el espacio del autor presenta una mayor estabilidad, al seleccionarse el 100% de las veces la solución de 5 conglomerados, la presencia de los reincidentes múltiples que se agrupan formando conglomerados muy definidos, nos deja de nuevo con soluciones

muy poco descriptivas. En el espacio del autor, observamos en la Figura 18 que, al recorrer las soluciones por orden decreciente desde la de 15 conglomerados, sí existen dos soluciones anteriores a la de 5 que presentan un salto un poco mayor en la distancia relativa, las soluciones de 7 y 11 conglomerados.

El estudio de los dendrogramas obtenidos en el análisis jerárquico sugiere que las soluciones recomendadas en el espacio del hecho podrían estar formadas por 2, 3, 5, 7 o 9 conglomerados; y en el espacio de autor podrían seleccionarse 2, 3, 4, 6, 7, 8 u 11 conglomerados; dependiendo del grado de discriminación entre conglomerados que se desee asumir.

Finalmente , y tras el análisis de varias soluciones donde se ha tenido en cuenta tanto los resultados de los estadísticos analizados, como el aporte teórico y la funcionalidad de los modelos, proponemos inicialmente las solución de 7 conglomerados para el espacio del hecho y 11 conglomerados para el espacio de autor.

La solución propuesta para el espacio del hecho clasifica los incendios forestales en 7 grupos, a los que se les ha puesto un nombre identificativo según sus centroides: Imprudente agrícola presente, Imprudente agrícola que escapa, Imprudente recreativo forestal, Sin sentido cinegético, Sin sentido forestal, Sin sentido agrícola e Instrumental ganadero.

En el caso del espacio de autor, la solución propuesta distribuye a los incendiarios en 11 grupos denominados: Joven trabajador, Joven consumidor con antecedentes, Jubilado propietario, Soltero propietario, Asocial consumidor sin ingresos, Parado sociable propietario, Información incompleta, Reincidente Múltiple 1, Reincidente Múltiple 2, Reincidente Múltiple 3 y Reincidente Múltiple 4.

En cuanto a la relación entre los dos espacios de agrupación, el análisis de correspondencias nos permite afirmar que existe una relación entre los espacios del hecho y del autor, y que con dos dimensiones podemos explicar el 80% de la inercia disponible en los datos.

En la dimensión1 se aprecia que está fuertemente influenciada por el sujeto Reincidente Múltiple 1 y por los hechos cometidos por este autor, lo que provoca la concentración del resto de hechos y sujetos en una zona restringida de esa dimensión, lo que dificulta su interpretación.

En la dimensión 2 se observa que los valores altos corresponden a hechos sin sentido realizados por sujetos reincidentes, los valores intermedios son hechos instrumentales realizados por sujetos consumidores de sustancias, y los valores inferiores son hechos imprudentes cometidos por personas “normales”. Destaca la presencia de propietarios en la zona interme-

dia de los hechos instrumentales e imprudentes, lo que se podría interpretar como un intento de enmascaramiento de los incendios instrumentales con los imprudentes. Esta dimensión sugiere la motivación del autor en la comisión del hecho.

Para analizar la estabilidad del modelo hemos utilizado la campaña de recogida de datos del 2011, empleándola como muestra de validación. Dado el reducido número de casos con el que hemos contado para el estudio, hubiera sido razonable volver a estimar la estructura de los conglomerados incorporando los nuevos casos y por tanto volver a realizar el análisis sobre el número idóneo de conglomerados y dando también la oportunidad a las variables descartadas inicialmente a contribuir en el modelo. Debemos tener en cuenta que estamos incorporando un 41,5% de casos nuevos, lo que supone que en la nueva masa de datos el 29% de los casos serán nuevos. Por ello, es esperable que la nueva solución pueda variar de manera sensible. No obstante, tan solo vamos a comentar la variación manteniendo las condiciones de variables y conglomerados constantes para poder comparar con mayor facilidad.

Los resultados de validación obtenidos sugieren que la solución para el espacio del hecho es relativamente estable. Excepto para alguno de los nuevos conglomerados, como el 2, que muestran una tendencia a desplazarse para mezclarse con otros conglomerados, lo que posiblemente sea debido a que ha variado el número de casos en cada perfil (ha aumentado en la muestra de validación el perfil *Imprudente-Recreativo-Forestal* de un 13,5% a un 24,8% en el período de validación), así como a la ausencia en la muestra de validación de los cuatro principales incendiarios reincidentes.

La solución para el espacio del autor es más estable, excepto para alguno de los nuevos conglomerados, como el 1 o el 7, que muestran una tendencia a desplazarse para mezclarse con otros conglomerados, lo que posiblemente sea debido a la desaparición de los cuatro perfiles de reincidentes múltiples.

Los índices de concordancia obtenidos en las comparaciones de las soluciones obtenidas con el procedimiento bietápico y con el de k-medias, que alcanzan un 76% en el espacio del hecho y un 79% en el espacio del autor respaldan la relativa estabilidad del modelo.

Las principales limitaciones de este estudio son el reducido tamaño muestral, el incumplimiento de los supuestos del análisis y el empleo de una muestra incidental.

La medida de distancia empleada por el análisis de conglomerados bietápico del SPSS es el incremento en el logaritmo de la verosimilitud. Sin embargo, el procedimiento no hace referencia en ningún momento a la relación entre el tamaño muestral y el número de paráme-

tros a estimar, ni a los grados de libertad del modelo. En el espacio del hecho partimos inicialmente de 19 variables con un total de 81 categorías, dando lugar a la necesidad de estimar 62*K parámetros, por lo que con una muestra de 267 estarían infraidentificados todos los modelos a partir del modelo de 5 conglomerados. En el espacio del autor la situación es aún peor, pues partimos inicialmente de 36 variables con un total de 158 categorías, dando lugar a la necesidad de estimar 122*K parámetros, por lo que los modelos de más de tres conglomerados estarían infraidentificados. A pesar de la reducción del número de variables realizada, la solución final de 7 conglomerados del hecho y 11 de autor son modelos infraidentificados.

Por otra parte en nuestra muestra se incumple el supuesto de independencia de las variables en el modelo, si bien el incumplimiento de este supuesto no es crítico (Chiu et al 2001). El estudio se ha realizado sobre una muestra incidental, y de tamaño reducido, lo que limita la posibilidad de generalización de los resultados a otras poblaciones.

A pesar de estas limitaciones, que nos hacen tomar estos resultados como provisionales, consideramos que la metodología de trabajo propuesta es adecuada para enfrentar bases de datos de delitos y delincuentes en las que las variables son categóricas. No obstante recomendamos tomar las siguientes acciones: 1) estudiar el cuestionario de recogida de datos para reducir el número final de variables y número de categorías de cada variable con el objetivo de minimizar el numero de parámetros a estimar por el modelo; 2) continuar con la recogida de datos en cada campaña para aumentar la muestra de estudio, estas dos medidas redundarían en el aumento de los grados de libertad y en una mejora de la generalizabilidad y estabilidad del modelo; 3) implementar la figura del "Monitor del trabajo de campo" lo que mejoraría la calidad de la muestra en su recogida; y 4) repetir toda la metodología desde la selección de variables hasta la exploración de la solución final cuando se obtenga una muestra de mayor tamaño.

Consideramos interesante abrir como líneas de investigación a partir de este estudio, la elaboración de reglas de producción que permitan poner a prueba el modelo en la próxima campaña de incendios forestales, así como la extrapolación de la metodología propuesta a bases de datos de delitos diferentes al incendio forestal, como los abusos sexuales o el crimen organizado, delitos donde las bases de datos actuales cuentan con muestras muy amplias pues la tasa de detenciones en estos delitos es sensiblemente mayor. También consideramos interesante el ensayo con modelos de redes neuronales cuando el tamaño muestral sea mayor.

Referencias

Alison, L., Bennell, C., Mokros, A., y Ormerod, D. (2002). The personality paradox in offender profiling: A theoretical review of the processes involved in deriving background characteristics from crime scene actions. *Psychology, Public Policy, and Law, 8*, 115-135.

Alison, L., Goodwill, A., y West, A. (2004). The Academic and the practitioner. Pragmatists' views of Offender Profiling. *Psychology, Public Policy and Law, 10*, 71-101.

Alison, L., Goodwill, A., Almond, L., Van de Heuvel, C., y Winter, J. (2010). Pragmatic solutions to offender profiling and behavioural investigative advice. *Legal and Criminological Psichology 15*, 115-132.

Bacher, J., Wenzig, K. y Vogler, M. (2004). SPSS TwoStep Cluster – A First Evaluation. Arbeits- und Diskussionspapiere 2004-2, 2, korr. Aufl. Erlangen-Nürnberg: Friedrich-Alexander Universität. Extraído el 27 de octubre de 2011 de http://www.soziologie.wiso.uni-erlangen.de/ publikationen/a-u-d-papiere/archiv-.shtml

Bellmann, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ., USA: Princeton University Press.

Bennell, C., Jones N., Taylor P.J., Snook B (2006). Validities and abilities in criminal profiling: a critique of the studies conducted by Richard Kocsis and his colleagues. *International Journal of Offender Therapy and Comparative Criminology 50(3)*, 344-360.

Beyer, K., Goldstein, J., Ramakrishnan, R., y Shaft, U. (1999). When is nearest neighbor meaningful? In C. Beeri, P. Buneman (Eds.), *Lecture Notes in Computer Science: Vol. 1540.* Database Theory - ICDT 1999 (pp. 217-235). Berlin, Germany: Springer-Verlag. doi: dx.doi.org/10.1007/3-540-49257-7_15

Burnham, K.P. y Anderson D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research 33(2)*, 261–304.

Canter, D., y Fritzon, K. (1998). Diferentiating arsonists: A model of firesetting actions and characteristics. *Legal and Criminological Psychology, 3*, 73–96.

Canter, D. (2000). Offender profiling and criminal differentiation. *Legal and Criminological Psychology 5*, 23–46.

Canter, D. (2004). Offender profiling and Investigative Psychology. *Journal of Investigative Psychology and Offender Profiling 1*, 1–15.

Canter, D., y Youngs, D. (2009). *Investigative psychology: Offender profiling and the analysis of criminal action.* Chichester, England: Wiley.

Canter, D. (2011). Resolving the offender "Profiling Equations" and the emergence of an Investigative Psychology. *Current Directions in Psychological Science, 20*, 5-10.

Chiu, T., Fang, D., Chen, J., Wang, Y., y Jeris, C. (2001). A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2001*, 263–268. doi: 10.1145/502512.502549

Chizi, B., y Maimon, O. (2005). Dimension Reduction and Feature Selection. In Maimon, O. y Rokach, L. (eds) *Data Mining and Knowledge Discovery Handbook*. (2º Edition). New York: Springer.

Devery, C. (2010). Criminal Profiling and Criminal Investigation. *Journal of Contemporary Criminal Justice. 26(4),* 393-409.

Doan, B. y Snook, B. (2008). A failure to find empirical support for the homology assumption in criminal profiling. *Journal of Police and Criminal Psychology, 23(2),* 61–70.

Doley, R. (2003). Making sense of arson through classification. *Psychiatry, Psychology and Law 10 (2)*, 346–352.

Dowden, C., Bennell, C., Bloomfield, S. (2007). Advances in offender profiling: A systematic review of the profiling literature published over the past three decades. *Journal of Police and Criminal Psychology 22*, 44–56.

Eastwood, J., Cullen, R., Kavanagh, J., y Snook, S. (2006). A review of the validity of criminal profiling. *The Canadian Journal of Police & Security Services, 4*, 118-124.

Fielding, A.H. (2007). *Cluster and Classification Techniques for the Biosciences.* Cambridge, U.K.: Cambridge University Press.

Fritzon, K., Canter, D., y Wilton, Z. (2001). The application of an action systems model to destructive behaviour: The examples of arson and terrorism. *Behavioral Science and the Law, 19*, 657–690.

González, J.L., Sotoca, A., Martínez, J.M. y Martín, M.J. (2010). Perfil Psicosocial del Incendiario Forestal. En Jiménez, J. (2010). *Manual Práctico del Perfil Criminológico* 345-370. Valladolid, España: Lex-Nova.

Guha, S., Rastogi, R. y Shim, K. (1999). ROCK: A Robust Clustering Algorithm for Categorical Atributes. In M. Kitsuregawa, M.P. Papazoglou & C. Pu (Eds.), *Proceedings of the 15th International Conference on Data Engineering, (ICDE),* (pp. 512-521), Sydney, Australia: IEEE Press. doi.ieeecomputersociety.org/10.1109/ICDE.1999.754967

Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W. (1995). *Análisis Multivariante.* (5ª edición). Madrid, España: Prentice Hall Iberia.

Hakkanen, H., Puolakka, P., y Santtila, P. (2004). Crime scene actions and offender characteristics in arsons. *Legal and Criminological Psychology, 9*, 197–214.

Hicks, S. J., y Sales, B. D. (2006). *Criminal profiling: Developing an effective science and practice.* Washington, DC, USA: American Psychological Association.

IBM SPSS, Inc. (2010a). *IBM SPSS Statistics 19 Algorithms.* IBM SPSS Inc., Chicago, Il.

IBM SPSS, Inc. (2010b). *IBM SPSS Statistics Base 19.* IBM SPSS Inc., Chicago, Il.

Jaworska N. y Chupetlovska-Anastasova A. (2009). A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains. *Tutorials in Quantitative Methods for Psychology, 2009, Vol. 5*(1), 1-10.

Kocsis, R.N., Irwin H.J. y Hayes A.F. (1998). Organized and disorganized criminal behavior syndromes in arsonists: a validation study of a psychological profiling concept. *Psychiatry, Psychology and Law 5*, 117–131.

Kocsis, R.N. y Cooksey, R.W. (2002). Criminal psychological profiling of serial arson crimes. *International Journal of Offender Therapy and Comparative Criminology 46*, 631–656.

Kocsis, R.N. (2004). Psychological profiling of serial arson offenses: an assessment of skills and accuracy. *Criminal Justice and Behavior 31(3)*, 341–361.

Kocsis, R.N. (2006a). *Criminal profiling: Principles and practice*. Totowa, NJ, USA: Humana Press.

Kocsis R.N. (2006b). Validities and abilities in criminal profiling: The dilemma for David Canter's investigative psychology. *International Journal of Offender Therapy and Comparative Criminology 50*, 458–477.

Kocsis, R. N., Middledorp, J., y Karpin, A. (2008). Taking stock of accuracy in criminal profiling: The theoretical quandary for investigative psychology. *Journal of Forensic Psychology Practice, 8(3)*, 244–261.

Ministerio de Medio Ambiente y Medio Rural y Marino (MARM, 2011). Los incendios forestales en España año 2010. Extraído el 17 de febrero de 2012 de http://www.magrama.gob.es/es/biodiversidad/temas/defensa-contra-incendios-forestales/estadisticas-de-incendios-forestales

Moise, G., Sander, J., y Ester, M. (2008). Robust projected clustering. *Knowledge and Information Systems, vol. 14, no. 3*, 273–298.

Mokros, A. y Alison, L.J. (2002). Is offender profiling possible? testing the predicted homology of crime scene actions and background characteristics in a sample of rapists. *Legal and Criminological Psychology, 7(1),* 25-43.

Muller, D.A. (2008). Offending and reoffending patterns of arsonists and bushfire arsonists in New South Wales. *Trends & Issues in crime and criminal justice*, 348.

Müller, E., Günnemann, S., Assent, I. y Seidl, T. (2009). Evaluating clustering in subspace projections of high dimensional data. In PVLDB, pp. 1270–1281. Extraído el 5 de noviembre de 2011 de http://www.vldb.org/pvldb/2/vldb09-600.pdf

Pastor D.A. (2010). Cluster analysis. In Hancock, G.R. y Mueller, R.O. (Eds.), *The reviewer's guide to quantitative methods in the Social Sciences* (41-54). New York, USA: Routledge.

Santtila, P., Häkkänen, H. y Fritzon, K. (2003). Inferring the characteristics of an arsonist from crime scene behaviour: a case study in offender profiling. *International Journal of Police Science and Management, 5(1)* 1-15.

Snook, B., Cullen, R.M., Bennell, C., Taylor, P. J., y Gendreau, P. (2008). The criminal profiling illusion: What's behind the smoke and mirrors? *Criminal Justice and Behavior, 35(10),* 1257–1276.

Viegas Ferreira, E., y Soeiro, C. (2007). Perfis psicossociais dos incendiários portugueses. Propostas para a prevenção. Informe de investigación. *Jornadas sobre Investigación Criminal de Incendios Forestales,* marzo 2007, Universidad de Santiago de Compostela.

Woodhams, J. y Toye, K. (2007). An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychology, Public Policy, and Law, 13(1),* 59-85.

Zhang, T., Ramakrishnan, R., y Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In Jennifer Widom (Ed.), *Proceedings of the 1996 ACM SIGMOD international conference on Management of data* 103-114. New York, NY, USA. doi: doi.acm.org/10.1145/233269.233324

Tabla 1. Descripción de las variables del hecho.

Variable	Categoría	Recuento	% de N =267
1. Quincena	1ª quincena	107	40,1%
	2ª quincena	160	59,9%
	no se sabe	0	,0%
2. Estación del año	Primavera	49	18,4%
	Verano	159	59,6%
	Otoño	23	8,6%
	Invierno	36	13,5%
	No se sabe	0	,0%
3. Tipo de día de la semana del hecho	Laborable	187	70,0%
	Sábado o víspera	41	15,4%
	Festivo	39	14,6%
	No se sabe	0	,0%
4. Franja horaria de inicio	Mañana (7 a 14)	74	27,7%
	Tarde (15 a 20)	127	47,6%
	Noche (21 a 6)	58	21,7%
	No se sabe	8	3,0%
5. Nivel de riesgo del incendio	Bajo	39	14,6%
	Medio	91	34,1%
	Alto	131	49,1%
	No se sabe	6	2,2%
6. Persona que denuncia	Propio autor	36	13,5%
	Familiares, amigos	6	2,2%
	Vecinos	99	37,1%
	Testigos	31	11,6%
	Agentes autoridad	33	12,4%
	Medios de vigilancia / extinción	61	22,8%
	No se sabe	1	,4%
7. Delito asociado	Sí	8	3,0%
	No	251	94,0%
	No se sabe	8	3,0%
8. Número de focos	Uno	219	82,0%
	Más de uno	46	17,2%
	No se sabe	2	,7%
9. Punto de inicio R	Carretera, vial	54	20,2%
	Pista forestal	66	24,7%
	Camino, senda	38	14,2%
	Casas	14	5,2%
	Cultivos	32	12,0%
	Interior masa vegetal alejada	34	12,7%
	Otros lugares	29	10,9%
	No se sabe	0	,0%

10. Tipo de superficie cerca del punto de inicio	Masa forestal	108	40,4%
	Dehesa	30	11,2%
	Matorral	89	33,3%
	Pastizales	16	6,0%
	Zona húmeda	2	,7%
	Agrícola	18	6,7%
	Urbana	3	1,1%
	No se sabe	1	,4%
11. Uso principal de la zona afectada	Aprovechamiento forestal	99	37,1%
	Aprovechamiento ganadero	30	11,2%
	Aprovechamiento agrícola	91	34,1%
	Aprovechamiento cinegético	12	4,5%
	Uso recreativo / turismo rural	9	3,4%
	Interfase forestal-urbana	21	7,9%
	No se sabe	5	1,9%
12. Afectado urbano	residencial	43	16,1%
	industrial	2	,7%
	vehículos	6	2,2%
	varios de los anteriores	5	1,9%
	No se sabe	211	79,0%
13. Afectado rural	forestal	186	69,7%
	agrícola	31	11,6%
	sin cultivar	30	11,2%
	varios de los anteriores	15	5,6%
	No se sabe	5	1,9%
14. Hallazgo de vestigios	Si	53	19,9%
	No	205	76,8%
	No se sabe	9	3,4%
15. Causa atribuida	Instrumental	52	19,5%
	Sin sentido	97	36,3%
	Imprudencia	116	43,4%
	No se sabe	2	,7%
16. Obedece a un patrón anterior	Sí	121	45,3%
	No	144	53,9%
	No se sabe	2	,7%
17. Permanece en el lugar del hecho	Sí	155	58,1%
	No	111	41,6%
	No se sabe	1	,4%
18. Ayuda en la extinción	Sí	119	44,6%
	No	146	54,7%
	No se sabe	2	,7%
19. Hubo víctimas	Sí	37	13,9%
	No	228	85,4%
	No se sabe	2	,7%

Tabla 2. Descripción de las variables del autor.

Variable	Categoría	Recuento	% de N =267
1. Franjas de edad	Hasta 34 años	80	30,0%
	De 34 a 46 años	74	27,7%
	De 46 a 60 años	58	21,7%
	Más de 60 años	48	18,0%
	No se sabe	7	2,6%
2. Sexo	Varón	253	94,8%
	Mujer	14	5,2%
	No se sabe	0	,0%
3. Nacionalidad	Español	249	93,3%
	Extranjero	17	6,4%
	No se sabe	1	,4%
4. Estado civil	Casado - pareja	133	49,8%
	Soltero	107	40,1%
	Separado, divorciado, viudo	23	8,6%
	No se sabe	4	1,5%
5. Situación laboral	desempleado	82	30,7%
	empleado	67	25,1%
	autónomo	31	11,6%
	esporádico	21	7,9%
	pensionista, jubilado	63	23,6%
	No se sabe	3	1,1%
6. Sector laboral	agrícola	53	19,9%
	forestal	17	6,4%
	pesca	18	6,7%
	industria	17	6,4%
	administración	5	1,9%
	comercio-hostelería	4	1,5%
	construcción	46	17,2%
	otros servicios	42	15,7%
	variados	17	6,4%
	No se sabe	48	18,0%
7. Tipo de trabajo	manual	168	62,9%
	cualificado	49	18,4%
	No se sabe	50	18,7%
8. Asistencia al trabajo	nunca falta	133	49,8%
	falta poco	44	16,5%
	falta regularmente	11	4,1%
	No se sabe	79	29,6%
9. Adaptación al puesto de trabajo	normal, adaptado	156	58,4%
	regular, rendimiento bajo	11	4,1%
	malo: conflictivo	21	7,9%
	No se sabe	79	29,6%
10. Franjas ingresos	No ingresos	39	14,6%
	Menos de 600 euros mes	51	19,1%
	Entre 600 y 1200 euros mes	137	51,3%
	Más de 1200 euros mes	23	8,6%
	No se sabe	17	6,4%
11. Nivel educativo	analfabeto	50	18,7%
	elemental, FP1	125	46,8%
	EGB, ESO, FP2	73	27,3%
	BUP, bachillerato, FP3	6	2,2%
	Universidad	7	2,6%
	No se sabe	6	2,2%

12. Rendimiento académico	suspendía	29	10,9%
	aprobaba con dificultad	72	27,0%
	aprobaba sin dificultad	72	27,0%
	buenas notas	21	7,9%
	no escolarizado/NS	42	15,7%
	No se sabe	31	11,6%
13. Infancia	normal	212	79,4%
	sufrió algún tipo de trauma	2	,7%
	historia de problemas en la familia	15	5,6%
	No se sabe	38	14,2%
14. Crianza R	Normal	204	76,4%
	Difícil	27	10,1%
	No se sabe	36	13,5%
15. Estilo de vida R	Vive solo	43	16,1%
	Vive con pareja	119	44,6%
	Vive con padres	58	21,7%
	Vive con otros	43	16,1%
	No se sabe	4	1,5%
16. Lugar de residencia actual	Una casa aislada en el campo	13	4,9%
	Una aldea	83	31,1%
	Un pueblo	119	44,6%
	Una ciudad	43	16,1%
	No tiene domicilio fijo	6	2,2%
	No se sabe	3	1,1%
17. Relaciones sociales	no tiene amigos	26	9,7%
	tiene pocos	86	32,2%
	tiene muchos	131	49,1%
	No se sabe	24	9,0%
18. Tiempo libre	estar solo	81	30,3%
	estar con gente	157	58,8%
	No se sabe	29	10,9%
19. Tratamiento psicológico/psiquiátrico	Sí	44	16,5%
	No	196	73,4%
	No se sabe	27	10,1%
20. Otros problemas de salud	Sí	41	15,4%
	No	197	73,8%
	No se sabe	29	10,9%
21. Abuso de sustancias R	Si	76	28,5%
	No	177	66,3%
	No se sabe	14	5,2%
22. Incendio bajo el efecto de sustancias	Sí	44	16,5%
	No	194	72,7%
	No se sabe	29	10,9%
23. Localización del incendio y domicilio	misma localidad	201	75,3%
	otra localidad misma provincia	44	16,5%
	otra localidad otra provincia	20	7,5%
	No se sabe	2	,7%
24. Localización del incendio y trabajo	misma localidad	123	46,1%
	otra localidad misma provincia	37	13,9%
	otra localidad otra provincia	13	4,9%
	No se sabe	94	35,2%
25. Conoce al propietario	es el mismo	54	20,2%
	nada	75	28,1%
	poco	35	13,1%
	mucho	96	36,0%
	No se sabe	7	2,6%

26. Relación con el propietario	familiar	18	6,7%
	empleado	8	3,0%
	compañeros de trabajo	1	,4%
	vecinos	80	30,0%
	amigos	7	2,6%
	enemigos	2	,7%
	relación sentimental	0	,0%
	el mismo	54	20,2%
	no hay relación	86	32,2%
	No se sabe	11	4,1%
27. Actitud durante la detención	asustado, nervioso	82	30,7%
	desafiante	14	5,2%
	presume situación	1	,4%
	tranquilo / normal	166	62,2%
	No se sabe	4	1,5%
28. Asume la responsabilidad	si del fuego, no incendio	31	11,6%
	si fuego e incendio	125	46,8%
	no	106	39,7%
	No se sabe	5	1,9%
29. Medio de transporte	A pie	118	44,2%
	Turismo	83	31,1%
	Todo terreno	27	10,1%
	Otros	28	10,5%
	No se sabe	11	4,1%
30. Medio de ignición	fósforos	7	2,6%
	mechero	192	71,9%
	velas	0	,0%
	bomba incendiaria	0	,0%
	artefacto artesanal	9	3,4%
	pirotécnico	8	3,0%
	cigarros	1	,4%
	vidrio, lupa	0	,0%
	maquinaria, chispas	24	9,0%
	No se sabe	26	9,7%
31. Cómplices	sí	5	1,9%
	no	253	94,8%
	No se sabe	9	3,4%
32. Coautores	sí	20	7,5%
	no	240	89,9%
	No se sabe	7	2,6%
33. Grupo	sí	2	,7%
	no	257	96,3%
	No se sabe	8	3,0%
34. Vigilancia policial	no controlado, vigilado ni investigado	184	68,9%
	control, contactos esporádicos	2	,7%
	sometido a vigilancia policial	19	7,1%
	investigado como supuesto autor	50	18,7%
	No se sabe	12	4,5%
35. Detención anterior por motivo distinto al incendio	sí	61	22,8%
	no	203	76,0%
	No se sabe	3	1,1%
36. Incendio en serie	sí	70	26,2%
	no	194	72,7%
	No se sabe	3	1,1%