Análisis comparativo de diferentes métodos para la obtención de modelos de riesgo humano de incendios forestales.

Vilar del Hoyo, L., Gómez Nieto, I., Martín Isabel, M.P., Martínez Vega, F.J[1].

Resumen

La prevención y planificación son fundamentales en la lucha contra los incendios forestales. Los sistemas integrados de riesgo constituyen una herramienta muy eficaz para la toma de decisiones en el este ámbito. Los avances logrados en los últimos años en el desarrollo de estos sistemas integrados de prevención han sido notables, no obstante, todavía la mayoría de ellos no incorporan, o lo hacen sólo de forma parcial, los factores relacionados con la actividad humana, a pesar de que su papel resulta clave ya que explican más de un 90 por ciento de la ocurrencia de incendios en España.

En la presente comunicación proponemos un análisis comparativo de tres métodos diferentes para la modelización del riesgo de incendio asociado a la actividad humana: Regresión Logística, Árboles de clasificación y Redes Neuronales Artificiales. El objetivo es obtener un modelo predictivo que permita estimar la probabilidad de ocurrencia vinculada a factores humanos, de cara a integrar la información en un modelo que incluya también otros aspectos del riesgo. Las áreas de estudio seleccionadas son la Comunidad de Madrid y la provincia de Huelva. Las variables utilizadas para la elaboración de los modelos se relacionan con la actividad humana e integran los usos del territorio y aspectos socioeconómicos.

Introducción

En España se producen unos 20.000 incendios forestales cada año, lo que supone una media de unas 152.000 ha de superficie quemada (período 1961-2004) (DGB, 2006). La incidencia de este fenómeno en nuestro país se relaciona con las características climatológicas propias de la región mediterránea, pero también con la acción del hombre, ya que, según las estadísticas oficiales el 96,1 % de los incendios que ocurren en España obedecen a causas humanas (DGB, 2006).

Actualmente se está asistiendo, en el entorno europeo, a cambios socioeconómicos, culturales y políticos que han dado lugar a importantes transformaciones económico-productivas y socioculturales en el mundo rural (Moyano, 2006). En España, en los años 60, el desarrollo industrial dio lugar al despoblamiento de las áreas rurales (Pausas, 2004) provocando un abandono del monte y de las actividades tradicionales de gestión del territorio. Ha desaparecido el uso del bosque como fuente de producción y la actividad ganadera en el sotobosque, dando lugar a una acumulación de biomasa combustible disponible para el incendio.

Respecto a las causas de incendio en nuestro país, en el período 1994-2003 el porcentaje más alto corresponde a los incendios intencionados (61,9%). Aunque las motivaciones de estos incendios intencionados se desconocen en más de un 50%, de los que sí se tiene un conocimiento

[1] Instituto de Economía y Geografía, Consejo Superior de Investigaciones Científicas (IEG-CSIC), C/ Pinar, 25, 28006, Madrid, tel.:91 411 10 98, lvilar@ieg.csic.es, israel@ieg.csic.es, mpilar.martin@ieg.csic.es, vega@ieg.csic.es

cierto de su origen destacan las quemas agrícolas sin control (17,4%) y la conversión del matorral en pasto (14,8%). El resto de motivaciones conocidas (pirómanos, modificación de usos del suelo, etc.) no alcanzan en ningún caso el 10% (APAS[2], 2004). Por tanto, resulta evidente, dada la importancia de las consecuencias de los incendios forestales a todos los niveles (ecológico, económico, social), el interés de contar con mecanismos para el establecimiento de acciones permanentes y eficaces de prevención. Con este objetivo se aborda el estudio del riesgo de incendio. De entre los diversos planteamientos conceptuales del riesgo que encontramos en la literatura, quizá los mas completos son aquellos que estructuran el mismo en tres componentes relacionados con el inicio de fuego, la propagación y los daños potenciales que produce (Chuvieco y col, 2004). Este planteamiento es objeto de estudio del proyecto *Firemap* "Análisis Integrado de Incendios Forestales mediante Teledetección y Sistemas de Información Geográfica"[3] (CGL2004-06049-C04-02/CLI), en el que se inscribe este trabajo. . El proyecto aborda diversos aspectos relacionados con la generación un índice de riesgo integrado. En esta comunicación se presentan los resultados obtenidos del análisis y modelización de los factores socio-económicos relacionados con el riesgo humano de incendios. Se utilizan técnicas de Regresión Logística, Árboles de Decisión y Redes neuronales para obtener un estimación de la ocurrencia de incendios a nivel de cuadrícula (1x1 km) en la Comunidad de Madrid y en la provincia de Huelva, con el objetivo de valorar qué técnica da lugar a un modelo de mayor capacidad predictiva y explicativa.

Material y métodos

Áreas de estudio

Las áreas de estudio para la obtención de mapas de predicción de riesgo humano de incendio forestal son la Comunidad de Madrid y la provincia de Huelva en España (Figura 1). El período de estudio comprende los años 1990 a 2004. Se ha elegido este período para asegurar la consistencia de los datos estadísticos de incendios utilizados y para garantizar la robustez de los análisis estadísticos efectuados.

[2] Asociación para la Promoción de Actividades Socioculturales
[3] http://www.geogra.uah.es/firemap/

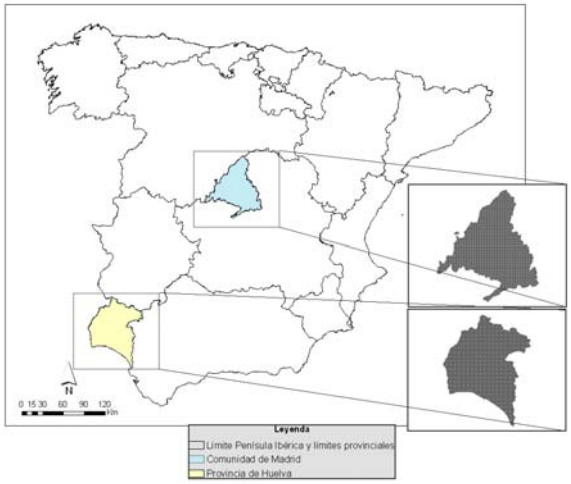
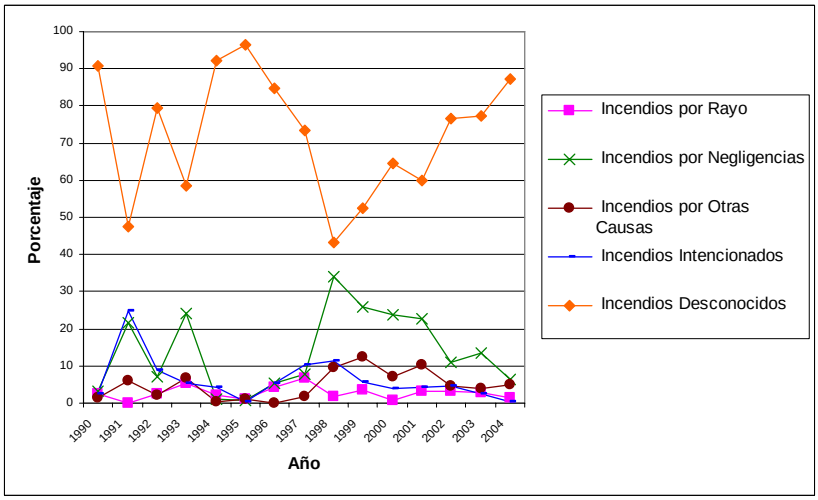


Figura 1—Zonas de estudio. Comunidad de Madrid y provincia de Huelva (España)

La Comunidad de Madrid es una de las regiones más pobladas de España, con unos 6 millones de habitantes (a 1 de enero de 2005, según el padrón municipal de habitantes del INE[4]), lo que supone una tasa media de densidad de unos 748 habitantes/km^2. Destaca su alto grado de urbanización (8,6% de su superficie es dedicada a suelo urbano en 2002,- Instituto de Estadística de la Comunidad de Madrid-), cobrando especial importancia el contacto entre las zonas urbanas y forestales. Posee una alta densidad de vías de comunicación, y su actividad económica se basa en el sector terciario. Las zonas forestales se distribuyen fundamentalmente del Noroeste a Suroeste de la comunidad y presentan un importante uso recreativo.

En la C. de Madrid la ocurrencia de incendios es relativamente baja si la comparamos con otras regiones española, sin embargo, la alta densidad de población y uso recreativo de sus masas forestales la convierten en un área de especial interés para este estudio. En la Figura 2 se observa la distribución y evolución temporal de las principales causas de incendios en la región en los últimos 15 años. Destaca el alto porcentaje de causas desconocidas y la notable proporción de incendios por negligencia.

[4] Población total española a 1 de enero de 2005: 44.108.530 (Padrón municipal de habitantes, INE)

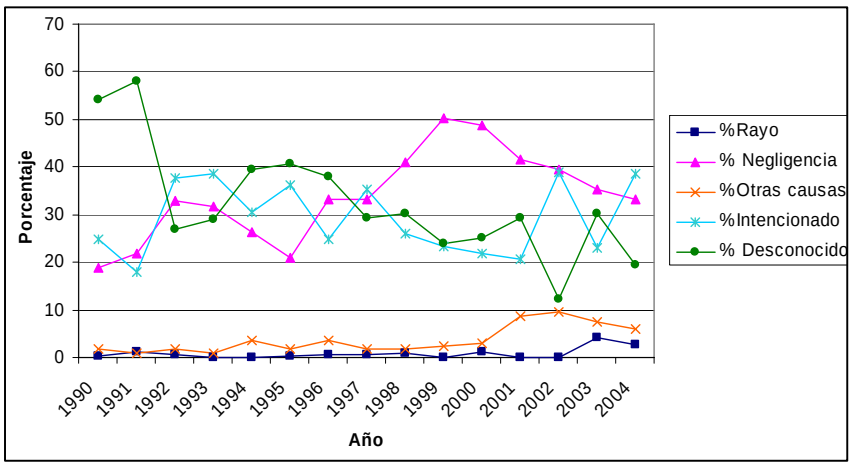


Causa	Nº	pct
Rayo	86	2,36
Negligencia	461	12,65
Otras causas	175	4,80
Intencionados	177	4,86
Desconocidos	2745	75,33
Total	3644	100

Figura 2. Tendencias de Incendios forestales según tipo de causa. C. de Madrid. Período 1990-2004

La provincia de Huelva cuenta con una población total de 483.792 habitantes en 2005 (IEA, 2007, a partir de la revisión del Padrón de Habitantes del INE 2005). La densidad de habitantes por km^2 es de 47,67 (IEA, 2007), y la mayor concentración de población se localiza en la zona costera. La actividad económica se basa principalmente en el sector servicios (IEA, 2007) aunque también es importante la actividad agrícola (el porcentaje de población activa ocupada en el sector agrícola en el mismo año era del 15,9 por ciento). Las zonas forestales se localizan principalmente en el sector centro-norte de la provincia, y destacan los espacios protegidos de la Sierra de Aracena y Picos de Aroche al Norte y un sector del Parque Nacional y Natural de Doñana al Sureste.

La ocurrencia de incendios en la provincia de Huelva, en el ámbito de la C. de Andalucía, se sitúa en primer lugar en número de siniestros en el año 2004, un 22 por ciento del total de incendios en ese año se produjeron en dicha provincia (DGB). Las causas de incendio forestal en el período 1990-2004 se recogen en la Figura 3. Destacan de nuevo los incendios debidos a negligencias unido, en este caso a un alto porcentaje de incendios intencionados.



Causa	Nº	pct
Rayo	32	0,9%
Negligencia	1123	32,0%
Otras causas	118	3,4%
Intencionados	1006	28,6%
Desconocidos	1235	35,1%
Total	3514	100

Figura 3. Tendencias de Incendios forestales según tipo de causa. Provincia de Huelva. Período 1990-2004

Variables independientes

Para la generación de variables independientes que dan lugar al modelo de riesgo humano se identificaron, en primer lugar, los factores de riesgo asociados a la actividad humana. A continuación se definieron las variables que mejor los representaban, permitiendo su cuantificación y espacialización. A partir de diversas fuentes bibliográficas (Leone y col 2003, Martínez 2004, Martínez y col 2004, Pew y col 2001, Vega-García y col 1995) y los estudios realizados sobre el tema en diversos proyectos a nivel local y regional (*Firerisk*, 2003; *Spread*, 2003; *Megafires*, 2002) se establecieron 6 grupos de factores de riesgo vinculados a la actividad humana: accidentes y/o negligencias, transformaciones socioeconómicas, actividades tradicionales en áreas rurales, conflictos y factores de disuasión de la ignición. Para cada grupo de factores se definieron una serie de variables (estadísticas y cartográficas) que permitían su representación espacial en cuadrículas de 1x1 km... La espacialización de las variables se ha realizado siguiendo dos metodologías, una para las cartográficas y otra para las estadísticas. En el caso de las variables cartográficas van a estar referidas a la superficie de la cuadrícula UTM como un cociente entre el valor del área de la variable en cuestión y el área de la cuadrícula UTM. En el caso de las variables estadísticas el proceso fue distinto, pues todas ellas estaban referidas a la unidad espacial municipio. Para asignar un valor a cada cuadrícula de 1 km^2 se intersectaron los polígonos de los municipios con la cuadrícula 1x1 km, asignándole a cada una de las cuadrículas incluidas en cada municipio el valor de la variable estadística en cuestión para ese municipio. En las cuadrículas UTM en las que coincidían varios municipios se asignó una media ponderada por la superficie ocupada por cada municipio en la cuadrícula. En la tabla 1 se recogen las variables independientes generadas para los distintos grupos de factores.

Tipo	Factor	Variable
CARTOGRÁFICAS	**Incendios por accidente o negligencia**	Áreas de influencia (buffer) de vías sin pistas forestales[5] (CARRET, CARRET_FOR)
		Índice de IMD por segmento de carretera (Longitud vía*IMD vía*factor de ponderación) (INDICE_IMD, INDICE_IMD_FOR)[6]
		Áreas de influencia (buffer) de vías de ferrocarril (B_FFCC, B_FFCC_FOR)
		Áreas de influencia (buffer) de pistas forestales (B_PISTAS, B_PISTAS_FOR)
		Áreas de influencia (buffer) de líneas eléctricas (B_LLEE, B_LLEE_FOR)
		Campos de tiro y canteras (P_A_TIROCANTERAS)
	Transformaciones socioeconómicas	Área de influencia (buffer) de áreas recreativas ponderadas por presencia de barbacoa (AREA_RECRE)
		Potencial demográfico (POT_DEM)
		Índice de cambio en Superficie Forestal (ICC)

[5] Las variables de vías sin pistas, ferrocarril, líneas eléctricas y pistas se obtienen también sólo en zona forestal, utilizando como fuente de referencia el Mapa Forestal

[6] Caso de Madrid

		Interfaz Urbano-Forestal (I_UFOR)
		Áreas de influencia (buffer) de vertederos (VERTEDEROS)
		Interfaz Cultivo-Forestal (I_CULT_FOR)
		Interfaz Pasto-Forestal (I_PASTO_FOR)
	Conflictos que pueden desencadenar el inicio intencionado de incendios	Espacios naturales protegidos (ENP)
		ZEPAS (ZEPAS)
		Montes de Utilidad Pública y Preservados (MUP_PRESER)
		Montes de Consorciados (MONTES_CONSOR)
ESTADÍSTICAS	**Transformaciones socioeconómicas**	Variación de la población entre 1970-2004 (VAR_POB)
		Infraestructuras hoteleras totales (PLAZAS_HOTEL)
		Variación de la población agraria (VAR_POB_AGRA)
	Actividades tradicionales en áreas rurales	Porcentaje de Jefes de explotaciones agrarias mayores de 55 años (JEFES55)
		Carga ganadera (cabezas de ganado ovino y caprino en superficie de pastos y matorral) (CARGA_GANADERA)
		Densidad de maquinaria agrícola (MAQUINA)
	Conflictos que pueden desencadenar el inicio intencionado de incendios	Renta per capita (RENTA)
		Tasa de paro 2001 (TASA_PARO)

Tabla 1—*Variables independientes modelo de riesgo humano.*

Variable dependiente

La variable dependiente empleada es la ocurrencia de incendios por causa humana[7] en el período 1990-2004, obtenida a partir de la información contenida en los partes de incendio de la Dirección General para la Biodiversidad del Ministerio de Medio Ambiente donde la localización espacial de los incendios se refiere a una cuadrícula 10x10 km. No se cuenta, por tanto, con información precisa que nos permita conocer con exactitud la localización de los puntos de ignición. Teniendo en cuenta que la unidad de análisis elegida en este estudio es la cuadrícula de 1x1 km, resulta necesario aplicar algún procedimiento que permita especializar la ocurrencia a la resolución elegida reduciendo en lo posible la incertidumbre en la localización espacial de los incendio. Para ello, comenzamos combinando la información que los partes ofrecen sobre la

[7] Dado el elevado número de incendios de causas desconocidas se decidió incluir en el análisis una parte de los incendios desconocidos proporcional al número de incendios de causa humana en cada región.

localización de los incendios a nivel municipal con la localización por cuadrículas 10x10 km. De esta forma se consigue acotar la localización de los incendios en polígonos de superficie inferior a la de las cuadrículas de referencia. Para afinar aún más esta localización se cruzaron los polígonos resultantes con el mapa forestal, eliminando las zonas sin superficie forestal. En esos polígonos finales se generan mediante el *script* de *ArcView 3.2 Random Point Generator v. 1.3*[8], tantos puntos aleatorios como incendios de causa humana ocurridos en el período de estudio. A partir de esta distribución aleatoria de "puntos de ignición", y con objeto de reducir la imprecisión en la localización de los puntos (Amatulli y col, 2005) se transformaron las observaciones puntuales en superficies continuas. Para ello se ha utilizado la técnica de interpolación de estimación de densidad de kernel adaptativo, propuesta por de la Riva y col (2004). Esta técnica consiste en posicionar una probabilidad de densidad sobre cada punto y estimar la densidad en cada intersección de una malla superpuesta al conjunto de puntos (Leone y col, 2003 citando a Seaman y Powell, 1996; Levine, 2002):

$$f(x) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left\{\frac{(x - X_i)}{h}\right\}$$

Siendo n el número de puntos, h el parámetro de suavizado ó *bandwidth*, x el vector de coordenadas que define la localización donde se estima la función y X_i el vector de coordenadas que define cada observación i. De entre las funciones diferentes que existen (distribución normal, función cuártica, triangular), se emplea la normal, que es la más utilizada (Levine, 2004).

En cuanto al procedimiento para fijar el kernel, este puede ser fijo (*bandwidth* constante) ó adaptativo (*bandwidth* varía dependiendo de la concentración de puntos) (Leone y col 2003 citando a Worton, 1989). Este último ofrece una mayor flexibilidad en la estimación de densidad, dado que el *bandwidth* se calcula como una función inversa a la concentración de puntos. En áreas con alta concentración será menor, mientras que con poca presencia de puntos será mayor (Amatulli y col, 2005). Debido a que los incendios no se distribuyen de manera regular, se emplea el modo adaptativo. Se establece un tamaño de intervalo de *bandwidth* de 5 puntos utilizando para llevar acabo la interpolación *Crimestat© 3.0 (*Levine, *2004).* La elección de este valor para llevar a cabo la interpolación resulta de la minimización del *goodness-of-fit criteria* propuesto por Breiman en 1977. En este ajuste se ensayan distintos órdenes de vecino próximo para dar con el que minimiza la curva de ajuste.

Las variables dependientes finalmente obtenidas par las dos zonas de estudio se muestran en la Figura 4:

[8] *Random Point Generator v. 1.3. Autor:* Jeff Jenness. Wildlife Biologist, GIS Analyst. Jenness Enterprises. jeffj@jennessent.com

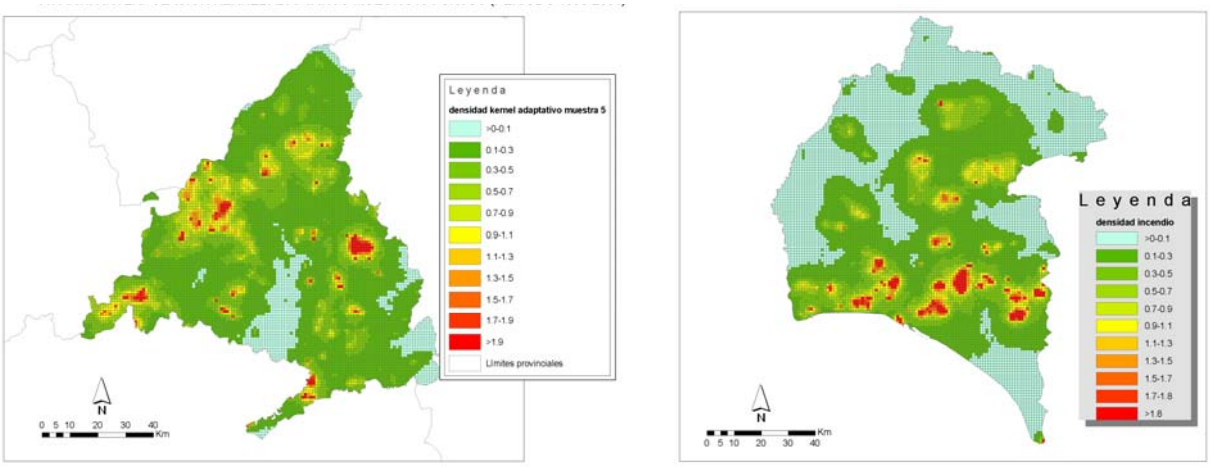


Figura 4. Variable dependiente obtenida a partir del método de interpolación Kernel adaptativo muestra 5 puntos. C. de Madrid, provincia de Huelva

Desarrollo de los modelos

Los modelos de riesgo humano han sido generados mediante las técnicas de regresión logística, árboles de decisión y redes neuronales en ambas zonas de estudio.

Regresión Logística

El método de regresión logística ha dado buenos resultados en anteriores análisis de ocurrencia de riesgo humano de incendios forestales a escalas tanto regionales como locales, permitiendo establecer modelos de tipo predictivo y a la vez explicativos, al conocer cuáles de las variables son las de mayor importancia en el fenómeno (Carvacho, 1998).

El objetivo de la regresión logística es estimar la probabilidad de ocurrencia de la variable dependiente dicotómica (en nuestro caso, alta o baja incidencia de incendio) a partir de las variables independientes, es decir, obtener la probabilidad de que cada individuo pertenezca a cada uno de los grupos que define la variable dependiente (González, 2004). De igual forma, se comprueba la relación entre la variable dependiente y las independientes seleccionadas en el modelo.

El modelo de regresión logística se define:

$$P_i = \frac{1}{1+e^{-z}}$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_\rho X_\rho$$

Donde P_i es la probabilidad de ocurrencia de incendio, z la combinación de variables independientes con sus coeficientes de regresión (β), X el valor de cada variable independiente y e la base del logaritmo natural (Pew y Larsen, 2001 citando a Afifi y Clarck. 1990; McGrew y Monroe, 1993).

De entre las posibilidades de modelos de regresión logística binaria se aplica el modelo *logit*:

$$\log\left(\frac{p}{1-p}\right) = x^T \beta$$

Siendo x^T el vector de las variables explicativas y β el vector de los parámetros (González, 2004).

Antes de construir el modelo es conveniente eliminar variables innecesarias o redundantes, que no aporten información. Cuando las variables independientes tienen mucha relación entre sí, el modelo no puede distinguir que parte de la variable dependiente es explicada por una u otra variable. Esto se conoce como multicolinealidad (Villagarcía, 2004). Para estudiar la incidencia de este fenómeno en los datos se han aplicado diagnósticos de colinealidad propios de la técnica de regresión multivariante Mediante coeficientes de correlación no paramétricos de *Spearman* se identificaron y eliminaron del modelo las variables que presentaban correlación superior a 0,9 en el caso de la Comunidad de Madrid y 0,7 en el caso de la provincia de Huelva Posteriormente, se estudió el fenómeno de multicolinealidad mediante diagnósticos propios de la técnica de regresión multivariante, eliminando del análisis aquellas variables en las que estos diagnósticos señalaban problemas de colinealidad. Finalmente, se aplicaron tests no paramétricos de estadística comparativa que proporcionan una medida de la diferencia entre dos conjuntos de datos (Martínez y col, 2005), la prueba de la U-Mann-Whitney y la prueba de Kruskal-Wallis. El objetivo era comprobar si existía o no diferencia significativa entre los valores de las variables seleccionadas correspondientes a dos muestras de cuadrículas, unas con alta ocurrencia y otras con baja ocurrencia de incendios.

Como hemos indicado, el modelo de regresión logística requiere una variable dependiente dicotómica así pues, fue necesario transformar la variable número de incendios de causa humana de continua a dicotómica. Esto se hizo dividiendo la variable ordenada en 3 grupos con el mismo número de casos (grupo 1, baja incidencia, y 3, alta incidencia y grupo 2 de incidencia intermedia). A los casos incluidos en el primer grupo se les da valor 0 y a los del grupo 3 valor 1. Se eliminan del análisis los valores intermedios que quedarían en el grupo 2.

Al aplicar la regresión logística se ha empleado el método por *pasos hacia delante de Wald*, con el valor 0,5 como punto de corte para la clasificación. El modelo se obtuvo empleando una muestra aleatoria del 60% de los casos, utilizando el 40% restante para validar la calidad de las estimaciones. Una vez validado el modelo se aplica a la totalidad de los casos, para posteriormente obtener la probabilidad de ocurrencia de incendio en el total del área de estudio. Para la obtención de la variación real de la variable dependiente en relación a cada independiente se aplica regresión logística con las variables del modelo normalizadas.

Árboles de Decisión

La técnica de Árboles de Decisión es una técnica de minería de datos ó *Data Mining*. Éstas se definen como un conjunto de técnicas que permiten la generación de modelos a partir de datos históricos. Estos modelos son de tipo empírico, capaces de extraer patrones y tendencias de una gran cantidad de datos (Zhang, B y col, 2005). El resultado del análisis es una estructura llamada árbol, con ramas y hojas que contienen las reglas para predecir nuevos casos. Esta técnica presenta una serie de ventajas respecto a la estadística tradicional, ya que funciona cuando las variables independientes cualitativas o cuantitativas presentan problemas, permite clasificar

individuos con información incompleta y su interpretación es sencilla. Sin embargo, no es una técnica robusta frente a valores atípicos Además, no puede expresar relaciones lineales ni producir un resultado en forma de variable continua, y no tiene una única solución (Zhang, B y col, 2005 citando a Iverson y Prasad, 1998; Scheffer, 2002). En este estudio proponemos el empleo del algoritmo C&RT, de *SPSS Answer Tree*. Este algoritmo identifica subconjuntos homogéneos en los datos. Crea árboles binarios y la variable criterio (dependiente) puede ser nominal, ordinal o continua (González, C., 2004). Se fijan nodos parentales de 100 casos y nodos hijos de 50 casos. A partir del primer nodo se desarrollan como máximo 5 niveles. Al igual que en el modelo de regresión logística, se utiliza el 60% de los casos para el entrenamiento y el 40% para validar la calidad del modelo (muestra de comprobación). Como variable dependiente se emplea la utilizada en Regresión Logística, la variable dicotómica de alta y de baja incidencia de incendios.

Redes neuronales

Las redes neuronales emulan el sistema biológico de un modo simplificado (Bischof y col., 1992). Están formadas por numerosos elementos procesadores de información (PEs, los equivalentes artificiales de las neuronas biológicas), interconectados entre sí; aunque capaces sólo de realizar operaciones relativamente simples. Los PEs se estructuran en niveles de capas (Vega, 1996). Existe un nivel de entrada que introduce los datos a la red; un nivel de salida, que proporciona la respuesta a los datos de entrada; y uno o más niveles ocultos que procesan los datos. Aprenden la relación entre los datos de entrada y de salida, por lo cual, todo lo que se necesita para entrenar una red neuronal artificial (RNA), es un conjunto de datos que contengan la relación entrada-salida (Carvacho, 2002).

Esta estructura otorga a una RNA gran capacidad para procesar datos y la habilidad para realizar procesos inteligentes como: aprender a partir de ejemplos, generalizar el conocimiento adquirido a nuevos casos y reconocer tendencias y patrones en los datos. Los componentes que definen un modelo de red neuronal son: tipo de PEs o neuronas, los pesos de sus conexiones con otras neuronas, la regla de aprendizaje, el número de niveles y neuronas por nivel, patrones de conexión entre niveles y flujo de información.

La neurona artificial, al igual que la biológica, se define por encontrarse en todo momento en un estado de activación que se expresa mediante un valor numérico. Este valor numérico responde a la siguiente fórmula:

$$a=S_{i=1,n}w_i x_i$$

Siendo x_i es el valor de activación proveniente de cada neurona de la capa anterior, y w_i es el peso asignado a dicho valor.

Una función de transferencia o salida transforma este valor en una señal de salida que viaja a través de las conexiones a otras neuronas de los niveles subsiguientes, eliminando la linealidad de la red y acotando los valores en un intervalo determinado (Carvacho, 2002). La función más extendida y la que utiliza el programa PCI *Geomatics* v10.0 utilizado en este estudio es la *sigmoide*, que acota los valores de salida entre 0 y 1 y tiene la siguiente forma:

$$x=1/(1+e)^{-a/p}$$

Siendo a es el valor de activación de la neurona, y p es un modificador de la función *sigmoide*, habitualmente 1.

Las señales enviadas a una neurona desde varias otras se modifican de acuerdo al peso de la conexión, w_i, y se combinan al llegar a la de destino de acuerdo a una regla de propagación que produce la entrada total (Vega, 1996). Las redes *backpropagation* de neuronas con función de transferencia se han convertido en la elección más frecuente para los diferentes modelos de redes (Bichof y col., 1992), y es el caso también de la utilizada en nuestro análisis. El flujo de datos procede del nivel de entrada y se difunde al/os ocultos y al de salida. La regla de aprendizaje de este tipo de redes es la *regla delta generalizada*, derivada de la regla *perceptron*, que responde a la siguiente fórmula:

$$Dw_i=h(D-Y)$$

Siendo w_i es el peso otorgado a una neurona, h es la tasa de aprendizaje, que controla la velocidad de aprendizaje (0,1 es nuestro caso); D es el resultado esperado e Y es el obtenido en cada iteración de la red. Esta regla al aplicarse a cada conexión entre las neuronas de la red pasa a denominarse *regla delta generalizada* (Carvacho, 2002).

El aprendizaje se produce en la etapa de entrenamiento y los pesos permanecerán inalterables posteriormente, durante la explotación de la red, es decir, cuando se aplica a otro conjunto de datos diferentes para predecir nuevos resultados.

Para generar un modelo de riesgo humano de incendios forestales el método utilizado se articula en dos fases sucesivas. En la primera, se diseña y entrena una red con capacidad de predicción de potenciales puntos de ocurrencia o de no ocurrencia, así como, un método de validación de los resultados de este proceso. En la segunda, se lleva a cabo un análisis de sensibilidad para establecer el grado de importancia de cada una de las variables independientes implicadas en el análisis.

a) Diseño, entrenamiento y validación de la RNA.

La cuestión esencial para el uso de redes neuronales en este modelo es definir la arquitectura de la red, es decir, el número y características de las unidades de entrada y de salida así como, el número de capas ocultas y sus unidades. En este sentido, no existe ningún tipo de consenso, excepto la constatación de que no hay una fórmula única para el diseño de una red (Hilera y Martínez, 1995). Por tanto se tratará de ensayar con diferentes arquitecturas hasta encontrar aquélla que mejores resultados arroje. En cualquier caso, la experiencia ha ido seleccionando una serie de estructuras orientadoras, que nosotros tomaremos como punto de partida y referencia (Klimasauskas, 1991c):

$$H=(I+O)/2$$
$$H=I*O$$
$$H=(I+O)^{1/2}$$
$$H=(I+O)^{2}$$

Donde, I es el número de unidades de entrada y O el de salida.

Como unidades de entrada se utiliza el conjunto de variables independientes, eliminando aquéllas que muestran un alto grado de correlación con otras sí consideradas, según los procedimientos descritos en regresión logística.

En el caso de la Comunidad de Madrid, tras definir diferentes arquitecturas y analizar los resultados con los datos de validación se optó por una arquitectura 22-4-2. La capa de entrada incluyó un total de 22 unidades de entrada. En el nivel oculto se consignaron 4 unidades en una sola capa. En el nivel de salida se definieron dos unidades, correspondientes a celdas de de alta y

baja ocurrencia de incendio (variable dependiente dicotómica empleada en las dos técnicas anteriores). La red se entrenó con 20.000 iteraciones.

La muestra total de puntos de alta ocurrencia de incendio para el caso de Madrid es de 2689; la de baja ocurrencia 2692. Para seleccionar las celdas de una y otra categoría a partir de los datos contenidos en la "variable dependiente" se usaron las siguientes condiciones:

- si "variable dependiente">0.334169 entonces "alta ocurrencia de incendio"
- si "variable dependiente"<0.173281 entonces "baja ocurrencia de incendio"

Para el entrenamiento de la red se seleccionaron dos grupos de píxeles de esta muestra, uno para el propio entrenamiento de la red, de un 30% del total, (compuesta por 807 puntos de "alta densidad" y 793 de "baja densidad") y otro para comprobar la solidez del entrenamiento, evitando un *sobre* o *infra* entrenamiento, también de un 30 % de la muestra (807 puntos de "alta densidad" y 793 de "baja densidad"). Una vez finalizado el proceso de entrenamiento se procedió a aplicar la red sobre una muestra de validación de 1075 puntos de "alta densidad" y 1056 de "baja densidad" (equivalente a un 40 % del total de la muestra), midiéndose los errores de comisión y omisión mediante una tabulación cruzada entre los datos de validación y los estimados por la red.

En la provincia de Huelva se procedió de la misma manera, empleando una capa de entrada compuesta por 21 unidades, correspondientes a las variables independientes después de eliminarse las variables altamente correlacionadas con otras.

La arquitectura final respondía al esquema 21-12-2. Por tanto, se trabajó con una capa oculta de 12 neuronas y una de salida con dos ("alta densidad"-"baja densidad"). La red se entrenó con 20.000 iteraciones.

Para la obtención de la muestra "alta/baja ocurrencia" de la variable dependiente se procedió de la misma forma que en la C. de Madrid, utilizando las siguientes condiciones:

- si "variable dependiente">0.222222 entonces "alta ocurrencia de incendio"
- si "variable dependiente"<0.083501 entonces "baja ocurrencia de incendio"

Al igual que en el caso de la Comunidad de Madrid la muestra total (de 3465 puntos de "alta ocurrencia" y 3249 de "baja ocurrencia") se dividió en tres grupos: uno para el entrenamiento de la red (1039 de "alta ocurrencia" y 974 de "baja ocurrencia"); un segundo para comprobar el entrenamiento de la red (1039 de "alta ocurrencia" y 974 de "baja ocurrencia"); y un tercero para validar el resultado (1387 de "alta ocurrencia" y 1301 de "baja ocurrencia"); nuevamente representaban un 30, 30 y 40 % de la muestra, respectivamente.

b) Análisis de sensibilidad para el establecimiento de la importancia de las variables independientes.

Si bien las redes neuronales artificiales no están pensadas para determinar un grupo de variables significativas, a diferencia de los modelos de regresión, es posible estimar qué variables han tenido mayor importancia a la hora de entrenar la red a través de un análisis de sensibilidad. Dicho análisis consiste en evaluar la variación de la medida del error medio cuadrático de la red diseñada cada vez que se sustituyen todos los valores de una variable por 0 y se vuelve a entrenar dicha RNA (Carvacho, 2002). Este proceso se repite con todas las variables (las 21 en el caso de la Comunidad de Madrid y las 22 de la provincia de Huelva), una a una. De este modo, en función de la entidad de dicha variación se podrá establecer la importancia que tuvo cada variable de entrada en el entrenamiento de la red. Así, si después de cambiar los valores de una variable por 0 el valor del RMS se aleja mucho del arrojado inicialmente durante el entrenamiento de la

red significará que esa variable pesó mucho en el proceso; mientras que si la variación es pequeña significará todo lo contrario.

Resultados

A continuación se exponen los resultados obtenidos a partir de las técnicas anteriormente mencionadas en las dos áreas de estudio propuestas.

Comunidad de Madrid

Regresión Logística

Los resultados obtenidos tras llevar a cabo las correlaciones no paramétricas de Spearman señalan que no han de incluirse en el análisis las variables buffer *de carreteras, pistas* y *máquina* por su alta correlación con otras variables. A partir de test no paramétricos de estadística comparativa se observa que las variables *buffer líneas de ferrocarril, buffer líneas eléctricas, campos de tiro-canteras* y *montes consorciados* no presentan diferencias significativas al 95 por ciento de confianza (p-valor mayor de 0,05) para dos muestras independientes del primer y cuarto cuartil (resultados del test de la U-Mann-Whitney) y que la variable *buffer líneas eléctricas* no es significativa en la comparación de las 4 muestras independientes, al 95 por ciento de confianza (resultados de la prueba de Kruskal-Wallis). Por estos motivos las variables señaladas se excluirían del análisis posterior. Los diagnósticos de colinealidad propios de la técnica de regresión múltiple muestran que la variable *renta* presenta problemas de colinealidad, por lo que de igual modo se excluiría del análisis. Por tanto, los análisis previos en la Comunidad de Madrid para estudiar el efecto de la colinealidad y de la relación entre variables indican que las variables *buffer carreteras, buffer carreteras en zona forestal, buffer pistas, maquinaria agrícola, renta, buffer líneas de ferrocarril, buffer líneas eléctricas, campos de tiro y canteras, montes consorciados* y *renta* presentan problemas, por lo que no van a ser incluidas en el análisis.

Mediante la técnica de regresión logística binaria con las variables excluidas por colinealidad y con la variable dependiente obtenida a partir de interpolación mediante kernel adaptativo (muestra de 5 puntos) se obtuvieron17 modelos, de ellos elegimos el modelo 7º. Los porcentajes globales de acierto de clasificación de la muestra de elaboración del modelo (60 por ciento) y de validación del mismo (40 por ciento) son 71,6 y 70,3 por ciento respectivamente. Los parámetros del modelo seleccionado se recogen en la tabla 2, siendo su ecuación la que se muestra a continuación:

Z= -1,012+1,582* Buffer_Pistas_Forestal+1,952* ENP+44,196* Interfaz_Urbano_Forestal-0,011* Variación_Pobalción_Agraria-0,018* Jefes mayores 55 años+0,175* Tasa_Paro -0,0003184* Hotel

		B	E.T.	Wald	Gl	Sig.	Exp(B)	I.C. 95,0 pct para EXP(B)	
								Inferior	Superior
Paso 7	B_pistas_for	1,582	,274	33,427	1	,000	4,864	2,845	8,315
	Enp	1,952	,146	177,944	1	,000	7,043	5,287	9,382

	I_urb_for	44,196	4,028	120,389	1	,000	1562817101497831 0000,000	5825112017197180, 000	4192876094954938 0000000,000
	Var_pob_agra	-,011	,001	99,847	1	,000	,989	,987	,991
	Jefes	-,018	,002	94,955	1	,000	,982	,979	,986
	Tasa_paro	,175	,017	109,451	1	,000	1,192	1,153	1,232
	Hotel	-0,0003	,000	44,244	1	,000	1,000	1,000	1,000
	Constante	-1,012	,217	21,671	1	,000	,364		

Tabla 2—*Resultados del modelo 7 obtenido por Regresión Logística.*

Las siete variables seleccionadas por el modelo son significativas al 95 por ciento de confianza (significatividad menor de 0,05), y es la variable *interfaz urbano-forestal* la que mayor peso tiene en el modelo (coeficiente B de 44,196) a priori. Las siguientes variables en importancia son *espacios naturales protegidos* (ENP) y *buffer de pistas en zona forestal*. La variable menos significativa es *hotel*, porque en el intervalo de confianza al 95 por ciento contiene el valor 1.

Al aplicar la ecuación del modelo elegido al 100 por ciento de la muestra se obtiene un 70,6 por ciento correcto de clasificación global de la misma, estando la baja incidencia correctamente clasificada en un 75,4 por ciento y la alta incidencia en un 65,7 por ciento.

Una vez normalizadas las variables del modelo elegido los resultados de la regresión arrojan las variaciones de la variable dependiente respecto de cada independiente recogidas en la tabla 3:

Paso 7	dx/dy
z_b_pistas	0.0642239
z_enp	0.1559056
z_i_urb_for	0.1902964
z_var_pob_agra	-0.1051724
z_jefes	-0.1052761
z_tasa_paro	0.113988
z_hotel	-0.2080815

Tabla 3—*Efectos marginales del modelo 7. Variación de la variable dependiente x con cada variable independiente y (dx/dy).*

La variable *interfaz urbano-forestal* es la que más influye en la variación de la variable dependiente. Como muestra la tabla 3, si se aumenta en una unidad la variable interfaz urbano-forestal, la variable dependiente aumenta 0,19 en desviación típica. Le sigue la variable *ENP* (0,15).

A continuación se muestra el mapa de los aciertos y errores para la muestra de comprobación y validación del modelo así como el mapa de probabilidad estimada (Figura 5):

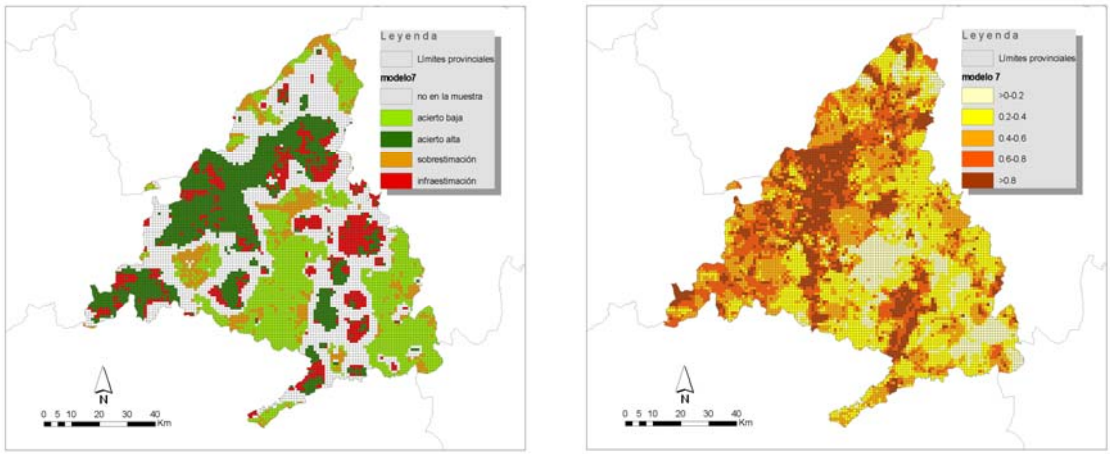


Figura 5—Mapas de aciertos y errores y de probabilidad estimada de riesgo humano (modelo 7)

En el mapa de acierto y error se observan zonas de infraestimación en el Norte, Noreste y Sureste (zonas de la Sierra de Madrid, Alcalá de Henares y Aranjuez), mientras que los errores de sobrestimación se comenten en la zona centro y Suroeste en mayor medida. El modelo predice acertadamente la alta incidencia de Noroeste a Suroeste y en las zonas centro-Sur y Este del área de estudio. El modelo estima valores más altos de probabilidad de ocurrencia en la zona Oeste del área de estudio (Sierra de Madrid), que se corresponde con la zona de mayor superficie forestal. Otra zona con alta probabilidad estimada de incendio forestal es la zona del Sureste, la que coincide con un área protegida, el Parque Regional del Sureste. Las zonas de probabilidad de incendio más bajas son el centro y Este en líneas generales.

Árboles de Decisión

A continuación se muestran los resultados obtenidos a partir de la técnica de Árboles de Decisión en la Comunidad de Madrid.

El árbol de clasificación de la figura 6 señala a las variables *interfaz urbano-forestal*, *ZEPA*, *ENP*, *carga ganadera*, *buffer pistas en zona forestal* y *potencial demográfico* como responsables del establecimiento de grupos de alta y baja incidencia de incendios forestales por causa humana. El porcentaje global correcto del modelo obtenido es de un 75,5 por ciento, clasificando correctamente tanto la baja incidencia de incendio como la alta en un 75,5 por ciento.

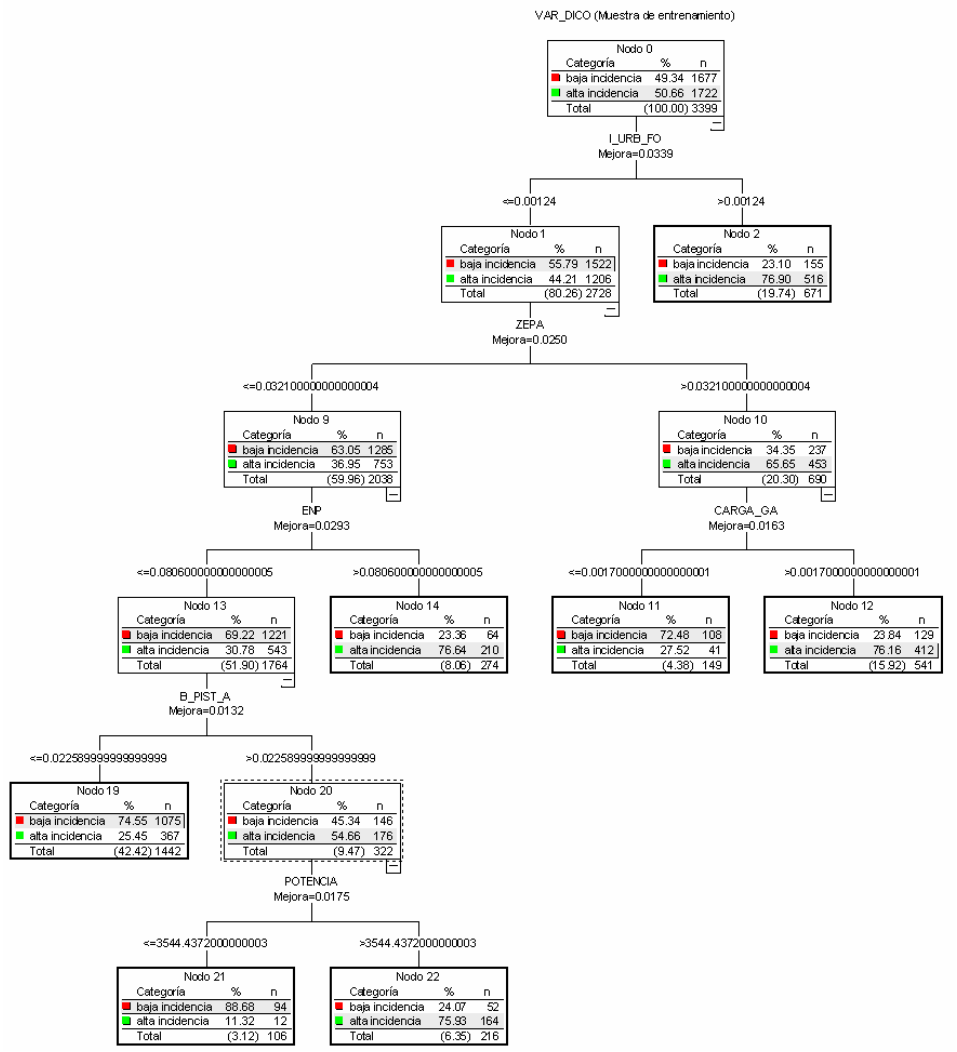


Figura 6— Árbol de decisión en la Comunidad de Madrid

La variable *interfaz urbano-forestal* es la que determina la primera división del árbol, con una mejora del 3 por ciento. Si es mayor de 0,001, las cuadrículas que cumplan esta condición quedarán clasificadas en un 76,9 por ciento como celdas de alta incidencia de incendio. Si es menor ó igual a este valor señalado, es la variable *ZEPA* la que determina el siguiente nivel, con una mejora del 2 por ciento. Si es mayor de 0,03, es la variable *carga ganadera* la que clasifica en un 76,2 por ciento las celdas con alta incidencia si se cumplen las reglas de ramas y niveles anteriores. Si la variable *ZEPA* es menor ó igual a 0,03, la variable *ENP* clasifica en alta incidencia (si es mayor de 0,08) en un 76,6 por ciento y con una mejora del 3 por ciento. Si *ENP* es menor ó igual a 0,08 y *buffer de pistas en zona forestal* menor o igual a 0,02, las celdas quedarán clasificadas en un 74,5 por ciento de baja incidencia cumpliendo todas las condiciones anteriores. Finalmente, el último nivel desarrollado del árbol viene determinado por la variable *potencial demográfico*, si la variable *pistas* es mayor de 0,02 y el potencial mayor de 3544,4, las celdas serán de alta incidencia en un 75,9 por ciento, mientras que serán clasificadas de baja incidencia (valor de potencial menor ó igual a 3544,4) en un 88,6 por ciento.

El método de árboles de decisión clasifica correctamente el 100 por cien de la muestra en un 75,5 por ciento, estando la baja y la alta incidencia bien clasificadas en un 75,5 por ciento.

La figura 7 muestra los aciertos y errores del área de estudio así como la probabilidad estimada a partir del árbol desarrollado:

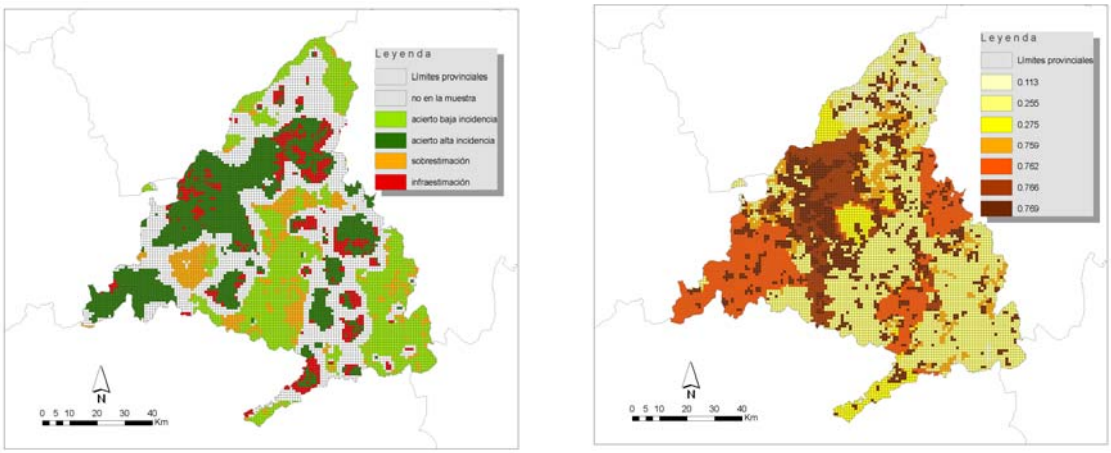


Figura 7— Aciertos y errores Árbol de Decisión y probabilidad estimada en la Comunidad de Madrid

El modelo desarrollado a partir de las reglas del árbol infraestima el riesgo en zonas del Norte, Noreste y Sureste, mientras que sobrestima en el Suroeste y centro principalmente. La distribución de las zonas de acierto y error sigue los mismos patrones que en el caso del modelo obtenido mediante regresión logística para la Comunidad de Madrid. El mapa de probabilidad obtenido mediante el árbol de clasificación señala que las zonas de mayor probabilidad de incendio se encuentran en el Oeste, Suroeste, Noreste y una mancha al Sureste, coincidiendo con las áreas protegidas de la comunidad (ENP y ZEPA), variables que determinan reglas de decisión.

Redes neuronales

Los resultados obtenidos mediante la técnica de Redes Neuronales Artificiales (RNA) se sintetizan en las tablas 4 y 5 de *validación* de resultados y de *acierto y error*, así como en la tabla 6 de *análisis de sensibilidad*.

El entrenamiento de la RNA en la Comunidad de Madrid, arrojó un RMS (error medio cuadrático) de 0,1772.

		Validación	
		Alta ocurrencia	Baja ocurrencia
Resultado RNA	Alta ocurrencia	769	309
	Baja ocurrencia	393	663

Tabla 4—*Validación Redes Neuronales Comunidad de Madrid.*

	Alta ocurrencia	Baja ocurrencia
Acierto RNA	71,34 pct	62,78 pct
Error comisión	28,66 pct	37,22 pct
Error omisión	33,82 pct	31,79 pct

Tabla 5—*Aciertos y errores método redes neuronales Comunidad de Madrid.*

El acierto global de la Red en la C. de Madrid en la alta ocurrencia es de un 71,34 por ciento mientras en la baja ocurrencia el método clasifica correctamente en un 62,78 por ciento. Los errores de comisión son 28,66 y 37, 22 por ciento, respectivamente, mientras que los de omisión superan el 30 por ciento en ambos casos.

Variable	**Valor RMS**	**Diferencia con el de la RNA inicial en valor absoluto**	**Orden de importancia en el entrenamiento de la RNA**
Areas recreativas	0,1799	0,0027	16º
Carga ganadera	0,1682	0,009	10º
Enp	0,2256	0,0484	1º
Ffcc._for	0,1654	0,118	6º
Hotel	0,1701	0,0071	13º
Icc	0,1669	0,0103	8º
Icf	0,2082	0,031	3º
Imd	0,2153	0,0381	2º
Imd_for	0,1844	0,0072	12º
Ipf	0,1705	0,0067	14º
Iuf	0,1625	0,0147	4º
Jefes	0,1835	0,0063	15º
Llee_for	0,1771	0,0001	19º
Medios_vig	0,1852	0,008	11º
Mup_preser	0,1657	0,0115	7º bis
Paro	0,1744	0,0028	17º
Pistas_for	0,1657	0,0115	7º
Pot_dem	0,1772	0	20º
Var_pob	0,1772	0	20º bis
Var_pob_agra	0,1784	0,0012	18º
Zepa	0,1877	0,0105	5º
Vertederos	0,1675	0,0101	9º

Tabla 6—*Análisis de sensibilidad método redes neuronales Comunidad de Madrid.*

Como ya se ha mencionado, el peso de las diferentes variables se estimó a partir de un *análisis de sensibilidad*, basado en la comparación del valor del RMS ofrecido por la red diseñada y el que se obtiene cuando se anula el valor, una a una, de cada variable.. Observando la tabla 6 donde se muestra el resultado del *análisis de sensibilidad* comprobamos que las variables *buffer de ferrocarril en zona forestal* (*Ffcc._for*), *espacios naturales protegidos* (*ENP*), *índice de intensidad media diaria de tráfico* (*Imd*), *montes de utilidad pública y preservados* (*mup_preser*), *pistas en zona forestal* (*pistas_for*), así como, *interfaz urbano-forestal* (*Iuf*), son las que mayor peso tienen.

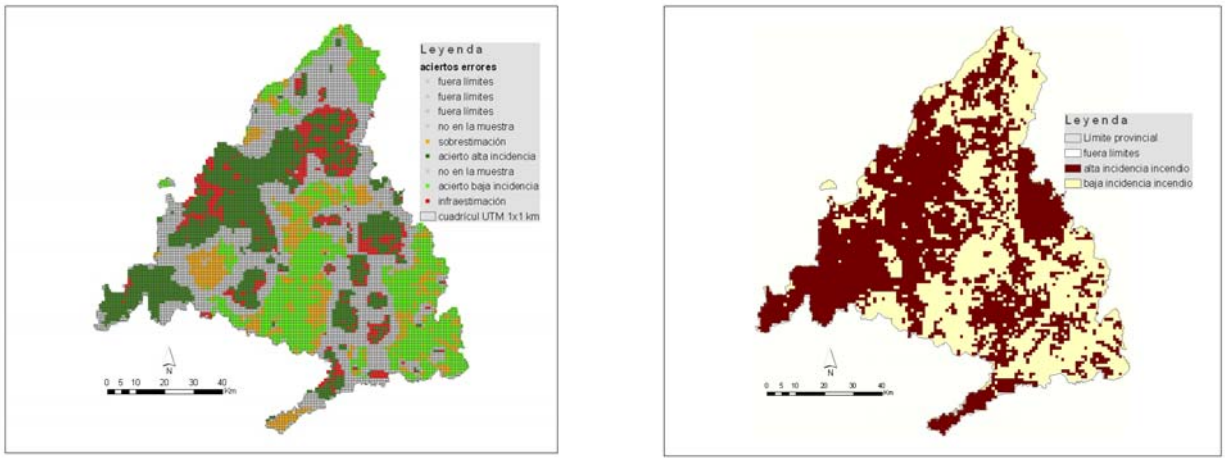


Figura 8— Aciertos y errores redes neuronales y ocurrencia estimada en la Comunidad de Madrid

En el mapa obtenido (figura 8) a partir del método de redes neuronales las zonas en las que la red predice alta incidencia de incendio se localizan en el Oeste (zona de la Sierra de Madrid), NE, coincidiendo con una ZEPA y al Sur, Sur-este, coincidiendo con ZEPA y el Espacio Natural Protegido del Parque Regional del Sureste. En el resto de celdas el método predice se va a dar baja incidencia de incendios. Las zonas en las que el método infraestima se localizan al Norte y Noreste principalmente.

Provincia de Huelva

Regresión Logística

Al igual que en la Comunidad de Madrid, mediante la técnica de regresión logística binaria con las variables excluidas por colinealidad y con la variable dependiente obtenida a partir de interpolación mediante kernel adaptativo (muestra de 5 puntos) se obtienen 12 modelos, eligiendo el 5º. Los porcentajes globales de acierto de clasificación de la muestra de elaboración del modelo (60 por ciento) y de validación del mismo (40 por ciento) son 84,2 y 85,2 por ciento respectivamente. Los parámetros del modelo seleccionado se recogen en la tabla 7, siendo su ecuación la que se muestra a continuación:

Z_5= -4,177+0,004* Variación_Población+0,015* Variación_Población_Agraria+0,002* Potencial_Demográfico-2,422* ENP+26,627* Buffer_Carreteras

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0 pct para EXP(B)	
								Inferior	Superior
Paso 5	Var_pob	,004	,001	9,797	1	,002	1,004	1,002	1,007
	Var_pob_agra	,015	,002	92,806	1	,000	1,015	1,012	1,019
	Potencial_dem	,002	,000	579,250	1	,000	1,002	1,002	1,002
	Enp	-2,422	,123	386,762	1	,000	,089	,070	,113
	B_carret	26,627	4,612	33,328	1	,000	36626896	43436214,	30885047

							5584,715	205	78896969,000
	Constante	-4,177	,188	495,168	1	,000	,015		

Tabla 7— *Resultados del modelo 5 obtenido por Regresión Logística.*

Del mismo modo que sucedía en el área de estudio de la Comunidad de Madrid, las cinco variables seleccionadas por el modelo son significativas al 95 por ciento de confianza (significatividad menor de 0,05), y es la variable *buffer de carreteras* la que mayor peso tiene en el modelo (coeficiente B de 26,627) a priori. Las siguientes variables en importancia son *variación de la población agraria* y *variación de la población*.

Al aplicar la ecuación del modelo elegido al 100 por ciento de la muestra se obtiene un 84,4 por ciento correcto de clasificación global de la misma, estando la baja incidencia correctamente clasificada en un 92,4 por ciento y la alta incidencia en un 76,4 por ciento.

Una vez normalizadas las variables del modelo elegido los resultados de la regresión arrojan las variaciones de la variable dependiente respecto de cada independiente recogidas en la tabla 8:

Paso 5	**dx/dy**
z_var_pob	0.0100406
z_var_pob_agra	0.0166597
z_potencial_dem	0.6741251
z_enp	-0.0407765
z_b_carret	0.0148861

Tabla 8— *Efectos marginales del modelo 5. Variación de la variable dependiente x con cada variable independiente y (dx/dy).*

Al estudiar los efectos marginales la variación de la variable dependiente respecto de cada independiente del modelo indica que *potencial demográfico* es la que mayor cambio produce: si se aumenta en una unidad la variable *potencial*, la variable dependiente aumenta 0,67 en desviación típica. Le siguen en importancia *variación de la población a agraria* y *buffer de carreteras*. Al normalizar las variables se observa el verdadero efecto de cada variable independiente sobre la dependiente.

A continuación se muestra el mapa de los aciertos y errores para la muestra de comprobación y validación del modelo así como el mapa de probabilidad estimada (Figura 9) para la provincia de Huelva:

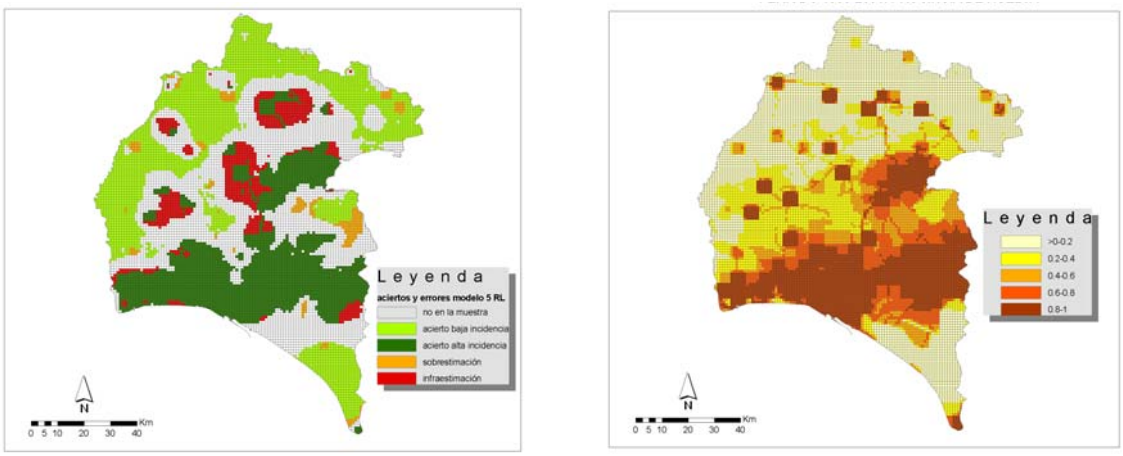


Figura 9— Mapas de aciertos y errores y de probabilidad estimada de riesgo humano (modelo 5)

El mapa de aciertos y errores indica que las zonas de infraestimación del modelo se encuentran en el Norte y centro del área de estudio especialmente, mientras que la sobrestimación se da en el Oeste de la misma, pero en muy pocas celdas. El modelo acierta en la alta incidencia en la franja que atraviesa la provincia de Este a Oeste, y en la baja incidencia en el Norte y Oeste. El mapa de probabilidad obtenido indica que las zonas de alta probabilidad de ocurrencia se encuentran en la franja que atraviesa la provincia, coincidente con la zona de valle de Huelva y las zonas más pobladas. Las zonas de menor probabilidad coinciden con los espacios naturales protegidos del Parque de Doñana, al Sureste, y la Sierra de Aracena, al Norte.

Árboles de Decisión

En el caso de la provincia de Huelva el árbol de decisión de la figura 10 señala a las variables *potencial demográfico*, *ENP*, *renta*, *maquinaria agrícola* y *carga ganadera* como las responsables del establecimiento de grupos de alta y baja incidencia de incendios forestales por causa humana. El porcentaje global correcto del modelo obtenido es de un 84,5 por ciento, clasificando correctamente la baja incidencia de incendio en un 87,4 por ciento y la alta en un 81,7 por ciento.

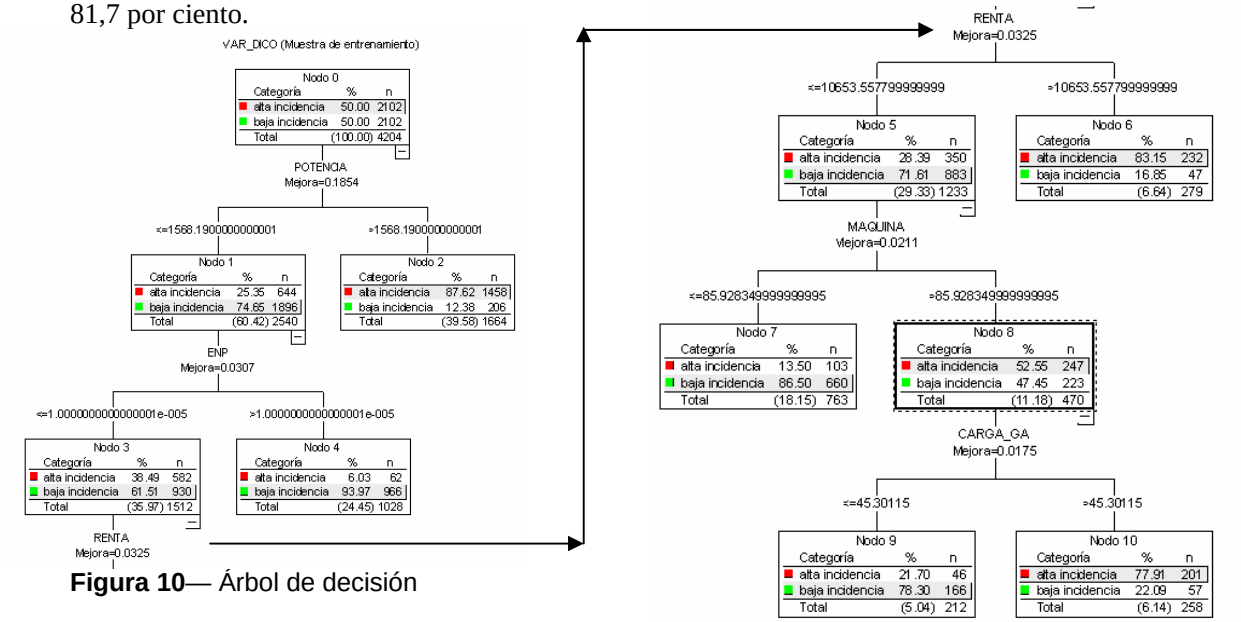


Figura 10— Árbol de decisión

La variable *potencial demográfico* es la que determina el primer nivel, con una mejora del 18,5 por ciento clasifica en alta incidencia en un 87,6 por ciento aquellas celdas con valores de potencial mayores a 1568,19. Cuadrículas con valores menores ó iguales a 1568,19 y *ENP* mayor de 1e-5 estarán clasificadas en baja incidencia en un 93,9 por ciento. Descendiendo otro nivel, cumpliendo las reglas de decisión anteriores y el valor de *ENP* menor ó igual a 1e-5 y *renta* mayor de 10653, las celdas serán de alta incidencia (83,1 por ciento). Si la renta es menor ó igual a 10653, con una mejora del 2 por ciento, es la variable *maquinaria agrícola* la que clasifica las celdas en baja incidencia (86,5 por ciento) si esta variable es menor ó igual a 85,9. El último nivel está definido por la carga ganadera; si la maquinaria agrícola es mayor de 85,9 y la carga ganadera supera 45,3, las cuadrículas serán de alta incidencia (77,9 por ciento).

El método de árboles de decisión clasifica correctamente en un 84,5 por ciento, estando la baja incidencia bien clasificada un 87,4 por ciento y la alta incidencia un 81,7 por ciento.

El mapa de acierto y error del modelo generado mediante la técnica de árboles de decisión así como el mapa de probabilidad estimada se muestra en la figura 11:

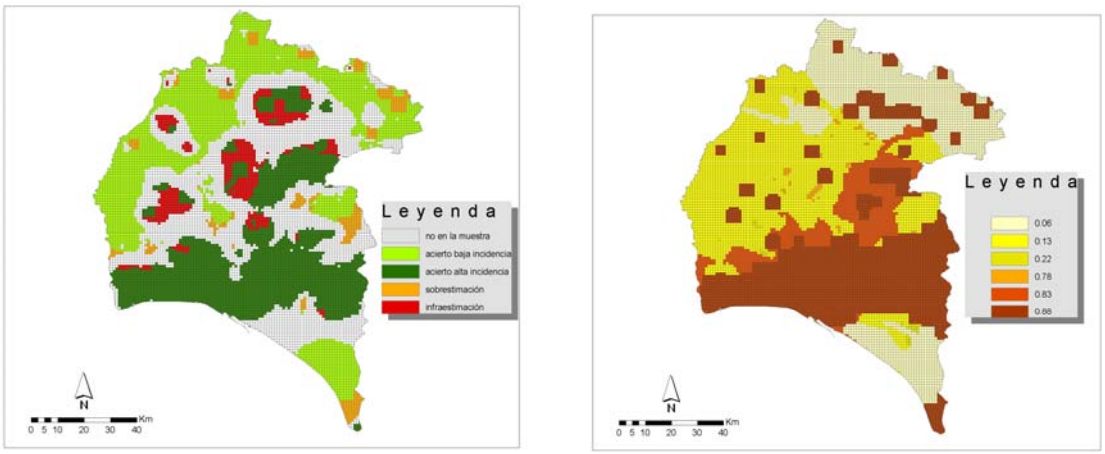


Figura 11— Aciertos y errores Árbol de decisión y probabilidad estimada en la Provincia de Huelva

Se observa que la distribución espacial de las zonas de infraestimación y sobrestimación coincide con el resultado obtenido mediante regresión logística, produciéndose errores de infraestimación en zonas al Norte y centro principalmente. La alta incidencia está correctamente clasificada en la zona central. La probabilidad estimada a partir de reglas de decisión señala como zonas de alta probabilidad de incendio la franja del valle de Este a Oeste así como cuadrículas de alto valor de potencial demográfico. Las zonas de baja probabilidad coinciden con los espacios naturales protegidos, la Sierra de Aracena al Norte y el parque de Doñana al Sureste.

Redes neuronales

Los resultados obtenidos mediante la técnica de Redes Neuronales se sintetizan el las tablas 9 y 10 de *validación* de resultados y de *acierto y error*, así como en la tabla 11 de *análisis de sensibilidad*. En el caso de la provincia de Huelva el RMS fue de 0,198.

		Validación	
		Alta ocurrencia	Baja ocurrencia
Resultado RNA	Alta ocurrencia	1359	28
	Baja ocurrencia	718	583

Tabla 9—*Validación Redes Neuronales provincia de Huelva.*

	Alta ocurrencia	Baja ocurrencia
Acierto RNA	97,98 pct	44,81 pct
Error comisión	2,02 pct	55,19 pct
Error omisión	34,57 pct	4,58 pct

Tabla 10—*Aciertos y errores método redes neuronales provincia de Huelva.*

En la provincia de Huelva el acierto global de la red en la alta densidad de incendio es del 97,98 por cierto, mientras que la baja incidencia se clasifica correctamente en un 44,81 por ciento. Los errores de comisión son 2,02 y 55,19 por ciento respectivamente mientras que los de omisión son del 34,57 y del 4,58 por ciento. En definitiva se trata de un modelo un tanto deficiente ya que sobrevalora las áreas de alta incidencia, mientras que, complementariamente, infravalora las de baja incidencia; de ahí, que el porcentaje de acierto en las primeras sea muy alto.

Variable	**Valor RMS**	**Diferencia con el de la RNA inicial en valor absoluto**	**Orden de importancia en el entrenamiento de la RNA**
Areas recreativas	0,2019	0,0039	12º
Cantera_tiro	0,2285	0,0305	3º
Carga_gana	0,1980	0	15º
carrete	0,2168	0,0188	7º
Enp	0,2417	0,0437	1º
Ffcc	0,1873	0,0107	9º
Hotel	0,2170	0,019	6º
Icc	0,2060	0,008	10º
Icf	0,2300	0,032	2º
Ipf	0,2030	0,005	11º
Iuf	0,1966	0,0014	15º
Llee	0,2268	0,0288	4º
Maquina	0,1995	0,0015	14º
Mconsor	0,2002	0,0022	13º
Mup	0,2029	0,0049	12º
Paro	0,2116	0,0136	8º
Pistas	0,1980	0	16º
Pot_dem	0,1980	0	16º bis
Var_pob	0,1980	0	16º bis
Var_pob_agra	0,1980	0	16º bis
Vertederos	0,2235	0,0255	5º

Tabla 11—*Análisis de sensibilidad método redes neuronales provincia de Huelva.*

En la provincia de Huelva las variables que presentan mayor diferencia en RMS con el valor de referencia de la red son *Espacios Naturales Protegidos* (*Enp*), *interfaz cultivo-forestal* (*Icf*), *presencia de campos de tiro y canteras* (*cantera-tiro*), *buffer de líneas eléctricas* (*Llee)* y

vertederos. En la figura 12 se muestra el mapa de aciertos y errores así como la ocurrencia estimada.

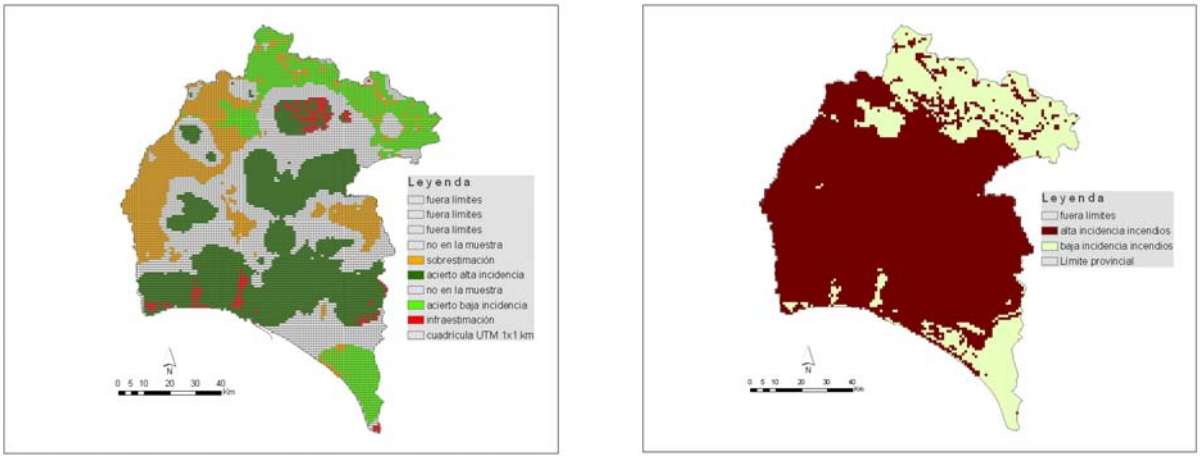


Figura 12— Aciertos y errores Redes neuronales y ocurrencia estimada en la Provincia de Huelva

En el mapa obtenido mediante el modelo de redes neuronales en la provincia de Huelva se observa que en la práctica totalidad del área de estudio el modelo predice alta incidencia de incendio salvo en la zona Norte de la Sierra de Aracena y en el Sureste del Parque Nacional de Doñana. La sobrestimación en este caso es muy elevada, observándose grandes manchas en la zona del Oeste de la provincia principalmente.

Discusión

Los resultados obtenidos a partir de las distintas técnicas muestran, en la Comunidad de Madrid, que los Árboles de Decisión logran la mejor clasificación global del modelo (75,5 por ciento), aunque en las tres los porcentajes de clasificación son similares, tanto el global como la alta y la baja incidencia (alrededor del 70 por ciento). Los aciertos y errores se localizan geográficamente en áreas similares, infraestimando los modelos algunas zonas del Norte y Este del área de estudio principalmente. Las variables que explican la probabilidad de incendio en los tres métodos coinciden y son la *interfaz urbano-forestal* y *Espacios Naturales Protegidos*, estando representadas también las variables *ZEPA, buffer de líneas de ferrocarril en zona forestal* e *índice de intensidad media de tráfico.*

En la provincia de Huelva los modelos obtenidos por regresión logística y árboles de decisión clasifican correctamente la probabilidad de riesgo humano en un 84 por ciento, clasificando correctamente la alta incidencia en un 80 por ciento y la baja en un 90 por ciento. Sin embargo, la técnica de redes neuronales clasifica correctamente la alta incidencia en un 97 por ciento y la baja tan solo en un 44,8 por ciento, dando lugar a una importantes sobreestimación del riesgo en la práctica totalidad del área de estudio. Con las dos primeras técnicas (regresión logística y árboles de decisión) las zonas donde se produce infraestimación se localizan en sectores del centro del área de estudio, clasificando correctamente las zonas de alta incidencia de la franja Este-Oeste. La baja incidencia del Norte y Sureste se clasifica correctamente. Respecto a la capacidad explicativa de los modelos obtenidos, las variables seleccionadas tanto en regresión logística como en árboles de decisión están relacionadas con la población, como el *potencial*

demográfico, *variación de la población agraria*, así como otras relacionadas con el uso del territorio, *maquinaria*, *buffer de carreteras*, *carga ganadera*, *espacios naturales protegidos*. Las redes neuronales seleccionan otras como *interfaz cultivo-forestal*, *buffer de líneas eléctricas* ó *vertederos*. El modelo obtenido mediante redes neuronales no es satisfactorio probablemente debido a que las muestras de entrenamiento están descompensadas, dada la distribución de zonas de alta y baja ocurrencia de incendios. La alta ocurrencia predomina en el centro de la provincia (salvo un pequeño sector de baja ocurrencia) y la baja en zonas del Norte y Sureste, y límites occidental y centro-oriental. Al tomar muestras proporcionales al tamaño de las manchas de alta y baja ocurrencia durante el proceso de entrenamiento, los píxeles han tendido a agruparse en las zonas donde predominaba cada tipo de ocurrencia, con lo que el resultado final ha presentado una generalización siguiendo los patrones espaciales definidos por la distribución de las celdas de alta y baja ocurrencia. Quizá podría haberse forzado la selección de un mayor número de píxeles en las áreas con menor tamaño, pero preferimos mantener el criterio de una selección aleatoria para que los resultados respondieran a un método lo más homogéneo posible para su posterior comparación.

Salvo el caso de las redes neuronales aplicadas en la provincia de Huelva, las distintas técnicas utilizadas predicen y explican de manera similar, en porcentaje de acierto y en las variables seleccionadas para dicha predicción. La variable *interfaz urbano-forestal* seleccionada en la Comunidad de Madrid se ajusta al conocimiento previo que se tiene acerca de las causas de incendio proporcionado por los gestores de dicha área de estudio, aunque otras variables como los *Espacios Naturales Protegidos* parecen estar explicando otro fenómeno, como la localización de masas vegetales, al no darse problemas reales de riesgo de incendio en estas zonas. En la provincia de Huelva las variables poblacionales explican el fenómeno, así como la *interfaz cultivo-forestal* en el caso de las redes. Comparando las dos áreas de estudio, con las dos primeras técnicas la capacidad predictiva es mayor en la provincia de Huelva, un 10 por ciento aproximadamente.

Las técnicas de regresión logística y redes neuronales exigen una mayor preparación previa de las variables así como de manejo del método (caso de las redes neuronales), mientras que los árboles de decisión permiten la entrada de todo tipo de variables. Sin embargo, esta última técnica se suele recomendar para el conocimiento previo del conjunto de variables, cuáles influyen más en la variable dependiente y su orden de importancia, para a continuación analizar estas variables con otros métodos, como los estadísticos tradicionales, y obtener modelos de riesgo. Por otro lado, la obtención de la variable dependiente con el método detallado anteriormente implica la incertidumbre en la localización de los puntos de ignición así como el empleo de superficies continuas de densidad de incendio creemos que este hecho puede estar influyendo en el modelo final obtenido, independientemente del método empleado.

Conclusiones

La obtención de modelos de riesgo humano de incendio forestal para su posterior integración en modelos de riesgo más complejos resulta de gran interés para lograr una mejora en la capacidad predictiva de los mismos. Debido a la dificultad que entraña la obtención de estos modelos, a menudo los sistemas integrados de predicción de riesgo no incluyen el factor humano.

En esta comunicación se ha llevado a cabo un ensayo comparativo de la potencialidad de tres técnicas para la obtención de modelos de riesgo, y, a pesar de las limitaciones, puede decirse que se obtienen aproximaciones similares y aceptables con las tres, exceptuando el modelo obtenido con redes neuronales en la provincia de Huelva. Los resultados obtenidos sugieren

profundizar en la técnica de redes neuronales así como en el empleo de otras variables dependientes que ofrezcan una mayor precisión en la localización espacial de los incendios.

Agradecimientos

Esta comunicación forma parte de la investigación desarrollada por el grupo de Tecnologías de la Información Geográfica del Instituto de Economía y Geografía (IEG) del CSIC en el marco del Proyecto *Firemap*, "Análisis Integrado de Incendios Forestales mediante Teledetección y Sistemas de Información Geográfica" (CGL2004-06049-C04-02/CLI)[9] y ha sido parcialmente financiada por el programa de Formación de Personal Investigador FPI BES-2005-7712 del Ministerio de Educación y Ciencia. Deseamos expresar nuestro agradecimiento a todas las instituciones que nos han facilitado información para la realización del estudio: *Dirección General para la Biodiversidad del Ministerio de Medio Ambiente;* en la Comunidad de Madrid, a la *Dirección General de Medio Natural; Dirección General de Carreteras; Dirección General de Agricultura y Desarrollo Rural; Servicio Cartográfico regional; Jefatura Cuerpo de Bomberos; Departamento Geografía UAH*. En Andalucía, a la *Universidad de Córdoba.*

Referencias bibliográficas

Amatulli, G., Pérez-Cabello, F., de la Riva, J. (2005). **Mapping lightning/human-caused wildfires ocurrence under ignition point location uncertainty.** Ecollogical modelling. Manuscript

Asociación para la Promoción de Actividades Socioculturales (APAS) (2004). **Estado del Conocimiento sobre las Causas de los Incendios Forestales.** *Proyecto financiado por la Dirección General para la Biodiversidad del Ministerio de Medio Ambiente*.

L. Breiman, W. Meisel and E. Purcell, **Variable kernel estimates of multivariate densities**. Technometrics 19 (1977), pp. 135–144.

Bischof, H. et al. (1992). **Multispectral classification of Landsat images using neural networks**. IEEE Transactions on Geoscience and Remote Sensing 30: 482-489.

Carvacho, L. (2002). **Aplicación de redes neuronales al análisis de datos en teledetección: predicción y cartografía de incendios forestales**. Departamento de Geografía. Alcalá de Henares, Universidad de Alcalá**:** 206.

Chuvieco, E., Salas, J., de la Riva, J., Pérez, F., Lana-Renault, N. (2004). **Métodos para la integración de variables de riesgo: el papel de los sistemas de información geográfica**, en Chuvieco, E., Martín, P. (Ed.): Nuevas tecnologías para la estimación del riesgo de incendios forestales. Madrid, CSIC, Instituto de Economía y Geografía, pp. 144-158.

[9] Proyecto financiado por la CICYT (diciembre 2004 -diciembre 2007). Entidades participantes: Universidad de Alcalá, Universidad de Córdoba, IEG-CSIC, INM, Universidad Castilla la Mancha, Universidad Politécnica de Madrid, CEAM, Universidad de Zaragoza

Dirección General para la Biodiversidad (2006). **Estadísticas de Incendios Forestales**. http://www.incendiosforestales.org/estadisticas.htm. Ministerio de Medio Ambiente.

Garson, D., 2006. **Statnotes: Topics in Multivariate Analysis**. http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

González, C. (2006). **Análisis de Datos Cualitativos. Curso de Metodología de Investigación Cuantitativa**. Técnicas Estadísticas. CSIC

Hilera, J.R. y Martínez, V.J. (1995). **Redes Neuronales artificiales: Fundamentos, modelos y aplicaciones**. Serie Paradigma, RA-MA Editorial, Madrid.

Instituto de Estadística de Andalucía (2007). **Huelva, Datos Básicos 2007**. http://www.juntadeandalucia.es/institutodeestadistica/dtbas

Klimasauskas, C.C. (1991c). **Applying neural networks. Part III: Training a neural network**. PC Artificial Intelligence: 20-24.

Leone, V., Koutsias, N., Martínez, J., Vega-García, C., Allgöwer, B., Lovreglio, R. (2003). **The human factor in fire danger assessment**, en Chuvieco, E. (Ed): Wildland fire Danger estimation and mapping. The role of remote sensing data. Series in Remote Sensing. World scientific Publishing Co. Pp. 143-194

Levine, N. (2004). **Kernel density interpolation**, en Crimestat 3.0, capítulo 8

Martín, P., Chuvieco, E., Aguado, I. (1998). **La incidencia de los Incendios Forestales en España**. Serie Geográfica, 7, pp. 23-36

Martínez, J. (2004). **Análisis, Estimación y Cartografía del Riesgo Humano de Incendios Forestales**. Tesis Doctoral. Facultad de Filosofía y Letras. Departamento de Geografía. Universidad de Alcalá

Martínez, J., Martínez, J., Martín, P. (2004). **El factor humano en los incendios forestales: Análisis de factores socio-económicos relacionados con la incidencia de incendios forestales en España**, en Chuvieco, E., Martín, P. (Eds.): Nuevas tecnologías para la estimación del riesgo de incendios forestales. Madrid, CSIC, Instituto de Economía y Geografía, pp. 101-142

Moyano, E. (2006). **Procesos de cambio en la agricultura y el mundo rural. Algunas reflexiones para el debate**. Jornada sobre Incendios Forestales. Fundación Biodiversidad. Fundación Santander-Central Hispano

Pausas, J. (2004). **Changes in fire and climate in the eastern Iberian Peninsula (mediterranean basin).** Climatic Change 63**:** 337–350, 2004.

Pew, K.L., Larsen, C.P.S (2001). **GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rain forest of Vancouver Island, Canada**. Forest Ecology and Management, 140, pp. 1-18

De la Riva, J., Pérez-Cabello, F., Lana-Renault, N., Koutsias, N. (2004). **Mapping wildfire ocurrence at regional scale**. Remote Sensing of Environment, 92, pp. 363-369.

Vega-Garcia, C., Woodard, P. M., Titus, S. J., Adamowicz, W. L., and Lee, B. S. (1995). **A Logit Model for Predicting the Daily Occurrence of Human Caused Forest Fires**', Int. J. of WildlandFire 5 2, pp 101–112.

Vega-García (1996). **Predicción de incendios forestales de causalidad humana en Whitecourt forest, Alberta**. Departamento de Geografía. Alcalá de Henares, Madrid., Universidad de Alcalá**:** 108.

Villagarcía, T. (2006). **Regresión, Curso de Metodología de Investigación Cuantitativa**. Técnicas Estadísticas. CSIC

Zhang, B., Valentine, I., Kemp, P. (2005). **Modelling the productivity of naturalised pasture in the North Island, New Zealand: a decisión tree aproach**, Ecological Modelling, 186, pp. 299-311.