**UNIVERSIDAD DE ALCALÁ**

**DEPARTAMENTO DE GEOGRAFÍA**

# ELABORACIÓN DE MODELOS ESPACIALES PREDICTIVOS DE OCURRENCIA DE INCENDIOS FORESTALES ASOCIADA A LA ACTIVIDAD HUMANA

Tesis doctoral presentada por

**Lara VILAR DEL HOYO**

Bajo la dirección de la

Dra. Mª Pilar Martín Isabel

Científica Titular CSIC, Profesora Asociada Departamento de Geografía

Programa de Doctorado en Cartografía, SIG y Teledetección del Departamento de Geografía

Alcalá de Henares, Abril de 2009

A mi familia y amigos

## AGRADECIMIENTOS

En primer lugar quiero agradecer a mi directora de tesis, Pilar Martín, la confianza depositada en mí desde el principio, su amabilidad, buenos consejos y ayuda para tomar cualquier tipo de decisión. Destaco el buen ambiente de trabajo compartido, las oportunidades que me ha ofrecido para mi formación y experiencia (congresos, cursos…) y, sobretodo, su apoyo en todo momento. Ha sido una oportunidad inmejorable para aprender un método de trabajo eficaz y práctico, de una persona dedicada con entusiasmo a la labor de investigación. Me atrevería a decir que sin su optimismo y mensaje resumido en una frase, "Lara, esto ya está" hubiese sido difícil llegar al final.

Agradecer a Javier Martínez Vega su apoyo, saber estar y capacidad de trabajo, consejos y puntos de vista, sin olvidar su trato profesional y personal realmente bueno, siempre con una palabra amable, que han facilitado enormemente la inmersión en el trabajo científico y mundo del CSIC.

Por otro lado, mi agradecimiento a Emilio Chuvieco, el cuál me ofreció la oportunidad de ampliar mi formación académica y profesional en el departamento de Geografía, en un grupo de gran calidad, tanto profesional como humana.

La realización de esta tesis ha sido posible gracias al disfrute de una beca FPI del Ministerio de Ciencia e Innovación, asociada al proyecto *Firemap*.

De forma muy especial, agradecer a mi familia su apoyo incondicional en los buenos y malos momentos, siempre constante, aguantando el tirón de mis cambios de ánimo en las etapas de mayor intensidad. Sabéis que, sin vuestro cariño, consejos, ideas, sugerencias, no estaría escribiendo estas líneas. Os debo sin duda el haber llegado hasta esta página, y, en definitiva, lo que soy, gracias de verdad por ofrecer lo mejor de vosotros mismos siempre, confiando sin cuestionar las decisiones que he ido tomando en estos años.

Y qué puedo decir de mis compañeros de viaje del CSIC; desde los comienzos en Pinar con grandes momentos como "que bien, salimos a las 17h", los cafés de después de comer en "nuestro" saloncito de sofás verdes, los vinos de Navidad entre incunables, algún que otro picnic en la terraza…al traslado a Albasanz, con interesantes conversaciones con olor a paella en "El Bulli" (del corazón, de actualidad, de recetas de cocina, de tipos de "taper", wikipedia...) y descubriendo juntos las novedades "científicas" (la terraza "polivalente" de la cafetería, los de historia, los demás, los de abajo, los del CAU, el pingi-pongi…entre muchas otras). Como no quiero dejarme a nadie los citaré como "los 300 del ieg", y decir que, el ánimo en todo momento de becarios y no becarios, consejos, puntos de vista, ayuda, comprensión y, sobretodo, buen humor, han sido fundamentales para llegar al final. El buen ambiente de trabajo y compañerismo dentro y fuera de la cueva bien se merece un párrafo en esta tesis.

Por otro lado, en esta etapa también han estado presentes los compañeros del departamento de Geografía de la Universidad de Alcalá, los del "sótano" y los del "torreón", sin olvidar a los profesores del departamento. Han sido numerosas las reuniones, congresos y horas de trabajo compartidas, de nuevo puedo decir que con un alto nivel profesional. En lo personal, aunque me dicen que soy demasiado "diplomática", no puedo dejar de mencionar a Héctor, Marta, Mariano, Angela, Patricia…siempre dispuestos para cualquier cosa.

De ambos ambientes de trabajo quedarán en mi memoria grandes reflexiones como "la normalidad no existe", "el universo es un tema", "la vida está aquí", "que muera la mortalidad", "de todo se sale"…entre muchas otras.

El transcurso de elaboración de tesis me ha permitido la gran oportunidad realizar estancias en otros centros de investigación, conociendo otros métodos de trabajo, enfoques, opiniones, conocimientos, maneras de ver la vida… No me gustaría olvidar hacer mención de la gente que he conocido y con la que he compartido interesantes momentos, recibiendo una muy buena acogida y cuyas aportaciones han sido fundamentales para el trabajo realizado: del *Firelab* en la *Faculty of Forestry* de la Universidad de Toronto (Canadá) nombrar de forma especial a Douglas Wooldford y Dave Martell (acompañados de un grupo muy bueno de canadienses y europeos), y, del *Joint Research Centre* en Ispra (Italia), nombrar a Andrea Camia, Jesús San-Miguel Ayanz,

Giuseppe Amatulli, Tracy Durrant y demás compañeros del IES, así como al numeroso grupo de *stagiers*.

No puedo acabar estos agradecimientos sin citar a mis amigos de "fuera del mundo de la investigación", de los cuales alguno sigue pensando que voy en bata blanca a trabajar o que me dedico a hacer "escuchas" en el CSIC. Bien sabemos todos que los modelos de riesgo de incendio deberían incluir otro factor humano de origen madrileño con nombre y apellidos... Agradezco vuestro cariño, y, sobretodo, apoyo sin condiciones estando ahí siempre, algo que ningún modelo puede estimar.

Como diría un buen amigo, "quiero dedicar esta canción"…a todos, gracias por estar ahí y haber formado parte de esta intensa etapa de trabajo científico. Con mucho cariño (y puedo decir que, "sentía"),

Lara

Nuestras horas son minutos cuando esperamos saber,
y siglos cuando sabemos lo que se puede aprender

*Antonio Machado*

# ÍNDICE

## INTRODUCCIÓN

Los incendios forestales son aquellos que suceden en ambientes naturales (Bachmann, 2001), definidos también como los fuegos que se propagan, sin control, en un sistema forestal y cuya quema no cumple funciones ni objetivos de gestión, por lo que requieren trabajos de extinción (Salas y Cocero, 2004). A nivel global, el papel del fuego en algunos ecosistemas es esencial para el mantenimiento de sus dinámicas, productividad y biodiversidad, constituyendo además una herramienta necesaria para la gestión del territorio (FAO, 2007). Por otro lado, cada año los incendios forestales destruyen millones de hectáreas de bosque, con importantes consecuencias ecológicas, ambientales, económicas y sociales. Además de los daños directos (impactos en la salud, pérdidas en biodiversidad, emisiones de $CO_2$ y otros gases de efecto invernadero, entre otros) los incendios producen efectos secundarios tales como erosión ó deslizamientos del terreno (cuando se producen lluvias en zonas quemadas donde no hay vegetación), así como plagas de insectos que suceden a dichos incendios (FAO, 2007). La mayor parte de los datos estadísticos disponibles sobre incendios forestales a nivel mundial señalan como causa principal de los incendios la acción del hombre, salvo en zonas remotas de Canadá o Rusia donde los rayos provocan el mayor porcentaje de los incendios. En la región Mediterránea más del 95% de los incendios forestales responden a causas humanas (FAO, 2007).

A pesar de la importancia de la acción del hombre en la ocurrencia de los incendios forestales hoy en día la disponibilidad de datos es todavía escasa a nivel global. La FAO (*Food and Agriculture Organization of the United Nations*) recopila estadísticas de causas de incendio desde 1980 que incluyen información sobre el número total de incendios y área quemada por tipo de uso del suelo y causa (FAO, 2002). Estas estadísticas están incompletas y son bastante heterogéneas. A nivel europeo, la Comisión Europea a partir de la norma ECC 2158/92 de la protección de los bosques frente al fuego, en 1994 estableció que los países miembros debían recoger un mínimo de información relativa a los incendios forestales, que fuese comparable y accesible a intervalos regulares. Estas estadísticas deben incluir, entre otros, datos del origen del incendio a nivel municipal (y sucesivas unidades territoriales), así como la causa de incendio según cuatro grandes categorías: desconocido, natural, intencionado y accidente/negligencia. En España, los datos de incendios se llevan recogiendo desde 1968 en los llamados Partes de Incendio (Ministerio de Medio Ambiente y Medio Rural y Marino-MARM-). Estos partes

contienen más de 150 campos de información relativa a los incendios forestales, incluyendo información de la localización espacial del incendio, causas, motivaciones, área quemada y evaluación de daños, entre otros. Por lo que respecta a la causalidad, los partes españoles clasifican los incendios en: producidos por rayos, negligencias/accidentes, intencionados, desconocidos o reproducidos.

Para llevar a cabo unas adecuadas labores de prevención en la lucha contra los incendios forestales, es necesario contar con la máxima cantidad de información posible sobre la incidencia del fenómeno. Como ocurre en la evaluación de otros riesgos naturales como terremotos, inundaciones, sequías, ciclones, tornados, tsunamis, deslizamientos de tierras, etc., la evaluación del riesgo de incendios forestales debe realizarse sobre la base de información espacial referenciada y de calidad que permita el establecimiento de prácticas para la toma de decisiones en la gestión de los mismos (Chen et al., 2003). Los Sistemas de Información Geográfica (SIG) reúnen funciones para la entrada de información, salida y/o representación gráfica y cartográfica, de gestión de la información espacial así como funciones analíticas (Bosque, 2000), por lo que constituyen una herramienta fundamental para el análisis de riesgos.

El término "riesgo" asociado a los incendios forestales hace referencia a la probabilidad de que se produzca un incendio en un lugar y momento dado, considerando la naturaleza e incidencia de los agentes causantes. Los índices de riesgo de incendio pueden clasificarse en índices a corto y a largo plazo (Chuvieco et al., 2003). Los índices de riesgo a corto plazo están relacionados con las variables dinámicas que afectan al peligro de incendio (parámetros meteorológicos y humedad del combustible) mientras que los índices a largo plazo se relacionan con variables que influyen en la ignición y/o propagación del fuego, tales como topografía, estructura del combustible, actividades humanas o patrones climáticos. Chuvieco et al. (2003, 2009) proponen que en la elaboración de índices de riesgo de incendio debe considerarse, tanto la probabilidad de ignición, como la evaluación de los daños potenciales (vulnerabilidad). La probabilidad de ignición estará relacionada con los agentes causales (factores humanos o naturales) así como con el estado del combustible. La integración de estas variables dará lugar al riesgo de ignición.

Aunque no es común la integración de variables relacionadas con los factores socioeconómicos que influyen en el riesgo de ignición en los índices de riesgo utilizados de forma operativa por los diversos organismos encargados de la gestión contra incendios,

diversos trabajos han elaborado modelos para la predicción de la ocurrencia de incendios por causa humana, empleando distintas técnicas y a diferentes escalas. Entre los primeros trabajos publicados en esta línea encontramos modelos basados en distribuciones binomiales negativas (Bruce, 1963) y de Poisson (Cunningham y Martell, 1973). Más frecuentes son los trabajos que optan por la utilización de métodos de regresión logística. Así encontramos los trabajos de Martell et al. (1987), Lorenzo y Pérez (1995); Chuvieco et al. (1999, 2009), Martínez et al. (2004, 2009) o Prasad et al. (2008) a nivel regional. A escala local encontramos los trabajos de Vega-García et al. (1995), Lin (1999), Pew y Larsen (2001), Vasconcelos et al. (2001). Otros autores han elaborado modelos empleando técnicas como la regresión lineal (Chao-Chin, 2002; Shypard et al., 2007), árboles de regresión (Amatulli et al., 2006, 2007; Amatulli y Camia, 2007), redes neuronales (Vega-García et al., 2007) o métodos de probabilidad bayesiana (Romero-Calcerrada et al., 2008).

En esta línea y debido a la importancia de la acción del hombre en los incendios forestales en las áreas Mediterráneas y, más concretamente, en España, esta tesis propone la elaboración de modelos de predicción de la ocurrencia de incendios forestales por causa humana, a una resolución espacial de 1km$^2$ que consideramos adecuada a las labores de gestión a escala regional. Una vez alcanzado este objetivo general, se propone la integración de modelos de ocurrencia por causa humana y rayo, componentes del riesgo de ignición en sistemas integrados de riesgo de incendio forestal. Las áreas de estudio han sido la C. de Madrid, la C. Valenciana, la C. de Aragón y la provincia de Huelva, áreas evaluadas en el proyecto *Firemap* (CGL2004-06049-C04-02/CLI) (Chuvieco et al., 2009) en el que se enmarca el trabajo realizado. De estas zonas, la principal área de estudio ha sido la C. de Madrid, debido a que se encuentra disponible mayor cantidad de información acerca de la ocurrencia de incendios. Para lograr el objetivo general, el trabajo se ha llevado a cabo en cuatro fases u objetivos específicos. Las metodologías propuestas y resultados obtenidos en cada fase se recogen en los 4 capítulos de los que consta la tesis. Cada capítulo corresponde a un artículo aceptado y/o en fase de evaluación en revistas de ámbito nacional e internacional con sistema de revisión externa. Tres de los capítulos están redactados en inglés, para cumplir con las condiciones requeridas para acceder a la mención de "Doctorado Europeo". Estos capítulos corresponden a artículos en fase de evaluación en revistas científicas de ámbito internacional. A continuación se describe el contenido de cada capítulo, recogido en síntesis en la Tabla 1:

*i.* **Capítulo I.** Mediante el empleo de herramientas SIG se han generado variables de tipo socioeconómico en las áreas de estudio, a partir de diversas fuentes de información de tipo cartográfico y estadístico. Aplicando análisis estadísticos de regresión logística se han obtenido modelos de predicción de ocurrencia de incendios por causa humana en las cuatro áreas de estudio. Para la elaboración de estos modelos se ha utilizado como variable dependiente la ocurrencia de incendios por causa humana recogida en las estadísticas oficiales (partes de incendio) y especializada a la escala requerida en el estudio a partir de técnicas de interpolación espacial (método kernel adaptativo). El artículo *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2008) Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional (Boletín de la AGE, 47, 5-29)* recoge la descripción de la metodología propuesta y los resultados obtenidos para la C. de Madrid y C. Valenciana. En ambas regiones el acierto global de los modelos es similar, en torno al 70%. La baja incidencia de incendio en este caso logra un mayor acierto en la predicción. Las variables seleccionadas que explican la ocurrencia en Madrid están relacionadas con el contacto entre zonas urbanas y forestales (interfaz urbano-forestal) mientras que en la C. Valenciana se relacionan con la variación de la población y la intencionalidad y/o negligencia del empleo de fuego en usos agrícolas y ganaderos.

En el transcurso de esta primera fase se han llevado a cabo análisis comparativos de estos modelos de ocurrencia basados en la regresión logística con los obtenidos a partir de Árboles de Decisión (CART) y redes neuronales, en este caso en la C. de Madrid y en la provincia de Huelva. El resultado de estos análisis se recoge en la comunicación presentada a la *IV Conferencia Internacional sobre Incendios Forestales* (*Wildfire 2007, Sevilla*, disponible en http://www.fire.uni-freiburg.de/sevilla-2007/contributions/doc/cd/SESIONES_TEMATICAS/ST1/VilardelHoyo_et_al_SPAIN.pdf, que se incluye en la presente tesis doctoral bajo el epígrafe de Anexo I con el propósito de mostrar las vías alternativas a la regresión logística que se exploraron en las etapas iniciales de la investigación. Las distintas técnicas utilizadas predicen y

explican de manera similar, seleccionando la variable interfaz urbano-forestal en la C. de Madrid y variables poblacionales en la provincia de Huelva. En esta área los resultados obtenidos mediante redes neuronales no son del todo satisfactorios. Las redes neuronales exigen una mayor preparación previa de las variables así como un trabajo exploratorio más complejo para la definición de la arquitectura de la red. Por otro lado, los Árboles de Decisión parecen más recomendables como técnica para una exploración previa del conjunto de variables a utilizar en el modelo de cara a analizar cuáles influyen más en la variable dependiente y su orden de importancia.

*ii.* **Capítulo II.** Una vez obtenidos los modelos de predicción de ocurrencia por causa humana el siguiente objetivo/fase es comprobar si una localización más precisa de la ocurrencia observada conduce a una mejora en los modelos. La incertidumbre en la localización espacial del inicio de los incendios pensamos que podría estar ejerciendo una influencia en los modelos obtenidos. Por este motivo, se decidió elaborar dichos modelos predictivos empleando otras variables dependientes obtenidas de fuentes de información distintas a las disponibles a nivel nacional. En este caso el ensayo se realizó sobre el territorio de la C. de Madrid, donde se encuentra disponible información espacialmente más detallada sobre la localización de los incendios. En este trabajo se ha empleado la técnica de regresión logística utilizando como variable dependiente los puntos de ignición a partir de coordenadas x, y. Estos resultados se han comparado con los obtenidos en otro modelo en el que la variable dependiente proviene de un conjunto de puntos aleatorios generados a partir de la información recogida en los Partes de Incendio y referida a una cuadrícula de 10x10 km. El artículo *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J. (2009) Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables (Journal of Environmental Management, en revisión)*, recoge los resultados obtenidos en la C. de Madrid. Ambos modelos difieren en su ajuste y en las variables explicativas, obteniéndose distintas distribuciones espaciales en la probabilidad de ocurrencia. El modelo elaborado a partir de coordenadas x, y de los puntos de ignición reales se ajusta mejor a la ocurrencia del fenómeno en el área de estudio, confirmado mediante una validación independiente y la opinión de

expertos. En esta validación el modelo propuesto alcanza un 75% de acierto mientras que el obtenido a partir de puntos aleatorios un 65%. La interfaz urbano-forestal es la variable con mayor influencia en la ocurrencia predicha.

*iii.* **Capítulo III.** Comprobado que el empleo de información espacial más precisa como variable dependiente da lugar a modelos mejor ajustados al fenómeno se planteó el siguiente objetivo: elaborar modelos para la predicción de la ocurrencia incluyendo la componente temporal. El propósito es comprobar si la inclusión de variables dinámicas mejora los modelos predictivos basados exclusivamente en variables de tipo estructural. Para ello, se emplea la técnica estadística de *Generalized Additive Models* (*GAMs*), extensión de los modelos lineales generales entre los que se incluye la regresión logística. Como variables independientes se emplean una selección de variables socioeconómicas procedentes de los análisis anteriores a las que se añaden otras variables dinámicas de tipo meteorológico y, como variable dependiente, los puntos de ignición localizados por coordenadas x, y. Este análisis se lleva a cabo en la C. de Madrid. El artículo *Vilar, L., Wooldford, D., Martell, D., Martín, M.P. A Generalized Additive Model for Predicting People-Caused Wildfire Occurrence in the Region of Madrid, Spain ( International Journal of Wildland Fire, enviado)* recoge los resultados obtenidos. Las variables independientes incluidas en el modelo son altamente significativas, mostrando tendencias esperadas. El modelo obtenido tiene un grado de ajuste satisfactorio a pesar de contar con una serie corta de datos de incendio. Su dimensión es espacio-temporal, permitiendo predicciones de la ocurrencia en el tiempo.

*iv.* **Capítulo IV.** Finalmente, se ha llevado a cabo la integración de modelos de ocurrencia de incendios de causa humana y producidos por rayo, a la resolución de 1km$^2$. Para ello se han empleado modelos de ocurrencia por causa humana elaborados a partir de puntos de ignición (coordenadas x, y) en la C. de Madrid y C. de Aragón, regiones con diferentes condiciones socioeconómicas y de ocurrencia de incendios. Para la integración de los valores de probabilidad se ha empleado un método probabilístico así como una media ponderada por la ocurrencia histórica,

comparando ambos con la media simple. El artículo *Vilar, L., Nieto, H., Martín, M.P. Integration of lightning and human-caused wildfire occurrence models (Human and Ecological Risk Assessment, en evaluación)* recoge los resultados obtenidos. Se ha llevado a cabo la validación de los resultados empleando el *Area Under the Curve* (AUC) obtenida a partir de la técnica *Receiver Operating System* (ROC) y la distancia de Mahalanobis. En la C. de Madrid la validación de los resultados es satisfactoria mientras que en la C. de Aragón requeriría una revisión de los modelos originales. En ambas regiones el empleo de una media ponderada por la ocurrencia histórica obtiene mejores resultados en la integración de los modelos de causalidad humano y natural. La distribución espacial del riesgo integrado coincide con la distribución de causas en ambas áreas de estudio.

**Tabla 1.** Síntesis contenido tesis doctoral

| Capítulo | Objetivos específicos | Publicación derivada (*) |
|---|---|---|
| **I** | Generación de variables de tipo socioeconómico y elaboración de modelos predictivos de ocurrencia de incendios por causa humana | Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2008). Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional (Boletín de la AGE, 47, 5-29) |
| **II** | Elaboración de modelos predictivos de ocurrencia de incendios por causa humana a partir de información más precisa de localización de inicio de incendios | Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2009). Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables (Journal of Environmental Management, en revisión) |
| **III** | Elaboración de modelos predictivos de ocurrencia de incendios por causa humana incluyendo el componente temporal | Vilar, L., Wooldford, D., Martell, D., Martín, M.P. A Generalized Additive Model for Predicting People-Caused Wildfire Occurrence in the Region of Madrid, Spain (International Journal of Wildland Fire, enviado) |
| **IV** | Integración modelos de ocurrencia de incendios por causa humana y rayos | Vilar, L., Nieto, H., Martín, M.P. Integration of lightning and human-caused wildfire occurrence models (Human and Ecological Risk Assessment, en revisión) |
| **Anexo I** | Comparación de técnicas para la generación de modelos predictivos de ocurrencia de incendios por causa humana | Vilar del Hoyo, L., Gómez Nieto, I., Martín Isabel, M.P., Martínez Vega, F.J (2007). Análisis comparativo de diferentes métodos para la obtención de modelos de riesgo humano de incendios forestales. Wildfire 2007, 4th International Wildland Fire Conference, 13-17 Mayo 2007, Sevilla, España |

(*) Lara Vilar es primera autora de todas las publicaciones derivadas del trabajo de tesis doctoral, y ha sido responsable de la elaboración de las propuestas metodológicas, ejecución de los procesos estadísticos y de análisis espacial para la generación, validación e integración de los modelos así como de la redacción de los documentos. M. Pilar Martín, directora del trabajo de tesis doctoral, es coautora de todas las publicaciones y ha participado en la elaboración de las propuestas metodológicas así como en la discusión de resultados y la revisión de los documentos. En el capítulo III, Douglas Wooldford se ocupó de la programación del proceso estadístico aplicado así como de la revisión del manuscrito. En el capítulo IV Héctor Nieto ha llevado a cabo desarrollos de la metodología para la integración, discusión de resultados así como colaborado en la revisión del manuscrito. En el Anexo I Israel Gómez

ha llevado a cabo el análisis de redes neuronales. Javier Martínez Vega aparece como coautor en los capítulos I, II y Anexo I por su participación en los planteamientos teóricos, discusión de resultados y revisiones de los documentos. David Martell es coautor en el capítulo III por sus aportaciones en la discusión de los resultados y revisión del manuscrito.

El trabajo realizado propone modelos espaciales de ocurrencia de incendios por causa humana en diferentes áreas de España a una resolución de 1km$^2$, comprobando diferentes alternativas de variables dependientes para la reducción de la incertidumbre espacial. Se propone la aplicación de *GAMs* para la obtención de modelos predictivos, incluyendo la componente espacio-temporal del riesgo. Trabajos previos han empleado esta técnica para la obtención de modelos de riesgo de incendio en función de variables meteorológicas, índices de riesgo o topografía (Brillinger et al. 2003, 2006; Preisler et al. 2004, 2007, 2008), resultando novedosa la inclusión de variables de tipo socioeconómico. Finalmente se han propuesto y validado métodos para la obtención de índices integrados de ignición. Se estima que la contribución metodológica es novedosa y los resultados relevantes tanto desde el punto de vista científico como por su interés social como aportación a la mejora de la gestión de las zonas forestales en la lucha contra incendios.

La línea de investigación queda abierta para seguir profundizando en la obtención de modelos predictivos más ajustados, por lo que se propone el empleo de *GAMs* en otras áreas, el uso de otras escalas espaciales, la aplicación de técnicas de regresión espacial o el análisis de cambios en el territorio y su relación con la causalidad de incendios. En el apartado de Líneas Futuras se describen con más detalle estas propuestas.

## INTRODUCTION

Wildfires are those which happen in natural environments (Bachmann, 2001) and are also defined as those that propagate, without control, in forest areas and require extinction activities because they don't have management purposes (Salas and Cocero, 2004). On one hand, at global level, fire is essential in some ecosystems to maintain their dynamics, productivity and biodiversity, also being a tool for land management (FAO, 2007). On the other hand, wildfires burn every year millions of forest hectares, with important ecological, environmental, economical and sociological consequences. Besides the direct damages (on the health, biodiversity losses, $CO_2$ emissions and other greenhouse gases) wildfires produce secondary effects such as landslides (when it rains in burned areas where there is not vegetation cover) or insect plagues after those fires (FAO, 2007). Wildfire statistics at global level show that the human action is the main fire cause, except for remote areas in Canada or Russia, where lightning starts the main percentage of wildfires. In the Mediterranean area more than 95% of wildfires are due to human causes (FAO, 2007).

Despite the importance of the human action in the wildfire occurrence, nowadays the available information on fire causes is still insufficient at global level. FAO (Food and Agriculture Organization of the United Nations) collects fire cause statistics since 1980 which include information related to the total number of fires, burned area by land use and fire cause (FAO, 2002). These statistics are incomplete and rather heterogeneous. In Europe, the European Commission through out the ECC 2158/92 regulation on forest protection against fire, established in 1994 that the EU members should collect a minimum set of comparable information about wildfires in regular time periods. These statistics should include data relate to fire origin at a municipal level (and successive territorial units) as well as fire cause, classified in four categories: unknown, natural, deliberate and accident/negligence. In Spain, wildfire data is collected since 1968 in the National Fire Statistics (Ministry of Environment and Rural and Marine Affairs-MARM). Those records have more than 150 fields, including information related to the fire location, causes, motivation, burned area, damages evaluation, etc. Regarding fire cause, the Spanish National Fire Records classification includes the following categories: lightning, negligence/accident, deliberate, unknown and re-ignited.

To carry out adequate fire prevention actions, it is needed to count on reliable fire data information in order to achieve an effective risk evaluation assessment. As in other natural risks, such as earthquakes, floods, droughts, cyclones, tornados, tsunamis, landslides, etc., wildfire risk evaluation assessment should be done based on accurate referenced spatial information. It allows the establishment of appropriate preventive actions for the decision making in the risk management (Chen et al., 2003). The Geographic Information Systems (GIS) include functions for the input, output and/or graphic or cartographic information representation, data management and analytical functions (Bosque, 2000) therefore, are essential tools in fire risk assessment.

The risk term linked to wildfires is referred to the probability of a wildfire to happen in a given place and moment. It also considers the type and frequency of the causative agents. Fire risk indices can be classified in short and long-term indices (Chuvieco et al., 2003). Short-term fire risk indices are related to dynamic variables that affect fire danger (meteorological and fuel moisture content parameters), while long-term fire risk indices are related to variables that have an influence in the ignition/propagation of the fire, such as topography, fuel load, human activities or climatic patterns. Chuvieco et al. (2003, 2009) suggest that fire risk indices should include not only the probability of ignition but also the evaluation of potential damages (vulnerability). The probability of ignition is related to causative agents (human and natural) as well as the fuel conditions. The integration of these variables will lead to the wildfire risk.

The integration into operational fire risk systems of socioeconomic variables related to fire risk is not common. However, some works can be found in the literature that propose models for the fire occurrence prediction due to human causes using different techniques and at different scales. First works used models based on negative binomial (Bruce, 1963) and Poisson distributions (Cunningham and Martell, 1973). More frequently those works use logistic regression technique such as Martell et al. (1987), Lorenzo and Pérez (1995), Chuvieco et al. (1999, 2009), Martínez et al. (2004, 2009) or Prasad et al. (2008) at a regional level. At a local level, we find the works by Vega-García et al. (1995), Lin (1999), Pew and Larsen (2001), Vasconcelos et al. (2001). Other authors have proposed models using linear regression (Chao-Chin, 2002; Shypard et al., 2007), regression trees (Amatulli et al., 2006, 2007; Amatulli and Camia, 2007), neural networks (Vega-García et al., 2007) or Bayesian probability techniques (Romero-Calcerrada et al., 2008).

Due to the important role of the human action in wildfires in the Mediterranean areas and, more specifically, in Spain, this thesis proposes the elaboration of models to predict the wildfire occurrence due to human actions at 1km$^2$ grid cell resolution. This scale has been considered appropriate for fire management at regional scale. The work also proposes a methodology to integrate these models with those obtained to predict the fire occurrence due to lightning. The study areas have been the regions of Madrid, Valencia, Aragón and Huelva (Spain). Those areas were analyzed in the framework of the *Firemap* project (Chuvieco et al., 2009). Madrid has been the main study area due to data availability on historical fire occurrence. To reach the main goal, the work has been elaborated in four steps or specific objectives. The proposed methodology and obtained results are described in four chapters. Each chapter has lead to an accepted and/or in revision publication in national or international journals. Three of the chapters are written in English in order to achieve the "European Doctorate" and to get more dissemination of the obtained results. In the following section it is described each chapter, summarized in Table 1:

*i.* **Chapter I.** Using GIS tools, socioeconomic variables have been generated in the study areas from statistical and cartographic information sources. Applying logistic regression analysis human-caused wildfire occurrence models have been obtained. As response variable to elaborate these models has been used the human fire occurrence (National Fire Records) spatially mapped using kernel adaptative interpolation techniques. In *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2008) Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional (Boletín de la AGE, 47, 5-29)* the proposed methodology and the obtained results for the regions of Madrid and Valencia are shown. In both regions the global accuracy is similar, around 70%. Lower fire occurrence incidence gets better accuracy than higher. The selected explanatory variables in Madrid are related to the contact between urban and forest areas (wildland urban interface) whereas in Valencia are related to the variation of the population and/or the negligence actions when using the fire for agricultural or livestock activities.

In the first stages of this work there have been carried out comparative analysis of these logistic regression fire occurrence models with those obtained by using Decision Trees (CART algorithm) and Neural Networks in the regions of Madrid and Huelva. The results were presented to the *IV International Conference of Wildland Fire* in Seville -Spain- (2007), available at http://www.fire.uni-freiburg.de/sevilla-2007/contributions/doc/cd/SESIONES_TEMATICAS/ST1/VilardelHoyo_et_al_SPAIN.pdf. With the aim of showing alternatives to logistic regression techniques explored in the first steps of the research, the results are presented in the Annex I. Those techniques offered similar results in comparison with logistic regression, selecting the wildland-urban interface variable in Madrid and population variables in Huelva. In the region of Huelva the models obtained by using neural networks need further improvement. On one hand, neural networks need a more intensive preparation of input data. On the other hand, decision trees are recommended as exploratory data analysis for a set of variables. This explanatory analysis shows which variables have more influence in the response variable. Also, it shows the order of importance of each variable.

*ii.* **Chapter II.** Once the fire occurrence predicted models are obtained, the next objective is to test if a more accurate fire occurrence data would improve the models. We consider that the uncertainty in the spatial location of the ignition of fires is influencing the models. Therefore new models were obtained by using other response variables from different available fire data sources. The test was made in the region of Madrid, where more accurate fire ignition location information is available. In this step it has been used logistic regression, using as response variable fire ignition points from x, y coordinates. These results have been compared with those from a model where the response variable was calculated as a set of random fire ignition points within a 10x10 km polygon which is the spatial unit used in the Spanish official statistics to locate fire occurrence. The paper *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2009) Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables (Journal of Environmental Management, under revision)* shows the obtained results. Both models

differ in their accuracy and selected explanatory variables, as well as in the spatial distribution of the obtained probability. The model from the x, y coordinates of the real fire ignition points as response variable shows a better fit related to the fire occurrence. It is confirmed by an independent validation as well as with the experience of fire managers of the area studied. In the independent validation this model presents 75% of fit whereas the one from the set of random points, 65%. The wildland-urban interface is the variable with the highest influence.

*iii.* **Chapter III.** Once tested that a more accurate spatial location of the response variable leads to better models, the temporal dimension has been included. The aim is to test if the inclusion of dynamic variables improves the predictive models based on structural variables. Generalized Additive Models (GAMs) technique has been applied. As explanatory variables have been included a selection of socioeconomic data coming from the previous analysis. Also, dynamic variables (meteorological) have been included. As response variable x, y coordinates of fire ignition points have been used. This analysis has been carried out in the region of Madrid. The paper *Vilar, L., Wooldford, D., Martell, D., Martín, M.P. A Generalized Additive Model for Predicting People-Caused Wildfire Occurrence in the Region of Madrid, Spain ( International Journal of Wildland Fire, submitted )* shows the obtained results. The included explanatory variables are significant and show the expected trends. The obtained model reach a satisfactory fit, even tough it has been made with a short time fire occurrence data set. It has a spatial and a temporal dimension, allowing us to obtain fire occurrence predictions in a specific period of time.

*iv.* **Chapter IV.** Finally, it has been carried out the integration of human and lighting-cause fire occurrence models at 1km$^2$ grid cell resolution. Human-caused predictive models calculated from fire ignition points (x, y coordinates) in the regions of Madrid and Aragón have been used. These regions differ in their socioeconomic conditions and fire occurrence. To integrate probability values it has been used a probabilistic method and a weighted average by the historic fire occurrence. Both have been compared with a simple integration based on a single average. The paper

*Vilar, L., Nieto, H., Martín, M.P. Integration of lightning and human-caused wildfire occurrence models (Human and Ecological Risk Assessment, in evaluation)* shows the obtained results. To validate the results it has been used the Area Under the Curve (AUC) from Receiver Operating System (ROC) and Mahalanobis distance. In the region of Madrid the results are satisfactory whereas in Aragón would require a revision of the original models. In both regions the weighted average show better results. The spatial distribution of the integrated risk is coincident with fire cause distribution in both study areas.

**Table 1.** Summarized thesis organization

| Chapter | Specific objectives | Publication (*) |
|---|---|---|
| **I** | Socioeconomic variables generation and elaboration of predictive models of human fire occurrence | Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2008). Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional (Boletín de la AGE, 47, 5-29) |
| **II** | Elaboration of predictive models of human fire occurrence from accurate spatial location of fire origin | Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J (2009). Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables (Journal of Environmental Management, under revision) |
| **III** | Elaboration of predictive models of human fire occurrence including temporal component | Vilar, L., Wooldford, D., Martell, D., Martín, M.P. A Generalized Additive Model for Predicting People-Caused Wildfire Occurrence in the Region of Madrid, Spain (International Journal of Wildland Fire, submitted) |
| **IV** | Integration of predictive models of human and lightning fire occurrence | Vilar, L., Nieto, H., Martín, M.P. Integration of lightning and human-caused wildfire occurrence models (Human and Ecological Risk Assessment, under revision) |
| **Annex I** | Comparison of techniques to elaborate predictive models of human fire occurrence | Vilar del Hoyo, L., Gómez Nieto, I., Martín Isabel, M.P., Martínez Vega, F.J (2007). Análisis comparativo de diferentes métodos para la obtención de modelos de riesgo humano de incendios forestales. Wildfire 2007, 4th International Wildland Fire Conference, 13-17 Mayo 2007, Sevilla, España |

(*) Lara Vilar is the first author in all the publications derived from the thesis and has been responsible for the methodological proposals, elaboration and validation of spatial analysis and statistical models and also in the elaboration of the manuscripts. M. Pilar Martin is the director of the thesis and co-author in the publications participating in the methodological proposal, results discussion and manuscripts revision. In Chapter III Douglas Wooldford has been in charge of the programming development and the revision of the manuscript. In Chapter IV, Héctor Nieto has contributed to methodology, results discussion and the revision of the manuscript. In Annex I, Israel Gómez contributed to the neural network analysis. Javier Martínez Vega is co-author in papers I and II and his contribution was in the theoretical analysis, results discussion and revision of the manuscripts. David Martell is co-author in paper III and his contribution was in the results discussion and revision of the manuscript.

The presented work propose spatial models of fire occurrence due to human causes in different study areas in Spain, at 1km$^2$ grid cell resolution. There have been tested different response variables with the aim of reducing the spatial location uncertainty. *GAMs* models have been proposed to include the temporal dimension in fire risk assessment. Previous works have used this technique to obtain fire risk models by using meteorological variables, danger fire indices or topography (Brillinger et al. 2003, 2006; Preisler et al. 2004, 2007, 2008), so it would be innovative to add socioeconomic variables. Finally, it has been proposed methods to obtain integrated fire risk ignition indices. Methodological contributions are original and the results are relevant both from the scientific and also from the social perspective taking into account their potential application to fire risk management.

Future work in this research line would include further exploration of predictive models in order to improve estimation accuracy. We suggest applying *GAMs* in other study areas, to explore other spatial resolutions, to apply spatial regression techniques or to analyse changes in the land use related to fire causality as future lines of research. These proposals are described in the future research lines section.

**Referencias**

Amatulli, G., Camia, A., 2007. Exploring the relationships of fire occurrence variables by means of CART and MARS models. Wildfire 2007. IV International Wildfire Conference, Seville, Spain, 13-17 May.

Amatulli, G., Pérez-Cabello, F., de la Riva, J., 2007. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. Ecological modelling 200, 321-333.

Amatulli, G., Rodrigues, M.J., Trombetti, M., Lovreglio, R., 2006. Assessing long-term fire risk at local scale by means of decision tree technique. Journal of Geophysical Research 111, G04S05, doi:10.1029/2005JG000133.

Bachmann, A. 2001. GIS-based Wildland Fire Risk Analysis. Mathematics. Zurich, Universidad de Zurich.

Bruce, D. 1963. How many fires?. Fire Control Notes 24(2), 45-50.

Bosque, J. 2000. Sistemas de Información Geográfica. Madrid, Ed. Rialp S.A. 451 pp.

Brillinger, D.R., Preisler, H.K., Benoit, J.W. 2003. Risk assessment: a forest fire example. In DR. Goldstein (Ed.), Science and Statistics. A Festschrift for Terry Speed. Beechwood, OH: Institute of Mathematical Statistics Lecture Notes 40, pp.177-196.

Brillinger, D.R., Preisler, H.K., Benoit, J.W. 2006. Probabilistic risk assessment for wildfires. Environmetrics 17, 622-633. doi: 10.1002/env.768.

Chao-Chin, L., 2002. A Preliminary Test of A Human caused Fire Danger Prediction Model. Taiwan Journal Forest Science 17(4), 525-529.

Chen, K., Blong, R., Jacobson, C. 2003. Towards an Integrated Approach to Natural Hazards Risk Assessment Using GIS: With Reference to Bushfires. Environmental Management 31 (4), 546-560.

Chuvieco E., Salas, F.J, Carvacho, L., Rodríguez Silva, F., 1999. Integrated fire risk mapping. In Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin .Berlin: Springer-Verlag, pp. 61-84.

Chuvieco, E., Allgower, B., Salas, J. 2003. Integration of Physical and Human Factors in Fire Danger Assessment. In Chuvieco, E. (Ed.), The role of remote sensing data. New Jersey, US: World Scientific Publishing, vol. 4, pp. 197-218.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M.P., Vilar, L., Martínez, J., Martín, S., Ibarra, P., De la Riva, J., Baeza, J., Rodríguez, F., Molina, J.R., Herrera, M.A.,

Zamora, R. 2009. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling (In press).

Cunningham, A.A. and Martell, D.L. 1973. A stochastic model for the occurrence of man-caused forest fires. Canadian Journal of Forest Research 3(2), 282-287.

Food and Agricultural Organization of the United Nations (FAO), 1986. Wildland fire management terminology. FAO Forestry Paper 70. Rome, Italy. 257 pp.

Food and Agricultural Organization of the United Nations (FAO), 2002. Forest fire statistics 1999-2001. ECE/TIM/BULL/2002/4. Volume LV.No.4. Available at http://www.unece.org/timber/ff-stats.html (Last accessed December 23, 2008).

Food and Agricultural Organization of the United Nations (FAO), 2007. Fire Management Global Assessment. A thematic study prepared in the framework of the Global Forest Resources Assessment 2005. FAO Forestry Paper 151. Rome, Italy. 320 pp. Available at http://www.fao.org/forestry/fra2005/en/ (Last accessed December 23, 2008).

Lin, C., 1999. Modelling probability of ignition in Taiwan Red Pine Forests. Taiwan Journal.

Lorenzo, M.C., Pérez, M.C. 1995. Modelos de probabilidad para el estudio de la ocurrencia de incendios forestales. In 'IX reunión ASEPELT España. (Santiago de Compostela, Spain).

MARM. Ministry of the Environment, Rural and Marine Affairs. *Subsecretaría General de política forestal y desertificación. Área de defensa contra incendios forestales*. 2006. Los incendios forestales en España. Decenio 1996-2005. Available at http://www.mma.es/portal/secciones/biodiversidad/defensa_incendios/estadisticas_incendios/pdf/estadisticasdecenio_1996-2005.pdf (Last accessed December 23, 2008).

Martell, D.L., Otukol, S., Stocks, B.J., 1987. A logistic model for predicting daily people caused forest fire occurrence in Ontario. Caumar.

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

Martínez, J., Vega-García, C., Chuvieco, E., 2009. Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90, 1241–1252.

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, 1-18.

Prasad, V.K., Badarinath, K.V.S., Eaturu, A., 2008. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. Journal of Environmental Management 86, 1, 1-13.

Preisler, H.K., Brillinger, D.R., Burgan, R.E., Benoit, J.W. 2004. Probability based models for estimation of wildfire risk. International Journal of Wildland Fire 13, 133-142.

Preisler, H.K., Chen, S., Fujioka, F., Benoit, J.W., Westerling, A.L. 2008. Wildland fire probabilities estimated from weather model-deduced monthly mean fire danger indices. International Journal of Wildland Fire 17, 305-316.

Preisler, H.K., Westerling, A.L. 2007. Statistical model for forecasting monthly large wildfire events in western United States. Journal of Applied Meteorology and Climatology 46, 1020-1030.

Romero-Calcerrada, R. N., C. J., Millington, J. D. A., Gomez-Jimenez I., 2008. GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). Landscape Ecology 23, 341-354.

Salas, J., Cocero, D. 2004. El concepto de peligro de incendio. Sistemas actuales de estimación del peligro. In Chuvieco, E., Martín, M.P. (Eds.), Nuevas tecnologías para la estimación del riesgo de incendios forestales. CSIC, Instituto de Economía y Geografía, Madrid, pp.23-32.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B., 2007. Human influence on California Fire Regimes. Ecological Applications 17, 5, 1388-1402.

Vasconcelos, M.P.P., Silva, S., Tomé, M., Alvim, M., Pereira, J.M.C., 2001. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. Photogrammetric Engineering and Remote Sensing 67 (1), 73-81.

Vega García, C., Woodard, P.M, Titus, S.J., Adamowicz, W.L., Lee, B.S., 1995. A Logit Model for predicting the Daily Occurrence of Human Caused Forest Fires. International Journal Wildland Fire 5 (2), 101-111.

Vega-García, C., 2007. Propuesta metodológica para la predicción diaria de incendios forestales. Wildfire 2007. IV International Wildfire Conference, Seville, Spain, 13-17 May.

# Capítulo I

## Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional

**Publicación derivada:** *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J. 2008. Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional. Boletín de la AGE, 47, 5-29*

# Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional

**Extended Abstract:** ***Application of logistic regression techniques to obtain human wildfire risk models at regional scale***

## 1. Introduction

Wildfires in Spain are related to Mediterranean meteorological conditions but also to the human activity, which is responsible of more than 96% of fires (DGB, 2006). Traditionally, Mediterranean wildfires have had an ecological significance, to the point in which some ecosystems can be explained by wildfire occurrence. However, in the last decades the historical equilibrium that explains the relationship between fires and Mediterranean ecosystems has been broken (Martínez et al., 2004). Some authors have found that the fire occurrence is less severe in southern African Mediterranean countries than in the northern Mediterranean counterpart, even though the former have the same or worse climate conditions and the same kind of vegetation. Therefore, these differences could be explained because of the different socioeconomic conditions (Estirado and Molina, 2005). In recent times major changes have taken place in the European area. Socioeconomic, cultural and political changes have brought important economic, production and social transformations in rural areas (Moyano, 2007). Although 52% of Spain's surface is forested (MAPA, 2004) its contribution to the national economy is only 0.15% (Vélez, 2005). In addition, the spread of residences into rural areas has increased the wildland urban interface. Besides, there are new uses for forested areas, such as recreational activities (Izquierdo, 2007). These changes imply new ecological problems and most of all (despite the investment in fire management) an increase in fire risk. Therefore, it is necessary to improve fire prevention tasks. For this purpose, it is essential to know the fire causes. Fire records, which have been compiled in Spain since 1968, show that over 90% of fires are human related. These fire records classify the typology of fire causes into *lightning*, *negligence*, *deliberate*, *others* and *unknown*. The motivation behind deliberate fires, which account for over 59% of the total from 1991 to 2004, is also recorded. In contrast, less than 4% of fires were due to natural causes (lightning) during the same period (DGB, 2006).

Given the serious consequences of wildfires, it is important to study the wildfire risk in order to improve the prevention systems or actions. One of the most complete approaches to the analysis of fire risk includes three components: ignition, behaviour and vulnerability (Chuvieco et al., 2004). This is the conceptual framework of the *Firemap* project 'Integrated Analysis of Wildland Fire with Remote Sensing and GIS' (CGL2004-06049-C04-02/CLI). The outcome of *Firemap* will be a fire risk index which will integrate the human and natural factors related to fire ignition. This paper focuses on the analysis of the human factors.

The human component in fire risk is difficult to model since it involves attempting to quantify and to map human behaviour. However, it is possible to make some relevant approximations based on the analysis of explanatory variables that allow us to represent socioeconomic factors. These factors have a direct or indirect influence on fire occurrence and they are related to (Leone et al., 2003):

- socioeconomic changes
- traditional activities in rural areas
- accidents or negligence
- fire prevention activities
- deliberate fires because of conflicts

A wildfire risk model can be generated by analysing these factors. First, it is necessary to obtain input explanatory variables from cartographic or statistical sources. GIS tools are widely used to integrate the spatial information for wildfire risk analysis, and various authors (Chuvieco and Salas, 1994, 1996; Castro and Chuvieco, 1998; Gouma and Chronopoulou-Sereli, 1998; Pew and Larsen, 2001; Cardille et al., 2001) have applied GIS to work on predictive models that explain the phenomenon.

The present work examines how to obtain human wildfire risk models from socioeconomic explanatory variables. Logistic regression techniques are used to generate predictive models at 1 $km^2$ grid level in the regions of Madrid and Valencia. The aim is to integrate these models in a complex risk system that includes other factors (vegetation and climate) related to fire occurrence. The specific objectives are:

- to identify and to represent the explanatory variables from each factor
- to define the variable response (fire occurrence due to human causes)
- to apply a statistical exploratory analysis of the explanatory variables

- to propose, generate and validate a logistic regression model

**2. The study areas**

The study areas are the regions of Madrid and Valencia. The former is located in central Spain and although it represents only 1.6% of the nation's surface, it is one of the most densely populated areas with more than 6 million inhabitants. Urban areas have increased in the last decades, spreading into agricultural and forested areas. The contact boundary between urban and forest areas (wildland urban interface) is one of the main concerns for fire managers. Despite the large urban sprawl, Madrid still keeps an important network of protected areas which have an intensive recreational use, and are therefore, very vulnerable to forest fires. The main causes of fire in the region of Madrid from 1990 to 2004 were attributed to *unknown* causes followed by negligence.

The region of Valencia suffers wildland fires every year due to its extreme climate conditions, rugged terrain and human pressure. It has a Mediterranean climate where severe summer droughts are followed by a maximum rainfall during the autumn. For the fire phenomenon it is very important to take into account the wind factor (Ferrando, 2004). Wildfires in the last 20 years have decreased the tree formations which have been replaced by more inflammable shrubland. Society has been using fires as a traditional tool, mainly to dispose of agricultural waste. Tourism in this region is very important, and many visitors arrive every year. From 1990 to 2004 the main causes of fire were attributed to *negligence*, followed by *deliberate* and *lightning*. It is remarkable the low percentage of *unknown* fires, which are fewer than 10% since 1995.

**3. Methods**

To obtain the predictive models first of all it is necessary to generate the explanatory variables which represent the human factors in space. Explanatory variables are structural variables, related to permanent elements in the territory. For each human factor group, explanatory variables have been spatially mapped from cartographic or statistical sources using

GIS tools, and represented on a 1km² UTM grid which has been considered appropriated by fire experts to be operationally useful for fire management at the regional level.

The response variable (fire occurrence caused by human activities) has been obtained from the Spanish fire database (DGB, Ministry of Environment). In this database, the spatial location of fire records are entered both at municipal and 10×10 km grid level. Therefore, since the exact position (x, y coordinates) of the fire ignition points remains unknown, a methodology has been applied in this work to reduce its inaccuracy. This method is based on the kernel density estimation approach applied by Amatulli et al. (2007) to map lightning/human-caused wildfires under ignition point location uncertainty. The final goal here is to have an approximation on fire location at the 1 km² grid resolution selected for the model. An additional problem in the region of Madrid was the high percentage of unknown fires. To avoid the loss of information, unknown fires in Madrid were proportionally assigned to human or natural causes.

Logistic Regression techniques require a dichotomous response variable (0, 1 values). For that reason, the previous response variable has been transformed into a dichotomous one. The continuous variable has been ranged and divided into three groups with the same number of records. The records in the first group (low fire occurrence) have been assigned as zero, and the records in the last group (high fire occurrence) have been assigned as one.

The result of the logistic regression indicates the probability of fire occurrence and also the relationship between the response and the explanatory variables. It is defined as follows:

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

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_\rho X_\rho \tag{2}$$

Where $P_i$ is the probability of a fire occurring in a grid cell, $z$ is the linear combination of the explanatory variables weighted by their regression coefficients ($\beta$) and $X$ the value of and independent variable for any cell (Afifi and Clark, 1990 in Pew and Larsen, 2001; McGrew and Monroe, 1993).

Logistic Regression techniques assume that, when irrelevant or highly intercorrelated explanatory variables are present, the model cannot distinguish which part of the response variable is explained by each explanatory variable (Villagarcía, 2006). This is what is known as multicolineality, which has been avoided by previously carrying out an explanatory analysis aimed at excluding unnecessary variables.

The *Wald step forward* Logistic Regression method was applied. A random sample of 60% of the grid cells was used to generate the model and the remaining 40 % was used to validate it. The model was then applied on 100% of the sample, thus obtaining the probability for each grid cell. Cells with more than 50% of urban area were not included in the analysis.

**4. Results**

Out of the 17 models obtained in the region of Madrid, number seven was selected for having the best balance between the level of complexity (number of independent variables) and the prediction ability. The percentage of correct predictions in the final model was 70.6%, of which 75.4% was for low fire occurrence and 65.7% for high fire occurrence. In the region of Valencia the total percentage of correct predictions in the model was 68.4%, where 79.4% was for low fire occurrence and 57.4% for high fire occurrence. In the region of Madrid the variables with the highest influence in the model were the *Wildland Urban Interface* and the *Natural Protected Areas*, followed by *unemployment rate* and *buffer of paths*. In the region of Valencia variables with the highest influence were the *variation of the population* and the *demographic potential*.

Over and underestimations on the prediction of fire occurrence were mapped. In the region of Madrid underestimations are located in the north, north-east and south-east (*Sierra de Madrid*, *Alcalá de Henares* and *Aranjuez*). The overestimation is located in the central and south-western areas. The highest probability values are located to the west (*Sierra de Madrid*) and the south-east (a Natural Protected Area called *Parque Regional del Sureste*). The lowest probability values are located in the central and eastern areas. In the region of Valencia the underestimation areas are in the North and in the South and also in some West areas. The high fire occurrence is well predicted in the South-East area close to *Alicante*. The highest probability values are located

in the areas close to the Mediterranean Sea, matching up with the most inhabited areas. The lowest fire probability is located inside the region.

**5. Discussion and conclusions**

The correct prediction percentage of the fire occurrence is near 70% in both study areas. Low fire occurrence is better predicted also in both Valencia (57.4%) and especially in Madrid (76.4%).

Experienced fire managers agree with results concerning the explanatory variables obtained from the model in Madrid. The fire phenomenon in this region is related to negligence or accidents that happen in the Wildland Urban Interface (WUI), recreational areas, roads, etc. The results of the model match the characteristics of these land uses. In the region of Valencia the explanatory variables that explain the phenomenon are related to the population. Other important factors are changes in the forested area index and the pasture-urban interface. In the region of Valencia, as well as in other Mediterranean areas, there has been a major urban sprawl mainly boosted by tourism, which is more intense during the summer, when extreme meteorological conditions favour fire ignition. The highest probability values were predicted in the most populated areas. However, the model does not consider the WUI variable to be relevant. The results of prediction are less accurate than in the region of Madrid, maybe due to the bigger size of the region of Valencia, as well as its greater socioeconomic differences. It could be done regional sub models for this area.

By applying logistic regression, different kinds of explanatory variables can be included. The calculation of the response variable implies the uncertainty of the location of the fire ignition points and a continuous surface of fire density applied. This procedure could influence the final results. It would be interesting compare this information with the real fire ignition points and also to have more detailed spatial information about some explanatory variables. In any case it is useful to obtain results of fire probability in order to integrate them in a general risk system as the one proposed in the *Firemap* project.

Despite its limitations, this technique can be applied to different study areas so long as the socioeconomic variables fit the factors related to fire occurrence. Obtaining prediction models for fire occurrence could be very useful for the fire risk managers since it allows

identifying high fire occurrence areas and what kinds of variables have an influence on the wildfire. This work shows the importance of land use distribution and how the fire phenomenon is influenced by human activity. Finally, it shows the relevance of including socioeconomic factors in fire risk prevention systems in general.

# Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional

**Resumen**

Se aborda la realización de modelos de riesgo humano de incendio forestal mediante el empleo de técnicas de regresión logística, estimando la probabilidad de ocurrencia del fenómeno a partir de variables de tipo socioeconómico relacionadas con la ocurrencia de incendios forestales en las Comunidades Autónomas de Madrid y Valencia. Las variables independientes de riesgo se generan a partir de herramientas de Sistemas de Información Geográfica (SIG), a una resolución de 1km$^2$.

Palabras clave: Incendio forestal, Regresión Logística, Riesgo humano

**Abstract**

*Application of Logistic Regression techniques to obtain human wildfire risk models at regional scale*

The objective of this work is to develop human wildfire risk models using Logistic Regression techniques, estimating the probability of occurrence from socioeconomic explanatory variables in the regions of Madrid and Valencia. The explanatory variables are generated thanks to GIS at 1km$^2$ grid level.

Keywords: Forest fire, Logistic Regression, Human wildfire risk

## 1. Introducción

Los incendios forestales pueden definirse como el fuego que se propaga, sin control, en un sistema forestal y cuya quema no cumple funciones ni objetivos de gestión, por lo que requiere trabajos de extinción. Es un suceso no deseado en el que se producen una serie de consecuencias económicas y ecológicas calificadas como daños y perjuicios (Salas y Cocero 2004 citando a Martínez, 2001). 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).

El fuego es un elemento propio de los ecosistemas mediterráneos, con efectos incluso positivos en un ciclo de recurrencia suficientemente largo. De hecho, buena parte de estos ecosistemas sólo se explican por una presencia recurrente del fuego (Martínez et al, 2004 citando a Moreno, 1989). Sin embargo, este equilibrio se ha roto en las últimas décadas al acortarse los ciclos de recurrencia (Martínez et al, 2004 citando a Vélez, 1986). Las transformaciones socio-económicas guardan estrecha relación con este fenómeno. Así, por ejemplo, diversos autores señalan que el problema de los incendios es mucho menor en la ribera Sur de los países del Mediterráneo, a pesar de que su clima es similar o incluso más severo que el de los países mediterráneos del Sur de Europa y la vegetación no es muy distinta, por lo que estiman que las diferencias se deben, fundamentalmente, a las condiciones socioeconómicas (Estirado y Molina, 2005). Según Vélez (2005), los factores condicionantes de esta situación en el conjunto de los países del arco norte del Mediterráneo son ecológicos (grandes períodos de sequía con alta inflamabilidad de la vegetación), económicos (baja renta del sector forestal), demográficos (éxodo rural) y políticos (atención de lo *urgente* –extinción- y no de lo *importante* –prevención-).

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, 2007). La superficie forestal española ha aumentado un 6% desde el período 1986-1995 (Segundo Inventario Forestal Nacional) al período 1997-2000 (Tercer Inventario Forestal Nacional), ocupando un 52% del territorio (MAPA, 2004). A pesar de este alto porcentaje, la contribución al PIB del sector forestal es tan sólo del 0,15% (Vélez, 2005). Por otro lado, se está produciendo una "urbanización de lo rural",

con una difusión de la ciudad hacia el territorio rural por medio de la urbanización, ampliándose la interfaz urbano-forestal. Igualmente se produce el desarrollo de nuevas actividades y usos en las zonas forestales, tales como el recreativo (Izquierdo, 2007). Estos cambios dan lugar a diversos problemas ecológicos. En el caso de los incendios forestales, buena parte de estos cambios han tenido como efecto inmediato un aumento del riesgo de incendios, además de crear las condiciones idóneas para su propagación (Martínez, 2004).

En las dos últimas décadas se han mejorado los recursos de extinción realizándose grandes inversiones y obtenido resultados aparentemente aceptables. Sin embargo, el problema de los incendios sigue e incluso se agrava. Según muestran las estadísticas, la tendencia en el número de incendios es creciente (Estirado y Molina, 2005). Resulta, por tanto, necesario dar un nuevo enfoque para mejorar las estrategias de gestión potenciando especialmente las labores de prevención.

Para poder llevar a cabo unas adecuadas labores de prevención es necesario conocer las causas de los incendios forestales. Estas se pueden dividir en dos grupos: estructurales si no inician el incendio pero incrementan el riesgo de que se produzca, e inmediatas si provocan el inicio del incendio (INFOCA, 2006). Como se ha citado anteriormente, las estadísticas de incendio forestal que se recogen en España desde 1968 (Base de Datos de Incendios Forestales-BDIF), muestran que el factor humano explica más de un 90% de los incendios que se producen en nuestro territorio. Los Partes de Incendio constituyen actualmente una valiosa fuente de información para la interpretación del fenómeno de los incendios en España. Estos partes se han ido modificando, enriqueciéndose y adecuándose a las necesidades marcadas por la evolución del fenómeno de los incendios y de los medios de detección y extinción disponibles (Martínez, 2004). Por lo que respecta a la causalidad, el modelo actual de Parte recoge información sobre los siguientes grupos de causa de incendio: rayo, negligencias, intencionado, desconocido, reproducido y otras causas. Asimismo, para los incendios de tipo intencionado se registra el tipo de motivación, entre los que se encuentran, por ejemplo, incendios provocados por agricultores, ganaderos, pirómanos, etc. De igual forma también se recoge información sobre el lugar de inicio del incendio (junto a caminos, pistas, vías férreas, entre otros).

En la Figura 1 se muestra la distribución de causas en el conjunto de España en el período 1996-2005 según datos de la DGB (2007). Como se puede observar, el porcentaje más alto corresponde a los incendios intencionados (60,04%). Las motivaciones de estos incendios

intencionados se desconocen en más de un 50%. De los que sí se tiene un conocimiento cierto de su origen, son las quemas agrícolas ilegales y abandonadas (42,96%) y las quemas para la regeneración de pastos (30,86%) los que ocupan los primeros lugares. Las negligencias son la segunda causa en importancia (17,57%). En cuanto a los incendios causados por rayo (la única causa natural de incendio en nuestro país) no alcanzan el 4%.

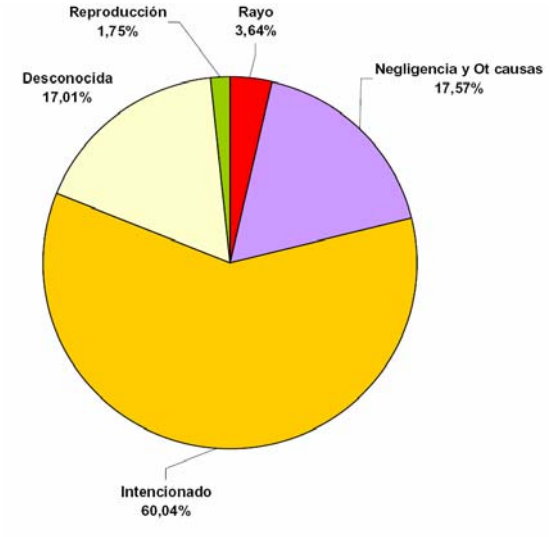



**Figura 1.** Distribución de causas de incendio forestal (%) en España. 1996-2005.
*Fuente: Los Incendios Forestales en España. Decenio 1996-2005. Área de Defensa Contra Incendios Forestales de la Dirección General para la Biodiversidad*

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 se encuentra el que estructura el mismo en tres componentes relacionados con el inicio de fuego, la propagación y los daños que produce en el medio (Chuvieco et al., 2004). Este planteamiento es objeto de estudio en el marco del Proyecto *Firemap:* "Análisis Integrado de Incendios Forestales mediante Teledetección y Sistemas de Información Geográfica" (CGL2004-06049-C04-02/CLI)[1]. En este proyecto se propone un esquema de integración de las variables

[1] 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

relacionadas con el riesgo de ignición, propagación y la vulnerabilidad (Figura 2). En este esquema se plantea la necesidad de obtener un índice de causalidad a partir del análisis de los factores humanos y naturales (rayo) que pueden desencadenar el inicio del fuego. Es precisamente en el estudio de los aspectos relacionados con la causalidad humana en el que se centra el presente trabajo.

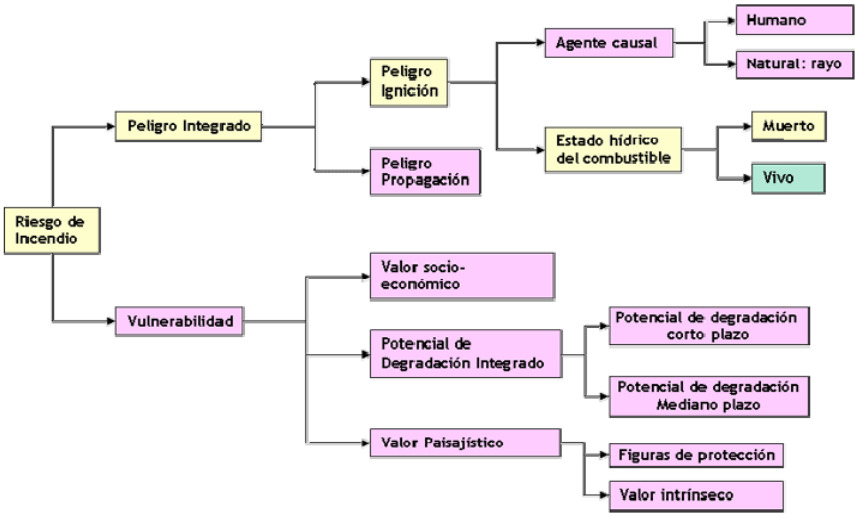



**Figura 2.** Esquema de obtención del Riesgo Integrado de incendio forestal.

*Fuente: Proyecto Firemap*

El componente humano del riesgo de incendio es difícil de modelar, debido, principalmente, a las dificultades para cuantificar y espacializar determinadas particularidades del comportamiento humano. No obstante, se pueden realizar aproximaciones interesantes a partir del análisis de ciertas variables o indicadores que nos permitan representar los factores de tipo socioeconómico que pueden influir directa o indirectamente en la ocurrencia de incendios. Algunos autores clasifican estos factores según el siguiente esquema (Leone et al., 2003):

- Factores relacionados con transformaciones socioeconómicas (abandono de actividades tradicionales, cambios demográficos, urbanización, nuevas actividades recreativas)

- Factores relacionados con actividades tradicionales en áreas rurales (población rural envejecida, quemas agrícolas y ganaderas)

- Factores que pueden causar incendios por accidente o negligencia (líneas eléctricas, vías de comunicación)

- Factores de disuasión frente a la ignición (recursos de extinción, medios de vigilancia)

- Factores que generan conflictos y que pueden desembocar en el inicio intencionado de incendios o facilitar su propagación (cambios de uso, declaración de zonas protegidas, conflictos en la propiedad forestal, venganzas contra la administración, "industria del fuego")

A partir de estos factores y de cara a generar un modelo espacializado del riesgo, es preciso obtener variables de tipo cartográfico o estadístico que permitan su representación. El manejo de toda esta información es posible gracias al empleo de herramientas de Sistemas de Información Geográfica (SIG). Son numerosos los estudios de riesgo de incendio que emplean el SIG para la integración espacial de variables, teniendo en cuenta las relaciones geográficas y analíticas entre los datos. Entre estos se encuentran los llevados a cabo por Chuvieco y Salas (1994 y 1996), Castro y Chuvieco (1998), Gouma y Chronopoulou-Sereli (1998), Pew y Larsen (2001), Cardille et al. (2001), entre otros. Estos trabajos tratan de generar modelos explicativos y fundamentalmente predictivos que permitan estimar la probabilidad de ocurrencia de incendios forestales a partir del análisis de un conjunto de variables. La probabilidad de ocurrencia de un incendio varía en el tiempo y en el espacio dependiendo de distintos factores de riesgo. Por este motivo los primeros modelos de riesgo humano se basaban en las distribuciones binomial y de Poisson, aptas para sucesos raros. Posteriormente, se profundiza en los factores de riesgo y en la explicación de la probabilidad de existencia del incendio mediante técnicas de regresión, principalmente regresión logística (Lorenzo y Pérez, 1995). Esta técnica permite describir las relaciones entre una variable dependiente nominal u ordinal y un conjunto de variables independientes continuas o categóricas, así como cuantificar las relaciones y clasificar. Diversos autores han empleado esta técnica para la obtención de modelos predictivos de ignición de incendio a escala regional, Chuvieco et al. (1999) y Martínez et al.

(2004); local, Vasconcelos et al. (2001), Vega-García et al. (1995), Lin (1999) (en Martín et al. 2002), Pew y Larsen (2001). Otros estudios utilizan esta técnica para la predicción de la ocurrencia diaria: Martell et al. (1985, 1987); Loftsgaarden y Andrews (1992) (en Martín et al. 2002); Vega-García et al. (1995).

## 2. Objetivos

El presente trabajo persigue la obtención de modelos predictivos de riesgo de incendios forestales a partir de variables de tipo socioeconómico, profundizando en el conocimiento de las relaciones entre factores humanos y la ocurrencia de incendios. Se propone el empleo de técnicas de Regresión Logística para generar modelos predictivos espacializados a una resolución de 1 km[2] en las Comunidades Autónomas de Madrid y Valencia[2]. El propósito es desarrollar modelos consistentes y espacialmente extrapolables que puedan integrarse con facilidad en un sistema de riesgo más complejo que incluya otros factores (vegetación, clima, etc.) relacionados con la ocurrencia de incendios forestales.

Para alcanzar este objetivo general se abordarán los siguientes objetivos específicos:

- Identificar y generar variables independientes que representen los diversos factores de riesgo de incendio vinculados a la actividad humana.

- Definir y generar la variable dependiente en el modelo: ocurrencia de incendios de causa humana.

- Proponer y aplicar los análisis estadísticos previos a la generación del modelo para la selección de las variables independientes a incluir en el mismo.

- Proponer, generar y validar el modelo de regresión logística binaria.

[2] La elección de esta unidad de análisis se basa en la demanda que los gestores vienen realizando para trabajar a un nivel local de riesgo

En este trabajo, el riesgo humano de ignición va a considerarse como un componente estructural, para una estimación del riesgo a largo plazo. La componente temporal del riesgo vendría matizada temporalmente por la variación de los factores del medio físico (rayos y humedad del combustible) en el modelo de riesgo integrado (Figura 2).

## 3. Material y métodos

### 3.1 Áreas de estudio

Los modelos de riesgo de incendio debidos a causa humana han sido elaborados para la C. Madrid y la C. Valenciana (Figura 3). El período de estudio comprende los años 1990 a 2004. Se ha elegido este período para asegurar la consistencia de los datos de incendios utilizados y para garantizar la robustez de los análisis estadísticos efectuados.

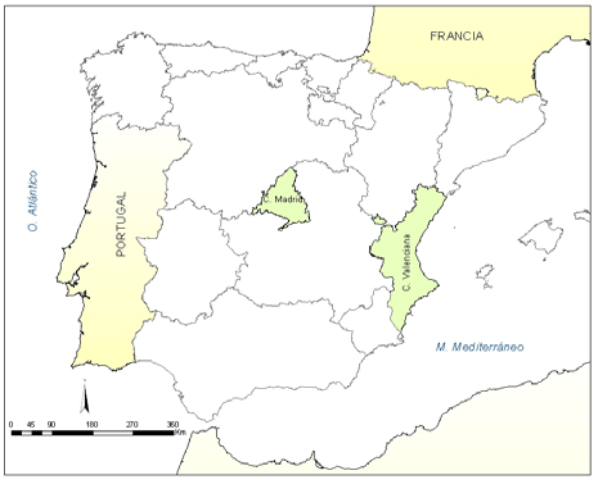



**Figura 3.** Áreas de estudio

#### 3.1.1 Comunidad de Madrid

La C. Madrid, con tan solo 8.027,9 km² (1,6 % de la superficie nacional), cuenta con una población de, aproximadamente, 6 millones de habitantes (Julio de 2007) (14% de la población total de nuestro país) distribuida en 179 municipios de características muy diversas. Se trata de la región más densamente poblada de España con unos 748 habitantes/km². En las últimas

décadas las áreas urbanas han experimentado un notable crecimiento en la región ocupando, primero, las áreas agrícolas vecinas y, después, las zonas forestales más distantes. La eficiente red de transportes de la región (más de 3.000 km de carreteras y aproximadamente 300 km de líneas de ferrocarril) ha contribuido a este proceso al facilitar la movilidad de la población entre zonas urbanas y peri-urbanas. Este hecho, unido al cambio en los modelos residenciales (predilección por zonas compuestas por agrupaciones de viviendas de baja densidad: casas, *chalets*), ha favorecido el crecimiento de áreas residenciales en zonas forestales tanto para uso recreativo (segunda residencia) como para vivienda habitual. Este modelo de urbanización ha dado lugar a que el contacto entre las áreas urbanas y las forestales (interfaz urbano-forestal) adquiera una gran importancia en la región y se convierta en una de las principales preocupaciones de los gestores contra incendios dado el alto riesgo al tiempo que eleva la vulnerabilidad que esta zonas presentan frente a este fenómeno (Leone et al., 2003 citando a Vélez). A pesar de su alto grado de urbanización la C. de Madrid cuenta con un gran número de espacios naturales preservados bajo diversas figuras de protección derivadas de la legislación estatal, autonómica y comunitaria (Parques naturales, regionales, ZEPAs, LICs, etc.). Estas zonas protegidas suponen aproximadamente el 13% del territorio regional (Consejería de Medio Ambiente y Ordenación del Territorio, Madrid) y son especialmente vulnerables al fenómeno de los incendios por la riqueza de ecosistemas, hábitat y especies vegetales y animales que albergan y también por el valor paisajístico y de uso recreativo que poseen.

En la C. de Madrid, la media de incendios por cada 10.000 ha de superficie forestal para el período 1991-2005 es de 6,7, siendo la media nacional 7,5 (WWF/ADENA, 2007). Las características especiales de alta densidad de población y uso recreativo de sus masas forestales la convierten en un área de especial interés para el estudio. En la Figura 4 se observa la distribución y evolución temporal de las principales causas de incendios en la región desde 1990 a 2004. Destaca el alto porcentaje de causas desconocidas y la notable proporción de incendios por negligencia.

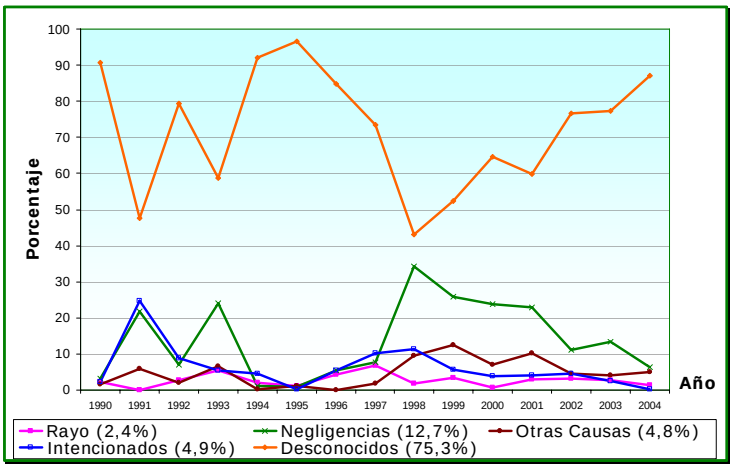



**Figura 4.** Tendencias de Incendios forestales según tipo de causa. C. Madrid. Período 1990-2004.

*Fuente: Elaboración propia a partir Partes de Incendio DGB 1990-2004*

Si se desglosan los incendios por estaciones del año y según la tipología detallada de causas, sin incluir los desconocidos, en primavera y otoño son los trabajos forestales la causa mayoritaria de incendios, mientras que en invierno es la quema de pastos y en verano los incendios intencionados.

### 3.1.2 Comunidad Valenciana

La Comunidad Valenciana, con una extensión de 23.255 km² (4,6% del total nacional), es la séptima comunidad española en superficie. Es un territorio especialmente castigado por los incendios forestales debido a sus características climáticas de extremo riesgo, su marcada orografía y la presión humana.

El clima de esta región es típicamente mediterráneo con una marcada sequía estival y un máximo pluviométrico otoñal. Climatológicamente existe en esta región un factor de capital importancia a efectos de la incidencia de incendios: los vientos terrales (ponientes). Si bien durante la época de peligro, coincidente con el verano, el régimen climatológico general es el establecido por los vientos de Levante o las brisas costeras que provocan una elevada humedad relativa del aire (50-70%) y vientos moderados; en determinadas ocasiones, la penetración por el oeste de frentes procedentes del Atlántico produce una situación bien diferente. En este caso llegan al este de la Península Ibérica masas de aire totalmente desprovistas de humedad y con

temperaturas muy elevadas, debido a efecto Foehn, que favorecen el inicio y la propagación de incendios (Ferrando, 2004).

La vegetación dominante en la región es típica de un ambiente mediterráneo. Se trata de una vegetación esclerófila cuya especie dominante es la encina con el roble y el alcornoque aunque también encontramos otras especies como el pino, que en la actualidad ocupa una importante superficie en la región debido a una intensa labor de repoblación que fue especialmente frecuente en las zonas afectadas por incendios. El matorral, de carácter termófilo, se debe a causas climáticas principalmente, pero también a la acción antrópica que favoreció áreas despejadas para la alimentación del ganado y a la frecuente ocurrencia de incendios. La reiteración de los grandes incendios en los últimos 20 años ha supuesto la disminución de las comunidades arboladas, sobre todo pinares, y un aumento de las comunidades dominadas por especies arbustivas de ciclo más corto y de elevada inflamabilidad (aulaga, romero, jaras). Los suelos han sufrido estas perturbaciones, presentando en algunas zonas riesgos altos de erosión y empobrecimiento (Ferrando, 2004).

La C. Valenciana tiene un total de 4.824.568 habitantes (Julio de 2007). La densidad de población en la provincia de Alicante es de 298 habitantes/km$^2$, en Valencia 224 habitantes/km$^2$ y en Castellón 82 habitantes/km$^2$ (INE, 2005). La población ocupada se dedica fundamentalmente al sector servicios (el 60,4% en 2001), trabajando el resto en la agricultura (4,1%), la construcción (11,4%) e industria (24,1%) (Plan General de Ordenación Forestal, 2004). La sociedad valenciana presenta un gran arraigo en el uso del fuego, y no tanto en actividades lúdicas, sino como herramienta tradicional de eliminación de residuos agrícolas (Suárez, 2000). El sector turístico juega un importante papel en la economía de este área de estudio, siendo visitada cada año, sobre todo en la época estival, por un gran número de turistas. En valores absolutos, en 2006, la región recibió a un total de unos 5 millones y medio de personas, frente a los 58 millones que recibió el conjunto de España (9% del total nacional aprox.) (INE, 2007).

La media de incendios por cada 10.000 ha de superficie forestal del período 1991-2005 en la C. Valenciana es de 4,6, menor que la media nacional 7,5 (WWF/ADENA, 2007).Las causas de incendio forestal en el período 1990-2004 se recogen en la Figura 5. Las causas más frecuentes son las negligencias, seguidos de los incendios intencionados y los producidos por rayo. Destaca el bajo porcentaje de incendios desconocidos (10,9%), el cual ha ido disminuyendo hasta ser inferior al 10% desde el año 1995.

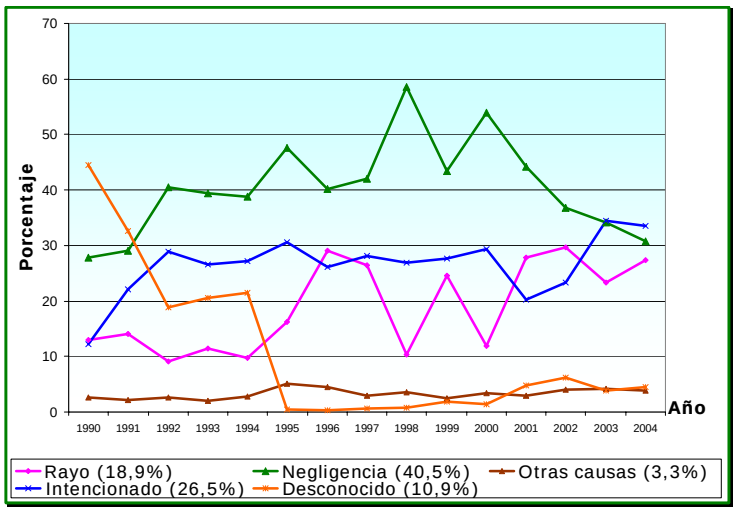



**Figura 5.** Tendencias de Incendios forestales según tipo de causa. C. Valenciana. Período 1990-2004. *Fuente: Elaboración propia a partir Partes de Incendio DGB 1990-2004*

En la C. Valenciana, si se desglosan los incendios por estaciones del año y por tipología detallada de causas, la mayoría de los incendios son de tipo intencionado en todas las estaciones salvo en el verano, estación en la que el porcentaje de intencionados es ligeramente inferior al de los incendios producidos por rayo (26% frente a un 27%). La segunda causa mayoritaria en primavera, otoño e invierno son las quemas agrícolas. Si no se desglosa la tipología de causas en todas las estaciones los incendios obedecen a negligencias (la suma de negligencias supera a los incendios de tipo intencionado).

## 3.2 Metodología

Para la obtención de los modelos de riesgo debidos a causa humana se ha empleado la metodología detallada en los apartados siguientes y recogida en síntesis en la Figura 6.

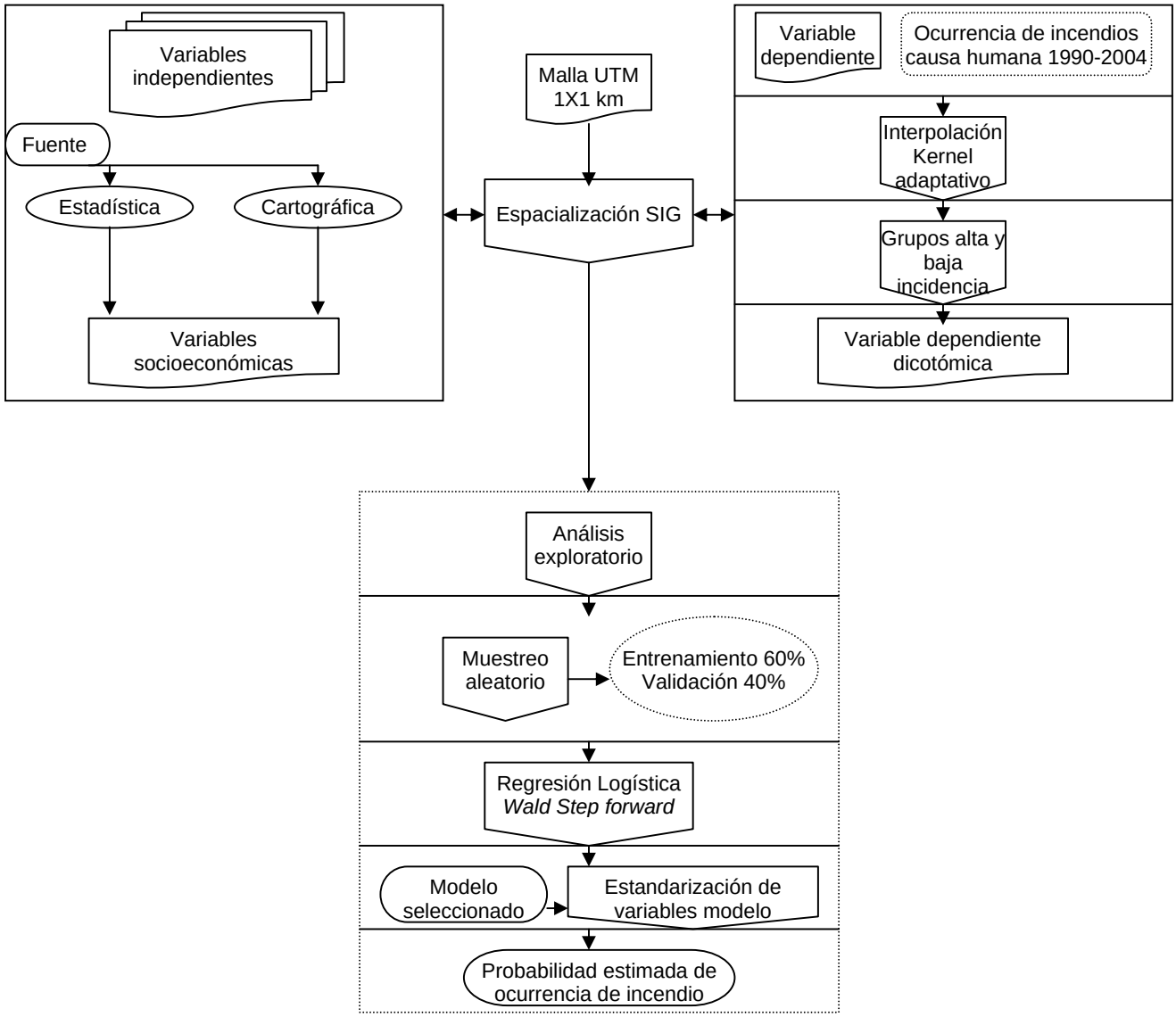



**Figura 6.** Diagrama de flujo de la metodología empleada en la obtención de la probabilidad de ocurrencia de incendio debido a causa humana

3.2.1 Generación de variables independientes

En primer lugar, se ha llevado a cabo la generación de las variables independientes que van a formar parte de los modelos predictivos. Estas variables deben ser representativas de los factores de riesgo vinculados a la actividad humana y, a su vez, permitir la cuantificación y representación en el espacio de los mismos. La identificación de estas variables se ha basado en

el análisis de fuentes bibliográficas especializadas (Leone et al., 2003, Martínez, 2004, Martínez et al., 2004, Pew y Larsen 2001, Vega-García et al., 1995), en las experiencias de proyectos a nivel local y regional de riesgo integrado de incendio forestal (*Firerisk, 2003; Spread, 2003; Megafires, 1998*) así como en la información obtenida en una encuesta a expertos realizada en el marco del proyecto *Firemap*. En general, se intentó considerar preferentemente aquellas variables de carácter estructural, relacionadas con elementos permanentes del territorio.

Se establecieron cinco 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 propuso una relación de variables que permitieran representarlos espacialmente de forma más o menos directa. Estas variables se elaboraron a partir de fuentes de información de tipo cartográfico y estadístico, representándolas espacialmente en la cuadrícula UTM de 1 km$^2$, mediante el empleo de herramientas SIG y de manejo de bases de datos y hojas de cálculo. La Tabla 1 recoge las variables independientes generadas dentro de cada factor de riesgo. El proceso de representación espacial ha variado en función de la fuente de información de cada variable definida; en el caso de las variables cartográficas éstas se han referido 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. Para las variables de tipo estadístico, al estar referidas a la unidad espacial de municipio, se intersectaron los polígonos de los municipios con la cuadrícula de 1 km$^2$, asignándole a todas las cuadrículas incluidas en cada municipio el mismo 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 el caso de variables definidas como *buffer*, se calculó un corredor de anchura variable en función de la distancia de seguridad establecida en las diferentes legislaciones de cada área de estudio, ya sea legislación en materia de incendios forestales ó de otro tipo, como por ejemplo, la relativa a las vías de comunicación. De igual modo, se han calculado las variables de *Interfaz*, estableciendo un corredor o área de influencia de la interfaz en cada cuadrícula, de una distancia definida en la legislación de cada área de estudio.

**Tabla 1.** Variables independientes de tipo socioeconómico

| Tipo | Factor | Nombre de la variable | Descripción |
|---|---|---|---|
| Cartográfica | **Accidente o negligencia** | b_carret | Buffer de carreteras |
| | | b_carret_for | Buffer de carreteras en zonas forestales |
| | | Indice_imd[3] | Índice de IMD[4] por segmento de carretera (Longitud vía × IMD vía × factor de ponderación) |
| | | Indice_imd _for | |
| | | b_ffcc | Buffer de vías de ferrocarril |
| | | b_ ffcc _for | Buffer de vías de ferrocarril en zonas forestales |
| | | b_pistas | Buffer de pistas |
| | | b_pistas_for | Buffer de pistas en zonas forestales |
| | | b_llee | Buffer de líneas eléctricas |
| | | b_llee_for | Buffer de líneas eléctricas en zonas forestales |
| | | Tiro_canteras | Campos de tiro y canteras |
| | **Transformaciones socioeconómicas** | Area_recre | Buffer de áreas recreativas ponderadas por presencia de barbacoa[5] |
| | | Pot_dem | Potencial Demográfico |
| | | ICC | Índice de cambio en superficie forestal<br>ICC = [(Cultivo a Forestal+ Improductivo a Forestal)-(Forestal a Cultivo+Forestal a Improductivo)] |
| | | IUF | Interfaz Urbano Forestal |
| | | vertederos | Buffer de vertederos |
| | | ICF | Interfaz Cultivo Forestal |
| | | IPF | Interfaz Pasto Forestal |
| | **Conflictos que pueden desencadenar incendios intencionados** | ENP | Espacios Naturales Protegidos |
| | | ZEPA | Zonas de Especial Protección de Aves |
| | | M_Preser_UP | Montes de Utilidad Pública y Preservados[6] |
| | | Consor | Montes Consorciados |
| | **Disuasión de la ignición** | Torres | Presencia o ausencia de torres o medios de vigilancia |
| Estadística | **Transformaciones socioeconómicas** | Var_pob | Variación de la población 1970-2004 |
| | | Hotel | Infraestructuras hoteleras |
| | | Var_pob_agra | Variación de la población agraria 1996-2001 (C. Madrid); 1999-2002 (C. Valenciana) |
| | **Actividades tradicionales en áreas rurales** | Jefes55 | Porcentaje de jefes de explotaciones agrícolas mayores de 55 años |
| | | Carga_gan | Carga ganadera<br>(número de cabezas ovinas y caprinas en superficie de pastos y matorral) |
| | | Maquina | Densidad de maquinaria agrícola |
| | **Conflictos que pueden desencadenar incendios intencionados** | Renta | Renta per capita |
| | | Tasa_paro | Tasa de Paro |

[3] Esta variable sólo ha sido calculada en la C. Madrid debido a la disponibilidad de la información de base requerida
[4] IMD: Intensidad Media Diaria de Tráfico
[5] En las C. Madrid y C. Valenciana ha sido llevada a cabo la ponderación por estar disponible el dato de presencia o ausencia de barbacoa
[6] La figura de montes Preservados es propia de la C. Madrid

### 3.2.2 Generación de la variable dependiente

La ocurrencia de incendios de causa humana en el período de estudio se obtiene a partir de los Partes de Incendio de la Dirección General para la Biodiversidad (DGB). Las estadísticas de incendio en España recogen los incendios ocurridos a nivel municipal y a nivel de cuadrícula de 10×10 km, por lo que no se conoce con exactitud la posición de los puntos de ignición[7]. Al ser la unidad de análisis de este trabajo la cuadrícula de 1 km$^2$ se ha aplicado un procedimiento para reducir la incertidumbre en la localización de los puntos de inicio del incendio. Para ello, se combina la información sobre la localización a nivel municipal con la localización por cuadrículas 10x10 km. De esta forma, se acota la localización de los incendios en polígonos de superficie inferior a la de las cuadrículas de referencia. Para afinar aún más esta localización se cruzan los polígonos resultantes con el mapa forestal, eliminando las zonas sin superficie forestal. Este proceso se aplica asumiendo que los incendios se inician en zonas forestales. De esta forma, se consiguen polígonos donde, *a priori,* la localización de los incendios es más precisa (Amatulli et al., 2007).

Teniendo en cuenta que se trata de generar un modelo predictivo de los incendios vinculados a actividades humanas, se incluyen en el análisis exclusivamente los incendios de causa humana en cada área de estudio. Dado que, en algunas regiones como la C. Madrid, el porcentaje de incendios de causa desconocida es muy elevado, se decidió asignar parte de los incendios desconocidos a causa humana, teniendo en cuenta, en cada zona de estudio, la proporción de incendios de causa conocida que corresponden a causa humana y a rayos. Al total de incendios desconocidos se le aplica dicha proporción, obteniéndose el número de incendios que se asignarían a rayos en cada área. Éstos se seleccionan de forma aleatoria de entre los incendios desconocidos. El resto de incendios desconocidos se asigna a causa humana. Para llevar a cabo la espacialización de la variable dependiente se han incluido en el análisis los incendios ocurridos en municipios limítrofes a las diferentes áreas de estudio. Este proceso tiene como objetivo reducir los posibles efectos de borde en la espacialización de dicha variable. Se obtienen finalmente 4.537 incendios de causa humana en la C. Madrid y 6.223 en la C. Valenciana para el período de estudio 1990-2004. A partir del número total de incendios

[7] Tan solo recientemente se han comenzado a recoger en los Partes datos relativos a las coordenadas de inicio del incendio. Estos datos son todavía escasos (la serie temporal es corta) y los datos no están disponibles para toda España.

referidos a una localización espacial más precisa de acuerdo al proceso anteriormente descrito, se generan puntos aleatorios de ignición mediante el *script* de *ArcView 3.2 Random Point Generator v. 1.3*[8]. Para obtener superficies continuas a partir de estos puntos de ignición se ha utilizado la técnica de interpolación de estimación de densidad de Kernel propuesta por De la Riva et al. (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 et al., 2003 citando a Seaman y Powell, 1996; Levine, 2004):

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

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 esta distribución, el *bandwidth* se corresponde con la desviación estándar de la misma. En cuanto al procedimiento para fijar el kernel, este puede ser fijo (*bandwidth* constante) o adaptativo (*bandwidth* varía dependiendo de la concentración de puntos) (Leone et al., 2003 citando a Worton, 1989). Este último procedimiento 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 et al., 2007). Debido a que los incendios no se distribuyen de manera regular, se emplea el modo adaptativo. El tamaño de intervalo de *bandwidth* establecido en cada zona de estudio para llevar a cabo la interpolación obedece a la elección del mismo según la minimización del *goodness-of-fit criteria* propuesto por Breiman *et al* (1977). Con este procedimiento se ensayan distintos órdenes de vecino próximo para dar con el que minimiza la curva de ajuste. En la C. Madrid es de 5 puntos, mientras que en la C. Valenciana es de 25 puntos. La interpolación se lleva a cabo con *Crimestat*© 3.0 (Levine, 2004). La Figura 7 muestra el resultado de la interpolación, en cada zona, que será utilizado como variable dependiente en el modelo.

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

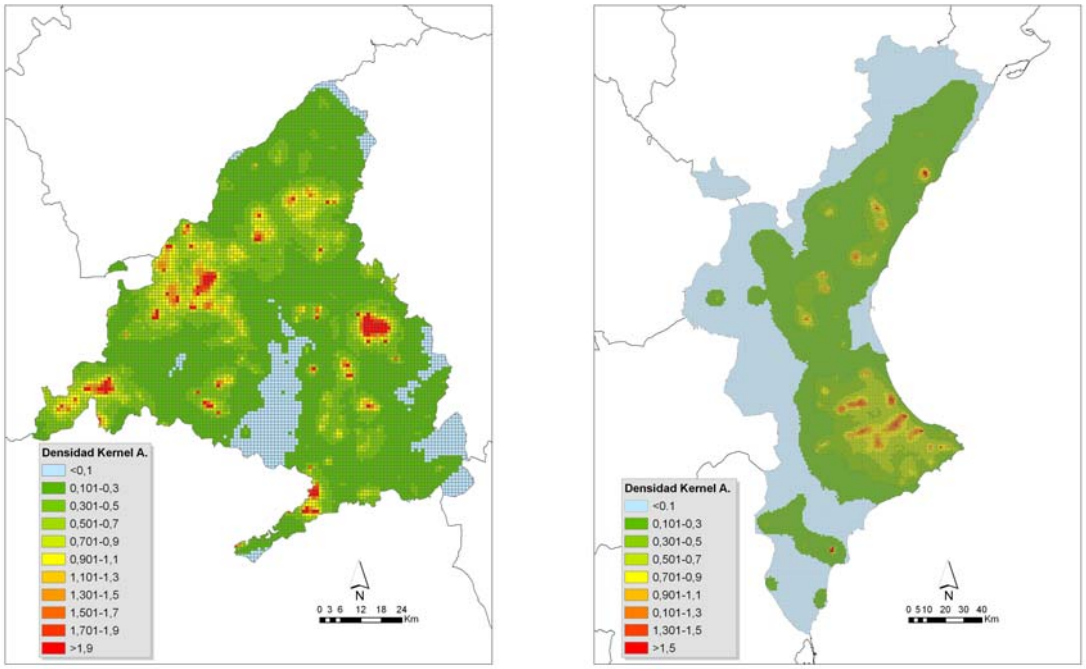



**Figura 7.** Interpolación Kernel Adaptativo. Variable dependiente continua ocurrencia de incendios debidos a causa humana

El método de regresión logística requiere una variable dependiente dicotómica. Así pues, fue necesario transformar la variable continua (número de incendios de causa humana) a dicotómica. Esto se hizo dividiendo la variable ordenada en 3 grupos con el mismo número de casos (grupo 1, cuadrículas con baja incidencia, grupo 2 de incidencia intermedia y 3 de alta incidencia). 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.

### 3.2.3 Generación de los modelos

El método de regresión logística empleado ha sido utilizado en análisis anteriores para la estimación de la ocurrencia de incendios forestales a escalas regionales y locales, obteniéndose modelos predictivos y explicativos, al conocer las variables de mayor importancia en el fenómeno (Carvacho, 2002).

El objetivo que se persigue con la aplicación de este modelo 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, 2006). 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:

$$P_i = \frac{1}{1+e^{-z}} \tag{2}$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_\rho X_\rho \tag{3}$$

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

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

$$\log\left(\frac{p}{1-p}\right) = x^T \beta \tag{4}$$

Siendo $x^T$ el vector de las variables explicativas y $\beta$ el vector de los parámetros (González, 2006).

La regresión logística tiene una serie de asunciones (Garson, 2006), como las de no asumir relación lineal entre la variable dependiente y las independientes. Por otra parte, la variable dependiente no necesita seguir una distribución normal así como ser homocedástica (homogeneidad de la varianza). Si se incluyen variables irrelevantes en el modelo, la varianza que comparten puede ser atribuida de forma errónea a estas variables irrelevantes. Esta técnica asume que los términos de error son independientes y no tiene en cuenta los efectos de

interacción entre las variables; no ha de darse multicolinealidad. Cuando las variables independientes tienen mucha relación entre si el modelo no puede distinguir qué parte de la variable dependiente es explicada por una u otra variable (Villagarcía, 2006). Según aumenta la correlación entre las variables, el error estándar de los coeficientes se incrementa. La multicolinealidad no cambia la estimación de los coeficientes, pero sí su seguridad.

Para evitar la inclusión de variables que causen problemas de colinealidad se exploró, previamente a la elaboración del modelo, el grado de correlación entre las variables independientes, mediante coeficientes de correlación no paramétricos de *Spearman*. En aquellas variables que presentan correlación superior a 0,9 (C. de Madrid) y 0,7 (C. Valenciana) se explora su correlación con la variable dependiente, excluyendo del modelo la que esté menos correlada de cada par. Posteriormente, con las variables que quedan tras esta primera selección, se estudia el fenómeno de multicolinealidad mediante diagnósticos propios de la técnica de regresión multivariante, como son el Coeficiente de Tolerancia, el Factor de Inflación de la Varianza y los Autovalores, Índice de Condición y Proporción de la Varianza. Finalmente, se aplicaron tests no paramétricos de estadística comparativa que proporcionan una medida de la diferencia entre dos conjuntos de datos (Martínez et al., 2005). El objetivo es comprobar si existe 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 (*Prueba de la U-Mann-Whitney* y *Kruskal-Wallis*).

A partir de los resultados obtenidos en los diagnósticos de colinealidad, correlación y tests de estadística comparativa se excluyen del modelo las variables que den lugar a problemas de colinealidad y que no sean significativas al comparar muestras independientes.

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. Quedan sin valorar aquellas celdas en las que el uso urbano ocupe una superficie superior al 50%. 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.

## 4. Resultados

Los resultados obtenidos en la C. de Madrid 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 tests 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% 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% de confianza (resultados de la prueba de *Kruskal-Wallis*). Por estos motivos las variables señaladas se excluyen 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 excluye del análisis. Por tanto, los análisis previos en la C. 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. En la C. Valenciana los mismos análisis estadísticos señalaron la conveniencia de excluir las variables *buffer pistas, máquina, vertederos* y *tasa de paro*.

Mediante la técnica de regresión logística binaria, una vez eliminadas las variables que presentaban problemas de colinealidad y, utilizando la variable dependiente obtenida a partir de interpolación mediante kernel adaptativo, se obtuvieron 17 modelos en la C. Madrid entre los que, finalmente, se seleccionó el modelo 7 por ser el que ofrecía una mejor relación entre complejidad (número de variables independientes) y acierto en la clasificación. Los porcentajes globales de acierto de clasificación de la muestra de calibración (60%) y de validación (40%) son 71,6 y 70,3%, respectivamente. En la C. Valenciana se obtuvieron 20 modelos, eligiendo el 9. Los porcentajes globales de acierto de clasificación de la muestra de elaboración del modelo (60%) y de validación del mismo (40%) son 69,6 y 70,6%, respectivamente.

Al aplicar la ecuación del modelo elegido al 100 % de la muestra se obtiene un 70,6% correcto de clasificación global de la misma, estando la baja incidencia correctamente clasificada

en un 75,4% y la alta incidencia en un 65,7% en la C. Madrid. Los resultados fueron muy similares en la C. Valenciana donde se obtuvo un 68,4% correcto de clasificación global, estando la baja incidencia correctamente clasificada en un 79,4% y la alta incidencia en un 57,4%.

En Madrid el modelo incluyó un total de 7 variables de las cuales, las que más influyen en la variación de la variable dependiente fueron la *interfaz urbano-forestal* seguida por la variable *ENP*, la *tasa de paro* y el *buffer de pistas en zona forestal*. Las que menos influyen fueron las *infraestructuras hoteleras*, la *variación de la población agraria* y los *jefes mayores de 55 años*, todas ellas variables de tipo estadístico. En la C. Valenciana es la variable *variación de la población* la que más influye en la variación de la variable dependiente. Le sigue la variable *Potencial demográfico*. En esta región las variables que producen menos variación en la variable dependiente y, por tanto, las que menor peso tienen en el modelo, son la *interfaz pasto-forestal*, las *infraestructuras hoteleras* y los *jefes mayores de 55 años*.

A continuación se muestran los mapas de los aciertos y errores para la muestra de comprobación y validación de los modelos, así como los mapas de probabilidad estimada en los que no se ha valorado aquellas celdas con una superficie superior al 50% de uso urbano (Figura 8).

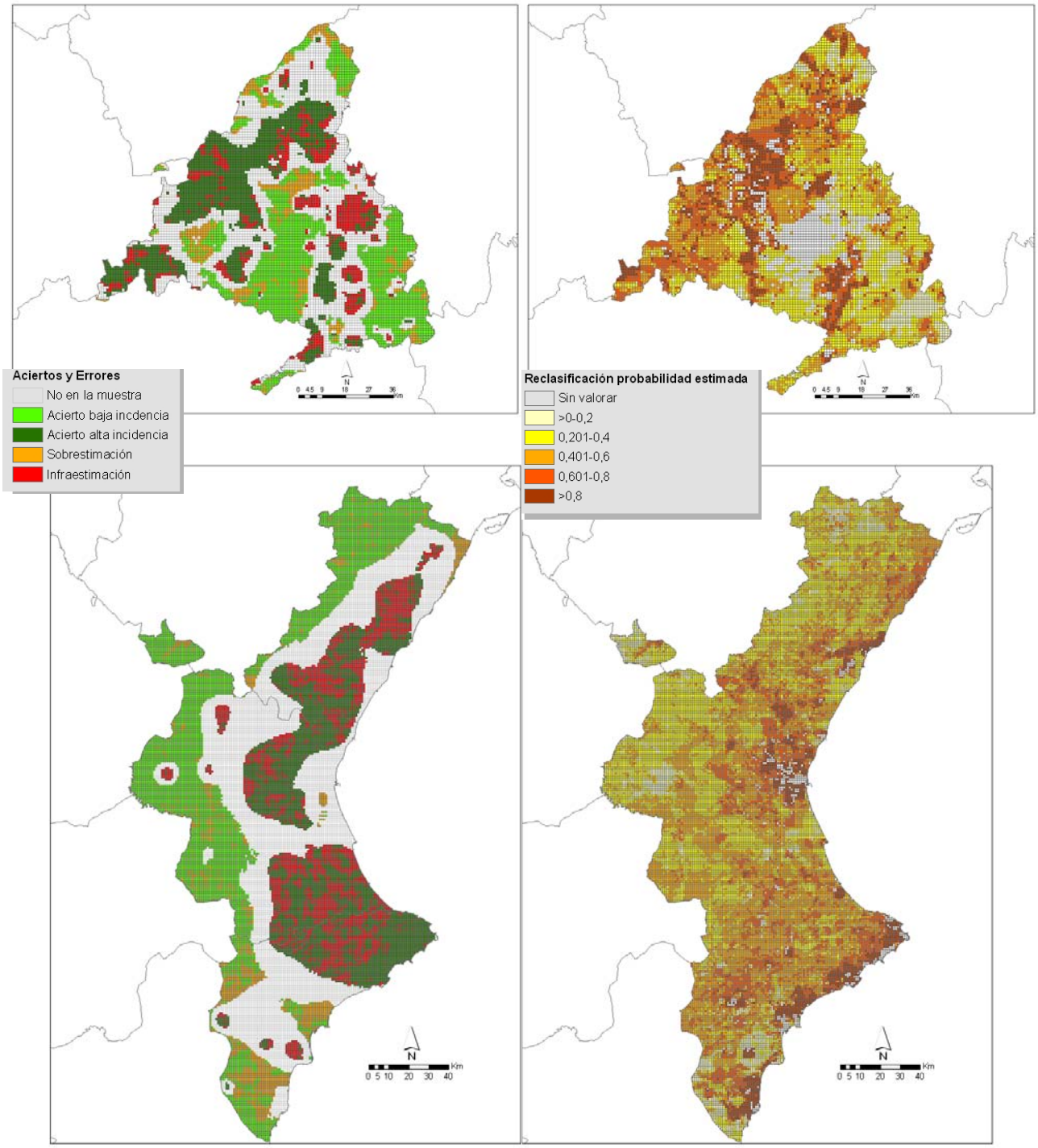



**Figura 8.** Mapas de aciertos y errores y de probabilidad estimada de riesgo humano C. Madrid (modelo 7) y C. Valenciana (modelo 9)

En el mapa de acierto y error de la C. Madrid 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 aparecen 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, que coincide con un área protegida, el *Parque Regional del Sureste*. Las zonas de probabilidad de incendio más bajas se localizan en el centro y Este, en líneas generales.

En la C. Valenciana el mapa de aciertos y errores indica que las zonas de infraestimación del modelo se representan de Norte a Sur así como en varias manchas al Oeste. El modelo acierta en la predicción de la alta incidencia de incendios en la franja citada anteriormente, especialmente en las cuadrículas de la zona Sureste coincidente con la provincia de Alicante. Las zonas de alta probabilidad de incendio se localizan a lo largo de la franja costera coincidiendo con las zonas más pobladas, mientras que la probabilidad va disminuyendo hacia el interior del área de estudio.

**5. Discusión y conclusiones**

Los modelos obtenidos en las dos áreas de estudio han ofrecido resultados muy similares en cuanto al porcentaje de acierto, si bien en la C. Madrid se alcanza un porcentaje de acierto global ligeramente superior (70,6%). En ambas zonas los modelos logran una mejor predicción de la baja incidencia de incendio, alcanzando mayor valor de acierto en la C. Madrid (76,4%) que en la C. Valenciana (57,4 %).

Las variables seleccionadas por el modelo generado para la C. Madrid, y en especial la importancia de la *interfaz urbano-forestal,* coinciden con la opinión de los expertos de esta región consultados en el marco del proyecto *Firemap*. Según éstos, los incendios forestales en la C. de Madrid están fundamentalmente relacionados con las negligencias y/o imprudencias en la interfaz urbano-forestal, carreteras y por usos recreativos. En menor medida por usos agrícolas y ganaderos y también por accidentes (chispas en líneas de ferrocarril, uso de maquinaria en zonas forestales, etc.). A la vista de los resultados el modelo explica de manera acertada la

influencia de la *interfaz urbano-forestal*. La variable de *ENP* fue introducida en el modelo con el propósito de representar el riesgo asociado a los posibles conflictos que la limitación de usos en estos espacios pudiera desencadenar y potencialmente derivar en incendios intencionados. Sin embargo el peso de la variable en el modelo parece indicar que estaría más bien recogiendo el riesgo asociado a la frecuentación para uso recreativo citado por los expertos. En definitiva, la predicción del riesgo que realiza el modelo generado parece considerar adecuadamente las características territoriales relacionadas con el mismo en el área de estudio, una de las más densamente pobladas y con un desarrollo urbanístico constante.

En la C. Valenciana son las variables *variación de la población* y *Potencial demográfico* las que más influyen en la variación de la variable dependiente. El modelo además señala como variables influyentes el *índice de cambio en superficie forestal* y la *interfaz cultivo-forestal*. También en este caso las variables seleccionadas por el modelo coinciden con la opinión de los expertos consultados que señalan la intencionalidad y la negligencia/imprudencia por usos agrícolas y ganaderos como las principales causas humanas de incendio en la región. La C. Valenciana, al igual que el resto de zonas costeras, sobre todo del levante y sur de la Península Ibérica, ha experimentado un gran desarrollo urbanístico ligado a la actividad turística en su franja costera. La afluencia de población en épocas estivales coincide con la estación de mayor riesgo de incendio desde el punto de vista de las condiciones meteorológicas. Los valores más altos de probabilidad obtenidos en el modelo coinciden con las zonas más pobladas y de mayor densidad de población. Sin embargo, el modelo no recoge la variable *interfaz-urbano forestal* como explicativa del riesgo, aunque sí la *interfaz-cultivo forestal* como se ha citado anteriormente. Esta variable señala también, en esta zona, la importancia de las actividades agrícolas relacionadas con el riesgo de ignición de incendios. La variable *índice de cambio de superficie forestal* indica la importancia de la acumulación del material combustible, debido al cambio en el aprovechamiento de las zonas forestales que se ha producido tras el abandono y despoblación de las zonas rurales. El modelo obtenido en este área no consigue resultados de clasificación tan satisfactorios como en el caso de Madrid. Esto puede deberse a la mayor extensión espacial de la región y su diversidad territorial en lo que respecta a las características socio-económicas. Esto hace que el fenómeno de los incendios no sea homogéneo ni en cuanto a la ocurrencia (frecuencia, tamaño de los incendios, etc.) ni, especialmente, respecto a la causalidad. Como consecuencia de ello, es posible que un modelo regional, como el que proponemos en este

trabajo no recoja adecuadamente las particularidades de la región y esto provoque un ajuste menor al esperado. Podrían llevarse a cabo modelos de tipo subregional sobre zonas homogéneas para mejorar la capacidad de predicción de los modelos.

Los modelos obtenidos en las dos áreas de estudio coinciden en señalar las variables más directamente relacionadas con la población (*potencial demográfico*, *variación de población*) como las de mayor importancia a la hora de explicar el riesgo de incendio por causa humana. De igual forma, aparecen destacadas las variables de *interfaz urbano* y *cultivo forestal*, las cuales son objeto de análisis en numerosos estudios de riesgo, y para las que hay legislación específica en materia de prevención de incendios en las comunidades autónomas.

La técnica de R. Logística permite la inclusión de variables de distinta naturaleza por lo que diversos autores la han empleado para el análisis de riesgo de incendio debido a causa humana. 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. Este hecho puede estar influyendo en el modelo final obtenido, independientemente del método empleado. Sería interesante contar con la localización precisa del inicio de los incendios, resolviendo los problemas de incertidumbre espacial que pueden ser decisivos en los resultados del modelo a la resolución empleada para este estudio. De igual forma, las variables obtenidas a partir de fuentes estadísticas en el proceso de espacialización a la unidad requerida de 1km² pueden estar suministrando cierto ruido, al replicar la información de partida a nivel municipal. En este caso, la solución sería contar con información más detallada lo que no siempre es fácil, especialmente cuando se abordan análisis a escala regional. La regresión logística tiene restricciones propias de la estadística tradicional, y la necesidad de convertir la variable dependiente en dicotómica hace que no pueda obtenerse una estimación del número de incendios sino únicamente de la probabilidad de ocurrencia o no ocurrencia. A pesar de ello, resulta muy conveniente la posibilidad de obtener resultados expresados en forma de probabilidad pues ello facilita la interpretación de los mismos y la posible integración con otros factores en un sistema de riesgo como el propuesto en el proyecto *Firemap* (figura 2).

A pesar de las limitaciones se ha demostrado que la metodología propuesta es aplicable a ámbitos geográficos muy diversos siempre que el conjunto de variables independientes representen adecuadamente los factores relacionados con la ocurrencia de incendios en la zona

de interés. La consecución de modelos a una resolución espacial como la que se propone en este trabajo puede ser de gran interés para los gestores, permitiendo identificar zonas de alta ocurrencia de incendios y tipos de variables de riesgo humano influyentes en el mismo. Este análisis refleja la importancia de la distribución de usos en los diferentes territorios, y de cómo la acción del hombre está influyendo en el fenómeno de los incendios forestales. Indica la importancia de estos factores socioeconómicos y del interés en incluirlos en los sistemas generales de riesgo de incendio forestal.

**Agradecimientos**

Este trabajo 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) 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 la C. Valenciana, al *CEAM*, *VAERSA*, *Consellería de Territori i Habitatge*. También expresar nuestro agradecimiento a Patrick Vaughan por sus valiosas correcciones al resumen presentado en inglés.

**Referencias**

Amatulli, G., Pérez-Cabello, F., de la Riva, J. 2007. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. Ecological modelling 200, 321-333.

Área De Defensa Contra Incendios Forestales. Centro de coordinación de la información nacional sobre incendios forestales. 2007. Los Incendios Forestales en España. Decenio 1996-2005. Ministerio de Medio Ambiente. Dirección General para la Biodiversidad, v 1.2, 67.

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

Cardille, J. A., Ventura, S.J., Turner, M.G. 2001. Environmental and Social Factors influencing Wildfires in the Upper Midwest, United States. Ecological Applications 11(1), 111-127.

Carvacho Bart, L. 2002. Aplicación de redes neuronales al análisis de datos en teledetección: predicción y cartografía de incendios forestales. Tesis Doctoral. Facultad de Filosofía y Letras. Departamento de Geografía. Universidad de Alcalá.

Castro, R., Chuvieco, E. 1998. Modeling Forest Fire Danger from Geographic Information Systems. Geocarto International. Vol 13 (1), 15-23.

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, M.P. (Eds.). Nuevas tecnologías para la estimación del riesgo de incendios forestales. Madrid, CSIC, Instituto de Economía y Geografía, 144-158.

Chuvieco E., Salas, F.J, Carvacho, L., Rodríguez Silva, F., 1999. Integrated fire risk mapping. En Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin .Berlin: Springer-Verlag, pp. 61-84.

Chuvieco, E., Martín, P., Martínez, J., Salas, J. 1998. Geografía e Incendios Forestales. Serie Geográfica 7, 11-17.

Chuvieco, E., Salas, J. 1996. Mapping the spatial distribution of forest fire danger using GIS. International Journal Geographical Information Systems, vol. 10 (3) 333-345.

Chuvieco, E., Salas, J. 1994. Sistemas de Información Geográfica y teledetección en la prevención de incendios forestales: un ensayo en el macizo oriental de la Sierra de Gredos. Estudios Geográficos, Tomo LV, 217.

Dirección General Para la Biodiversidad. 2006. Estadísticas de Incendios Forestales. Ministerio de Medio Ambiente. Disponible en http://www.incendiosforestales.org/estadisticas.htm (15 Noviembre, 2007).

Estirado, F., Molina, V. 2005. El problema de los incendios forestales en España. Documento de Trabajo, *Fundación Alternativas*.

Ferrando, J.F. 2004. Restauración de funciones ambientales de una zona forestal quemada en La Hoya de Buñol. Trabajo Profesional Fin de Carrera. ETSIAM, Universidad de Córdoba.

Fidalgo García, P., Martín Espinosa, A. 2005. Atlas Estadístico de la Comunidad de Madrid 2005. Consejería de Economía e Innovación Tecnológica. Instituto de Estadística de la Comunidad de Madrid. Madrid.

Garson, D. 2006. Statnotes: Topics in Multivariate Analysis. Disponible en http://www2.chass.ncsu.edu/garson/pa765/statnote.htm (23 Abril, 2009).

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

Gouma, V., Chronopoulou-Sereli, A. 1998. Wildland Fire Danger Zoning-A Methodology. International Journal of Wildland Fire, 8(1), 37-43.

INFOCA. 2006. Clima e información meteorológica. Capítulo III. Junta de Andalucía. Consejería de Medio Ambiente.

Instituto de Estadística de la C. Madrid. 2006. Datos municipales. Disponible en http://www.madrid.org/iestadis/ (23 Abril, 2009).

Instituto de Estadística de la C. Valenciana 2006. Revisión del Padrón 2006. Disponible en http://ive.infocentre.gva.es/ (23 Abril, 2009).

Instituto Nacional de Estadística (INE) 2007. Estimaciones de población a partir censo 2001. Disponible en http://www.ine.es/ (23 Abril, 2009).

Izquierdo, J. 2007. Instrumentos económicos para la prevención y la lucha contra incendios. En Fundación Santander-Central Hispano (Ed.), Hacia la viabilidad económica del medio rural y de los bosques. Antes del Fuego. Soluciones a los incendios forestales en España. Madrid.

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.), The role of remote sensing data. New Jersey, US: World Scientific Publishing, vol. 4, pp. 143-194.

Levine, N. 2004. Kernel density interpolation. In Crimestat 3.0, 8.

Lorenzo MC, Pérez MC. 1995. Modelos de probabilidad para el estudio de la ocurrencia de incendios forestales. In 'IX reunión ASEPELT España'. (Santiago de Compostela, España).

Martell, D.L., Otukol, S., Stocks, B.J. 1985. A daily people-caused forest fire occurrence prediction model. In 8th National Conference on Fire and Forest Meteorology. (Detroit, Michigan).

Martell, D.L., Otukol, S., Stocks, B.J. 1987. A logistic model for predicing daily people-caused forest fire occurrence in Ontario. Caumar.

Martín, P. Bonora, L., Conese, C., Lampin, C., Martínez, J., Salas, J. 2002. Towards methods for investigating on wildland fire causes. *Deliverable D-05-02*. EUFIRELAB. Disponible en http://eufirelab.org (23 Abril, 2009).

Martínez, J., Martín, M.P., Romero, R., Martínez, J., Echavarría, P. 2005. Aplicación de los SIG a los modelos de riesgo de incendios forestales: riesgo humano a escala regional. De lo local a lo global: nuevas tecnologías de la información geográfica para el desarrollo. En Gurría Gascón, L., Hernández Carretero, A., Nieto Masot, A. (Eds.). Servicio de Publicaciones Universidad de Extremadura, pp. 329-345.

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, M.P. (Eds.), Nuevas tecnologías para la estimación del riesgo de incendios forestales. CSIC, Instituto de Economía y Geografía, Madrid, pp. 101-142.

Ministerio de Agricultura, Pesca Y Alimentación (2004): Hechos y cifras de la Agricultura, la Pesca y la Alimentación en España. Disponible en http://www.mapa.es/es/ministerio/pags/hechoscifras/introhechos.htm (15 Noviembre, 2007).

Moyano, E. 2007. Incendios forestales en España. Diagnóstico de las causas. En Fundación Santander-Central Hispano (Ed.), Hacia la viabilidad económica del medio rural y de los bosques. Antes del Fuego. Soluciones a los incendios forestales en España. Madrid

Pew, K.L. and 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, 1-18.

Plan General de Ordenación Forestal de la Comunidad Valenciana. 2004. Decreto 106/2004, de 25 de junio, del Consell de la Generalitat Valenciana.

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

Salas, J., Cocero, D. 2004. El concepto de peligro de incendio. Sistemas actuales de estimación del peligro. En Chuvieco, E., Martín, M.P. (Eds.), Nuevas tecnologías para la estimación del riesgo de incendios forestales. CSIC, Instituto de Economía y Geografía, Madrid, pp.23-32.

Suárez Torres, J. 2000. La prevención de incendios forestales en la Comunidad Valenciana, en: publicaciones de la Real Sociedad Económica de Amigos del País. Valencia.

Vasconcelos, M.P.P., Silva, S., Tomé, M., Alvim, M., Pereira, J.M.C., 2001. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. Photogrammetric Engineering and Remote Sensing 67 (1), 73-81.

Vega García, C., Woodard, P.M, Titus, S.J., Adamowicz, W.L., Lee, B.S., 1995. A Logit Model for predicting the Daily Occurrence of Human Caused Forest Fires. International Journal Wildland Fire 5 (2), 101-111.

Vélez, R. 2005. Defensa contra incendios forestales: estrategias, recursos, organización. Disponible en http://forum.europa.eu.int (15 Noviembre, 2007).

Villagarcía, T. 2006. Regresión, en: Curso de Metodología de Investigación Cuantitativa. Técnicas Estadísticas. CSIC.

WWW/ADENA. 2007. Incendiómetro. Incendios forestales. Verano 2007. Disponible en http://www.wwf.es/incendios07.php (15 Noviembre, 2007).

# Capítulo II

## Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables

**Publicación derivada:** *Vilar del Hoyo, L., Martín Isabel, M.P., Martínez Vega, F.J. 2009. Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables. Journal of Environmental Management (en revision/under revision)*

# Logistic regression models for human-caused wildfire risk estimation: Analyzing the effects of different response variables

**Abstract**

As reported in other Mediterranean areas, such as California, where human-caused fires account for at least 95% of the fires in the last century, in Southern Europe, 90% of wildland fires is caused by human activities. In spite of these figures, the human factor hardly ever appears in the definition of operational fire risk systems due to the difficulty in characterising it. This paper describes the use of logistic regression models to estimate the probability of fire occurrence due to human causes at 1×1 km grid resolution in the region of Madrid, a highly populated area in the centre of Spain. Socioeconomic data have been used as predictive variables to spatially represent anthropogenic factors related to fire risk in Euro-Mediterranean countries such as Spain. Historical fire occurrence from 2000-2005 was used as the response variable. In order to analyze the effects of the spatial accuracy of the response variable on the model performance (significant variables and classification accuracy) two different models were defined. In the first model (model 1) real fire ignition points (x, y coordinates) reported by the regional fire management agency were used as response variable. This model was compared with another one (model 2) where the response variable was obtained from fire ignition points randomly located within the 10×10 km grid to which those fires were spatially referred to in Spanish official statistics. Both models have been validated using and independent set of real fire ignition points for 2006 and 2007. Model 1 performs better since it correctly predicts 74.8% of the fires, as opposed to 65.4% for the random points model. The explanatory component of model 1 also fits better than the random points model with the official version of the human factors most directly connected with fire occurrence in this region. According to official reports on historical fire causes confirmed by fire managers' experience the Wildland-Urban Interface (*WUI*) is the main driving factor affecting fire occurrence in the region. Although this variable was selected by both models, its relative importance is very high in model 1 and low in model 2. Results show that accurate information on fire ignition coordinates is necessary to obtain reliable spatial models at regional level which can benefit resources management in operational fire prevention activities.

**Keywords**
Euro-Mediterranean region, Fire ignition points, GIS, Socioeconomic, Spatial location, Wildland-Urban Interface

## 1. Introduction

Mediterranean ecosystems can not be fully understood without the role of fires. Natural fires have been essential to maintain biodiversity. Fire has been also a widely used tool to manage the territory. However, in the last decades, natural fire regimes have experienced significant alterations (fire frequency, intensity and severity) which have aggravated their ecological, social and economic consequences (Westerling et al., 2006; FAO, 2007). Between 1991 and 2001, a total of 1 034 133 fires affected six million hectares (ha) in Europe. In five of the southern European member states (Portugal, Spain, France, Italy and Greece) a yearly average of 59 683 fires and 476 879 ha of burned area were reported for the 2000-2007 period. During 2007 in those five countries total burned area was 575 531 ha (above the average of the last 28 years) while the number of fires (45 623) was lower than average for the same period (Camia et al., 2008).

The importance of having accurate and reliable information on wildfire incidence and also in fire causes is evident. Knowing the mechanisms behind fire occurrence is expected to improve fire prevention activities (i.e. fire risk estimation) and to reduce its negative impacts. This fact contrasts with the limited information available on fire causes and its motivations, especially at regional and global scales. Unknown causes are still too frequent in many wildfire statistics including those from most European countries/regions. Some global organizations and networks such as the Food and Agricultural Organization of the United Nations (FAO) are making a great effort to collect fire data from different countries or regions. FAO has worked in the compilation of fire statistics since 1980 by carrying out enquiries in the countries of the United Nations Economic Commission for Europe. The information compiled includes the total number of fires and total area burned by type of land and cause. Fire causes are classified in three general categories: *arson*, *negligence* and *natural causes*. The negligence category is divided in fires due to *agricultural operations*, *logging and forest operations*, *other industrial activities*,

*communications*, *general public* and *other* (FAO, 2002). Despite this, regional/global datasets are still incomplete and rather heterogeneous.

In Europe, the European Commission, on the basis of Council Regulation (ECC) 2158/92 on protection of the Community's forest against fire, decided, in 1994, that member states must, at least, collect a set of data consisting on fire statistical information, comparable at community level and accessible at specified regular intervals. Annex I of this regulation established the minimum core information on forest fires. According to that, fire location statistics should include the name and the code of the municipality and the successive territorial units (province or department, region, state) in which the fire was reported. Concerning the fire causes, they should be classified in the following categories: *unknown*, *natural*, *deliberate and accidental or negligence.*

In Spain fire data is collected since 1968 in a national fire database (Spanish Ministry of the Environment, Rural and Marine Affairs) (MARM, 2006). Historical fire records contain more than 150 fields of information related to fire including, among others, information on the spatial location of the fire, fire causes and motivations, burned area, evaluation of damages in forest products, etc. These fire records reveal that between 1996 and 2005 over 90% of all wildland fire ignitions were caused by humans (MARM, 2006). Fire causes are classified in the following main categories: *lightning*, *negligences/accidents, deliberate, unknown* and *re-ignited fires*. In the last decade 60% of fires were deliberate while negligence and accidents caused 17.5% and lightning caused 3.6%. In this period 17% of the fires causes were unknown. In the Spanish database fires are spatially referred to a 10×10 km grid and also at the municipality level. Some regions additionally register x, y coordinates of the ignition points. In our study area (the region of Madrid) x, y coordinates of the ignition points have been collected since 2000.

Despite the lack of comprehensive information on fire motivations it is obvious that the human factor is critical in fire incidence. However, there is still a long way to go in terms of predicting and modelling fire risk of human origin and also in the integration of human and natural factors in a comprehensive risk scheme, which also includes an adequate estimation of the vulnerability associated to humans. This was the approach of the *Firemap* research project (*'Integrated Analysis of Wildland Fire with Remote Sensing and GIS'*), where an integrated wildland fire risk index was proposed which includes, among other factors, an estimation of fire probability associated to human activities (Chuvieco et al., 2007). This paper presents the

methodological approach followed in the framework of the *Firemap* project to characterize fire ignition causes related with socioeconomic conditions and to derive spatially explicit information on fire probability associated to human activity.

Scientific approaches which include human factors in fire risk assessment have been commonly based on statistical models that attempt to explain historical, human-caused fire occurrence from a set of independent variables. Logistic regression analysis has been frequently used both to predict and also to explain human-caused fires (Martell et al., 1987; Vega-García et al., 1995; Chuvieco et al., 1999; Lin, 1999; Pew and Larsen, 2001; Vasconcelos et al., 2001; Martínez et al., 2004; Prasad et al., 2008; Martínez et al., 2009). Other statistical methods such as linear regression, classification regression trees, neural networks or Bayesian probability have also been used in fire risk mapping to generate local risk models (Chao-Chin, 2002; Koutsias et al., 2004; Robin et al., 2006; Amatulli et al., 2006, 2007; Amatulli and Camia, 2007; Syphard et al., 2007; Vega-García, 2007; Yang et al., 2007; Romero-Calcerrada et al., 2008).

There are a number of studies which analyze the human influence on fire occurrence in Mediterranean areas as the one examined in this paper. However, the conclusions drawn from these studies can hardly be generalized since different explanatory variables have been used as indicators of the human activity and they cover different spatial and temporal scales. The response variable in those models always refers to fire occurrence, commonly obtained from information available in forest fire databases; however, the precision on the spatial location of fires significantly varies depending on the data source. As a result, a wide range of possibilities can be found in the literature. Some authors have used as a response variable in their statistical models the exact location (x, y coordinates) of the fire ignition points (Cardille et al., 2001; Pew and Larsen, 2001; Vasconcelos et al., 2001; Koutsias et al., 2004; Robin et al., 2006; Kalabokidis et al., 2007; Yang et al., 2007). Even though this information can be considered the most accurate data on fire location, some authors highlight other limitations of the fire data that can affect the model results. Vasconcelos et al. (2001), for instance, used in their analysis not all the fires but only a sample of those occurring in each fire season. This limitation is imposed by the data source, the Portuguese Forest Service, which collects fire data in the field trying to maximize the number of fires studied; therefore, field teams collect more data in areas with a high concentration of fire events. The authors pointed out that the spatial patterns they found in the model results could be a consequence of the sampling procedure. Pew and Larsen (2001)

mentioned that many of the false positives their model produced were due to the relative brevity of their observation period relative to the fire cycle. Others have used as response variable in their models less accurate fire information spatially referred to a grid and/or different administrative units (municipality, province, etc.). Shypard et al. (2007) studied humans influence in California fires by relating contemporary and historic fire data with both human and biophysical variables. They concluded that some variables that would be expected to influence fire occurrence were not selected by the model as consequence of the use of aggregated fire data at a county level. Martínez et al. (2009) used an ignition danger index, defined as cumulative number of fires in a period divided by forest area in each municipality (Vélez, 2000). The use of a density index reduced the bias of comparing absolute number of fires per municipality, since they had a wide range of sizes in the study area. A different approach has been based on transformations of the original fire locations to continuous surfaces. Amatulli et al. (2006) used latitude and longitude coordinates of fire ignition points provided by the Italian National Fire Database to derive a continuous fire occurrence map through kernel density approach. Authors recognize that resulting kernel density surfaces might be influenced by the ignition point positions in the highest density zones. Amatulli et al. (2007) calculated also a continuous map of fire occurrence in Aragon (Spain) by interpolation of fire events referred to the municipalities and to the UTM 10×10 km grid using adaptative kernel technique. In the analysis they concluded that the model seemed to be more sensitive to the number of ignition points rather that to their precise position.

This paper describe the model that has been built for predicting human-caused fire occurrence in the context of wildfire risk analysis using (1) socioeconomic data as predictor variables and (2) real fire ignition points as the response variable. Logistic regression is used to generate a model that produces fire occurrence probability spatially mapped at a 1×1 km grid level. This spatial resolution has been regarded by experts as the most useful for fire management on a regional level in Euro-Mediterranean ecosystems. Results were compared with those from a model using the same predictors but a different response variable. In this case x, y fire coordinates has been obtained from a set of random points located within the 10×10 km grid in accordance with Spanish national fire database. This paper describes the methodology that has been followed to select and generate the response variables, the statistical analysis

applied to build the predictive model and the results obtained with the two different response variables.

## 2. Materials and methods

### 2.1 Study Area

The region of Madrid is located in central Spain (Fig. 1) and covers an area of 8028 km$^2$. It is primarily composed of two sub regions: the mountain range, that follows a NE-SW direction (a third part of the region, with its highest peak having an elevation of 2428 m and a width that ranges from 15 to 35 km), and the Tagus river basin, which runs from the fault of the mountain range (800 m above the sea level) to the Tagus' river bed (Fidalgo and Martín, 2005). Although the climatic conditions in the region are Mediterranean, its altitude, distance to the sea and the barrier effect of the surrounding mountain ranges are the causes for lower annual rainfalls (300 to 1000 mm -in the mountain range-) and higher thermal amplitude compared to other Mediterranean regions. Pastures and shrublands occupy vast areas in the region, whereas forest areas are mainly located in the mountain range, two thirds of which are broadleaves. Holm oak (*Quercus ilex*), which is the predominant specie, can be commonly found in the typical Mediterranean *dehesa* formation, a sparsely forested meadow highly regarded from the point of view of its landscape as well as their recreational, environmental and livestock value. Urban areas in the Madrid region have been increasing in size and population since the 1960s, spreading into agricultural and forested areas. Due to the high human density and pressure, land use has changed significantly, affecting the agricultural and forested areas. Forest fires in Madrid region are spatially associated to roads, railways, dump sites and urban areas (Nicolás and Caballero, 2001). The contact boundary between urban and forest areas, which is commonly referred to as the Wildland-Urban Interface (*WUI*), is an area of major concern for regional fire managers. The main areas of *WUI* follow the radial pattern of the main motorway network, especially in the western part of the region and disperse areas in the North-East and South-East (Caballero, 2001). Fig. 2(a) shows the main land uses and Fig. 2(b) the *WUI* distribution and road infrastructure in the region of Madrid.

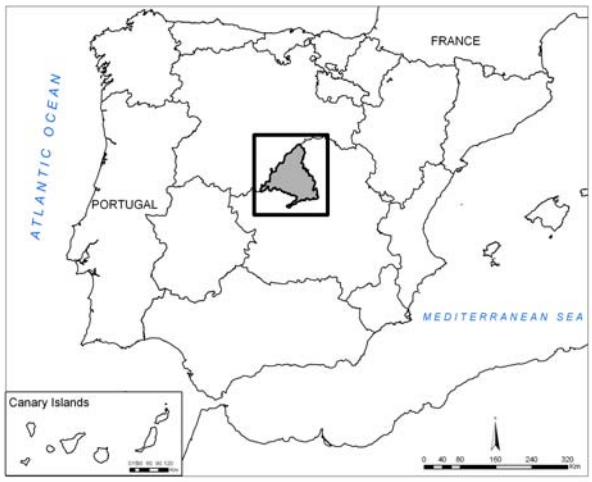



**Fig.1.** Study area. The region of Madrid (Spain).

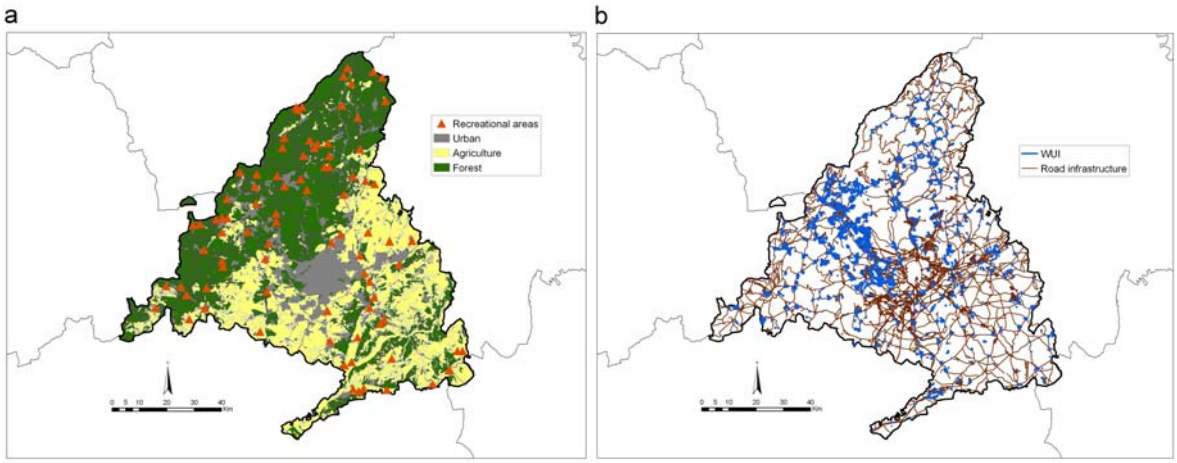



**Fig.2.** (a) Main land use categories in the region of Madrid: urban, agricultural and forest areas plus recreational areas (*source*: Corine Land Cover 2000). (b) WUI and road infrastructure.

The study period was reduced to five years (2000-2005) due to the lack of x, y coordinates of fire ignition points prior to 2000. During this period more than 68% of the fires were due to unknown causes (MARM, 2006). Of the known fires, 24% were caused by negligence or accidents while 5% were deliberate (Fig. 3). In this region over 90% of the fires are caused by human activities, a figure which is in agreement with the average data for the whole country. The detailed cause categories of the negligence or accident fires, available in the fire records for the study period, show that most fires were caused by forest works (12.3%), railway

infrastructure (11.9%), accidents by machinery (10.4%) and agricultural burning (9.6%). Also, over 28% of the fires started close to the roads, followed by those in other places in the forest (19.9%), forest trails (12.2%) and houses (10.5%) (MARM, 2006).

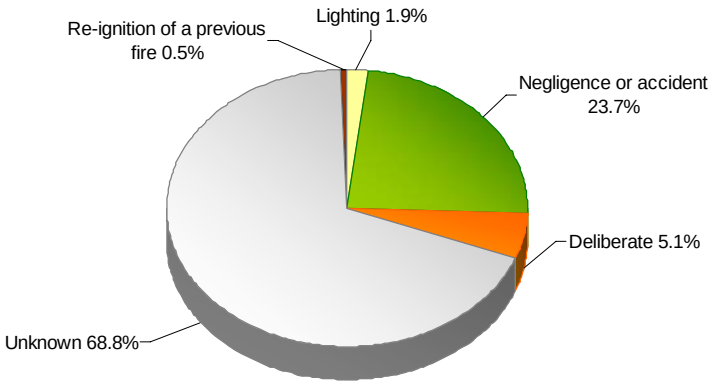



**Fig.3.** Fire causes from 2000 to 2005 in the region of Madrid (MARM, 2006).

2.2 Response variable

Two different response variables were used in order to evaluate the importance of having accurate information on fire location for the production of spatially explicit fire risk models at a regional scale. As previously explained, in the Spanish national fire database, fire locations are referred to a 10×10 km grid and also to the municipality. In recent years and depending on the region, the fire database has included the exact location (x, y coordinates) of fire ignition points. In the Madrid region both levels of information are available since 2000.

The database including information on fire ignition coordinates provided by the Fire Department of the Madrid region contains a total of 1700 fire ignition points during the study period (2000-2005). This database does not include information on fire causes. Since the objective is to build a model to predict only human-caused fires, just those historical fire records identified as human-caused, should be used. However, due to the lack of information on fire causes it was impossible to discriminate between natural and human caused fires; therefore it was assumed that all fires were due to human activity, given that historical fire statistics show that over 90% of fires in this region are human-caused. Fig. 4(a) shows the spatial location of these points in the study area. Fire points incorrectly located, (i.e. out of the regional

boundaries, 0.4% of the total dataset) were omitted in the analysis. The dichotomic response variable used in the model is the presence or absence of fire ignition points in each cell of the 1×1 km UTM reference grid during the study period (Fig. 4b). A total of 1203 cells out of 8452 were affected by fire during this period.

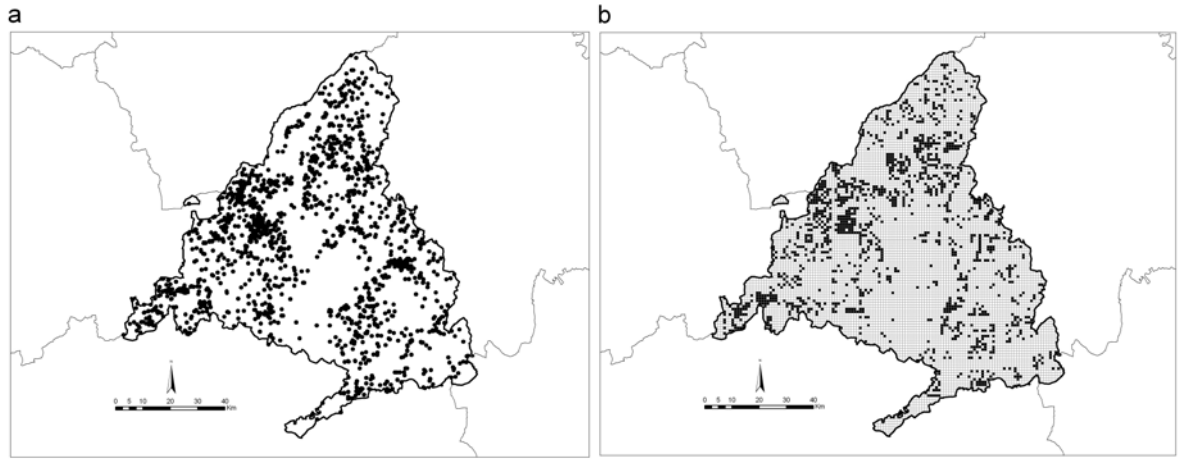



**Fig.4.** Response variable for model 1 (real fire ignition points). (a) Location of the fire ignition points (b). Dichotomic variable in 1×1 km UTM grid from real fire points.

The response variable with the random fire ignition points (model 2) was generated using also official fire statistics for the same period (2000 to 2005). In this case they were obtained from the national fire database provided by MARM. In this database the minimum spatial unit to which forest fires is spatially referred to is a 10×10 km grid. However, fires are also referred to a coarser spatial unit: the municipality. Although this database contains some information on fire causes, a large proportion of fires are labelled as unknown. Therefore, the same criteria was followed as with the x, y fire ignition points and all the unknown fires were attributed to human causes. In this case, fires identified as natural caused (lightning) in the database were excluded from the analysis (1.9% for the study period). From this database it is impossible to know the exact position of each ignition point within the 10×10 km grid. However, some spatial analysis can be done in order to reduce the uncertainty in the fire point location within the grid. The approach chosen in this study was to combine the information on fire location at the municipality level with the one referred to the 10×10 km grid. Applying a spatial overlay of this information using GIS tools it is possible, in some cases, to locate the fire within the smaller area where the municipality and the 10×10 km cell overlay. In addition to

that, a forest mask was applied assuming that forest fires only can occur in forest areas. The exact positions of the fires in the resulting polygons were obtained by random distribution of the total number of fires assigned to the original 10×10 km cell in the fire database. Fig. 5 shows the spatial analysis described and the final distribution of the random fire ignition points in the region. Again, a dichotomic response variable was generated by recording the presence or absence of fire ignition points in each cell of the 1×1 km UTM reference grid (Fig. 6). In this case a total of 1321 cells out of 8452 were affected by fires during the study period.

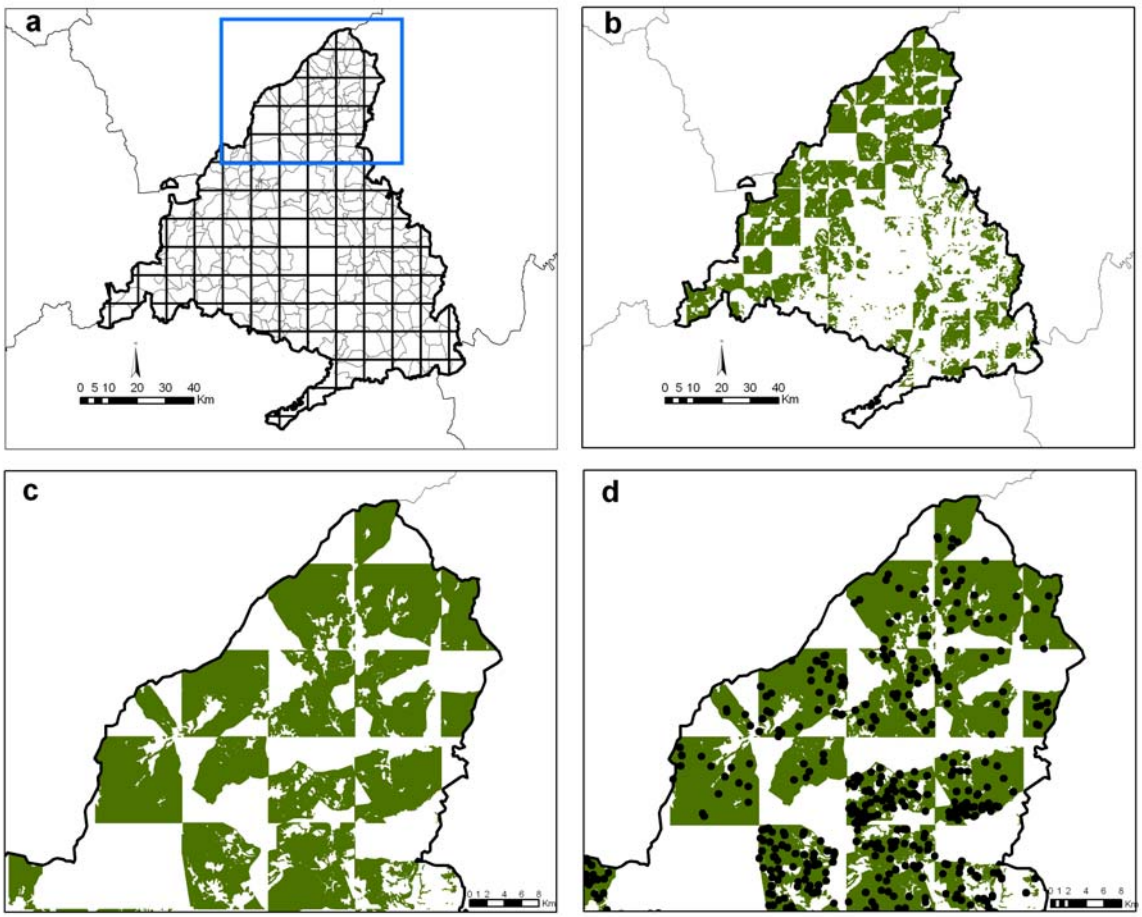



**Fig.5.** Response variable from a set of random fire points. (a) Municipalities and 10×10 km grid overlay. (b) Municipalities and 10 km² grid overlay without non forest areas. (c) Zoom of final polygons for fire location. (d) Final polygons with random fire ignition points.

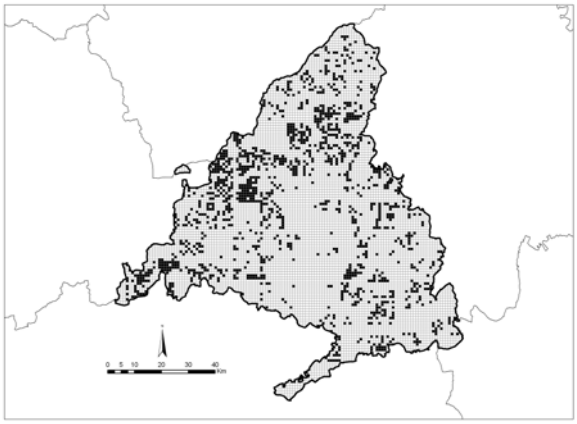


**Fig.6.** Dichotomic variable in 1×1 km UTM from a set of random fire points.

2.3 Explanatory variables: socioeconomic data

Tables 1 and 2 show the socioeconomic variables used in this analysis. These explanatory variables represent different social factors related to human fire risk in Spain (Martínez et al., 2009). Those factors have been reported by the literature and historical fire databases to have a direct or indirect influence in fire occurrence in Spain and could be considered representative of Euro-Mediterranean countries. They are related with socioeconomic changes in rural and urban areas, traditional activities in rural areas, accidents or negligence, fire prevention activities and factors which can lead to social unrest (e.g., land use disputes or high unemployment rates). A set of variables, which were used as explaining variables in the model, were selected to represent those factors. Those variables were obtained from diverse cartographic and statistical sources and were spatially mapped using GIS tools to the 1×1 km UTM reference grid (Vilar, 2006; Vilar et al., 2007).

**Table 1.** Socioeconomic predictors. Type of source (cartographic), human factor which affect fire occurrence and predictor variables that represent each factor

| Type | Factor | Name of the variable | Description | Range of values |
|---|---|---|---|---|
| Cartographic source | **Accident or negligence** | road_b | Buffer of roads[9] | 0-0.51 |
| | | road_b_for | | 0-0.31 |
| | | imd_index | IMD = Daily Average Traffic Intensity | 0-40,1273.52 |
| | | imd_index_for | Index by road segment (road length * road IMD * weight factor) | 0-194,115.9 |
| | | railway_b | Buffer of railways | 0-0.36 |
| | | railway _b_for | | 0-0.24 |
| | | trails_b | Buffer of forest trails | 0-1.77 |
| | | trails_b_for | | 0-1.73 |
| | | electric_line_b | Buffer of electric lines | 0-0.02 |
| | | electric_line_b_for | | 0-0.01 |
| | | shot_quarry | Shooting fields and quarries | Presence (1) Absence (0) |
| | **Socioeconomic changes in rural and urban areas** | recreational_area | Recreational Areas weighted by barbecue presence | 0-0.49 |
| | | demograp_potential | Demographic potential | 1,440.75-1,994,000 |
| | | fuel_incre | Fuel increase | -833,373.4-753,969.9 |
| | | WUI | WUI | 0-0.19 |
| | | Dump | Buffer of dump sites | 0-0.09 |
| | | for_agricult_i | Forest-culture interface | 0-0.98 |
| | | for_pasture_i | Forest-pasture interface | 0-0.75 |
| | **Conflicts that can break the beginning of deliberated fires** | protected_area | Natural Protected Areas | 0-1.0 |
| | | zepas | Special Protection Areas | 0-1.0 |
| | | preser_pu_mount | Public Utility and Preserved Mountains | 0-1.0 |
| | | consor_mount | Consorcied Mountains | 0-1.0 |
| | **Fire prevention activities** | fire_watch | Presence or Absence of the fire watch towers | Presence (1) Absence (0) |
| | | viewsheed | Viewshed obtained from the fire watch towers | 0-1.78 |

[9] Buffers of roads, railways and electric lines were also calculated only in forest areas

**Table 2.** Socioeconomic predictors. Type of source (statistical), human factor which affect fire occurrence and predictor variables that represent each factor

| Type | Factor | Name of the variable | Description | Range of values |
|---|---|---|---|---|
| Statistical source | **Socioeconomic changes in rural and urban areas** | population_var | Variation of the population between 1970-2004 | 38.16-1026.91 |
| | | hotel_sites | Hotel sites | 0-14811 |
| | | agra_pop_var | Variation of the agrarian population 1996-2001 | 0-250.02 |
| | **Traditional activities in rural areas** | owners55 | Percentage of owners of agrarian holdings older than 55 | 0-92.04 |
| | | cattle_charge | Cattle grazing pressure (number of sheep and goat cattle in pastures and shrublands) | 0-117.91 |
| | | agr_machinery | Agricultural machinery | 0-3.62 |
| | **Conflicts that can break the beginning of deliberated fires** | income | Income | 6521.9-16877.16 |
| | | unemploy_rate | Unemployement rate 2001 | 0-20.63 |

2.4 Statistical analysis

Logistic regression was used to estimate the probability of the occurrence of a fire in a 1×1 km cell given the set of observed explanatory variables (Vilar, 2006; Vilar et al., 2007). Furthermore, logistic regression estimates coefficients for each explanatory variable which can be used to assess the relative influence that each covariate has on the response. The logistic regression model is defined as follows:

$$P_i = \frac{1}{1+e^{-z_i}} \quad (1)$$

$$z_i = \alpha \ + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_\rho X_{i\rho} \quad (2)$$

where $P_i$ is the probability of a fire occurring in grid cell $i$; $z_i$ is the linear combination of the independent variables weighted by their regression coefficients ($\beta$); $x_{ij}$ represents the value of the independent variable $j$ in the cell $i$, $\alpha$ is the constant, and $e$ is the base of the natural log (Afifi and Clark. 1990; McGrew and Monroe, 1993; Pew and Larsen. 2001). The logit model applied was:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^{\ T}\beta \quad (3)$$

where $X_i^T$ is the vector of the explanatory variables and $\beta$ is the vector of the parameters.

The *Wald step forward* method was applied to build the logistic regression model. Models were fit using the statistical package SPSS v.15. A random sample of 60% of the grid cells was used to generate the model and the remaining 40% of the cells were used for the validation of the fitted model. Predictions were finally obtained for 100% of the data, thus obtaining the probability for each grid cell. In addition, cut-off points were used to obtain the response variable (presence or absence of fire). When determining a cut-off threshold, two statistics should be considered: sensitivity and specificity. Sensitivity is the proportion of true positives that are predicted as events and specificity is the proportion of true negatives that are predicted as non-events (Vasconcelos et al., 2001). In this model, sensitivity is the proportion of fitted values that have been correctly classified as fire and specificity is the proportion of fitted values obtained from the model that have been correctly classified as non-fire. For the response variables sensitivity and specificity values were obtained using the Receiver Operating Characteristic (ROC) analysis (Fawcett, 2006). With this analysis, the sensitivity (Y axis) and 1-specificity (X axis) was obtained. The area under the curve represents the probability that the assay result for a randomly chosen positive case (fire) will exceed the result for a randomly chosen negative case (no fire). The ROC analysis table reports the sensitivity and 1-specificity for every possible cut-off for possible classification purposes (SPSS Statistical support). The optimal cut-off point corresponds to the intersection of the two lines (sensitivity vs. specificity) (Vasconcelos et al., 2001). Fig. 7 shows the sensitivity and specificity of the calibration models. Based on this method, the optimal cut-off point for classification was set to 13% for model 1 and 17% for model 2.

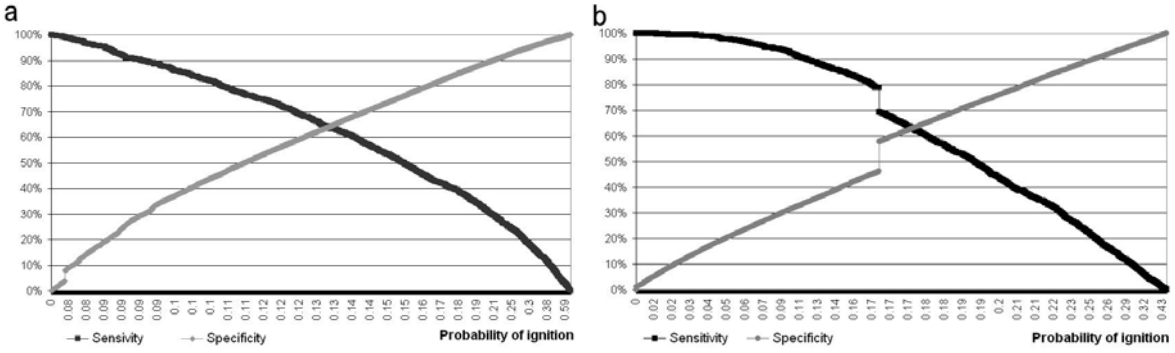



**Fig.7.** Sensitivity and specificity of the calibration models. Cut-off points of the response variables. (a) Real fire ignition points. (b) Random fire ignition points.

The statistical methods applied assume that the explanatory variables are not correlated. Therefore, as in previous works (Vilar, 2006; Syphard et al., 2007; Vilar et al., 2007), *Spearman* correlation coefficients between the explanatory variables were calculated in order to avoid multicollinearity problems. All pairs of explanatory variables with a correlation over 0.9 were identified. The next step was to take these pairs of variables, and to calculate the correlation between each variable and the response variable and exclude from the model those variables less correlated with the response variable. Other common multicollinearity diagnostics that were applied were: *coefficient of tolerance*, VIF (*variance inflation factor*), *eigenvalues*, *condition indices* and *variance proportions*. Based on these diagnostics it was possible to identify and to exclude those explanatory variables that still present any multicollinearity problem. Non-parametric tests were also applied to analyse differences between groups, namely the Mann-Withney-U and the Kruskal-Wallis. The aim was to compare if there were differences between the variables from two samples, one with a high fire occurrence and the other one with a low fire occurrence (Martínez et al., 2005, 2009).

An 'urban mask' was applied to the final probability maps using the Corine Land Cover 2000 (European Environment Agency, http://dataservice.eea.europa.eu/, April 2009) as the land use data source. Cells with more than 50% of urban use were excluded from the final fire probability map.

An independent validation of the results was carried out using official fire ignition points from 2006 and 2007 provided by the Fire Department of the Madrid region in order to test whether the predicted probability of fire occurrence agrees with real fire occurrence. 2006 and 2007 fire ignition points were spatially overlaid on the probability maps generated by the two models. The distribution of the fires/non fires in the predicted probability values was then analyzed. The probability values were divided into two groups, using the cut-off values previously obtained for the statistical analysis (0.13 for the '*real fire ignition points model*' –model 1-and 0.17 for the '*set of random fire ignition points model*'-model 2- ). These two groups were subdivided into five intervals (20% probability values per interval). Firstly, the 2006-2007 fire ignition points that occurred above the cut-off point in each model (0.13 and 0.17 for model 1 and 2 respectively) were quantified and labelled as correct fire predictions. The second step was to verify the absence of fire in the cells with probability values below the cut-off point, which would mean that the models were correct in predicting 'no fire'. The next step was to check the

presence of fires in those cells with predicted probabilities values below the cut-off points (therefore it is assumed there is an absence of fires), those would be labelled as omission errors. A further step was to check the predicted probabilities above the cut-off point where any fire has occurred. This could not be considered as an error of the model because it is plausible that the probability of ignition is high and yet no fire may occur if there is not an ignition source. Therefore, we concentrate on analysing the fires in the predicted probabilities below the cut-off points because omission errors or underestimation of ignition probability is the most critical for an operational application of the model.

**3. Results**

3.1 Logistic Regression model using real fire ignition points as the response variable (model 1)

Results of the multicollinearity analysis applied to the explaining variables showed that the *Spearman* correlation coefficient was over 0.9 for the following pairs: *IMD indices/road buffers*, *IMD indices in forests/road buffers in forests*, *forest trail buffers/forest trail buffers in forest*s *agricultural machinery/owners of agrarian holdings older than 55*. From those variables, the ones that were finally excluded from the model because they had the lowest correlation with the dependent variables were: *IMD indices*, *forest trail buffers* and *owners of agrarian holdings older than 55.* In addition, the multicollinearity diagnostics revealed that the *income* variable has some collinearity problems and accordingly it was also excluded.

Using the *Wald step forward,* 13 candidate models were obtained. Out of these, the model at the 6$^{th}$ step was chosen as the final model, having used as a criterion the acceptable balance between the number of variables (complexity) and the prediction ability (following a parsimonious criterion). Table 3 shows the logistic regression results for the model built with the calibration sample. The variables selected were: *recreational areas*, *road buffers in forests*, *railway buffers in forests*, *WUI*, *forest agriculture interface* and *hotel sites*. *WUI* has the highest logistic coefficient B, followed by *railway buffers in forests* and *road buffers in forests* The EXP (B) is the odds ratio which refers to the predicted change in odds for a unit increase in the corresponding independent variable (Garson, 2006). In this model a unit change in *hotel sites* (odds equal to 1) does not affect the response variable. The rest of the variables of the model

have an odds ratio larger than 1, meaning that there was an increase in the odds. Consequently, an increase in a unit in each of these variables affects the response variable. The final equation of model 1 is as follows:

$$z = -2.25 + 4.48 \times \textit{recreational area} + 11.03\times \textit{road\_b\_for} + 11.08 \times \textit{railway\_b\_for} + 22.20 \times \textit{WUI} + 1.04 \times \textit{for\_agricultur\_i} - 0.0003 \times \textit{hotel\_sites} \quad (4)$$

Table 3 also shows the marginal effects (dx/dy) of the standardized variables of the models. The marginal effects are the partial derivative of the prediction function *f* with respect to each covariate *x*, which means it is the variation of the response variable with each covariate or independent variable. The *WUI* variable has the greatest influence on the response variable (dx/dy = 0.039) followed by *road buffers in forests* (0.024), *railway buffers in forests* (0.023) and *forest-agriculture interface* (0.023). The variables with the smallest influence on the response variable are *hotel sites* and *recreational areas*.

**Table 3.** Logistic Regression results for model 1. Response variable: from real fire ignition points

| | **B** | **S.E** | **Sig.** | **Exp(B)** | **C. I. 95.0% for EXP(B)** | | **Marginal effects** |
|---|---|---|---|---|---|---|---|
| | | | | | **Lower** | **Upper** | |
| *recreational_area* | 4.48 | 1.27 | 0.00 | 88.31 | 7.27 | 1072.82 | 0.011 |
| *road_b_for* | 11.03 | 1.74 | 0.00 | 61978.71 | 2020.63 | 1901064.00 | 0.024 |
| *railway_b_for* | 11.08 | 1.70 | 0.00 | 65225.52 | 2289.84 | 1857931.82 | 0.022 |
| *WUI* | 22.20 | 2.12 | 0.00 | 4387626982.27 | 68360321.69 | 281614686115.15 | 0.039 |
| *for_agricult_i* | 1.04 | 0.19 | 0.00 | 2.83 | 1.92 | 4.18 | 0.023 |
| *hotel_sites* | 0.0003 | 0.00 | 0.00 | 1.00 | 0.99 | 1.00 | -0.075 |
| Constant | -2.25 | 0.06 | 0.00 | 0.10 | | | |

The final model correctly predicts 61.7% of the fires in the calibration data set, and 61.3% of the fire occurrences in the validation data set. The overall model adjustment to all observations is 61.6%, where the correct 'no fire' adjustment is 60.7% and correct 'fire' adjustment is 67.1%. Table 4 shows the frequency and the percentage of observed fires and those predicted by the Logistic Regression model.

**Table 4.** Classification of observed fires *versus* fires predicted by the Logistic Regression model. Response variable: real fire ignition points (Model 1)

| | **Predicted No fire** | **Predicted fire** |
|---|---|---|
| **Observed No fire** | 4401 (60.7%) | 2848 (39.3%) |
| **Observed fire** | 396 (32.4%) | 807 (67.6%) |
| **Global** | 61.6% | |

Fig. 8 shows the predicted fire occurrence probability for all the study area (a) and the correct predictions and errors (overprediction and underprediction) in the classification for each 1×1 km cell (b). In the map of the predicted probabilities of fire occurrence green cells show probability values below 0.13; cells in orange and red show high probability values, over 0.13. In the map of correct predictions and errors light green cells show where the 'no fire' events have been correctly predicted; dark green cells show where the 'fire' events have been correctly predicted; in orange, cells where there have been overprediction and red cells show where there have been underprediction.

As it can be observed in Fig. 8(a), low probability values dominate the study region, which agrees with historical fire occurrence data. High probability values are located to the west, southeast and following a line in a northern direction. Cell patterns with the highest probability values follow the main road and railway networks, as well as the *WUI* and the urban-agricultural interface. The areas where the model accurately predicts fire occurrences are mainly located to the west in the mountain range where the most important forest areas are located (*Sierra of Madrid*). Fire incidence is also correctly predicted in the north and in mid-eastern areas. Overestimations (probability values higher than 0.13 in non fire-affected cells) are sparsely located to the west and east. Finally, underestimation areas (probability values lower than 0.13 in fire affected cells) are quite dispersed throughout the study region (Fig. 8(b)).

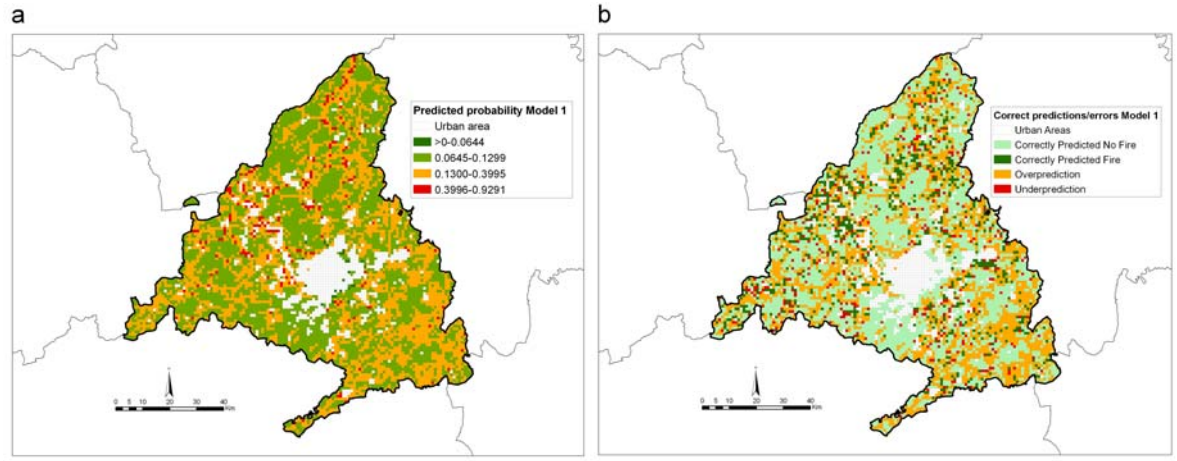



**Fig.8.** (a) Predicted probabilities of fire occurrence for model 1. (b) Correct predictions and errors for model 1. In (a) green cells show low probability values (below cut-off point); orange and red cells show high probability values (over cut-off point). In (b) light green cells show where the 'no fire' events have been correctly predicted; dark green cells show where the 'fire' events have been correctly predicted; orange cells show where there have been overprediction and red cells where there have been underprediction.

3.2 Logistic Regression model using the random fire ignition points as the response variable (Model 2)

Results of the multicollinearity analysis applied on the explaining variables showed that the variables with collinearity problems were the same ones as in model 1. Therefore, the following variables were excluded from this model: *IMD indices*, *forest trail buffers agricultural machinery* and *income*.

As in model 1, the *Wald step forward* approach was used and 10 candidate models were obtained. Among the candidate models, the one showing the best balance between number of variables (complexity) and prediction ability (in this case the model generated in step 6), was selected. Table 5 shows the logistic regression results for the model built using a 60% calibration sample. The variables selected were: *forest trail buffers in forests*, *WUI*, *forest-agriculture interface, forest-pasture interface*, *owners of agrarian holdings older than 55* and *hotel sites*. *WUI* has the highest logistic coefficient B, followed by *forest-agriculture interface* and *forest-pasture interface*. In this model a unit change in *owners of agrarian holdings older than 55* (odds lower than 1) does not

affect the response variable. The rest of the variables of the model have an odds ratio above 1, which means that there is an increase in the odds. The increase in a unit in each of these variables affects the response variable. The final equation for model 2 is as follows:

$$z = -1.58+0.98 \times trail\_b\_for + 12.11 \times WUI + 2.74 \times for\_agricult\_i + 2.16 \times for\_pasture\_i - 0.036 \times owners55 - 0.0002 \times hotel\_sites \quad (5)$$

Table 5 also shows the marginal effects (dx/dy) of the standardized variables of the models. *Forest-agriculture interface* is the variable that has a highest influence in the response variable (dx/dy = 0.063) followed by *forest-pasture interface* (0.022) and *WUI* (0.022). The variables that have the least influence in the response variable are *hotel sites* and *owners of agrarian holdings older than 55*.

**Table 5.** Logistic Regression results for model 2. Response variable: random fire ignition points

| | **B** | **S.E** | **Sig.** | **Exp(B)** | **C. I. 95.0% for EXP(B)** | | **Marginal effects** |
|---|---|---|---|---|---|---|---|
| | | | | | **Lower** | **Upper** | |
| *trails_b_for* | 0.99 | 0.21 | 0.00 | 2.682 | 1.77 | 4.06 | 0.018 |
| *for_agricult_i* | 2.75 | 0.29 | 0.00 | 15.60 | 8.92 | 27.28 | 0.063 |
| *WUI* | 12.11 | 1.96 | 0.00 | 181671.1 | 3939.60 | 8377588.93 | 0.022 |
| *for_pasture_i* | 2.16 | 0.40 | 0.00 | 8.69 | 3.99 | 18.94 | 0.022 |
| *owners55* | -0.036 | 0.003 | 0.00 | 0.97 | 0.96 | 0.97 | -0.094 |
| *hotel_sites* | -0.0002 | 5.31e-005 | 0.00 | 1.00 | 1.00 | 1.00 | -0.053 |
| Constant | -1.59 | .068 | 0.00 | 0.21 | | | |

The final model correctly predicts 59.5% of the fires in the calibration data set, and 59.3% of fire occurrence in the validation data set. Overall, the final model correctly classifies 59.8% of all observations, where the correct 'no fire' adjustment is 58.0% and the correct 'fire' adjustment is 69.1%. Table 6 shows the frequency and the percentage in the final model of the predicted fires.

**Table 6.** Classification of observed fires *versus* fires predicted by the Logistic Regression model. Response variable: random fire ignition points (Model 2)

| | **Predicted No fire** | **Predicted fire** |
|---|---|---|
| **Observed No fire** | 4138 (58.0%) | 2993 (42.0%) |
| **Observed fire** | 408 (30.9%) | 913 (69.1%) |
| **Global** | 59.8% | |

Fig. 9 shows the predicted probability for all study region (a) and the correct predictions and errors (overprediction and underprediction) in the classification for each 1×1 km cell (b). In the map of the predicted probabilities of fire occurrence green cells show low probability values below the cut-off point; cells in orange and red show high probability values, over the cut-off point. In the map of correct predictions and errors light green cells show where the 'no fire' events have been correctly predicted; dark green cells show where the 'fire' events have been correctly predicted; in orange, cells where there have been overprediction and red cells show where there have been underprediction. In the final map of predicted probability, low probability values are, as in model 1, dominant in the study region. The highest probability values are located in the north-west, following the main road infrastructures and the *WUI* in the *Sierra of Madrid*. High probability values are located in the south-east but in more isolated groups of cells. Regarding model accuracy (Fig 9(b)), the cells where the model accurately predicts fires are located in the northern and western areas of the region of Madrid (*Sierra of Madrid*). Fires are also correctly predicted in the central and southern areas. Cells where fire occurrence has been underestimated (probability values under 0.17 in fire affected cells) are dispersedly distributed throughout the study region but they are mostly located in the north and in the east, while others are in the south-west.

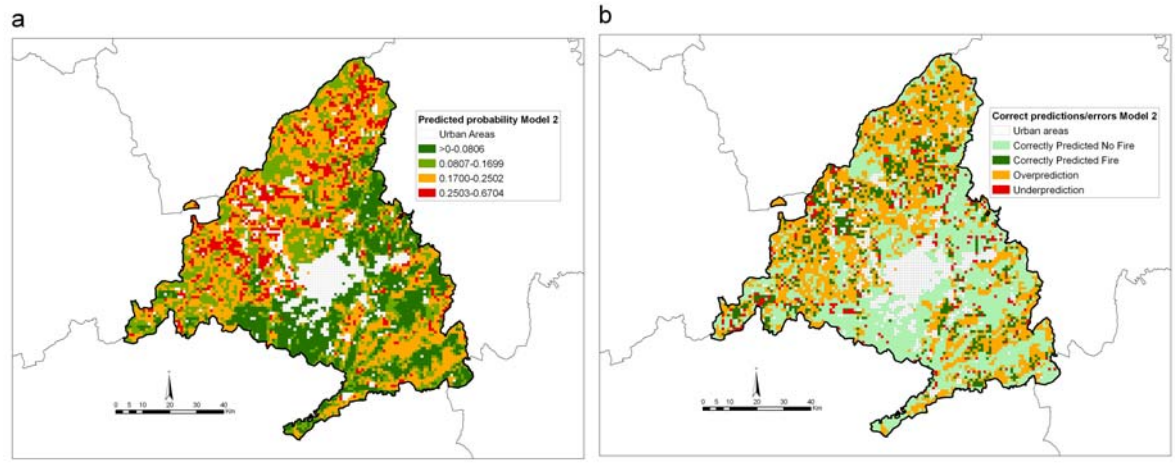



**Fig.9.** (a) Predicted probabilities of fire occurrence for model 2. (b) Correct predictions and errors for model 2. In (a) green cells show low probability values (below cut-off point); orange and red cells show high probability values (over cut-off point). In (b) light green cells show where the 'no fire' events have been correctly predicted; dark green cells show where the 'fire' events have been correctly predicted; orange cells show where there have been overprediction and red cells where there have been underprediction.

Table 7 shows a summary of the results of the statistical models using the two different response variables. It includes the explanatory variables of each model and their adjustments. The variables *WUI*, *forest-agriculture interface* and *hotel sites* explain both models. In model 1 *WUI*, *road buffers in forests* and *forest-agriculture interface* are the variables that have more influence in the response variable. In model 2 *forest-agriculture interface*, *forest-pasture interface* and *WUI* are the main ones. *Hotel sites* have little influence in both predictive models. Model 1 presents better global prediction adjustment, 61.6%, versus 59.8% of model 2. 'Fire' events have better adjustment in model 2 (69.1%) than in model 1 (67.6%). 'No fire' events have better adjustment in model 1 (60.7%) than in model 2 (58.0%). Probability values in model 1 are higher than in model 2.

**Table 7.** Summarized results of the Logistic Regression models

| **Model** | | Model 1 | Model 2 |
|---|---|---|---|
| **Explanatory variables** *(arranged according to the relative influence in the response variable)* | | WUI<br>Road buffers in forests<br>Forest-agriculture interface<br>Railway buffers in forests<br>Recreational areas<br>Hotel sites | Forest-agriculture interface<br>Forest-pasture interface<br>WUI<br>Trail buffers in forests<br>Owners older than 55<br>Hotel sites |
| **Classification adjustment** | No Fire | 60.7% | 58.0% |
| | Fire | 67.6% | 69.1% |
| | Global | 61.6% | 59.8% |

3.3 Independent Validation

Table 8 shows the independent validation results from the two models. In model 1, 74.8% of the 2006-2007 fires occurred in probability cells with values above 0.13, hence they were correctly predicted by the model. In model 2 only 65.4% of the fires occurred in cells with probability values above 0.17. Consequently, model 1 underestimated the occurrence of about 25 % of the fires in that period while the underestimation in model 2 reached 35 %. Taking into account that this is the most important error from an operational point of view because it represents an underestimation of the fire risk, it has been analyzed in more detail. Table 8 shows the percentage of fires that occurred in cells with different probability intervals: from 0 to 0.13 (predicted 'no fire' in model 1) and from 0 to 0.17 (predicted 'no fire' in model 2). In model 1, 'no fire' was located in cells with the two lowest probability intervals (P1, P2), which means that fires never occurred in areas identified as less risky by the model. On the contrary, in model 2, 12.3% of the fires occurred in low probability intervals which should be regarded as a serious underestimation of the fire ignition risk.

**Table 8.** Validation results of models 1 and 2 using real fire ignition points from 2006-2007. Predicted probability divided in 10 groups (P1-P10) with 20% of probability values each

| **Model 1** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Predicted** | | | | | | | | | |
| | No Fire | | | | | Fire | | | | |
| **Observed** | **P1 (0.0011-0.0268)** | **P2 (0.0269-0.0526)** | **P3 (0.0527-0.0783)** | **P4 (0.0784-0.1042)** | **P5 (0.1043-0.1299)** | **P6 (0.1300-0.2898)** | **P7 (0.2899-0.4496)** | **P8 (0.4497-0.6094)** | **P9 (0.6095-0.7692)** | **P10 (0.7693-0.9291)** |
| No Fire | 0.1% | 0.2% | 0.2% | 35.9% | 19.9% | 38.2% | 3.8% | 1.1% | 0.4% | 0.1% |
| Fire | 0.0% | 0.0% | 0.3% | 10.6% | 14.3% | 54.2% | 12.9% | 5.3% | 1.7% | 0.7% |
| Global | 59.4% | | | | | | | | | |
| **Model 2** | | | | | | | | | | |
| | **Predicted** | | | | | | | | | |
| | No Fire | | | | | Fire | | | | |
| **Observed** | **P1 (0.0086-0.0408)** | **P2 (0.0409-0.0731)** | **P3 (0.0732-0.1053)** | **P4 (0.1054-0.1290)** | **P5 (0.1291-0.1699)** | **P6 (0.1700-0.2700)** | **P7 (0.2701-0.3701)** | **P8 (0.3702-0.4702)** | **P9 (0.4703-0.5703)** | **P10 (0.5704-0.6704)** |
| No Fire | 14.5% | 7.7% | 5.5% | 0.0% | 24.4% | 40.0% | 6.4% | 1.3% | 0.3% | 0.0% |
| Fire | 4.3% | 8.0% | 8.0% | 0.0% | 14.2% | 45.1% | 17.3% | 3.1% | 0.0% | 0.0% |
| Global | 53.5% | | | | | | | | | |

## 4. Discussion

Two models of predicted fire occurrence from socioeconomic variables were obtained using the same explicative variables but a response variable (historical fire occurrence) which differs in its spatial accuracy. Model results differ both in their classification accuracy and in the variables they selected. In model 1, some of the variables that explain the fire occurrence (*WUI*, *forest-agriculture interface, recreational areas, hotel sites*) are related to socioeconomic changes which are bringing about changes in the influences between rural and forest use (*forest-agriculture interface*) and between urban and forest activities (*WUI*); others are related to accidents or negligence in forest areas (*road buffers*, *railway buffers*). The relationships between the response variable and the explanatory variables selected by the model are as expected, that is, positive for all variables except for *hotel sites*. However, this variable has little influence on the phenomenon as indicated by the odds ratio and marginal effects values. Regarding the variation of the response variable with each independent variable (the marginal effects) in model 1 *WUI* is the most important variable to explain fire occurrence in the region. This is in accordance with the opinion of regional fire managers who identified the increasing *WUI* as one of the main fire causes in the region of Madrid. In this region population growth has resulted in rapid development in the outlying fringe of metropolitan areas and also in rural areas with

attractive recreational and aesthetic amenities, especially forests. This has increased the number of people living in or near areas prone to wildfire. According to fire managers, these new wildland/urban inhabitants give little thought to the wildfire hazard and these results in an increasing number of negligence and/or accidental fires.

Other causes highlighted by the experts in this area were negligence, both in the recreational areas and as well as in agricultural activities, livestock farming and also accidents such as sparks coming from passing trains, agricultural machinery, etc. Official fire statistics confirm that a large percentage of fires start close to the roads and houses and an important number of accidental fires are associated to the railway infrastructure.

In model 2 the explanatory variables are also related to socioeconomic changes (*WUI, forest-agriculture interface, forest-pasture interface, hotel sites*), accidents or negligence in forest areas (*forest trail buffers in forests*) and traditional activities in rural areas (*owners of agrarian holdings older than 55*). Some of the model 2 variables such as *WUI* and *forest-agriculture interface* concur with the ones selected by model 1 while their relative influence on the response variable differs. Some other variables such as *railway buffers* or *recreational areas* have not been selected in model 2. Those variables have been replaced by others such as *forest-pasture interface* and *owners of agrarian holdings older than 55* with less apparent relationship with forest fires in Madrid region according to historical fire data and fire experts' opinion.

In model 2 the relationships between the dependent variable and the predictors are positive and significant for *forest trail buffers in forests*, *WUI*, *forest-agriculture interface* and *forest-pasture interface*. It is negative for *owners older than 55* and *hotel sites*. As indicated by their marginal effects those variables have the smallest influence on the response variable.

Although direct comparison with other similar works is not possible due to the different variables used, similarities have been found with the results of other authors working in Mediterranean areas. For instance, Martínez et al. (2009) found also positive and significant relationships between fire occurrence (number of cumulated fires per forest areas in the municipalities) in Spain and *WUI, road and railway density* among other variables. Shypard et al. (2007) observed that *population density, proportion of intermix WUI* and *mean distance to WUI* explained the greatest amount of variability related to the number of fires in California.

The probability values and its spatial distribution in the region differ between model 1 and 2 even though they have some common predictors. The distribution of the obtained

probability values in model 1 is less homogeneous, having more extreme probability values, than in model 2. In both models the highest probability values are located in the mountain range (*Sierra of Madrid*) where the most important forest areas can be found. However in model 2 probability values are above the cut-off point in most of the area while in model 1 those values are restricted to specific zones whose spatial distribution correspond with the forest areas more affected by urban spread in the last decades. Also, these cells are coincident with the road network infrastructure represented by the variable *road buffers*.

Regarding classification accuracy, model 1 reached an overall correct classification of 61.6%, where 60.7% of 'no fire' events and 67.1% of the 'fire' events were correctly classified. In model 2 the overall correct classification was 59.8%, of which 58.0% of 'no fire' events and 69.1% of 'fire' events were correctly classified. These values are in accordance with similar models proposed by different authors with classification accuracies between 68% (no fire events) to 71% (fire events) (Pew and Larsen, 2001). However, they are lower than logistic models results obtained by others, which range between 71% and 89 % for no fire events, unburned areas or low fire occurrence and up to 89% for fire events, burned areas or high fire occurrence (Vasconcelos et al., 2001; Kalabokidis et al., 2007; Martínez et al., 2009).

The overestimation in the predicted fire occurrence may be produced by the probability threshold (cut-off point) used to classify between the 'fire' and 'no fire' events. In our case we have selected the optimal cut-off point constructing the ROC curve from the training data set and intersecting the lines of the two computed statistics: sensitivity and specificity. In spite of the overestimation problems, this methodology allows a reliable model classification when the number of 'no fire' and 'fire' cases largely differs as in our study. This proportion may be affecting the accuracy of the models and the spatial distribution of the probability values.

The independent validation using real fire ignition points from 2006-2007 give us the opportunity of testing model prediction ability and compare model 1 and 2 performance. Besides, this validation strengthens the reliability of the results which showed that model 1 perform significantly better than model 2 in fire prediction. This is related with the lack of accuracy in fire location which affects model 2 response variable. Although the spatial analysis applied to the data reduced the uncertainty of fire location within the 10×10 km grid cell to a smaller polygon, the information is still too coarse to approach the required 1×1 km grid analysis. Setting a set of random points within these polygons introduce a level of uncertainty

that affects the final fire occurrence prediction model. Similar spatial has been previously suggested by other authors as a method to reduce fire location uncertainty to produce spatially continuous fire occurrence maps through kernel density approach (De la Riva et al., 2004; Koutsias et al., 2004; Amatulli et al., 2006, 2007). However, the obtained results in this work emphasize the necessity of having accurate statistics on fire location if reliable fire prediction models are to be produce and subsequently used in operational fire prevention activities.

The proposed models predict fire occurrence based on structural socioeconomic data linked to human activity. A wide range of variables that represent the human fire risk factors have been included. In this context, high probability values in a cell do not necessarily imply fire occurrence which would require an additional combination of physical or meteorological conditions for the fire to start. These variables have not been included because the aim of this work was to isolate the human influence in the prediction of the fire occurrence in the wider framework of the *Firemap* project, where the natural conditions that influence in the ignition of a fire are integrated afterwards with the human fire occurrence probability map produced by our model (Chuvieco et al., 2007).

**5. Conclusions**

FAO recent reports indicates that about 95% of wildland fires in the Mediterranean regions are due to human activities, therefore human factors should have an increasing role in fire risk systems which focus on prevention activities. Given the difficulties in modelling human behaviour, one approach is to spatially represent the human activities and factors that are directly or indirectly related with fire occurrence. The proposed work comes together to a wide group of variables for the fire occurrence modelling. This collection of variables introduced to construct the models may get most of the human component in the fire risk analysis in Mediterranean areas. This paper proposes a methodology to predict wildland fire occurrence linked to human activities in the region of Madrid at the spatial resolution which fire managers demand for operational purposes. It is evident that fire managers demand models which should be spatially accurate and that require to improve the spatial information in fire location provided by most current fire databases by including fire ignition x, y coordinates. Two models were constructed using the same predictors but different response variables. Model 1 based on

real fire ignition points performed better than model 2 (where less accurate response variable was used) not only in its predictive ability, but also in the selection of explanatory variables. The model 1 shows a more reliable fire occurrence prediction. These results were confirmed by an independent validation and by the fire experts' opinion. In this model, *WUI* is the predictor with the greatest effect on fire occurrence which seems reasonable given the high urban and recreational pressure of Madrid, a mega-cephalic region dominated by a metropolitan area of more than 4 million inhabitants, on its closest woodlands. Longer series of fire ignition points would be desirable to test the temporal and spatial strength of the model. Using GIS tools and logistic regression statistical analysis allowed characterizing human fire occurrence in the region of Madrid (Spain). These results can be included in operational fire risk systems and help in the fire risk management at regional scale.

**Acknowledgements**

This research has been partially supported by the *Firemap* project CGL2004-06049-C04-01/CLI, funded by the Spanish Ministry of Education, through the FPI scholarship BES-2005-7712. Historic fire data was provided by the Fire Department of the Community of Madrid and the Spanish Ministry of Environment, Rural and Marine Areas. Other essential data has been provided by the Regional Environmental Office of Madrid.

**References**

Afifi, A., Clark, V., (Eds.), 1990. Computer-Aided Multivariate Analysis. Van Nostrand, NY.

Amatulli, G., Camia, A., 2007. Exploring the relationships of fire occurrence variables by means of CART and MARS models. Wildfire 2007. IV International Wildfire Conference, Seville, Spain, 13-17 May.

Amatulli, G., Pérez-Cabello, F., de la Riva, J., 2007. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. Ecological modelling 200, 321-333.

Amatulli, G., Rodrigues, M.J., Trombetti, M., Lovreglio, R., 2006. Assessing long-term fire risk at local scale by means of decision tree technique. Journal of Geophysical Research 111, G04S05, doi:10.1029/2005JG000133.

Caballero, D., 2001. Particularidades del incendio forestal en el interfaz urbano. Caso de estudio en la Comunidad de Madrid. In 'II Seminario de Prevención de Incendios Forestales. Planes de Defensa contra Incendios Forestales', 28th March, ETSIM, Madrid. Available at http://www.gnomusy.com/publications/20010328_Caballero_Interfaz_UF.pdf (Last accessed December 23, 2008).

Camia, A. San-Miguel-Ayanz, J., Kucera, J., Amatulli, G. Boca, R., Libertà, G., Durrant, T., Schmuck, G., Schulte, E., Bucki, M., 2008. Forest Fires in Europe 2007. Joint Research Centre, Institute for Environment and Sustainability. EUR 23492 EN. Luxembourg: Office for Official Publications of the European Communities, 77 pp.

Cardille, J.A., Ventura, S.J., Turner, M.G., 2001. Environmental and Social Factors influencing Wildfires in the Upper Midwest, United States. Ecological Applications 11 (1), 111-127.

Chao-Chin, L., 2002. A Preliminary Test of A Human caused Fire Danger Prediction Model. Taiwan Journal Forest Science 17(4), 525-529.

Chuvieco E., Salas, F.J, Carvacho, L., Rodríguez Silva, F., 1999. Integrated fire risk mapping. In Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin .Berlin: Springer-Verlag, pp. 61-84.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Martín, M.P., Vilar, L., Martínez, J., Padrón, D., Martín, S., Salas, J., 2007. Generación de un modelo de peligro de incendios forestales mediante teledetección y SIG. Teledetección. In Martín, M.P. (Ed.) Hacia un mejor entendimiento de la dinámica global y regional, pp 19-26.

European Commission Regulation (EC) No 804/94 of 11 April 1994 laying down certain detailed rules for the application of Council Regulation (EEC) No 2158/92 as regards forest-fire information systems. Available at http://eur-lex.europa.eu (Last accessed December 23, 2008).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Fidalgo García, P., Martín Espinosa, A., 2005. Atlas Estadístico de la Comunidad de Madrid 2005. Consejería de Economía e Innovación Tecnológica. Instituto de Estadística de la Comunidad de Madrid.

Food and Agricultural Organization of the United Nations (FAO), 2007. Fire Management Global Assessment. A thematic study prepared in the framework of the Global Forest Resources Assessment 2005. FAO Forestry Paper 151. Rome, Italy. 320 pp. Available at http://www.fao.org/forestry/fra2005/en/ (Last accessed December 23, 2008).

Food and Agricultural Organization of the United Nations (FAO), 2002. Forest fire statistics 1999-2001. ECE/TIM/BULL/2002/4. Volume LV.No.4. Available at http://www.unece.org/timber/ff-stats.html (Last accessed December 23, 2008).

Garson, D., 2006. Statnotes: Topics in Multivariate Analysis. Available at http://www2.chass.ncsu.edu/garson/pa765/statnote.htm (Last accessed December 23, 2008).

Kalabokidis, K. D., Koutsias, N., Konstantinidis, P., Vasilakos, C., 2007. Multivariate analysis of landscape wildfire dynamics in a Mediterranean ecosystem of Greece. Area 39 (3), 392-402.

Koutsias, N., Kalabokidis, K.D., Allgöwer, B., 2004. Fire occurrence patterns at landscape level: beyond positional accuracy of ignition points with kernel density estimation methods. Natural Resource Modeling 17 (4), 359-375.

Lin, C., 1999. Modelling probability of ignition in Taiwan Red Pine Forests. Taiwan Journal Forest Science 14 (3), 339-344.

MARM, Spanish Ministry of Environment, Rural and Marine Affairs, 2006. Subsecretaría General de política forestal y desertificación. Área de defensa contra incendios forestales. Los incendios forestales en España. Decenio 1996-2005. Available at http://www.mma.es/portal/secciones/biodiversidad/defensa_incendios/estadisticas_incendios/pdf/estadisticasdecenio_1996-2005.pdf (Last accessed December 23, 2008).

Martell, D.L., Otukol, S., Stocks, B.J., 1987. A logistic model for predicting daily people caused forest fire occurrence in Ontario. Caumar.

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. In Chuvieco, E., Martín, M.P. (Eds.), Nuevas tecnologías para la estimación del riesgo de incendios forestales. CSIC, Instituto de Economía y Geografía, Madrid, pp. 101-142.

Martínez, J., Martín, M.P., Romero, R., Martínez, F.J., Echavarría, P., 2005. Aplicación de los SIG a los modelos de riesgo de incendios forestales: riesgo humano a escala regional. In Gurría Gascón, J.L., Hernández Carretero, A., Nieto Masot, A. (Eds.), De lo local a lo global: nuevas tecnologías de la información geográfica para el desarrollo, pp. 329-345.

Martínez, J., Vega-García, C., Chuvieco, E., 2009. Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90, 1241–1252.

McGrew, J. Jr., Monroe, C., 1993. Statistical problem solving in geography. Wm. C. Brown Communications Inc., Dubuque.

Nicolás, J.M. Caballero, D., 2001. Demanda territorial de defensa contra incendios forestales. Un caso de estudio: Comunidad de Madrid. Proceedings of Spanish National Forest Congress. Palacio de Congresos y Exposiciones, Granada, Spain, 25-28 September. Available at http://www.gnomusy.com/publications/20021025_Caballero_Geodatabases_Poster.pdf (Last accessed December 23, 2008).

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, 1-18.

Prasad, V.K., Badarinath, K.V.S., Eaturu, A., 2008. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. Journal of Environmental Management 86, 1, 1-13.

Robin, J.G., Carrega, P., Fox, D., 2006. Modelling fire ignition in the Alpes-Maritimes Department, France. A comparison. In Viegas D.X. (Ed.) V International Conference on Forest Fire Research, Figueira da Foz, Portugal, 27-30 November.

Romero-Calcerrada, R. N., C. J., Millington, J. D. A., Gomez-Jimenez I., 2008. GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). Landscape Ecology 23, 341-354.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B., 2007. Human influence on California Fire Regimes. Ecological Applications 17, 5, 1388-1402.

Vasconcelos, M.P.P., Silva, S., Tomé, M., Alvim, M., Pereira, J.M.C., 2001. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. Photogrammetric Engineering and Remote Sensing 67 (1), 73-81.

Vega García, C., Woodard, P.M, Titus, S.J., Adamowicz, W.L., Lee, B.S., 1995. A Logit Model for predicting the Daily Occurrence of Human Caused Forest Fires. International Journal Wildland Fire 5 (2), 101-111.

Vega-García, C., 2007. Propuesta metodológica para la predicción diaria de incendios forestales. Proceedings of IV International Wildfire Conference, Seville, Spain, 13-17 May.

Vélez, R., 2000. La Defensa Contra Incendios Forestales. Fundamentos y Experiencias. McGraw-Hill, Interamericana de España S.A.U, Madrid.

Vilar del Hoyo, L., Gómez Nieto, I., Martín Isabel, M.P., Martínez Vega, F.J., 2007. Análisis comparativo de diferentes métodos para la obtención de modelos de riesgo humano de incendios forestales. Wildfire 2007. IV International Wildfire Conference. Seville, Spain, 13-17 May. Available at http://www.fire.uni-freiburg.de/sevilla-2007/contributions/doc/cd/SESIONES_TEMATICAS/ST1/VilardelHoyo_et_al_SPAIN.pdf (Last accessed December 23, 2008).

Vilar del Hoyo, L., 2006. Empleo de Regresión Logística para la obtención de modelos de riesgo humano de incendios forestales. Proceeding of XII National Conference of Geographic Information Technologies. Granada, Spain, 19-22 September.

Westerling, A.L., Hidalgo, H.G., Cayan, D.R. and Swetnam, T.W., 2006. Warming and earlier spring increase western US forest wildfire activity. Science 313, 940-943.

Yang, J., He, H.S., Shifley, S.R., Gustafson, E.J., 2007. Spatial Patterns of Modern Period Human-Caused Fire Occurrence in the Missouri Ozark Highlands. Forest Science 53, 1-15.

# Capítulo III

## A Generalized Additive Model for predicting people-caused wildfire occurrence in the region of Madrid, Spain

**Publicación derivada**: *Vilar, L., Wooldford, D., Martell, D., Martín, M.P. A Generalized Additive Model for Predicting People-Caused Wildfire Occurrence in the Region of Madrid, Spain. International Journal of Wildland Fire (enviado/submitted)*

# A Generalized Additive Model for predicting people-caused wildfire occurrence in the region of Madrid, Spain

**Abstract**

This paper describes the development and validation of a spatio-temporal model for people-caused wildfire occurrence prediction at a regional scale. The study area is the 8028 km$^2$ region of Madrid, located in central Spain, where more than 90% of wildfires are caused by humans. We construct a logistic generalized additive model to estimate daily fire ignition risk at a 1km$^2$ grid spatial resolution. Spatially referenced socioeconomic and weather variables appear as covariates in the model. Spatial and temporal effects are also included. The model was selected using an iterative approach, based on the generalized cross-validation. We use the model to predict the expected number of fires in our study area during the 2002-2005 period, by aggregating the estimated probabilities on space-time scales of interest. The estimated partial effects of the amounts of railways and of roads in forested areas, and the amount of wildland urban interface were highly significant, as were the maximum temperature and precipitation.

**Additional keywords:** Fire risk, GIS, Logistic, Nonparametric spline smoothing, Socioeconomic variables, Wildland fire, Wildland Urban Interface.

## 1. Introduction

Wildland fires occur in non-urban settings and can be a significant disturbance factor in many ecosystems, including the Mediterranean (Pausas and Vallejo, 1999; Leone et al., 2003). Wildland fires can be a natural disturbance, such as those fires ignited by lightning, or they can be caused by humans. For the latter case, some authors have identified anthropogenic activity as playing an important role in altering natural functioning of ecosystems during the 20th century (Campbell and Liegel, 1996). Conversely, wildland fires can significantly impact human behavior. For example, due to their perceived potential to cause social, economic and ecological damage, wildland fires are commonly fought and suppressed by fire management agencies. Fires which occur at the interface of urban and/or agricultural developments and wildland settings are of particular concern, since they have the potential to cause significant damage to infrastructure and housing, and can threaten human safety.

Recent socioeconomic, cultural and political changes in Europe have brought important economic, production and social transformations to rural areas (Moyano, 2007). Although 52% of Spain is forested (Ministry of Agriculture, Fisheries and Food 2004), forestry related industry only contributes 0.15% to the national economy (Vélez, 2005). In addition, the spread of residences into rural areas has increased the size of the wildland urban interface (which is commonly referred as WUI). Furthermore, urban individuals are increasingly using forested land for recreational purposes (Izquierdo, 2007). This increase of human activity on the landscape has the potential to intensify wildfire occurrence in these areas and consequently has led to increased management concerns. Fire prevention is an important component of any fire management program and a sound understanding of the phenomenon and the ability to predict people-caused fire occurrence is essential to wildland fire management decision support systems.

Historical wildfire records have been compiled in Spain since 1968. These records reveal that recently (from 1996 to 2005) more than 90% of all wildland fire ignitions were caused by humans (Ministry of the Environment, Rural and Marine Affairs -MARM-, 2006). Those fire records classify the fires by their source of ignition into the following main categories: *lightning*, *negligence (accidents) and others*, *fires that are deliberately set (arson)*, *unknown* and *re-ignited fires*[10].

[10] Reignited fire: : the re-occurrence of a wildfire that was previously under control.

During this period, 60.6% of fires were deliberately set (e.g., arson, illegal agricultural burning, burning to create/regenerate pasture lands, fires related to hunting, etc.), while negligence and accidents accounted for an additional 17.7% of wildfires. Lightning accounted for only 3.6% of all fires and 1.8% of fires were re-ignitions. The cause of the remaining 16.2% of fires is unknown.

The human component of wildfire ignition risk is difficult to model since, in many cases, it requires the identification, quantification and mapping of behavioral factors. As noted by Leone et al. (2003), of the most important factors that have a direct or indirect influence on fire occurrence in Euro-Mediterranean environments are socioeconomic changes, traditional activities in rural areas, accidents or negligence, fire prevention activities, and factors which can lead to social unrest (e.g., land use disputes or high unemployment rates). Models that estimate the probability of human-caused wildfires can be generated by analyzing such factors along with relevant weather information.

Within the framework of the *Firemap* research project ('Integrated Analysis of Wildland Fire with Remote Sensing and GIS') a wildland fire risk index was developed which integrates the human and environmental factors related to fire ignition as well as moisture content of fuels (Chuvieco et al., 2009). The human factors that were identified in the conceptual framework of the *Firemap* project were included in our analysis.

The first human caused predictive models proposed in the context of fire risk estimation were based on negative binomial (Bruce, 1963) and Poisson distributions (Cunningham and Martell, 1973). Then, the fire factors were widely studied and the probability of ignition was calculated using logistic regression techniques (Martell et al., 1987; Lorenzo and Pérez, 1995). This technique quantifies the relationships between a dichotomous response variable (i.e., the presence or absence of a fire ignition) and explanatory variables. Some authors have employed logistic regression to obtain fire risk models at regional (Chuvieco et al., 1999; Martínez et al., 2004) and local scales (Vega-García et al., 1995; Lin, 1999; Pew and Larsen, 2001; Vasconcelos et al., 2001). Others have used this technique for temporal models that predict daily human-caused occurrence such as Martell et al. (1987), Loftsgaarden and Andrews (1992) and Vega-García et al. (1995).

Many studies attempt to quantify the risk of human caused wildfires in terms of socioeconomic, demographic and other human land-use covariates. The advent of relevant GIS

tools has facilitated the integration of spatially referenced variables into such models. For example, some authors have built models for the probability of human-caused wildfires using spatial data regarding transportation infrastructure: distance to roads (Vasconcelos et al., 2001), railway and road density (Cardille et al., 2001). Other authors have taken into account variables related to the population: population density (Cardille et al., 2001), demographic potential, variation of the population (Martínez et al., 2009), rural population density and rural population to forest area ratio (Prasad et al., 2008). There are also studies that use variables related to human infrastructure or activities and urban spread: distance to campsites (Vega García et al., 1995), distance to human-built infrastructures (Pew and Larsen, 2001), distance to city (Cardille et al., 2001), percentage of urbanized area (Martínez et al., 2009), industrial units, continuous/discontinuous urban and industrial development, port areas (Amatulli et al., 2006), WUI, low-density housing, change in housing density (Syphard et al., 2007). Further studies employ variables related to land and property: total area privately owned (Vega García et al., 1995), proportion of owner-occupied units (Cardille et al., 2001). A final set of related research investigates fire risk due to agricultural activity: distance to agricultural fields (Vasconcelos et al., 2001), forest-culture interface (Martínez et al., 2009), density of livestock (Kalabokidis et al., 2007).

An important statistical development of the last thirty years has been an advance in statistical modelling methods. In particular, generalized linear regression models have been extended to generalized additive models (GAMs), which allow for non-linear relationships between the response variable and covariates. Models which employ this methodology have a great potential for application in many fields of scientific research due to their ability to deal with the multitude of distributions that define ecological data and due to the fact that they blend in very well with traditional practices used in linear modelling and analysis of variance (Guisan et al., 2002).

A recent group of papers used GAM-based methods to develop fine-scale spatial-temporal models of wildfire ignition risk for regions in the United States (Brillinger et al., 2003, 2006; Preisler et al., 2004) and for forecasting large fire events at a coarser scale (Preisler et al., 2007, 2008). Although the modelling we present parallels the methods in the former set of papers, we focus on the incorporation and estimation of nonlinear relationships between wildland fire ignition risk and spatially referenced socioeconomic variables, including road,

railway and wildland urban interface densities. To our knowledge, this is the first time the nonlinear effects of such human-land use characteristics have been quantified.

The objective of our analysis is to construct a spatio-temporal model that predicts the probability of people-caused wildfire ignitions given these, and other, relevant variables. The covariate effects included in our model were identified by applying an iterative model selection approach that is based on the generalized cross-validation score, and is analogous to model selection using AIC (Akaike, 1973). The paper is structured as follows: the following section describes the study area, the data, and the statistical methods in more detail. In section 3 we outline our model and examine its fit and predictive ability. Section 4 discusses and compares the results to previous studies. The paper ends with a brief concluding discussion.

## 2. Methods

### 2.1 Study Area

The region of Madrid is located in central Spain and is illustrated in Fig. 1. It is primarily composed of two subregions: a mountain range that runs from the northeast (NE) to the southwest (SW) (this part of the region, with its highest peak having an elevation of 2428 m and a width that ranges from 15-35 km), and the Tagus river basin, which runs from the fault of the mountain range (800 m above the sea level) to the Tagus' bed. (Fidalgo and Martín, 2005). The region has a Mediterranean climate, but due to the distance to the sea and the orographic barrier, the annual precipitation is lower and the thermal amplitude is higher, with warm summers and cold winters. Pastures and shrublands are also common features on this landscape. The forested areas are primarily located in the mountain range, where deciduous trees occupy more than 75% of the total forest land. The holm oak (*Quercus ilex*) is the most abundant specie, represented by the *dehesa* formation. This vegetal formation has important landscape, recreational, environmental and livestock values.

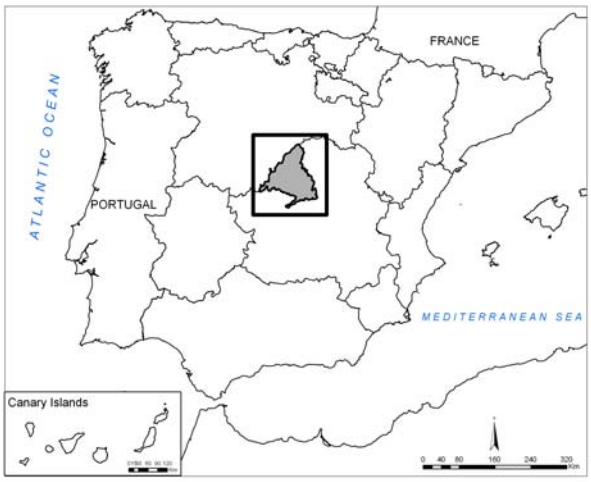



**Fig.1.** Administrative community borders in Spain and the location of the study area

Although it represents only 1.6% of the nation's area, the region of Madrid with more than 6 million inhabitants is one of the most populated regions in Spain. Its economy is based on the service sector and its transport structure it is composed of more than 3000 km of roads and 300 km of railway lines. These land uses of Madrid are illustrated in Fig. 2. Fig. 2a shows the main land uses from Corine Land Cover 2000 source (http://etc-lusi.eionet.europa.eu/CLC2000, April 2009): urban, agricultural and forest areas. It also shows the spatial location of the recreational areas, distributed mainly in the forest areas. Fig. 2b illustrates the dense road network, the WUI and the Natural Protected Areas in the region.

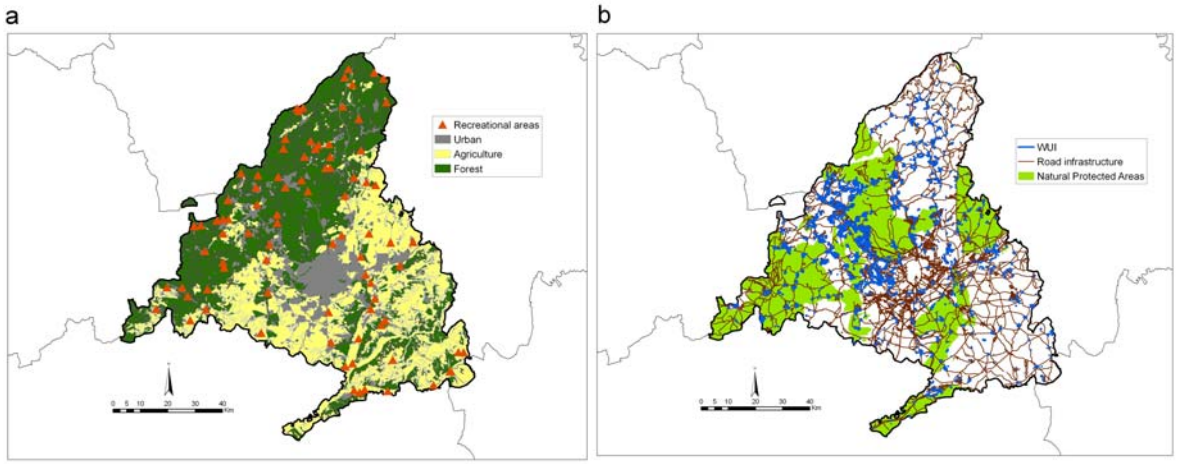



**Fig.2**. (a) Land Uses in the region of Madrid. Location of recreational areas, urban, agricultural, forest and others land uses from *Corine Land Cover 2000* (http://etc-lusi.eionet.europa.eu/CLC2000, April 2009); (b) Road infrastructure, WUI and natural protected areas.

Urban areas in the region of Madrid have been increasing in both size and population since the 1960s, and have been spreading into agricultural and forested areas in dispersed urban developments. Due to high population density and pressure, land uses have changed, affecting agricultural and forested areas. In general, the spatial distribution of wildfire ignitions in this region commonly associated with these urban areas as well as proximity to roads, railways and dump sites (Nicolás and Caballero, 2001). The contact boundary between urban and forest (WUI) is an area of major concern for fire managers in this region because wildfires which occur in this interface have the potential to endanger human life, housing, infrastructure and other values at risk. Despite the large urban sprawl, the region of Madrid maintains a network of protected areas which are intensively used as recreational areas and consequently are also very vulnerable to forest fires. The national fire records for our study period reveal that most of the accidental/negligent wildfires were caused by forest works, railway infrastructure, accidents by machinery and agricultural burning. In addition, over 70% of ignitions occurred close to roads, trails or houses near or in forested areas (MARM, 2006).

## 2.2 The Data

### 2.2.1 The fire data

The fire occurrence records available for the study area describe the ignition location coordinates and dates for all fires that occurred in the region of Madrid during the fire season (from May through October) between 2002 and 2005, inclusive. Months outside of this time interval have been excluded due to low wildland fire activity (less than 9% of annual fire ignitions on average). We are restricted to data from 2002 onwards, since weather data has only been available since that year. The locations of these ignitions are illustrated in Fig. 3.

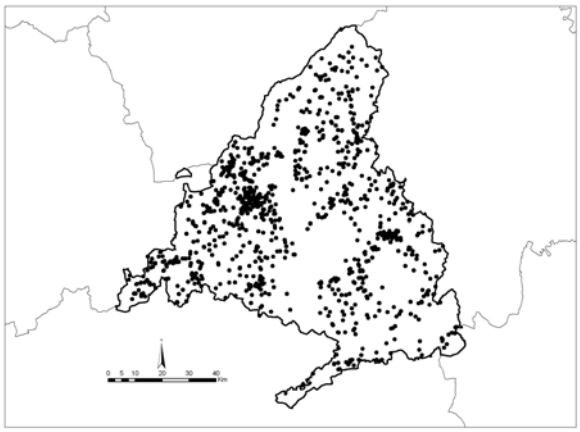


**Fig.3**. Location of fire ignition points from May to October of 2002-2005 (*Source*: Fire Department of the region of Madrid)

The total number of fires on each day in our data set appears in Fig. 4, which illustrates the increase in the number of fires during the summer season. Fire occurrence peaks in July. The database including information on fire ignition coordinates provided by the Fire Department of the Madrid region contains a total of 1026 fire ignition points during our study period (2002-2005). This database does not include information on the source of ignition. However, given that historical statistics indicate that over 90% of wildland fires were due to human activity, we make the simplifying assumption that all these fires were people-caused.

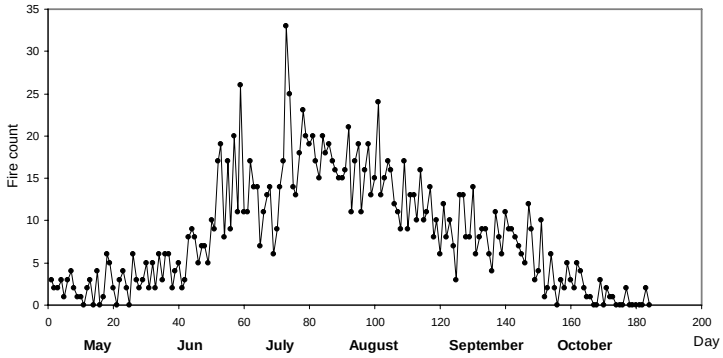



**Fig.4**. Total Fires by day from May to October 2002-2005 (*Source*: Fire Department of the region of Madrid)

2.2.2 Explanatory variables

The set of potential explanatory variables that were spatially referenced for our analysis can be divided into two broad categories: socioeconomic and environmental. The socioeconomic variables are static and can be assumed constant during the study period, while the environmental variables include elevation and dynamic weather data.

As we have indicated, anthropogenic activity has influenced fire occurrence in our study region and there are several variables that can be used to represent factors related to people-caused wildfire ignitions in Euro-Mediterranean environments. In this work we consider a set of variables selected from a list of socioeconomic variables identified in Vilar et al. (2008). These include variables such as the road and railway network, the WUI, electric lines, population density, recreational areas, etc. All variables were spatially mapped as densities for each 1 $km^2$ grid cell (Table 1).

**Table 1.** Socioeconomic predictors: human factor which affect fire occurrence and predictor variables that represent each factor. The units are density of each variable by 1 km² voxel

| Factor | Variable Name | Description | Range |
|---|---|---|---|
| **Accident or negligence** | buffer_r_for | Buffer of roads in forested areas | [0.00, 0.31] |
| | buffer_ rail_for | Buffer of railways in forested areas | [0.00, 0.24] |
| **Socioeconomic changes** | WUI | Wildland Urban Interface | [0.00, 0.19] |

It is well known that weather is a driving factor that influences fuel moisture and fire risk, regardless of the source of ignition. For example, previous research has included environmental variables such as precipitation, temperature or elevation along with human predictors (Pew and Larsen, 2001; Amatulli et al., 2006; Kalabokidis et al., 2007; Maingi et al., 2007; Prasad et al., 2008). Consequently, we also considered weather variables (see Table 2) supplied by Meteológica (Spanish company for meteorological information). This processed data had been interpolated in a 3×3 km UTM grid from the European Centre for Medium range Weather Forecasting (ECMWF) outputs. After, the data has been referred to the 1 km² resolution using GIS tools. Similar to Preisler et al. (2004), the spatial location of each grid cell was also included to account for dependencies amongst nearby points as well as spatially explicit topography or vegetation characteristics not captured by any variable(s) in the model. Also, we include the elevation variable, coming from a DEM (Digital Elevation Model) 100×100 m of the Iberian Peninsula. Finally, we include a temporal effect as a function of the day of the year, to account for temporal dependencies and hence, identify the general seasonal trend within each fire season.

**Table 2.** Natural predictors: weather and elevation data

| Variable Name | Description | Range of values (units) |
|---|---|---|
| Weather data | | |
| tmax | Maximum temperature | [3.6, 41.3] ºC |
| tmin | Minimum temperature | [-25.5, -3.1] ºC |
| p24 | 24 h rain | [0.0, 59.8] mm |
| Windspeed | Wind speed | [0.04, 8.46] m/s |
| SunRad | Solar Radiation | [2 165, 32 262] kJ/m²/day |
| Pres_vap | Steam pressure | [0.48, 3.26] kPa |
| Elevation | | [439.032, 2291.67] m |

## 2.3 Statistical methods

Generalized additive models (GAMs) extend the well known and frequently used class of generalized linear models (GLMs), by incorporating smooth, non-linear covariate effect(s) into the model structure. Similar to a GLM, a GAM uses a link function to establish an additive relationship between the mean of the response variable and the explanatory variables (Guisan et al., 2002; Wood and Augustin, 2002). A detailed description of how smooth functions can be expressed as linear combinations of basis functions and how this leads to GAMs constructed using penalized regression splines appears in Wood and Augustin (2002). This is the methodology employed herein.

For the case of human-caused wildland fire occurrence in the region of Madrid, a logistic-GAM model can be used to produce maps of predicted ignition probabilities. An estimate of the total number of expected fires in a given region and time period can then be obtained by aggregating these probabilities over a space-time neighborhood of, or by choosing an appropriate threshold limit and then classifying fitted probabilities as fire ignition events or non-events. We employ the former method as in our work as both goodness of fit measure and in our variable selection technique (we found that the latter method led to too high a proportion of "false positives").

Following Brillinger et al. (2003), we constructed a dichotomous response variable, by defining the following binary variable on our 1km$^2$ × 1 day set of voxels: $N_{xyt}= 1$ if there is at least one fire at location $(X, Y)$ during time period $t$ and 0 otherwise. Then, similar to Preisler et al. (2004) the probability of fire occurrence is defined as

$$P= Prob\{N_{xyt}=1 \mid U_{xyt}\} = \exp(\theta_{xyt}) / [1+\exp(\theta_{xyt})] \qquad (1)$$

where $P$ is the probability of fire occurrence in the cell located at $(x, y)$ on day $t$; $U_{xyt}$ is the collection of the values of explanatory variables; and $\theta_{xyt}$ is a vector of parameters to be estimated.

One complication from such a setup is the sheer volume of data that arises. For example, the grid of our study area consists of 8452 one square kilometer cells, and there are 184 daily observations for each cell during each of the four years in question. This results in a large

number of observations [number of grid cells (8452) × number of years (4) × number of days (184) = 6 220 672]. However, on this fine of a space-time scale, fires are a rare event since nearly all of the voxels do not contain a fire event. Consequently, a sample of the data from the zero-fire voxels yields sufficient covariate information on the non-ignitions for model building. To reduce the data set to a manageable (i.e., computationally feasible) size, previous authors (e.g., Vega Garcia et al., 1995; Brillinger et al., 2003; Preisler et al., 2004) retained all observations corresponding to a fire event and randomly selected data from a small proportion of the non-fire observations. This introduces a deterministic offset term into the model, which can easily be incorporated into this modelling framework and hence, does not bias the analysis. For our modelling a sample of 1% of the zero-fire voxels appears to be sufficient.

After employing such a sampling scheme, we partitioned our data into two sets. Observations from 2002 through 2004 inclusive were used to build and select a model, while observations from 2005 were reserved for validation of the fitted model. To test if there is any difference between the average number of fires by day between this period (2002-2005) and recent historical data (1990-2001) we applied the t-test for two independent samples. We found that at the 95% confidence level there were no differences between the average number of fires per day during the study period versus historical period. Hence, we postulate that our results may be applicable to historical data sets that extend further into the past. Model selection was based on the approach discussed by Wood and Augustin (2002, section 3.3), where they suggest selecting model terms that 1. have estimated degrees of freedom above the lower limit of one for a univariate smooth; 2. have confidence regions that do not include zero over the entire range of the covariate; and, 3. reduce the Generalized Cross Validation (GCV) score. Note that the GCV score (Craven and Wahba ,1979; Golub et al., 1979) is essentially a modification of the ordinary cross validation score, which estimates the average squared prediction error over all data sets where a single observation is omitted from the model fitting, and then predicted.

Starting with a model containing the additive partial effects of space and of time, we fit a sequence of nested models containing the preceding effects and one additional term. From the set of candidates that satisfied conditions 1 through 3 above, we selected the model that had the smallest GCV score. Hypothesis testing, based on analysis of deviance, was then performed to determine whether there was a statistically significant improvement in fit (between the updated model and the model from the previous step). This process was iterated until we arrived at the

model we present herein. We remark that a similar step-forward technique was employed by Preisler et al. (2008) to construct a model for fire ignition risk, where mutual information (MI) was used as their model selection criteria. Moreover, both the GCV score and the MI statistic have similarities with the AIC score, which is a widely used model selection criterion. Also, in the context of GAMs, the use of the GCV score is appealing due to its consistency with its use for smoothing parameter selection (Wood and Augustin, 2008). Models were fit using the gam function in the mgcv package (Wood, 2006) for R (R Development Core Team, 2007).

## 3. Results

As a preliminary analysis, we explored the empirical relationships between fire ignitions and each of the covariates, as well as correlations between covariate pairs. This was done to identify a set of candidate variables to include in our model selection process. Candidate variables were identified as those which appeared to be both associated with the response variable and supported by our understanding of the processes underlying wildfire ignitions. For example, we examined plots of the logit of the empirical proportions of fire ignitions across bins that partitioned the range of each covariate. Other visual summary methods comparing covariate values for the fire and non-fire events, such as histograms and quantile-quantile plots were also examined. The set of candidate variables (besides space and time) we identified were *temperature, precipitation, the buffers of railways, and of roads, in forested areas, as well as buffers of the interfaces between forested areas and urban, agricultural and pastured areas.* In particular, we observed increasing trends between temperature, road and railways buffers in forested areas, as well as buffers for the interfaces between urban, agricultural and pasture areas. Decreasing trends in ignition rates were observed for precipitation and elevation. Empirical associations did not appear as strong for the remaining potential explanatory variables.

### 3.1 A Model for People-Caused Wildfire Ignitions

Similar to Preisler et al. (2004) for notational simplicity, we let $k$ index the set of space-time voxels over our study area and period. Then let $p_k$ denote the probability of ignition at voxel $k$ given the corresponding set of observed covariates. Using the variables identified in our exploratory analysis we followed the model selection technique of Wood and Augustin (2002)

outlined in section 2.4 and obtained the following logistic GAM as a model for people-caused wildfire ignitions in the region of Madrid:

$$\begin{aligned} logit\,(p_k) = \beta_0 &+ f_1\,(day\ of\ year_k) + f_2\,(X_k, Y_k) + f_3\,(WUI_k) \\ &+ f_4\,(buffer\ roads\ in\ forested\ areas_k) + f_5\,(elevation_k) \\ &+ f_6\,(maximum\ temperature_k) + f_7\,(buffer\ of\ railways\ in\ forested\ areas_k) \\ &+ \beta_2 I\,(precipitation_k > 0) \end{aligned} \quad (2)$$

Our selected model contains an intercept $\beta_0$, along with eight additive covariate effects. Non-linear effects were estimated for the day of year, the spatial location (X, Y), the buffers of WUI, and of roads and of railways in forested areas, maximum temperature and elevation. The effect of precipitation was best modelled via a jump process, with a coefficient of $\beta_2$ when precipitation occurs. The p-values and estimates for the intercept and precipitation coefficient appear in Table 3(a). The estimated p-values for the non-linear effects appear in Table 3(b). All terms in our model were highly statistically significant. The GCV score decreased at each step of our iterative model selection process, and analysis of deviance tests indicated that each step led to statistically significant improvements in fit.

**Table 3(a).** Estimated linear coefficients and their significance

| Variable | Estimated Coefficient | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| Intercept | -7.16 | 0.096 | -74.665 | $< 2 \times 10^{-16}$ |
| I(precipitation > 0 ) | -0.48 | 0.171 | -2.867 | 0.004 |

**Table 3(b).** Estimated significance of the smooth terms

| Variable | edf | Est.rank | Chi.sq | p-value |
|---|---|---|---|---|
| day of year | 6.5 | 7 | 145.31 | $< 2 \times 10^{-16}$ |
| space | 21.4 | 24 | 381.70 | $< 2 \times 10^{-16}$ |
| WUI | 2.9 | 3 | 273.05 | $< 2 \times 10^{-16}$ |
| buffer of roadways in forested areas | 2.8 | 3 | 122.08 | $< 2 \times 10^{-16}$ |
| elevation | 1.9 | 2 | 76.10 | $< 2 \times 10^{-16}$ |
| maximum temperature | 1.9 | 2 | 38.91 | $1.5 \times 10^{-08}$ |
| buffer of railways in forested areas | 1.9 | 3 | 49.33 | $1.0 \times 10^{-10}$ |

Fig. 5 illustrates the estimated partial effects plots for each term in the model, plotted on the logit scale with point-wise 95% confidence bands. Fig. 5a displays the seasonality of fire ignition risk as a function of day of year. Note that, it increases from early spring to mid summer, peaking around day 200 (approximately late July), and then decreases as summer turns to fall. The partial effect of elevation (Fig. 5b) exhibits a negative trend, which is postulated to occur due to less human land use and activity at high elevations. Temperature (Fig. 5c) and wildland urban interface (Fig. 5d) have, as expected, increasing trends. Similarly, the road and railway density effects (Figs. 5e and 5f) also follow increasing trends. However, the road density buffer effect increases and appears to plateau, after which the estimated effect is quite uncertain as indicated by the widening confidence bands. The estimated spatial effect (Fig. 6) is relatively flat, except for a large drop in risk at the southernmost tip, and in the center of the region where the city of Madrid is located.

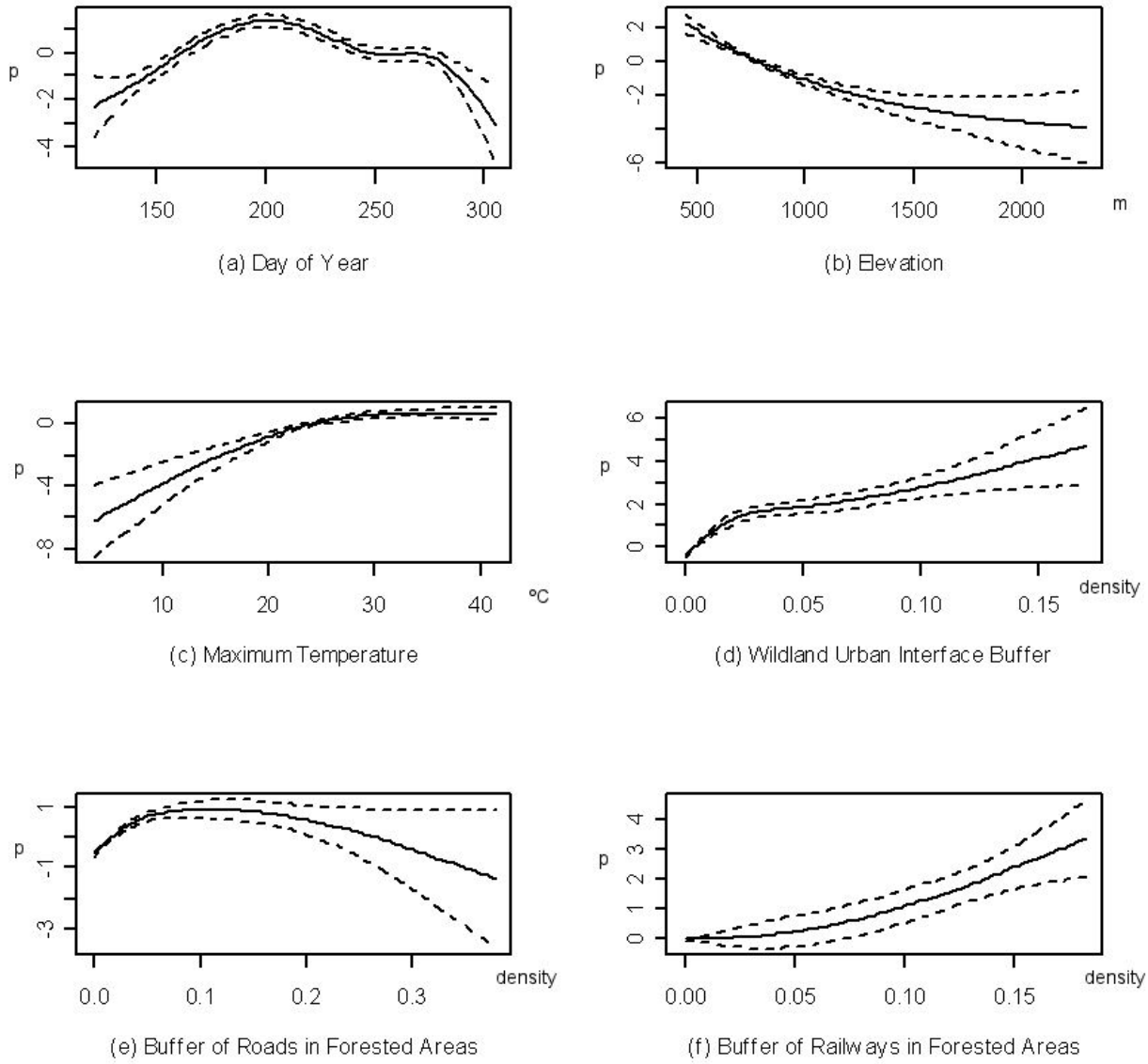



**Fig.5**. Estimated partial effects of explanatory variables (solid line) with 95% confidence bands (dashed lines): (a) day of year; (b) elevation; (c) maximum temperature; (d) WUI buffer; (e) buffer of roads in forested area; (f) buffer of railways in forested areas.

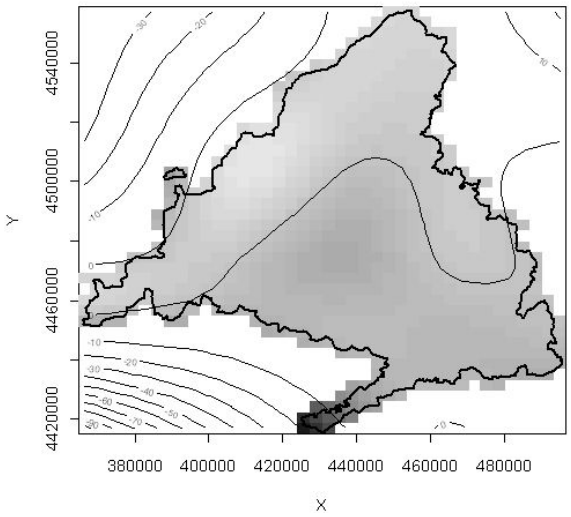



**Fig.6.** Estimated partial effects of explanatory: spatial effect.

Table 4 illustrates the predictive ability of the model for both the calibration sample (a), and the validation sample (b). Similarly, the overall monthly predictive ability is summarized in Table 5.

**Table 4.** Classification results of the fitted model: Calibration sample, years 2002-2004 from May to October (a), Validation sample, year 2005 from May to October (b)

| **(a) Calibration sample** | | |
|---|---|---|
| **Year** | **Observed fire** | **Predicted fire** |
| 2002 | 304 | 267 |
| 2003 | 177 | 254 |
| 2004 | 289 | 248 |
| **(b) Validation sample** | | |
| **Year** | **Observed fire** | **Predicted fire** |
| 2005 | 245 | 276 |

**Table 5.** Classification results of the fitted model by month: Calibration sample, years 2002-2004 from May to October (a), Validation sample, year 2005 from May to October (b)

| **(a) Calibration sample, years 2002-2004** | | |
|---|---|---|
| **Month** | **Observed fire** | **Predicted fire** |
| May | 31 | 31 |
| June | 147 | 134 |
| July | 261 | 275 |
| August | 193 | 195 |
| September | 116 | 107 |
| October | 22 | 25 |
| **(b) Validation sample, year 2005** | | |
| **Month** | **Observed fire** | **Predicted fire** |
| May | 30 | 14 |
| June | 30 | 49 |
| July | 84 | 98 |
| August | 58 | 64 |
| September | 36 | 36 |
| October | 7 | 12 |

A visual assessment of the model's predictive ability appears in Fig. 7, where observed and expected number of fire ignitions are plotted across months year-by-year.

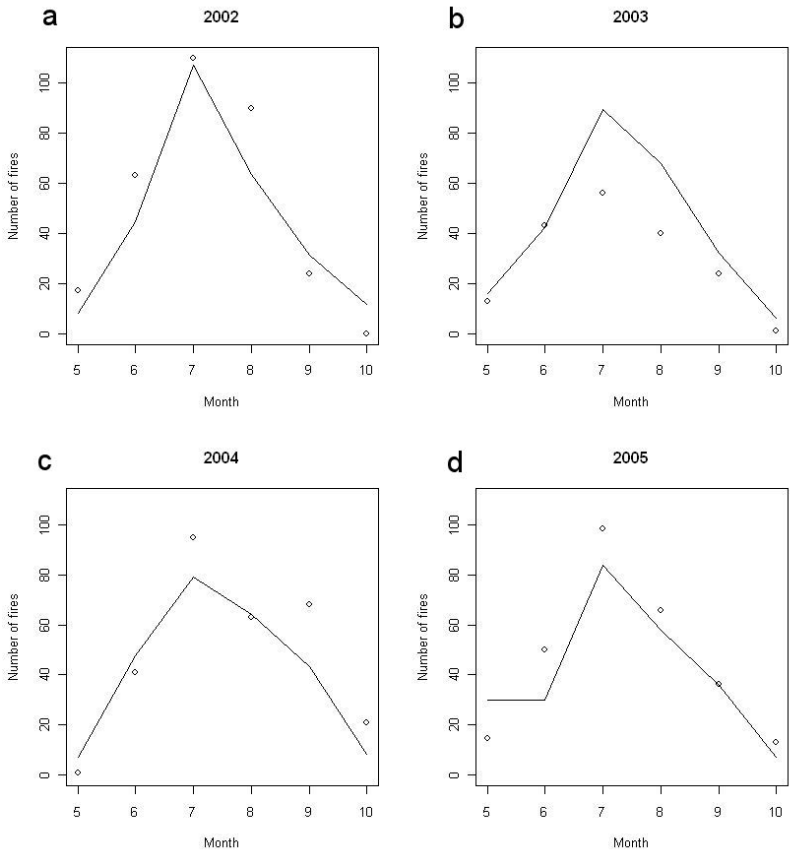



**Fig.7.** Classification accuracy of the fitted model for the calibration and validation samples by month: (a) year 2002; (b) 2003; (c) 2004; (d) 2005. Lines represent expected fires and dots observed fires.

For all of the preceding, the expected number of fires was calculated by summing the fitted (or predicted) probabilities over the time-period of interest. Note that these probabilities are easily obtained via an inverse-logistic transformation of values returned by the model. We note that although there are inaccuracies, such as a somewhat large over-prediction in 2003, as well as other over/under predictions elsewhere, we feel that the model fits quite well especially given the small amount of data used for model fitting. We suspect that as more data becomes available, the model fit will continue to improve, especially in terms of the width of confidence intervals for the estimated coefficients and partial effects.

## 4. Discussion

The GAM framework is preferred over traditional logistic generalized linear models, since we are able to estimate nonlinear relationships between ignition risk and covariates. This has a distinct advantage when estimating intra-annual seasonality effects, which previously had to be incorporated via polynomial or periodic functions (see e.g., Martell and Belivacqua, 1989). Moreover, the advantage when estimating non-linear relationships between weather and/or socioeconomic variables should also be clear: for example, although fire risk should increase as *WUI* increases, one would not expect it to do so linearly.

The model selection technique we employed illustrates a stepwise method that follows the suggestions of Wood and Augustin (2002), and demonstrated that one can use this algorithm to obtain a sensible model while limiting the number of statistically significant independent variables to only those that lead to an improvement in model fit and/or prediction. Hence, we are able to determine an adequate, yet parsimonious model.

The set of explanatory variables included in our model had estimated effects that were highly significant. In addition, the coefficients and partial effects estimated all exhibit intuitively sensible trends. The estimated intercept is negative and large, relative to the other range of the other estimated effects, reflecting the fact that fire ignitions are a rare event on the fine spatio-temporal scale of daily 1 $km^2$ voxels. Precipitation and increasing elevation also have dampening effects on ignition risk, with the latter effect likely explained by decreased human land use and activity at very high elevations. Moreover, it is widely accepted that increasing temperatures are associated with decreases in fuel moisture content and hence, an increase in fire ignition risk. This effect is clearly visible in the estimate partial effect for maximum temperature. Recalling that more than 90% of fires are due to human activities in this region, we observed that ignition risk increases as the density of urban, road or railway developments increase. However, the uncertainty of these effects, as measured by the width of the corresponding confidence bands, also increases over the range of each covariate, especially when considering the effect of road density. Note that increases in uncertainty are due to limited numbers of observations at the higher ranges of these covariates. Finally, although the estimated spatial effect was relatively flat, it decreased in regions where there were little-to-no fire ignitions, such as in the southernmost tip of the study area, and near the city of Madrid in

the highly urbanized center of the region. The estimated effects can be used to identify when and where fire ignition risk is higher than normal. For example, if a location has *buffers of railways in forested areas* density values greater than 0.05 and *WUI* density values greater than 0.01, then on days when the *maximum temperature* is greater than 30ºC, the model would predict higher number of fires than the average for that location and time of the year.

Previous related research using GAMs for ignition risk has focused on quantifying temporal, spatial and/or spatio-temporal ignition risk as a function of weather, fire danger indices or topography (e.g., Brillinger et al., 2003, 2006; Preisler et al., 2004, 2007, 2008). While other researchers have used logistic GLMs to look for associations between ignition risk and anthropogenic variables (e.g., Pew and Larsen, 2001; Vasconcelos et al., 2001; Prasad et al., 2008), we have illustrated how GIS tools can be used to create a set of spatially-referenced human land-use variables, and how logistic GAMs can also be employed to quantify non-linear relationships between ignition risk and these anthropogenic characteristics. To our knowledge, this is the first such study.

In spite of the lack of ignition source information and the limited temporal data set used for model building, we found that the model fit quite well overall. Goodness of fit was evaluated by comparing the observed number of fire ignition events to expected, where the expected number of fires was estimated by summing probabilities predicted by the model of the period of interest. Further improvements in goodness of fit would likely be attained as more data becomes available for analysis, especially if such data specifically identifies the source of ignition. The future goal of this model would be its integration in a risk system for the fire management, where the human risk factors are not currently adequately modelled.

## 5. Conclusions

Human activities are the main cause of fires in the Mediterranean countries. Historically, in our study region over 90% of the forest fires are due to human ignitions. We illustrated how this human component can be modelled and predicted in order to integrate it in the general fire ignition risk estimation systems with a spatial-temporal dimension. The proposed model includes spatially-referenced socioeconomic variables that are used to represent human land use patterns in the study region, the usefulness of which was demonstrated in previous related

research. The model we present shows a reasonable accuracy by incorporating dynamic temporal effects such as the seasonality modelled using a nonlinear intra-annual effect and the maximum temperature or precipitation. Further refinement and assessment of the model is anticipated as more and better data becomes available.

**Acknowledgements**

This research has received partial support from the Firemap project CGL2004-06049-C04-01/CLI, funded by the Spanish Ministry of Education, through the FPI scholarship BES-2005-7712. Additional funding from the National Institute for Complex Data Structures (NICDS) and Geomatics for Informed Decisions (GEOIDE SII Project 51) is also gratefully acknowledged. We would like to thank *Inmaculada Aguado, Mariano García, Héctor Nieto, Marta Yebra and Felipe Verdú* from the Department of Geography of the University of Alcalá (Spain), for their advice and provided information. We would like also thank *Robert Kruus* from the Fire Management Systems Laboratory in the Faculty of Forestry at the University of Toronto (Canada) for his assistance in data base preparation. Historic fire data has been provided by the Fire Department of the region of Madrid and the Spanish Ministry of Environment, while other data was provided by the Madrid Regional Environmental Office.

**References**

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. Petran and F. Csaaki (Eds.), International Symposium on Information Theory. Akadeemiai Kiadi, Budapest, Hungary, pp. 267-281.

Amatulli, G., Rodrigues, M.J., Trombetti, M., Lovreglio, R. 2006. Assessing long-term fire risk at local scale by means of decision tree technique. Journal of Geophysical Research 111, 1-15. G04S05, doi: 10.1029/2005JG000133.

Brillinger, D.R., Preisler, H.K., Benoit, J.W. 2003. Risk assessment: a forest fire example. In Goldstein, D.R. (Ed.), Science and Statistics. A Festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes 40, Beechwood, OH, pp. 177-196.

Brillinger, D.R., Preisler, H.K., Benoit, J.W. 2006. Probabilistic risk assessment for wildfires. Environmetrics 17, 622-633. doi: 10.1002/env.768.

Bruce, D. 1963. How many fires?. Fire Control Notes 24(2), 45-50.

Campbell, S., Liegel, L. 1996. Disturbance and forest health in Oregon and Washington. USDA Forest Service, General Technical Report PNW-381, 105 pp.

Cardille, J.A., Ventura, S.J., Turner, M.G. 2001. Environmental and Social Factors influencing Wildfires in the Upper Midwest, United States. Ecological Applications 11 (1), 111-127.

Chuvieco E., Salas, F.J, Carvacho, L., Rodríguez Silva, F., 1999. Integrated fire risk mapping. In Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin. Berlin: Springer-Verlag, pp. 61-84.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M.P., Vilar, L., Martínez, J., Martín, S., Ibarra, P., De la Riva, J., Baeza, J., Rodríguez, F., Molina, J.R., Herrera, M.A., Zamora, R. 2009. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling (In press).

Craven, P., Wahba, G. 1979. Smoothing noisy data with spline functions. Numerische Mathematik 31, 377-403.

Cunningham, A.A. and Martell, D.L. 1973. A stochastic model for the occurrence of man-caused forest fires. Canadian Journal of Forest Research 3(2), 282-287.

Fidalgo García P., Martín Espinosa, A. 2005. Atlas Estadístico de la Comunidad de Madrid 2005. Consejería de Economía e Innovación Tecnológica. Instituto de Estadística de la Comunidad de Madrid.

Golub, G.H., Heath, M., Wahba, G. 1979. Generalized cross validation as a method for choosing a good ridge parameter. Technometrics 21(2), 215-223.

Guisan, A., Edwards, T.C., Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89-100.

Izquierdo, J. 2007. Instrumentos económicos para la prevención y la lucha contra incendios. In Fundación Santander-Central Hispano (Ed.), Hacia la viabilidad económica del medio rural y de los bosques. Antes del Fuego. Soluciones a los incendios forestales en España. Madrid, Spain.

Kalabokidis, K.D., Koutsias, N., Konstantinidis, P., Vasilakos, C. 2007. Multivariate analysis of landscape wildfire dynamics in a Mediterranean ecosystem of Greece. Area 39(3), 392-402.

Leone, V., Koutsias, N., Martínez, J., Vega-García, C., Allgöwer, B., Lovreglio, R. 2003. The human factor in fire danger assessment. In Chuvieco, E. (Ed.), The role of remote sensing data. New Jersey, US: World Scientific Publishing, vol. 4, pp. 143-194.

Lin, C. 1999. Modelling probability of ignition in Taiwan Red Pine Forests. Taiwan Journal Forest Science 14 (3), 339-344.

Lorenzo, M.C., Pérez, M.C. 1995. Modelos de probabilidad para el estudio de la ocurrencia de incendios forestales. In 'IX reunión ASEPELT España', Santiago de Compostela, Spain.

Loftsgaarden, D. and Andrews, P.L. 1992. Constructing and testing logistic regression models for binary data: applications to the National Fire Danger Rating System. USDA Forest Service, General Technical Report INT-286 (Ogden, UT).

Maingi, J.K. and Henry, M.C. 2007. Factors influencing wildfire occurrence and distribution in eastern Kentucky, USA. International Journal of Wildland Fire 16, 23-33.

MARM. Ministry of the Environment, Rural and Marine Affairs. *Subsecretaría General de política forestal y desertificación. Área de defensa contra incendios forestales*. Los incendios forestales en España. Decenio 1996-2005. 2006. Available at http://www.mma.es/secciones/biodiversidad/defensa_incendios/estadisticas_incendios/pdf/estadisticas_decenio_1996-2005.pdf (Last accessed April 23, 2009).

Martell, D.L., Otukol, S., Stocks, B.J. 1987. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. Canadian Journal of Forest Research 17(5), 394-401.

Martell, D.L., Belivacqua, E. 1989. Modelling seasonal variation in daily people-caused forest fire occurrence. Canadian Journal of Forest Research 19(2), 1555-1563.

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

Martínez, J., Vega-García, C., Chuvieco, E. 2009. Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90, 1241-1252.

Ministry of Agriculture, Fisheries and Food (2004) Hechos y cifras de la Agricultura, la Pesca y la Alimentación en España. Available at http://www.mapa.es/es/ministerio/pags/hechoscifras/introhechos (Last accessed March 19, 2009).

Moyano, E. 2007. Incendios forestales en España. Diagnóstico de las causas. . In Fundación Santander-Central Hispano (Ed.), Hacia la viabilidad económica del medio rural y de los bosques. Antes del Fuego. Soluciones a los incendios forestales en España. Madrid, Spain.

Nicolás, J.M. and Caballero, D. 2001. Demanda territorial de defensa contra incendios forestales. Un caso de estudio: Comunidad de Madrid. Proceedings of the Spanish National Forest Congress. Granada, Spain, 25-28 September.

Pausas, J.L. and Vallejo, R. 1999. The role of fire in European Mediterranean ecosystems. In Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin. Berlin: Springer-Verlag, pp. 3-16.

Pew, K.L. and 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, 1-18.

Prasad, V.K., Badarinath, K.V.S., Eaturu, A. 2008. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. Journal of Environmental Management 86 (1), 1-13.

Preisler, H.K., Brillinger, D.R., Burgan, R.E., Benoit, J.W. 2004. Probability based models for estimation of wildfire risk. International Journal of Wildland Fire 13, 133-142.

Preisler, H.K., Westerling, A.L. 2007. Statistical model for forecasting monthly large wildfire events in western United States. Journal of Applied Meteorology and Climatology 46, 1020-1030.

Preisler, H.K., Chen, S., Fujioka, F., Benoit, J.W., Westerling, A.L. 2008. Wildland fire probabilities estimated from weather model-deduced monthly mean fire danger indices. International Journal of Wildland Fire 17, 305-316.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B. 2007. Human influence on California Fire Regimes. Ecological Applications 17(5), 1388-1402.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org (Last accessed April 23, 2009).

Vasconcelos, M.P.P., Silva, S., Tomé, M., Alvim, M., Pereira, J.M.C. 2001. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. Photogrammetric Engineering and Remote Sensing 67 (1), 73-81.

Vega García, C., Woodard, P.M., Titus, S.J., Adamowicz, W.L., Lee, B.S. 1995. A Logit Model for predicting the Daily Occurrence of Human Caused Forest Fires. International Journal Wildland Fire 5 (2), 101-111.

Vélez, R. 2005. Defensa contra incendios forestales: estrategias, recursos, organización. Available at http://forum.europa.eu.int (Last accessed March 19, 2009).

Vilar del Hoyo, L., Martín Isabel M.P., Martínez Vega, F.J. 2008. Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional. Boletín de la Asociación de Geógrafos Españoles 47, 5-29.

Word, S.N., Augustin, N.H. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling 157, 157-177.

Wood, S.N .2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC press: Boca Raton, FL, 392 pp.

# Capítulo IV

## Integration of lightning and human-caused wildfire occurrence models

**Publicación derivada:** *Vilar, L., Nieto, H., Martín, M.P. Integration of lightning and human-caused wildfire occurrence models. Human and Ecological Risk Assessment (en revision/under revision)*

# Integration of lightning and human-caused wildfire occurrence models

**Abstract**

The development of fire risk indices helps in the establishment of fire prevention actions. The integration of lightning and human fire probability models into those indices would improve fire prevention systems. Fire risk indices can be classified in short-term and long-term. The latter generally involves the integration of the most stable variables that affect fire ignition and/or propagation such as topography, fuel load, human activity, or climate patterns. The ignition of a fire can be the result of natural-lightning- or human causes. In European Mediterranean countries more than 90% of forest fires are due to human activities. We present two methods for the integration of lightning and human fire occurrence probability models at 1×1 km grid cell resolution in two regions of Spain: Madrid and Aragón, with very different fire cause typology. The two integration methods (probabilistic and weighted average considering historical fire causes) have been compared with a single probability average. Validation has been accomplished using an independent set of fire ignition points. For the validation we have used the Area Under de Curve (AUC) obtained by a Receiver Operating System (ROC) and the Mahalanobis distance. The use of a weighted average fits better in both tested regions. Validation results in Madrid are satisfactory (AUC~0.7) whereas in Aragón the fit are less suitable (AUC~0.6).

**Keywords**

AUC, Fire risk, Mahalanobis distance, Probabilistic, ROC

## 1. Introduction

Wildland fires can be considered as a significant disturbance factor in many ecosystems (Mooney, 1981; Pyne, 1982). In European Mediterranean countries wildfires play an important role in the landscape featuring. Even thought vegetation communities in these ecosystems are adapted to fire (Moreno, 1989), in the last decades socioeconomic, cultural and political changes have brought important transformations in these areas, where more than 90% of wildfires are due to human activity (FAO, 2007), that have caused significant changes in wildfire activity.

Wildfires have important consequences at environmental, ecological, economic and sociological levels. Because of that, the development of tools to predict wildfires such as fire risk indices is essential to the establishment of fire prevention actions in order to reduce fire impacts.

According to FAO (Food and Agriculture Organization for the United Nations) terminology, fire risk is defined as the probability of fire starting determined by the presence and activities of causative agencies, while fire danger is defined as considering both fixed and variable factors of the fire environment that determine the ease of ignition, rate of spread, difficulty of control, and fire impact, often expressed as an index (http://www.fao.org/forestry/site/firemanagement/en/, April 2009). In opposition with other natural phenomena, fires can theoretically start in any point of space (in the zones covered with vegetation). The probability of ignition depends primarily on the fuel conditions (flammability, moisture content) and the causal agents which can be human or natural (lightning). Fire Danger Rating indices can be classified in short-term and long-term indices (Chuvieco et al., 2003). The former are primarily concerned with the most dynamic variables involved in fire danger (weather parameters and fuel moisture content) while the latter refer to the integration of the most stable variables that affect fire ignition and/or fire propagation, such as topography, fuel load and structure, human activities and climate patters. Chuvieco et al. (2003, 2009 in press) proposed that an integrated assessment of fire risk should consider both fire ignition probability as well as the assessment of potential damages (vulnerability of the affected areas). Within the framework of the *Firemap* research project ('Integrated Analysis of Wildland Fire with Remote Sensing and GIS') a wildland fire risk index was developed which integrates the human and environmental factors related to fire ignition (Chuvieco et al., 2009 in press).

Assuming that the variables to be included in a synthetic fire risk index can be generated at the required temporal and spatial resolution, the most critical problem is to establish a coherent criterion to properly combine those variables. Since the goal is to obtain a single fire risk index, the component variables (vegetation, topography, climate, socioeconomic factors etc.) should first be classified in a numerical scale of risk and then combined into a single index. In some cases, the proposal of risk levels implies changing the nominal-categorical scale of the original variables to an ordinal scale. On the other hand, the integration of the different components in a single risk index requires that a weight be applied to each variable according to its relative importance on the fire occurrence in an specific area (i.e. how much importance has the human than the natural factors in the fire ignition probability?).

There are several methods for the integration of fire danger variables. Chuvieco et al. (2003) distinguish five groups of techniques for the integration of danger variables in a single fire danger index: (*i*) qualitative models, where arbitrary weights are based on the judgement of an expert; (*ii*) quantitative indices, based on multicriteria evaluation; (*iii*) regression techniques, where statistical estimation methods are applied to explain fire occurrence; (*iv*) neural networks, similar conceptually to the regression model and (*v*) physical models, based on meteorological conditions or on fire propagation models. In addition, probabilistic models can be used. These models show the advantage that (*i*) they deal with probability values, which are intrinsically normalised (0-1), and (*ii*) they are based in algebraic relationships such as Kolomogorov axioms and Bayes theorem. Probabilistic models have been widely used in ecological applications e.g. habitat modelling studies (Skidmore, 1989; Fischer, 1990; Aspinall, 1992; Brzeziecki et al., 1993; Hutley et al., 1995) or population dynamic modelling (Soudant et al., 1997; Wilson et al., 2008; Renken et al., 2009 in press).

Logistic regression analysis has been extensively used both to predict and also to explain human and/or lightning-caused fires by integrating geophysical, environmental or socioeconomic variables (e.g. related to topography, vegetation, land uses, climate and meteorological conditions, environmental parameters, fire danger indices, human factors) with observed fire occurrence (Martell et al., 1987; Vega-García et al., 1995; Lin, 1999; Pew and Larsen, 2001; Vasconcelos et al., 2001; Martínez et al., 2004; Wotton and Martell, 2005; Kalabokidis et al., 2007; Prasad et al., 2008; Martínez et al., 2009; Modugno et al., 2008; Vilar et al., 2008; Nieto et al., in revision). Other statistical methods such as linear regression, classification regression trees, neural networks, generalized additive models or Bayesian

probability have also been used in fire risk mapping to generate risk models (Chuvieco et al., 1999; McKenzie et al., 2000; Sebastián et al., 2001; Chao-Chin, 2002; Koutsias et al., 2004; Preisler et al., 2004; Robin et al., 2006; Amatulli et al., 2006, 2007; Amatulli and Camia, 2007; Syphard et al., 2007; Vega-García, 2007; Yang et al., 2007; Romero-Calcerrada et al., 2008).

Following Guisan and Zimmermann (2000) model development includes (*i*) the underlying of the conceptual model; (*ii*) the statistical formulation and the sampling design, where a calibration and evaluation (= validation) datasets are also defined; (*iii*) the calibration of the model, where fitted values are obtained and tested for the quality of the fit; (*iv*) the obtaining of predicted values using the model formulation and the validation dataset; (*v*) and the validation of the model using validation tables. Models are a representation of reality and, to ensure its accuracy, models should be tested and improved. Validation compares simulated system outputs with real system observations using data not used in model development (Mazzotti and Vinci, 2007). Regarding Rykiel (1996), validation is a demonstration that a model, within its domain of applicability, possesses a satisfactory range of accuracy consistent with the intended application of the model. Validation therefore refers to model performance. It describes a testing process on which to base on an opinion of how well a model performs so that a user can decide whether is acceptable for its intended purpose. To validate the predictive power of a model two different approaches can be followed: (*i*) to use a single data set to calibrate the model and then validate it by cross validation, leave-one-out jack-knife or bootstrap methods and (*ii*) to use two independent datasets, one to calibrate the model and another one to validate it (Guisan and Zimmermann, 2000). When using two independent datasets, any discrete measure of association between predicted and observed values can be used (e.g. Fielding and Bell, 1997; Stehman, 1999; Guisan and Harrel, 2000). If the predictions of a statistical model are probabilistic, they need to be transformed back to the scale of the real observations. For binary data, it can be done by truncating probabilities at a given threshold (Guisan and Zimmermann, 2000) and computing the omission and commission errors. One of the drawbacks of this approach is that one must select a threshold value, sometimes in an arbitrary way. Receiver Operating Characteristic (ROC) is a threshold independent measure that computes the model performance over all possible thresholds (Fawcett, 2006). ROC is frequently used for the evaluation of species distribution models (Fielding and Bell, 1997). It compares a rank map (e.g. predicted probability) against a Boolean map (e.g. fire presence/absence). ROC indicates how well the events of the Boolean map falls within the high suitability values in the rank map (Pontius and Scheneider, 2001).

Human, lightning and integrated wildfire risk models have been validated using different approaches. Authors usually have randomly partitioned the sample and used one subset to fit the model and the other one to validate it (Sebastián et al., 2001; Vasconcelos et al., 2001; Vega-García, 2007). Some authors used the ROC area under the curve (Modugno et al., 2008; Nieto et al., in revision) for validating. Other authors, however, employed residual analysis and Akaike Information Criterion (AIC), or the Hosmer and Lemeshow test to select the best model to fit the data (Yang et al., 2007). McKenzie et al., (2000) validated the model with standard diagnostics such as plot of residuals against fitted values and quantile-quantile plots. Also, they calculated a bootstrap estimate of prediction error for the regression models and compared it to the model's error sum of squares. Annual summaries of the modeled number of fires were compared to observations to examine the fit of the model in Wotton and Martell (2005). Amatulli et al. (2006) carried out a spatial validation in their fire risk resulting map from a CART model. They analyzed the correlation between the predicted fire risk value and the observed fire occurrence values. Also, they performed a second correlation analysis looking at the predicted and observed density values. The ignition points for validation falling in each fire management unit, expressed as observed density, were plotted against the predicted density values of the fire risk map. Dong et al. (2006) used linear regression to analyze the relationships between the value of area weighted of forest fire risk and the frequency of historical forest fires.

Short-term indices (fire danger rating systems) have been also validated, e.g. Andrews et al. (2003) evaluated fire danger rating systems (US NFDRS) using logistic regressions and percentile comparisons, to examine the relationships between fire danger indices and fire activity. As independent variable in the logistic regression they used the index, whereas the dependent variables were fire-day, large fire-day and multiple fire-day. Chuvieco et al. (2009 in press) based the assessment of daily integrated indices on fire statistics (ignition points collected within 4 months of daily data), evaluating the existence of significant differences between the risk values of fire and no fire cells. They computed (*i*) the Mahalanobis distance; (*ii*) the Mann-Whitney U-test; (*iii*) and the Nagelkerke $R^2$ coefficient from logistic regression fittings for each integrated index.

In the present study, we propose three methods to integrate human and lightning-caused wildfire probability models in the regions of Madrid and Aragón (Spain) at 1$km^2$ grid cell resolution. Then, we validate the human and lightning-caused models as well as the integrated ones using wildfire occurrence data (x, y ignition points) as reported by official statistics and spanning the years 2005-2007.

## 2. Methods

### 2.1 Study areas

We have selected two Mediterranean regions with different fire causality conditions in Spain: Aragón and Madrid. More than 90% of fires in the region of Madrid have human cause whereas in the region of Aragón ~30% are due to lightning. The region of Madrid is located in central Spain and the region of Aragón in the North-East as shown in Figure 1.

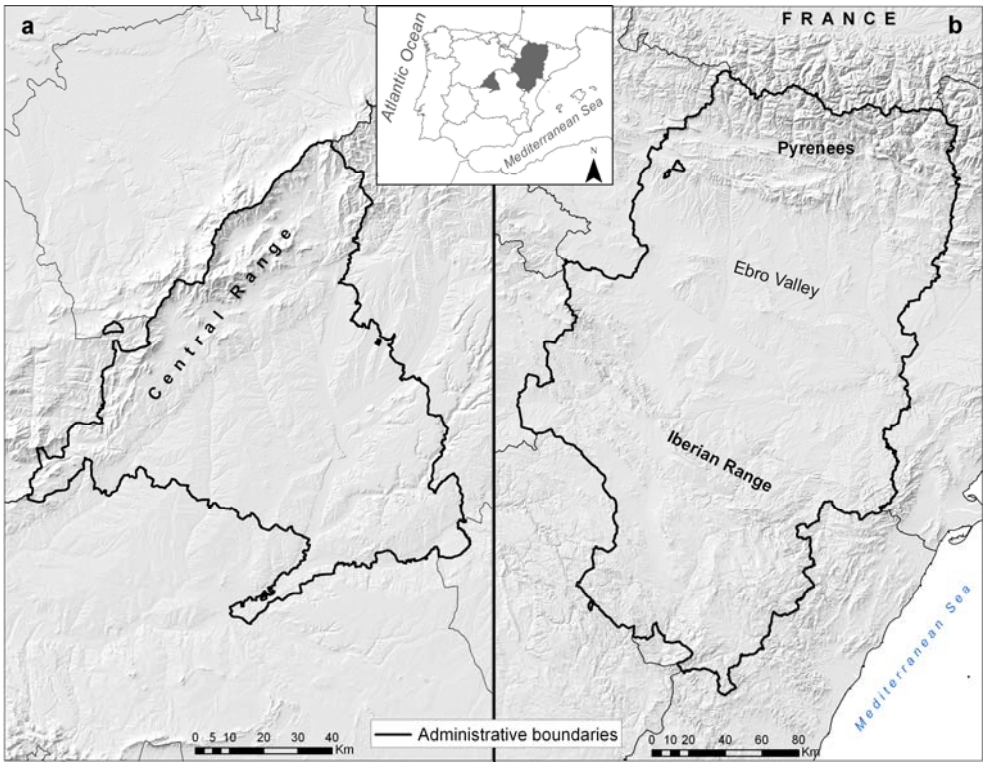



**Fig. 1.** Study areas. The regions of Madrid (a) and Aragón (b) (Spain).

The region of Aragón has 47 500 km² distinguishing three geomorphologic units, the Pyrenees Range in the North, the Iberian Ranges extended at the South, and between them the Ebro Valley, a topographic depression crossed by the Ebro River. Due to its topographic characteristics it has a wide range of climatic conditions, from arid Mediterranean conditions in the valley to permanent snow in the highest peaks of the Pyrenees. In the central area (200 m above mean sea level) the annual precipitations are below 400 mm with an annual average temperature around 14-15ºC. In higher altitudes, in the mountains (600-1000 m above mean sea level) the average temperature is below 10ºC. In the Pyrenees Range the annual precipitations are around 1000-2000 mm while in the Iberian Range 1000 mm

(Cuadrat, 2004). Highest altitudes are mainly composed by natural vegetation (forest, shrublands and pastures) whereas the central depression is mainly composed by crop lands. Dry mountainous areas are occupied primarily by conifer woodlands. In more humid areas, on the other hand, broadleaved species become dominant. In the Iberian Range (dryer conditions) conifer species are the most abundant (Gobierno de Aragón, 1997). The human pressure has given to the territory a high level of landscape fragmentation leaving few wildland areas especially in the poor and non-productive agricultural lands (Lasanta et al., 2006). The area of the Iberian Range is one of the lowest populated in Spain. In overall, the region of Aragón has one the lowest value of population density in Spain, 25 people by $km^2$. Regarding the wildfire cause typology in the last decades, without taking into account the fires without a known reported cause, ~30% of the fires were due to lightning. This region has the highest percentage of fires caused by lightning in Spain, due to its orographic and climatic conditions.

The region of Madrid has 8028 $km^2$. It is composed by a mountain range that follows a NE-SW direction (with its highest peak having an elevation of 2428 m) and the Tagus river basin, which runs from the fault of the mountain range (800 m above mean sea level) to the Tagus' river bed (Fidalgo and Martín, 2005). Madrid has Mediterranean climatic conditions but due to its altitude, the distance to the sea, and the barrier effect of the surrounding mountain ranges, it has a lower annual rainfalls (300 to 1000 mm-in the mountain range) and higher thermal amplitude compared to other Mediterranean regions. Pastures and shrublands occupy vast areas in the region, whereas forest areas are mainly located in the mountain range, two thirds of which are broadleaved woodlands. Holm oak (*Quercus ilex*), which is the predominant specie, can be commonly found in the typical Mediterranean *dehesa* formation, a sparsely meadow highly regarded from the point of view of its landscape as well as their recreational, environmental and livestock value. This region is the most populated in Spain, where urban areas have been increasing in size and population since the 1960s and spreading into agricultural and forest areas. Forest fires in Madrid are spatially associated to roads, railways, dump sites and urban areas (Nicolás and Caballero, 2001; Vilar et al., 2008). The contact boundary between urban and forest areas, which is commonly referred to as the Wildland-Urban Interface (*WUI*), is an area of major concern for regional fire managers. Regarding the known cause typology fires in the last decades, more than 90% of fires were due to human activities and therefore, only 2.36% of fires were due to lightning (MARM, Spanish Ministry of Environment, Rural and Marine Affairs, 2006). Figure 2 shows the main land uses in both study regions. As we can see, forest areas in the study regions are

mainly located in the mountain ranges. In the region of Madrid the urban areas are occupying big surfaces mainly located in the centre of the region. Croplands appears in the east and south areas, whereas in Aragón are occupying a large extension in the Ebro valley.

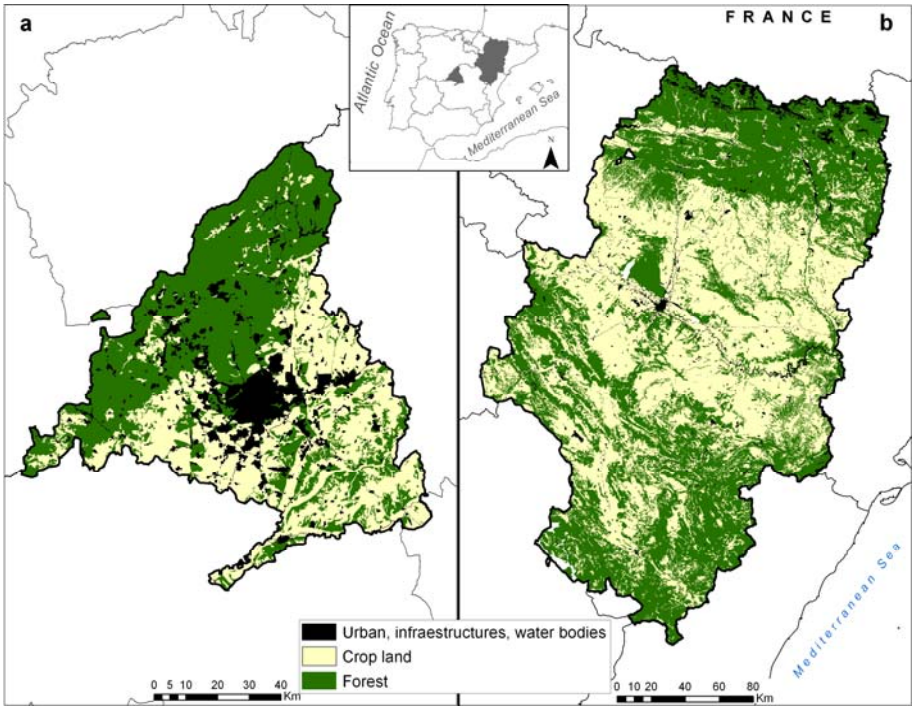



**Fig.2.** Main Land uses categories in Madrid (a) and Aragón (b) (*source*: Corine Land Cover 2000)

## 2.2 Data

### 2.2.1 Lightning and human-caused wildfire probability of occurrence models

The probability of occurrence models for lightning and human-caused wildfires have been obtained in the two study regions within the framework of the *Firemap* project. More details on the fire risk scheme proposed in this project as well as on the methodologies applied to model the ignition potential from human and natural factors can be found in Chuvieco et al. (2009 in press).

The lightning-caused models in the study regions were obtained with three years interval from 2002-2004 in a 3×3 km UTM grid. The independent variables used were those related to topography, vegetation and weather patterns. The logistic regression results showed that the *number of thunderstorms* was the most significant variable in terms of

probability of occurrence in both regions, achieving best results in Aragón by taking into account only the dry storms. Lightning-caused models for the two study areas are shown in Figure 3. Highest probability values are located over mountainous areas. The probability values range from 0.0046 to 0.2157 in Madrid and 0.0001 to 0.3883 in Aragón. More information about this model can be found in Nieto et al. (in revision).

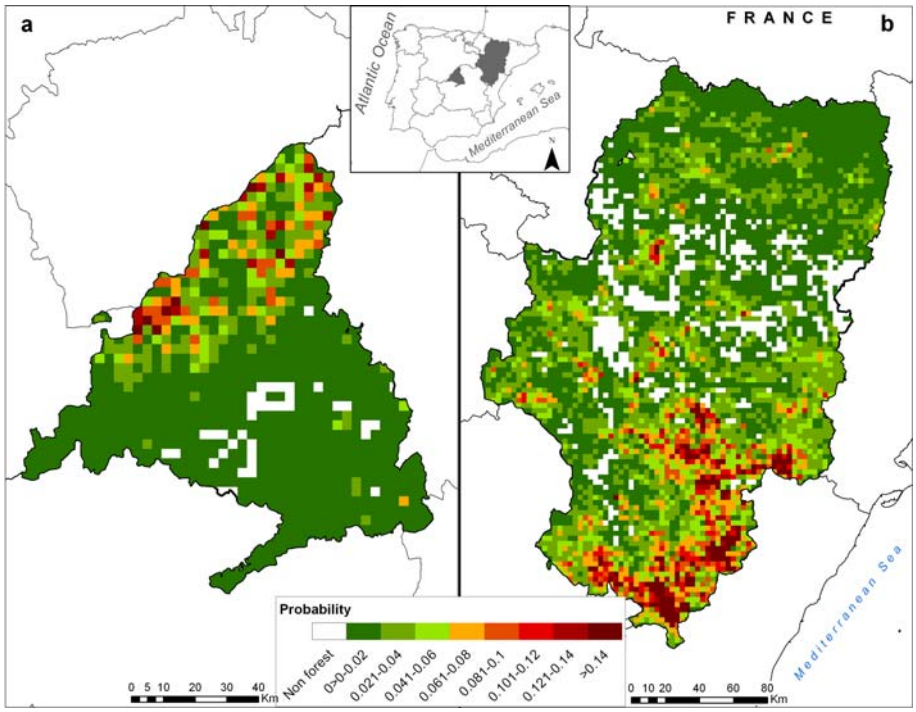



**Fig.3.** Lightning-caused wildfire probability of occurrence maps. Region of Madrid (a) and region of Aragón (b)

Logistic regression was used to produce human-caused ignition probability maps at 1km$^2$ grid cell resolution, using 60% of the sample to fit the model and the remaining 40% to calibrate it. As independent variables were used those representing socioeconomic issues related to the ignition of a fire which were obtained from diverse cartographic and statistical sources. This set of variables represents social factors related to human fire risk in Spain: socioeconomic changes, traditional activities in rural areas, accidents or negligence, fire prevention activities, and factors which can lead to social unrest (e.g. land use disputes or high unemployment rates) (Martínez *et al.* 2009). Those factors have been reported by the literature and historical fire databases to have a direct or indirect influence in fire occurrence in Spain and could be considered representative of Euro-Mediterranean countries. To homogenise lightning and human-caused models prior to the integration, the latter was

adapted from Vilar *et al.* (2008) using as response variable fires due to human causes from 2002-2004 (x, y coordinates of the fire ignition points, from the Fire Department in Madrid and from The National Fire records in Aragón). In order to account for potential inaccuracies in the location of the fire ignition points, their position was compared with ancillary data contained in the National Fire records. In this database the fire locations are referred to both a 10×10 km grid and the municipality. Those ignition points whose coordinates were not included in the corresponding 10×10 km grid and municipality of the National Fire records have been filtered and, therefore, omitted from the analysis. The logistic regression results showed that in Madrid the *Wildland Urban Interface* (*WUI*) was the most significative variable, along with the *roads in forested areas*. In Aragón, *WUI* was followed by *electric lines* variable. Human-caused models are shown in Figure 4. The probability values range from 0.0385 to 0.7351 in Madrid and 0.00002 to 0.9527 in Aragón. From both lightning and human-caused models non forest areas were masked because we are considering only the probability of ignition of wildland fires, that is, those that occurs in forested areas.

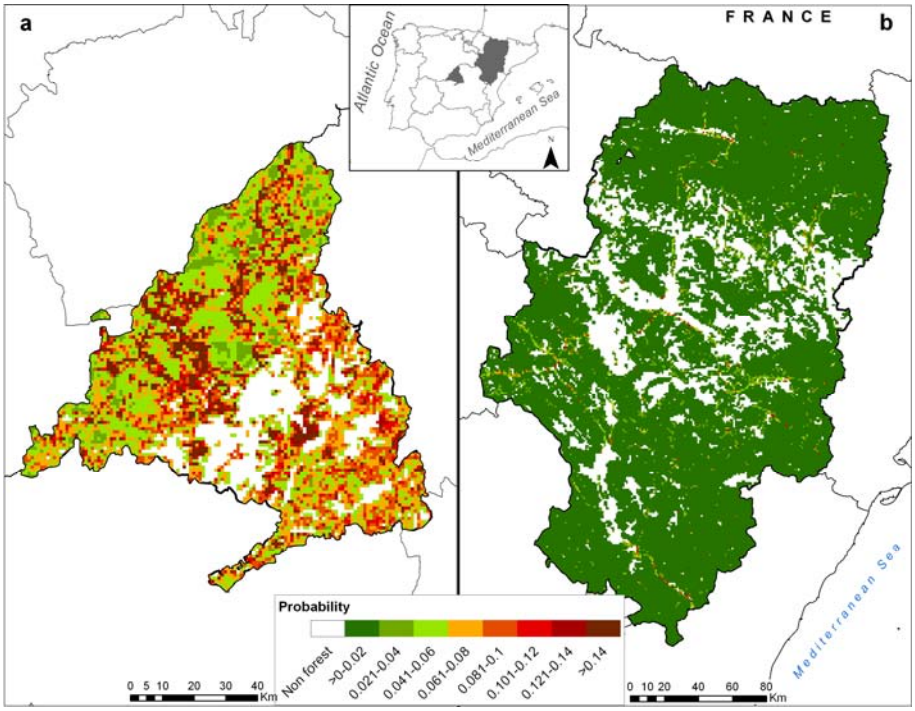



**Fig.4.** Human-caused wildfire probability of occurrence maps. Region of Madrid (a) and region of Aragón (b)

### 2.2.2 Validation data: Fire ignition points

Fire ignition points from 2005-2007 were used to validate the results of the methods to integrate lightning and human-caused wildfire occurrence models in the two study areas. As previously mentioned, in the National Fire records database fire locations are referred to a 10×10 km grid and also at a municipality level. In recent years and depending on the region, the fire database has included the exact location (x, y coordinates) of the fire ignition points. In Madrid, the database including information on fire ignition coordinates provided by the Fire Department of the Madrid region contains a total of 728 fire ignition points during the chosen validation period (2005-2007). In Aragón, this database is provided by the National Fire records, and contains 1481 fire ignition points in 2005-2007. Figure 5 shows the spatial location of these points in the study areas.

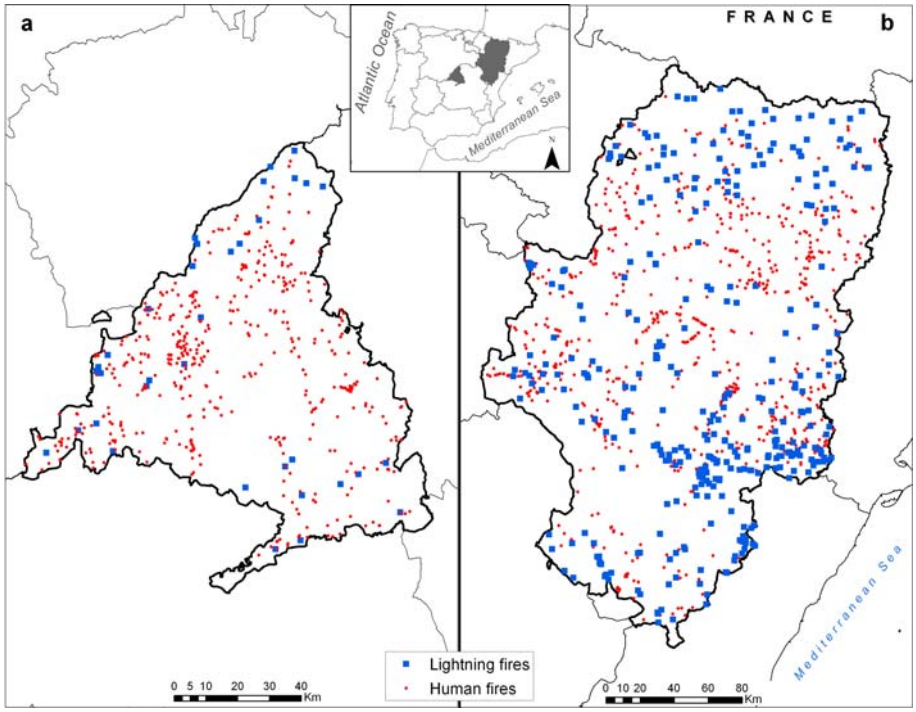



**Fig.5.** Fire ignition points for validation. Years 2005-2007. Region of Madrid (a) and region of Aragón (b)

## 2.3 Integration methods

Since both models (lightning and human) have the same probabilistic scale (0-1) no transformation of the original values was needed prior the integration. This integration has

been performed following three different methodologies. 1) Probabilistic integration, based on Kolmogorov axioms (Tarantola, 2005); 2) average weighted by the historic occurrence of wildfires; 3) and a simple average as reference method. Our hypothesis is that the two first methods should perform better than a single average.

In the probabilistic approach, we assume that the integrated probability of the occurrence model should explain the probability of occurrence of either of a human-caused or a lightning-caused fire. Based on the Kolmogorov axioms, the integrated probability of occurrence can be expressed following the addition law of probability:

$$P\,(A \cup B) = P\,(A) + P\,(B) - P\,(A \cap B) \qquad (1)$$

where $P\,(A \cup B)$ is the integrated probability, $P\,(A)$ and $P\,(B)$ are the probability of both events (human and lightning-caused fires) and $P\,(A \cap B)$ is the probability that both lightning and human-fires will happen. Equation (1) can be simplified in (2) assuming that lightning-caused and human-caused fires are independent events at the grid resolution and in long-term

$$P\,(A \cup B) = P\,(A) + P\,(B) - P\,(A)\,P\,(B) \qquad (2)$$

The second integration approach is based in a weighted average of the probability of occurrence of lightning and human caused fires. The weights are based in the historic fire causes from 1990-2004 (National Fire records source) and is expressed as:

[$P\,(A)$ × historic human-caused fire occurrence] + [$P\,(B)$ × historic lightning-caused fire occurrence] (3)

Where $P\,(A)$ is the probability of human-caused fires and $P\,(B)$ is the probability of human-caused fires. The historic fire occurrence of human or lightning fire causes is expressed in terms of proportion of fires due to human or lightning cause by grid cell in the period 1990-2004.

Finally, it has been carried out the average between lightning and human-caused probability values by cell in the study areas

$$[P\,(A) + P\,(B)]/2 \qquad (4)$$

2.4 Validation

To carry out the validation of the lightning and human-caused models as well as the integrated ones we count on the fire ignition points (x, y coordinates) from 2005-2007 in each region as mentioned before. First of all, in order to validate the lightning and human-caused models separately, the validation data (fire ignition points) were classified by human or lightning cause. In Madrid there were 557 fires due to human causes (94%) and 35 due to lightning (6%) in the validation period, while in Aragón the figures were 690 (68%) and 321 (31%) respectively. To validate the resulting integrated models we used the total set of fire ignition points (excluding unknown causes). The total ignitions in Madrid were 592 and 1011 in Aragón. We evaluated the existence of significant differences between the predicted fire occurrence probability values on fire and no fire cells (Chuvieco et al., 2009 in press). To evaluate these differences we used Receiver Operating Characteristic (ROC) analysis (Fawcett, 2006) and Mahalanobis Distance (Viegas et al., 1999). In ROC analysis the Area Under the Curve (AUC) represents the probability that the assay result for a randomly chosen positive case (fire) will exceed the result for a randomly chosen negative case (no fire). The asymptotic significance was set to less than 0.05, which means that using the assay is better than guessing. AUC with a 0.5 value mean non discrimination; between 0.7-0.79 mean a reasonable discrimination; 0.8-0.89 excellent; 0.9 or higher exceptional.

From Viegas et al. (1999) and Andrews et al. (2003), the Mahalanobis Distance is expressed as

$$M_d = [(X_1 - X_2)/\sigma]^2 \qquad (5)$$

where $X$ is the probability value from the integrated results, $X_1$ is the average of the probability values on no-fire cells, $X_2$ is the average of the probability values on fire cells, and σ is the standard deviation of the probability values on all cells.

**3. Results**

3.1 Integration of lightning and human-caused models at 1×1 km grid cell resolution

Figure 6 shows the results from the probabilistic integration, and the weighted and single average integration in Aragón. The spatial distribution of the probability values

shows similar trends using the three integration methods. The highest values are located in the southeast (Iberian range). This probability distribution is in accordance to the lightning-caused model (Figure 3b), where the highest probability values were located in the Iberian range. However, high probability values from the human-caused model (e.g. cells following *WUI* or *electric lines* located in north, centre and west areas of the region) are also included in the integrated result (Figure 6). The probability values of the integrated maps using the probabilistic method and the weighted average are higher than the ones using the single average. The probabilistic method shows the highest values, located in cells at the southeast of the region.

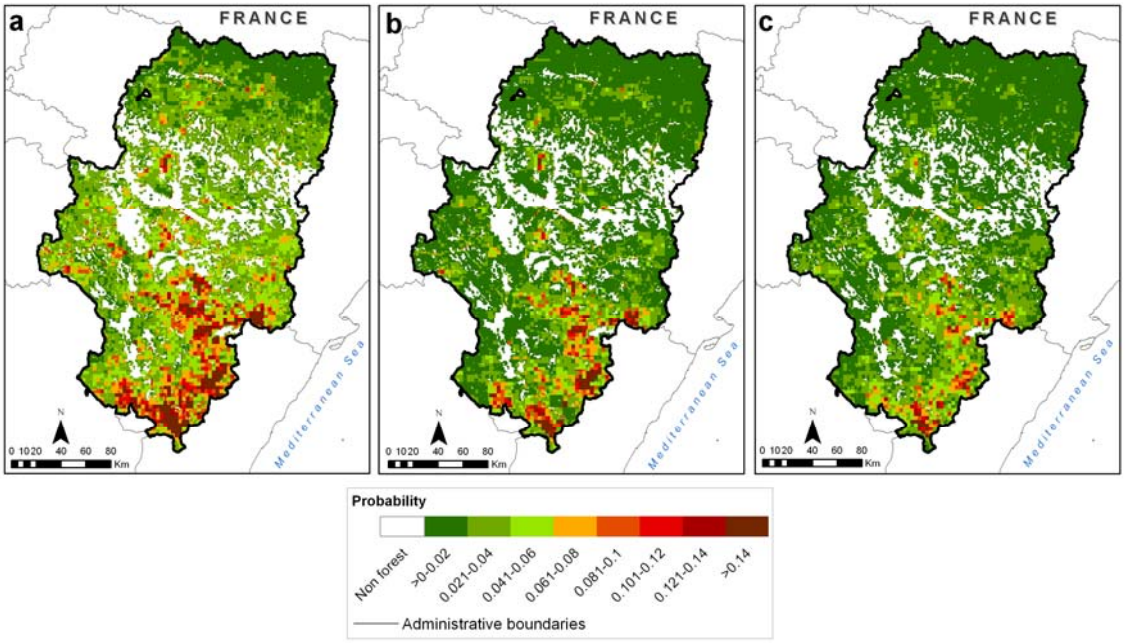



**Fig.6.** Integrated probability models in the region of Aragón: Kolmogorov probabilistic integration (a), weighted by the historic fire occurrence average of human and lightning-caused integration (b), single average probability values (c)

Figure 7 shows the integration results in Madrid. The three integration methods show the same pattern in the spatial distribution of probability values. The highest values are located to the west, southeast and following the northern direction in the central range. Cells with higher probability values follow the roads and areas with high density of *WUI*. The human component dominates the integrated results in this region which is in accordance to historical fire causes (Figure 4a). In north areas that belong to the central range, probability values are related to the lightning-caused model which enhances the

integrated result. In areas where the lightning-caused probability values are low is the human model which is influencing the final result (east and south-eastern areas). Regarding the differences between the three maps, the probability values obtained by probabilistic methodology and the weighted average probability are higher than using the single average approach. The probabilistic method presents the highest values.

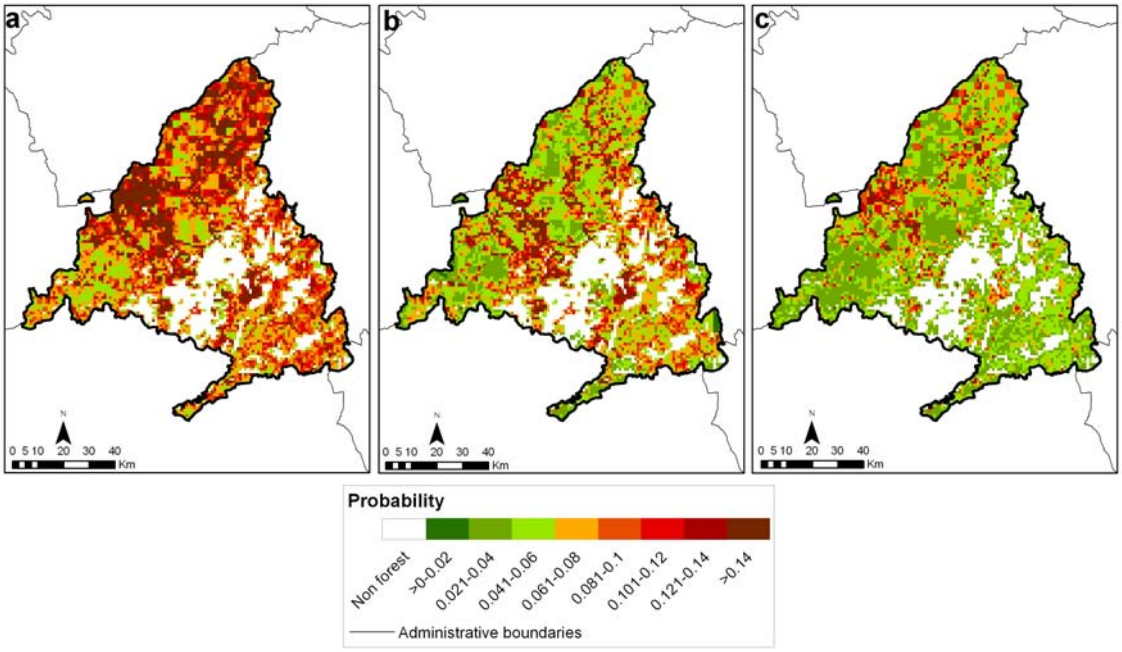



**Fig.7.** Integrated probability models in the region of Madrid: Kolmogorov probabilistic integration (a), weighted by the historic fire occurrence average of human and lightning-caused integration (b), single average probability values (c)

Table 1 shows the descriptive statistics for the three methods used in the integration for all the cells in the study region as well as in cells with and without fire.

In the region of Aragón, the probabilistic method and the weighted average show higher maximum values (0.952 and 0.886 respectively) than the single average (0.476). Lower mean belongs to both single and weighted average methods (0.022 and 0.024 respectively) than the probabilistic integration method (0.044). In the no-fire set, the values are quite similar to the ones in the whole dataset. In those cells where a fire has been observed, the mean values of probability obtained in the three methods are higher than in cells without a fire. In the region of Madrid, the probabilistic method and weighted average show higher maximum probability values (0.737 and 0.735 respectively) than the single average approach (0.372). The lowest mean value belongs to the average probability (0.052) whereas the

probabilistic integration method has the highest (0.104). In the cells with no fire the values are mostly the same as in the all sample. In those cells with an observed fire the mean values are higher for the probabilistic and weighted probabilities (0.141 and 0.111 respectively). In both regions, the highest variability in the integrated probability is obtained using the weighted average (variation coefficients of 125 % and 73% for Aragón and Madrid respectively).

**Table 1.** Descriptive statistics the integrated models: probabilistic, weighted average and single average in the region of Aragón (a); and in the region of Madrid (b) at 1×1 km grid cell resolution

| **a. Aragón** | | | | | | |
|---|---|---|---|---|---|---|
| **Total** | **N** | **Minimum** | **Maximum** | **Mean** | **Std. Deviation** | **Coefficient of Variation** |
| Probabilistic | 39842 | 0.000 | 0.952 | 0.044 | 0.041 | 93% |
| Weighted average | 39842 | 0.000 | 0.886 | 0.024 | 0.030 | 125% |
| Single average | 39842 | 0.000 | 0.476 | 0.022 | 0.020 | 91% |
| **Cells without fire** | | | | | | |
| Probabilistic | 38797 | 0.000 | 0.952 | 0.043 | 0.041 | 95% |
| Weighted average | 38797 | 0.000 | 0.886 | 0.024 | 0.030 | 125% |
| Single average | 38797 | 0.000 | 0.476 | 0.022 | 0.020 | 91% |
| **Cells with fire** | | | | | | |
| Probabilistic | 1045 | 0.002 | 0.602 | 0.054 | 0.051 | 94% |
| Weighted average | 1045 | 0.001 | 0.333 | 0.031 | 0.036 | 116% |
| Single average | 1045 | 0.000 | 0.311 | 0.027 | 0.026 | 96% |
| **b. Madrid** | | | | | | |
| **Total** | **N** | **Minimum** | **Maximum** | **Mean** | **Std. Deviation** | **Coefficient of Variation** |
| Probabilistic | 7052 | 0.043 | 0.737 | 0.104 | 0.062 | 60% |
| Weighted average | 7052 | 0.004 | 0.735 | 0.075 | 0.055 | 73% |
| Single average | 7052 | 0.021 | 0.372 | 0.052 | 0.032 | 62% |
| **Cells without fire** | | | | | | |
| Probabilistic | 6467 | 0.043 | 0.737 | 0.100 | 0.058 | 58% |
| Weighted average | 6467 | 0.004 | 0.735 | 0.072 | 0.051 | 71% |
| Single average | 6467 | 0.021 | 0.372 | 0.051 | 0.030 | 59% |
| **Cells with fire** | | | | | | |
| Probabilistic | 585 | 0.043 | 0.686 | 0.141 | 0.092 | 65% |
| Weighted average | 585 | 0.022 | 0.627 | 0.111 | 0.082 | 74% |
| Single average | 585 | 0.022 | 0.349 | 0.072 | 0.048 | 67% |

## 3.2 Validation

We carried out the validation of the lightning and human-caused models as well as the integrated results using the three mentioned techniques at 1×1 km grid cell resolution. Figure 8 shows the ROC for the lightning and human-caused probability models in the study areas. In Aragón, the human-caused model presents a regular ROC curve similar to

the one from the lightning-caused model but with lower values. In Madrid, lightning ROC curve presents a segmented shape.

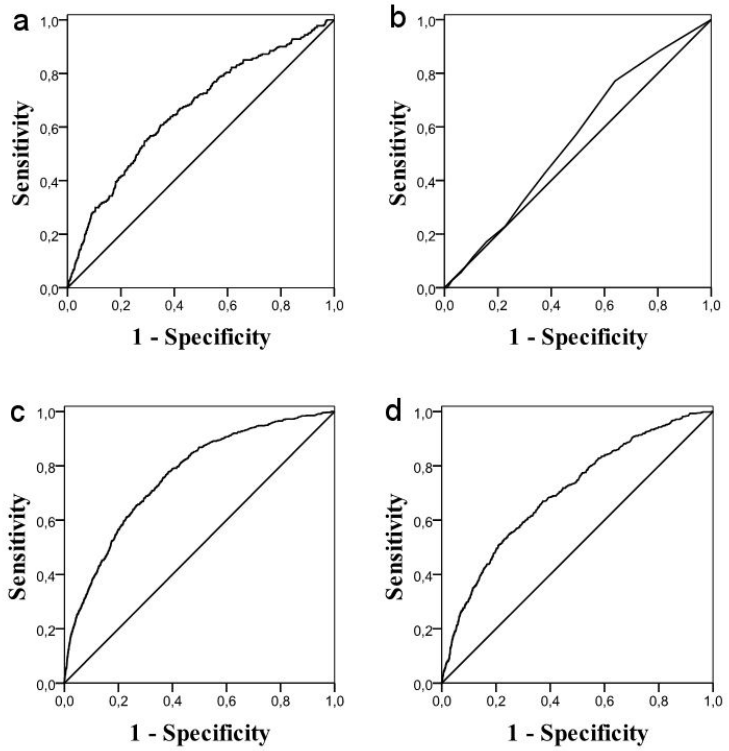



**Fig.8.** ROC for the lightning-caused probability of occurrence fire models in Aragón (a) and Madrid (b); ROC for the human-caused probability of occurrence fire models in Aragón (c) and Madrid (d) at 1×1 km grid cell resolution

Figure 9 shows the ROC for the integrated-caused probability models.

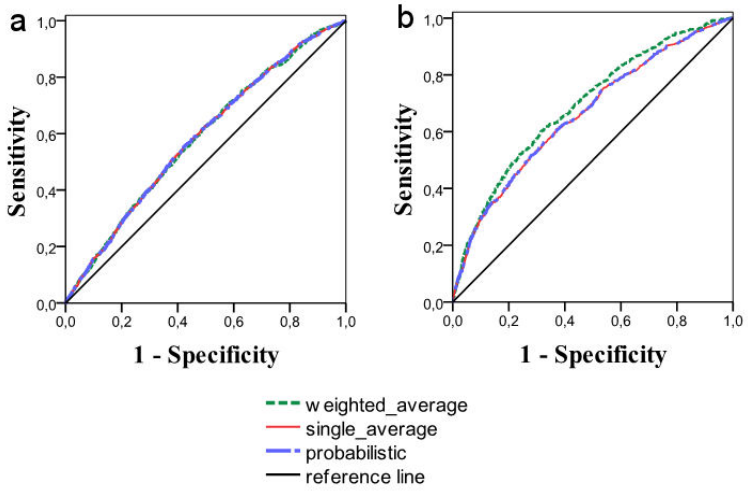



**Fig.9.** ROC for the integrated probability methods in the study areas: Aragón (a), Madrid (b)

In Aragón, the curves from the three methods have similar values whereas in Madrid the curves from the probabilistic method and the single average have similar values and the resulting curve for the weighted average probability values presents higher values. Table 2 shows the values of the AUC for the human and lightning-caused models as well as the integrated results at 1×1km grid cell resolution in the study areas.

**Table 2.** Area Under the Curve, standard error, p-value and 95% confidence interval for the lightning and human-caused probability of occurrence models and integrated models at 1×1km grid cell resolution in Aragón (a) and Madrid (b)

| **a.Aragón** | | | | | |
|---|---|---|---|---|---|
| **Model** | **Area** | **Std. Error (a)** | **Asymptotic Sig.(b)** | **Asymptotic 95% Confidence Interval** | |
| | | | | **Upper Bound** | **Lower Bound** |
| Lightning-caused | 0.663 | 0.016 | 0.000 | 0.632 | 0.694 |
| Human-caused | 0.762 | 0.009 | 0.000 | 0.745 | 0.780 |
| Probabilistic | 0.585 | 0.009 | 0.000 | 0.568 | 0.602 |
| Weighted average | 0.584 | 0.009 | 0.000 | 0.567 | 0.601 |
| Single average | 0.585 | 0.009 | 0.000 | 0.568 | 0.602 |
| **a.Madrid** | | | | | |
| Lightning-caused | 0.553 | 0.044 | 0.278 | 0.466 | 0.640 |
| Human-caused | 0.702 | 0.012 | 0.000 | 0.680 | 0.725 |
| Probabilistic | 0.662 | 0.012 | 0.000 | 0.639 | 0.686 |
| Weighted average | 0.693 | 0.011 | 0.000 | 0.671 | 0.715 |
| Single average | 0.662 | 0.012 | 0.000 | 0.638 | 0.685 |

a Under the nonparametric assumption , b Null hypothesis: true area = 0.5

In Aragón the AUC for the lightning and human-caused model is 0.663 and 0.762 respectively, which means that for both models the discrimination is reasonable. This value decreases to 0.584 for probabilistic and single average integration and 0.585 for the weighted average, meaning that the integration is not very satisfactory. In Madrid, the AUC for the lightning and human-caused model is 0.553 and 0.702 respectively, which means that for the human model the discrimination is reasonable whereas is not completely satisfactory for the lightning model. The AUC for both the probabilistic and the single average is 0.662, whereas AUC for the weighted average is slightly higher (0.693), showing that the integration performs reasonably better than the lightning model in all three methods.

Regarding Mahalanobis distance results in Aragón, the probabilistic method presents the highest value (0.067) followed by the single average (0.062) and the weighted average (0.052). It means that, in the probabilistic integration method, the probability of cells with and without fire differs more than when using the other methods. In Madrid, the weighted average shows the highest value (0.490) followed by the probabilistic method (0.435) and the

single average method (0.422), which is in accordance with the obtained AUCs. In this region, the weighted average presents the highest difference between the probability in the fire and no-fire cells.

**4. Discussion**

The spatial distribution of the probabilities is almost the same for the three methods, varying only their values ranges. Comparing the integrated results with the original maps (lightning and human-caused maps) we find that in Aragón the integrated results are dominated by the lightning-caused wildfire model, while in Madrid by the human-caused model. However, in Aragón, cells where the human-caused model presents high values enrich the integrated probability results. In the region of Madrid, we find that the lightning-caused model is influencing the final result in north areas, where there have been high probability values coming from this model. This behaviour is in accordance to the historical causality of fires in both regions: the majority of fires in Madrid are human-caused (~90%), whereas in Aragón, the causality of lightning-fires is more important, representing the ~30% of total reported fires. The descriptive statistic showed that the weighted average integration has a higher variability of values than the other two approaches, and therefore it seems that this approach has a better capability of representing the different factors affecting ignition.

Regarding the validation of the obtained results using the independent ignition points sample from 2005-2007, in both study areas the human-caused probability model reaches a reasonable discrimination by using ROC (~0.7). The lightning-caused models showed a poorer performance, being satisfactory in the region of Aragón (0.66) but not getting a significant discrimination in the region of Madrid (0.55). Related to the integrated results, in the region of Aragón the discrimination is less satisfactory, while in the region of Madrid it reaches reasonable discrimination being better when using the weighted probability. Despite validation results show a satisfactory discrimination in this area the AUC of the integrated models is lower than for the human-caused model. The results of the Mahalanobis distances agree with the ROC validation, where significant highest distances are found when using the weighted average in Madrid. In the region of Aragón the human-caused model has been obtained using a small number of fire ignition points as dependent variable. This was due to the location errors that were found in the original dataset. This may be causing an underestimation of the wildfire occurrence due to human causes. Also, in Aragón there are more fires due to lightning than in the region of Madrid. In this region the

lightning-caused model hardly ever influences the final integrated result (except for north areas of the Central range), due to its low probability values. Indeed, the integrated model performs similarly but poorer than the human model alone due to the influence of the lightning model. We have tested these results with the ones obtained at a coarser scale, 3×3 km grid cell resolution. The obtained AUC follow the same trends as in the previous scale but with higher values, meaning that the models perform better at coarse resolution. At 3×3 km grid cell resolution the human-caused and the weighted average integrated models obtain the higher AUC values, being for the integrated model 0.639 in Aragón and 0.719 in Madrid. This better fit indicates that the change in the scale is affecting the results. With a coarser scale the uncertainty in the fire ignition spatial location clearly decreases. In addition, the individual models have been calibrated only with three years of fire data, due to restrictions imposed by the availability of meteorological data (from 2002 to 2004), and this may be influencing in the final result (Nieto et al. in revision). As more data become available the fitting and validation of the individual models could be improved as well as the integrated results. In addition, a longer time-series of fire data should be required to improve independent validation and assessment of results. Despite of this, the probabilistic and weighted average methods tested to integrate lightning and human-caused models show satisfactory results compared with the single average method. Those integration alternatives try to be an objective method to weight the variables (human and natural causative agents) according to its relative importance on fire occurrence. Comparing the obtained results with related works we find that AUC values in the region of Madrid are close to those from Modugno et al. (2008) in their human-caused models (~0.7).

## 5. Conclusions

We have tested two methods (probabilistic and a weighted average) for the integration of lightning and human fire caused models at 1×1 km grid cell resolution. Two different regions regarding fire causality were compared to get a better performance assessment of both methodologies. The proposed methods show to be adequate if we compare them to the single average results but they reach a lower accuracy than the original lightning or human-caused models in the region of Aragón or than the human-caused model in the region of Madrid. The use of a weighted average by the historic fire occurrence fits better in both tested regions. Regarding the spatial distribution of the integrated fire risk, it fits with the spatial distribution of the phenomenon related to the fire causes of each

region. Validation is an essential step in the model performing, to test the sustainability of the predicted results and if their accuracy is consistent for the application of the model. Validation results in Madrid are satisfactory whereas Aragón models would require a revision of original data. Despite of the effort we made to homogenise the ignition fire models due to human and natural causes, better original fire data should lead to better models and conclusions. The spatial inaccuracy of fire statistics may be having an effect in the final results. As we have tested, the integrated models perform better at a coarser scale. Also, when a longer time series data became available these results could be improved.

Long-term fire risk indices require the integration of fire ignition components (human and lightning). Those components have different impacts on fire risk conditions. Identifying which are more relevant and how they should be weighted to generate synthetic indices is a critical phase in risk assessment for fire prevention and mitigation management. The results obtained in this work show that proposed methods can be considered an appropriate alternative to a simple average to integrate the information of causative agents in order to estimate fire ignition probability. However, further assessment is required in other periods and regions to check consistency and generalization potential.

**Acknowledgements**

This research has been partially supported by the *Firemap* project CGL2004-06049-C04-01/CLI, funded by the Spanish Ministry of Education, through the FPI scholarship BES-2005-7712. Historic fire data was provided by the Fire Department of the Community of Madrid and the Spanish Ministry of Environment, Rural and Marine Areas. Other essential data has been provided by the Regional Environmental Office of Madrid, the Department of Geography of the University of Zaragoza and the firm Meteológica, and the Spanish Meteorological Agency.

**References**

Amatulli, G., Rodrigues, M.J., Trombetti, M., Lovreglio, R., 2006. Assessing long-term fire risk at local scale by means of decision tree technique. Journal of Geophysical Research 111, G04S05, doi:10.1029/2005JG000133.

Amatulli, G., Pérez-Cabello, F., de la Riva, J., 2007. Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty. Ecological modelling 200, 321-333.

Amatulli, G., Camia, A., 2007. Exploring the relationships of fire occurrence variables by means of CART and MARS models. Wildfire 2007. IV International Wildfire Conference, Seville, Spain, 13-17 May.

Aspinall, R., 1992. An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. Int.J. Geogr. Inf. Syst. 6 (2), 105–121.

Brzeziecki, B., Kienast, F., Wildi, O., 1993. A simulated map of the potential natural forest vegetation of Switzerland. J.Veg. Sci. 4, 499–508.

Chao-Chin, L., 2002. A Preliminary Test of A Human caused Fire Danger Prediction Model. Taiwan Journal Forest Science 17(4), 525-529.

Chuvieco E., Salas, F.J, Carvacho, L., Rodríguez Silva, F., 1999. Integrated fire risk mapping. In Chuvieco, E. (Ed.), Remote Sensing of Large Wildfires in the European Mediterranean Basin .Berlin: Springer-Verlag, pp. 61-84.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M.P., Vilar, L., Martínez, J., Martín, S., Ibarra, P., De la Riva, J., Baeza, J., Rodríguez, F., Molina, J.R, Herrera, M.A., Zamora, R. 2009. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling In press doi:10.1016/j.ecolmodel.2008.11.017.

Cuadrat Prats, J.M. 2004. El clima de Aragón. In J.L Peña, L.A Longares and M.Sánchez (Eds.): Geografía Física de Aragón. Aspectos generales y temáticos. Universidad de Zaragoza e Institución Fernando el Católico. Zaragoza. Available at http://age.ieg.csic.es/fisica/XIXJornadas/Documentos/003.pdf (Last accessed April 23, 2009).

FAO (Food and Agricultural Organization of the United Nations) (2007). Fire Management Global Assessment. A thematic study prepared in the framework of the Global Forest Resources Assessment 2005. *FAO Forestry Paper 151*. FAO, Rome. Available at http://www.fao.org/forestry/fra2005/en/ (Last accessed April 23, 2009).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Fielding, A.H., Bell, J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24 (1), 38-49.

Fischer, H.S., 1990. Simulating the distribution of plant communities in an alpine landscape. Coenoses 5, 37–43.

Gobierno de Aragón. Available at http://www.Aragón.es/pre/cido/vegeta.htm (Last accessed April 23, 2009).

Guisan, A., Harrell, F.E. 2000. Ordinal response regression models in ecology. Journal of Vegetation Science 11, 617-626.

Guisan, A., Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135, 147-186.

Huntley, B., Berry, P.M., Cramer, W., McDonald, A.P., 1995. Modeling present and potential future ranges of some European higher plants using climate response surfaces. J. Biogeogr. 22, 967–1001.

Kalabokidis, K. D., Koutsias, N., Konstantinidis, P., Vasilakos, C., 2007. Multivariate analysis of landscape wildfire dynamics in a Mediterranean ecosystem of Greece. Area 39 (3), 392-402.

Koutsias, N., Kalabokidis, K.D., Allgöwer, B., 2004. Fire occurrence patterns at landscape level: beyond positional accuracy of ignition points with kernel density estimation methods. Natural Resource Modeling 17 (4), 359-375.

Lasanta, T., González-Hidalgo, J.C., Vicente-Serrano, S.M., Sferi, E. 2006. Using landscape ecology to evaluate an alternative management scenario in abandoned Mediterranean mountain areas. Landscape Urban Planning 78 (1–2), 101–114.

Lin, C., 1999. Modelling probability of ignition in Taiwan Red Pine Forests. Taiwan Journal Forest Science 14 (3), 339-344.

MARM, Spanish Ministry of Environment, Rural and Marine Affairs, 2006. Subsecretaría General de política forestal y desertificación. Área de defensa contra incendios forestales. Los incendios forestales en España. Decenio 1996-2005. Available at http://www.mma.es/portal/secciones/biodiversidad/defensa_incendios/estadisticas_incendios/pdf/estadisticasdecenio_1996-2005.pdf (Last accessed April 23, 2009).

Martell, D.L., Otukol, S., Stocks, B.J., 1987. A logistic model for predicting daily people caused forest fire occurrence in Ontario. Caumar.

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. In Chuvieco, E., Martín, M.P. (Eds.), Nuevas tecnologías para la

estimación del riesgo de incendios forestales. CSIC, Instituto de Economía y Geografía, Madrid, pp. 101-142.

Martínez, J., Vega-García, C., Chuvieco, E., 2009. Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90, 1241–1252.

Mazzotti, F.J., Vinci, J.J. 2007. Validation, Verification, and Calibration: Using Standardized Terminology When Describing Ecological Models. WEC 216, Wildlife Ecology and Conservation Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida. Available at http://edis.ifas.ufl.edu (Last accessed April 23, 2009).

McKenzie, D., Peterson, D.L., Agee, J.K. 2000. Fire frequency in the interior Columbia River Basin: building regional models from fire history data. Ecological Applications 10 (5), 1497-1516.

Modugno, S., Serra, P., Badia, A. 2008. Dinámica del riesgo de ignición en un área de interfase urbano-forestal. In Hernández, L., Parreño J.M. (Eds.). XIII Congreso Nacional de Tecnologías de la Información Geográfica, Las Palmas de Gran Canaria, Spain, 15-19 September.

Mooney, H.A., Bonnicksen, T.M., Christensen, N.L., Lotan, J.E., Reiners, W.A. 1981. Fire regimes and ecosystem properties. Proceedings of the Conference USDA For. Serv. Gen. Tech. Rep. WO-26, 594 pp.

Moreno, J.M. 1989. Los ecosistemas terrestres mediterráneos y el fuego. Política científica 18, 46-50.

Nicolás, J.M. Caballero, D., 2001. Demanda territorial de defensa contra incendios forestales. Un caso de estudio: Comunidad de Madrid. Proceedings of Spanish National Forest Congress. Palacio de Congresos y Exposiciones, Granada, Spain, 25-28 September. Available at http://www.gnomusy.com/publications/20021025_Caballero_Geodatabases_Poster.pdf (Last accessed April 23, 2009).

Nieto, H., Aguado, I., Chuvieco, E. García, M.. (in revision) Lightning-caused fires in Central Spain: development of a probability of occurrence model in two Spanish regions. International Journal of Wildland Fire.

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, 1-18.

Pontius Jr, Robert Gilmore and Laura Schneider. 2001. Land-use change model validation by a ROC method for the Ipswich watershed, Massachusetts, USA. Agriculture, Ecosystems & Environment 85(1-3), 239-248.

Prasad, V.K., Badarinath, K.V.S., Eaturu, A., 2008. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. Journal of Environmental Management 86, 1, 1-13.

Preisler, H. K., Brillinger, D.R, Burgan, R.E., Benoit, J.W. (2004). Probability based models for estimation of wildfire risk. International Journal of Wildland Fire133-142.

Pyne, S.J., 1982. Fire in America: A Cultural History of Wildland and Rural Fire. Princeton University Press, Princeton, NJ, 654 pp.

Renken, H., Mumby, P. J. 2009 Modelling the dynamics of coral reef macroalgae using a Bayesian belief network approach. Ecological modelling, in press. Doi: 10.1016/j.ecolmodel.2009.02.022.

Robin, J.G., Carrega, P., Fox, D., 2006. Modelling fire ignition in the Alpes-Maritimes Department, France. A comparison. In Viegas D.X (Ed.) V International Conference on Forest Fire Research, Figueira da Foz, Portugal, 27-30 November.

Romero-Calcerrada, R. N., C. J., Millington, J. D. A., Gomez-Jimenez I., 2008. GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). Landscape Ecology 23, 341-354.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229-244.

Sebastián López, A., Burgan, R.E., San-Miguel Ayanz, J. 2001. Assessment of fire potential in Southern Europe. V. e. Forest Fire Research & Wildland Fire Safety. Rotterdam, Millpress.

Skidmore, A.K., 1989. An expert system classifies eucalypt forest types using Landsat Thematic Mapper data and a digital terrain model. Photogramm. Eng. Remote Sen. 55, 1449–1464.

Soudant D., Beliaeff B., Thomas G. 1997. Dynamic linear Bayesian models in phytoplankton ecology. Ecological Modelling 99(2-3), 161-169.

Syphard, A.D., Radeloff, V.C., Keeley, J.E., Hawbaker, T.J., Clayton, M.K., Stewart, S.I., Hammer, R.B., 2007. Human influence on California Fire Regimes. Ecological Applications 17, 5, 1388-1402.

Stehman, S.V. 1999. Comparing thematic maps based on map value. International Journal of Remote Sensing 20, 2347-2366.

Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. Society for Industrial and Applied Mathematics, Philadelphia, 342 pp.

Vasconcelos, M.P.P., Silva, S., Tomé, M., Alvim, M., Pereira, J.M.C., 2001. Spatial Prediction of Fire Ignition Probabilities: Comparing Logistic Regression and Neural Networks. Photogrammetric Engineering and Remote Sensing 67 (1), 73-81.

Vega-García, C., 2007. Propuesta metodológica para la predicción diaria de incendios forestales. Proceedings of IV International Wildfire Conference, Seville, Spain, 13-17 May.

Viegas, D.X., Bovio, G., Ferreira, A., Nosenzo, A., Sol, B. 1999. Comparative study of various methods of fire danger evaluation in Southern Europe. International Journal of Wildland Fire 9(4), 235-246.

Vilar del Hoyo, L., Martín Isabel, M.P, Martínez Vega, F.J. 2008. Empleo de técnicas de regresión logística para la obtención de modelos de riesgo humano de incendio forestal a escala regional. Boletín de la Asociación de Geógrafos Españoles 47, 5-29.

Wilson, D. S., Stoddard, M.A, Puettmanna, K. J. 2008. Monitoring amphibian populations with incomplete survey information using a Bayesian probabilistic model term. Ecological Modelling 214 (2-4), 210-218.

Wotton, B. M., Martell, D.L (2005). A lightning fire occurrence model for Ontario. Canadian Journal of Forest Research 35, 1389-1401.

Yang, J., He, H.S., Shifley, S.R., Gustafson, E.J., 2007. Spatial Patterns of Modern Period Human-Caused Fire Occurrence in the Missouri Ozark Highlands. Forest Science 53, 1-15.

# LÍNEAS FUTURAS

# LÍNEAS FUTURAS

El desarrollo de modelos espaciales de ocurrencia para la predicción de incendios por causa humana resulta de interés para su inclusión en sistemas generales de riesgo de incendio. A pesar del la importancia del factor humano en la ignición de los incendios forestales no suele ser incluido en los mismos debido a la dificultad que supone modelizar el comportamiento humano. Esta tesis propone una serie de metodologías y aproximaciones para la consecución de modelos de riesgo de incendio forestal basados en factores socioeconómicos. Sin embargo, la línea de investigación se encuentra abierta para seguir profundizando en el logro de modelos mejor ajustados. En este sentido se plantean las siguientes líneas de investigación futuras

- Se propone seguir trabajando en la inclusión del factor temporal.

  - Aplicando *Generalizad Additive Models* (*GAMs*) en otras áreas de estudio.
  - Integrando los modelos obtenidos mediante esta técnica con modelos de incendios producidos por rayo.

- Se propone la obtención de modelos a escalas más groseras, que, según el área de estudio, puede resultar también de interés para la gestión (3×3 km o 10×10 km). De igual forma, se propone el desarrollo de modelos regionales dentro de una misma área de estudio, con el objetivo de detectar diferencias locales.

- En la línea de los análisis espaciales, se propone el desarrollo de modelos que incluyan el factor socioeconómico empleando técnicas de regresión espacial (*Geographical Weighted Regression*, *GWR*).

- Por último, teniendo en cuenta el interés de conocer la influencia de los incendios forestales en los cambios que se han producido en el territorio de cara a entender los cambios pasados y modelizar escenarios futuros, se propone la aplicación de *GAMs* para hallar relaciones entre los cambios en el territorio con los cambios en el patrón de incendios, así como con los factores socioeconómicos.

# ANEXO I

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

**Publicación derivada:** *Vilar del Hoyo, L., Gómez Nieto, I., Martín Isabel, M.P., Martínez Vega, F.J (2007). Análisis comparativo de diferentes métodos para la obtención de modelos de riesgo humano de incendios forestales. Comunicación presentada a Wildfire 2007, 4th International Wildland Fire Conference, 13-17 Mayo 2007, Sevilla, España*

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

**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 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 % 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 Decisión 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.

## 1. 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 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[11], 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 et al., 2004). Este planteamiento es objeto de estudio del proyecto *Firemap* "Análisis Integrado de Incendios Forestales mediante Teledetección y Sistemas de Información Geográfica"[12] (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 socioeconómicos 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.

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

## 2. Material y métodos

### 2.1 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.

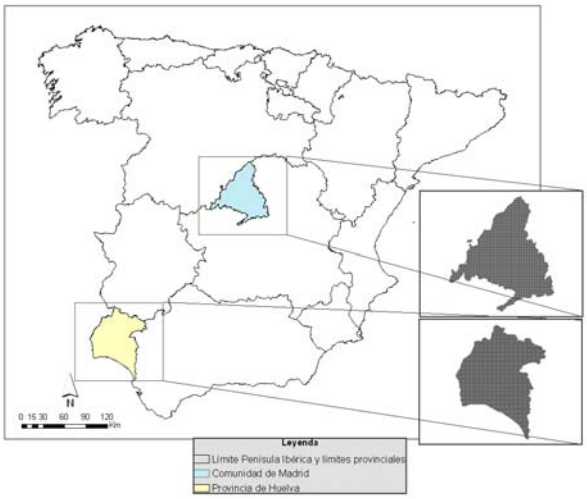


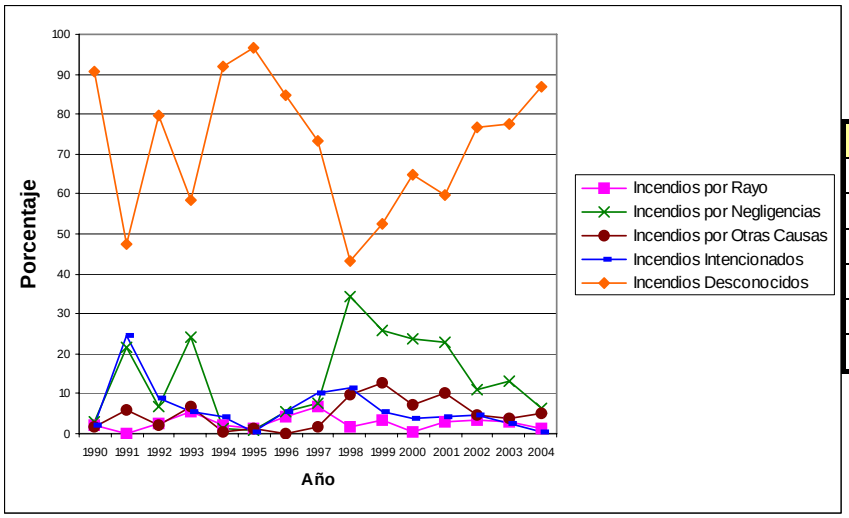

**Figura 1**—Zonas de estudio. Comunidad de Madrid y provincia de Huelva (España)

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[13]), 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

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

incendios en la región en los últimos 15 años. Destaca el alto porcentaje de causas desconocidas y la notable proporción de incendios por negligencia.

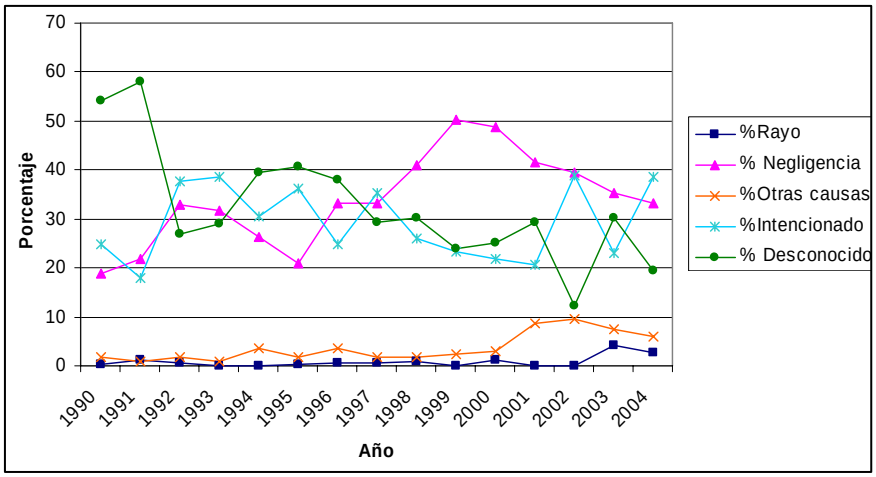



| Causa | Nº | % |
|---|---|---|
| Rayo | 86 | 2,36 |
| Negligencia | 461 | 12,65 |
| Otras causas | 175 | 4,80 |
| Intencionados | 177 | 4,86 |
| Desconocidos | 2745 | 75,33 |
| Total | 3644 | 100 |

**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 km² 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%). 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 % 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 unidos, en este caso a un alto porcentaje de incendios intencionados.




| Causa | Nº | % |
|---|---|---|
| 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

2.2 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 et al., 2003, Martínez, 2004, Martínez et al., 2004, Pew et al., 2001, Vega-García et al., 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.

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

| Tipo | Factor | Variable |
|---|---|---|
| CARTOGRÁFICAS | **Incendios por accidente o negligencia** | Áreas de influencia (buffer) de vías sin pistas forestales[14] (CARRET, CARRET_FOR) |
| | | Índice de IMD por segmento de carretera (Longitud vía*IMD vía*factor de ponderación) (INDICE_IMD, INDICE_IMD_FOR)[15] |
| | | 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) |
| | | 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) |

[14] 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

[15] Caso de Madrid

2.3 Variable dependiente

La variable dependiente empleada es la ocurrencia de incendios por causa humana[16] 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 espacializar 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 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*[17], 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 et al., 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 et al., (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 et al., 2003 citando a Seaman y Powell, 1996; Levine, 2002):

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

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

[16] 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.

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

(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 et al., 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 et al., 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.

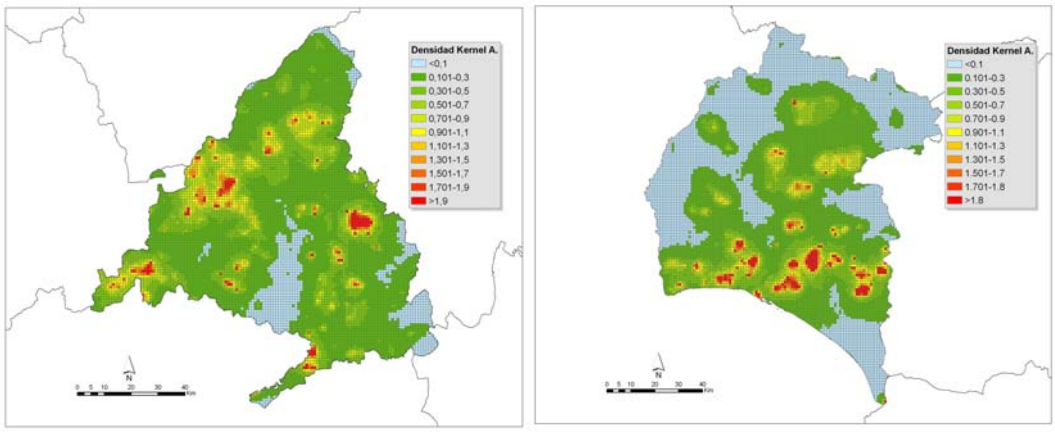



**Figura 4**. Variable dependiente obtenida a partir del método de interpolación Kernel adaptativo muestra 5 puntos. C. de Madrid, provincia de Huelva

## 2.4 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.

2.4.1 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:

$$P_i = \frac{1}{1+e^{-z}} \tag{2}$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_\rho X_\rho \tag{3}$$

Donde $P_i$ es la probabilidad de ocurrencia de incendio, $z$ la combinación de variables independientes con sus coeficientes de regresión ($\beta$), $X$ 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*:

$$\log\left(\frac{p}{1-p}\right) = x^T \beta \tag{4}$$

Siendo $x^T$ el vector de las variables explicativas y $\beta$ 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 et al., 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.

### 2.4.2 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 et al., 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 et al., 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.

### 2.4.3 Redes neuronales

Las redes neuronales emulan el sistema biológico de un modo simplificado (Bischof et al., 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:

$$a=S_{i=1,n}w_ix_i \quad (5)$$

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:

$$x=1/(1+e)^{-a/p} \quad (6)$$

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 et al., 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:

$$Dw_i=h(D-Y) \quad (7)$$

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):

$$H=(I+O)/2$$
$$H=I*O \qquad (8)$$
$$H=(I+O)^{1/2}$$
$$H=(I+O)^{2}$$

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 2 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 2 ("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.

**3. Resultados**

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

3.1 Comunidad de Madrid

3.1.1 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% 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% 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 obtuvieron 17 modelos, de ellos elegimos el modelo7º. Los porcentajes globales de acierto de clasificación de la muestra de elaboración del modelo (60%) y de validación del mismo (40%) son 71,6 y 70,3% 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 (9)

| | B | E.T. | Wald | Gl | Sig. | Exp(B) | I.C. 95,0 % para EXP(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Inferior | Superior |
| 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 | 1562817101497831000,000 | 5825112017197180,000 | 419287609495493800000000,000 |
| 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% 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% contiene el valor 1.

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

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).

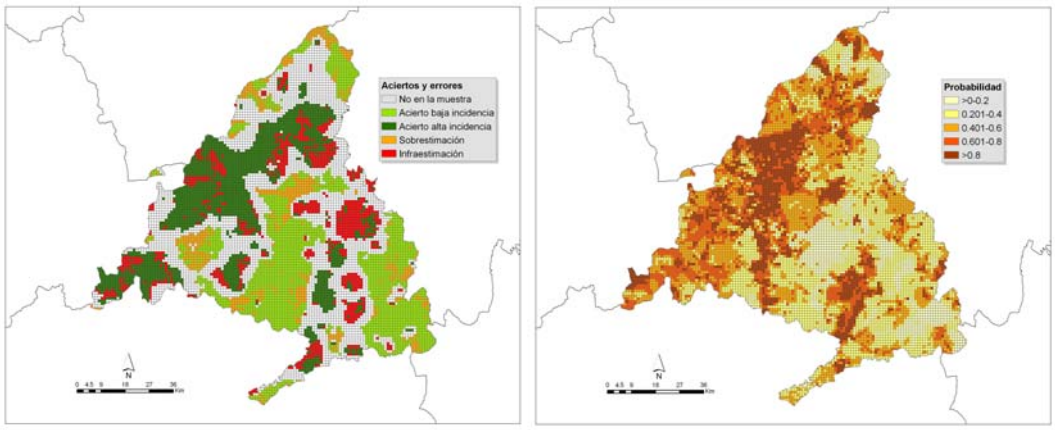



**Figura 5—**Mapas de aciertos y errores y de probabilidad estimada de riesgo humano (modelo 7)

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.

### 3.1.2 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%, clasificando correctamente tanto la baja incidencia de incendio como la alta en un 75,5%.

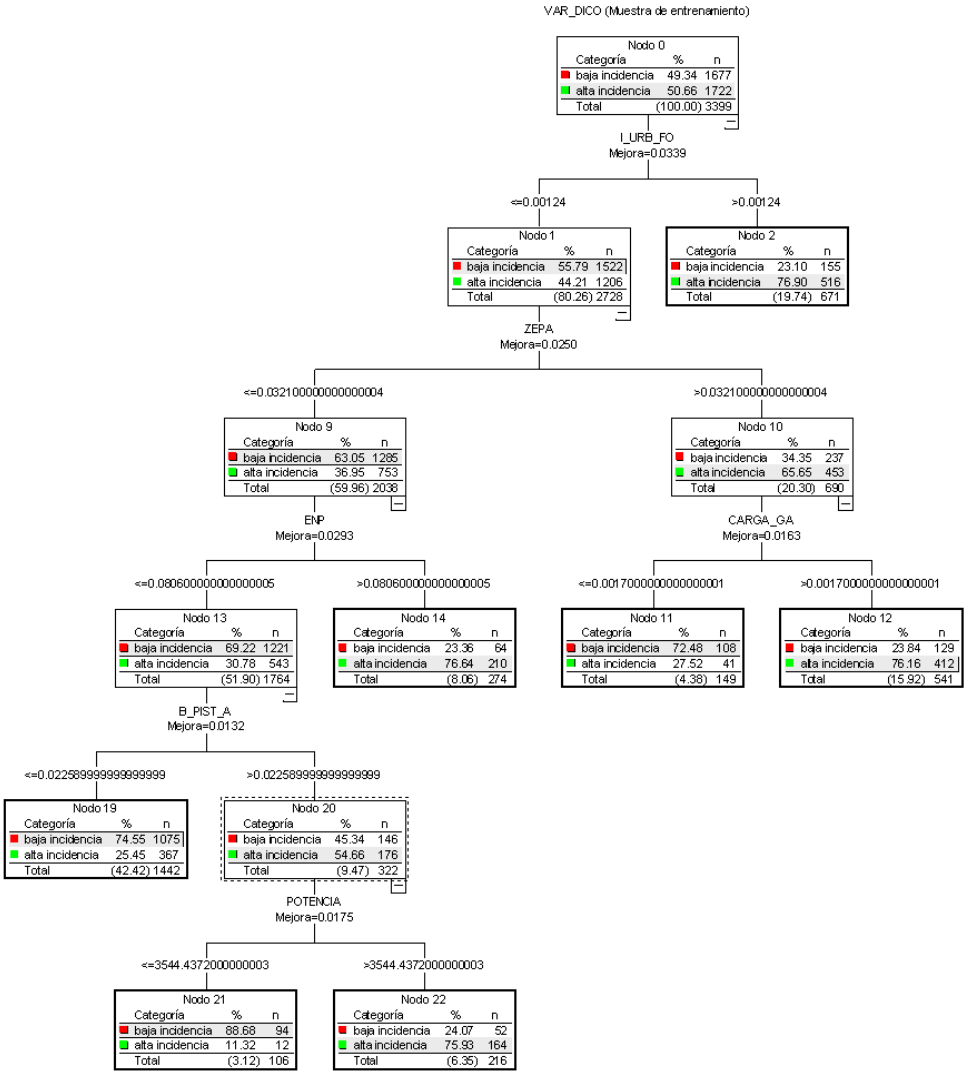



**Figura 6—** Árbol de Decisión en la Comunidad de Madrid

La variable *interfaz urbano-forestal* es la que determina la primera división del árbol, con una mejora del 3%. Si es mayor de 0,001, las cuadrículas que cumplan esta condición quedarán clasificadas en un 76,9% 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%. Si es mayor de 0,03, es la variable *carga ganadera* la que clasifica en un 76,2% 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 % y con una mejora del 3%. 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% 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%, mientras que serán clasificadas de baja incidencia (valor de potencial menor ó igual a 3544,4) en un 88,6%.

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

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

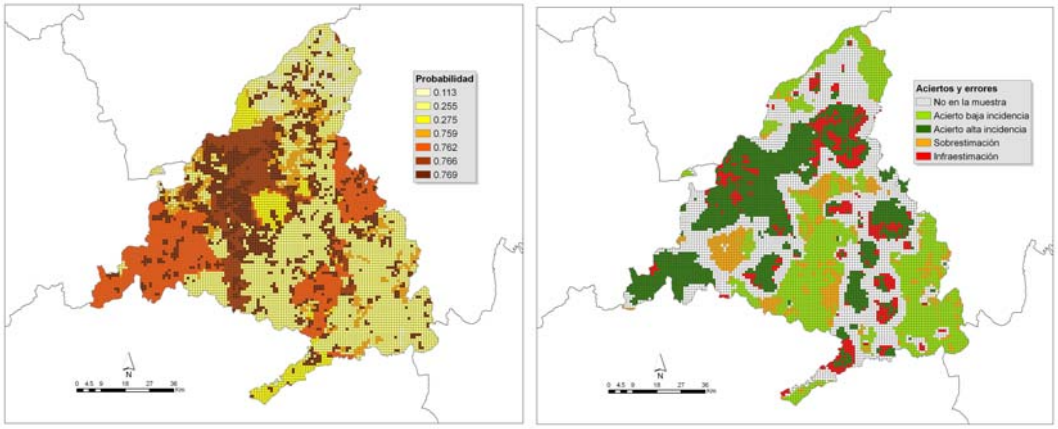



**Figura 7—** Aciertos y errores Árbol de Decisión y probabilidad estimada en la Comunidad de Madrid

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.

### 3.1.3 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 % | 62,78 % |
| Error comisión | 28,66 % | 37,22 % |
| Error omisión | 33,82 % | 31,79 % |

**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% mientras en la baja ocurrencia el método clasifica correctamente en un 62,78%. Los errores de comisión son 28,66 y 37, 22%, respectivamente, mientras que los de omisión superan el 30% 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.

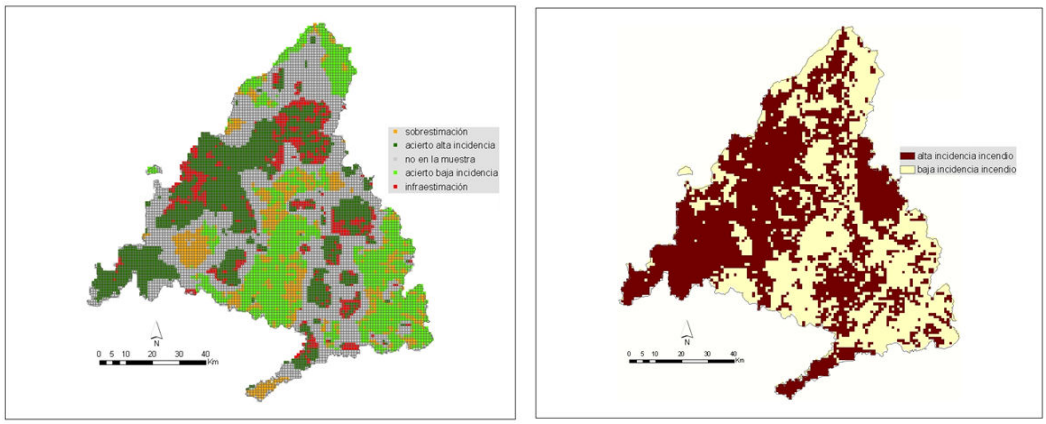



**Figura 8—** Aciertos y errores redes neuronales y ocurrencia estimada en la Comunidad de Madrid

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.

3.2 Provincia de Huelva

3.2.1 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%) y de validación del mismo (40%) son 84,2 y 85,2% 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 (10)

| | B | E.T. | Wald | gl | Sig. | Exp(B) | I.C. 95,0 % para EXP(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Inferior | Superior |
| 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 | 366268965584,715 | 43436214,205 | 3088504778896969,000 |
| 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% 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% de la muestra se obtiene un 84,4% correcto de clasificación global de la misma, estando la baja incidencia correctamente clasificada en un 92,4% y la alta incidencia en un 76,4%.

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.

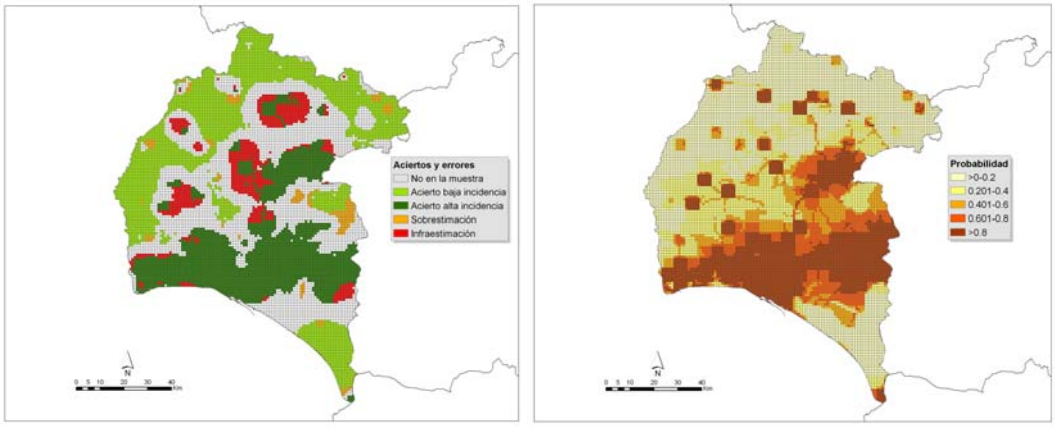



**Figura 9—** Mapas de aciertos y errores y de probabilidad estimada de riesgo humano (modelo 5)

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.

### 3.2.2 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%, clasificando correctamente la baja incidencia de incendio en un 87,4% y la alta en un 81,7%.

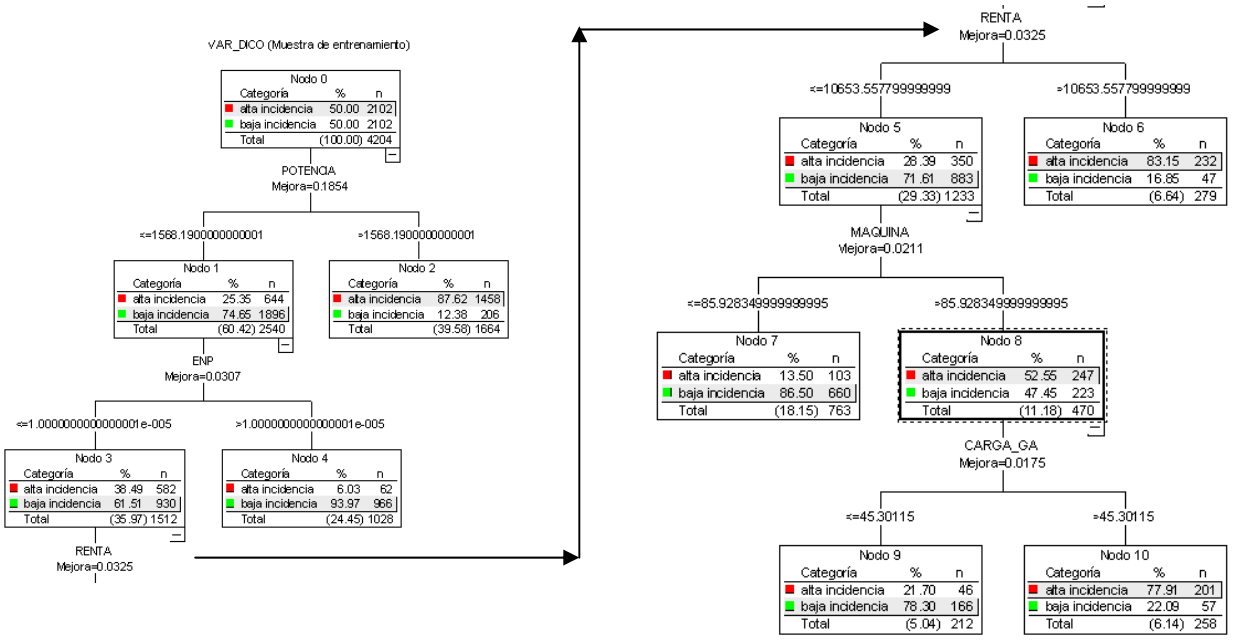



**Figura 10—** Árbol de Decisión

La variable *potencial demográfico* es la que determina el primer nivel, con una mejora del 18,5% clasifica en alta incidencia en un 87,6% 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%. 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%). Si la renta es menor ó igual a 10653, con una mejora del 2%, es la variable *maquinaria agrícola* la que clasifica las celdas en baja incidencia (86,5%) 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%).

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

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.

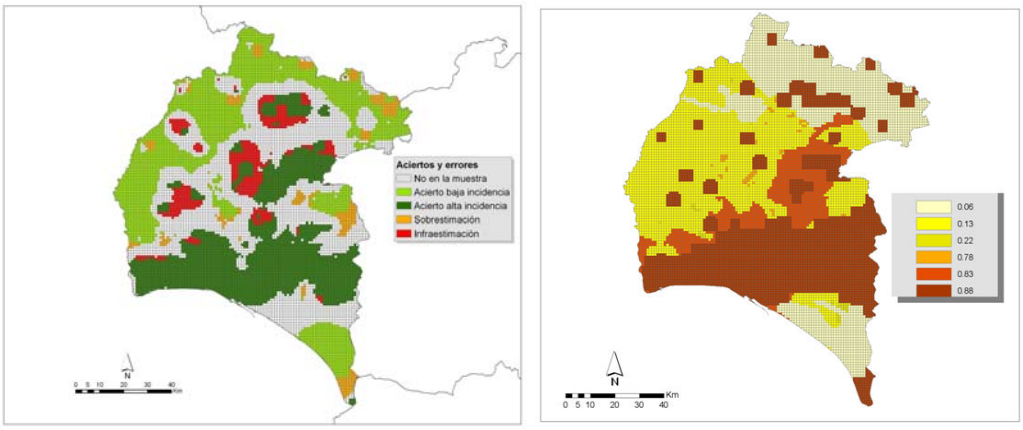



**Figura 11 –** Aciertos y errores Árbol de Decisión y probabilidad estimada en la Provincia de Huelva

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.

### 3.2.3 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 % | 44,81 % |
| Error comisión | 2,02 % | 55,19 % |
| Error omisión | 34,57 % | 4,58 % |

**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%, mientras que la baja incidencia se clasifica correctamente en un 44,81%. Los errores de comisión son 2,02 y 55,19% respectivamente mientras que los de omisión son del 34,57 y del 4,58%. 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.

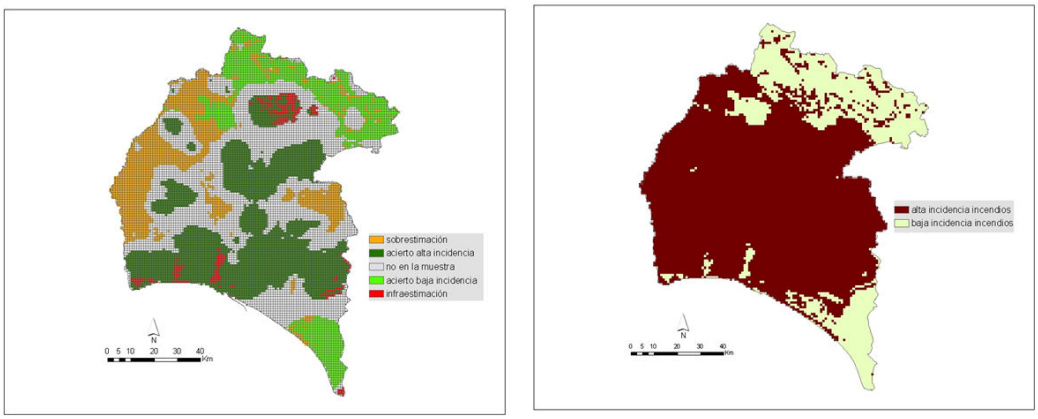



**Figura 12—** Aciertos y errores Redes neuronales y ocurrencia estimada en la Provincia de Huelva

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.

**4. 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%), aunque en las tres los porcentajes de clasificación son similares, tanto el global como la alta y la baja incidencia (alrededor del 70%). 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%, clasificando correctamente la alta incidencia en un 80% y la baja en un 90%. Sin embargo, la técnica de redes neuronales clasifica correctamente la alta incidencia en un 97% y la baja tan solo en un 44,8%, 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% 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.

**5. 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)[18] 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

[18] 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

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. Ecological 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*.

Breiman, L., Meisel W. and E. Purcell. 1997. Variable kernel estimates of multivariate densities. Technometrics 19, 135–144.

Bischof, H. Schneider, W. Pinz, A.J. 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, M.P. (Ed.): Nuevas tecnologías para la estimación del riesgo de incendios forestales. Madrid, CSIC, Instituto de Economía y Geografía, pp. 144-158.

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. Disponible en http://www2.chass.ncsu.edu/garson/pa765/statnote.htm (23 Abril 2009)

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. Disponible en http://www.juntadeandalucia.es/institutodeestadistica/dtbas (23 April 2009)

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, M.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.

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, 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, 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), 101–112.

Vega-García, C. 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 decision tree approach, Ecological Modelling, 186, 299-311.