DETERMINACIÓN DE VARIABLES DE RIESGO HUMANO EN LA GENERACIÓN DE INCENDIOS FORESTALES MEDIANTE REGRESIÓN LOGÍSTICA Y REDES NEURONALES

Luis Carvacho Bart[1], José Ignacio Barredo Cano[2], Emilio Chuvieco Salinero[2], Javier Martínez Vega[3], Javier Salas Rey[2]
[1] Instituto de Geografía, Pontificia Universidad Católica de Chile.
[2]Departamento de Geografía, Universidad de Alcalá
[3]Instituto de Geografía Aplicada, C.S.I.C.

RESUMEN

Se analizan diversas variables relacionadas con la generación de incendios forestales mediante dos pruebas estadísticas, regresión logística y redes neuronales. Se comparan los resultados y el rendimiento de ambas pruebas en la predicción de ocurrencia/no ocurrencia de incendios forestales en el área mediterránea europea. Finalmente, se indican las fortalezas y debilidades de cada una para este análisis.

1 Introducción

Dentro de las causas que originan los incendios forestales, aquellas relacionadas de una u otra forma con la intervención humana sobre el medio vegetal, aparecen muy destacadas en relación con aquellas de origen natural. El identificar cuáles son estas variables y el grado de responsabilidad que se puede asignar a cada una de ellas como influyentes en el proceso de generación, e incluso de propagación de los incendios forestales, es el objetivo del trabajo que se presenta.

Sin embargo, resulta evidente que circunstancias no imputables a la acción humana influyen también decisivamente en el grado de destrucción que un incendio forestal puede ocasionar. De esta forma, es también interesante analizar las condiciones naturales y las variables de actividad o de acción humana en el medio que hacen más probable la aparición de Incendios forestales, cuyo concepto quisiéramos establecer claramente. "Incendio forestal, es el fuego que se extiende sin control sobre un terreno forestal" (ICONA, 1982). De esta forma, el antiguo "*Instituto para la Conservación de la Naturaleza*" español definía el concepto. Se trata, según esto, de un *fuego sin control*, o de un accidente provocado por el fuego (Salas, 1994). Precisamente el hecho de que se trate de un fuego *sin control* es lo que diferencia a un incendio forestal de otras prácticas forestales o agrícolas que utilizan el fuego como herramienta. No se consideran incendios las quemas de pastos o en general, el empleo del fuego para la eliminación de residuos forestales si éste no se extiende más allá de la zona que se pretendía quemar. (Salas, 1994)

Aunque es posible reconocer también al incendio forestal como un proceso natural e integral de los ecosistemas (Evangelia y Costa, 1992, Aguado, 1997), es preciso tener presente que aquellos fuegos con una *causa* natural en sus orígenes, ha descendido notablemente en relación con el total de incendios, hasta el punto que menos del 5% de los incendios registrados en España en los últimos años, son imputables a causas

naturales, en especial a rayos durante las tormentas (Tárrega y Luis, 1992). En otros países de la cuenca mediterránea, las causas naturales de los incendios forestales virtualmente no existen. En Grecia, la totalidad de los incendios forestales tienen una causalidad humana, por ejemplo (Papastavrou, 1997).

De la gran cantidad de variables involucradas de una u otra forma en la generación de incendios forestales, se hace necesario encontrar aquellas que con mayor probabilidad sean las que originan los siniestros, de forma tal de buscar las soluciones adecuadas que ayuden a prevenir la ocurrencia de los incendios atacando sus orígenes. A la búsqueda de estas variables, se dirige este trabajo.

2 Área de Estudio

Se aplicará el estudio sobre 164 provincias de la cuenca mediterránea, 48 españolas, 20 italianas, 52 griegas, 18 portuguesas y 26 francesas.

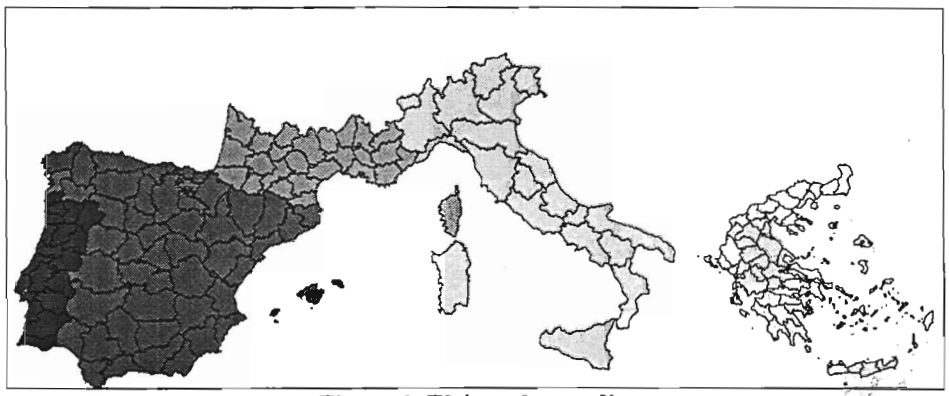

Figura 1. El área de estudio

3 Metodología

El estudio aplicará dos técnicas estadísticas para buscar los siguientes objetivos:

a. Una buena predicción en la ocurrencia – no ocurrencia de incendios
b. La selección de las variables más significativas en esta predicción

Se aplicará un modelo de regresión logística y un modelo de redes neuronales para cada uno de los objetivos planteados y se compararán los resultados obtenidos por cada método y las debilidades y fortalezas de ellos. El primero de estos modelos ha sido utilizado previamente en el análisis de ocurrencia de incendios logrando buenos resultados (Martell et al., 1987; Loftsgaarden y Andrews, 1992; Chou et al., 1993; (Vega, 1996). Las redes neuronales se están utilizando de forma creciente en una gran variedad de estudios donde los valores esperados se obtienen a partir de muestras de valores conocidos. (Benediktsson et al., 1990; Civco, 1993).

3.1 El modelo logit

La regresión logística, permite obtener una salida de tipo binario (0/1) para un conjunto de variables para las cuales no es posible determinar *a priori* si la relación entre ellas es lineal o no lineal. Puesto que éste es nuestro caso, y lo que se pretende modelar es la "ocurrencia o no ocurrencia de incendios" por provincia, parece conveniente utilizar este modelo. El análisis de regresión logística, se basa en la siguiente función:

$$f(z) = \frac{1}{1+e^{-z}}$$

donde z se obtiene por una combinación lineal estimada de las variables independientes mediante una ajuste de máxima probabilidad:

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

donde α es la constante y β_n el factor de ponderación de la variable n. La función $f(z)$ convierte los valores de z en una función continua, cuyo rango oscila entre 0 y 1. De este modo, los valores obtenidos de z menores a 0.5 se asignan a no ocurrencia del fenómeno, y los iguales o superiores a este valor, a ocurrencia.

Para evitar el efecto de colinealidad en las variables, se efectuó un análisis previo para buscar correlaciones no deseadas, eliminando aquellas que no aportaran significación al estudio. Algunas variables se normalizaron utilizando transformaciones logarítmicas, que aunque no es un requerimiento para la regresión logística, asegura una estimación más robusta, así como la posibilidad de comparar los resultados con otros métodos.

3.2 Redes neuronales

Procuraremos, en primer lugar, entregar una definición de redes neuronales con el fin de comprender su principio de funcionamiento y el tipo de respuesta que entregan.

> *"Las redes neuronales artificiales son sistemas de trazado no lineal con una estructura basada ligeramente en los principios observados en los sistemas nerviosos biológicos. (...) Se pueden utilizar para aprender la relación que existe entre un conjunto de datos de entrada y otro de salida. (...) Todo lo que se requiere para entrenar una red neuronal, es un conjunto de datos que contengan la relación entrada/salida"* (Jensen, 1997).

El fundamento de las redes neuronales se encuentra entonces en los siguientes aspectos:

1. La existencia de una relación entre los datos de entrada y de salida.
2. La capacidad de entrenamiento o de "aprendizaje" de la red.
3. La capacidad de utilizar ese entrenamiento para aplicarlo a otro conjunto de datos y predecir nuevos resultados, lo que se conoce como *explotación* de la red.

Una de las grandes ventajas de las redes neuronales como elemento de análisis, es que no está constreñida a la existencia de una distribución normal de los datos de entrada, y de hecho, de ningún tipo de distribución (Benediktsson,1990, Civco, 1993). Tampoco la presencia de valores extremos afecta a una red neuronal, ni las relaciones lineales entre las variables, ni la existencia o no existencia de autocorrelación espacial (Openshaw, 1994). Ello permite combinar variables de distinto origen, sin la necesidad de tomar las precauciones que se precisan en los métodos más convencionales.

El entrenamiento de la red se realiza sobre un 80% de las muestras identificadas en los datos de partida. El 20% restante se emplea para comprobar la precisión de los valores que se van obteniendo en el proceso en cada iteración. Este procedimiento de comprobación es conceptualmente simple, y consiste en aplicar el método aprendido por la red a ese 20% de datos, comparando sus valores de salida con los observados en la realidad para ellos.

La selección del algoritmo de entrenamiento de la red se realiza mediante distintas pruebas de ensayo y error, buscando aquel que entregue el menor error medio cuadrático y que a la vez no produzca el fenómeno de "sobre-entrenamiento" de la red. En cuanto el error general de los valores de salida estimados por la red en comparación a los observados se hace menor que un umbral previamente establecido, se puede decir que se ha obtenido un modelo ajustado perfectamente, que la red "ha aprendido" las relaciones que existen entre las variables de acuerdo a los datos con que se ha calibrado.

La tabla 1 muestra las variables que se procesarán por los métodos ya descritos:

Nombre variable	Explicación
Sup_ForKm2	Superficie forestal
SupFor_pc	Superficie forestal relativa
Densi91(hab/Km2)	Densidad de población en 1991
PobAct90_pc	Población económicamente activa en 1990, relativa
DifPobAct_pc	Relación de población activa entre 1960/1990
Desem90_pc	Población en desempleo 1990, relativa
BSh+Csa_PC	Porcentaje de la provincia bajo este tipo climático
BSk_PC	
Cfa_PC	
Cfb+Cfc_PC	
Csb+Csc_PC	
Dfb+Dfc_PC	
AltitMed	Altitud media de la provincia
BRLEAF	Porcentaje de bosque de frondosas
DisCar_m	Distancia a carreteras en metros
Agricult_pc	Porcentaje del suelo destinado a la agricultura (siembra)
DensLuc(promed)	Densidad de luces
DenEmp90	Relación de empleo agrícola respecto a la superficie provincial
SupAgr90_pc	Porcentaje de la superficie provincial destinado a agricultura
DenBov90	Densidad de la masa bovina de la provincia
DenCap90	Densidad de la masa caprina de la provincia

Tabla 1. Variables utilizadas en el estudio

4 Resultados

4.1 Regresión logística

Tras generar una matriz de correlación para el conjunto de variables, y combinar las variables climáticas, se escogieron 10 de ellas, cuya correlación con otras era inferior a 0.5. Recordemos que para que una regresión sea significativa, las variables independientes no deben estar correlacionadas entre sí.

El modelo final fue calculado sobre una base de 161 provincias, descartándose 3 que tenían un comportamiento anómalo.

El modelo final generado por la regresión logística fue:

z = 0.8349 LDENSI91 -0.2168 POBACT91 + 0.6631 ALTIMED – 0.3524 BRLEAF + .0229 CLIM_BS

donde:

DENSI91	densidad de población
POBACT91	población económicamente activa de 1991
ALTIMED	altitud media
BRLEAF	proporción de bosque de frondosas
CLIM_BS	proporción de climas B y Cs (Köppen)

Los signos de los coeficientes son lógicos, ya que se esperan más incendios a mayor densidad de población, altitudes mayores, climas más áridos con menos población activa y menor área cubierta por bosque de frondosas.

El nivel de significación es mayor a un 99% para la población activa, clima y densidad de población, que deben considerarse las variables más relacionadas con la ocurrencia de grandes incendios.

La tabla siguiente muestra la precisión general del modelo de acuerdo a la ecuación enunciada anteriormente:

	Predicho		Porcentaje correcto
Observado	0	1	
0	38	25	60.32 %
1	10	88	89.80 %
General			78.26 %

Tabla 2. Precisión del modelo logístico

Como se puede ver, el porcentaje de acierto es muy alto, considerando que se aplica una ecuación única a un área de estudio de gran diversidad, tanto geográfica como de aspectos nacionales. Este acierto es aún mayor cuando se consideran sólo los errores de omisión, pues sólo un 10.2% de los incendios predichos son erróneos. Si bien los errores de comisión son más altos, cerca de un 40% (incendios predichos donde no los hubo), ellos son de menor importancia desde el punto de vista de la prevención de incendios. La figura 2 muestra la distribución del acierto/no acierto en el área de estudio.

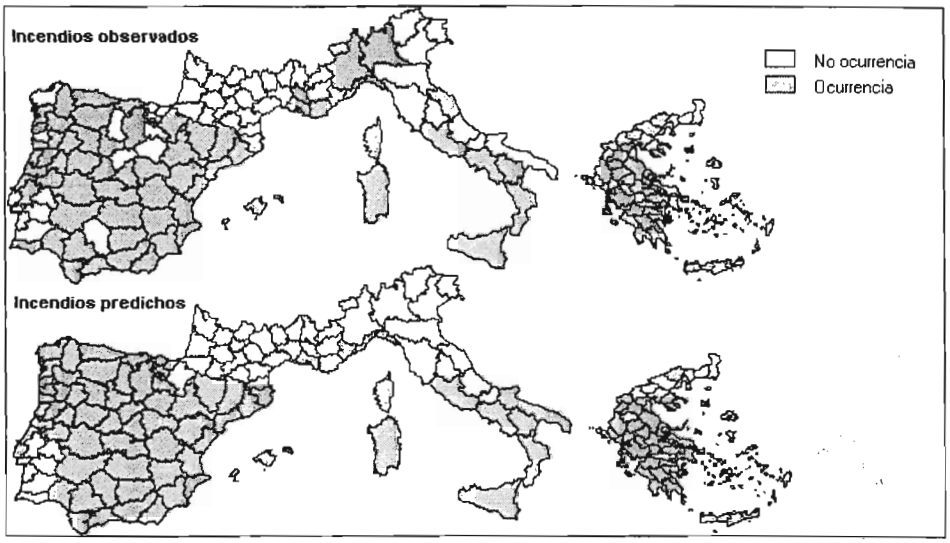

Figura 2. Distribución de incendios observados y predichos por regresión logística

4.2 Red neuronal

El ensayo y error de distintos algoritmos de entrenamiento de la red neuronal, dio como resultado que el mejor método para entrenar las variables de entrada para este caso, es el denominado "*QuickProp*". La topología de la red se construyó sobre la base de 3 capas ocultas, 6, 1 y 1 neurona, respectivamente.

De acuerdo con lo anterior, puede establecerse que el entrenamiento de la red, con el algoritmo y la topología seleccionada, ha sido satisfactorio para el intento de encontrar relaciones lógicas entre las variables de entrada.

A diferencia del modelo de regresión logística, la red neuronal no necesita asumir ningún tipo de supuesto previo respecto a las variables de entrada, de modo que no se ha efectuado ningún tipo de búsqueda o filtrado de ellas. Han entrado todas al análisis. La red debe discriminar por sí sola las variables relevantes.

Asumiendo, que un valor de salida inferior a 0.5 indica "no ocurrencia" y uno igual o superior a este valor como "ocurrencia", el acierto de la red es pleno (figura 3).

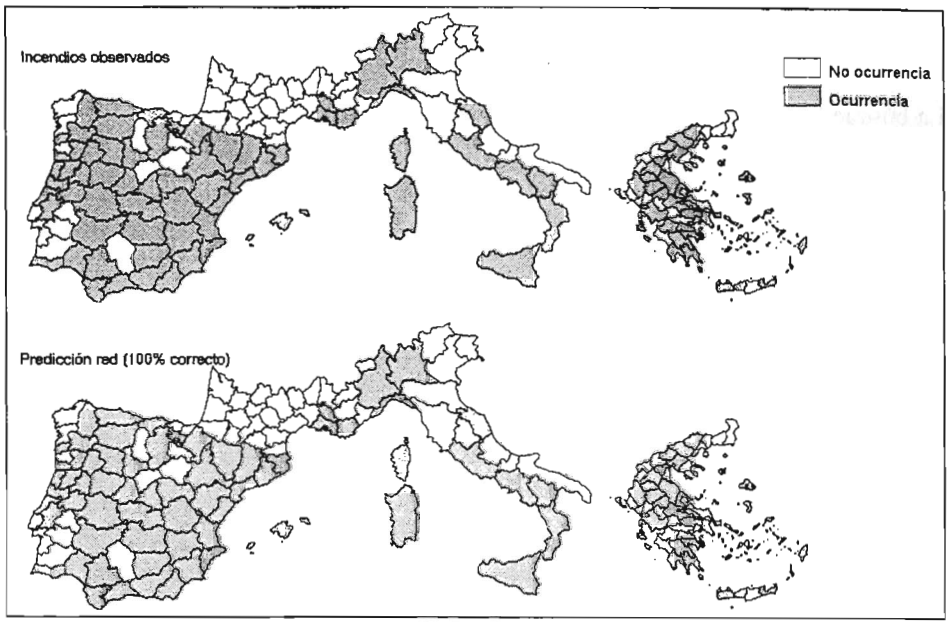

Figura 3. Acierto en la predicción de ocurrencia/no ocurrencia de incendios por la red neuronal

4.2.1 Variables predictoras

Si bien la idea de las redes neuronales es encontrar la forma en que las distintas variables se combinan para explicar los valores de aprendizaje, no es posible obtener directamente de ellas ni la "función de ajuste" (que puede no existir) ni las variables con mayor influencia en el comportamiento aprendido. Es posible, sin embargo, buscar estas últimas mediante un artificio.

Ya se ha dicho que lo que se obtiene de la red neuronal es un aprendizaje, representado por los pesos que combinan las neuronas de las diferentes capas de la topología de red escogida. Al aplicar esos pesos a las variables de entrada se obtienen los valores predichos por la red. Si se modifican los valores de partida, introduciendo valores aleatorios en una variable, se incrementará el Error Cuadrático Medio (ECM), en proporción equivalente a la importancia de la variable que estemos modificando. En otras palabras, asumimos que una variable será tanto más destacada en el ajuste conseguido por la red, cuanto mayor peso tenga en el incremento del ECM.

Los resultados de dichas pesquisas deben tomarse con extrema precaución, ya que las relaciones que las redes establecen entre las variables no son necesariamente lineales, y pueden ser totalmente diferentes entre distintas variables. Ello implica que si una variable que en la realidad tiene poca influencia en el resultado final debido (digamos) a su escasa varianza, al transformarla en una variable aleatoria, puede disparar la influencia de una tercera variable con la que ésta pudiera estar virtualmente ligada (con

una función de orden desconocido) y sobre la que no se tiene control en el experimento. Por otra parte, a partir de la medida del ECM no podemos conocer los signos de la relación entre la variables independiente y la que queramos estimar.

La búsqueda de las variables más significativas, se resume en el cuadro siguiente, que muestra los distintos ECM para cada reemplazo. Como se ha dicho antes, los errores mayores podrían indicar una mayor influencia de esa variable en el comportamiento general de la red.

Variable	ECM
Densi91(hab/Km2)	0.3418
Sup_ForKm2	0.3398
Cfa_PC	0.3323
DenCap90	0.2863
AltitMed	0.2101
DenBov90	0.2093
DensLuc(promed)	0.2022
BSk_PC	0.1978
Dfb+Dfc_PC	0.1912
BSh+Csa_PC	0.1899
PobAct90_pc	0.1782
Desem90_pc	0.1769
Decidu_pc	0.1696
SupAgr90_pc	0.1678
Csb+Csc_PC	0.1474
SupFor_pc	0.1410
Agricult_pc	0.1254
Cfb+Cfc_PC	0.1027
DisCar_m	0.1018
DifPobAct_pc	0.0972
DenEmp90	0.0967

Tabla 3. Errores Cuadráticos Medios tras sustitución de variables por valores aleatorios y número de incendios equivalentes

Como se puede ver en la Tabla 3, la sustitución de cualquier variable por valores aleatorios, dispara ostensiblemente el ECM (el valor original es 0.015), incluso en aquellas que aparentan tener una menor significación en el problema. En efecto, el ECM mínimo que se obtiene en este escenario es 0.0967, cifra superior al máximo error convencional de 3 ECM. De todos modos, la tabla nos entrega un indicio de cuáles podrían ser las variables de mayor influencia en el comportamiento general de la red, que corresponden a Densidad de Población 1991 (ECM 0.34), Superficie forestal (ECM 0.34), Tipo de clima Cfa (ECM 0.33) y Densidad de caprinos (ECM 0.29).

5 Conclusiones

Las redes neuronales se presentan como un muy buen método de predicción de incendios forestales de acuerdo a las variables seleccionadas para este estudio. Los grados de precisión obtenidos tras el entrenamiento de la red, son considerablemente

altos, lo que indica que la red puede modelar con éxito las relaciones entre las variables de forma que éstas expliquen los incendios producidos.

Debe considerarse, sin embargo, que en el mejor de los casos, al ser la red neuronal un procedimiento de análisis no lineal, se podría obtener como producto del aprendizaje de ella, valores estimados idénticos a los valores observados. Recuérdese la precisión absoluta en la predicción de ocurrencia/no ocurrencia de incendios, lo que significa que la red fue capaz de encontrar y modelar las diferencias regionales que no puede considerar la regresión logística. Por otra parte, y a diferencia de otros procedimientos estadísticos aquí no se obtiene una ecuación que con cierto grado de probabilidad "explique" el comportamiento general de las observaciones, razón por la cual, insistimos, el error residual de la red podría llegar a ser 0.

En cuanto a ventajas y desventajas de los métodos comparados, red neuronal y regresión logística, podemos decir a favor de la red neuronal que es en teoría imposible de superar en cuanto al nivel de ajuste del modelo con los datos de entrenamiento, que los datos no necesitan asumir ningún tipo de distribución especial y que los ítems de entrada pueden ser una combinación de variables discretas o continuas. En general sus ventajas son las mismas que las redes neuronales ofrecen como características más atractivas. El problema, y he aquí la gran ventaja de la regresión, es que no es posible determinar fiablemente de todas las variables de entrada cuáles son las de mayor importancia y cuáles son prescindibles. La red neuronal debe utilizarse entonces en alguno de los dos casos siguientes:

- Que el usuario tenga la certeza que las variables utilizadas son las relevantes
- Que no sea importante la determinación de qué variables son las mayor importancia en la explicación del fenómeno, sino predecir el fenómeno en sí.

6 Agradecimientos

Este trabajo se ha realizado en el marco del proyecto europeo Megafires (**ENV4-CT96-0256**), financiado por el programa de Medioambiente y Clima de la Comisión europea (DG-XII). También se ha obtenido financiamiento parcial de la CICYT (AGF96-2094-CE).

7 Referencias

Aguado, I. (1997). Utilización de índices meteorológicos e imágenes de satélite NOAA en la previsión del peligro de incendios forestales. Departamento de Geografía. Alcalá de Henares, Universidad de Alcalá de Henares.

Benediktsson, J. A., P. H. Swain, *et al* . (1990). "Neural network approaches versus statistical methods in classification of multisource remote sensing data." IEEE Transactions on Geoscience and Remote Sensing **28**(4): 540-552.

Chou, Y. H., R. A. Minnich, *et al* . (1993). "Mapping probability of fire occurrence in San Jacinto Mountains, California, USA." Environmental Management **17**(1): 129-140.

Civco, D. L. (1993). "Artificial neural networks for land-cover classification and mapping." International Journal of Geographical Information Systems **7**: 173-186.

Evangelia, N. y A. Costas (1997). Post-fire establishment and survival of Aleppo Pine seedlings. Forest fire risk and management. Halkidiki, Greece, Office for Official Publications of the European Comission.

ICONA (1982). Manual de valoración de pérdidas por incendios forestales. Madrid, Instituto para la Conservación de Recursos Naturales.

Jensen, C. (1997). Quiknet, 32 bit Artificial Neural Network software for Windows 95/NT. Kirkland, WA, Craig Jensen.

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

Martell, D. L., S. Otukol, *et al* . (1987). "A logistic model for predicting daily people-caused forest fire occurrence in Ontario." Canadian Journal of Forest Research **17**: 394-401.

Openshaw, S. (1994). Neuroclassification of spatial data. Neural Nets: Applications in Geography. B. Hewitson y R. Crane. Dordrecht, Kluwer Academic Publishers.

Papastavrou, A. (1997). Social, economic, cultural and legislative aspects of forest and wildland fires in Greece. Forest fire risk and management. Halkidiki, Greece, Office for Official Publications of the European Comission.

Salas, F. J. (1994). Detección de áreas de riesgo de incendio forestal a partir de los Sistemas de Información Geográfica y la Teledetección. Departamento de Geografía. Alcalá de Henares, UNiversidad de Alcalá de Henares: 471.

Tárrega, R. y E. Luis (1992). Los incendios forestales en León. León, UNiversidad de León, Secretariado de Publicaciones.

Vega-Garcia, C., P. M. Woodard, *et al* . (1993). Geographic and temporal factors that seem to explain human-caused fire occurrence in Withecourt Forest, Alberta. GIS'93 Symposium, Vancouver, British Columbia.