Serie Geográfica
Vol. 7 1998: 73 - 85 ISSN: 1136-5277

EVALUACIÓN DE LA ESTIMACIÓN DE GRANDES INCENDIOS FORESTALES EN LA CUENCA MEDITERRÁNEA EUROPEA POR REDES NEURONALES Y REGRESIÓN LOGÍSTICA

Luis Carvacho Bart[1]

Pontificia Universidad Católica de Chile, Instituto de Geografía
Av. Vicuña Mackenna 4860, Macul. Santiago de Chile

Resumen. Este artículo muestra algunos de los resultados obtenidos hasta la fecha en el marco de una investigación más amplia relacionada con el proyecto MEGAFiReS. Se comparan métodos de Redes Neuronales Artificiales y Regresión logística para determinar cuál de ellos se presenta como un mejor estimador de la ocurrencia de grandes incendios forestales en la cuenca mediterránea europea. Finalmente se comentan algunas de las ventajas y desventajas observadas en cada método.

Palabras clave: Redes neuronales, regresión logística, logit, incendios forestales, Megafires.

Abstract. *This paper presents some of the results currently obtained in a broader investigation related with the MEGAFiReS project. Artificial Neural Networks and Logistic Regression are compared to determine which one produces better estimations in the prediction of large wildfires in the european Mediterranean basin. Finally some pros and cons of both methods are discussed.*

Key words: *Neural networks, logistic regression, logit, forest fires, Megafires.*

INTRODUCCIÓN

Dentro del gran espectro de instituciones que se dedican a la investigación del problema de los incendios forestales a nivel mundial y continental, en España, la Universidad de Alcalá ha participado y participa activamente en proyectos relacionados con este problema, de forma tal que actualmente a un equipo del Departamento de Geografía de esta institución le cabe la responsabilidad de la coordinación del proyecto europeo MEGAFiReS, *Remote Sensing of large wildland fires in the European Mediterranean Basin*, que persigue mejorar la información cartográfica y estadística considerada crítica en las tres fases del manejo de los incendios: riesgo, detección y combate, y evaluación post-incendio. Este proyecto es el que sirve de marco general al presente trabajo de investigación.

El trabajo se centra, en relación con lo anterior, en la fase de determinación del riesgo de incendio, a través de la evaluación de técnicas que permitan mejorar y evaluar lo más certeramente posible la información estadística disponible, de manera de realizar una certera predicción del riesgo de que en determinado lugar se produzcan incendios forestales. La exactitud de estos pronósticos permitirá eventualmente una focalización más adecuada de los recursos que se relacionan con las otras dos fases ya mencionadas del manejo de los incendios forestales.

[1] Doctorando en el Departamento de Geografía de la Universidad de Alcalá.

OBJETIVO

Determinar, entre redes neuronales y regresión logística, el estadístico más adecuado para estimar la probabilidad de ocurrencia de grandes incendios forestales en algunos países de la cuenca mediterránea europea.

METODOLOGÍA

El objeto de estudio de este trabajo son los "grandes incendios forestales", tal como se definen en el proyecto Megafires, es decir, aquellos que afectan a más de 500 hectáreas. Las razones para escoger allí los grandes incendios como objeto del estudio son básicamente dos: en primer lugar, los grandes incendios son más fácilmente observables por sensores instalados a bordo de satélites que aquellos incendios de menor magnitud, y en segundo lugar, porque los grandes incendios son los más críticos desde una perspectiva ecológica y económica en los países mediterráneos (Chuvieco, 1998).

El área de estudio a tratar corresponde a la cuenca europea mediterránea, específicamente a aquellos países miembros de la Unión Europea (Portugal, España, sur de Francia, Italia y Grecia), salvo algunos territorios insulares. El nivel territorial escogido para cada país es el provincial, con el fin de disponer de información estadística con una base común para cada uno de ellos. En general se han escogido todas las provincias de los países miembros, excepto en el caso francés, del que sólo se han seleccionado aquellas que corresponden a la cuenca del mediterráneo. El total de casos es de 163.

Variables

Las variables utilizadas son un subconjunto de aquellas que se definieron en el proyecto citado para generar su base de datos estadística, un total de 152 variables, de las que se escogieron 26, eliminando variables colineales, que no cumplieran con el requisito de normalidad en su distribución o que no aportaran información adicional.

Las variables que finalmente se seleccionaron para este estudio pertenecen a tres grandes categorías: agrícolas, naturales y socioeconómicas. La tabla 1 muestra los códigos mnemónicos asignados a las variables y su descripción.

El período a estudiar está comprendido entre los años 1991 y 1995, por lo que el cálculo de las

Tabla 1. Códigos asignados a las variables y descripción de las mismas

Variable	Descripción
Agricult_pc	Porcentaje de la provincia con tipo de combustible 5 (*)
Arren90_pc	Porcentaje de arrendatarios agrícolas en 1990
DenBov90	Densidad de la masa bovina en 1990
DenCap90	Densidad de la masa caprina en 1990
DenEmp90	Densidad de población empleada en 1990
DenOvi90	Densidad de la masa ovina en 1990
DifArren_pc	Diferencia porcentual de arrendatarios agrícolas 1960/1990
DifEmp_pc	Diferencia porcentual de población empleada 1960/1990
DifSupAgr_pc	Diferencia porcentual en la superficie agrícola provincial 1960/1990
DifTamExp_pc	Diferencia porcentual del tamaño de las explotaciones agrícolas 1960/1990
SupAgr90	Superficie de la provincia destinada a la agricultura
AltitMed	Altitud media
Bsh+Csa_PC	Porcentaje de la provincia bajo clima Bsh o Csa
Bsk_PC	Porcentaje de la provincia bajo clima Bsk
Csb+Csc_PC	Porcentaje de la provincia bajo clima Csb o Csc
Decidu_pc	Porcentaje de bosque con hoja caduca
Densi91(hab/Km2)	Densidad de población en 1991
DensLuc(promed)	Promedio de densidad de luces
Desem90_pc	Porcentaje de población desempleada en 1990
DIF_IE	Diferencia porcentual de índice de ancianidad 1960/1990
DIF_IJ	Diferencia porcentual de índice de juventud 1960/1990
DifPobAct_pc	Diferencia porcentual de población activa 1960/1990
DifPobSec_pc	Diferencia porcentual de población en el sector secundario 1960/1990
DifPobTer_pc	Diferencia porcentual de población en el sector terciario 1960/1990
DisCar_m	Distancia a las carreteras en metros
PobAct90_pc	Porcentaje de población económicamente activa en 1990

(*) Se define como tipo de combustible 5 las cubiertas de cultivos de secano, cultivos de regadío permanentes, arrozales, viñedos, frutales en regadío y olivares

redes neuronales y de las regresiones se basa en la ocurrencia o no ocurrencia de grandes incendios forestales por provincia entre los años 1991 y 1995. La variable dependiente es, por tanto, la observación o no observación de al menos un gran incendio forestal en el período estudiado.

Una vez escogido el grupo de variables a analizar, se procesan mediante los dos métodos mencionados, redes neuronales y regresión logística. Para comparar su eficacia como estimadores, se ha obtenido con cada uno de ellos la estimación de ocurrencia o no ocurrencia de incendios, es decir, la probabilidad de que se produzca al menos un incendio en un área específica en el período 1991-1995.

Modelos

Redes neuronales

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). Esencialmente estos modelos se han venido empleando como uno de los métodos de clasificación digital de imágenes de satélite (Clark y Cañas 1995; Foody, 1996), o como clasificadores espaciales, en general, originando el nuevo concepto de "neuroclasificación de datos espaciales" (Openshaw, 1994). En el caso específico de estimación de incendios forestales, esta técnica ha sido utilizada previamente por Vega-García et al.(1993)

Con respecto al concepto de red neuronal, se puede decir que son sistemas de trazado no lineal con una estructura basada ligeramente en los principios observados en los sistemas nerviosos biológicos, y pueden ser utilizadas para aprender la relación que existe entre un conjunto de datos de entrada y otro de salida. Todo lo que se requiere entonces para entrenar una red neuronal, es un conjunto de datos que contengan la relación entrada/salida. Una vez que la red ha conseguido determinar un modelo que sea capaz de relacionar los datos de entrada y que explica la distribución de los datos de salida, ese entrenamiento puede ser aplicado sobre otros valores de entrada para estimar, esta vez, valores de salida desconocidos.

De lo anterior, podemos deducir que el fundamento de las redes neuronales se encuentra en:

- La existencia de una relación entre los datos de entrada y de salida
- La capacidad de entrenamiento o de "aprendizaje" de la red
- 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.

De aquí se desprende que la gran utilidad de esta técnica sea su intento de replicar la habilidad del cerebro de "aprender por el ejemplo", de generalizar de lo específico a lo abstracto y de manejar datos "ruidosos" (Openshaw y Openshaw, 1997).

Entre las ventajas de las redes neuronales como elemento de análisis, es que a diferencia de otros procedimientos estadísticos, no está constreñida a la existencia de una distribución normal de los datos de entrada, y de hecho, de ningún tipo de distribución. (Benediktsson et al., 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 funcionamiento de las redes, tal como se ha anticipado en las definiciones precedentes, se basa en unidades procesadoras independientes, llamadas "nodos" o "neuronas", agrupados en capas. Estos nodos reciben información ya procesada por aquellos de otras capas y envían sus propios resultados a los nodos de la capa siguiente a través de canales o enlaces con un "peso" propio. Esquemáticamente una red neuronal se puede representar de la manera que se muestra en la figura 1.

La estructura o topología de la red que se muestra, corresponde a una red de cuatro capas o de topología 2 5 3 1: una capa de entrada con dos neuronas, dos capas ocultas de cinco y 3 neuronas respectivamente, y una capa de salida. El concepto de "capa oculta" ("*hidden layer*") queda mejor definido como "capa intermedia", ya que en realidad corresponde a un estrato de neuronas que no es de entrada ni es de salida.

Existen diversas maneras en que una red neuronal calcula los pesos de los enlaces entre las neuronas. Estos métodos o "algoritmos" tienen diversa eficacia según el tipo de datos de entrada que se procesen, pero en general la regla que se utiliza para escoger un algoritmo u otro es "*si funciona, ese es el correcto*". Entre los algoritmos más habituales está *Backprop* y sus derivados, RPROP y *Quickprop*. La discusión de las características de cada uno de ellos escapa al objeti-

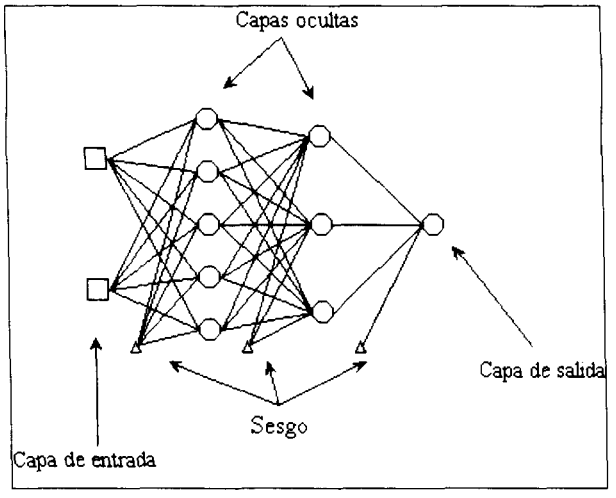

***Figura 1**. Esquema de una red neuronal. Además de la topología de las capas,se aprecia el sesgo de cada capa, un valor auxiliar para mejorar el entrenamiento*

vo de este artículo, pero una buena fuente de información es Serle (1997).

Regresión logística

El modelo de regresión logística 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-Garcia et al., 1993).

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. 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 un ajuste de máxima probabilidad:

$$z = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

donde ***a*** es la constante y b_n el factor de ponderación de la variable ***n***. Los valores ***z*** se pueden interpretar como la probabilidad de ocurrencia del fenómeno. 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 obtener muestras de entrenamiento de la red o de cálculo de la ecuación de regresión logística, así como de muestras de control para ambos métodos, fueron seleccionados aleatoriamente el 60% de los casos (98 provincias) con los que se obtuvieron entrenamiento de la red y la ecuación de regresión. Luego, los casos restantes, el 40% aquel que no se incluyó en los cálculos previos (65 provincias), fueron utilizados para medir la calidad de las estimaciones de ambos métodos.

La figura 2 ilustra la distribución de las muestras de entrenamiento y control para toda el área estudiada.

RESULTADOS

Redes neuronales

Entrenamiento

Utilizando los valores correspondientes de las 26 variables en uso hasta el momento en los 98 casos del grupo de entrenamiento, se ha obtenido una red entrenada satisfactoriamente por medio de una topología 1, 3, 1 con el algoritmo llamado "*quickprop*". La bondad de las estimaciones para este grupo de datos se refleja en la tabla 2 en términos del error cuadrático medio (ECM) de las estimaciones.

Como se puede ver, el entrenamiento consiguió encontrar un patrón de comportamiento entre las variables que consigue estimar correctamente al 100% de los casos. El error cuadrático medio máximo encontrado, indica que el mayor error de acierto para los casos de ocurrencia y no ocurrencia, sería de 1-0.29 y 0+0.29 respectivamente. Como el umbral para decidir si una estimación indica ocurrencia (1) o no ocurrencia (0) es de 0.5, puede decirse que la precisión alcanzada es muy buena.

Control

Corresponde en este punto determinar el grado de acierto que consigue la red recién entrenada utilizando esta vez, como datos de prueba, el 40% de los casos que no se incluyeron en el entrenamiento. La tabla 3 muestra los resultados obtenidos tras el experimento, tanto en la precisión general como en las distintas combinaciones de acierto entre estimados/observados y no estimados/no observados, cuyo resumen se muestra en la tabla de contingencia respectiva (Tabla 4). Es necesario comentar aquí, respecto a la tabla 3, que debido al proceso de escalamiento de los

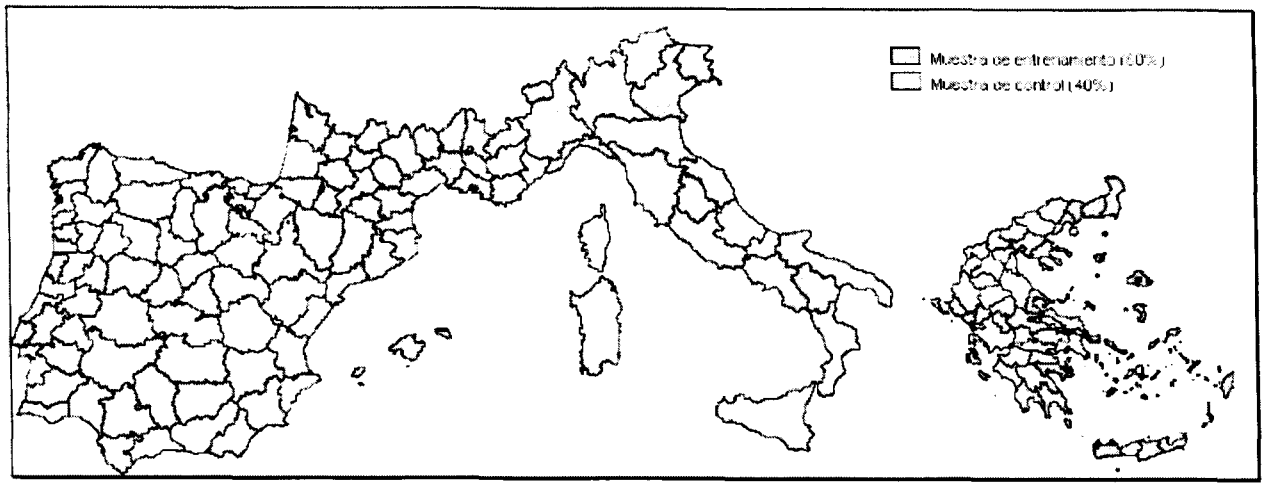

***Figura 2**. Distribución de las muestras de entrenamiento y control*

datos que se suele realizar antes de entrenar una red, es que se asignan valores de 0,9 para "observado" (en lugar de 1) y 0,1 para "no observado" (en lugar de 0). Los valores estimados se representan, eso sí, en el rango 0 - 1.

Como se puede ver en el cuadro, el acierto general es bastante bueno, un 69,23%, aunque el error de omisión (es decir, donde no se estimaron incendios, pero se produjeron) cercano al 41% de los casos, puede considerarse relativamente alto.

El mapa de la figura 3 muestra la distribución de estas estimaciones para las 65 provincias utilizadas como caso de estudio en el experimento ya comentado.

Como se puede apreciar, los mayores aciertos se logran en la Península Ibérica, con una buena relación de incendios estimados respecto a los observados. Estos casos, sumados a aquellos que representan incendios correctamente no estimados hacen de esta zona la de mejor pronóstico. Las provincias francesas también presentan una buena relación de aciertos, lo mismo que Italia, aunque en este último caso no parece demasiado prudente esbozar una conclusión dado el insuficiente número de casos que representan a este país. En el caso de Grecia, los errores y los aciertos están relativamente equilibrados, aunque los primeros superan ligeramente a los segundos.

***Tabla 2**. Indicadores de la precisión de la estimación de la red neuronal, casos de entrenamiento*

Indicador	Valor
ECM general	0.059
ECM máximo	0.292
Casos correctos	98
% correcto	100

Búsqueda de variables significativas.

Aunque las redes neuronales no buscan determinar un grupo de variables significativas, a diferencia de los modelos de regresión, puede ser posible mediante análisis de sensibilidad buscar a qué variables la red ha asignado mayor importancia a la hora de encontrar el entrenamiento adecuado que explica el comportamiento general de las variables. Para comprender el procedimiento que se ha seguido, es preciso entender el grupo de variables de entrenamiento como una matriz, en la que cada variable representa un vector determinado. Se reemplazan entonces los valores de uno de los vectores por una constante (cero) y se procesa a continuación esa nueva matriz utilizando el entrenamiento que la red consiguió con los valores originales. Se obtiene el error cuadrático medio de este procesamiento, se restauran los valores originales del vector, y se repite el procedimiento hasta realizarlo con todas las variables. El esquema de la figura 4 representa este proceso.

El resultado obtenido tras realizar los reemplazos respectivos para las 26 variables en estudio, arrojó los Errores Medios Cuadráticos que se muestran en la tabla 5.

De acuerdo a la tabla, entonces, las variables que más se ven afectadas por el cambio de sus valores, y por consiguiente aquellas a las que mayor peso asignó la red en la fase de entrenamiento son: áreas bajo clima Csb o Csc, distancia a carreteras, áreas bajo clima Bsh o Csa, población activa, superficie agrícola, diferencia de población en actividades terciarias y densidad de población en 1991. El gráfico de la figura 5 explica la razón por la cual se han escogido las siete primeras variables como las significativas en este caso. Como se ve allí, se aprecia un claro quie-

Tabla 3. Acierto/error en las estimaciones de ocurrencia de incendios mediante la red neuronal

	Valor		Acierto			Error	
Provincia	Estimado	Observado	General	Estimado/ observado	No estimado/ no observado	Estimado/no observado	No estimado/ observado
Bragança	0.846	0.9	1	1			
Coimbra	0.814	0.9	1	1			
Leiria	0.827	0.9	1	1			
Portalegre	0.000	0.1	1		1		
Setúbal	0.061	0.1	1		1		
Vila Real	0.850	0.9	1	1			
Viseu	0.910	0.9	1	1			
Alicante	0.916	0.9	1	1			
Castellón	0.916	0.9	1	1			
Cuenca	0.915	0.9	1	1			
Gerona	0.914	0.9	1	1			
Granada	0.915	0.9	1	1			
Logroño	0.915	0.1	0			1	
Lugo	0.902	0.9	1	1			
Málaga	0.510	0.9	1	1			
Navarra	0.855	0.9	1	1			
Orense	0.893	0.9	1	1			
Oviedo	0.028	0.9	0				1
Palencia	0.915	0.1	0			1	
Pontevedra	0.507	0.9	1	1			
Segovia	0.913	0.1	0			1	
Soria	0.915	0.1	0			1	
Teruel	0.916	0.9	1	1			
Toledo	0.912	0.9	1	1			
Vizcaya	0.103	0.1	1		1		
Zamora	0.911	0.9	1	1			
Alpes de Haute	0.000	0.1	1		1		
Hautes Alpes	0.000	0.1	1		1		
Alpes Maritimes	0.703	0.1	0			1	
Corse	0.636	0.9	1	1			
Haute-Garonne	0.003	0.1	1		1		
Hérault	0.899	0.1	0			1	
Landes	0.000	0.1	1		1		
Tarn-et-Garonne	0.000	0.1	1		1		
Var	0.831	0.9	1	1			
Vaucluse	0.025	0.9	0				1
Abruzzo	0.196	0.1	1		1		
Lombardia	0.004	0.9	0				1
Molise	0.070	0.1	1		1		
Trentino A.A.	0.061	0.1	1		1		
Etoloakarnania	0.587	0.9	1	1			
Arkadia	0.896	0.9	1	1			
Attiki	0.242	0.9	0				1
Viotia	0.841	0.9	1	1			
Grevena	0.000	0.9	0				1
Drama	0.000	0.9	0				1
Ilia	0.622	0.1	0			1	
Imathia	0.041	0.9	0				1
Ioannina	0.009	0.9	0				1
Karditsa	0.078	0.9	0				1
Kastoria	0.180	0.1	1		1		
Kerkyra	0.103	0.1	1		1		
Kilkis	0.000	0.9	0				1
Kozani	0.693	0.9	1	1			
Kyklades	0.897	0.1	0			1	
Larissa	0.797	0.9	1	1			
Lassithi	0.877	0.9	1	1			
Lefkada	0.495	0.1	1		1		
Pieria	0.059	0.1	1		1		
Rethymno	0.723	0.9	1	1			
Samos	0.793	0.9	1	1			
Serres	0.000	0.9	0				1
Trikala	0.225	0.9	0				1
Florina	0.124	0.1	1		1		
Fokida	0.002	0.1	1		1		

***Tabla 4**. Tabla de contingencia acierto/error en la estimación de incendios mediante la red neuronal*

	Observado		
Estimado	0	1	Acierto
0	17	12	58.62%
1	8	28	77.78%
		General	69.23%

bre en la tendencia del ECM a partir de la séptima variable, Densi91.

El problema que emerge en este punto, sin embargo, es la imposibilidad de determinar a través de un procedimiento "cuadrático" como el ECM, la tendencia del error obtenido, es decir si éste influye de manera directa o inversa en el comportamiento global de las variables. El valor de error que se obtiene, no puede interpretarse como "positivo" o "negativo". Recordemos que una regresión, por ejemplo, entrega el peso de cada variable en la ecuación en conjunto con el signo que le corresponde, es decir su "sentido" con respecto al resultado estimado. En el caso del análisis por redes neuronales, este sentido se puede desentrañar indirectamente mediante un nuevo análisis de sensibilidad, básicamente una variación del efectuado para encontrar las variables significativas. Se trata también, por tanto, de un reemplazo iterativo de los valores de las variables por constantes. El concepto que se intenta aplicar es el siguiente: se reemplazan los valores de una de las variables de la matriz por una constante baja (digamos 0,1) y se somete a esta nueva matriz al entrenamiento ya realizado, obteniendo el ECM de dicho cálculo. Luego, la misma variable se vuelve a modificar, esta vez con una constante alta (0,9) y se repite el proceso con lo que se obtiene un nuevo valor de ECM. Entonces, si con el primer reemplazo, aquel con valores bajos, se obtiene un ECM mayor que con el segundo reemplazo, de valores altos, la relación entre esa variable y las estimaciones, ha de ser inversa; y análogamente, ha de ser directa si con el reemplazo por valores bajos se obtiene un menor ECM que con el reemplazo por valores altos. El esquema de la figura 6 muestra gráficamente el proceso del análisis de sensibilidad aplicado a dos variables de un conjunto ficticio de cinco.

Los resultados del análisis antedicho, se muestran en la tabla 6. De acuerdo a ella, cuatro variables se presentan con una tendencia inversa con respecto a la existencia de incendios forestales en el área, es decir que a medida que el valor de dichas variables aumenta, los incendios tienden a no producirse, y viceversa. Estas variables de tendencia inversa fueron: superficie agrícola en 1990, población activa en 1990, área provincial bajo climas Csb o Csc y densidad de población en 1991. Como se ve, algunas de estas tendencias parecen lógicas, como la variable climática (a mayor humedad, menor riesgo de incendio) o la población activa (a menor población activa más riesgo de incendio), aunque la tendencia de la superficie agrícola no parece coherente ya que se espera en ese caso una relación directa entre esa variable y los incendios. Las tres variables que presentan una relación directa son área provincial bajo climas Bsh o Csa, distancia a carreteras y diferencia de población en el sector terciario entre 1960 y 1990. Nuevamente la variable climática se presenta con una tendencia

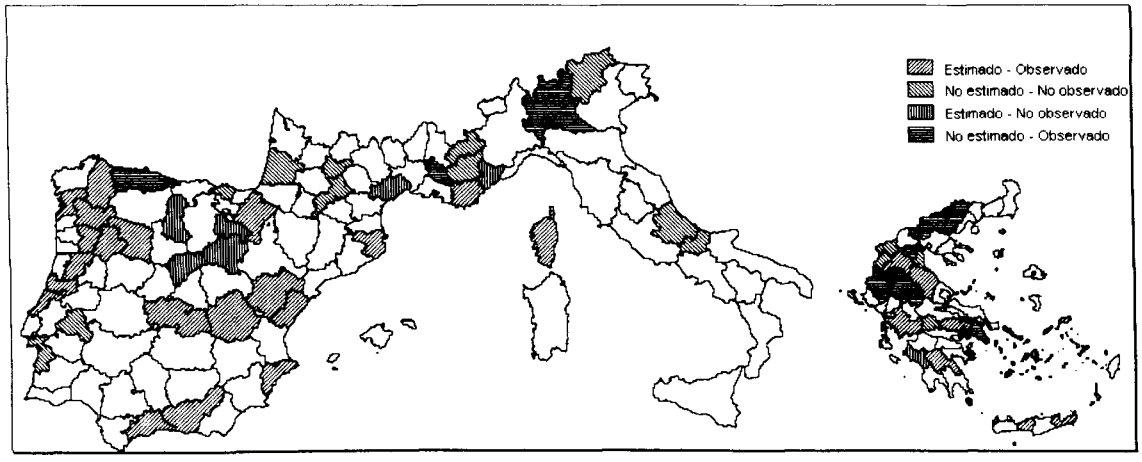

***Figura 3**. Distribución de acierto/error en la estimación de ocurrencia/no ocurrencia de incendios forestales para 65 provincias seleccionadas usando redes neuronales.*

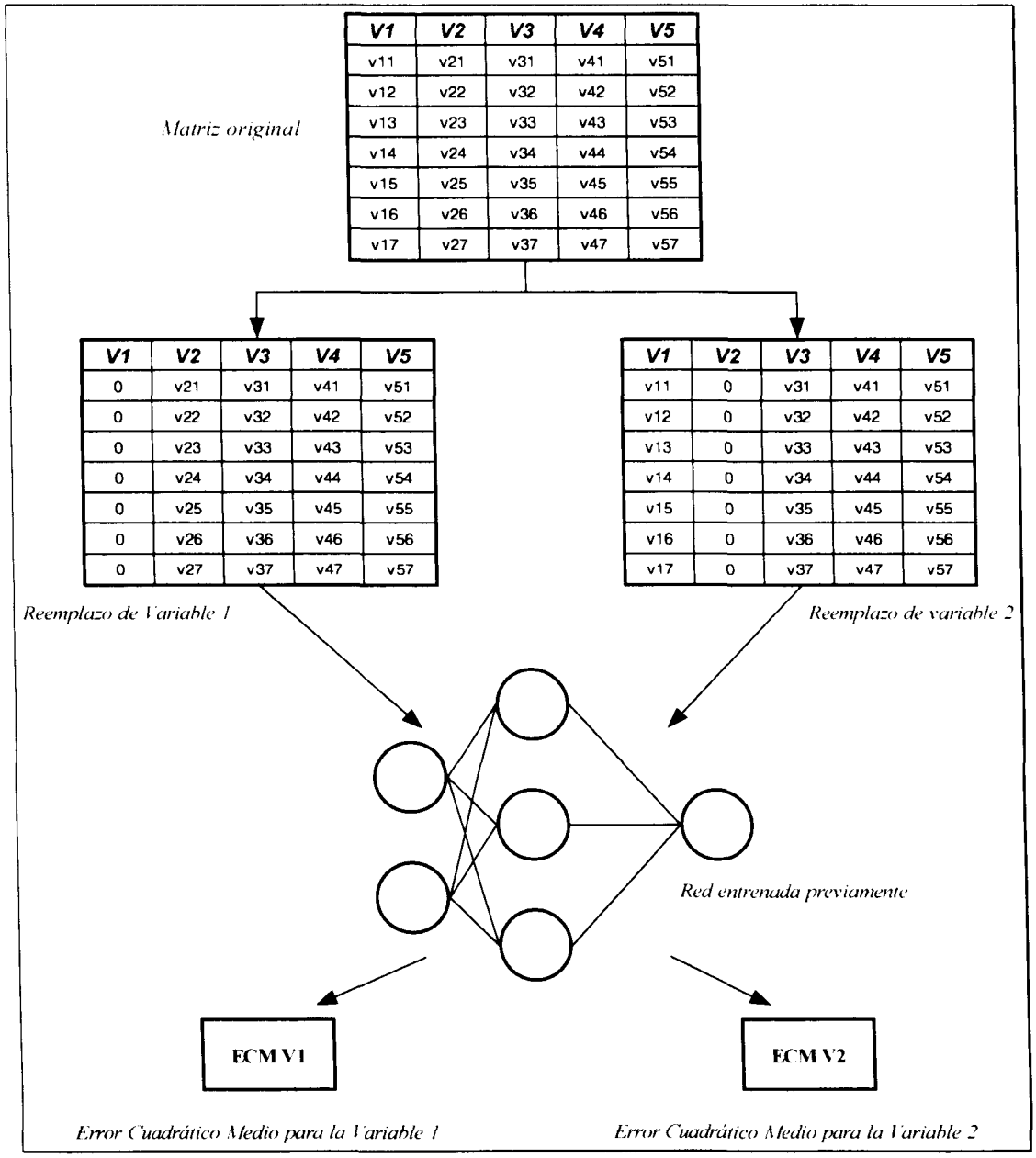

***Figura 4**. Esquema del proceso para buscar variables significativas con una red neuronal. Se muestra sólo para dos variables*

lógica, pues se espera que a mayor sequedad, la posibilidad de generación de incendios sea también mayor, aunque la variable distancia a carreteras presenta una tendencia opuesta a la esperada, pues las probabilidades de generación de incendios son mayores justamente en las proximidades de estas vías.

Regresión logística

Cálculo

Al igual que en el caso de la red neuronal, se ha utilizado una muestra aleatoria de un 60% de los casos para buscar la ecuación de regresión logística. Este grupo de casos es el mismo que se empleó en el análisis anterior.

Tras el cálculo correspondiente, la regresión logística obtenida por *SPSS*, nos ha entregado la tabla de contingencia que se observa en la tabla 7.

Las variables seleccionadas por la regresión son: superficie provincial bajo climas Bsh o Csa, diferencia de población en actividades del sector terciario entre 1960 y 1990, distancia a carreteras, porcentaje de población activa en 1990 y su-

***Tabla 5**. Valores de ECM obtenidos tras reemplazo de variables*

Variable	ECM
Csb_Csc	0.621026
Discar_m	0.603777
Bsh-Csa	0.599430
Pobact90	0.589695
Supagr90	0.583415
Difpobte	0.581250
Densi91	0.580525
Decidu_pc	0.531710
Denbov	0.527659
Bsk	0.518433
Desem90	0.506076
Diftamex	0.502487
Denovi	0.494765
Densluc	0.485160
Dif_ie	0.484124
Denenmp90	0.481625
Difpobact	0.477964
altimed	0.472521
agricult	0.466352
Arren90	0.463694
Difemp	0.461603
Difpobsec	0.458196
Difarren	0.457975
Dencap	0.457611
Difsupagr	0.448946
Dif_ij	0.442221

perficie agrícola en 1990. El porcentaje de precisión general es bueno, situándose en un 76,53%, en tanto que el error de omisión resulta sumamente bajo, con sólo 8 casos errados sobre 59, es decir un 13,55% de las estimaciones de ocurrencia de incendios.

La ecuación de regresión logística ha sido:

z = BSH_CSA_ × 0,0059 + DIFPOBTE × -0,701 + DISCAR_M × -0,004 + POBACT90 × -0,1994 + SUPAGR90 × 1,59E-06 + 9,0132

En este caso, el signo de los coeficientes de las variables seleccionadas por el procedimiento de regresión es coherente respecto a lo esperado para un fenómeno como el de los incendios forestales. Véase como la variable distancia a carreteras presenta esta vez un coeficiente negativo, más lógico con el problema de los incendios, como ya se discutió previamente. Análogamente, el signo de la variable de la superficie agrícola también es lógico, ya que se esperan más incendios donde hay más actividad agraria, lo que es consecuente con el coeficiente negativo que muestra la diferencia de población terciaria entre 1960 y 1990, pues ello implica que mientras más personas trabajen en actividades netamente urbanas, menos riesgo de incendio forestal existe.

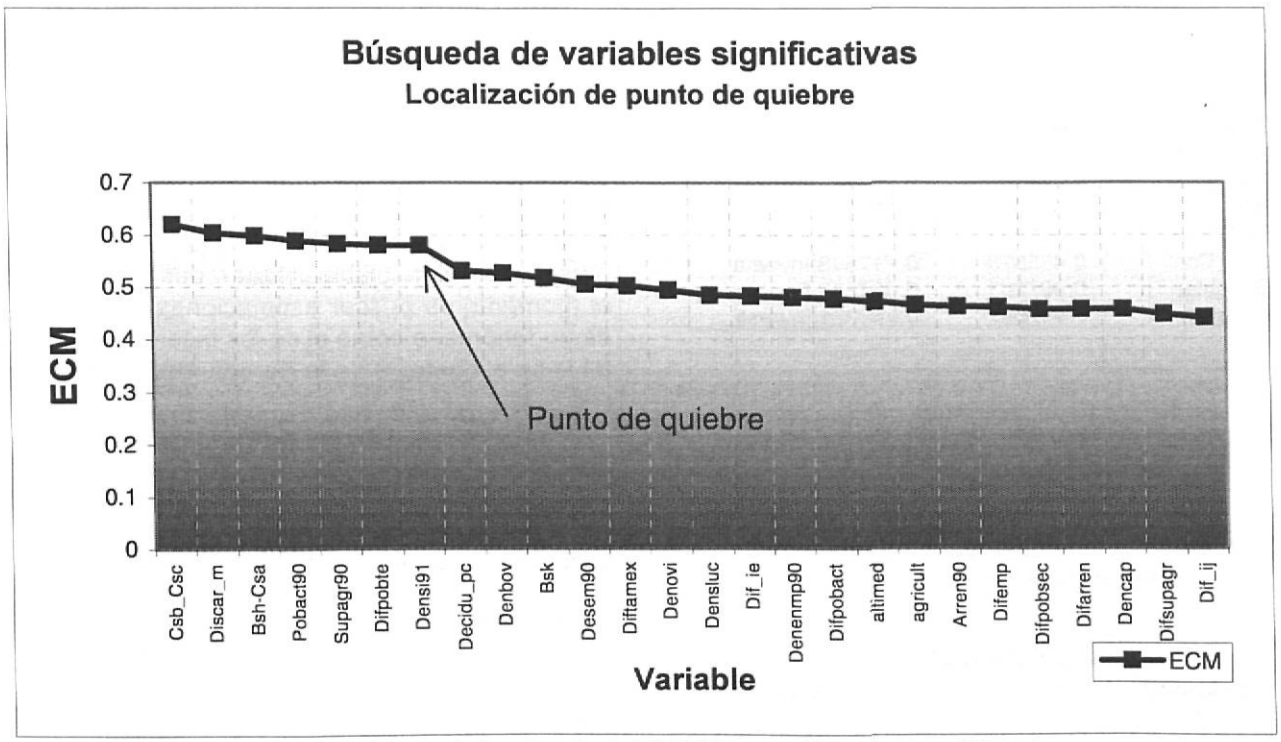

***Figura 5**. Línea de tendencia del ECM tras análisis de sensibilidad. Se ha identificado el primer punto de quiebre de la tendencia.*

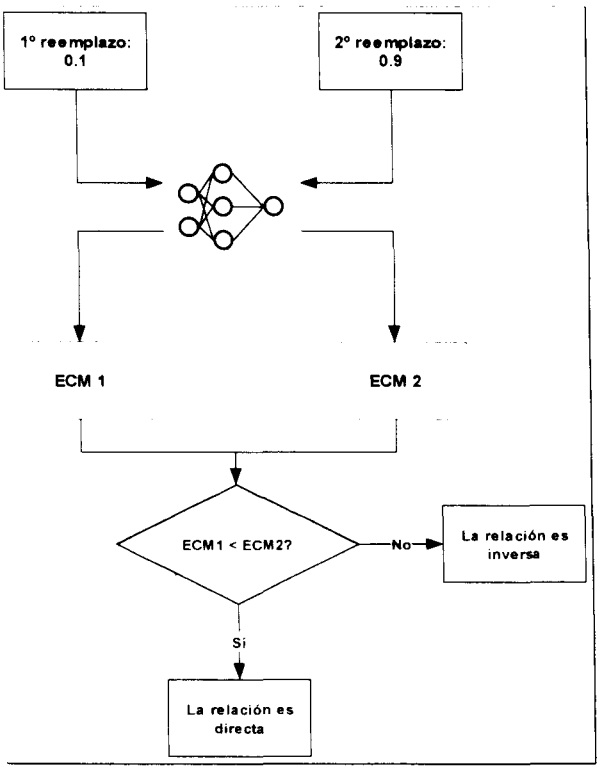

***Figura 6**. Esquema del análisis de sensibilidad para determinar la tendencia de las variables respecto a las estimaciones de la red neuronal*

***Tabla 6**. Valores de ECM obtenidos para la búsqueda de la tendencia de las variables significativas*

Variable	Reemplazo	Reemplazo	Tendencia
SupAgr90	0.449591	0.500538	Inversa
PobAct90	0.465701	0.537182	Inversa
Bsh_Csa	0.544912	0.500717	Directa
DisCar_m	0.473033	0.464068	Directa
Csb_Csc	0.455076	0.717863	Inversa
Difpobter	0.487203	0.437670	Directa
Densi91	0.447597	0.468021	Inversa

***Tabla 7**. Tabla de contingencia de la regresión logística sobre el 6M% de los casos.*

	Observado		
Estimado	0	1	Incorrecto
0	24	8	61.54%
1	15	51	86.44%
		General	76.53%

Control

La ecuación recientemente obtenida, se utilizará en este punto para probar la bondad de su ajuste utilizando el mismo procedimiento que para la red neuronal, es decir, sometiendo a ésta al 40% de los casos no utilizados en su cálculo. De este experimento obtenemos la tabla 8, que muestra la relación acierto/error para cada uno de los casos de prueba. La tabla de contingencia respectiva, puede verse en tanto, en la tabla 9.

Del cuadro anterior, se puede desprender que el acierto general es aceptable, con un 60% de precisión, aunque el error de omisión, incendios no estimados en provincias donde sí se observaron, es superior al 51%.

La distribución de estas estimaciones para las 65 provincias de prueba, se muestra en el mapa de la figura 7.

CONCLUSIONES

De acuerdo a lo anterior, las redes neuronales se han mostrado como un estimador más robusto que la regresión logística en el caso de estudio. Tanto en la bondad de las estimaciones con los datos de entrenamiento (60% de las observaciones) como con los datos de prueba (40% de las observaciones), se evidenció la mayor exactitud de las redes neuronales. Con los datos de entrenamiento, recordemos, la precisión fue del 100%, en tanto que en la regresión logística llegó sólo al 76,5%. Con los datos de prueba, la diferencia en la precisión entre ambos métodos es menor que con los anteriores, pero siempre con un sesgo favorable hacia las estimaciones de la red neuronal. Desde este punto de vista, parece más aconsejable utilizar redes neuronales al momento de buscar estimaciones confiables de un fenómeno como el de los incendios forestales en el contexto de la Europa mediterránea.

El problema con las redes neuronales estriba en que cuando además de precisar un buen estimador, se requiere obtener información adicional, como qué variables son las más determinantes como estimadoras, o conocer el sentido en que esas variables influyen en la predicción, es decir, si su relación con las estimaciones son directas o inversas. Los métodos desarrollados para intentar extraer esta información de la "caja negra" que es el entrenamiento logrado por la red, son muy laboriosos y como se observó en los resultados obtenidos aquí, poco confiables. Reflexionando además sobre los principios que operan

***Tabla 8**. Acierto/error en las estimaciones de ocurrencia de incendios mediante regresión logística*

	Valor		Acierto		Error	
Provincia	Estimado	Observado	General	Estimado/ observado	No estimado/ no observado	Estimado/no observado	No estimado/ observado
Braganca	0.65	1	1	1			
Coimbra	0.59	1	1	1			
Leiria	0.73	1	1	1			
Portalegre	0.47	0	1		1		
Setúbal	0.14	0	1		1		
Vila Real	0.86	1	1	1			
Viseu	0.17	1	0				1
Alicante	1.00	1	1	1			
Castellón	0.99	1	1	1			
Cuenca	0.99	1	1	1			
Gerona	0.97	1	1	1			
Granada	0.66	1	1	1			
Logroño	0.99	0	0			1	
Lugo	0.92	1	1	1			
Málaga	0.73	1	1	1			
Navarra	0.90	1	1	1			
Orense	0.93	1	1	1			
Oviedo	0.32	1	0				1
Palencia	0.99	0	0			1	
Pontevedra	0.92	1	1	1			
Segovia	0.99	0	0			1	
Soria	0.96	0	0			1	
Teruel	1.00	1	1	1			
Toledo	0.99	1	1	1			
Vizcaya	0.25	0	1		1		
Zamora	0.97	1	1	1			
Alpes de Haute	0.03	0	1		1		
Hautes Alpes	0.00	0	1		1		
Alpes Maritimes	1.00	0	0			1	
Corse	0.71	1	1	1			
Haute-Garonne	0.00	0	1		1		
Hérault	0.39	0	1		1		
Landes	0.00	0	1		1		
Tarn-et-	0.00	0	1		1		
Var	0.97	1	1	1			
Vaucluse	0.12	1	0				1
Abruzzo	0.53	0	0			1	
Lombardia	0.86	1	1	1			
Molise	0.12	0	1		1		
Trentino A.A.	0.30	0	1		1		
Etoloakarnania	0.14	1	0				1
Arkadia	0.51	1	1	1			
Attiki	1.00	1	1	1			
Viotia	0.62	1	1	1			
Grevena	0.02	1	0				1
Drama	0.03	1	0				1
Ilia	0.96	0	0			1	
Imathia	0.07	1	0				1
Ioannina	0.34	1	0				1
Karditsa	0.02	1	0				1
Kastoria	0.43	0	1		1		
Kerkvra	0.89	0	0			1	
Kilkis	0.09	1	0				1
Kozani	0.52	1	1	1			
Kvklades	1.00	0	0			1	
Larissa	0.43	1	0				1
Lassithi	0.04	1	0				1
Lefkada	0.98	0	0			1	
Pieria	0.10	0	1		1		
Rethvmno	0.45	1	0				1
Samos	1.00	1	1	1			
Serres	0.09	1	0				1
Trikala	0.08	1	0				1
Florina	0.97	0	0			1	
Fokida	0.11	0	1		1		

***Tabla 9**. Tabla de contingencia acierto/error en la estimación de incendios mediante regresión logística*

	Observado		
Estimado	0	1	Acierto
0	14	15	48.28%
1	11	25	69.44%
		General	60.00%

sobre el funcionamiento de las redes neuronales, estos procedimientos son discutibles, ya que con los análisis de sensibilidad efectuados, estamos asumiendo que la relación existente entre las variables entrenadas es lineal, pues esperamos un comportamiento lineal de los resultados con respecto a los cambios ingresados en las variables para probar la reacción de la red. En este punto debemos recordar que al usar redes neuronales, estamos asumiendo que es muy probable que exista una relación no lineal entre las variables en estudio, lo que parece confirmarse en este caso por el hecho de obtener mejores estimaciones con este método que con la regresión.

La técnica de la regresión logística entregó en este caso resultados aceptables, por lo que su uso parece indicado si es necesario conocer la información explicativa que las redes neuronales no entregan.

Es el investigador, por tanto, el que de acuerdo a sus necesidades deberá optar por obtener mejores estimaciones o mejores explicaciones. Como en todo orden de cosas, nada es perfecto.

AGRADECIMIENTOS

El autor desea dejar constancia de su expreso agradecimiento a todos los investigadores del proyecto MEGAFiReS por su apoyo e interés constante en el desarrollo de la investigación en que se encuentra inmerso, y del que este artículo es parte; en particular al coordinador del proyecto, Dr. Emilio Chuvieco y al Dr. José Ignacio Barredo, que en su momento sentó algunas de las premisas que ha seguido esta investigación.

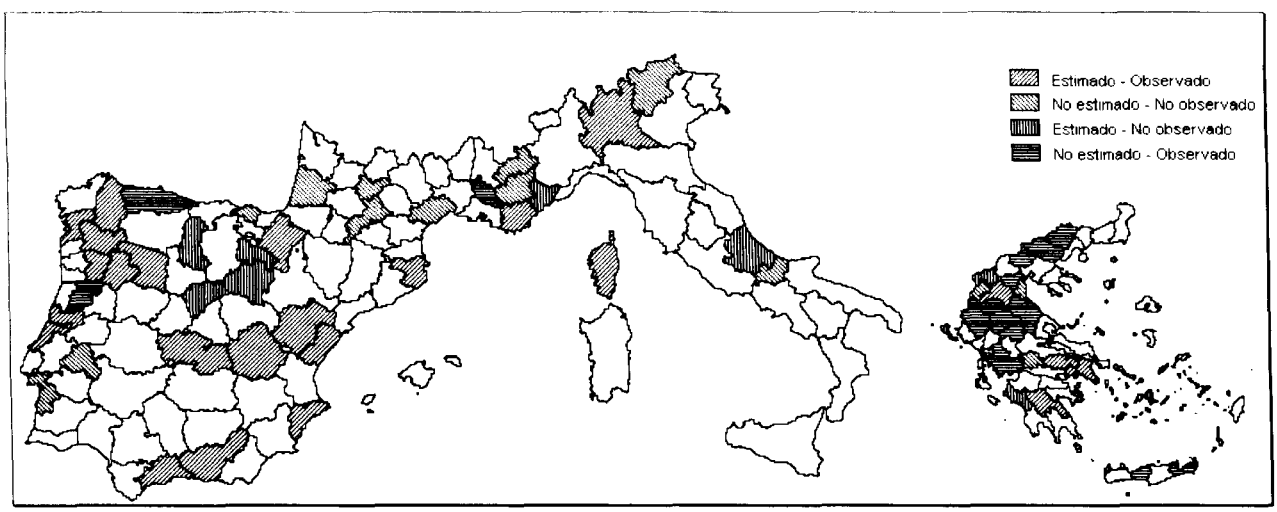

***Figura 7**. Distribución de acierto/error en la estimación de ocurrencia/no ocurrencia de incendios para 65 provincias seleccionadas usando regresión logística*

REFERENCIAS

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

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

Chuvieco, E. (1998). *Megafires Project, final report.* Alcalá de Henares.

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

Clark, C. y Cañas, A. (1995). Spectral identification by artificial neural network and genetic algorithm. *International Journal of Remote Sensing* **16**: 2256-2272.

Foody, G., M (1996). Relating the Land Cover Composition of Mixing Pixels to Artificial Neural Network Classification Output. *Photogrammetric Engineering and Remote Sensing* **62**: 491-499.

Loftsgaarden, D. y Andrews, P. L. (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., Otukol, S. y Stocks, B.J. (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 and R. Crane. Dordrecht, Kluwer Academic Publishers.

Openshaw, S. y Openshaw, C. (1997). *Artificial Intelligence in Geography*. Chichester, John Wiley & Sons.

Serle, W. S. (1997). Neural Network FAQ. Periodic posting to the USENET newsgroup comp.ai.neural-nets; URL: ftp://ftp.sas.com/pub/neural/faq.htm.

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