

EMPLEO DE REGRESIÓN LOGÍSTICA PARA LA OBTENCIÓN DE MODELOS DE RIESGO HUMANO DE INCENDIOS FORESTALES

VILAR DEL HOYO, L.

Instituto de Economía y Geografía. CSIC
c/Pinar 25. 28006, Madrid
lvilar@ieg.csic.es

RESUMEN

La simulación y modelización de fenómenos geoespaciales permiten entender el fenómeno observado, comprobar hipótesis y teorías, predecir su comportamiento en el espacio y en el tiempo bajo distintas condiciones y escenarios así como descubrir nuevos funcionamientos y comportamientos. Esta comunicación presenta el empleo de Regresión Logística para generar un modelo de riesgo de ignición de incendios forestales. Se incluyen en el modelo los factores relativos al riesgo producido por las actividades humanas y su relación con la ocurrencia de incendios debidos a esta causa en la Comunidad de Madrid, en el período 1990-2004 (resolución 1km²). Los factores vinculados a la actividad humana tienen especial relevancia aunque en la cuantificación del riesgo existe una relativa dificultad en valorarlos y espacializarlos de cara a obtener modelos que permitan predecir el inicio y la propagación del fuego. No obstante, la identificación de los distintos tipos de factores de riesgo vinculados a la actividad humana es posible generar variables de riesgo, relacionadas con los usos del territorio (vías de comunicación, vertederos, espacios naturales protegidos, etc.), aspectos socioeconómicos (población ocupada en la agricultura, nivel de renta, tasa de paro, etc.) y otras como zonas de contacto con el uso forestal (interfases) o medios de vigilancia de incendios forestales, que permitan obtener modelos predictivos suficientemente fiables. Estos modelos pueden ser de gran interés para la planificación y establecimiento de estrategias de prevención y gestión de los espacios forestales.

Palabras clave

Actividades humanas, Incendio forestal, Modelización, Regresión logística, Riesgo

ABSTRACT

The simulation and making models of geospatial phenomena allow to understand the observed phenomenon, to verify hypothesis and theories, to thus predict their behaviour in the space and the time under different conditions and scenes like discovering new operations and behaviours. This communication presents the use of Logistic Regression to generate a model of ignition wildfire risk. In this model contains the factors relative to the risk produced by the human activities and their relation with the fire occurrence due to this cause in the Community of Madrid, in period 1990-2004 (resolution 1km²). The tie factors to the human activity have special relevance although in the quantification of the risk a relative difficulty in valuing them exists and to spatialise them facing obtaining models that allow to predict the beginning and the propagation of the fire. However, the identification of the different types of factors from tie risk from the human activity is possible to generate variables of risk, related to the uses of the territory (protected routes of communication, garbage dumps, natural protected areas, etc.), socioeconomic aspects (population occupied in agriculture, level of rent, rate of unemployment, etc.) and others as zones of contact with the forest use (interface) or means of forest fire monitoring, that allows to obtain sufficiently trustworthy predictive models. These models can be of great interest for the planning and establishment of prevention strategies and management of the forest spaces.

Keywords

Human activities, Logistic Regression, Model, Risk, Wildfire

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). En nuestro país este fenómeno se relaciona con las características climatológicas propias de la región mediterránea pero también con la acción del hombre ya que el 96.1 % de los incendios obedecen a causas humanas (19.2 % es de causa desconocida. DGB, 2006).

El fenómeno del fuego es un factor más de los ecosistemas mediterráneos, con efectos positivos en un ciclo de recurrencia suficientemente largo. Buena parte de estos ecosistemas mediterráneos sólo se explican por una presencia recurrente del fuego (Martínez *et al.*, 2004 citando a Moreno). Sin embargo, este equilibrio se ha roto en las últimas décadas (Chuvienco *et al.*, 1998). El empleo del fuego ha sido práctica habitual a lo largo de la historia para el manejo del territorio. Sin embargo, 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). El fuerte éxodo ha provocado 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 (Martín *et al.*, 1998). 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 51.93% del territorio (MAPA, 2004).

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 y el desarrollo de nuevas actividades y usos en las zonas forestales, tales como el recreativo (Izquierdo, 2006). Estos cambios dan lugar a diversos problemas ecológicos y, 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).

Para poder llevar a cabo unas adecuadas labores de prevención es necesario conocer las causas de los incendios forestales así como acondicionar las zonas forestales para impedir la propagación. Como se ha citado anteriormente las estadísticas de incendio forestal que se recogen en España desde 1968 (Partes de Incendios Forestales), muestran que el factor humano explica más de un 90% de los incendios.

Las motivaciones que están detrás de las causas de origen humano pueden clasificarse en varias tipologías según el país. En el caso de España, se pueden distinguir (Leone *et al.* citando a Vélez, 2003)

- incendios en los que el causante espera un beneficio (quema de pastos, cambio de uso del territorio, generación de trabajos relacionados con extinción de incendios, etc.)
- incendios en los que el causante asume que no obtiene un beneficio tangible (conflicto de intereses, resentimiento contra la administración pública, vandalismo, etc.)
- incendios causados por razones políticas

La obtención de modelos de riesgo humano de incendio permite entender el fenómeno, comprobar hipótesis, predecir su comportamiento en el espacio y en el tiempo así como descubrir nuevos funcionamientos.

Para la modelización de esta ocurrencia de incendios debida a causa humana se utilizan métodos estadísticos tradicionales como la regresión logística, técnica que 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. La variable dependiente ha de ser dicotómica. Diversos autores han empleado esta técnica para la obtención de modelos predictivos de ignición de incendio: Chuvieco *et al.* (1999) y Martínez *et al.* (2004) a escala regional; Vasconcelos *et al.* (2001), Vega-García *et al.* (1993, 1995), Lin (1999), Pew *et al.* (2001) a escala local. Otros estudios utilizan esta técnica para la predicción de la ocurrencia diaria: Martell *et al.* (1985, 1987); Loftsgaarden y Andrews (1992); Vega-García *et al.* (1995). Se ha empleado para estudiar la interacción entre la ocurrencia de incendio y patrones de vegetación bajo la influencia de factores ambientales y las actividades humanas (Kalabokidis *et al.*, 2002). También para su comparación con los resultados obtenidos mediante otras técnicas como redes neuronales (Carvacho, 2002).

El presente trabajo se inscribe en el marco del proyecto *Firemap* de “Análisis Integrado de Incendios Forestales mediante Teledetección y Sistemas de Información Geográfica” (CGL2004-06049-C04-02/CLI), cuyo objetivo es la obtención de un índice de riesgo integrado. En esta línea el presente ensayo persigue la obtención de modelos de riesgo de incendios forestales a partir de variables riesgo humano, mediante técnicas de Regresión Logística, a nivel de cuadrícula (1*1 km) en la Comunidad de Madrid. La idea es desarrollar un modelo aplicable a otras regiones que pueda integrarse con facilidad en un modelo más complejo que incluya factores estructurales (vegetación, clima, etc.) relacionados con la ocurrencia de incendios forestales.

2. MATERIAL Y MÉTODOS

2.1 Área de estudio

La generación de modelos de riesgo humano de incendio mediante regresión logística se lleva a cabo en el conjunto de la Comunidad de Madrid. Esta área se divide en la unidad de análisis de 1*1 km, obteniendo un total de 8452 celdas.

La Comunidad de Madrid es una de las regiones más pobladas con unos 6 millones de habitantes (a 1 de enero de 2005, según el padrón municipal de habitantes del INE¹), lo que supone una tasa media de densidad de unos 748 habitantes/km². Destaca su alto grado de urbanización (8.6% de suelo urbano y 4.6% de suelo para sistemas generales 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, con gran diferencia, en el sector terciario. Las zonas forestales se distribuyen fundamentalmente del NO a SO de la comunidad y son frecuentadas con fines de tipo recreativo (existen actualmente 86 áreas recreativas en el área de estudio según datos de la *Dirección General de Medio Natural*. Según la opinión de los gestores de la región esta serie de características socioeconómicas y de usos del territorio entre otras influyen en el riesgo humano de incendio de manera notable (Martínez *et al.*, 2005. Así, por ejemplo, señalan a la aglomeración urbana como causa directa o indirecta de riesgo de incendio.

Se ha elegido como período de estudio el comprendido entre los años 1990 y 2004, período en el que se produjeron en la Comunidad de Madrid 3644 incendios, 899 de causa conocida. De los conocidos, 813 son de causa humana, un 90.43%.

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

Se muestran a continuación (Figura 1) las tendencias de causalidad de incendios forestales en la Comunidad de Madrid y en la Tabla 1 los porcentajes de cada causa respecto del total:

Tendencias de Causas de Incendios Forestales. 1990-2004. Comunidad de Madrid

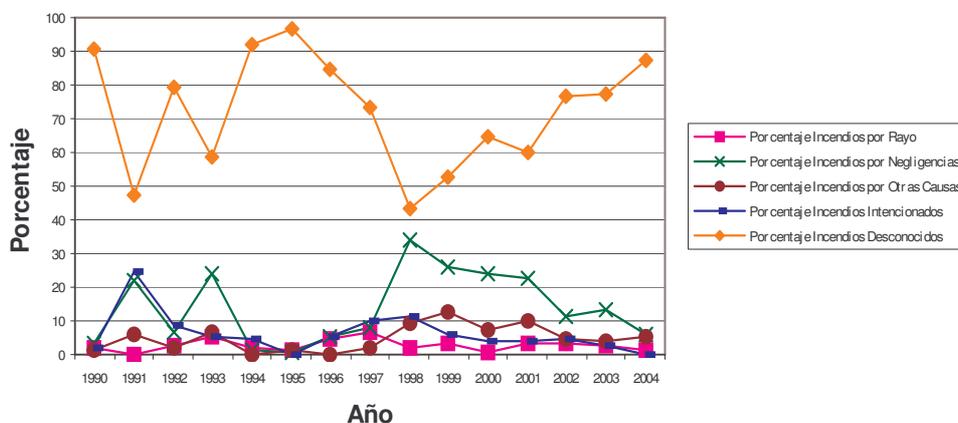


Figura 1. Tendencias de Incendios forestales por causa humana según causa. Período 1990-2004

Fuente: elaboración propia a partir Partes de Incendio DGB

Tabla 1. Nº de incendios según causa y porcentajes respecto al total de incendios período 1990-2004

Causa	Número incendios	Porcentaje respecto al total
Incendios por Rayo	86	2.36
Incendios por Negligencia	461	12.65
Incendios por Otras causas	175	4.80
Incendios Intencionados	177	4.86
Incendios Desconocidos	2745	75.33
Total	3644	100

Fuente: elaboración propia a partir Partes de Incendio DGB

2.2 Generación de Variables Independientes de riesgo humano

El primer paso en la generación del modelo de riesgo vinculado a la actividad humana consistió en la identificación de factores de riesgo y de las variables que podían permitir su cuantificación y especialización. Para ello se llevó a cabo una exhaustiva recopilación y de fuentes bibliográficas (Leone *et al* 2003, Martínez 2004, Martínez *et al* 2004, Pew *et al* 2001, Vega-García *et al* 1993) de proyectos a nivel local y regional de riesgo integrado de incendio forestal (*Firerisk*, 2003; *Spread*, 2003; *Megafires*, 2002) así como de la información obtenida en la realización de un cuestionario a expertos realizado en el marco del proyecto *Firemap*. En general se intentó considerar preferentemente aquellas variables de carácter estructural, relacionadas con elementos permanentes del territorio.

De entre diversas variables independientes de riesgo humano se han seleccionado las que se citan a continuación (Tabla 2). Estas variables han sido generadas a partir de fuentes cartográficas y estadísticas espacializándose a nivel de

cuadrícula UTM de resolución 1*1 km mediante el empleo de herramientas de Sistemas de Información Geográfica (*ArcMap 8.1, ArcInfo 8.1.2, ArcView 3.2 de ArcGis e Idrisi 3.2*), *Microstation SE* y de manejo de bases de datos y hojas de cálculo (*Access y Excel de Microsoft Office 2000*). La espacialización de las variables se ha realizado siguiendo dos metodologías, una para las cartográficas y otra para las estadísticas. En las primeras, las variables 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 las estadísticas, la fuente de datos es a nivel municipal. Por tanto, el valor en la cuadrícula será la media de los valores de los municipios en cada cuadrícula en función del área que ocupa cada municipio en esa cuadrícula.

Tabla 2. Variables independientes de riesgo humano de incendio forestal generadas

Tipo	Variable	Fuente
CARTOGRÁFICAS	Área de influencia (buffer) de áreas recreativas ponderadas por presencia de barbacoa (AREA_RECREE)	Mapa de coordenadas de áreas recreativas (2005)
	Áreas de influencia (buffer) de vías sin pistas forestales ² (CARRET, CARRET_FOR)	Mapa de carreteras a partir del Mapa de Tráfico 1:200.000 de la Comunidad de Madrid (2004)
	Índice de IMD por segmento de carretera (Longitud vía*IMD vía*factor de ponderación) (INDICE_IMD, INDICE_IMD_FOR)	Mapa de Tráfico 1:200.000 de la Comunidad de Madrid (2004)
	Áreas de influencia (buffer) de vías de ferrocarril (B_FFCC, B_FFCC_FOR)	Mapa de cartografía básica 1:50.000 de la Comunidad de Madrid (2000)
	Áreas de influencia (buffer) de pistas forestales (B_PISTAS, B_PISTAS_FOR)	Mapa de pistas a partir del Mapa de cartografía básica 1:50.000 (2000) de la Comunidad de Madrid actualizado con el Mapa de Tráfico 1:200.000 de la Comunidad de Madrid (2004)
	Áreas de influencia (buffer) de líneas eléctricas (B_LLEE, B_LLEE_FOR)	Mapa de cartografía básica 1:50.000 (2000) de la Comunidad de Madrid
	Potencial demográfico (POT_DEM)	Potencial demográfico de España por km ² asignado a cuadrícula de 25 km ² (2004)
	Índice de cambio en Superficie Forestal (ICC)	Mapa Forestal 1:50.000 de la Comunidad de Madrid (2000). Mapa de Cultivos y Aprovechamientos 1:50.000 (1974-1980) de la Comunidad de Madrid
Interfaz Urbano-Forestal Franja de 12.5 m (I_UFOR)	Mapa Forestal 1:50.000 de la Comunidad de Madrid (2000). Instituto de Economía y Geografía (CSIC). Ortofoto procedente proyecto SIGPAC (2002)	

² 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 1:50.000 de la Comunidad de Madrid (uso obtenido a partir de la reclasificación de categorías consideradas forestales, del campo "Tipo de Estructura" del Mapa Forestal)

Tipo	Variable	Fuente
	Interfaz Cultivo-Forestal Franja de 200 m (I_CULT_FOR)	Mapa Forestal 1:50.000 de la Comunidad de Madrid (2000). Ortofoto procedente proyecto SIGPAC (2002)
	Interfaz Pasto-Forestal Franja de 200 m (I_PASTO_FOR)	Mapa Forestal 1:50.000 de la Comunidad de Madrid (2000) Ortofoto procedente proyecto SIGPAC (2002)
	Áreas de influencia (buffer) de vertederos (VERTEDEROS)	Mapa Corine L&C2000 1:200.000 (2000) de la Comunidad de Madrid
	Campos de tiro y canteras (P_A_TIROCANTERAS)	Coordenadas de campos de tiro de la Comunidad de Madrid (Ministerio de Defensa, 2005) Mapa Corine L&C2000 1:200.000 (2000) de la Comunidad de Madrid
	Espacios naturales protegidos (ENP)	Espacios naturales protegidos 1:50.000 de la Comunidad de Madrid (2004)
	ZEPAS (ZEPAS)	ZEPAS 1:50.000 de la Comunidad de Madrid (2002)
	Montes de Utilidad Pública y Preservados (MUP_PRESER)	Mapa Montes de Utilidad Pública de la Comunidad de Madrid (2005) Mapa Montes Preservados ajustados a la cartografía 1:5000 de la Comunidad de Madrid
	Montes de Consorciados (MONTES_CONSOR)	Mapa Montes Consorciados de la Comunidad de Madrid (2005)
ESTADÍSTICAS	Variación de la población entre 1970-2004 (VAR_POB)	Padrón continuo de 2004 de la Comunidad de Madrid por municipios; Censo de Población 1970 (Anuario Estadístico de la Comunidad de Madrid citando al INE)
	Porcentaje de Jefes de explotaciones agrarias mayores de 55 años (JEFES55)	Censo Agrario 1999 (INE)
	Carga ganadera (cabezas de ganado ovino y caprino en superficie de pastos y matorral) (CARGA_GANADERA)	Censo Agrario 1999 (INE) Mapa Forestal 1:50.000 de la Comunidad de Madrid (2000)
	Densidad de maquinaria agrícola (MAQUINA)	Censo Agrario 1999 (INE)
	Variación de la población agraria entre 1996-2001 (VAR_POB_AGRA)	Población ocupada agraria de la Comunidad de Madrid 1996, 2001 por municipios (Anuario Estadístico de la Comunidad de Madrid 1996, 2005)
	Renta per. capita 2003 (RENTA)	Anuario Estadístico de la Comunidad de Madrid
	Plazas hoteleras totales 2004 (PLAZAS_HOTEL)	Base de datos municipal Almudena. Instituto de Estadística de la Comunidad de Madrid
	Tasa de paro 2001 (TASA_PARO)	Censo de Población y Vivienda 2001 (INE)

Fuente: elaboración propia

Las variables relacionadas con las vías de comunicación (buffer de carreteras, buffer de líneas de ferrocarril, etc.) así como las relacionadas con líneas eléctricas y campos de tiro y canteras, se agrupan en un factor de incendios causados por accidentes y negligencias. Por otro lado, las relacionadas con áreas recreativas, interfases (interfaz urbano-forestal), potencial demográfico, cambio en la superficie forestal y vertederos, variación de la población, porcentaje de jefes de explotaciones y plazas hoteleras, se agrupan en un factor de incendios producidos por cambios socioeconómicos. Las variables de interfaz cultivo-forestal y pasto-forestal, carga ganadera y maquinaria agrícola se agrupan por otro factor, de incendios producidos por actividades rurales tradicionales. Finalmente, el factor de incendios producidos por la generación de conflictos, agrupa las variables relacionadas con montes, espacios naturales protegidos y ZEPAS, tasa de paro y nivel de renta.

2.3 Generación de la Variable Dependiente (ocurrencia de incendios de causa humana en zona forestal)

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) donde la localización espacial de los incendios se recogen a nivel de cuadrícula 10*10 km y por municipio, sin saber con exactitud la posición de los puntos de ignición. Teniendo en cuenta la unidad de análisis elegida (cuadrícula 1*1 km) se decidió aplicar un procedimiento que permitiera reducir la incertidumbre de localización de los puntos de inicio del incendio. Así se precisó la localización espacial de los incendios cruzando la información de localización por cuadrículas y municipios y eliminando las zonas sin superficie forestal. De esta forma se consiguen polígonos donde *a priori* los incendios que se han producido van a estar mejor localizados. A partir del número total de incendios obtenidos (4537) referidos a una localización espacial más precisa se generan puntos aleatorios en esas zonas, mediante el *script* de ArcView 3.2 *Random Point Generator v. 1.3*³. A partir de la distribución aleatoria de puntos de ignición se convirtieron estas observaciones puntuales en superficies continuas para reducir la imprecisión en la localización de los puntos (Amatulli *et al.*, 2005). Una manera habitual para calcular densidades a partir de datos puntuales es dividir a la muestra en un número finito de intervalos del mismo tamaño y asignar un valor de densidad en función de los puntos que entran en ese intervalo. Para ello se superpone una malla regular. El inconveniente de este proceso es que el resultado depende del punto inicial del intervalo y del tamaño del mismo (Amatulli *et al.*, 2005). De igual forma, al superponer una malla regular sobre puntos imprecisos, el resultado puede ser muy dependiente de la magnitud de los errores de posición y de la resolución de la malla (de la Riva *et al.* 2004 citando a Koutsias *et al.*). Por estos motivos se propone utilizar métodos de interpolación para convertir puntos en áreas continuas (de la Riva *et al.*, 2004).

Para obtener superficies continuas partir de los puntos de ignición se utiliza la técnica de interpolación de estimación de densidad de Kernel. 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):

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

³ *Random Point Generator v. 1.3*. Autor: Jeff Jenness. Wildlife Biologist, GIS Analyst. Jenness Enterprises. jeffj@jennessent.com

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 modo, 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 más estrecho, mientras que con poca presencia de puntos será más ancho (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 10 puntos⁴ y una malla de resolución de 1000 m, utilizando *Crimestat*® 3.0 (Levine, 2004). La *Figura 2* muestra la variable dependiente obtenida:

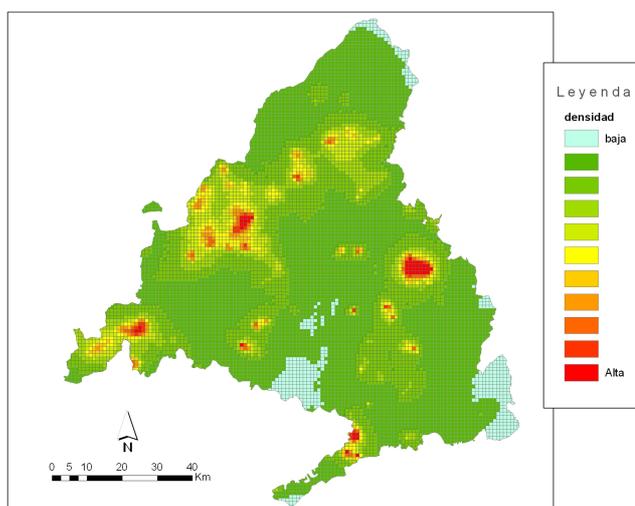


Figura 2. Densidad de incendios de causa humana por interpolación mediante método Kernel adaptativo de intervalo de bandwidth 10 puntos

Fuente: elaboración propia

2.4 Desarrollo de los modelos

2.4.1 Regresión logística

El objetivo que se persigue con la aplicación de este modelo es estimar la probabilidad de ocurrencia de incendios que se ha convertido en una variable dependiente dicotómica (alta incidencia de incendio, baja incidencia de incendio) a partir de las variables independientes. Se trata de obtener la probabilidad de que cada individuo pertenezca a cada uno de los grupos que define la variable dependiente (González, 2004).

El modelo de regresión logística se define:

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

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

⁴ El grado de precisión depende del tamaño de la muestra del intervalo de *bandwidth*. Por defecto se establece en 100 puntos dentro del radio del *bandwidth*. Cuanto menor sea el número de puntos de esta muestra más fina será la estimación (Levine, 2004)

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

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

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

Siendo x^T el vector de las variables explicativas y β el vector de los parámetros (González, C., 2004). Se emplea el método de “pasos hacia delante de *Wald*”, con punto de corte para la clasificación de 0.5.

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:

- Coeficiente de tolerancia: Indicador de la independencia de una variable respecto de otras. El porcentaje de esa variable que no es explicada por las otras independientes. Por debajo de 0.1 se considera que la multicolinealidad es alta (Martínez, 2004 citando a Vaquerizo, 2003)
- Factor de Inflación de la Varianza (FIV): El valor recíproco de la tolerancia (1/Tolerancia), siempre mayor que 1. Es un indicador de lo que aumenta la varianza del coeficiente de regresión de la variable. Cuanto mayor sea este factor mayor será la multicolinealidad (Martínez, 2004 citando a Etxebarria, 1999). Valores mayores de 30 indican problema serio y mayores de 15 posible inconveniente
- Autovalores, Índice de condición y proporción de la varianza de los coeficientes de regresión asociada a cada valor propio (Martínez, 2004 citando a Visauta, 1998): Los autovalores suministran indicación de cuantas dimensiones distintas hay entre las variables independientes. Cuando varios están cercanos a 0, las variables están altamente correlacionadas. El índice de condición es la raíz cuadrada del ratio entre el máximo autovalor y cada sucesivo autovalor. Cuanto mayor sea el índice, mayor dependencia entre las variables implicadas. Un valor mayor de 15 indica un posible problema y mayor de 30 un gran problema (SPSS)

De igual forma, se ha explorado el grado de correlación entre las variables independientes, mediante coeficientes de correlación no paramétricos de *Spearman*. Aquellas variables correladas entre si más de 0.7 se excluyen del modelo.

Finalmente, se ha aplicado un test de estadística comparativa que proporciona 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: prueba no paramétrica sobre la significación de la diferencia entre las medianas (Martínez *et al.*, 2005). Se seleccionan las muestras

dividiendo en cuatro cuartiles la variable dependiente y tomando para comparar el primero y el cuarto

Para transformar la variable dependiente de continua a dicotómica se siguió el criterio de ordenar los datos en función de la variable dependiente, dividiendo a continuación la variable en 3 grupos con el mismo número de casos. A los obtenidos en el primer grupo se les da valor 0 (baja incidencia de incendios) y a los del grupo 3 valor 1 (alta incidencia de incendios).

Para llevar a cabo una validación del modelo obtenido se seleccionó aleatoriamente el 60% de los casos para obtener la ecuación de regresión, utilizando el 40% restante para validar la calidad de las estimaciones. Posteriormente, el modelo se aplicó a toda la cuadrícula de la Comunidad de Madrid.

3. RESULTADOS

3.1. Análisis previos colinealidad, correlación y tests de estadística comparativa

Los diagnósticos de colinealidad obtenidos a partir de técnicas multivariantes de Tolerancia y FIV indican que las variables de pistas y pistas en zona forestal presentan posibles problemas de colinealidad (tabla 3):

Tabla 3. Coeficientes de Tolerancia próximos a 1 y FIV con mayor valor de las variables independientes

	Estadísticos de colinealidad	
	Tolerancia	FIV
B_PISTAS_FOR	0.118	8.463
B_PISTAS	0.125	8.019

Los resultados obtenidos de autovalores e índice de condición muestran que puede haber un problema de colinealidad, debido a que se obtienen autovalores cercanos a 0 en 4 dimensiones, por lo que hay variables que están altamente correlacionadas. Asociados a estas dimensiones hay valores de índice de condición cercanos a 15 y mayor a 15 en un caso. La proporción de la varianza para esas dimensiones posee el mayor valor en las variables de: VAR_POB_AGRA, BUFFER PISTAS, BUFFER PISTAS_FOR y RENTA, indicando un posible problema de correlación.

Las correlaciones bivariadas de *Spearman* muestran que las variables más correlacionadas entre sí, con un valor superior a 0.7 son las siguientes (tabla 4):

Tabla 3. Correlaciones bivariadas no paramétricas de Spearman superiores a 0.7 de las variables independientes

Variable	Variable	Coefficiente de correlación <i>Spearman</i>
INDICE_IMD_FOR	B_CARRET_FOR	0.966
INDICE_IMD	B_CARRET	0.908
B_PISTAS_FOR	B_PISTAS	0.956
B_FFCC_FOR	B_FFCC	0.72
B_LLEE_FOR	B_LLEE	0.75

El test de estadística comparativa de la U-Mann-Whitney muestra que las variables B_FFCC B_LLEE y P_A_TIROCANTERAS no presentan diferencias significativas al 95% de confianza (p-valor mayor de 0.05) para dos muestras independientes del primer y cuarto cuartil.

A partir de los resultados obtenidos en los análisis de colinealidad, correlación y tests de estadística comparativa se excluyen del modelo de regresión logística las variables de renta (RENTA), variación de la población agraria (VAR_POB_AGRA), buffer de vías de ferrocarril (B_FFCC), buffer de líneas eléctricas (B_LLEE), presencia o ausencia de Tiro-Canteras (P_A_TIROCANTERAS), carreteras totales (B_CARRET) y en zonas forestales (B_CARRET_FOR), pistas (PISTAS). De las 30 variables iniciales se seleccionaron 22.

3.2. Regresión logística

La construcción del modelo a partir de las variables anteriores y con una muestra del 60% de los casos da lugar a 17 modelos, con un porcentaje de acierto superior al 75% a partir del modelo 8. La baja incidencia de incendio en este modelo ha sido correctamente clasificada en un 80.9% mientras que la alta en un 68.8%. La muestra de validación tiene un porcentaje de acierto global del 74% (Tabla 4):

Tabla 4. Clasificación de la muestra de entrenamiento (60%) y muestra de validación (40%)

OBSERVADO		PRONOSTICADO					
		Casos seleccionados (60%)			Casos no seleccionados (40%)		
		Variable dependiente		Porcentaje correcto	Variable dependiente		Porcentaje correcto
		baja incidencia	alta incidencia		baja incidencia	alta incidencia	
Variable dependiente	baja incidencia	1395	329	80,9	877	216	80,2
	alta incidencia	519	1144	68,8	369	785	68,0
Porcentaje global				75,0			74,0

Las variables que han sido incluidas en el modelo son INDICE_IMD_FOR, B_PISTAS_FOR, ENP, ZEPA, ICC, I_UFOR, I_CULT_FOR, PLAZAS_HOTEL. La interfaz urbano-forestal es la variable que posee un mayor coeficiente en la regresión, seguida por las pistas en zona forestal y los espacios naturales protegidos. Las que poseen el coeficiente más bajo son el índice IMD en zona forestal, el índice de crecimiento de combustible y las plazas hoteleras. La interfaz cultivo forestal aparece con el signo contrario al esperado, con signo negativo, el cual puede deberse a que siguen existiendo problemas de colinealidad entre las variables.

Se aplica el modelo obtenido a la totalidad de los casos dando lugar a un mapa de ocurrencia de incendio en escala probabilística (entre 0 y 1) estimada que una cuadrícula cuente con alta incidencia de incendios provocados por la actividad humana (Figura 3):

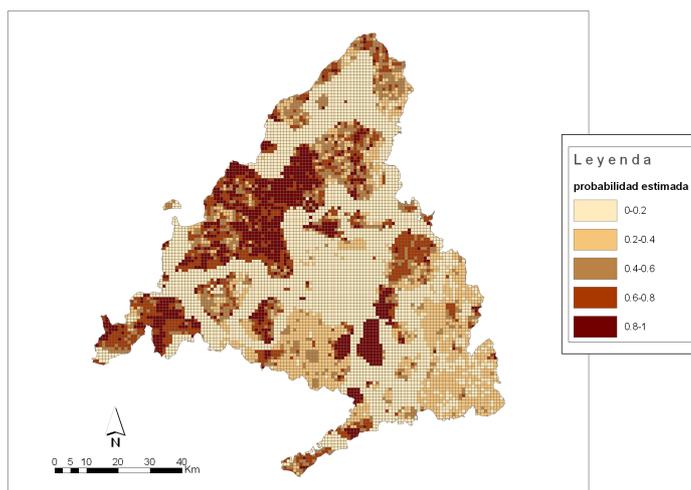


Figura 3. Probabilidad estimada de incendios de causa humana a partir del modelo de regresión logística

Fuente: elaboración propia

Este mapa muestra una probabilidad estimada de incendio que se asemeja a resultados anteriores a nivel municipal (Martínez, 2005), siendo las cuadrículas de mayor probabilidad las de la zona Oeste de la Comunidad de Madrid, así como un corredor de NE a SE.

4. CONCLUSIONES

La modelización del riesgo humano de incendios forestales es una herramienta que nos permite entender el fenómeno e intentar hacer predicciones. Los resultados obtenidos por regresión logística responden a lo señalado por los gestores de la Comunidad de Madrid, como es el caso de la problemática de la interfaz urbano-forestal. Aparecen en el modelo variables relacionadas con las vías de comunicación (índice de IMD en zona forestal y pistas en zona forestal), aunque con poco peso en el conjunto del modelo.

En próximos trabajos se valorará el empleo de técnicas de *Data Mining* en la generación de modelos de riesgo humano de incendio forestal, ya que son técnicas que permiten una mayor flexibilidad en la elección de variables de entrada, al no contar con las restricciones habituales de la estadística tradicional.

El conocimiento de las causas de incendios es necesario para el establecimiento de estrategias de prevención de incendios, estimando la probabilidad de que el fenómeno ocurra.

BIBLIOGRAFÍA

- Amatulli, G., Pérez-Cabello, F., de la Riva, J. (2005). *Mapping lightning/human-caused wildfires occurrence under ignition point location uncertainty*. Ecological modelling. Manuscript
- Chuvienco, E., Martín, P., Martínez, J., Salas, J. (1998): "Geografía e Incendios Forestales", *Serie Geográfica*, 7, pp. 11-17.
- Dirección General para la Biodiversidad (2006): Estadísticas de Incendios Forestales. <http://www.incendiosforestales.org/estadisticas.htm>. Ministerio de Medio Ambiente.

- González, C. (2006): “Análisis de Datos Cualitativos”, Curso de Metodología de Investigación Cuantitativa. Técnicas Estadísticas. CSIC
- Izquierdo, J., (2006): *Jornada sobre Incendios Forestales. Fundación Biodiversidad. Fundación Santander-Central Hispano*
- 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
- 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. <http://eufirelab.org>
- Martínez, J. (2004): *Análisis, Estimación y Cartografía del Riesgo Humano de Incendios Forestales*. Tesis Doctoral. Departamento de Geografía. Universidad de Alcalá
- Martínez, J., Martínez Vega, J., Martín, P. (2004): “El factor humano en los incendios forestales: Análisis de factores socio-económicos relacionados con la incidencia de incendios forestales en España”, en Chuvieco, E., Martín, P. (Ed.): Nuevas tecnologías para la estimación del riesgo de incendios forestales. Madrid, CSIC, Instituto de Economía y Geografía, pp. 101-142.
- 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, en 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. Servicio de Publicaciones Universidad de Extremadura, pp. 329-345
- Ministerio de Agricultura, Pesca y Alimentación (2004): “Hechos y cifras de la Agricultura, la Pesca y la Alimentación en España”. <http://www.mapa.es/es/ministerio/pags/hechoscifras/introhechos.htm>
- 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*
- Pew, K.L., Larsen, C.P.S (2001): “GIS analysis of spatial and temporal patterns of human-caused wildfires in the temperate rain forest of Vancouver Island, Canada”, *Forest Ecology and Management*, 140, pp. 1-18
- Pueyo Campos, A. *et al* (2006): Mapa de Potencial Demográfico 2004 de España. Universidad de Zaragoza
- De la Riva, J., Pérez-Cabello, F., Lana-Renault, N., Koutsias, N. (2004): “Mapping wildfire occurrence at regional scale”, *Remote Sensing of Environment*, 92, pp. 363-369.
- Villagarcía, T. (2006): “Regresión”, Curso de Metodología de Investigación Cuantitativa. Técnicas Estadísticas. CSIC

