UNIVERSIDAD DE OVIEDO

ESCUELA POLITÉCNICA DE MIERES

MÁSTER EN TELEDETECCIÓN Y SISTEMAS DE INFORMACIÓN GEOGRÁFICA

DETERMINACIÓN DE LA PROBABILIDAD DE OCURRENCIA DE INCENDIO POR RAYO EN CASTILLA Y LEÓN

TRABAJO FIN DE MÁSTER

AUTOR:
SUSANA EGEA TRAPIELLO

TUTOR:
CELESTINO ORDOÑEZ GALÁN

Asturias, Julio de 2015

INDICE

RESUMEN 3
1 INTRODUCCIÓN 4
2 OBJETIVOS 5
3 MÉTODOS Y MATERIALES 6
3.1 DATOS 6
3.1.1 ORIGEN Y CARACTERIZACIÓN 6
3.1.2 DISTRIBUCIÓN ESPACIAL 7
3.1.3 COMPARATIVA VISUAL ENTRE VARIABLES 20
3.2 METODOLOGÍA 27
3.2.1 COMPROBACIÓN DE LA NORMALIDAD DE LAS VARIABLES 28
3.2.2 CORRELACIÓN DE VARIABLES 32
3.2.3 CONSTRUCCIÓN DEL MODELO CON REGRESIÓN LOGÍSTICA 32
3.2.4 VALIDACIÓN DEL MODELO 37
4 RESULTADOS 40
6 CONCLUSIONES 41
BIBLIOGRAFÍA 43
ANEXO I. MAPA DE PROBABILIDAD DE OCURRENCIA DE INCENDIO POR RAYO EN CASTILLA Y LEÓN 45

RESUMEN

El objeto de este trabajo es construir un modelo que permita predecir la probabilidad de ocurrencia de incendios por rayo (causa natural) en Castilla y León, así como conocer qué factores son los que mayor incidencia tienen en la aparición de los mismos.

Inicialmente en nuestro modelo, el número de variables independientes o explicativas es muy elevado, lo que no permite interpretar los resultados con facilidad. Por ello debe realizarse una selección de aquellas que aporten mayor capacidad de discriminación. Para ello se usa el modelo de regresión logística binaria que permite estimar la probabilidad de ocurrencia de incendio a causa de rayos.

El modelo utiliza como variable dependiente la presencia o ausencia de incendio y a priori consta de 23 variables independientes relacionadas con la cobertura vegetal, datos fisiográficos, características intrínsecas a los rayos y datos sobre tormentas.

Los resultados obtenidos muestran que tres de ellas permiten construir el modelo, siendo la superficie de terreno ocupada por cultivos y prados, la altitud y el número de días de tormenta las variables que mejor explican el proceso. El modelo muestra una fiabilidad global aceptable (por encima del 64%).

Palabras clave: Ocurrencia, incendios, rayos, regresión logística.

ABSTRACT

The object of this work is to build a model in order to predict the probability of occurrence of lightning-induce fires in “Castilla y León”.

Firstly, our model have a huge number of independent or explanatory variables, which do not allow to interpret the result easily. Therefore, it is necessary choose those which provide the greater discrimination capacity. For that, it is use the binary logistic regression model with the aim of predict the probability of occurrence of lightning-induce fires.

The primary model have the presence or absence of fire as the dependent variable and 23 independent variables related with vegetation cover, physiographic data, lightning characteristics and data about thunderstorms.

The results show that three of them allow to build the model and are the best to explain the process. These are: the crops areas, the altitude and the number of thunderstorms days. The model shows an acceptable global reliability (above 64%).

Keywords: Occurrence, fire, lightning, logistic regression.

1 INTRODUCCIÓN

Una de las amenazas más graves del patrimonio forestal, material y humano son los incendios. Según la estadística general de incendios forestales publicada por la Dirección General del Desarrollo Rural y política Forestal, Subdirección General de Silvicultura y Montes del Ministerio de Agricultura, Alimentación y Medio Ambiente, en España la media es de 17.117 siniestros anuales y 113.847,72 ha afectadas. Esto hace que sea necesario conocer con precisión los incendios forestales para desarrollar actuaciones de defensa.

La Comunidad Autónoma de Castilla y León sufre anualmente gran número de incendios, principalmente debido a su gran superficie forestal. En el periodo 2000-2010, los incendios causados por rayo representaron de media un 6% de los totales que se producen en la Comunidad Autónoma, sin embargo, en algunos años como en 2003 y 2006, han llegado a representar más del 15% del total, y en otros años como 2003, 2005 y 2006 más del 10% de la superficie total quemada (Faba-Fernández, M. *et al.*, 2013).

Múltiples y variados son los agentes que generan un incendio, siendo los causados por rayo los más fácilmente reconocibles. Diversos investigadores determinan que éstos no ocurren aleatoriamente si no que tienden a iniciarse en lugares concretos (Vankat 1985, citado por Ordoñez, C. *et al.* 2013). Las variables intrínsecas de las descargas, variables fisiográficas y características del combustible (muy relacionadas con el tipo de cubierta vegetal), pueden explicar en gran medida la probabilidad de inicio (PACHECO *et al*., 2009; NIETO *et al*., 2012 citado por Faba-Fernández, M. *et al.*, 2013).

Con el objeto de prevenir, minimizar y mitigar los efectos de incendios por rayo inicialmente se realizan análisis y mapas de riesgos que indican las áreas más vulnerables (Bonazountas *et al.* 2005; citado por Ordoñez, C. *et al.* 2013).

Se han realizado diferentes estudios estadísticos siendo los análisis por regresión logística los que han sido especialmente exitosos en la predicción de ocurrencia de incendios y en el estudio de factores críticos involucrados proporcionando mayores casos de éxito (García *et al*., 1995; Vasconcelos *et al*. 2001; Andrews *et al*. 2003; Wotton and Martell, 2005; Martínez *et al.* 2009; Vilar *et al.* 2010; Ordoñez, C. *et al*. 2013).

Uno de los principales problemas a la hora de desarrollar un modelo de probabilidad de ocurrencia de incendio con regresión logística es el de identificar el conjunto de variables predictoras que den la mejor capacidad discriminatoria, evitando la inclusión de variables irrelevantes o redundantes.

2 OBJETIVOS

El objetivo principal del presente trabajo es el de desarrollar un modelo que permita determinar la probabilidad de ocurrencia de incendios forestales a causa de rayo en la comunidad autónoma de Castilla y León, además de conocer qué factores son los que mayor incidencia tienen en la aparición de los mismos.

Para ello se parte de datos de diversas variables fisiográficas, variables relacionadas con la cubierta vegetal existente y variables relacionadas con las descargas de los rayos, cuya descripción se encuentra en apartados posteriores.

Tras su tratamiento estadístico y mediante regresión logística se obtiene un modelo que permitirá finalmente realizar una cartografía de probabilidad de incendio por rayo mediante Sistemas de Información Geográfica, siendo utilizado en este caso el paquete ESRI, Arcmap 10.0.

A grandes rasgos, los pasos a seguir son:

- Filtrado de los datos.
- Visualización de la información que permite realizar una valoración inicial de las relaciones entre variables.
- Obtención de relaciones entre variables.
- Selección de las mejores variables.
- Combinación de variables mediante regresión logística para la obtención del modelo de probabilidad de incendio por rayo.
- Validación del modelo con una muestra de entrenamiento aleatoria del 70% y el 30% restante de test.
- Cálculo de la probabilidad de ocurrencia de incendio por rayo en la zona de estudio y su representación gráfica.

3 MÉTODOS Y MATERIALES

Una vez establecida la problemática del trabajo y el objetivo a conseguir, se procede a establecer la metodología que ayudará a dar respuesta al problema. Para ello y en primer lugar, hay que conocer la fuente de datos que permitirá obtener las variables necesarias para explicar y obtener el modelo de probabilidad de ocurrencia de incendio por rayo y proponer seguidamente el método de análisis que será fundamentado en la regresión logística.

3.1 DATOS

3.1.1 ORIGEN Y CARACTERIZACIÓN

El trabajo se ha realizado con una base de datos en formato Excel suministrada por la Agencia Estatal de Meteorología (AEMET) del Ministerio de Agricultura, Alimentación y Medio Ambiente.

Dicha base de datos contiene información estructurada en 6253 celdas según una malla de 4x4 kilómetros que abarca toda la comunidad autónoma, con datos acerca del tipo de vegetación, relieve, precipitación, características de los rayos, etc.

A continuación se muestra la tabla resumen de las variables contenidas en la base de datos:

Tabla 1. Variables de la base de datos

Campo	Abreviatura	Descripción
id_4km	ID	Identificador de las celdas que componen la malla
Incendios	FIRE	Indica la presencia (1) o ausencia (0) de incendios forestales causados por rayo en la celda
Número_incendios	FIRE_N	Número de incendios forestales causados por rayo en la celda.
Coníferas	CONIFEROUS	% de la superficie de la celda que ocupan las masas de coníferas.
CultivosPrados	CROPS	% de la superficie de la celda que ocupan los cultivos y prados.
Frondosas	BROADLEAF	% de la superficie de la celda que ocupan las masas de frondosas.
Matorrales	SHRUBLAND	% de la superficie de la celda que ocupan los matorrales y arbustos.
Mixtas	MIXED	% de la superficie de la celda que ocupan las masas mixtas.
Pastizales	MEADOW	% de la superficie de la celda que ocupan los pastizales
Otros	OTHER	% de la superficie de la celda que ocupan otros usos del suelo (zonas urbanas, de extracción minera, vertederos, roquedos, aguas continentales, etc.)
Altitud_media	ALTITUDE	Altitud media de la celda expresada en metros
Pendiente_media	ALTITUDE	Pendiente media de la celda expresada en porcentajes
Norte	NORTH	% de superficie de la celda con orientación norte
Este	EAST	% de superficie de la celda con orientación este

Tabla 1. Variables de la base de datos		
Campo	**Abreviatura**	**Descripción**
Sur	SOUTH	% de superficie de la celda con orientación sur
Oeste	WEST	% de superficie de la celda con orientación oeste
Llano	FLAT	% de superficie de la celda sin orientación, llano
Rayos_negativos/km^2	NEGATIVE_N	Cociente entre el número de rayos válidos con intensidad negativa que han caído en la celda y su superficie. Se mide en rayos/km^2
Intensidad_media_ negativa	NEGATIVE_I	Intensidad media en kA de los rayos con intensidad negativa que han caído en la celda
Duracion_media_ negativa	NEGATIVE_L	Duración media en ns de los rayos con intensidad negativa que han caído en la celda.
Rayos_positivos/km^2	POSITIVE_I	Cociente entre el número de rayos válidos con intensidad positiva que han caído en la celda y su superficie. Se mide en rayos/km2.
Intensidad_media_ positiva	POSITIVE_N	Intensidad media en kA de los rayos con intensidad positiva que han caído en la celda.
Duracion_media_ positiva	POSITIVE_L	Duración media en ns de los rayos con intensidad positiva que han caído en la celda.
Rayos_validos/km^2	STROKES_N	Cociente entre el número de rayos válidos (tanto los de carga positiva como negativa) que han caído en la celda y su superficie. Se mide en rayos/km^2.
Dias_tormenta	THUNDERSTORMS	Número de días de tormenta
TS25	THUNDERSTORMS_D	Número de días de tormenta en la que la precipitación ha sido inferior a 2,5 mm.

Según la información proporcionada, los rayos recogidos son los caídos entre los meses de Mayo y Septiembre, de primera descarga y válidos; en el periodo comprendido entre el año 2000 y 2011.

Se considera que un rayo es válido cuando cumple simultáneamente los siguientes criterios (Álvarez E., *et al*. 2011):

- Semieje mayor de la elipse de localización menor que 6 km
- Semieje menor de la elipse de localización menor que 3 km
- Valor del estadístico normalizado Chi^2 menor que 10.

Para 29 de las 6253 celdas se carece de información fisiográfica por lo que se ha decidido eliminarlas, para así trabajar con los datos de 6224 celdas.

3.1.2 DISTRIBUCIÓN ESPACIAL

Una vez eliminadas las celdas incompletas de datos se representan con ArcGis todas las variables en formato ráster.

A continuación se incluyen las figuras que muestran la distribución espacial de todas y cada una de las variables que componen la base de datos.

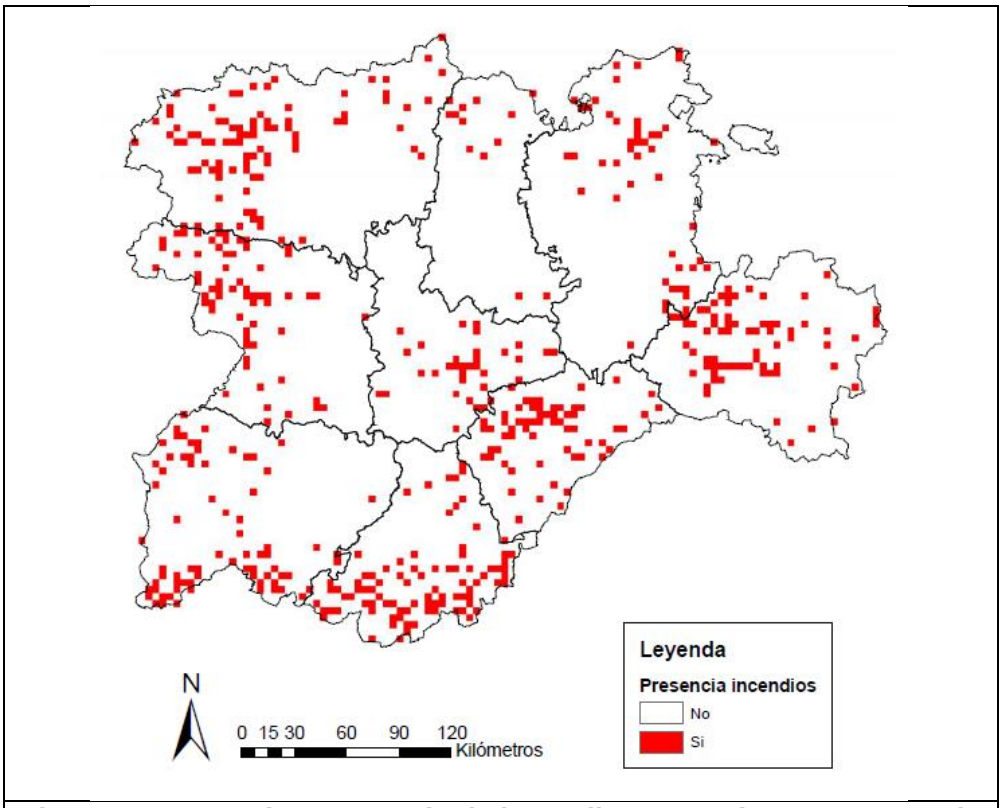


Figura 1. Presencia o ausencia de incendios causados por rayo en la celda

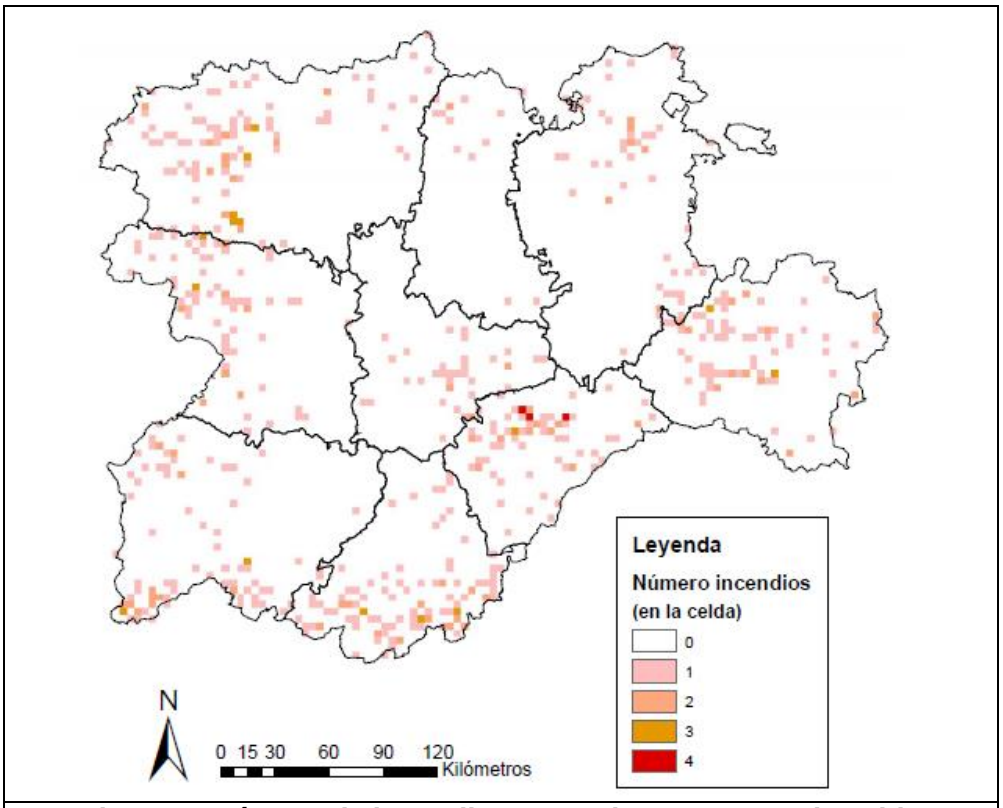


Figura 2. Número de incendios causado por rayo en la celda

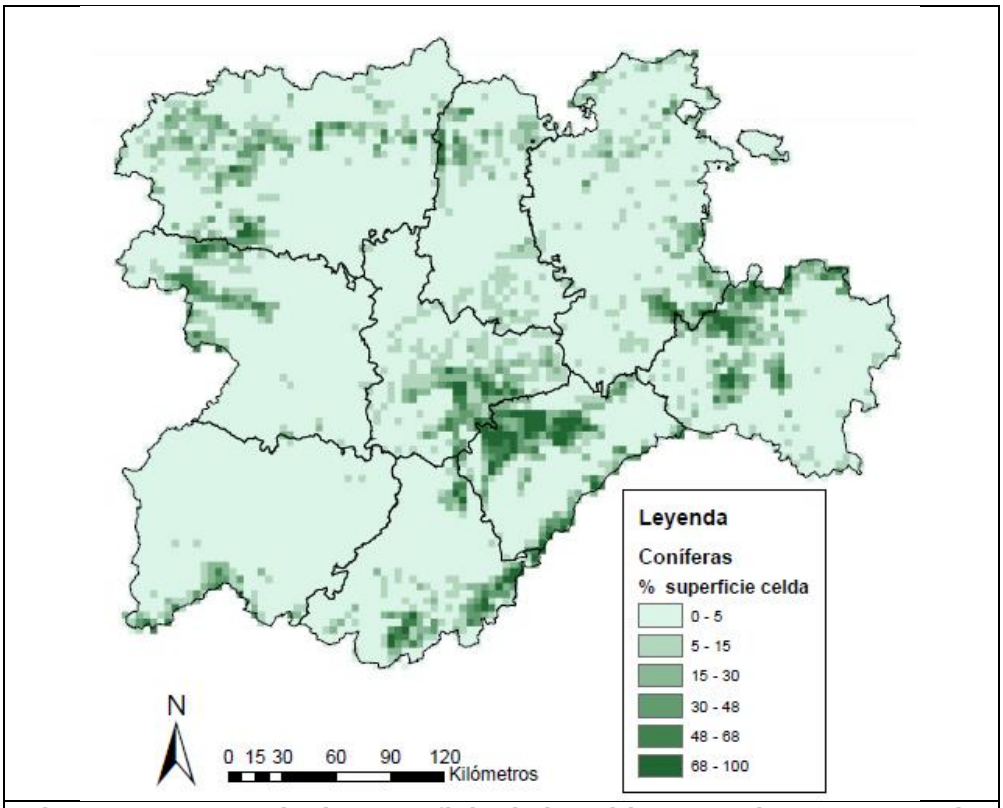


Figura 3. Porcentaje de superficie de la celda ocupado por masas de coníferas

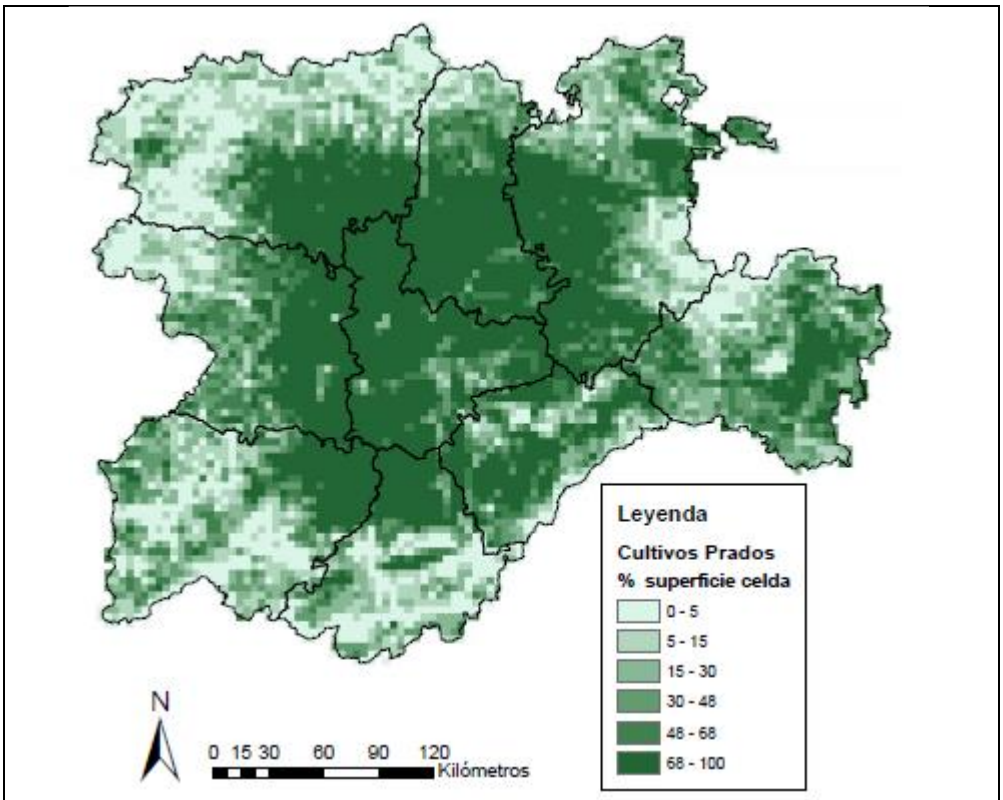


Figura 4. Porcentaje de superficie de la celda ocupado por cultivos y prados

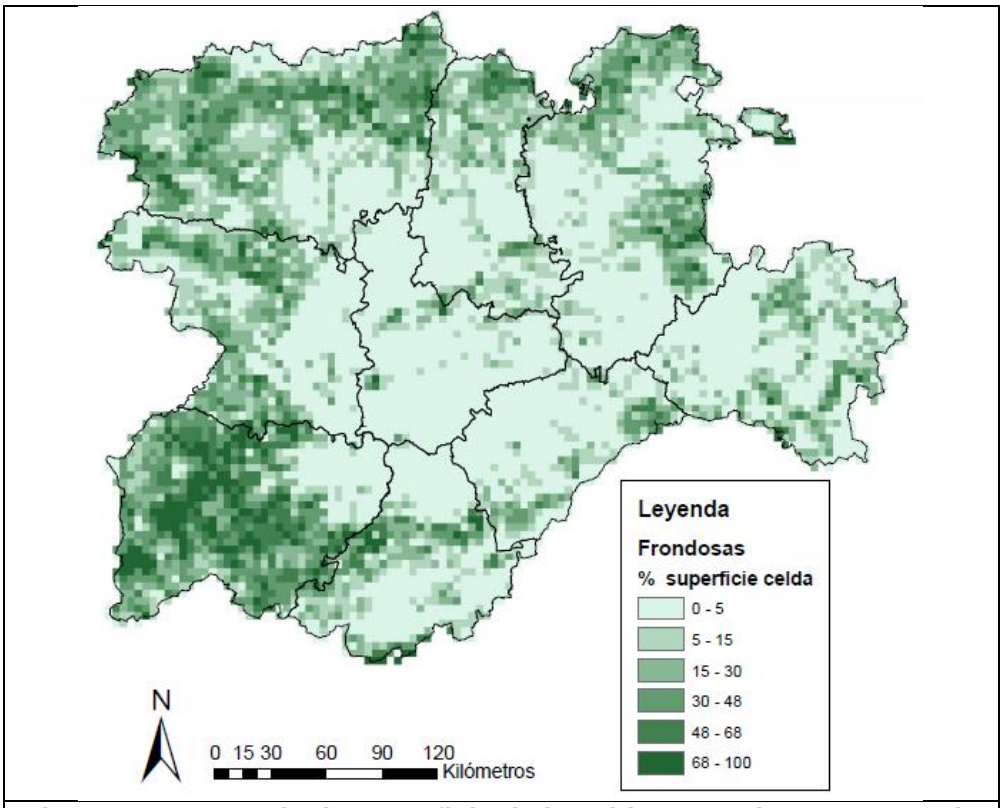


Figura 5. Porcentaje de superficie de la celda ocupado por masas de frondosas

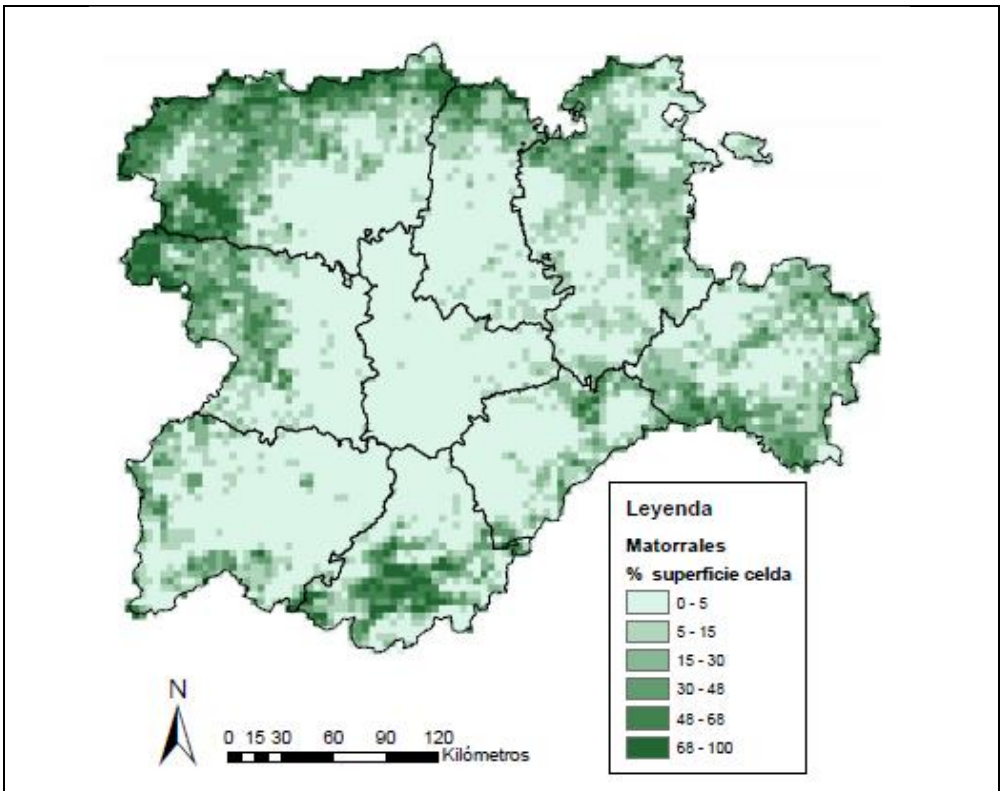


Figura 6. Porcentaje de superficie de la celda ocupado por matorrales y arbustedos

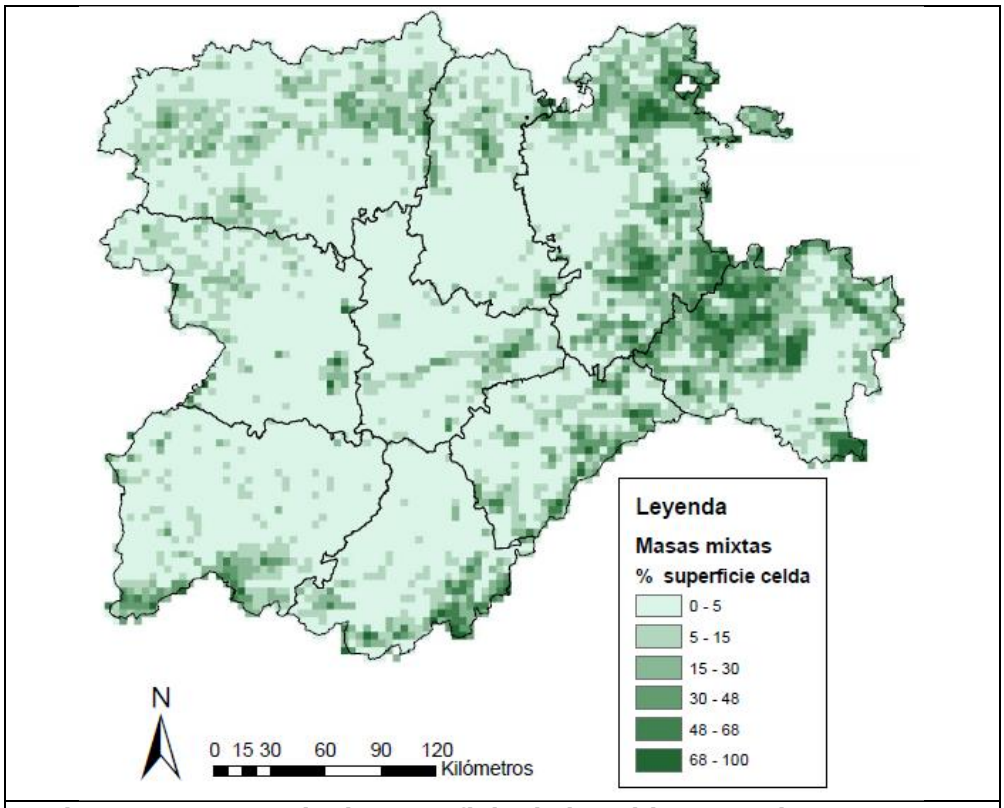


Figura 7. Porcentaje de superficie de la celda ocupado por masas mixtas

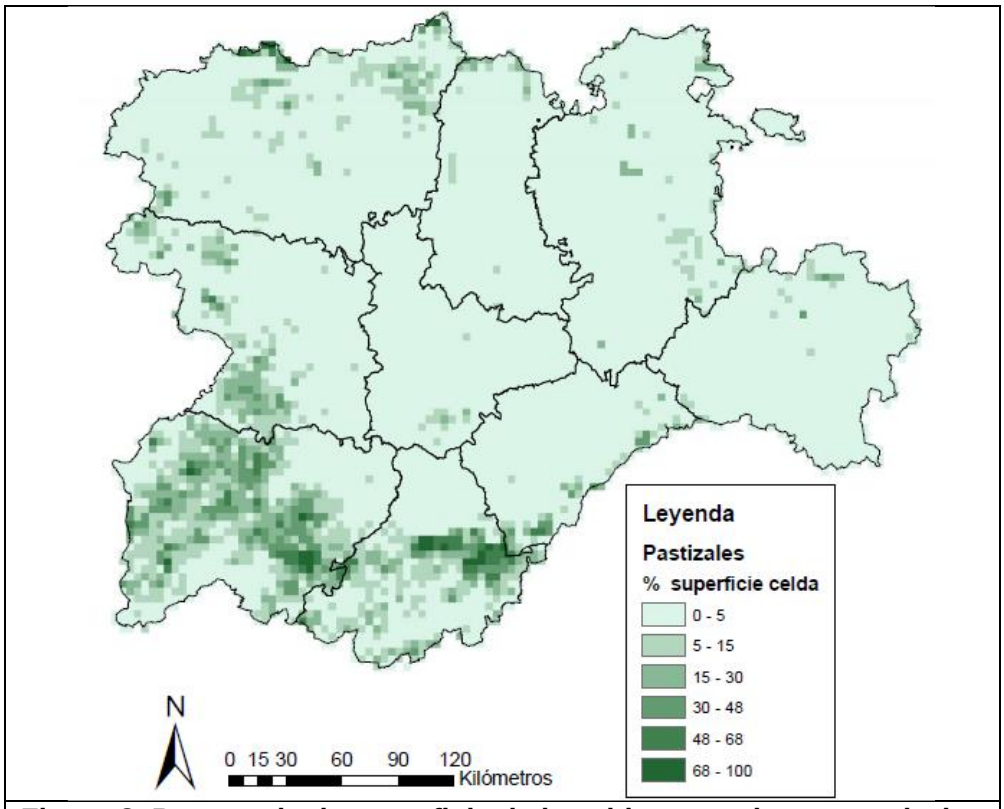


Figura 8. Porcentaje de superficie de la celda ocupado por pastizales

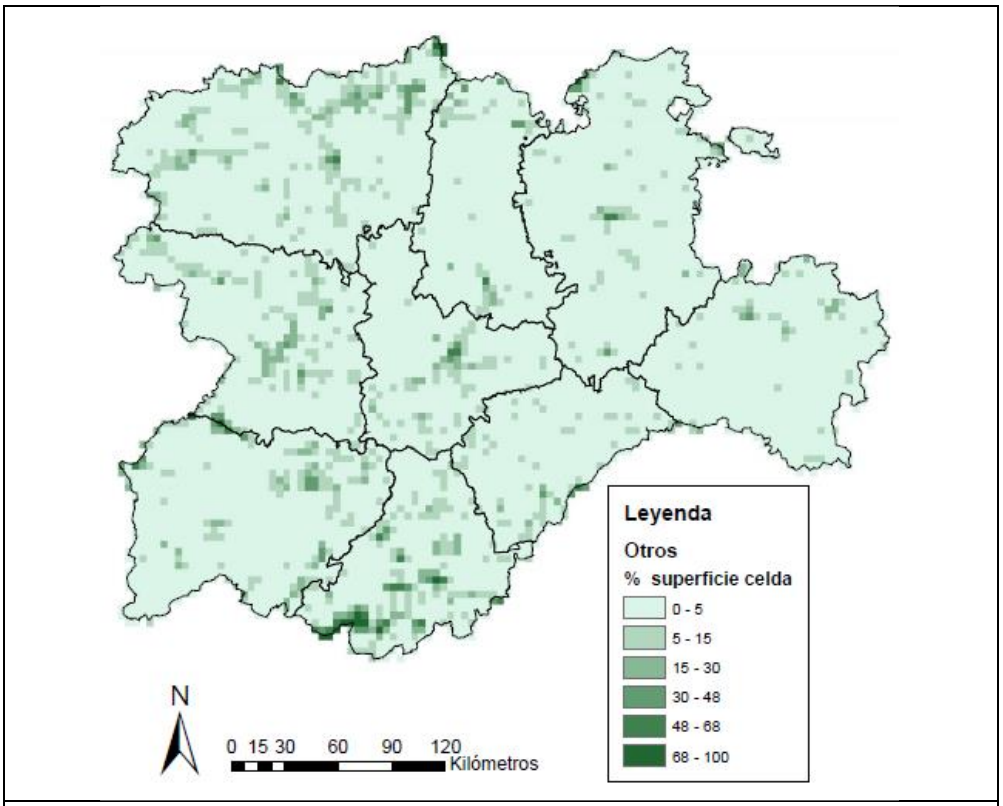


Figura 9. Porcentaje de superficie de la celda ocupado por otros usos del suelo

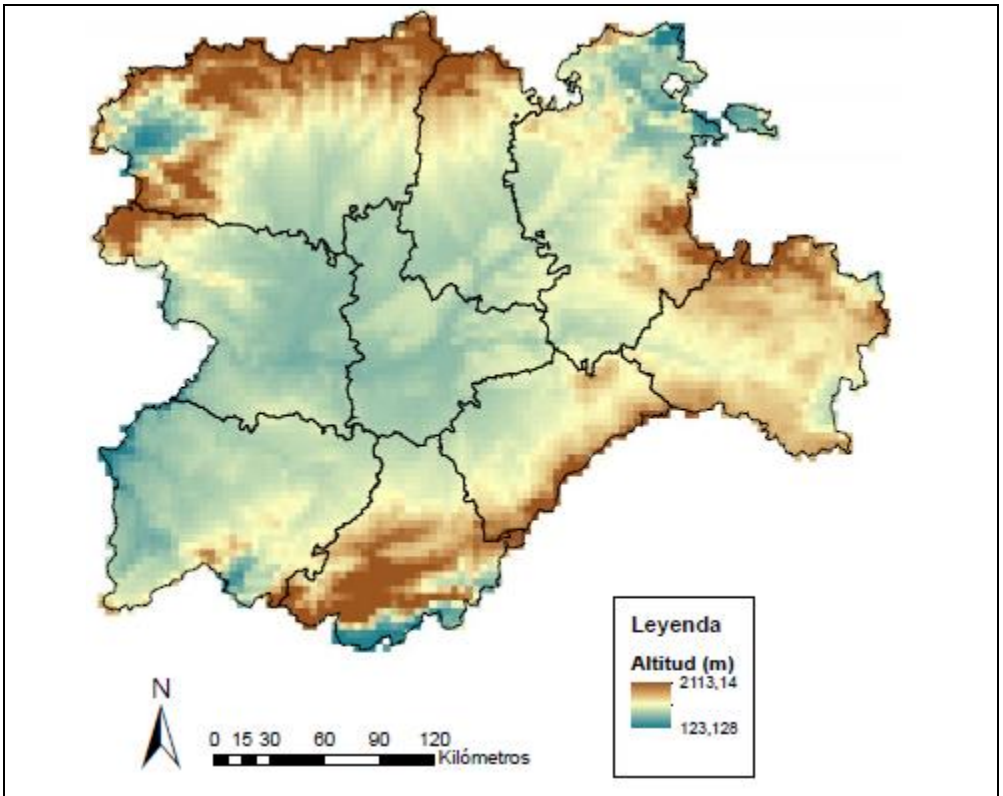


Figura 10. Altitud media de la celda en metros

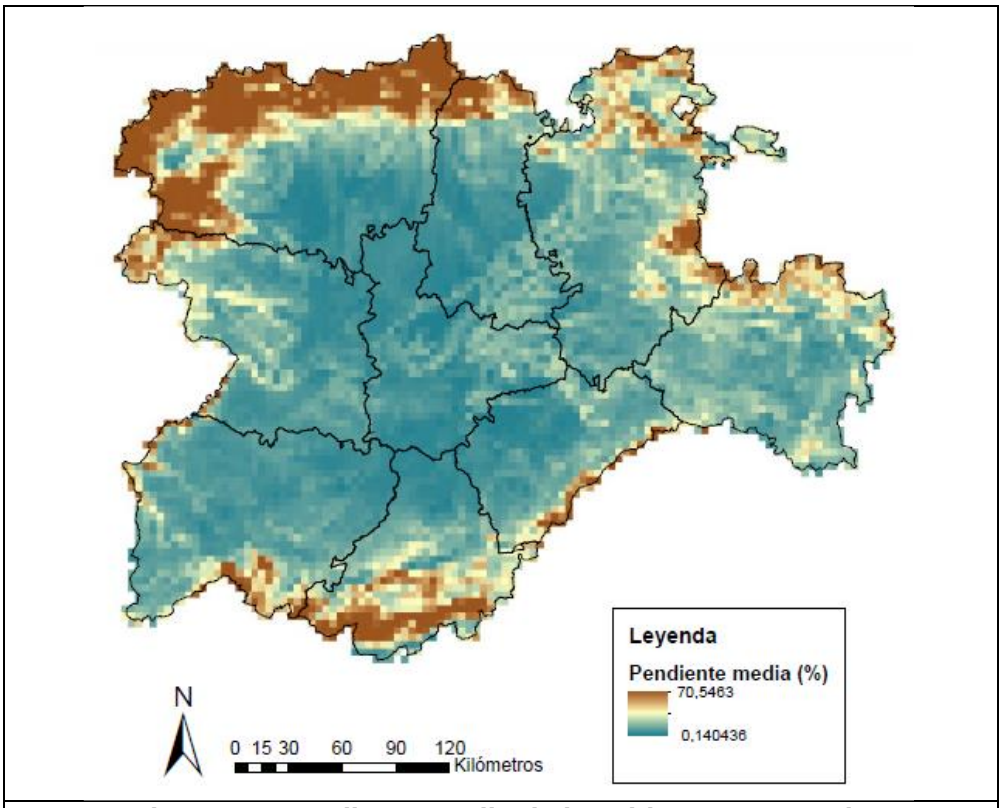


Figura 11. Pendiente media de la celda en porcentaje

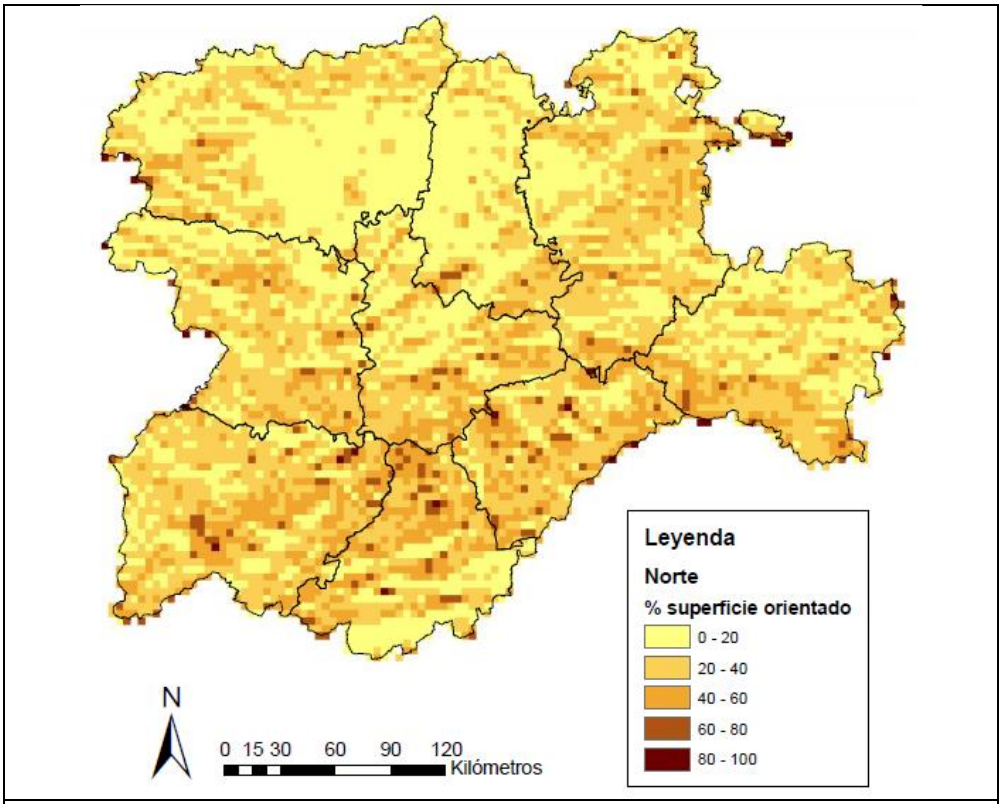


Figura 12. Porcentaje de superficie de la celda con orientación N

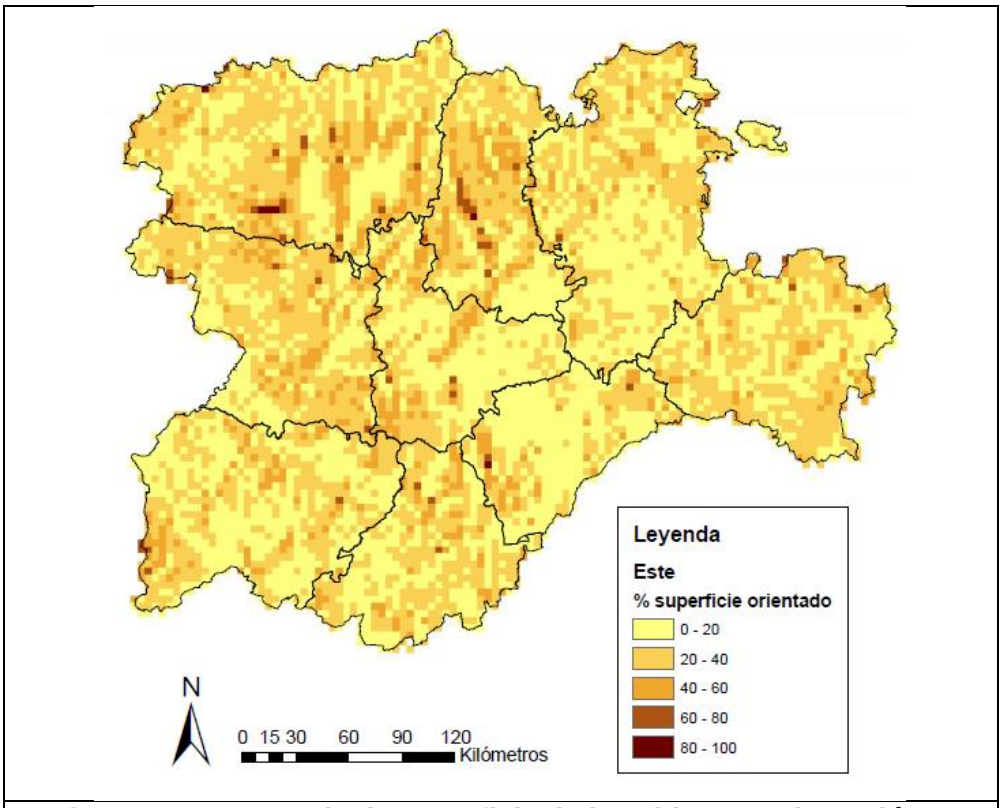


Figura 13. Porcentaje de superficie de la celda con orientación E

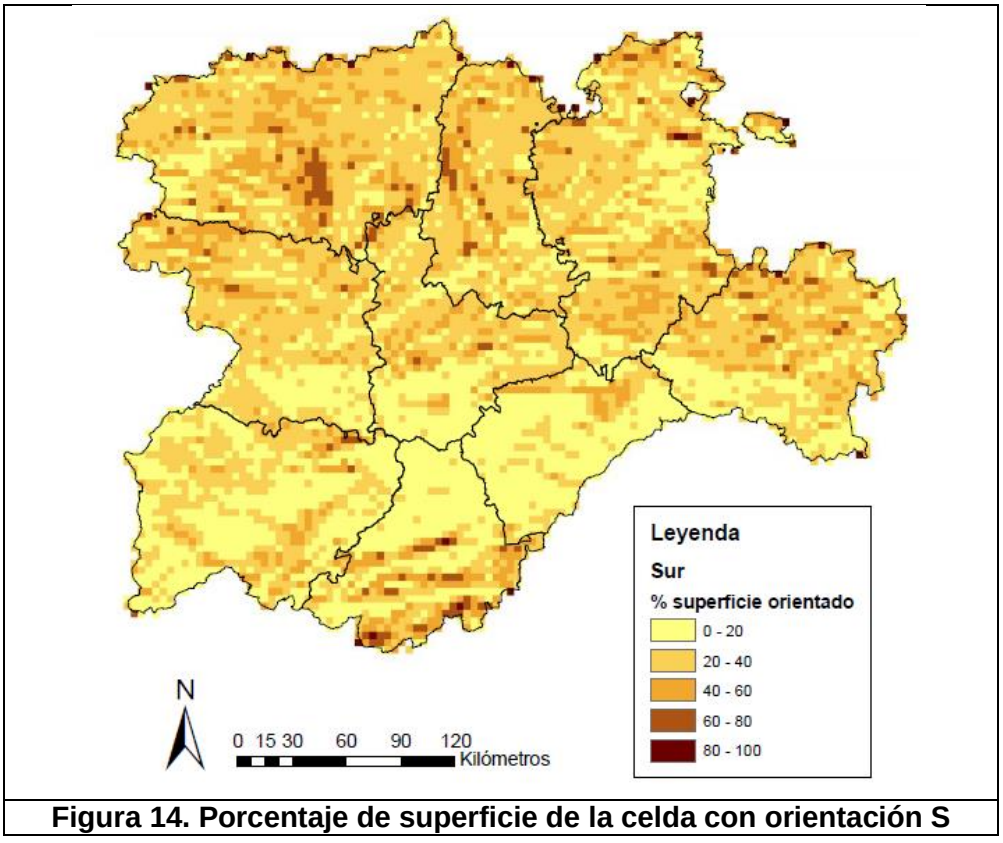


Figura 14. Porcentaje de superficie de la celda con orientación S

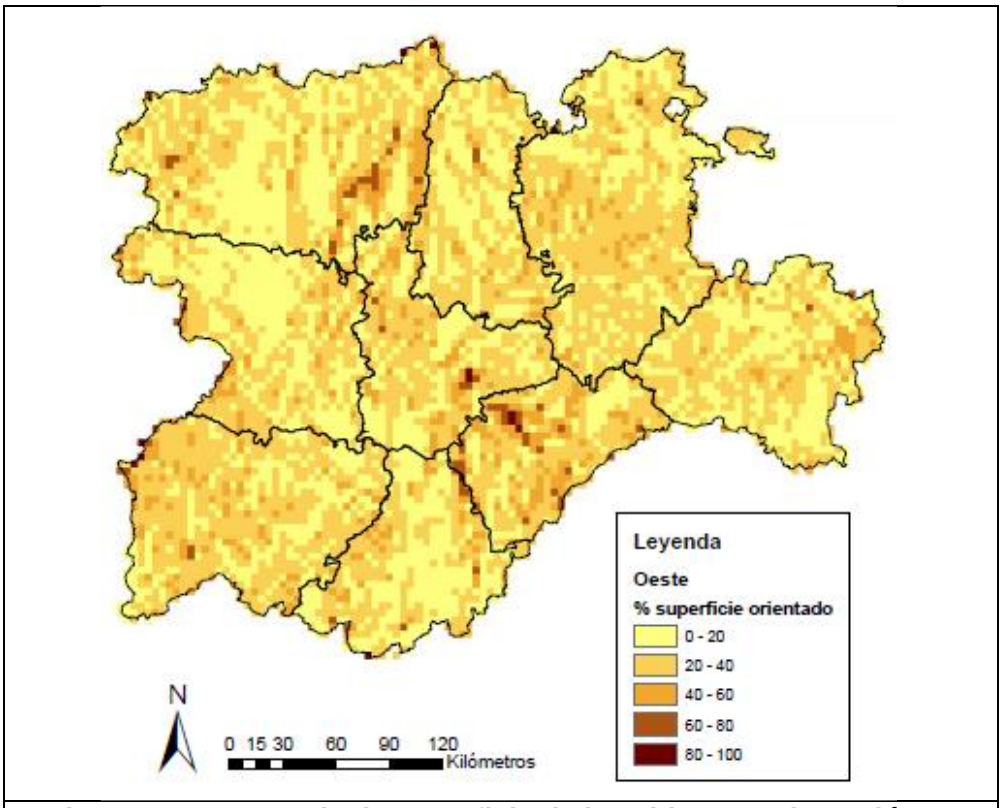


Figura 15. Porcentaje de superficie de la celda con orientación W

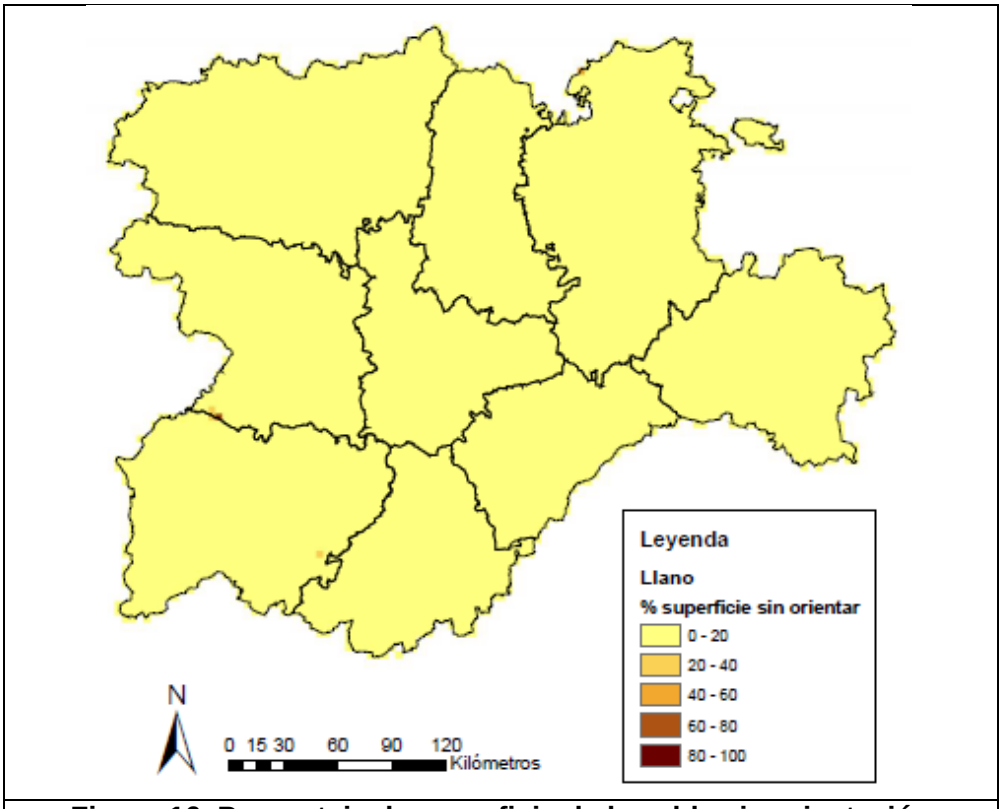


Figura 16. Porcentaje de superficie de la celda sin orientación

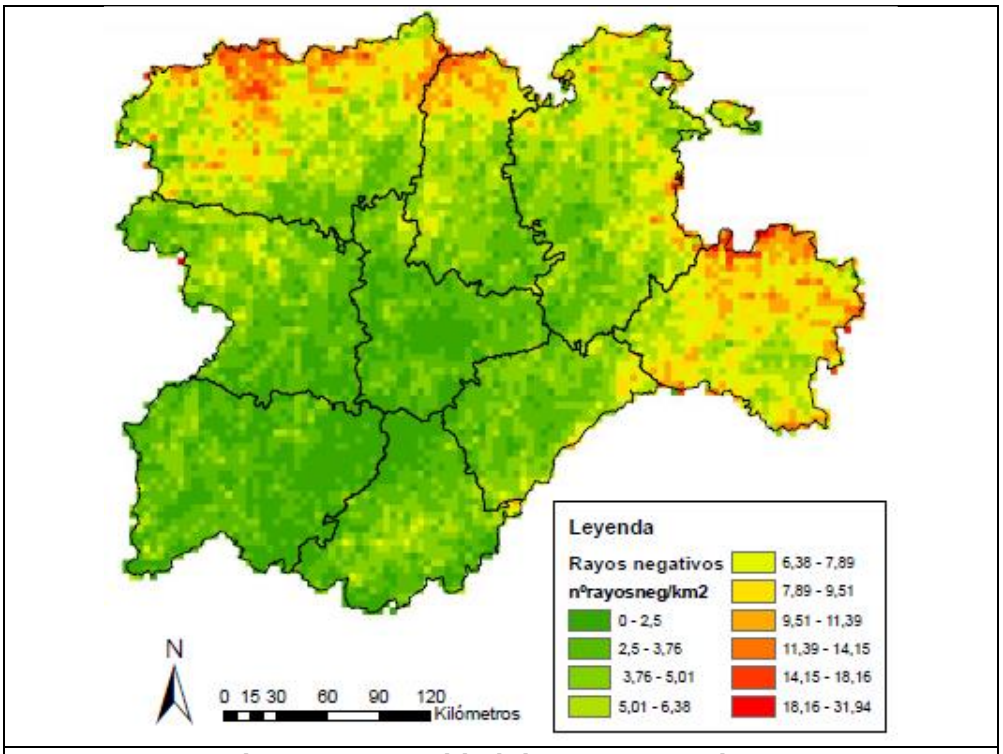


Figura 17. Densidad de rayos negativos

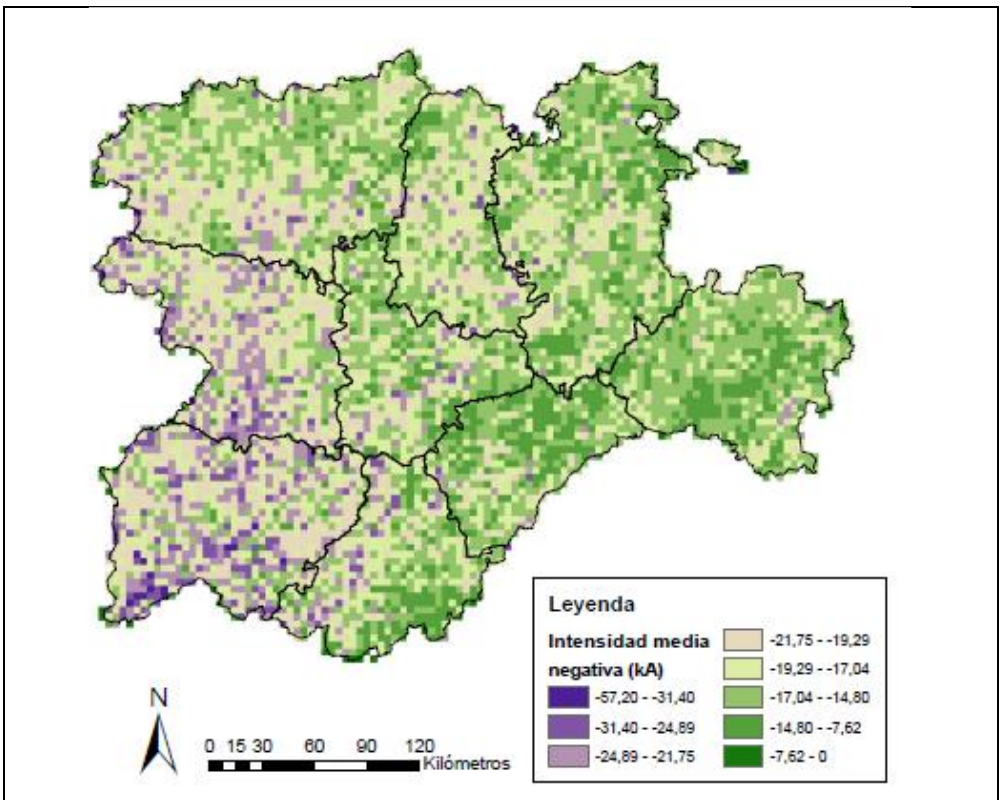


Figura 18. Intensidad media de rayos negativos

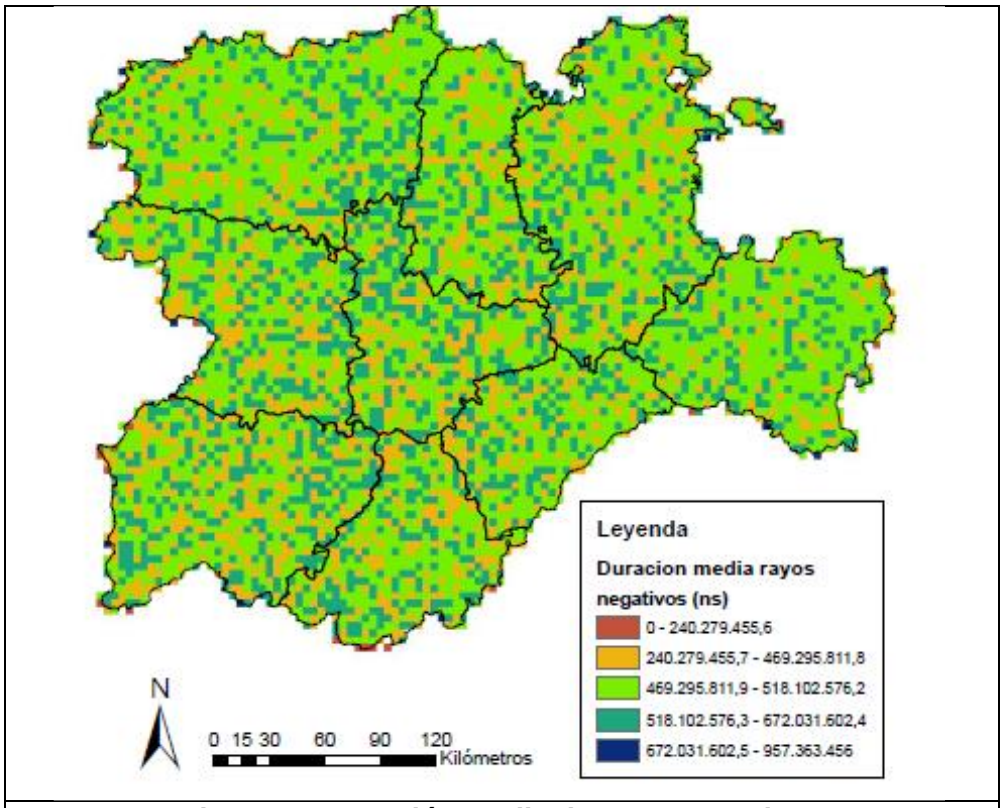


Figura 19. Duración media de rayos negativos

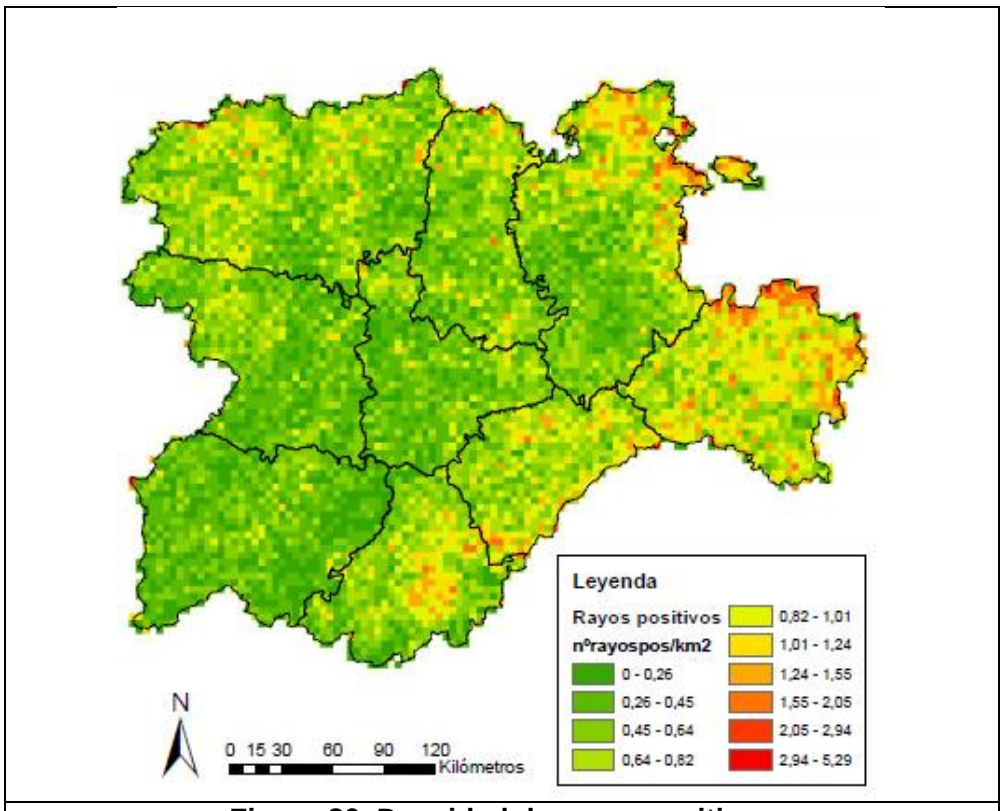


Figura 20. Densidad de rayos positivos

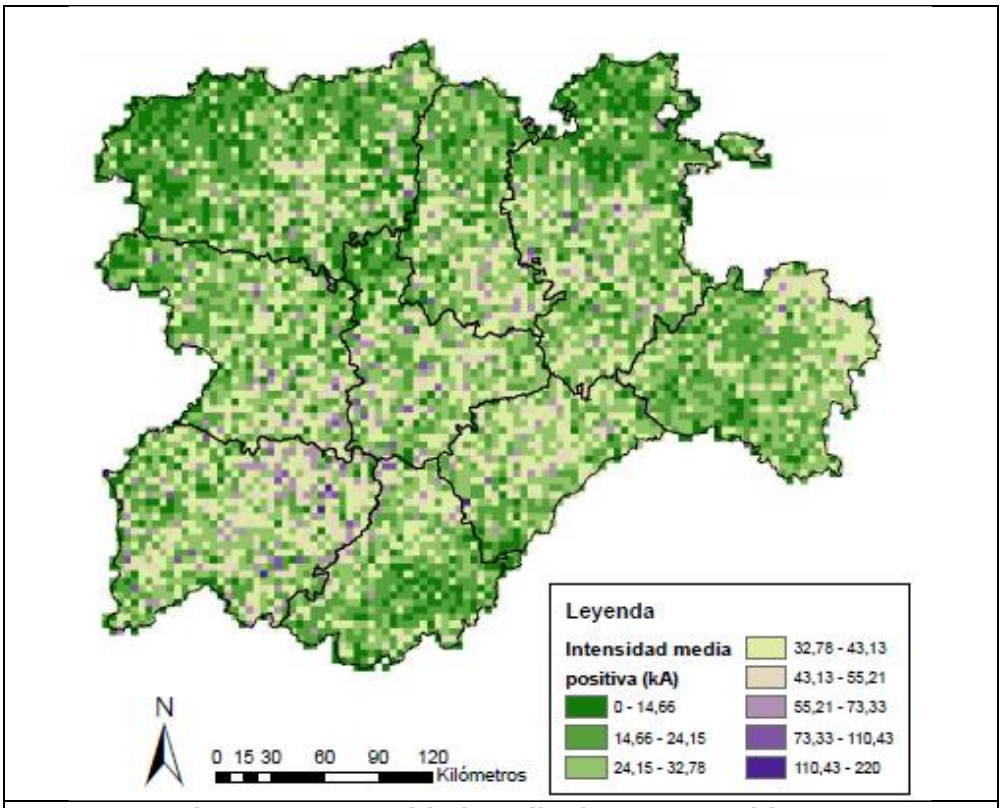


Figura 21. Intensidad media de rayos positivos

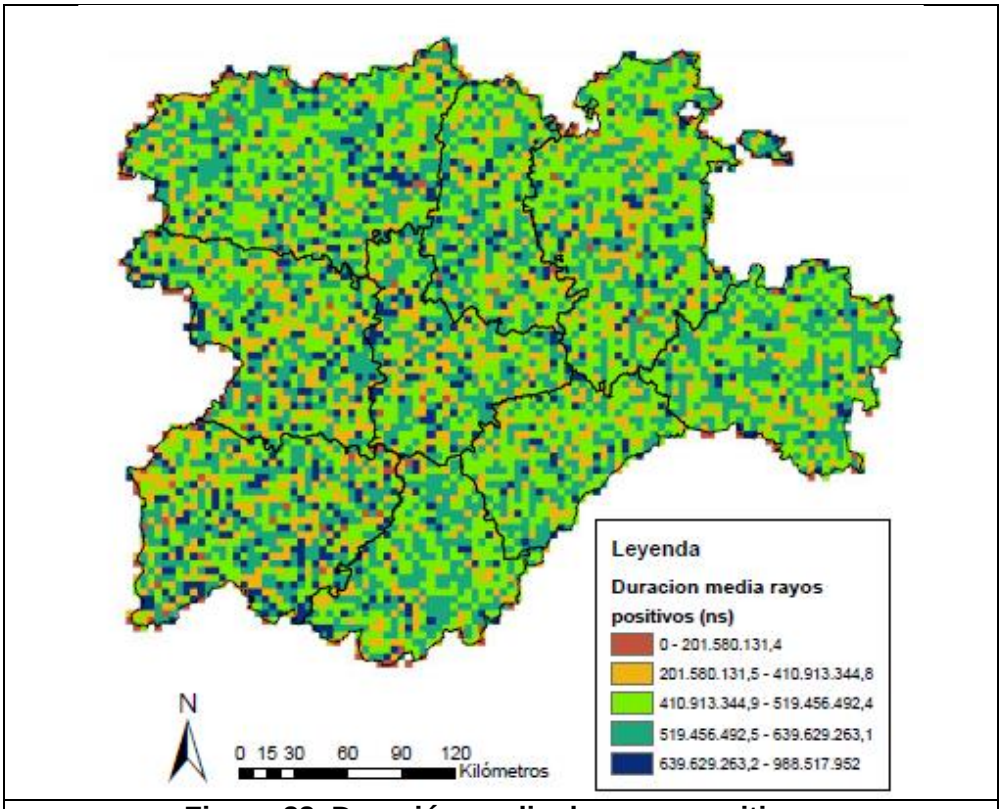


Figura 22. Duración media de rayos positivos

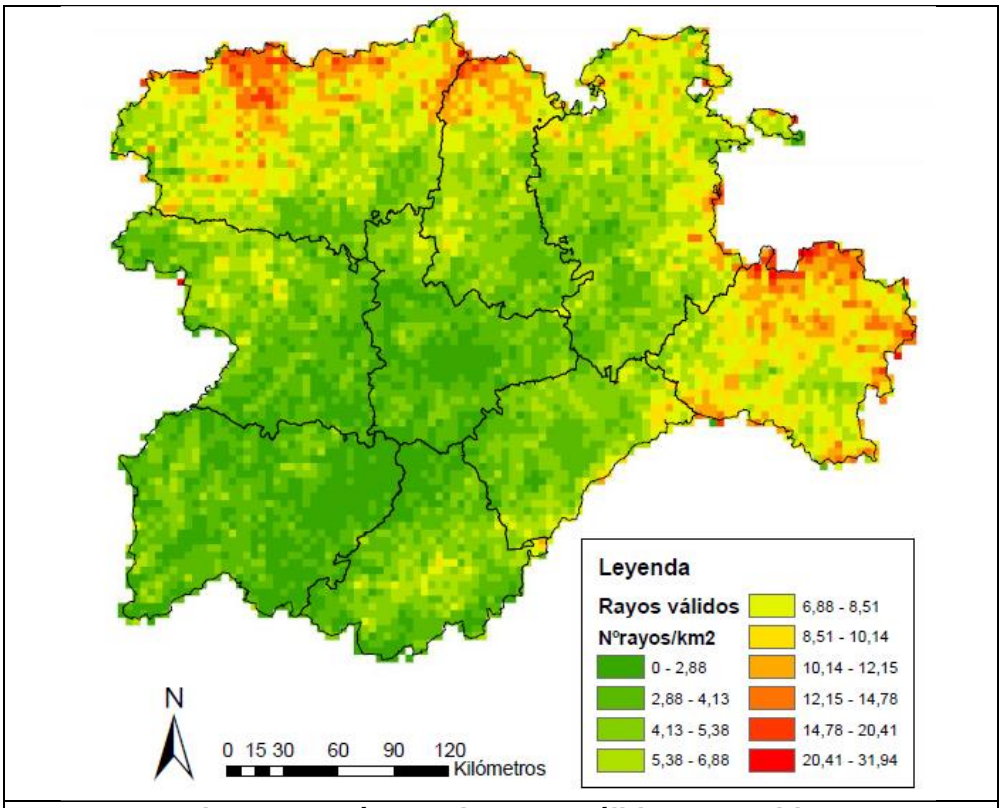


Figura 23. Número de rayos válidos por celda

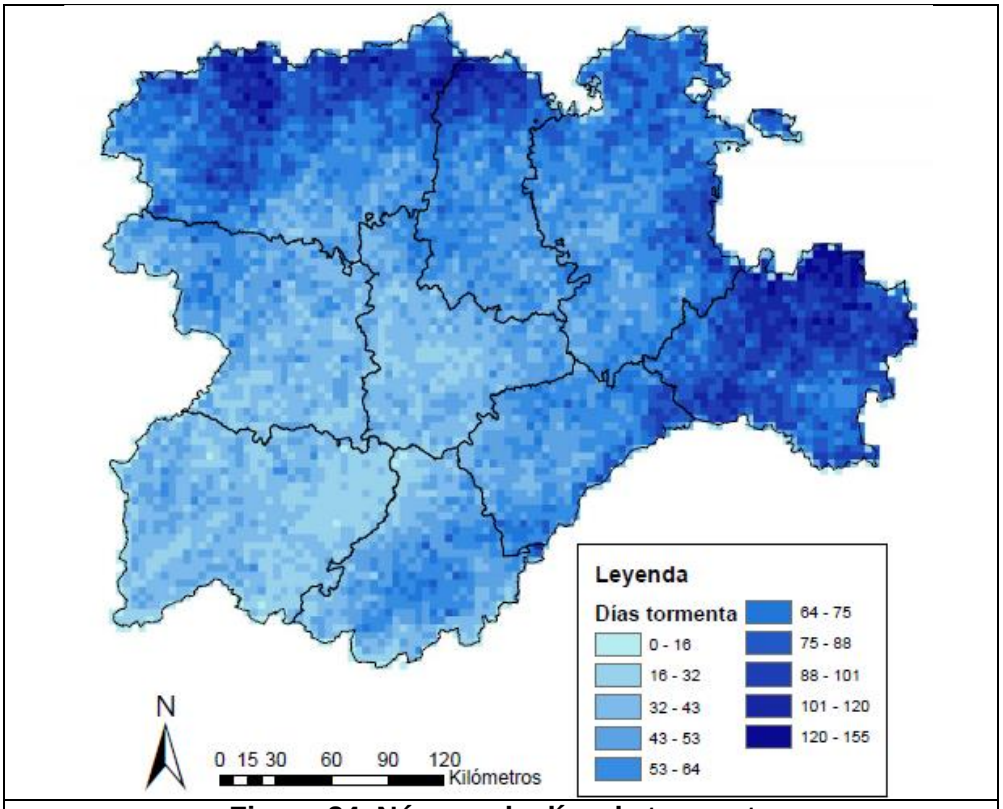


Figura 24. Número de días de tormenta

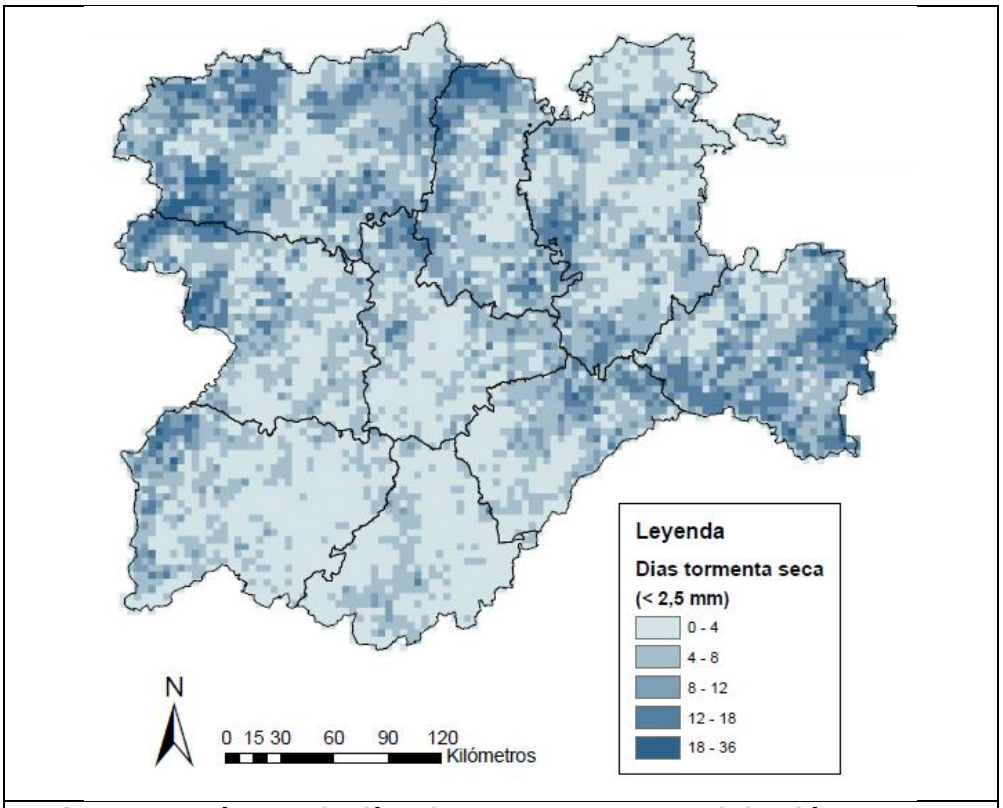


Figura 25. Número de días de tormenta con precipitación< 25 mm

3.1.3 COMPARATIVA VISUAL ENTRE VARIABLES

Previamente a la construcción del modelo matemático se realiza un pre-análisis visual con la intención de buscar asociaciones entre las variables explicativas y la presencia o ausencia de incendios forestales por rayo, se expone en las siguientes tablas:

Tabla 2. Comparativa presencia incendio/tipo vegetación

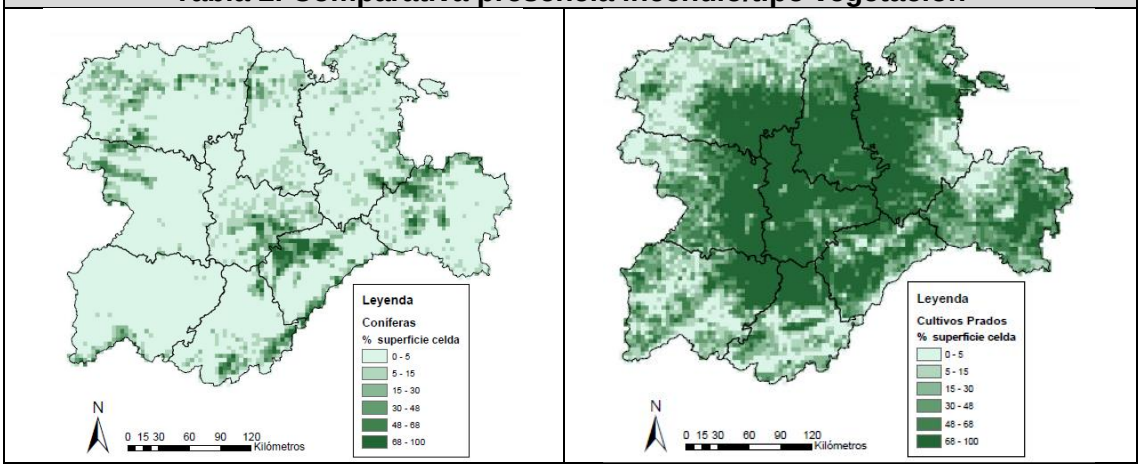
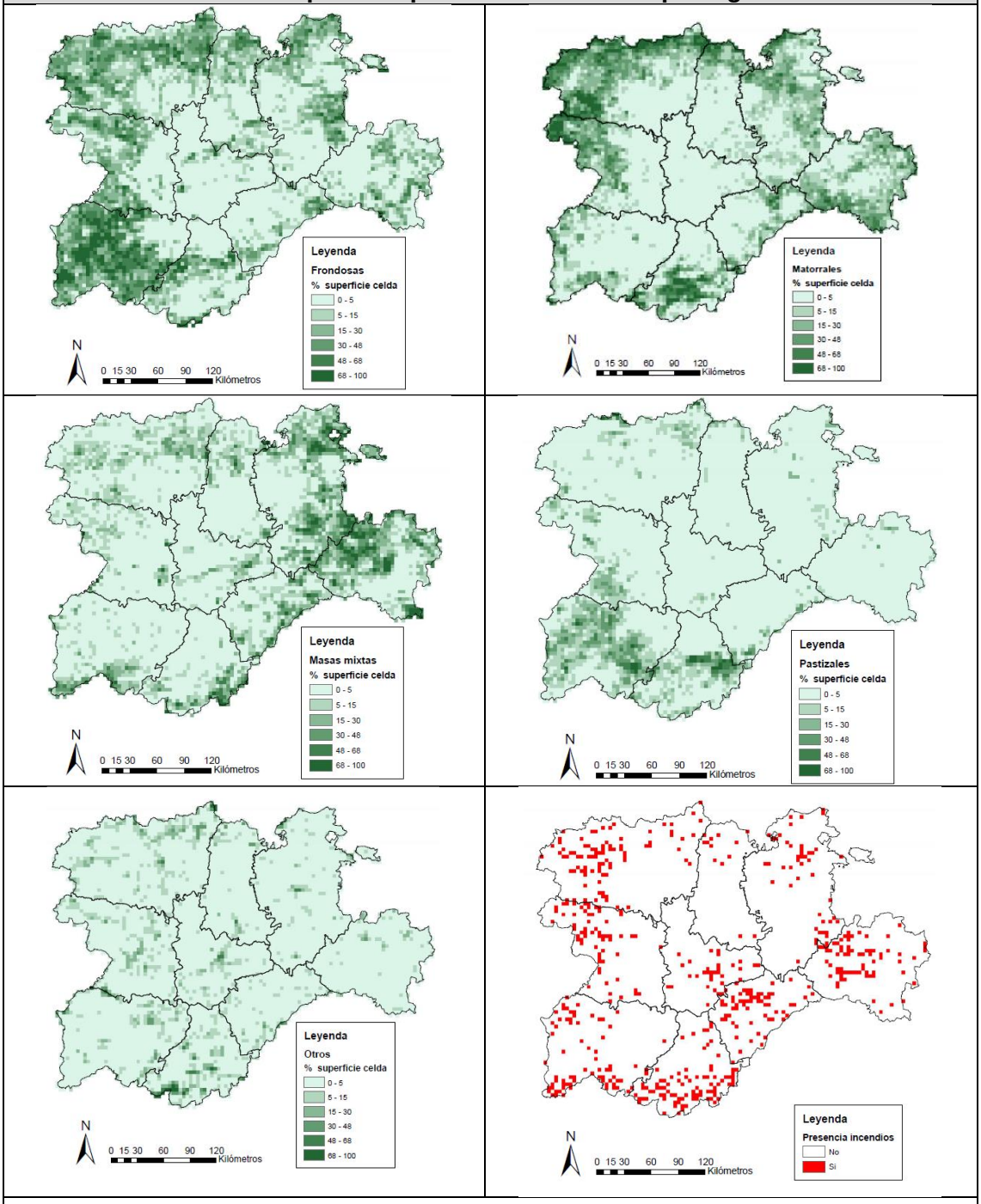
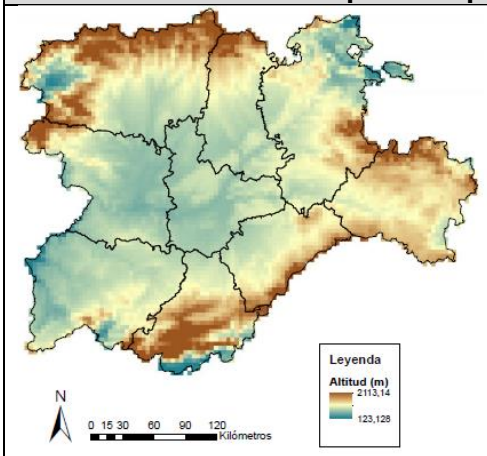


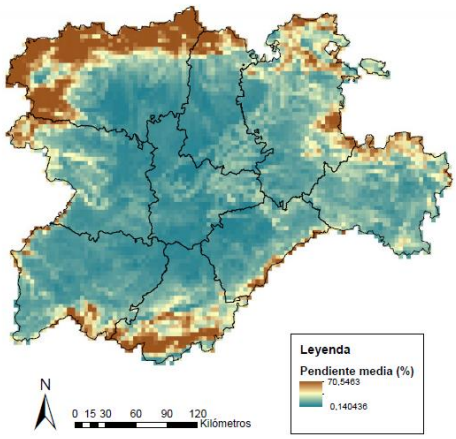
Tabla 2. Comparativa presencia incendio/tipo vegetación

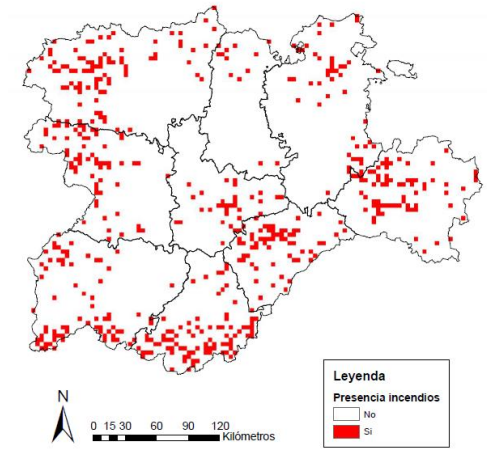


En un primer término y comparando el tipo de vegetación se puede afirmar que guardan cierta relación entre sí. Hay mayor proporción de incendios en zonas con vegetación “más robusta”, claramente se aprecia un descenso de la presencia de incendios en zonas de cultivos; así mismo, hay mayor presencia de incendios asociados a las superficies con mayor porcentaje de coníferas.

Tabla 3. Comparativa presencia incendio/altitud-pendiente

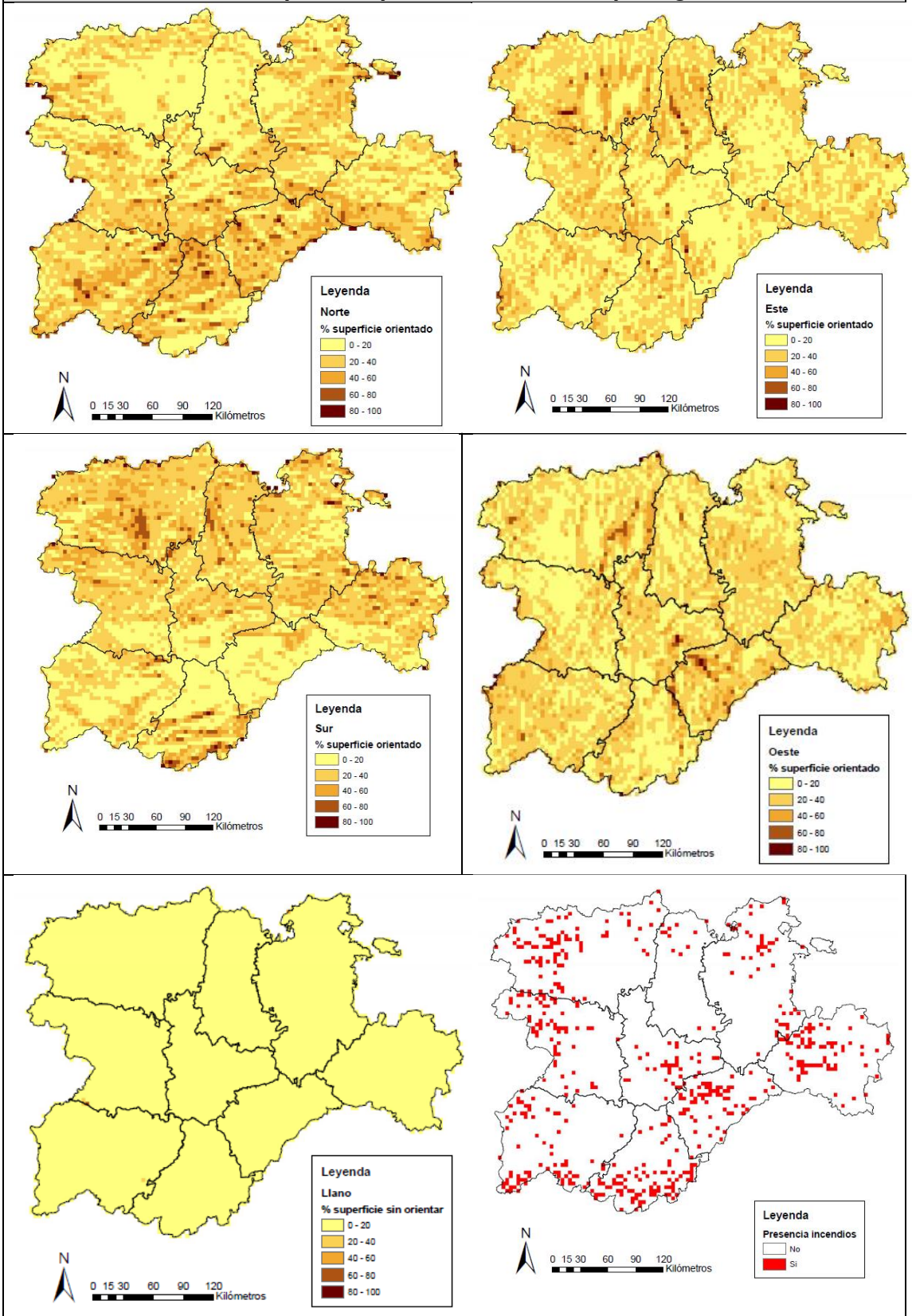






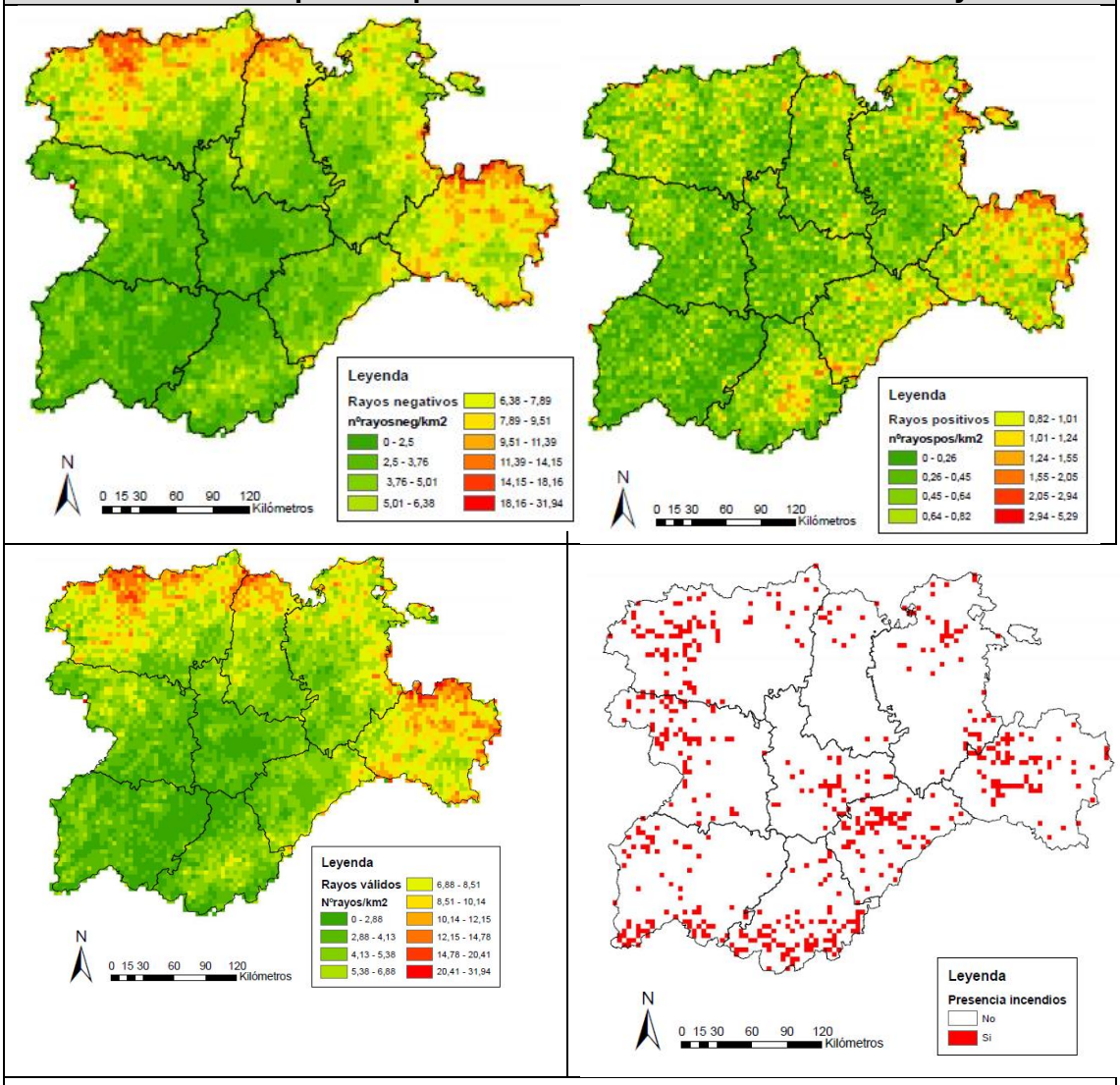
Las zonas de mayor altitud presentan mayores pendientes y parece que existe una cierta relación con la altitud. Existe mayor índice de incendios en zonas de mayor altitud a excepción de la zona central, entorno a Valladolid y Segovia, donde se producen incendios a menor altitud pero asociados a zonas con mayor proporción de coníferas.

Tabla 4. Comparativa presencia incendio/tipo vegetación



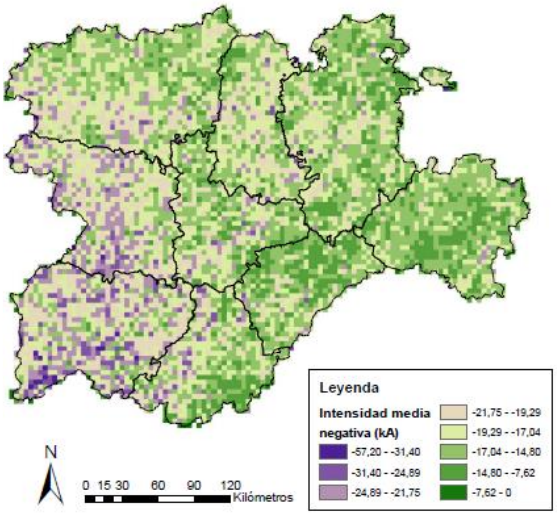
En un primer análisis visual no se detecta una relación especial entre la orientación del terreno y la presencia o ausencia de incendios forestales producidos por rayo.

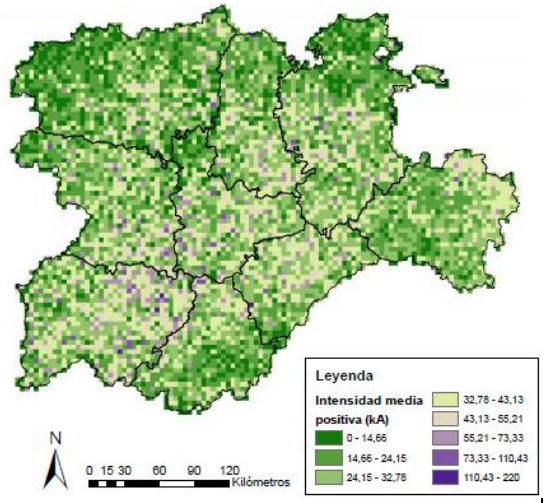
Tabla 5. Comparativa presencia de incendios/ Densidad de rayos

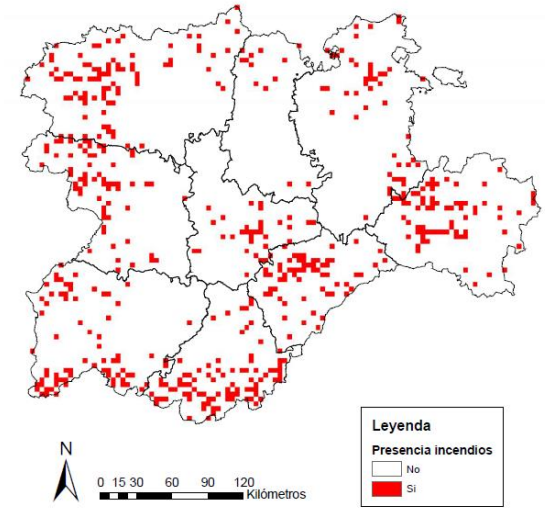


La mayor densidad de rayos válidos (en este caso negativos) se concentra en la zona más septentrional de la comunidad autónoma además de en la provincia más al oeste (Soria), extendiéndose la presencia de rayos positivos por zonas de Segovia y Ávila. Aunque no se observa una relación muy fuerte con la presencia de incendios sí se aprecia que, en las zonas más montañosas (hacia el norte), un aumento de densidad de rayos negativos se relaciona con la presencia de incendios.

Tabla 6. Comparativa de presencia de incendio/Intensidad media

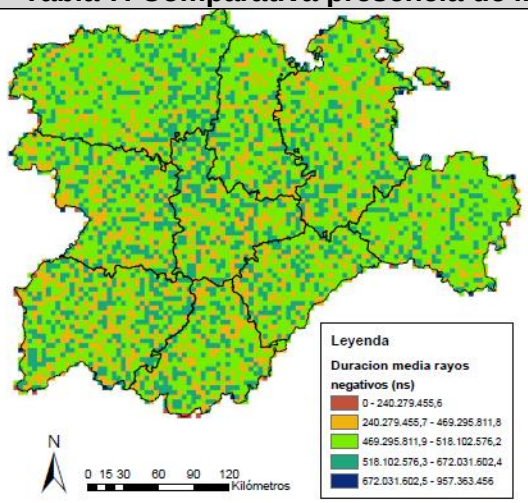


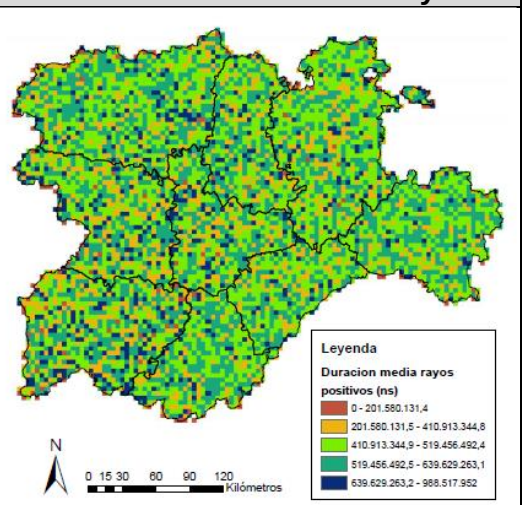


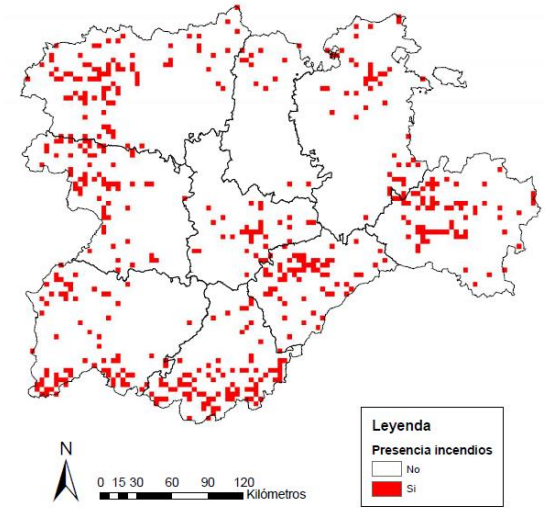


No se observa una relación fuerte entre la intensidad de los rayos y la presencia de incendio; si bien si se aprecia que existen mayores valores de intensidad media en zonas con presencia de incendios.

Tabla 7. Comparativa presencia de incendios/duración media de los rayos

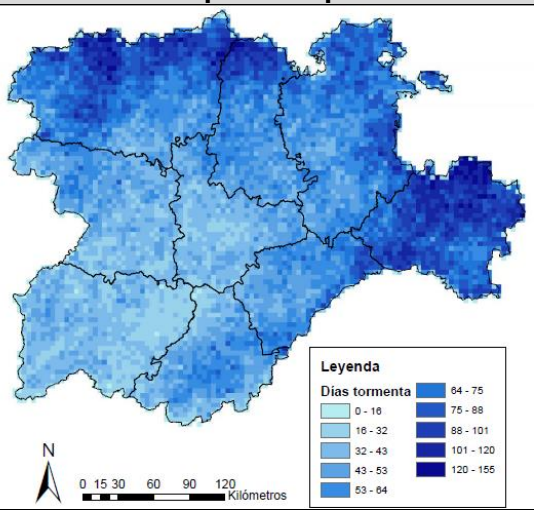


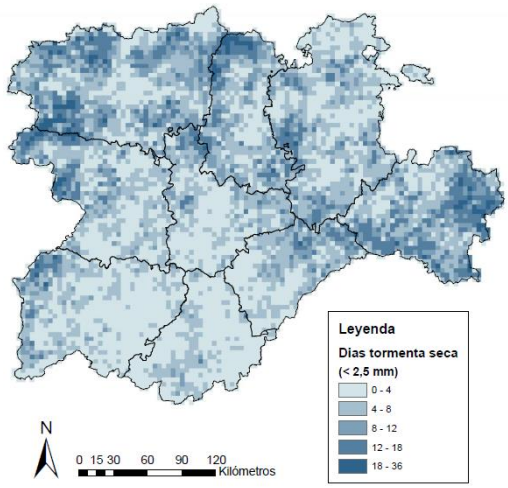


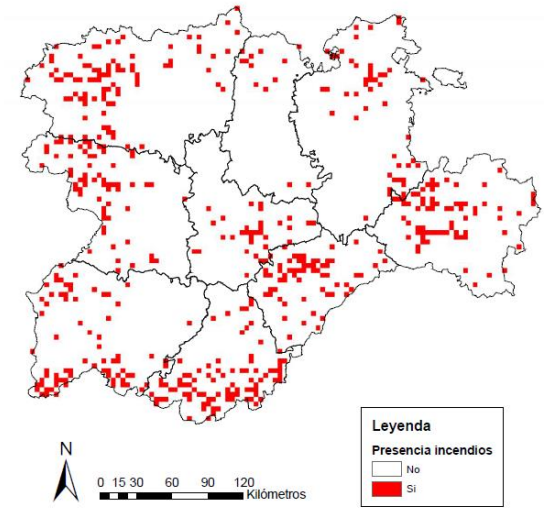


La duración media de los rayos es muy superior en el caso de los positivos pero para ninguno se observa relación con la presencia de incendio.

Tabla 8. Comparativa presencia de incendios/duración media de los rayos







Se observa una relación entre la presencia de tormentas e incendios, si bien la relación entre el número de días de tormenta y la presencia de incendio no es muy clara, sí parece que existe una relación más fuerte con el número de días de tormenta seca. Como ya se ha visto con otras variables, esto se observa o explica mejor en la zona noroeste de la comunidad.

3.2 METODOLOGÍA

Como se ha citado con anterioridad se busca obtener un modelo que explique o permita predecir la probabilidad de ocurrencia de incendio a causa de rayo. Para su consecución se realiza un tratamiento estadístico de las variables mediante el uso del paquete estadístico SPSS 22.

A grandes rasgos los trabajos a realizar son:

- Comprobación de la normalidad de las variables
- Correlación de las variables
- Regresión logística para extracción del modelo
- Validación del modelo

En los apartados a continuación se van detallando uno a uno.

Inicialmente se construyó un modelo con todos los datos obteniendo un modelo de muy baja sensibilidad que fallaba en la predicción de presencias de incendio. Esto es debido a que el número de celdas sin incendios está muy por encima del de celdas con presencia de incendios.

Por ello se procede a seleccionar una muestra aleatoria entre las celdas sin incendio del mismo tamaño que la muestra correspondiente a la presencia de incendios y se repite el modelo.

A continuación se incluyen los resultados obtenidos para la submuestra seleccionada que cuenta con un total de 1092 celdas de datos.

3.2.1 COMPROBACIÓN DE LA NORMALIDAD DE LAS VARIABLES

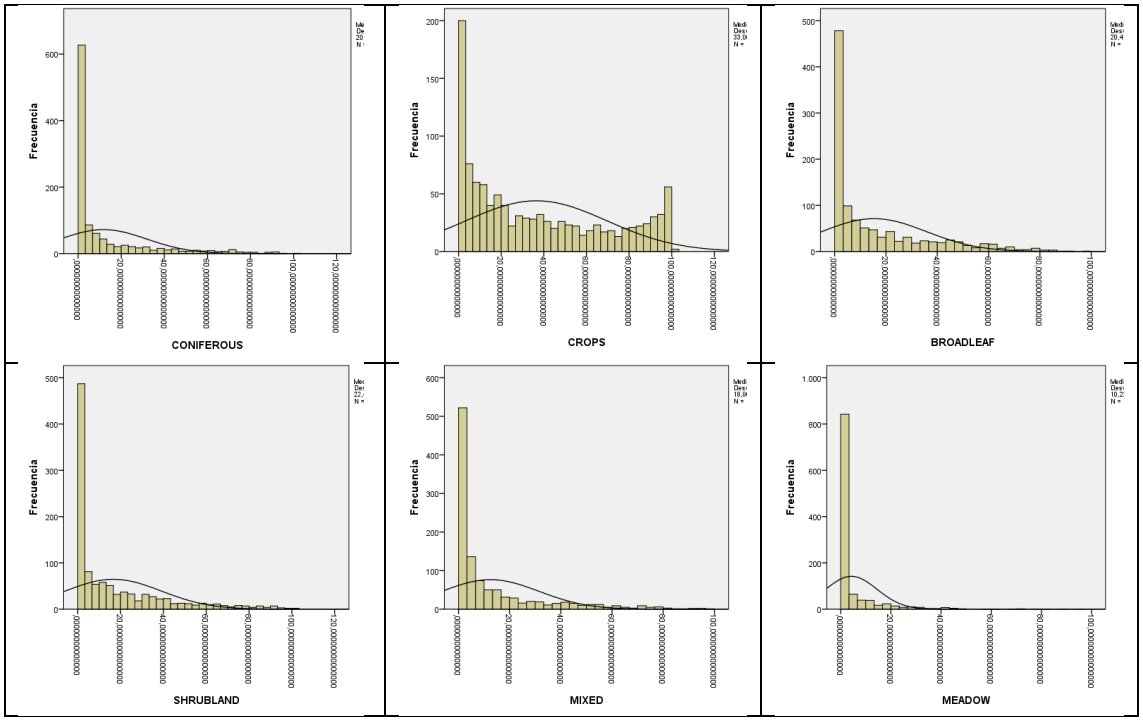
Para observar la distribución de los datos, en un primer término se determinan los estadísticos descriptivos:

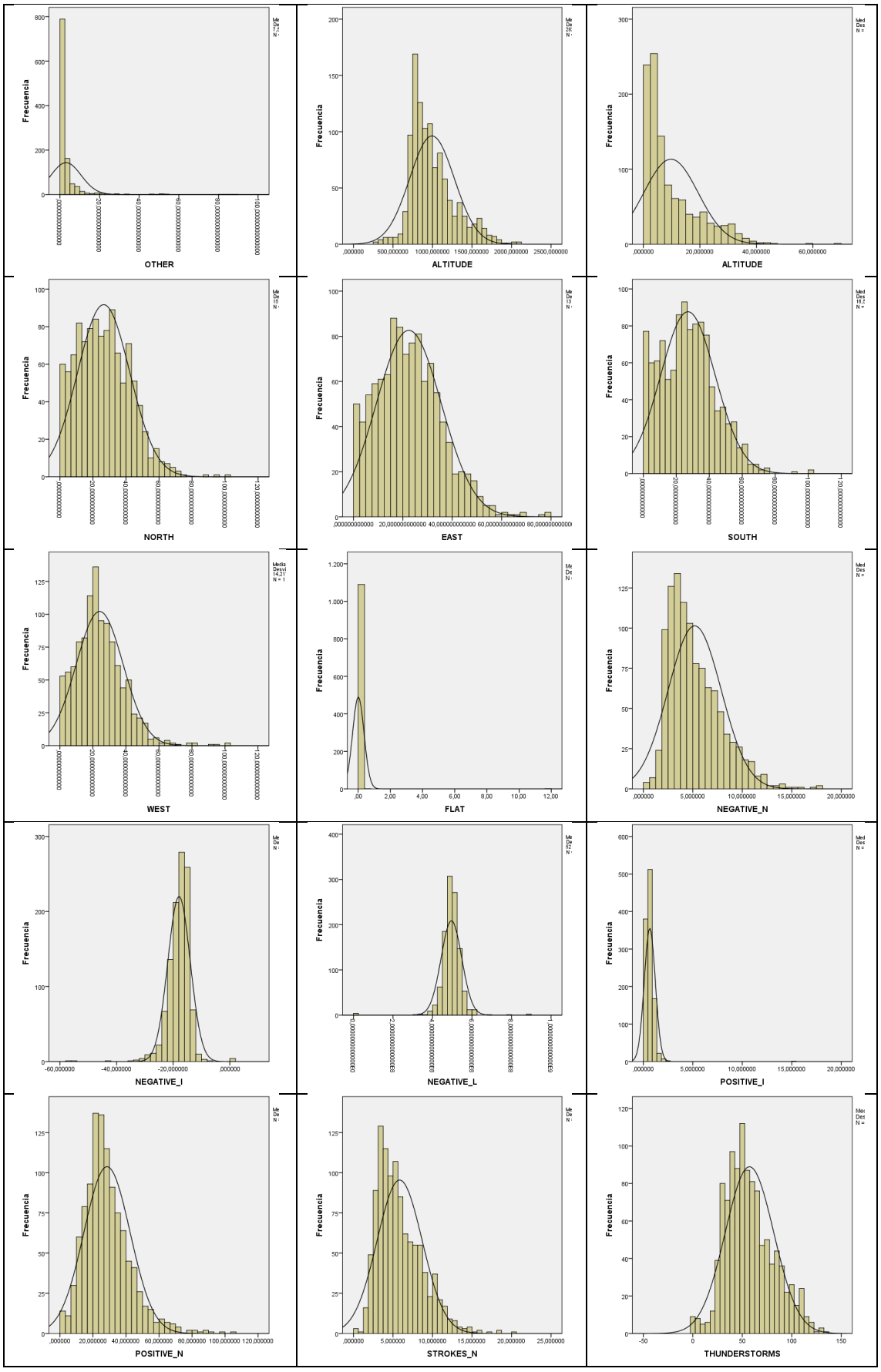
Tabla 9. Estadísticos descriptivos

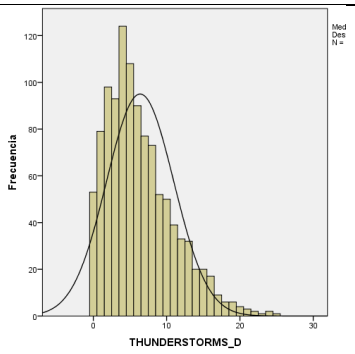
	N	Rango	Mínimo	Máximo	Media	Desviación estándar	Varianza	Asimetría	Curtosis
Coníferas	1092	100	0	100	11,75	20,15	406,040	2,105	3,841
Cultivos	1092	100	0	100	36,51	33,06	1093,079	,587	-1,061
Frondosas	1092	98,95	0	98,95	15,35	20,42	416,816	1,487	1,417
Matorrales	1092	100	0	100	16,59	22,47	505,242	1,629	2,027
Mixtas	1092	97,44	0	97,44	12,47	18,97	359,821	2,021	3,696
Pastizales	1092	94,96	0	94,96	4,25	10,22	104,568	4,240	23,163
Otros	1092	88,21	0	88,21	3,07	7,59	57,645	6,965	60,694
Altitud	1092	1801,42	271,45	2072,87	993,58	282,94	80058,919	,953	,990
Pendiente	1092	68,40	0,22	68,62	9,77	9,61	92,309	1,510	2,380
Norte	1092	100	0	100	26,48	15,82	250,438	,529	,252
Este	1092	78	0	78	22,40	13,18	173,633	,582	,468
Sur	1092	100	0	100	27,03	16,56	274,250	,484	,189
Oeste	1092	100	0	100	24,08	14,22	202,145	1,011	2,613
Llano	1092	11,75	0	11,75	0,01	0,35	0,127	32,958	1088,048
Densidad rayos negativos	1092	18	0	18	5,19	2,68	7,211	1,085	1,472
Intensidad rayos negativos	1092	56,40	-56,40	0,00	-17,92	3,96	15,672	-2,041	17,633
Duración rayos negativos	1092	880125100	0	880125100	495781072,88	52191013,25	2723901864466780	-2,273	34,510
Densidad rayos positivos	1092	15,37	0	15,367639	,64943584	0,56	0,311	17,089	444,509

Tabla 9. Estadísticos descriptivos									
	N	Rango	Mínimo	Máximo	Media	Desviación estándar	Varianza	Asimetría	Curtosis
Intensidad rayos positivos	1092	103,50	0	103,50	28,50	13,99	195,825	1,129	2,674
Duración rayos positivos	1092	969005800	0	969005800	498427589,702	122777003,05	150741924787561	-0,712	3,325
Densidad rayos válidos	1092	20,51	0	20,51	5,84	2,85	8,134	1,045	1,393
Días de tormenta	1092	137	0	137	57,30	24,50	600,266	0,482	-0,032
Días tormenta seca	1092	25	0	25	6,38	4,58	21,022	0,958	0,769

También se generan los histogramas de las distribuciones con curvas de densidad:




Frecuencia
OTHER
ALTITUDE
ALTITUDE
NORTH
EAST
SOUTH
WEST
FLAT
NEGATIVE_N
NEGATIVE_I
NEGATIVE_L
POSITIVE_I
POSITIVE_N
STROKES_N
THUNDERSTORMS



Para comprobar las características de la distribución de las variables se realiza la prueba de normalidad de Kolmogorov-Smirnov:

Tabla 10. Pruebas de normalidad

Variables	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
Coníferas	0,280	1092	0,000
Cultivos	0,136	1092	0,000
Frondosas	0,226	1092	0,000
Matorrales	0,230	1092	0,000
Mixtas	0,255	1092	0,000
Pastizales	0,339	1092	0,000
Otros	0,343	1092	0,000
Altitud	0,103	1092	0,000
Pendiente	0,178	1092	0,000
Norte	0,048	1092	0,000
Este	0,045	1092	0,000
Sur	0,051	1092	0,000
Oeste	0,062	1092	0,000
Llano	0,511	1092	0,000
Densidad rayos negativos	0,102	1092	0,000
Intensidad rayos negativos	0,084	1092	0,000
Duración rayos negativos	0,102	1092	0,000
Densidad rayos positivos	0,178	1092	0,000
Intensidad rayos positivos	0,077	1092	0,000
Duración rayos positivos	0,064	1092	0,000
Densidad rayos válidos	0,098	1092	0,000
Días de tormenta	0,075	1092	0,000
Días tormenta seca	0,127	1092	0,000

a. Corrección de significación de Lilliefors

La hipótesis nula del test de Kolmogorov-Smirnov asume que la variable a contrastar se distribuye de forma normal, frente a la alternativa que niega esta característica. El grado de significación que se obtiene (probabilidad de error al rechazar la hipótesis nula) al realizar el test para las variables es 0,000 por lo que se rechaza la hipótesis de

normalidad para todas (Álvarez E., *et al*. 2011), dato que nos será de interés para hacer la correlación de las variables.

3.2.2 CORRELACIÓN DE VARIABLES

Una vez comprobada la no-normalidad se considera usar el coeficiente de correlación de Spearman, ρ (rho), que describe la intensidad de relación entre dos conjuntos de variables aleatorias continuas. Los valores que resultan fluctúan entre -1 y +1, indicando el signo si las asociaciones entre variables son negativas o positivas; si el valor es 0 no existe correlación.

En términos generales se puede determinar que el tipo de asociación o correlación en función de rho es:

- $\rho < 0,3$ Asociación débil
- $0,3 \leq \rho \leq 0,7$ Asociación moderada
- $\rho > 0,7$ Asociación fuerte

Obtenidos y estudiados los resultados de ρ entre pares de variables, se determina eliminar una de las del par que tenga un coeficiente de correlación mayor o igual a 0,7. Los pares de variables que cumplen esta condición son:

Tabla 11. Coeficientes de correlación

Pares de variables	Coeficiente de Spearman (ρ)
Días de tormenta / Densidad de rayos negativos	0,850
Días de tormenta / Densidad de rayos válidos	0,857
Densidad de rayos negativos / Días de tormenta	0,987

Como el objetivo es obtener el mínimo número de variables que nos permitan definir el modelo, en este paso se decide eliminar el mayor número de variables por lo que a partir de este punto se elimina del estudio las variables: DENSIDAD DE RAYOS VÁLIDOS (STROKES_N) y DENSIDAD DE RAYOS NEGATIVOS (NEGATIVE_N).

3.2.3 CONSTRUCCIÓN DEL MODELO CON REGRESIÓN LOGÍSTICA

La regresión logística binaria resulta útil cuando se quiere predecir la presencia o ausencia de un suceso (como puede ser un incendio) en función de los valores de un conjunto de variables predictoras.

Permite introducir diversas variables para analizarlas una a una así como las interacciones entre ellas. Se trata de ajustar la ecuación en la que la variable dependiente ha de ser dicotómica (0 o 1, ausencia o presencia del incendio) y las variables independientes o explicativas pueden ser continuas o categóricas (en nuestro caso continuas), siendo β_0 y β_i los parámetros del modelo a ajustar. Se busca obtener de entre todos los modelos posibles, el que genere una predicción más precisa con el menor número de variables.

La ecuación que provee la probabilidad de ocurrencia de un suceso según una serie de variables de entrada es:

$$P\ (presencia\ o\ Y = 1) = \frac{1}{1 + e^{(-\beta 0 - \beta 1 * X1 - \cdots \beta i * Xi)}}$$

El proceso a realizar se indica a continuación, para su consecución se utiliza el programa SPSS 22.

3.2.3.1 VARIABLES CON SIGNIFICANCIA AL 95%

El primer paso es realizar el análisis de regresión logística binaria con el objetivo de encontrar las variables que son significativas al 95%. Para llevar a cabo el ajuste estadístico con SPSS se utiliza el procedimiento *INTRODUCIR.*

Este procedimiento se utiliza para que el operador pueda controlar las variables que se introducen o excluyen del modelo, pues existen otros métodos automáticos donde se incluyen todas las variables y luego va extrayendo una a una (*ATRÁS*), o al revés (*ADELANTE*).

Los motivos por los que se elige dicho método son:

- Para introducir en el modelo las variables realmente predictoras, aquí entra en juego el conocimiento del tema y la revisión de trabajos previos.
- Evitar introducir variables de confusión, que si bien son variables predictoras, son ajenas a la relación principal en análisis y conllevan un error al evaluar la relación entre la variable dependiente y las independientes. El conocimiento de estas variables también viene dado por el conocimiento y revisión de literatura.

Así se procede a introducir inicialmente todas las variables, y una vez analizados los resultados, se tiene que las variables con significancia al 95% son:

Tabla 12. Variables significativas en regresión logística

Variables	B	Error estándar	Wald	gl	Sig.
CROPS (Cultivos)	-0,029	0,010	8,522	1	0,004
ALTITUDE (Altitud)	-0,001	0,000	15,727	1	0,000
THUNDERSTORMS (Días tormenta)	0,011	0,004	7,767	1	0,005

3.2.3.2 ANÁLISIS DEL AJUSTE DEL MODELO

Se realiza otro análisis mediante regresión logística binaria donde la variable dependiente es la presencia o ausencia de incendio y las variables explicativas son las extraídas del paso anterior (cultivos, altitud y días de tormenta). El método de entrada de las variables en el modelo ha sido el denominado *Introducir* o *Entrar*, o sea que el operador lo realiza “manualmente”, no se deja de mano del programa la elección de las variables que compondrán el modelo.

A continuación se muestra la salida y descripción de resultados.

Tabla 13. Resumen de procesamiento de casos

Casos sin ponderar		N	Porcentaje
Casos seleccionados	Incluido en el análisis	1092	100,0
	Casos perdidos	0	0,0
	Total	1092	100,0

Tabla 13. Resumen de procesamiento de casos		
Casos sin ponderar	N	Porcentaje
Casos no seleccionados	0	0,0
Total	1092	100,0

En la Tabla 13 se indica el número de casos introducidos, los seleccionados para el análisis y los excluidos (casos perdidos).

Tabla 14. Codificación de variable dependiente	
Valor original	Valor interno
0 (Ausencia)	0
1 (Presencia)	1

Seguidamente proporciona una tabla (14) que indica la codificación que adopta el programa para la variable dependiente. En este caso coincide con la codificación de la base de datos de partida.

Otros datos de interés que nos proporciona el programa se observan en la tabla 15 donde se aportan tres medidas resumen del modelo, que servirán para evaluar su validez de forma global.

En primer lugar se tiene -2LL (-2 Logaritmo de la verosimilitud), que mide hasta qué punto un modelo se ajusta bien a los datos; y dos coeficientes de determinación R^2 que expresan la proporción en tanto por uno de la variación explicada por el modelo.

Tabla 15. Resumen del modelo			
Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1364,221[a]	0,128	0,171

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001

Un modelo perfecto tendría un -2LL pequeño y R^2 cercano a uno. En este caso la información que se extrae es que el 17,1% de la variación de la variable dependiente es explicada por las incluidas en el modelo.

A continuación se tiene una prueba de ajuste global del modelo, prueba de Hosmer y Lemeshow que valora la bondad del ajuste del modelo de regresión logística (Tabla 16). Nos da la significancia de R^2 para la variable dependiente (es significativo).

Tabla 16. Prueba de Hosmer y Lemeshow			
Escalón	Chi^2	gl	Sig.
1	20,054	8	0,010

Si el ajuste es bueno un valor alto de la probabilidad predicha se asocia con el resultado 1 de la variable binomial dependiente, mientras que uno bajo lo hará normalmente con el valor 0. Para cada observación del conjunto de datos, intenta calcular las probabilidades de la variable dependiente que predice el modelo, ordenarlas, agruparlas y calcular con ellas las frecuencias esperadas y compararlas con las observadas por una prueba Chi^2.

Esta prueba muestra inconvenientes, como que no se computa para grupos de valores esperados nulos o muy pequeños. Además, se busca que no exista significancia por lo que muchos autores proponen comparar los valores observados y los esperados mediante inspección visual (Tabla 17. De contingencia).

Tabla 17. Tabla de contingencia para la prueba de Hosmer y Lemeshow						
		FIRE = 0		FIRE = 1		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	94	88,922	15	20,078	109
	2	91	82,148	18	26,852	109
	3	56	71,173	53	37,827	109
	4	60	60,310	49	48,690	109
	5	46	52,951	63	56,049	109
	6	42	47,360	67	61,640	109
	7	48	42,111	61	66,889	109
	8	41	37,763	68	71,237	109
	9	35	33,920	74	75,080	109
	10	33	29,342	78	81,658	111

Una manera de evaluar la ecuación de regresión y el modelo obtenido es mediante la realización de una tabla 2x2 de clasificación o matriz de errores donde se clasifica toda la muestra según la correspondencia de valores observados con los estimados por el modelo.

Una ecuación sin poder de clasificación tendría una especificidad, sensibilidad y fiabilidad total de clasificación igual al 50% por simple azar (AGUAYO, M., 2007).

Tabla 18. Tabla de clasificación[a]					
			Pronosticado		
			FIRE		Corrección de porcentaje
	Observado		0	1	
Paso 1	FIRE	0	312	234	57,1
		1	149	397	72,7
	Porcentaje global				64,9

a. El valor de corte es 0,500

Con la tabla de clasificación (Tabla 18) se comprueba que el modelo predice bien el 57,1% de las ausencias de incendio (especificidad) y el 72,7% de las presencias de incendio (sensibilidad) y tiene una fiabilidad global del 64,9%.

3.2.3.3 ECUACIÓN DE PROBABILIDAD

Finalmente el programa proporciona los coeficientes de regresión de las variables de la ecuación (B en el programa β_i en nuestra ecuación), sus errores estándar, el valor del estadístico de Wald para evaluar la hipótesis nula ($\beta_i=0$), la significación estadística asociada y el valor odd ratio (Exp (B)) con sus intervalos de confianza (Tabla 19).

Tabla 19. Variables en la ecuación								
	B	Error estándar	Wald	gl	Sig.	Exp (B)	95% CI para Exp (B)	
							Inferior	Superior
CROPS (X_1)	-0,025	0,002	113,882	1	0,000	0,975	0,970	0,979
ALTITUDE (X_2)	-0,001	0,000	17,163	1	0,000	0,999	0,998	0,999
THUNDERSTORMS (X_3)	0,010	0,003	13,492	1	0,000	1,011	1,005	1,016
Constante	1,426	0,316	20,367	1	0,000	4,162		

Con estos datos se construye la ecuación de regresión logística que sirve para predecir la probabilidad de ocurrencia de incendio por rayo:

$$P\ (presencia\ o\ Y = 1) = \frac{1}{1 + e^{(-1{,}426 + 0{,}025 * X1 + 0{,}001 * X2 - 0{,}010 * X3)}}$$

En el anexo I se incluye el resultado cartográfico una vez aplicado el modelo a las celdas y determinada la probabilidad de ocurrencia de incendio en cada una de ellas.

3.2.3.4 CURVA COR

Al realizar la regresión logística se obtienen también los valores de probabilidad que se extraen como una nueva variable (denominada PRE- por defecto). PRE- se utiliza como variable de contraste que, junto con la variable de estado presencia o ausencia de incendio (FIRE) sirve para construir la curva COR (acrónimo de Receiver Operating Characteristic o Característica Operativa del Receptor) y calcular el área bajo la misma, lo que proporciona una medida de la fiabilidad del modelo construido.

La tabla 20 indica el número de casos procesados y la figura 26 muestra la curva COR para nuestro modelo.

Tabla 20. Resumen de procesamiento de casos

FIRE	N válido (por lista)
Positivo[a]	546
Negativo	546

Los valores más grandes de la(s) variable(s) de resultado de prueba indican una prueba mayor para un estado real positivo.
a. El estado real positivo es 1.

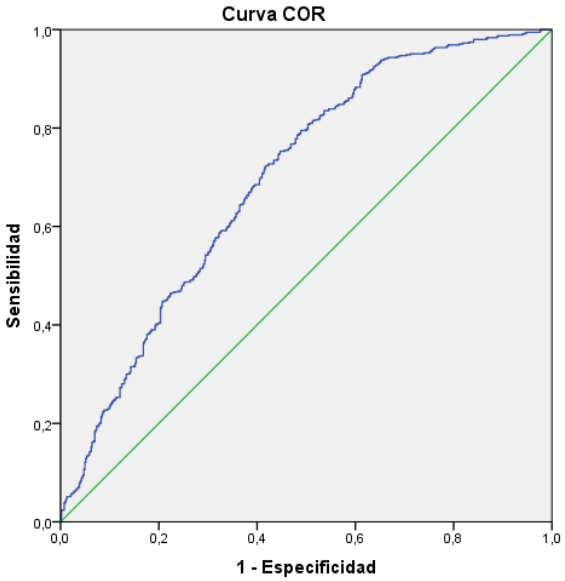


Figura 26. Curva COR para el modelo construido con el 100% de los datos

En la tabla 21 se muestran los resultados obtenidos, con un área bajo la curva de 0,698 lo cual se considera un valor válido para estimar que el modelo posee una capacidad de predicción aceptable.

Tabla 21. Área bajo la curva				
Variable(s) de resultado de prueba: Probabilidad pronosticada				
Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
0,698	0,016	0,000	0,667	0,729

La(s) variable(s) de resultado de prueba: Probabilidad pronosticada tiene, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.
a. Bajo el supuesto no paramétrico
b. Hipótesis nula: área verdadera = 0,5

3.2.4 VALIDACIÓN DEL MODELO

Para comprobar el funcionamiento del modelo se realiza una validación del mismo seleccionando al azar un 70% de la muestra (mediante el uso de Excel) que es nuestra muestra de entrenamiento y dejando el 30% restante como muestra test.

3.2.4.1 MUESTRA DE ENTRENAMIENTO

Se procede a realizar una regresión logística binaria a la muestra de entrenamiento (muestra aleatoria del 70% de los datos), calculando la variable probabilidad (PRE-), la curva COR y el área bajo la misma.

A continuación se incluyen los resultados obtenidos para evaluar la fiabilidad del modelo, primero mediante una tabla de clasificación o matriz de errores/confusión (Tabla 22) y posteriormente con la curva COR (Figura 28).

Tabla 22. Tabla de clasificación[a] (muestra entrenamiento)					
			Pronosticado		
			FIRE		Corrección de porcentaje
	Observado		0	1	
Paso 1	FIRE	0	214	163	56,8
		1	109	278	71,8
	Porcentaje global				64,4

a. El valor de corte es 0,500

El modelo predice bien el 56,8% de las ausencias de incendio (especificidad) y el 71,8% de las presencias de incendio (sensibilidad). Tiene una fiabilidad global del 64,4%.

La tabla 23 indica el número de casos procesados y la figura 27 muestra la curva COR para la muestra de entrenamiento.

Tabla 23. Resumen de procesamiento de casos (Curva COR muestra entrenamiento)	
FIRE	N válido (por lista)
Positivo[a]	387
Negativo	377

Los valores más grandes de la(s) variable(s) de resultado de prueba indican una prueba mayor para un estado real positivo.
a. El estado real positivo es 1.

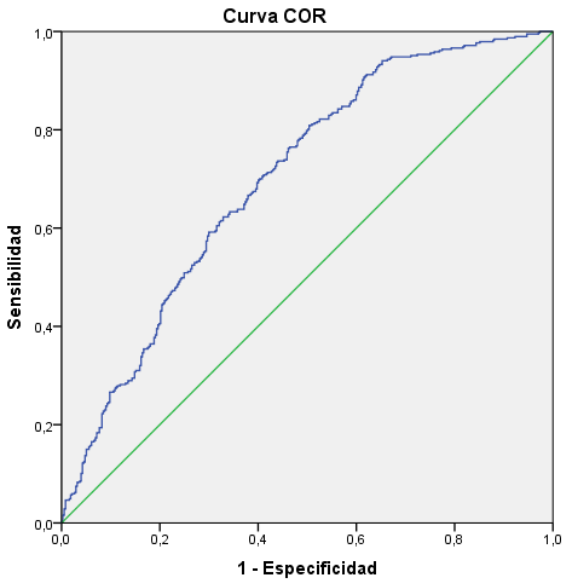


Figura 27. Curva COR para la muestra de entrenamiento

En la tabla 24 se muestran los valores obtenidos, con un área bajo la curva de 0,701, ligeramente superior a la obtenida con el 100% de los datos; por lo que, de igual manera, se considera un resultado válido para estimar que el modelo posee una capacidad de predicción aceptable.

Tabla 24. Área bajo la curva (muestra entrenamiento)				
Variable(s) de resultado de prueba: Probabilidad pronosticada				
Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
0,701	0,019	0,000	0,664	0,738

a. Bajo el supuesto no paramétrico
b. Hipótesis nula: área verdadera = 0,5

3.2.4.2 MUESTRA TEST

Se procede a seguir los mismos pasos que en la muestra de entrenamiento sobre el 30% restante. Los resultados son (Tabla 25):

Tabla 25. Tabla de clasificación[a] (muestra test)					
			Pronosticado		
			FIRE		Corrección de porcentaje
	Observado		0	1	
Paso 1	FIRE	0	101	68	59,8
		1	46	113	71,1
	Porcentaje global				65,2

a. El valor de corte es 0,500

El modelo predice bien el 59,8% de las ausencias de incendio (especificidad) y el 71,1% de las presencias de incendio (sensibilidad). Con una fiabilidad global del 65,2%.

La tabla 26 indica el número de casos procesados y la figura 28 muestra la curva COR para la muestra de entrenamiento.

Tabla 26. Resumen de procesamiento de casos (Curva COR muestra test)	
FIRE	N válido (por lista)
Positivo[a]	159

Tabla 26. Resumen de procesamiento de casos (Curva COR muestra test)	
FIRE	N válido (por lista)
Negativo	169

Los valores más grandes de la(s) variable(s) de resultado de prueba indican una prueba mayor para un estado real positivo.
a. El estado real positivo es 1

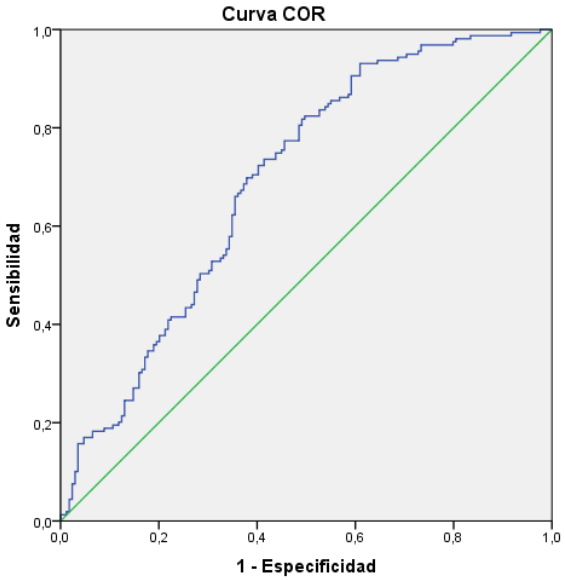


Figura 28. Curva COR para la muestra test

El área bajo la curva es de 0,692, valor ligeramente inferior a los mostrados anteriormente considerando asimismo que es un valor válido para estimar que el modelo posee una capacidad de predicción aceptable.

Tabla 27. Área bajo la curva (muestra test)				
Variable(s) de resultado de prueba: Probabilidad pronosticada				
Área	Error estándar[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
0,692	0,029	0,000	0,636	0,749

a. Bajo el supuesto no paramétrico
b. Hipótesis nula: área verdadera = 0,5

Por tanto, y según todo lo expuesto anteriormente se considera que el modelo de predicción propuesto es adecuado.

4 RESULTADOS

En el presente apartado se incluye de forma resumida los resultados de interés obtenidos para la construcción del modelo de probabilidad de ocurrencia de incendio por rayo en Castilla y León.

ECUACIÓN DE REGRESIÓN LOGÍSTICA PARA LA PREDICCIÓN

$$P\ (presencia\ o\ Y = 1) = \frac{1}{1 + e^{(-1{,}426 + 0{,}025 * X1 + 0{,}001 * X2 - 0{,}010 * X3)}}$$

Siendo:

- X_1: Porcentaje de superficie de celda que ocupan los cultivos y prados
- X_2: Altitud media de la celda en metros
- X_3: Número de días de tormenta

RESUMEN DE LA FIABILIDAD DEL MODELO

Se opta por incluir en una tabla (28) las medidas de especificidad, sensibilidad, fiabilidad y área bajo la curva COR para facilitar la comparación de los resultados obtenidos para el total de la muestra, la muestra de entrenamiento y la de validación o test.

Tabla 28. Resumen fiabilidad del modelo

	Muestra		
	Total (100%)	Entrenamiento (70%)	Test (30%)
Especificidad	57,1%	56,8%	59,8%
Sensibilidad	72,7%	71,8%	71,1%
Fiabilidad	64,9%	64,4%	65,2%
Área bajo la curva COR	0,698	0,701	0,692

6 CONCLUSIONES

Tras un estudio estadístico de los datos, se construye mediante regresión logística un modelo que sirva para calcular la probabilidad de ocurrencia de incendio por rayo en Castilla y León.

Se realiza el modelo incluyendo variables relacionadas con el tipo de vegetación y características de los rayos y tormentas acaecidos. De forma análoga a otros estudios similares se ha encontrado significancia en las variables:

- Porcentaje de superficie de la celda ocupada por cultivos y prados.
- Altitud media.
- Número de días de tormenta.

Que serán por tanto las variables consideradas para la construcción del modelo.

De acuerdo con el coeficiente β_i, las variables cultivos y altitud tienen un efecto negativo en la probabilidad de ocurrencia de incendios por rayo, es decir, a menor proporción de superficie de cultivos y altitud, menor probabilidad de ocurrencia de incendio. De forma contraria, la variable número de días de tormenta tiene un efecto positivo, por lo que cuanto mayor valor muestre mayor será la probabilidad de ocurrencia.

Para comprobar el funcionamiento del modelo se utiliza una muestra de entrenamiento y otra de test. Los resultados obtenidos de especificidad, sensibilidad, fiabilidad y área debajo de la curva COR indican que EL MODELO TIENE UNA CAPACIDAD DISCRIMINATORIA ACEPTABLE.

Con todo lo observado se concluye que, las zonas de mayor riesgo de incendio se sitúan en las zonas periféricas de la comunidad autónoma, especialmente en el norte-noroeste de la provincia de León, norte de Palencia y de Burgos. También en la zona central de Soria e incluso Segovia; el Sur de Ávila, gran parte de Salamanca (especialmente al oeste) así como el oeste de Zamora. Por el contrario, las zonas de menor riesgo se sitúan en la zona centro de Castilla y León.

Esto se observa claramente si comparamos el resultado obtenido con la representación visual de las variables utilizadas en el modelo. El porcentaje de superficie ocupada por cultivos (Figura 4) muestra una relación negativa con la presencia de incendio, situándose los mayores porcentajes en el centro de la zona de estudio. Igualmente con la altitud (Figura 10) se observa que las mayores altitudes se encuentran en las zonas periféricas de Castilla y León al igual que las zonas de mayor riesgo de incendio. Por último, si se comparan las zonas de mayor riesgo de incendio con el número de días de tormenta (figura 24), también se observa la relación entre ellos, puesto que los mayores valores de esta variable se sitúan también en la periferia de la comunidad autónoma, especialmente en el norte-noroeste y en la provincia de Soria.

BIBLIOGRAFÍA

CUBO, J.E., DEL MORAL, L., GALLAR, J.J., JEMES, V., LÓPEZ, M., MONDELO, R., MUÑOZ, A., PARRA, P.J., 2014: *"Incendios forestales en España. Año 2012".* Ministerio de Agricultura, alimentación y medio ambiente.

Dirección General de Desarrollo Rural y Política Forestal. Subdirección General de Silvicultura y Montes. Ministerio de Agricultura, Alimentación y Medio Ambiente. *"Estadística general de incendios forestales. Memoria".*

MORA, M., 2012 : "*La actividad tormentosa en Castilla y León: Análisis microescalar y modelos conceptuales".* Tesis leída en la Universidad de Salamanca para optar al doctorado en Física. http://www.tdx.cat/handle/10803/111017

ÁLVAREZ, E., ESPEJO, F., CORTÉS, F.J., LAFRAGÜETA, C. y SERRANO, R., 2011: *"Caracterización sinóptica de los procesos convectivos en el interior del NE peninsular".* Agencia Estatal de Meteorología (AEMET). Nota técnica 3. NIPO: 784-11-008-8.

FABA-FERNÁNDEZ, M., BLANCO-VÁZQUEZ, M.A., BLANCO-OVIEDO, J., CASTEDO-DORADO, F. RODRIGUEZ-PEREZ, J.R., 2013: *"Caracterización de los incendios forestales producidos por rayo en Castilla y León en el periodo 2000-2010"*

PACHECO, C. E., AGUADO, I., NIETO, H., 2009:*"Análisis de ocurrencia de incendios forestales causados por rayo en la España peninsular", GeoFocus (Artículos), nº 9, p. 232-249. ISSN: 1578-5157.*

CASTEDO-DORADO, F., RODRIGUEZ-PEREZ, J.R., MARCOS-MENENDEZ, J.L. y ÁLVAREZ-TABOADA, M.F., 2011: *"Modelling the probability of lightning-induced forest fire ocurrence in the province of León (NW Spain)".*

RODRIGUEZ-PEREZ, J.R., GÓMEZ-CUARESMA, M., ÁLVAREZ-TABOADA, M.F., MARCOS, J.L., RUIZ-PÉREZ, I., CASTEDO-DORADO, F., 2009: *"Modelización de la probabilidad espacial de ocurrencia de incendios forestales por rayo en la provincia de León".*

BISQUERT, M.M., 2011: *"Una metodología para la estimación del riesgo de incendio empleando imágenes del sensor MODIS/TERRA"*. Tesis leída en la Universidad de Valencia para optar al doctorado en Física.

ORDÓÑEZ, C., ROCA-PARDIÑAS, J., CASTEDO-DORADO, F. & RODRÍGUEZ-PÉREZ, J.R., (2013): "*A Bootstrap-Based Covariate Selection Method for Modeling the Risk of Lightning-Induced Fires at a Local Scale: A Case Study in Northwest Spain", Human and Ecological Risk Assessment: An International Journal, 19:1, 254-267*

"IBM SPSS Statistics 22 Core System". Guía del usuario.

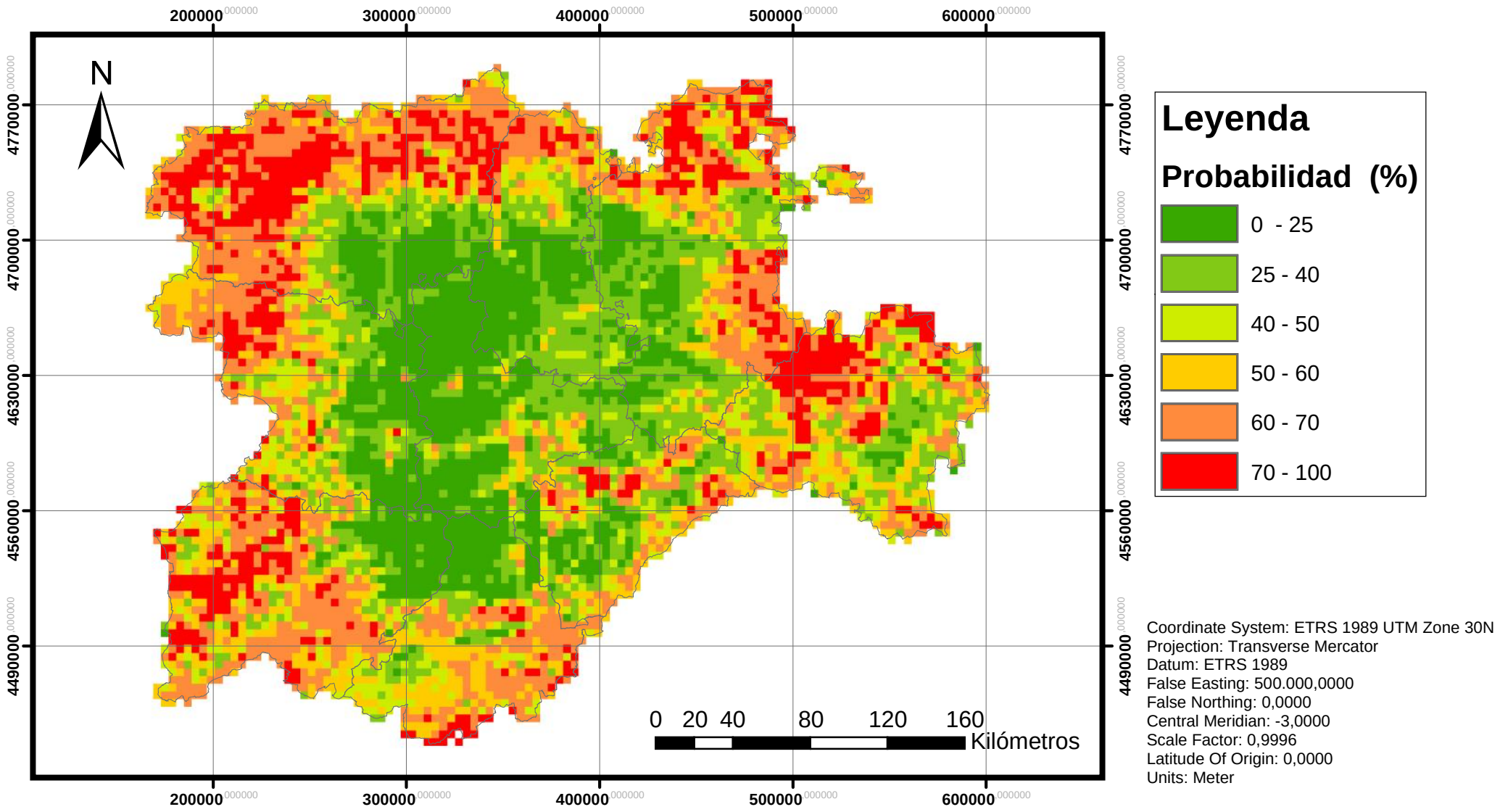
IBM SPSS REGRESSION 19. 2010. *Capítulo 17. Análisis de correlación lineal: Los procedimientos Correlaciones bivariadas y Correlaciones parciales*.

PARDO, A.; RUIZ, M. A., 2006: "*Análisis de datos con SPSS 13 Base*". ISBN 8448145364.

AGUAYO, M., 2007: "*Cómo hacer una Regresión Logística con SPSS © 'paso a paso'. (I)".* Docuweb FABIS (Fundación andaluza Beturia para la investigación en salud).

AGUAYO, M., LORA, E., 2007: "*Cómo hacer una Regresión Logística con SPSS © 'paso a paso'. (II):análisis multivariante".* Docuweb FABIS (Fundación andaluza Beturia para la investigación en salud).

ANEXO I. MAPA DE PROBABILIDAD DE OCURRENCIA DE INCENDIO POR RAYO EN CASTILLA Y LEÓN



Coordinate System: ETRS 1989 UTM Zone 30N
Projection: Transverse Mercator
Datum: ETRS 1989
False Easting: 500.000,0000
False Northing: 0,0000
Central Meridian: -3,0000
Scale Factor: 0,9996
Latitude Of Origin: 0,0000
Units: Meter

EPM ESCUELA POLITECNICA DE MIERES
UNIVERSIDAD DE OVIEDO

Mapa de probabilidad de ocurrencia de incendio por rayo en Castilla y León

Mapa 1.1	Máster en Teledetección y Sistemas de Información Geográfica
	Realizado por: Susana Egea Trapiello
Hoja 1 de 1	Revisado por: Celestino Ordoñez Galán
	Fecha: Julio de 2015