Índice general

Capítulo 1

Construcción del conjunto de datos

El primer paso a la hora de construir cualquier modelo de predicción es disponer de datos adecuados que permitan explicar correctamente el fenómeno en estudio, en este caso los incendios forestales en Andalucía. Con este fin, se ha llevado a cabo un extenso estudio previo del dominio del problema para conocer qué variables son relevantes de cara a la predicción de incendios forestales, analizando estudios similares realizados anteriormente así como otras fuentes relativas a la ecología del fuego, que nos permitiesen conocer el efecto que cabría esperar de estas variables.

Se ha querido adoptar un enfoque dinámico, es decir, el objetivo no es construir un modelo estacionario que nos indique si una determinada zona se verá afectada por un incendio forestal a lo largo de un amplio periodo temporal, si no que se pretende ser capaz de predecir si un determinado punto del territorio se verá afectado por un incendio forestal en un momento concreto, en base a las covariables correspondientes a ese lugar en ese momento. Es decir, se considera no solo la dimensión espacial de los datos si no también la temporal, al mayor nivel de desagregación disponible. Este es un enfoque mucho menos explorado, debido fundamentalmente a dos factores:

1. La dificultad de disponer de información fiable y de calidad desagregada espacio-temporalmente.
2. La dificultad de trabajar con datos de estas características de cara al análisis y principalmente a la modelización, ya que son datos correlados en el tiempo y en el espacio.

Queda claro, por tanto, que se trata de un problema complejo que requiere de simplificaciones para poder ser abordado, más aun dadas las limitaciones en los recursos disponibles y la enorme cantidad de de datos que se están considerando y que requieren de un procesamiento sumamente costoso desde un punto de vista computacional.

Por todo ello, esta sección es probablemente la de mayor importancia y dificultad de todo el trabajo. Por un lado, debido a que implica la toma de decisiones que serán determinantes de cara al correcto desempeño de los modelos que se construirán más adelante. Además, porque requiere de un vasto conocimiento del problema que permita un enfoque adecuado que posibilite la consecución de los objetivos que se esperan conseguir y necesita del uso de técnicas específicas de procesamiento de datos espaciales que no han sido tratadas durante el grado. Y por último, por la enorme demanda computacional que requiere el procesamiento de datos espaciales.

1.1. Determinación del marco del estudio

El primer paso ha sido limitar el área y la franja temporal que abarcará el estudio. Para ello, ha sido necesario basarse principalmente en la disponibilidad y consistencia de la información requerida para el proyecto y en las limitaciones computacionales impuestas por el equipo disponible.

En cuanto a la disponibilidad de información, hay que diferenciar entre la información de incendios forestales y la información de variables que permitan explicar este fenómeno considerando la mayor desagregación espacial y temporal posible.

1.1.1. Incendios forestales

En lo referente a los datos sobre incendios forestales cabe mencionar que España cuenta con una de las mayores y más completas bases de datos sobre incendios forestales a nivel europeo. Se trata de la Estadística General de Incendios Forestales (EGIF), que en su versión definitiva actualmente contiene toda la información que se recoge en cada parte de incendio forestal que ha tenido lugar en España desde 1983 hasta 2015, incluyendo su información espacial con sus coordenadas de origen. Se ha explorado extensamente el uso de esta base de datos para el proyecto, dada su exhaustividad y completitud. Sin embargo, lamentablemente no ha sido posible en este caso incorporarla al trabajo por diversas razones que se detallarán a continuación.

La principal de ellas fue que hasta marzo de 2024 la base de datos de la EGIF solo se encontraba disponible en el Catálogo de Datos del Gobierno de España en formato TURTLE[1] y esto conllevó numerosas dificultades. Se exploraron distintas librerías de R (y alguna de Python) para el manejo de datos en este formato, como *RDFlib* [2]. Sin embargo, al tratarse de una base de datos de un tamaño considerable (**aproximadamente 1GB y con más de una decena de millones de tripletas**), esta librería no era suficientemente eficiente para poder realizar consultas en un tiempo razonable al conjunto de datos. Tras explorar otras alternativas, se valoró la posibilidad de usar un *triplestore*, es decir, una base de datos especialmente diseñada para el almacenamiento y recuperación de tripletas a través de consultas semánticas. En este caso se usó *Apache Jena Fuseki*, ya que cuenta con una interfaz que facilita su uso. Sin embargo, aunque esto supuso una mejora considerable en la eficiencia y permitió realizar consultas sencillas a la base de datos, en este caso fue la complejidad del gráfico de datos (ontología) y la escasa documentación disponible sobre esta, la que impidió que se pudiesen realizar las consultas más complejas requeridas para llevar a cabo el proyecto. Además, se debe tener en cuenta que se trata de una base de datos muy heterogénea y con numerosos datos faltantes debida su naturaleza, por lo que requiere de un preprocesamiento que probablemente será complicado y costoso en tiempo y en recursos computacionales. Al no disponer de ninguno de estos, finalmente se optó por buscar una alternativa más abarcable dada las limitaciones con las que cuenta un Trabajo de Fin de Estudios, aunque queda abierta la posibilidad de explorar esta base de datos en futuros estudios, la cual podrá aportar nuevas dimensiones al estudio de los incendios forestales en España gracias a la enorme cantidad de información que ofrece.

[1] TURTLE es una sintaxis para RDF compatible con SPARQL. RDF (*Resource Description Framework*) es un estándar de semántica web utilizado para el intercambiar de datos en la Web.

Ante esta situación, la solución planteada fue limitar el área en estudio a la Comunidad Autónoma de Andalucía, aprovechando la enorme disponibilidad de información medioambiental que ofrece la Red de Información Ambiental de Andalucía (REDIAM). En particular, se emplea la cartografía generada por la REDIAM sobre las áreas recorridas por los incendios forestales entre 1975 y 2022. Esta contiene los perímetros de incendios forestales mayores de 100 ha en Andalucía obtenidos a partir de imágenes de satélite y datos de campo. Se trata por tanto de una información que no es exhaustiva, pues los incendios con una extensión inferior a 100 ha no han sido considerados. Sin embargo, frente a no disponer de otra información operativa de mayor calidad, se utilizará esta teniendo en cuenta que tendrá un efecto sobre las conclusiones que se puedan sacar de los modelos que se construyan.

De esta forma, se han recopilado los polígonos de 1090 incendios forestales ocurridos en Andalucía entre 2002 y 2022, junto con la fecha de inicio de cada uno de ellos (Figura 1.1). El motivo por el que se ha decidido limitar el estudio a solo 20 años se detalla en la siguiente sección.

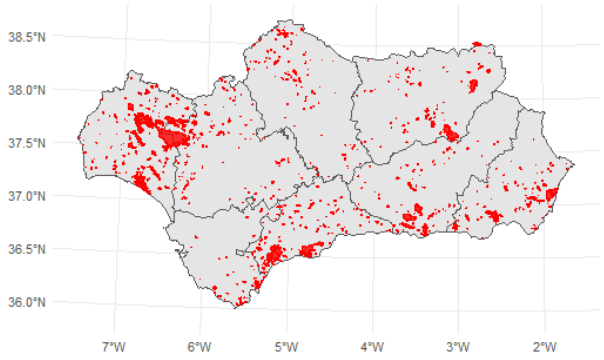

Figura 1.1: Áreas recorridas por el fuego entre 2002 y 2022 en incendios forestales mayores de 100 hectáreas en Andalucía. *Fuente: Elaboración propia a partir de las áreas recorridas por el fuego disponibles en la REDIAM.*

1.1.2. Variables predictoras

Una vez limitada la extensión territorial del estudio, el siguiente paso era acotar la franja temporal que abarcaría el estudio en base a la disponibilidad de datos adecuados para explicar el fenómeno en cuestión, desagregando espacial y temporalmente.

Los incendios forestales son un proceso sumamente complejo, en el que actúan numerosos factores de muy distinta índole (cau [1],Moreno et al. [5]). Además, dentro de un incendio forestal se pueden distinguir distintas fases que presentan características muy diversas y sobre las que actúan distintos agentes: ignición, propagación y extinción. Dada la información sobre incendios forestales disponible, se está obligado a adoptar un enfoque global, pues no se dispone de los puntos de ignición u origen de los incendios forestales. El enfoque será, por tanto, intentar predecir si una determinada localización se verá afectada por un incendio forestal (de más de 100 ha) en un momento concreto.

Además, es importante tener en cuenta que existen factores estructurales que tienen una influencia directa sobre los regímenes de incendios forestales como son las tendencias de uso y explotación de los bosques, la presencia de interfaz urbano forestal (terreno forestal

entremezclado con las viviendas), los tipos y técnicas de agricultura que se llevan a cabo, la presencia e intensidad del pastoreo, los cambios en los usos de suelo e incluso conductas sociales y tendencias demográficas diversas. Se trata de variables que cambian a lo largo de periodos relativamente largos de tiempo y que muy difícilmente pueden ser incluidos en los modelos, dada la falta de datos sobre ellas, así como su carácter transversal. Por ello, se ha considerado conveniente no extender en exceso el periodo de estudio, reconocida la imposibilidad de incluir en el modelo todas las variables que tienen un impacto relevante en la aparición de incendios y que son cambiantes en el tiempo.

Todo ello hace necesario que el conjunto de datos utilizado contenga información sobre todas las dimensiones (o al menos las principales) que influyen en cualquiera de las fases de un incendio forestal. Es decir, se deben incluir la dimensión antropogénica, la demográfica, la hidrográficas, la topográfica, la meteorológica y la vegetación. Es importante recalcar que siempre se hace referencia a datos geoespaciales pues debe ser la información relativa al lugar (y al momento) del incendio, con la dificultad posterior que esto supondrá.

Por último, es importante diferenciar entre características que se considerarán estructurales (y por tanto invariantes a lo largo del periodo de estudio) y aquellas que se considerarán variables en el tiempo. Dentro de las primeras se encuentran todas las características relacionadas con la topografía del terreno, las infraestructuras y los usos del suelo, como por ejemplo el modelo de elevaciones, la distribución de asentamientos de población, la red de carreteras y el uso de suelo. Todas las demás variables de carácter demográfico, meteorológico o de vegetación se considerarán, por tanto, desagregadas temporalmente.

En base a todo lo mencionado y a la disponibilidad de información de calidad de las categorías comentadas, se ha decidido limitar la franja temporal del estudio a los 20 años que van de 2002 a 2022, ambos inclusive.

1.2. Fuentes de datos

Como se ha comentado en la sección anterior, los datos sobre los incendios forestales se han obtenido de los perímetros de incendios forestales mayores de 100 ha en Andalucía entre 1975 y 2020 disponibles la REDIAM. De cada incendio registrado se dispone de su fecha de inicio y del polígono del área recorrida por el fuego, así como de otras variables que dependen del año de la campaña y que no serán relevantes de cara al presente estudio.

Tomando como base estudios similares [4, Sayad et al. [6], Stojanova et al. [8]] y partiendo de las 6 categorías ya mencionadas se han recopilado 23 conjuntos de datos de distinto tipo que se usarán para explicar y predecir los incendios forestales en Andalucía. Estos conjuntos se recogen en la Tabla 1.1, donde también se indica la fuente de la que ha sido obtenido cada uno de ellos, el tipo de datos que contiene (indicando su resolución en el caso de los datos ráster) y la frecuencia de las observaciones (o resolución temporal) en el caso de las variables temporales. En realidad, el número de archivos de datos que se manejan es mucho mayor, ya que, por ejemplo, para la variable *NDVI* se dispone de un archivo *tiff* para cada mes del periodo de estudio, resultando en un total de unos 240 archivos ráster diferentes solo para esta variable. Inevitablemente, esto añade cierta complejidad al manejo de los datos, al tener que combinarlos para poder emplearlos.

Es relevante la heterogeneidad de los datos recopilados, pues se dispone tanto de datos tabulares como de datos espaciales y dentro de estos últimos de datos vectoriales y datos

Categoría	Datos	Fuente	Tipo de dato	Frecuencia
Topográficas	Altitud	DERA[a]	TIFF (100m)	-
	Orientación	REDIAM[b]	TIFF (100m)	-
	Pendiente	REDIAM	TIFF (100m)	-
	Curvatura	REDIAM	TIFF (100m)	-
Vegetación	NDVI	REDIAM	TIFF (250m)	Mensual
Antropogénicas	Uso de suelo	DERA	Shapefile	-
	Red de carreteras	DERA	Shapefile	-
	Red de ferrocarril	DERA	Shapefile	-
	Línea eléctrica	DERA	Shapefile	-
	Espacio protegido	DERA	Shapefile	-
	Senderos / Vías Verde / Carriles Bici	DERA	Shapefile	-
	Caminos / Vías Pecuarias	DERA	Shapefile	-
Demográficas	Número de habitantes por municipio y año	IECA[c]	csv	Anual
	Extensión municipal	IECA[c]	csv	-
Hidrográficas	Principales Ríos	MAGRAMA[d]	Shapefile	-
Meteorológicas	Precipitación (mm/day)	NASA POWER[e]	df (0.5º x 0.625º)	Diaria
	Temperatura a 2m sobre la superficie (º)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Humedad del suelo (%)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Dirección del viento a 10 metros sobre la superficie terrestre(º)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Humedad relativa a 2m sobre la superficie (%)	NASA POWER	df (0.5º x 0.625º)	Diaria
	Cantidad de precipitaciones (mm/day)	NASA POWER	dfdf (0.5º x 0.625º)	Diaria

Fuente: Elaboración propia

[a] Datos Espaciales de Referencia de Andalucía (DERA)

[b] Descargas Rediam

[c] Instituto de Estadística y Cartografía de Andalucía (IECA)

[d] Ministerio de Agricultura, Alimentación y Medio Ambiente (MAGRAMA)

[e] NASA Prediction Of Worlwide Energy Resources (NASA POWER)

Tabla 1.1: Datos recopilados por categorías.

ráster, con distintas resoluciones, distintas frecuencias y distintos sistemas de referencia de coordenadas. Esto hará que el procesamiento de estos datos hasta obtener datos adecuados para el análisis estadístico sea costoso y que deban utilizarse técnicas específicas de geocumputación.

Cabe también mencionar que se ha optado por el uso de datos meteorológicos basados en modelos y en observaciones satelitares, en lugar del uso de datos provenientes de estaciones meteorológicas. Si bien la información de estaciones meteorológica puede ser más precisa, la dificultad de disponer de datos consistentes y continuos en el tiempo a lo largo del periodo de estudio de las variables meteorológicas seleccionadas ha hecho que este enfoque no sea viable. En esta dirección se ha explorado la API de la AEMET y algunos paquetes de R como *climate* [3], sin llegar a resultados satisfactorios. Por otro lado, el paquete *nasapower* [7] permite la descarga de una gran cantidad de variables meteorológicas con frecuencia diaria y con una resolución de aproximadamente 0.5×0.625 grados de latitud y longitud (unos 50 km). Se trata de información meteorológica basada en modelos y obtenida a partir de observaciones satelitares, por lo que no se recomienda su uso para trabajar a escalas pequeñas y a un gran nivel de detalle. Teniendo esto en cuenta, y tras explorar en profundidad todas las alternativas mencionadas, se ha decidido trabajar con esta fuente de información meteorológica al ser la única viable en estos momentos. Si quisiese extenderse el estudio, sería conveniente profundizar en la búsqueda de fuentes que permitan obtener información meteorológica de una mayor calidad y detalle, probablemente obtenida a partir de estaciones meteorológicas.

1.3. Procesamiento de los datos

Una vez se dispone de todos los conjuntos de datos que se usarán en el estudio, el siguiente paso será combinarlos de manera adecuada y transformarlos a un formato tabular apto para el análisis estadístico y la construcción de modelos predictivos. Dado que el objetivo que se persigue es predecir si, dada unas condiciones meteorológicas concretas en un momento dado, un punto del territorio andaluz se verá afectado o no por un incendio forestal, será necesario disponer de una cantidad suficiente de muestras negativas y positivas distribuidas espacial y temporalmente que tengan asociadas las variables explicativas correspondientes.

En las siguientes secciones se explicará en profundidad el proceso seguido para generar la muestra. Intuitivamente, las muestras positivas serán aquellas observaciones (puntos definidos en el tiempo y en el espacio) dentro del marco espacio-temporal del estudio en las que se ha detectado un incendio forestal en el día de la observación. Es decir, son observaciones dentro de los polígonos de incendios el día que estos se han producido. Por tanto, las muestras negativas serán observaciones dentro del marco espacio-temporal definido en las que no se ha detectado un incendio forestal. Es importante tener en cuenta que dado que solo se dispone de los incendios con una extensión mayor a 100 ha, la muestra cuenta con un importante sesgo, ya que los casos positivos están infrarepresentados. Por ello, no podremos hacer inferencia a todos los incendios forestales, si no solo a los de una extensión superior a 100 ha. Un comentario relevante es que al usar información meteorológica basada en modelos, la presencia de un incendio forestal no altera los valores meteorológicos de esa zona.

A continuación se detalla el proceso seguido para generar el conjunto de datos depurado sobre el que desarrollar el estudio a partir de los distintos conjuntos de datos en bruto:

1. Generación de una muestra balanceada de casos positivos y negativos.
2. Asignación de las variables descriptivas a cada observación.
3. Depuración de la muestra.

1.3.1. Generación de una muestra balanceada de casos positivos y negativos.

Para poder construir cualquier modelo de clasificación binaria se necesita disponer de una muestra que cuente con un número suficiente de casos positivos y negativo. Además, es aconsejable trabajar con conjuntos de datos balanceados para evitar sesgos en los modelos de clasificación. [9]

Como ya se ha comentado, se considerarán observaciones positivas aquellas que se hayan visto afectadas por un incendio forestal en el día y lugar de la observación. En cambio, serán observaciones negativas aquellas que no se hayan visto afectadas por un incendio forestal en el día y lugar de la observación. Estas observaciones deberán generarse a partir de los polígonos de incendios disponibles. Para ello, se usará un enfoque similar al utilizado en [8], con algunas diferencias importantes. En el estudio citado se usan los puntos de ignición como muestras positivas, con el objetivo de predecir los puntos de origen de los incendios forestales. En cambio, en el presente trabajo no se disponen de los puntos de ignición de los incendios, por lo que el enfoque adoptado es ligeramente diferente; el

objetivo es predecir las zonas que pueden verse afectadas por un incendio forestal (superior a 100 ha) bajo unas circunstancias concretas. De esta forma, para construir la muestra de casos positivos se han generado 10 puntos aleatorios dentro de cada polígono de incendio y se les ha asignado la fecha del día de inicio del incendio. En Stojanova et al. [8] la muestra de casos negativos se genera de la siguiente manera: se toman fechas aleatorias dentro del periodo de estudio y a cada una de ellas se le asocia una localización aleatoria dentro del área de estudio satisfaciendo que deben estar a al menos 15 km de de cualquier incendio detectado en un margen de $\pm$ 3 días. Esta forma de tomar los casos negativos asegura que estén lo suficientemente alejados de los incendios forestales para representar condiciones no influidas por estos, dando prioridad así a las áreas con una menor prioridad de ocurrencia de incendio en un período definido. En el presente trabajo se utilizarán los mismos parámetros (franja de $\pm$ 3 días y distancia superior a 15 km), sin embargo, sería conveniente en estudios futuros plantearse si esos parámetros son adecuados o tal vez sería más adecuado tomar otros valores, basados, por ejemplo, en la duración media de los incendios en Andalucía y otras características propias de los incendios en la región.

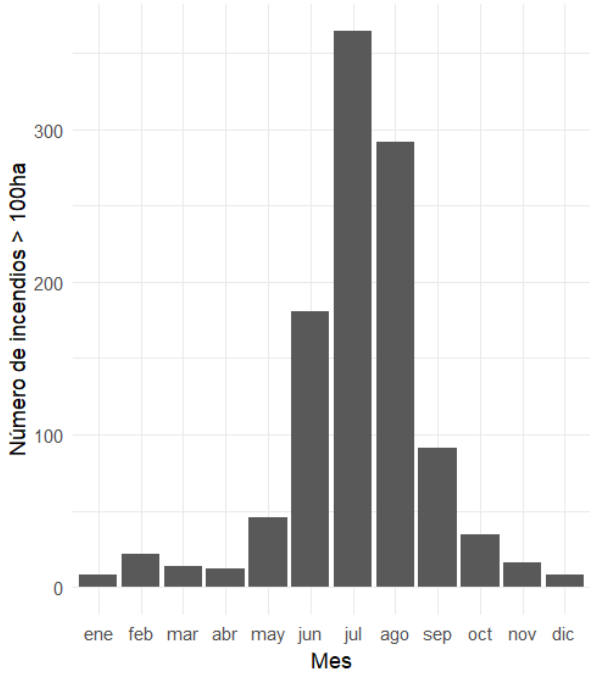

Figura 1.2: Número total de incendios por mes durante el periodo de estudio. *Fuente: Elaboración propia.*

Tanto en Stojanova et al. [8] como en otros estudios consultados se generan los casos negativos tomando fechas completamente aleatorias dentro de la franja temporal del estudio. Sin embargo, esto induce un claro sesgo en los datos, ya que las observaciones positivas no se distribuyen uniformemente entre los 12 meses, si no que se concentran marcadamente en los meses de verano, como puede observarse en la Figura 1.2. Generar el conjunto de datos sin tener en cuenta este hecho hace que las muestras positivas y negativas tengan características meteorológicas muy diferenciadas que no responden al verdadero proceso latente de aparición de incendios forestales sino al proceso de selección de la muestra, pudiendo provocar un marcado sesgo positivo en las medidas de evaluación de los modelos que en realidad no estarían reflejando la realidad si no los sesgos introducidos en los conjuntos de datos. O dicho de otro modo, los modelos acabarían dando la mayor parte

del peso a las variables meteorológicas y acabarían limitándose a identificar el mes y no las demás causas antropológicas que son las que verdaderamente desencadenan la gran mayoría de los incendios. Por ello, en el presente trabajo, para generar aleatoriamente los días de las muestras negativas se seguirá una distribución de probabilidad proporcional a la cantidad de incendios observados a lo largo del periodo de estudio en cada mes (véase la Figura **??**). Mediante este enfoque se espera obtener una muestra balanceada y estratificada por mes que permita construir modelos de clasificación capaces de captar los patrones latentes de aparición de incendios forestales. Además, este enfoque permite centrar el esfuerzo computacional en los periodos en los que más incendios se producen.

Pueden observarse todos los detalles del procedimiento seguido para generar la muestra en el código usado para ello, que se adjunta en la primera sección del [Apéndice: Código].

1.3.2. Asignación de las variables descriptivas a cada observación

Una vez generada una muestra balanceada de 22.170 observaciones dentro de Andalucía entre el 1 de enero de 2002 y el 31 de diciembre de 2022 siguiendo el enfoque apenas descrito, el siguiente paso es asignar a cada una de ellas los valores correspondientes a ese día y a esa localización concreta de todas las variables predictoras a partir de los conjuntos de datos que se han recopilado (recogidos en la Tabla 1.1). Dado que será necesario calcular distancias, se usa un Sistema de Referencia de Coordenadas proyectado al generar la muestra. En particular se usa una variante del UTM30N por ser el que traen los archivos *raster* de las variables topográficas (es conveniente, siempre que sea posible, no transformar el crs de los datos *raster* pues al hacerlo se deben interpolar los valores de los píxeles, produciendo una pérdida de información).

A continuación se detalla el proceso seguido para manipular construir cada variable:

- **Variables topográficas**: Simplemente se extrae el dato del píxel correspondiente a las coordenadas del punto de cada observación para cada uno de los conjuntos de datos.
- **Variables antropológicas**:
  - Se calculan las distancias de cada observación a la red de carreteras, a la red de ferrocarril, a las poblaciones y a la línea electrica, creando así las variables *dist_carreteras*, *dist_ferrocarril*, *dist_poblaciones* y *dist_electr*, respectivamente. Las geometrías de los senderos, de las vías verdes y de los carriles bicis se unen y se calcula la distancia de cada observación a este conjunto de geometrías, generando así la variable *dist_sendero*. De la misma forma se procede con los caminos y las vías pecuarias, que dan origen a la variable *dist_camino*. Estas uniones se han llevado a cabo por considerar que sus elementos tienen características similares. Siempre se hace referencia a la distancia euclídea.
  - Para construir una variable dicotómica *enp* que indique si la observación se encuentra o no dentro de un Espacio Natural Protegido primero se ha rasterizado el conjunto de polígonos de Espacios Naturales Protegidos de Andalucía (de forma que en cada píxel se indica 1 si el centro de este está dentro del polígono de un ENP o 0 si no) y posteriormente se ha extraído el valor del píxel

Nivel 1	Nivel 2	Código
Superficies artificiales	Zonas urbanas	11
	Zonas industriales, comerciales y de transporte	12
	Zonas de extracción minera, vertederos y de construcción	13
	Zonas verdes artificiales, no agrícolas	14
Zonas agrícolas	Tierras de labor	21
	Cultivos permanentes	22
	Prados y praderas	23
	Zonas agrícolas heterogéneas	24
Zonas forestales con vegetación natural y espacios abiertos	Bosque	31
	Espacios de vegetación arbustiva y/o herbácea	32
	Espacios abiertos con poca o sin vegetación	33
Zonas húmedas	Zonas húmedas continentales	41
	Zonas húmedas litorales	42
Superficies de agua	Aguas continentales	51
	Aguas marinas	52

Tabla 1.2: Códigos de uso de suelo. *Fuente: Elaboración propia a partir de https://land.copernicus.eu/content/corine-land-cover-nomenclature-guidelines/html/*

que contiene a cada observación. En la rasterización se ha usado como modelo el ráster de eleviaciones con una resolución de 100 m $\times$ 100 m. Proceder de este modo hace que se pierda algo de resolución pero resulta significativamente más eficiente computacionalmente que comprobar para cada observación la relación espacial *estar dentro del polígono de algún ENP* (este enfoque resultaba computacionalmente inviable dado el equipo disponible).

- Para construir la variable *uso_suelo* se ha procedido de manera similar. Primero se ha rasterizado, usando como modelo el mapa de elevaciones con una resolución de 100 m $\times$ 100 m, asignando a cada píxel la categoría de uso de suelo del polígono que cubriese el centro del píxel. Y posteriormente, para cada observación se ha extraído el valor del píxel sobre el que estuviese. De nuevo, se ha procedido así por cuestiones de eficiencia. Es necesario hacer algunos comentarios más sobre esta variable. La información de uso de suelo proviene del mapa de Ocupación de Uso de Suelo CORINE Land Cover 2018, donde se establece 3 niveles de clasificación con 5, 15 y 44 clases, respectivamente. La primera clase se corresponde con el primer dígito del código, las segunda con el segundo y la tercera con el tercero. En este trabajo se ha decidido trabajar con el segundo nivel de clasificación, por lo que se consideran solo los dos primeros dígitos del código de cada observación. En la Tabla 1.2 se recoge la clase correspondiente a cada código.

- Para construir la variable *poblacion* se ha asignado a cada observación el código del municipio en el que está y se ha hecho un *left_join* con el código del municipio y el año de la observación. Para la variable *dens_población* se ha procedido de la misma manera pero se ha divido por la extensión del municipio en km^2.

- La distancia a los ríos *dist_rios* se ha obtenido simplemente calculando la distancia de cada observación al conjunto de geometrías de los principales ríos de España.

- El NDVI viene en archivos *raster* mensuales (lo que supone un total de unos 240 archivos en formato *.tiff*). Para cada observación se ha extraído el valor del píxel correspondiente (en función de las coordenadas del punto) del archivo correspondiente (que depende del mes y año de la observación).

1.3.3. Depuración de la muestra

Una vez construido el conjunto de datos "en bruto", se tratan los valores perdidos y se ajustan adecuadamente los tipos de las variables. En primer lugar se convierten en factores las variables *fire*, *enp* y *uso_suelo*. A continuación, se codifican las variables numéricas *WD10M* y *orientacion* en los 4 puntos cárdinales y sus bisectrices, generando así 8 clases ("N", "NW", "W", "SW", "S", "SE", "E", "NE"). En el caso de la variable orientación se añade también la clase "plano", si la pendiente en ese punto es 0.

El conjunto de datos construido (formado por 21.746 observaciones de 27 variables) tiene 200 registros incompletos, lo cual supone un 0.1 % del total de registros. De estos, el 68 % son casos negativos y el 32 % son casos positivos. Los valores perdidos se encuentran en las variables demográficas (85), en *uso_suelo* (8), en NDVI (85) y en las variables topográficas (53). Las causas de los datos faltantes son:

- El dato no está disponible para esa observación. Esto sucede con las variables demográficas (hay años para los que no está disponible el número de habitantes de algunos municipios) y el NDVI (para algunos meses no se dispone del archivo correspondiente).
- En el caso de las variables topográficas los valores perdidos se encuentran todos en los límites de la comunidad (Figura 1.3). Al proceder de datos en formato ráster, los píxeles con información no se ajustan exactamente a los límites de Andalucía (ya que son cuadrados). Esto provoca que para algunos puntos situados en los bordes del polígono no esté disponible la información de las variables topográficas.

Tras explorar otras alternativas y teniendo en cuenta tanto el reducido número de registros incompletos como la naturaleza de los valores desconocidos, se opta simplemente por eliminar estos registros.

Figura 1.3: Observaciones para los que no está disponible alguna de las variables topográficas. *Fuente: Elaboración propia.*

El resultado de todo este proceso es un conjunto de datos con **21.546 registros** y **27 variables**, las cuales se detallan en la Tabla 1.3.

Categoría	Nombre	Descripción	Tipo
Topográficas	elevacion	Elevación sobre el nivel del mar (m)	numérica
	orientacion	Orientación de la pendiente descendiente	categórica
	pendiente	Pendiente del terreno (º)	numérica
	curvatura	Curvatura de la superfie	numérica
Vegetación	NDVI	Índice de vegetación de diferencia normalizada	numérica
Antropogénicas	uso_suelo	Clasificación del uso del suelo	categórica
	dist_carretera	Distancia a la carretera más cercana (m)	numérica
	dist_ferocarril	Distancia a la vía de ferrocarril más cercana (m)	numérica
	dist_electr	Distancia a la línea electrica más cercana (m)	numérica
	enp	Espacio Natural Protegido	categórica
	dist_sendero	Distancia a la vía verde, al carril bici o al sendero más cercano (m)	numérica
	dist_camino	Distancia al camino o a la vía pecuaria más cercano (m)	numérica
Demográficas	poblacion	Número de habitantes del municipio	numérica
	dens_poblacion	Densindad de población del municipio	numérica
Hidrográficas	dist_rios	Distancia al río más próximo (m)	numérica
Meteorológicas	PRECTRORCORR	Promedio corregido del total de precipitaciones en la superficie de la tierra en masa de agua (incluye el contenido de agua en la nieve) (mm/día)	numérica
	T2M	Temperatura promedio del aire a 2 metros sobre la superficie de la tierra (ºC)	numérica
	GWETTOP	Porcentaje de humedad del suelo	numérica
	WD10M	Promedio de la dirección del viento a 10 metros sobre la superficie de la tierra	categórica
	WS10M	Promedio de la velocidad del viento a 10 metros sobre la superficie de la tierra (m/s)	numérica
	RH2M	Humedad relativa a 2 metros sobre la superficie de la tierra	numérica
Variable Objetivo	fire	Incendio forestal	categórica
Identificadoras	date	Fecha de la observación	fecha
	municipio	Nombre del municipio	texto
	cod_municipio	Código del municipio	texto
	geometry	Geometría de los puntos	sfc

Tabla 1.3: Variables del conjunto de datos depurados. *Fuente: Elaboración propia.*

Bibliografía

[1] (2005). «Estudio sobre Motivaciones de los Incendios Forestales Intencionados en España». *Informe técnico 50920029 (15DGB-2005)*, Ministerio de Medio Ambiente. https://www.miteco.gob.es/content/dam/miteco/es/biodiversidad/publicaciones/investigacion_causas_tcm30-278882.pdf.

[2] Boettiger, Carl (2018). *rdflib: A high level wrapper around the redland package for common rdf applications*. doi: 10.5281/zenodo.1098478. https://doi.org/10.5281/zenodo.1098478.

[3] Czernecki, Bartosz; Głogowski, Arkadiusz y Nowosad, Jakub (2020). *Climate: An R Package to Access Free In-Situ Meteorological and Hydrological Datasets For Environmental Assessment*. doi: 10.3390/su12010394. https://github.com/bczernecki/climate/. R package version 0.9.1.

[4] Gutiérrez-Hernández, Oliver; Senciales-González, J. M. y García, Luis V. (2015). «Los incendios forestales en Andalucía: investigación exploratoria y modelos explicativos».

[5] Moreno, J. M.; Urbieta, I.R.; Bedia, J.; Gutiérrez, J.M. y Vallejo, V.R.. «Los incendios forestales en España ante el cambio climático». https://www.miteco.gob.es/content/dam/miteco/es/cambio-climatico/temas/impactos-vulnerabilidad-y-adaptacion/cap34-losincendiosforestalesenespanaantealcambioclimatico_tcm30-70236.pdf.

[6] Sayad, Younes Oulad; Mousannif, Hajar y Al Moatassime, Hassan (2019). «Predictive modeling of wildfires: A new dataset and machine learning approach». *Fire Safety Journal*, **104**, pp. 130–146. ISSN 0379-7112. doi: https://doi.org/10.1016/j.firesaf.2019.01.006. https://www.sciencedirect.com/science/article/pii/S0379711218303941.

[7] Sparks, Adam H. (2018). «nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R». *The Journal of Open Source Software*, **3(30)**, p. 1035. doi: 10.21105/joss.01035.

[8] Stojanova, Daniela; Kobler, Andrej; Ogrinc, Peter; Ženko, Bernard y Džeroski, Sašo (2012). «Estimating the risk of fire outbreaks in the natural environment». *Data mining and knowledge discovery*, **24**, pp. 411–442.

[9] Thabtah, Fadi; Hammoud, Suhel; Kamalov, Firuz y Gonsalves, Amanda (2020). «Data imbalance in classification: Experimental evaluation». *Information Sciences*, **513**, pp. 429–441. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.

11.004.
https://www.sciencedirect.com/science/article/pii/S0020025519310497.