Índice general

1. Construcción de la base de datos **3**

1.1. Determinación del marco del estudio 4

1.1.1. Incendios forestales 4

1.1.2. Variables predictoras 5

1.2. Fuentes de datos 6

1.3. Selección de variables 6

1.4. Fuentes de datos 7

1.5. Organización de los datos 7

1.6. Depuración de los datos 7

Capítulo 1

Construcción de la base 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 pueden ser 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 andaluz 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 computacionales 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, ya que implica la toma de decisiones que serán determinantes de cara al correcto desempeño de los modelos que se construirán más adelante, requiere de un vasto conocimiento del problema que permita un enfoque adecuado que haga posible la consecución de los objetivos que se esperan conseguir, necesita del uso de técnicas específicas de procesamiento de datos espaciales que no han sido tratadas durante el grado y se ve fuertemente limitada por los escasos recursos computacionales disponibles.

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.

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 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. 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 impidió que pudiese realizar las consultas más complejas que requería para llevar a cabo el proyecto. Además, se debeb 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 aportar nuevas dimensiones al estudio de los incendios forestales en España gracias a la enorme cantidad de información que ofrece.

TURTLE es una sintaxis para RDF con una sintaxis 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 100ha 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.

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 desagregados espacial y temporalmente.

Los incendios forestales son un proceso sumamente complejo, en el que actúan numerosos factores de muy distinta índole (...). 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 (da más de 100ha) 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, 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

Categoría	Dato	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	Población del municipio	IECA[c]	csv	Anual
Hidrográficas	Principales Ríos	MAGRAMA[d]	Shapefile	-
Meteorológicas	Precipitación (mm/day)	NASA POWER[e]	df / raster (0.5º)	Diaria
	Temperatura a 2m sobre la superficie (º)	NASA POWER	df / raster (0.5º)	Diaria
	Humedad del suelo (%)	NASA POWER	df / raster (0.5º)	Diaria
	Dirección del viento a 10 metros sobre la superficie terrestre	NASA POWER	df / raster (0.5º)	Diaria
	Humedad relativa a 2m sobre la superficie (%)	NASA POWER	df / raster (0.5º)	Diaria
	Cantidad de precipitaciones (mm/day)	NASA POWER	df / raster (0.5º)	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 brutos

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 20 años que van de 2002 a 2022, ambos inclusive.

1.2. Fuentes de datos

En la tabla 1.1 se muestra algo

1.3. Selección de variables

Tras una extensa revisión de estudios similares (. . .), se ha optado por usar un conjunto de [22]

Siguiendo el ejemplo de otros trabajos similares (. . .) se ha optado por usar un conjunto de [22] variables predictoras que pueden clasificarse en las siguientes 5 categorías:

1. Topográficas Altitud Orientación Pendiente Curvatura
2. Vegetación. NDVI

3. Antropogénicas Distancia a población Uso de suelo Distancia a carreteras Distancia a Ferrocarril Distancia a Línea eléctrica Distancia a Espacio protegido Distancia a Sendero / Vía Verde / Carril Bici Distancia a Camino / Vía Pecuaria

4º Demográficas Población (Nº habitantes del municipio)

4. Hidrográficas Precipitación (mm/day) Distancia a ríos

5. Meteorológicas Temperatura a 2m sobre la superficie (º) Humedad del suelo (%) Dirección del viento a 10 metros sobre la superficie terrestre Humedad relativa a 2m sobre la superficie (%) Cantidad de precipitaciones (mm/day) Radiación solar incidente al medio día (MJ/mˆ2/day)

Fecha Municipio y código municipio

1.4. Fuentes de datos

Para la información sobre incendios forestales en Andalucía se ha empleado la cartografía generada por la Red de Información Ambiental de Andalucía (REDIAM) sobre áreas recorridas por los incendios forestales en Andalucía entre 1975 y 2022. Se acota la el periodo de estudio, cambios estructurales en los regímenes de incendios y disponibilidad de la información

1.5. Organización de los datos

No disponibilidad puntos de ignición. Ri

1.6. Depuración de los datos