

Índice general

1. Cuerpo	3
1.1. Análisis exploratorio de datos	3
1.1.1. Distribución de la variable objetivo	4
1.1.2. Análisis univariantes variables numéricas	5
1.1.3. Análisis multivariantes de las variables numéricas	7
1.1.4. Análisis de las variables categóricas	8
1.2. Modelización	9
1.2.1. Modelo 1	9
1.2.2. Modelo 2	9
1.2.3.	9
1.3. Comparación	9

Capítulo 1

Cuerpo

1.1. Análisis exploratorio de datos

En esta sección se aplicarán distintos métodos numéricos y gráficos de análisis de datos a la muestra generada siguiendo el procedimiento detallado en el capítulo anterior. Se usarán principalmente técnicas de estadística descriptiva para comprender las características del conjunto de datos y extraer conocimiento útil para el problema que se intenta abordar, predecir incendios forestales. Es importante tener presente que se trata de datos correlados espacial y temporalmente, lo que hace necesario el uso de métodos específicos para este tipo de datos. Los objetivos de esta etapa son:

1. Generar conocimiento sobre el conjunto de datos que nos permita evaluar la calidad de este, sin olvidar las limitaciones que ya se han comentado en la sección anterior.
2. Conocer, al menos de forma descriptiva, el impacto de cada variable en la variable objetivo. Este conocimiento será necesario para evaluar e interpretar los modelos que se construirán en la próxima sección.
3. Analizar las características de las distintas variables, de cara a usar posteriormente técnicas de preprocesamiento adecuadas para cada modelo.

Antes de abordar el estudio detallado de cada una de las variables y las relaciones entre estas, en la Figura 1.1 se recoge un resumen de todo el conjunto de datos, sin incluir la columna de geometría. En este resumen se puede observar que en el conjunto de datos hay 4 tipos de variables (además de la variable *geometry* que es de tipo *simple feature column POINT*, abreviado como *sfc_POINT*): cadenas de caracteres, fechas, factores y variables numéricas.

Se puede observar que hay registros en 749 municipios diferentes (de los 785 municipios de que hay en Andalucía). Probablemente el hecho de que en algunos municipios no haya habido observaciones sea debido a los datos faltantes. Las variables *municipio* y *cod_municipio* no se incorporarán a los modelos. De la misma forma, se puede ver que hay observaciones en 3691 días diferentes. El conjunto cuenta con 5 variables de tipo factor: *fire* (la variable objetivo), *WD10M*, *orientacion*, *enp* y *uso_suelo*; y con 18 variables numéricas. Aunque cada una de ellas se analizará a continuación con detalle, ya cabe hacer algunos comentarios:

- El 38 % de las observaciones se encuentran en espacios de vegetación arbustiva y/o herbácea (código 32).
- Como era de esperar, por la forma en la que se ha tomado la muestra, el conjunto está balanceado.
- El 81 % de las observaciones se encuentran fuera de Espacios Naturales Protegidos.
- Todas las variables, salvo *T2M* y *curvatura*, son positivas y la mayoría de ellas presentan una marcada distribución asimétrica hacia la derecha.
- Las variables muestran escalas muy diversas entre ellas, siendo *GWETTOP* la que presenta menor desviación típica (0.145) y *poblacion* la que tiene una desviación típica mayor (64453). Se evidencia la necesidad de incluir algún método de normalización de las variables en el preprocesamiento de los datos.

Data Summary										
Name	Values									
Number of rows	datos									
Number of columns	26									
Column type frequency:										
character	2									
Date	1									
factor	5									
numeric	18									
Group variables										
None										
Variable type: character										
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace			
1 cod_municipio	0	1	5	5	0	749	0			
2 municipio	0	1	3	32	0	749	0			
Variable type: Date										
skim_variable	n_missing	complete_rate	min	max	median	n_unique				
1 date	0	1	2002-01-02	2022-11-29	2012-08-04	3691				
Variable type: factor										
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts					
1 fire	0	1	FALSE	2	1: 10794, 0: 10752					
2 WD10M	0	1	FALSE	8	SW: 4965, W: 4867, S: 3786, SE: 3316					
3 orientacion	0	1	FALSE	9	S: 3267, SW: 3090, SE: 2956, W: 2544					
4 enp	0	1	FALSE	2	0: 17393, 1: 4153					
5 uso_suelo	0	1	FALSE	15	32: 8068, 21: 3128, 24: 2798, 22: 2786					
Variable type: numeric										
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1 T2M	0	1	24.6	5.20	-0.8	22.7	25.6	28.0	36.0	
2 GWETTOP	0	1	0.318	0.145	0.09	0.2	0.29	0.42	0.91	
3 RH2M	0	1	46.4	16.1	9.23	33.8	44.6	57.5	96.8	
4 WS10M	0	1	3.64	1.42	0.99	2.67	3.34	4.25	13.3	
5 PRECTOTCORR	0	1	0.229	1.38	0	0	0	0	46.1	
6 elevacion	0	1	494.	400.	0	169.	425.	711.	3351.	
7 pendiente	0	1	12.3	11.8	0	3.46	8.24	17.9	98.0	
8 curvatura	0	1	0.00549	0.0563	-0.294	-0.0193	-0.000100	0.0255	0.420	
9 dist_carretera	0	1	1799.	1907.	0.0706	507.	1200.	2425.	17662.	
10 dist_poblacion	0	1	902.	800.	0	379.	713.	1170.	8215.	
11 dist_electr	0	1	5253.	5640.	0.220	1216.	3462.	7276.	48069.	
12 dist_ferrocarril	0	1	15311.	13521.	0.824	5086.	11855.	21467.	85242.	
13 dist_camino	0	1	762.	741.	0.0238	240.	539.	1060.	7090.	
14 dist_sendero	0	1	5619.	4805.	0.0465	1781.	4361.	8276.	25103.	
15 poblacion	0	1	22095.	64453.	114	2252	4860	16759	704414	
16 dens_poblacion	0	1	105.	320.	2.30	12.2	30.8	71.1	4974.	
17 dist_rios	0	1	6781.	5836.	1.94	2162.	5237.	10123.	37074.	
18 NDVI	0	1	0.412	0.136	0	0.314	0.393	0.495	0.944	

Figura 1.1: Incendios durante el periodo de estudio

1.1.1. Distribución de la variable objetivo

En primer lugar, se estudiará la distribución de la *fire* espacial y temporalmente.

En la Figura 1.2 se muestran los histogramas de la variable objetivo en función del día de la semana, del mes y del año, respectivamente. En primero de ellos se observa que mientras que la distribución de los casos negativos es uniforme entre los días de la semana, en los casos positivos se aprecia un ligero aumento en el fin de semana, especialmente en

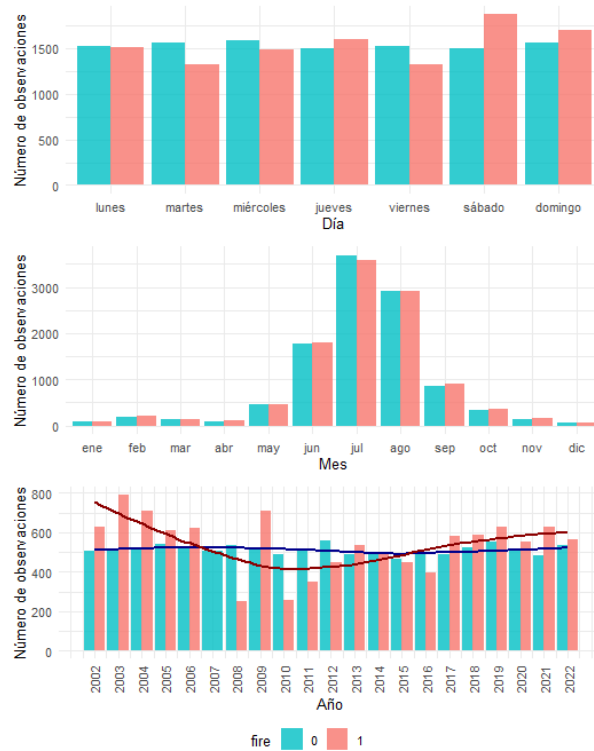


Figura 1.2: Distribución temporal de las observaciones en función de la variable objetivo.

el sábado. En el segundo histograma, se observa como las observaciones se concentran en los meses de verano y en cada mes hay una cantidad balanceada de muestras de ambas clases (esto es fruto del proceso de muestreo de las observaciones negativas, que como se ha explicado en la sección anterior, se ha llevado a cabo asegurando que la proporción de casos negativos en cada mes sea igual a la de los casos positivos). En el tercer histograma es remarcable que, mientras las observaciones negativas están uniformemente distribuidas entre los 20 años del estudio, las positivas muestran una disminución importante en los años 2008 y 2010. En el año 2007 no hay observaciones positivas, debido a que los 4 polígonos de incendios mayores de 100ha que había registrados ese año no disponían de la fecha de inicio del incendio, por lo que no pudieron usarse para el estudio. Se desconoce la causa del reducido número de incendios (mayores de 100ha) en 2007, 2008 y 2010.

Dada la clara influencia del mes y la aparente influencia del día de la semana en la aparición de incendios, estas variables serán incluidas en los modelos a través del procesamiento de la variable *date*.

En la Figura 1.3 se observa claramente como las 10752 muestras negativas están uniformemente distribuidas dentro de los límites de la Comunidad Autónoma de Andalucía, mientras que las 10794 muestras positivas se concentran a ambos lados de la cuenca del río Guadalquivir, con una mayor densidad de observaciones en la provincia de Huelva y en algunas zonas de la costa mediterránea (como ya se apreciaba en la Figura ??).

1.1.2. Análisis univariantes variables numéricas

El análisis univariante de las variables numéricas se lleva a cabo desde 3 enfoques complementarios:

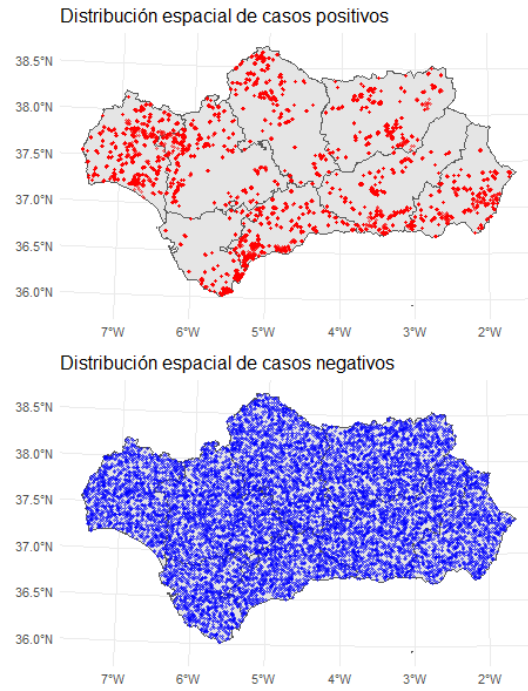


Figura 1.3: Distribución espacial de las observaciones en función de la variable objetivo.

1. A través de los resúmenes numéricos recogidos en la Figura 1.1 y del análisis gráfico de los diagramas de caja y bigote (1.4).
2. Estudiando la media mensual de cada variable en función de la variable *fire*.
3. Analizando la distribución espacial de cada variable separando por mes si corresponde.

En los *boxplots* de las variables numéricas en función de la variable *fire* (1.4) destacan varios aspectos. Por un lado, como ya se había comentado anteriormente, que las variables presentan escalas muy diferentes y que la mayoría de las variables tienen una marcada asimetría hacia la derecha. Por otro lado, es evidente la gran cantidad de valores *outliers*

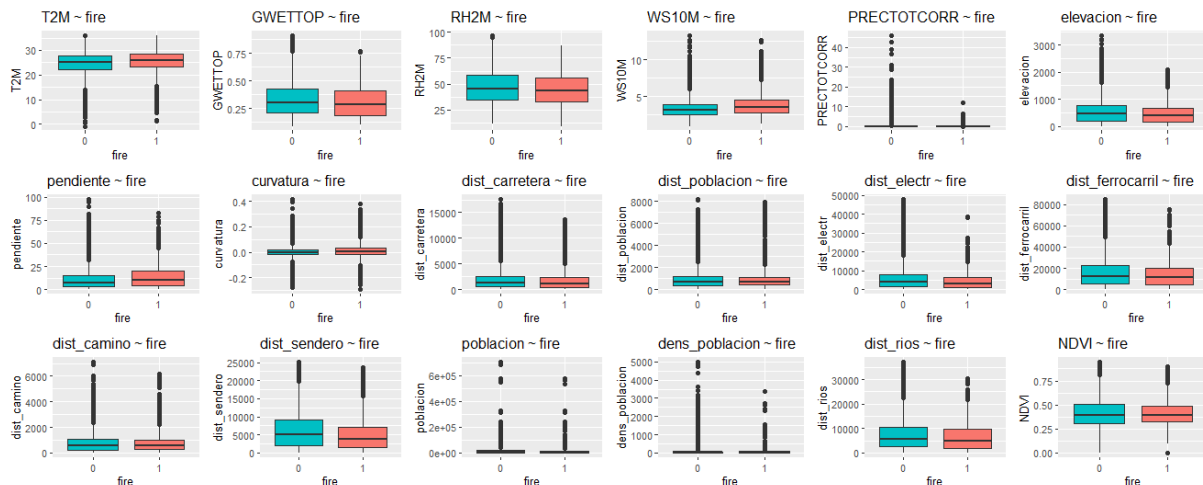


Figura 1.4: Boxplot de cada variable numérica en función de la variable objetivo.

que se observan en los datos, lo que tendrá implicaciones en los modelos que se construyan con ellos. Sin embargo, es importante destacar que no se trata de observaciones erróneas, si no que son inherentes a la naturaleza de los datos. Por ejemplo, en el caso de la variable PRECTOTCORR el valor máximo observado es 46.06mm en un día, un valor elevado que sin duda es atípico en esta región de clima seco, pero sin embargo, posible. Es también remarcable que todas las variables presentan una variabilidad similar en ambos niveles del factor *fire*, lo que indica que no será un problema de clasificación trivial. A priori, solo con los diagramas de caja y bigotes y los resúmenes numéricos es difícil llegar a abarcar más, sin embargo, sí pueden observarse sutiles diferencias entre las distribuciones de algunas variables para ambos niveles del factor *fire*.

Dada la naturaleza temporal de algunas variables, el análisis gráfico de los *boxplots* resulta insuficiente. Con el fin de considerar la componente estacional de las variables climáticas y de vegetación, a continuación, se estudiará la media mensual de cada una de estas variables función de la variable objetivo.

En la Figura 1.5 se puede observar como en casi todos los meses, la temperatura media mensual es superior en las observaciones en las que se ha registrado un incendio forestal.

En la Figura 1.6 se puede observar como en todos los meses la media mensual de la humedad relativa del aire a 2m sobre la superficie es menor en las observaciones en las que se ha registrado un incendio forestal. Sin embargo, las diferencias se reducen durante los meses de verano.

En la Figura 1.7 se observa una clara diferencia en la media mensual de las precipitaciones diarias en función del de si se ha registrado o no un incendio forestal en la observación, siendo significativamente mayor en este último caso.

En la Figura 1.8 se observa un gráfico similar al de la humedad relativa del aire, con valores medios más elevados en las observaciones en las que no se han registrado incendio forestal. Sin embargo, también parece que las diferencias son más reducidas durante la estación estival.

En la Figura 1.9 se observa como durante todos los meses, la media mensual de la velocidad del viento a 10m sobre la superficie es mayor en los registros afectados por incendio forestal.

Como se observa en la Figura 1.10, las diferencias entre los casos en los que se ha registrado incendio y los que no en términos del NDVI no están claras.

En el [Apéndice: Gráficos espaciales EDA] se recogen los gráficos espaciales y espacio_temporales de todas las variables numéricas. En ellos se refleja como la valores de las variables en estudio son coherentes con la realidad. Además, permiten una comprensión mayor de la distribución espacial de las variables en el área de estudio.

1.1.3. Análisis multivariantes de las variables numéricas

En la Figura 1.11 se muestra un gráfico con las correlaciones entre las variables. La interpretación es sencilla, cuanto más intenso sea el color y cuanto mayor sea la excentricidad de la elipse, mayor será la correlación (en valor absoluto) para ese par de variables. De esta forma, se observa que las variables más correlacionadas en la muestra son:

- *T2M* con *RH2M* (negativamente, -0.71)

- *T2M* con *GWETTOP* (negativamente, -0.69)
- *GWETTOP* con *R2HM* (positivamente, 0.68)
- *poblacion* con *dens_poblacion* (positivamente, 0.63)

En la Figura 1.12 se muestra el gráfico de coordenadas paralelas de las variables tipificadas a una normal estándar, es decir, restándoles la media y dividiendo por la desviación típica. Este gráfico complementa la información de los *boxplots*, pues refleja también las relaciones entre las variables. Si bien es cierto que al tener un número de observaciones el gráfico no es tan claro, pueden hacerse algunas observaciones importantes.

En primer lugar, se observa que la variable con mayor variabilidad (una vez tipificada) es *PRECTOTCORR*, que presenta bastantes valores atípicos, todos ellos en observaciones en las que no se ha registrado incendio. También destacan en este sentido *dens_poblacion* y *poblacion*, entre las que además puede observarse que no hay una relación lineal clara (hay municipios con un elevado número de habitantes pero con una densidad de población reducida y viceversa). Además, puede verse que todas las variables tienen una marcada asimetría positiva (salvo *curvatura*, *T2M* y *NDVI*). Este gráfico es útil también pues permite ver a que clase de *fire* corresponden los valores más atípicos de cada variable. Por ejemplo: la mayor parte de los valores más elevados de *WS10M*, *dist_poblacion*, *curvatura* y *dist_camino* se dan en observaciones positivas, mientras que en *PRECTOTCORR*, *elevacion*, *GWETTOP*, *dist_Carretera*, *dist_electr* y *dist_rios* sucede lo contrario.

Los resultados de aplicar análisis de componentes principales sobre la matriz de correlaciones de las 18 variables numéricas se muestran en la Figura 1.13. Como se puede observar, se necesitan al menos 11 componentes principales para lograr explicar el 80 % de la varianza de la muestra, y 14 para alcanzar el 90 % de la varianza de los datos. Estos resultados se aplicarán más adelante en los modelos, pero a nivel meramente explicativo ya indican que se trata de un conjunto de datos complejo en cuanto a la dimensión real de estos.

1.1.4. Análisis de las variables categóricas

Las variables categóricas se analizarán a través de los histogramas de cada variable en función de la variable *fire* (Figura 1.14).

En la variable *WD10M* cabe destacar la escasez de observaciones con dirección del viento norte. En el histograma no se observa una clara relación de esta variable con la variable objetivo, aunque entre las observaciones con viento con dirección sur o suroeste hay más observaciones negativas y entre las que tienen dirección noroeste o este hay una mayor presencia de observaciones positivas.

En el caso de la variable *orientación*, la relación tampoco está clara, aunque puede verse una mayor proporción de observaciones positivas en las superficies con orientación sur (sureste, sur y suroeste).

En términos de la variable *enp* por si sola no se observan diferencias significativas entre ambas clases.

La variable *uso_suelo* sí que muestra una distribución marcadamente diferenciada entre ambas clases. La mayoría de las observaciones positivas se dan en espacios de vegetación

arbustiva y/o herbácea, clase en la que hay casi el doble de observaciones positivas que negativas. En tierras de labor y cultivos permanentes la proporción de observaciones negativas es mucho mayor, mientras que en zonas agrícolas heterogéneas y en espacios abiertos con poca o sin vegetación hay una mayor presencia de observaciones positivas. También es relevante el hecho de que casi la totalidad de las observaciones se encuentran en zonas agrícolas y en zonas forestales, mientras que en las demás clases la proporción de observaciones es mucho menor (3.5 de total). Es por ello que antes de construir los modelos, todas las categorías de uso de suelo que no se corresponden con zonas agrícolas o forestales se agruparán en el nivel *Otro*. En ?? puede observarse la distribución espacial de esta variable.

1.2. Modelización

1.2.1. Modelo 1

1.2.2. Modelo 2

1.2.3. ...

1.3. Comparación

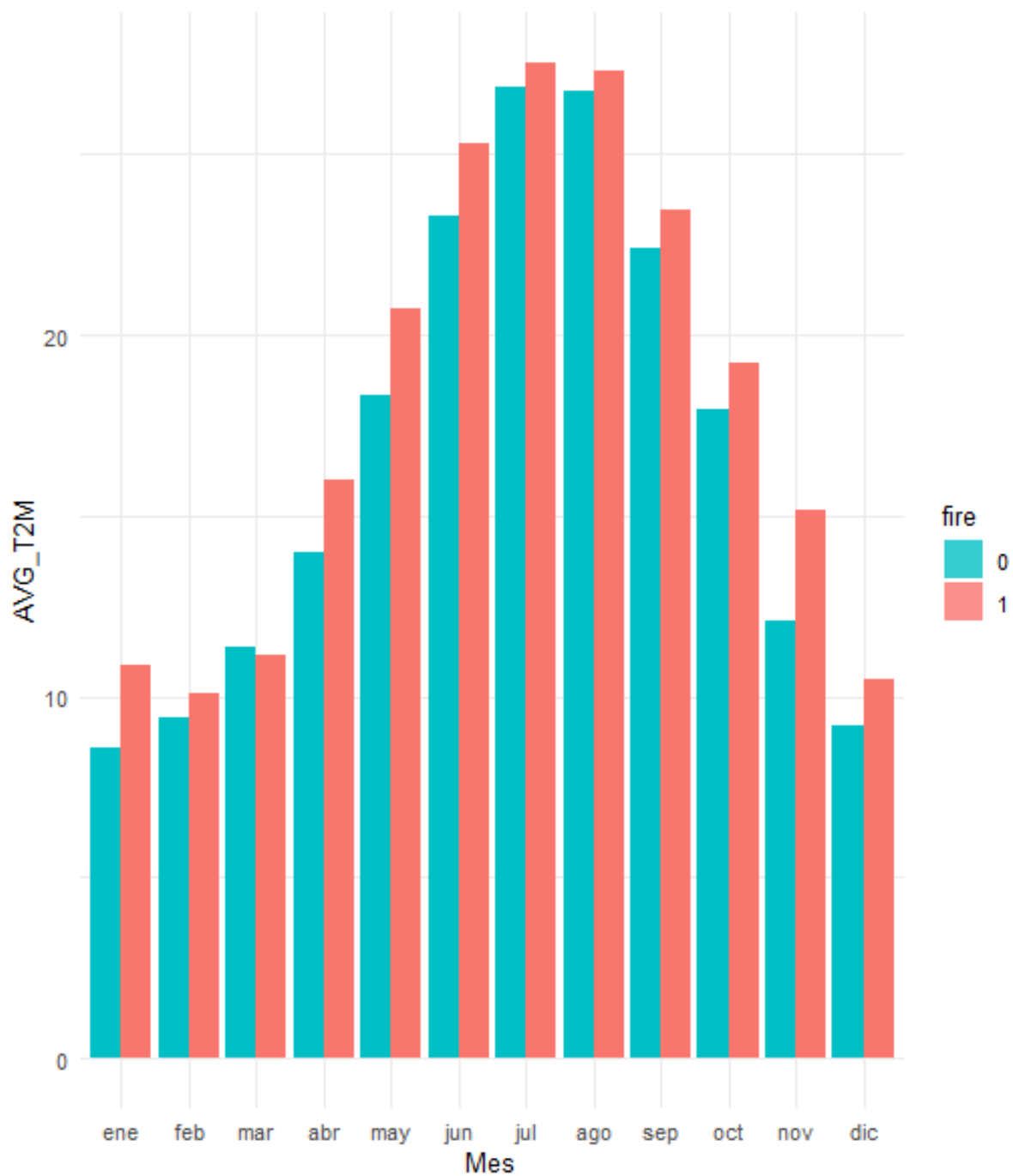


Figura 1.5: Media mensual de la temperatura en función de fire.

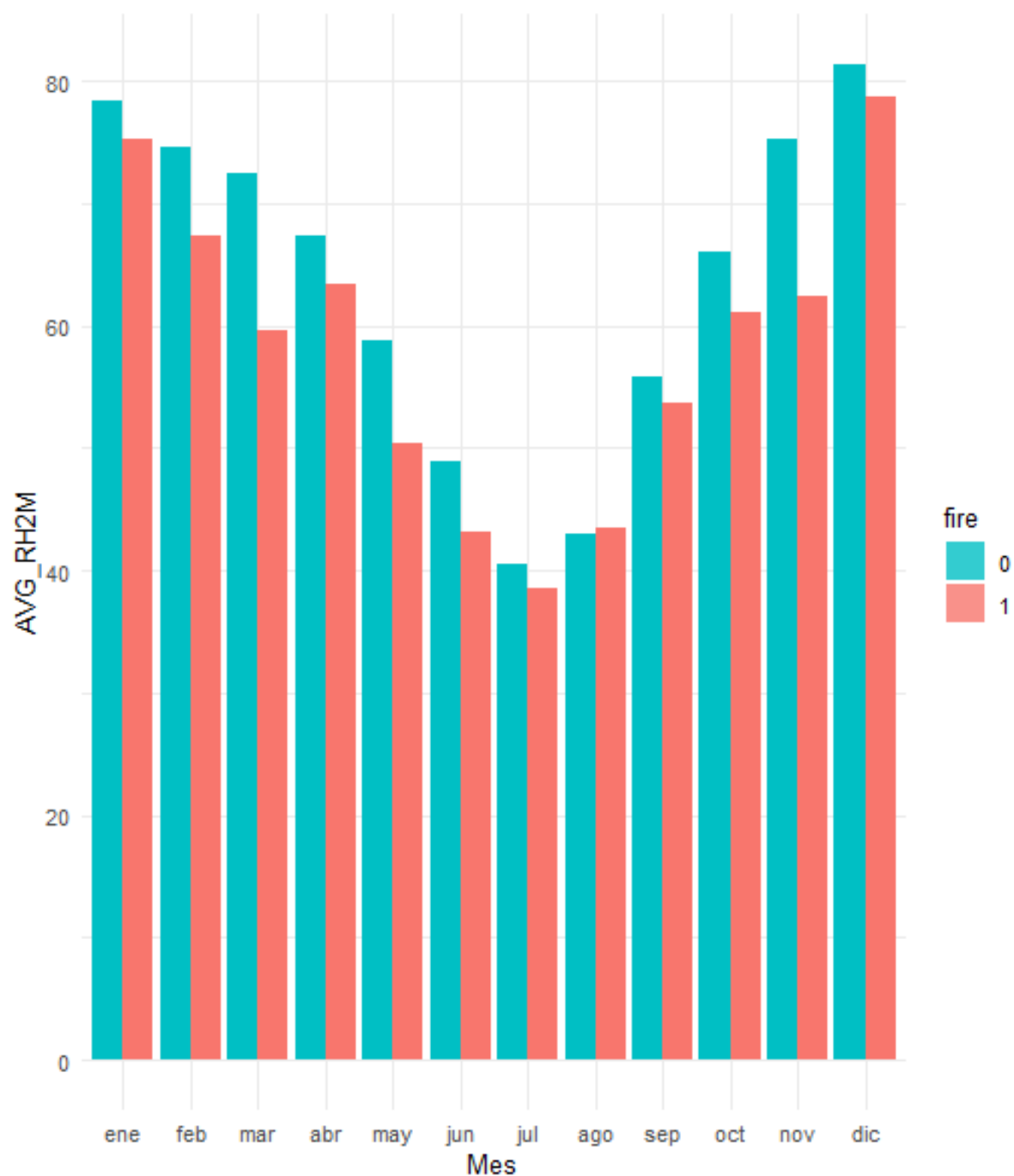


Figura 1.6: Media mensual de la humedad relativa en función de fire.

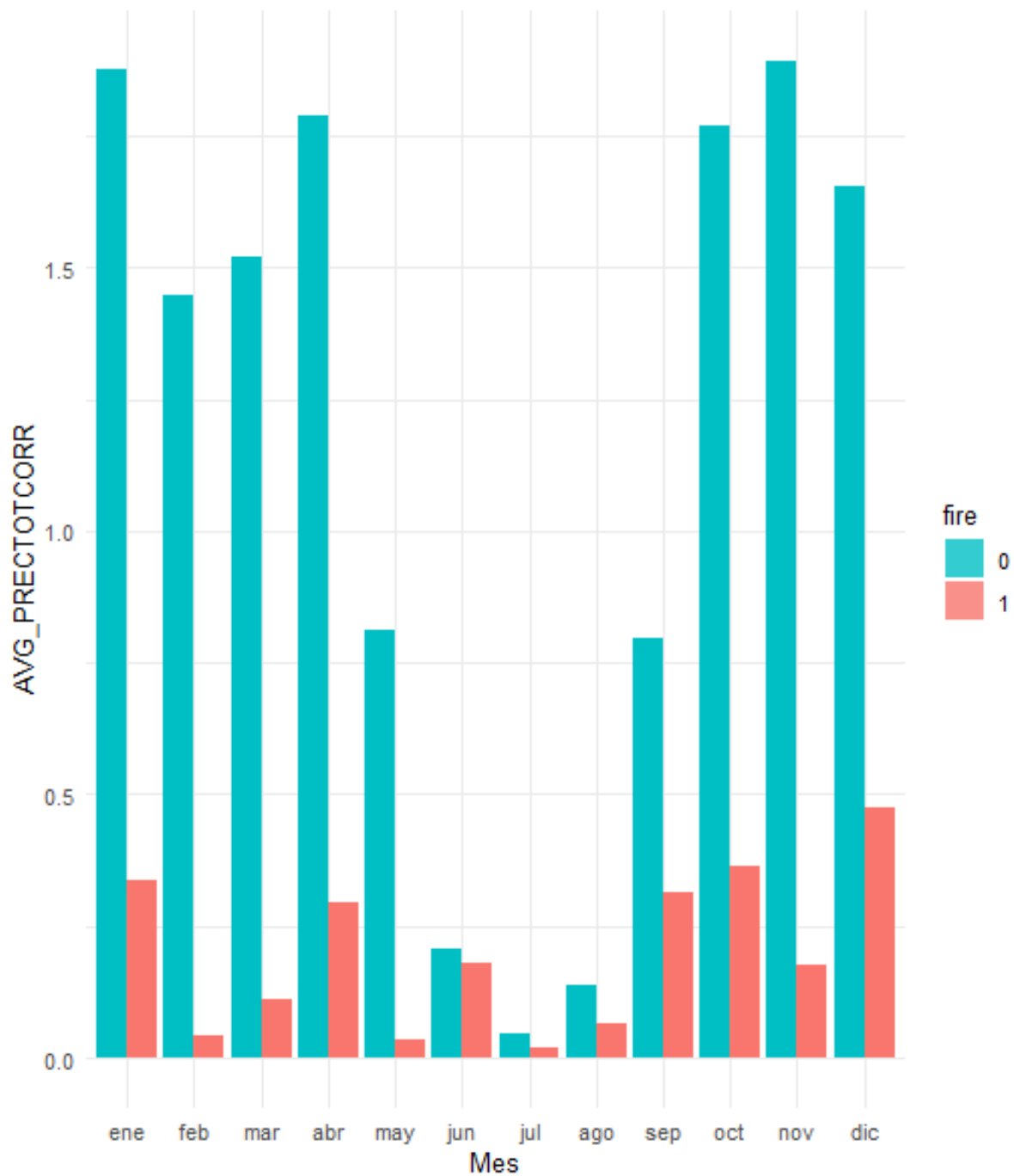


Figura 1.7: Media mensual de la precipitaciones en función de fire.

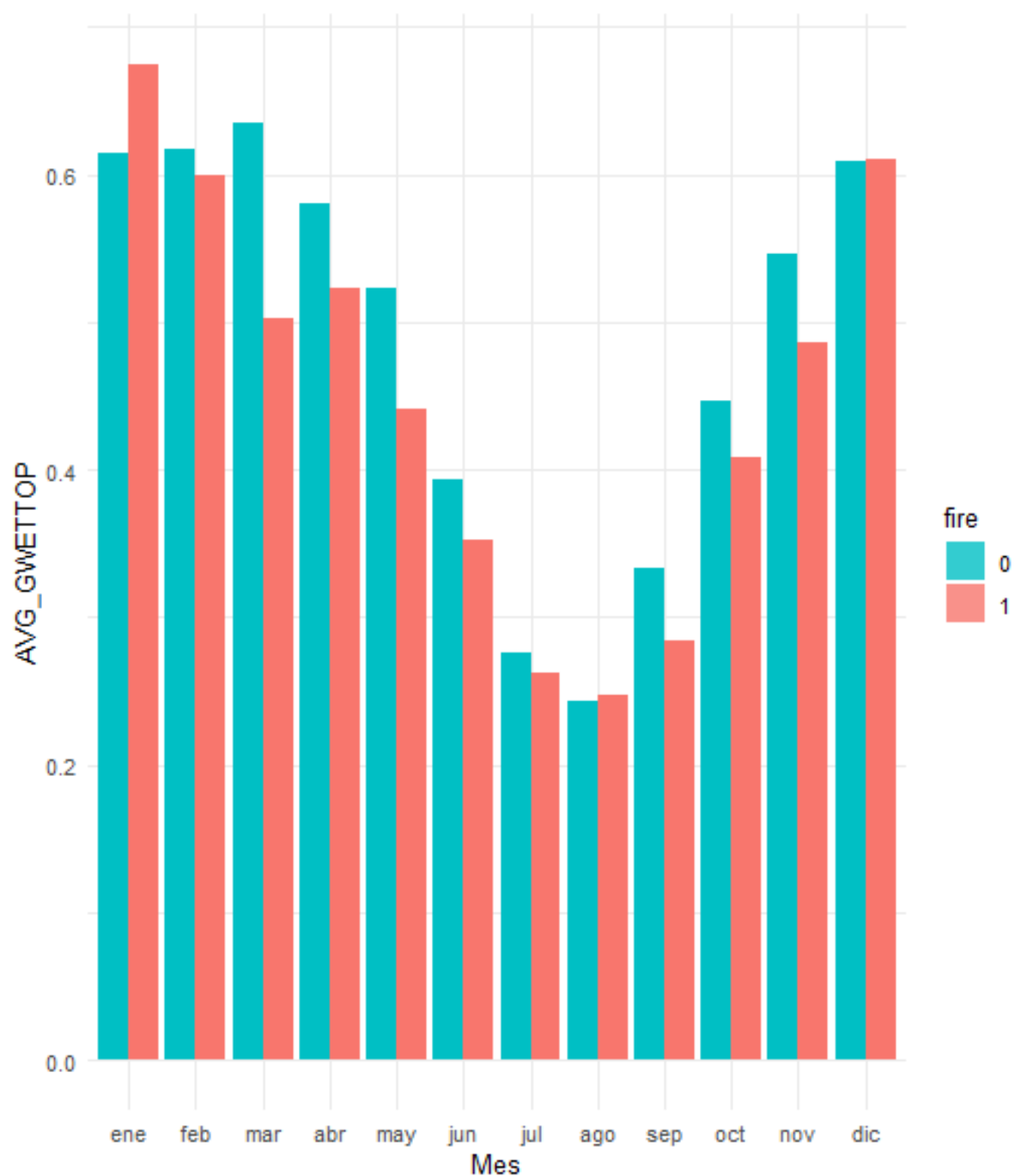


Figura 1.8: Media mensual de la humedad del suelo en función de fire.



Figura 1.9: Media mensual de la humedad del suelo en función de fire.

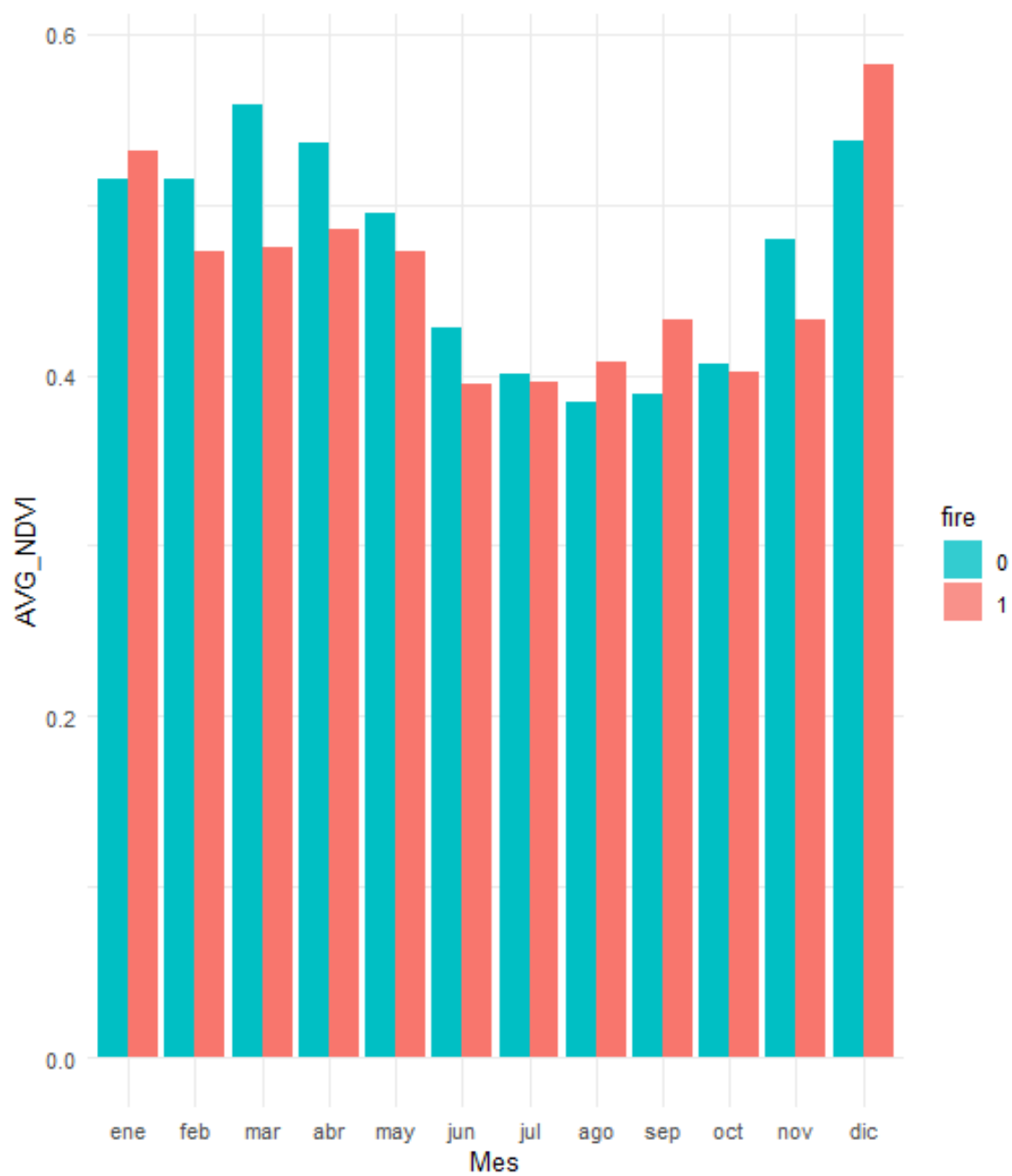


Figura 1.10: Media mensual de la humedad del suelo en función de fire.

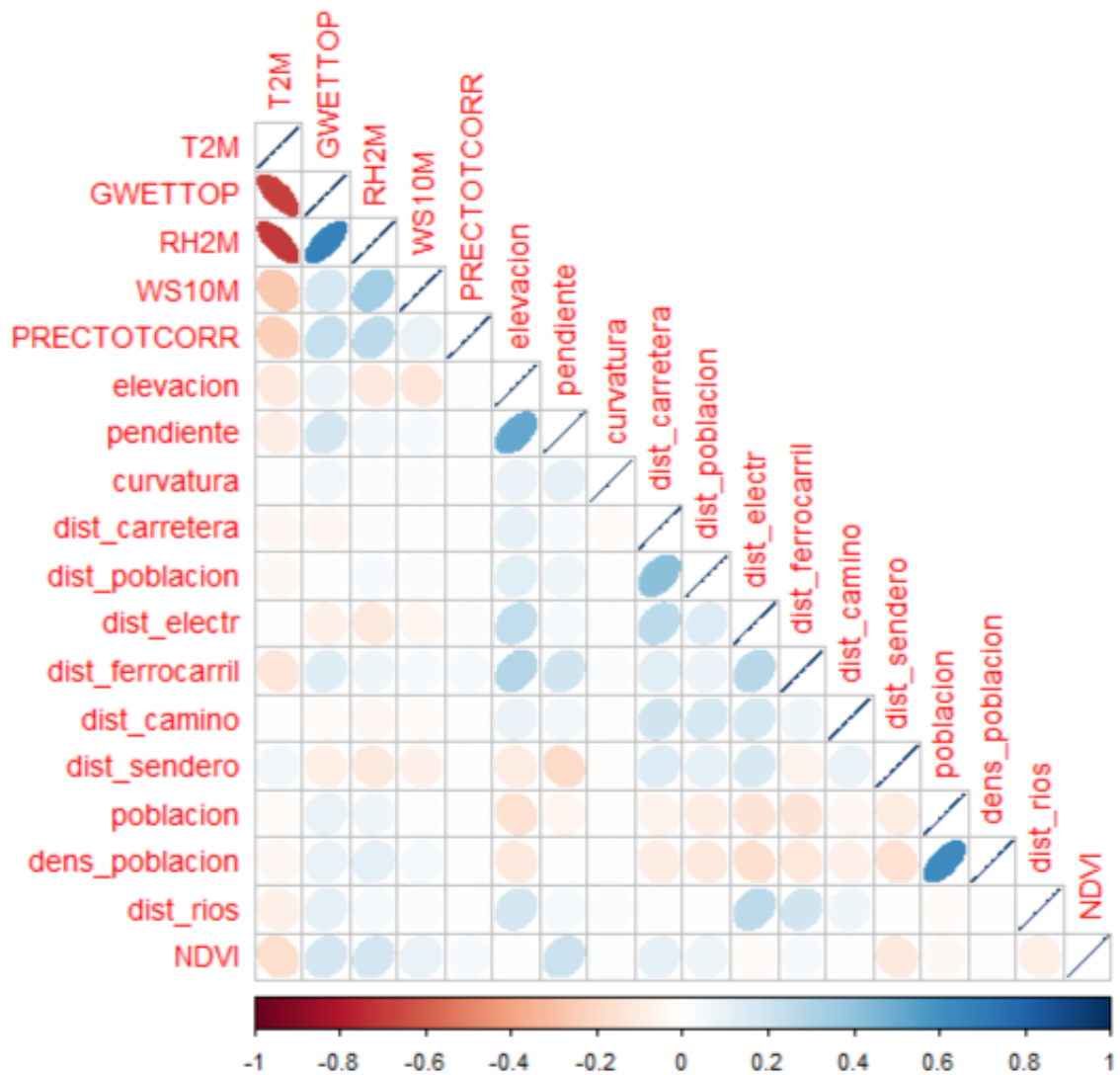


Figura 1.11: Correlaciones entre variables numéricas

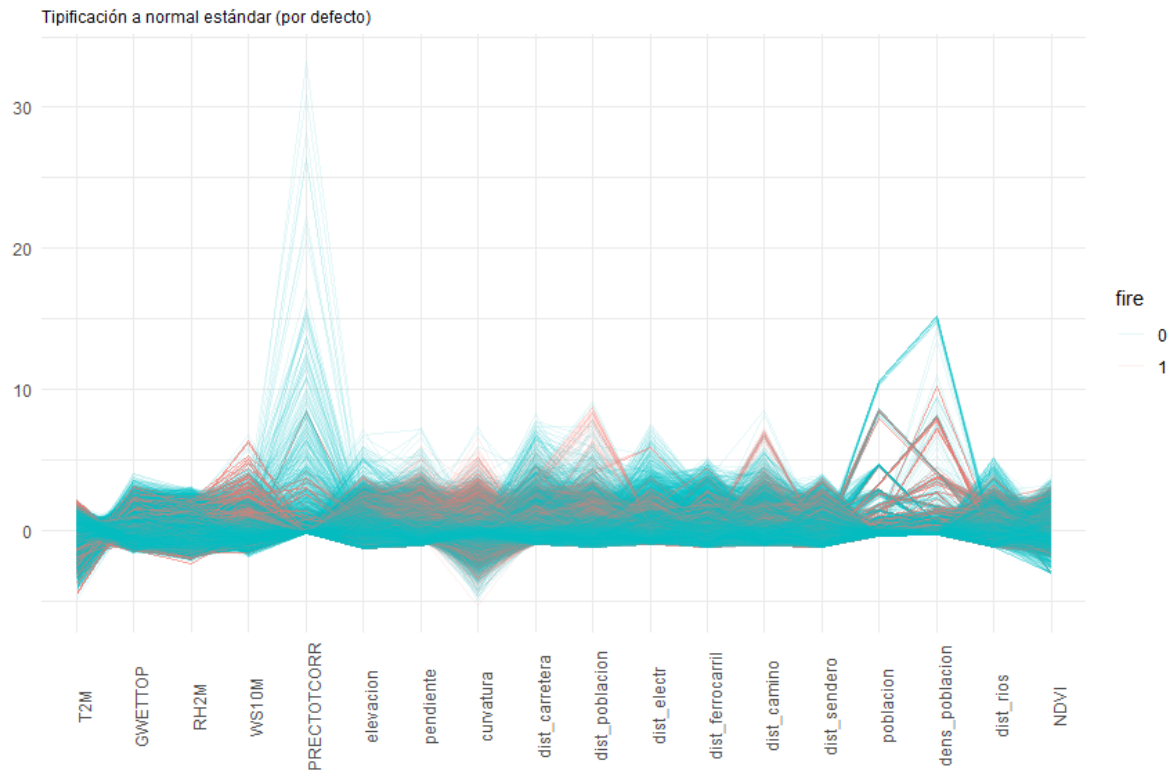


Figura 1.12: Gráfico de coordenadas paralelas.

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	1.6830901	1.5509109	1.26704790	1.18114361	1.13860921	1.00151777	0.98498153	0.92948032	0.92853341
Proportion of Variance	0.1573773	0.1336291	0.08918946	0.07750557	0.07202394	0.05572433	0.05389937	0.04799631	0.04789857
Cumulative Proportion	0.1573773	0.2910065	0.38019595	0.45770152	0.52972546	0.58544979	0.63934916	0.68734547	0.73524404
	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18
Standard deviation	0.91036573	0.86805852	0.85739620	0.78965012	0.72570700	0.65041713	0.60872397	0.52343451	0.48001944
Proportion of Variance	0.04604254	0.04186253	0.04084046	0.03464152	0.02925837	0.02350236	0.02058583	0.01522132	0.01280104
Cumulative Proportion	0.78128658	0.82314912	0.86398958	0.89863109	0.92788946	0.95139182	0.97197765	0.98719896	1.00000000

Figura 1.13: PCA sobre la matriz de correlaciones de las variables numéricas

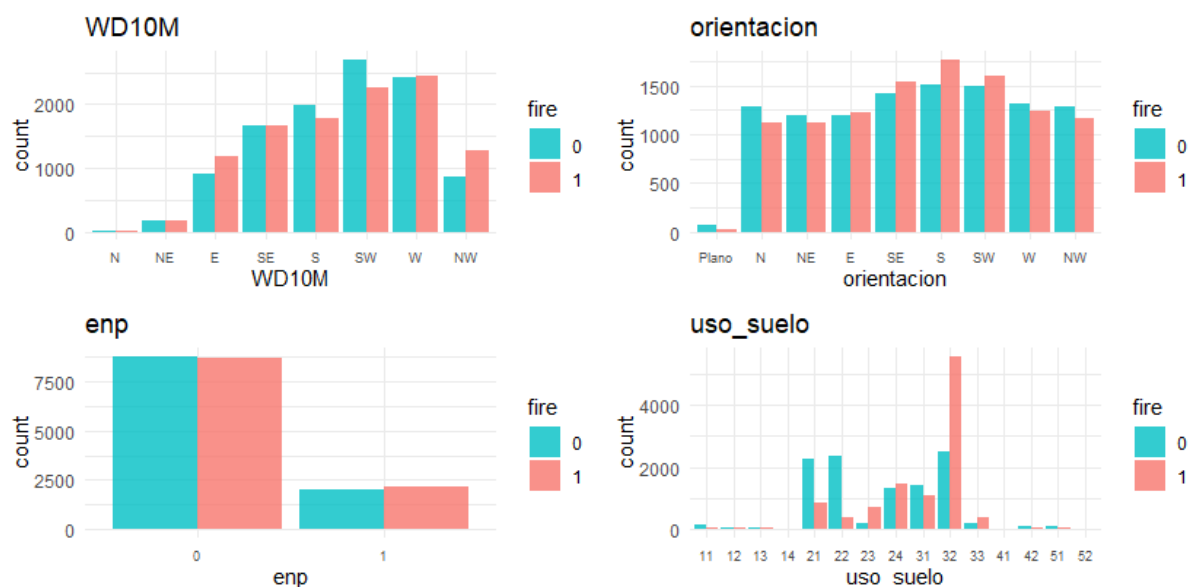


Figura 1.14: Histogramas de las variables categóricas en función de fire.