FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2017/2018

Trabajo de Fin de Máster

APLICACIÓN DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE ALERTAS POR NO_2 EN MADRID

Alumno: Ángel Luis Calzada Florenciano
Tutor: Juana María Alonso Revenga

Noviembre 2018

UNIVERSIDAD COMPLUTENSE
MADRID

ÍNDICE GENERAL

1. Introducción

La contaminación del aire impacta significativamente en la salud de la población europea, especialmente en las zonas urbanas. Asimismo, tiene un considerable impacto económico, puesto que aumenta los costes sanitarios derivados de las enfermedades que provoca. Esto ha llevado a la calidad del aire a convertirse en una de las principales cuestiones políticas de las últimas décadas en el ámbito europeo.

La importancia de este tema se ve reflejada en las múltiples Directivas Europeas que tratan de establecer unos mínimos de calidad de aire para proteger a la población de excesivos niveles de contaminación.

La progresiva regulación que se ha ido aprobando desde finales de los 70 se ha consolidado en la Directiva 2008/50/CE relativa a la calidad del aire ambiente y a una atmósfera más limpia en Europa. Haciendo una muy breve puntualización jurídica, una Directiva Europea es una norma de ámbito supranacional en la que se fija una serie de objetivos a alcanzar, pero deja libertad a los Estados miembros para elegir los medios que consideren más convenientes para alcanzarlos. Dicha norma comunitaria, en su artículo primero señala cuáles son sus objetivos:

- *Definir y establecer objetivos de calidad del aire ambiente para evitar, prevenir o reducir los efectos nocivos para la salud humana y el medio ambiente en su conjunto.*
- *Evaluar la calidad del aire ambiente en los Estados miembros basándose en métodos y criterios comunes.*
- *Obtener información sobre la calidad del aire ambiente con el fin de ayudar a combatir la contaminación atmosférica y otros perjuicios y controlar la evolución a largo plazo y las mejoras resultantes de las medidas nacionales y comunitarias.*
- *Asegurar que esa información sobre calidad del aire ambiente se halla a disposición de los ciudadanos.*
- *Mantener la calidad del aire, cuando sea buena, y mejorarla en los demás casos.*
- *Fomentar el incremento de la cooperación entre los Estados miembros para reducir la contaminación atmosférica.*

Los contaminantes más serios en términos de deterioro para la salud son las partículas en suspensión PM2.5 y los óxidos de Nitrógeno NOx. La Agencia Europea del Medio Ambiente ha estimado el número de muertes prematuras que tuvieron lugar en los 28

países miembros de la UE durante 2014, dando como resultado 399.000 fallecimientos derivados de una prolongada exposición a altas concentraciones de partículas en suspensión y 75.000 muertes al año relacionadas con el NO2 [1]. El presente trabajo se centra en el estudio del segundo de los citados contaminantes, el NO2.

1.1. Acerca del Dióxido de Nitrógeno (NO_2)

El NO_2 se forma a nivel de suelo a partir de emisiones relacionadas con la quema de combustibles fósiles de vehículos, centrales de energía, industrias, etc.

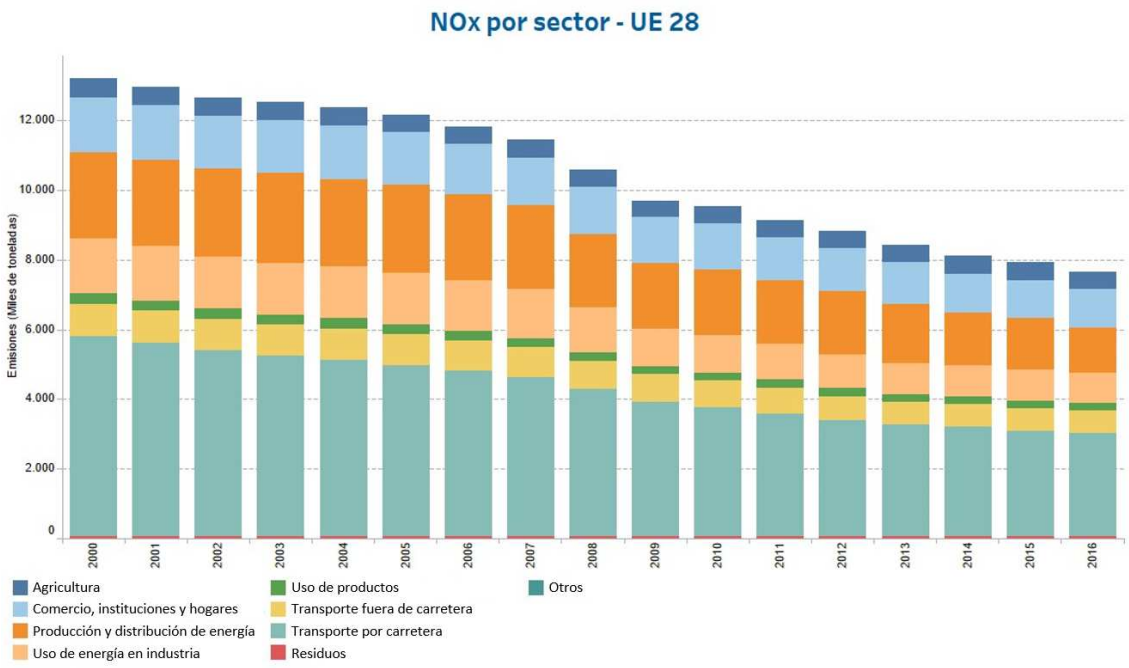


Figura 1. Evolución emisiones dióxidos de nitrógeno UE28. (Fuente: Agencia Europea del Medio Ambiente)

Además de contribuir en la formación de ozono, se relaciona al NO2 con efectos nocivos sobre el sistema respiratorio. Los dióxidos de nitrógeno reaccionan con el amoniaco, con la humedad y otros compuestos para formar pequeñas partículas. Estas pequeñas partículas pueden penetrar profundamente en las partes sensibles de los pulmones.

La evidencia científica vincula exposiciones cortas a NO2 (desde 30 minutos hasta 24 horas) con efectos adversos respiratorios, incluida la inflamación de las vías respiratorias en personas sanas y el aumento de los síntomas en personas que padecen de asma. Los estudios también muestran que existe conexión entre la exposición a corto plazo a este

[1] Fuente: Informe 13/2017 Agencia Europea Medio Ambiente sobre calidad del aire en Europa

contaminante y el aumento de visitas a las emergencias hospitalarias por problemas respiratorios.

El NO2 también es uno de los causantes de la conocida lluvia ácida, ya que al reaccionar con el vapor de agua produce ácido nítrico. Los efectos sobre la agricultura, la ganadería, los bosques, los suelos y las aguas son muy graves.

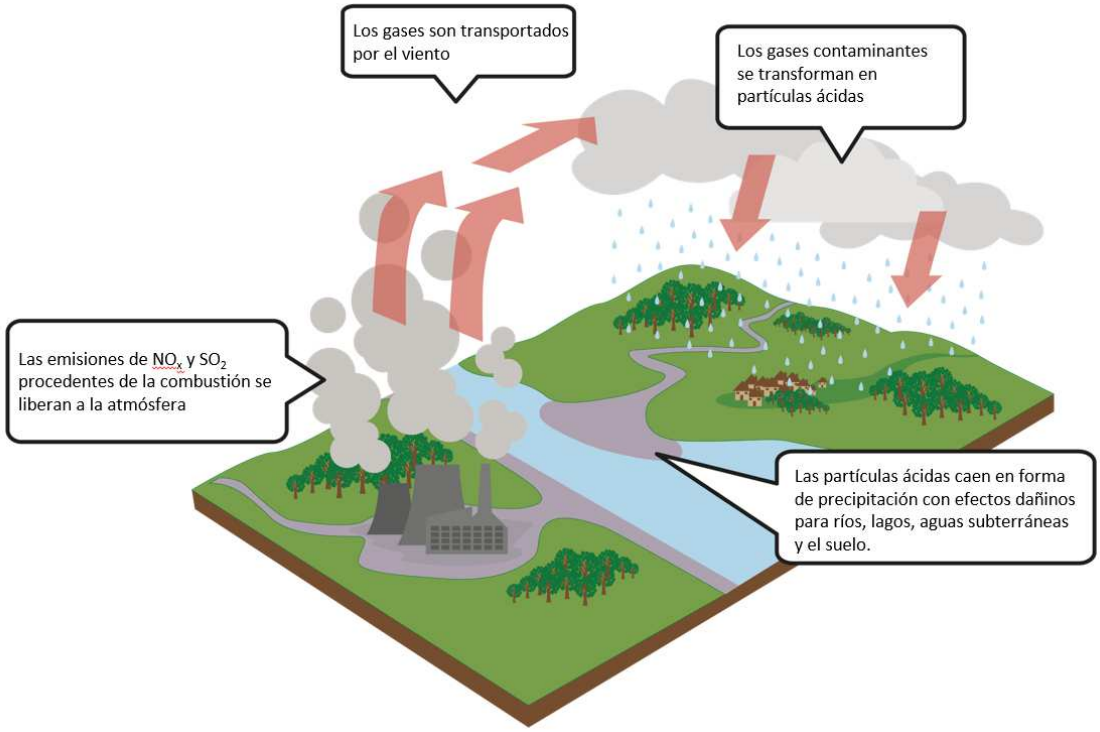


Figura 2. Esquema de formación de lluvia ácida

En este sentido, llevar a cabo acciones encaminadas a reducir el impacto de la contaminación atmosférica ha sido un objetivo primordial.

Los distintos estados miembros de la UE han ido incorporando progresivamente a su legislación distintas medidas para alcanzar los objetivos planteados en la Directiva 2008/50/CE. Algunas de estas medidas consisten en protocolos de actuación en materia de tráfico ante episodios de alta contaminación ya que, como se muestra en la Figura 1, el tráfico por carretera es uno de los principales orígenes del NO_2. Madrid, por ejemplo, ha puesto en marcha un protocolo de actuación en episodios de alta contaminación. Esta ha sido una de las motivaciones para realizar el presente trabajo: tratar de estudiar si, a través de la minería de datos podemos predecir si un día se dará o no un episodio de alta contaminación.

1.2. Adaptación a la Directiva 2008/50/CE en España

Centrándonos en España, el Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire, establece umbrales de alerta para algunos agentes contaminantes, entre ellos el dióxido de nitrógeno. Se define el umbral de alerta como *"el nivel a partir del cual una exposición de breve duración supone un riesgo para la salud humana, que afecta al conjunto de la población y que requiere la adopción de medidas inmediatas.*" El valor del umbral de alerta para el dióxido de nitrógeno está establecido en 400 microgramos/m3 durante tres horas consecutivas en lugares representativos de la calidad del aire, en un área de al menos 100 km2 o en una zona o aglomeración entera, si esta última superficie es menor. El citado Real Decreto establece asimismo un valor límite horario para la protección de la salud de dióxido de nitrógeno de 200 microgramos/m3 (nivel de aviso) que no debe superarse más de 18 horas al año en ninguna de las estaciones de la red.

1.3. Protocolo de Contaminación en Madrid

El Ayuntamiento de Madrid, para llevar a cabo el control de la calidad del aire de la ciudad, dispone del Sistema de Vigilancia, Predicción e Información de la Calidad del Aire que permite conocer, de forma continua y en tiempo real, las concentraciones de contaminantes, con el principal objetivo de proteger la salud de la población y reducir al máximo las situaciones de riesgo.

Las elevadas concentraciones son originadas fundamentalmente por las emisiones del tráfico, y tienen lugar en situaciones con condiciones meteorológicas especialmente adversas, que requieren la ejecución de medidas para reducir los niveles de contaminación y la duración de los episodios, y evitar que llegue a superarse el valor límite horario y que se llegue a alcanzar el umbral de alerta.

En el municipio de Madrid los valores límites horarios se han superado varias veces en distintas estaciones como puede verse en el siguiente gráfico:

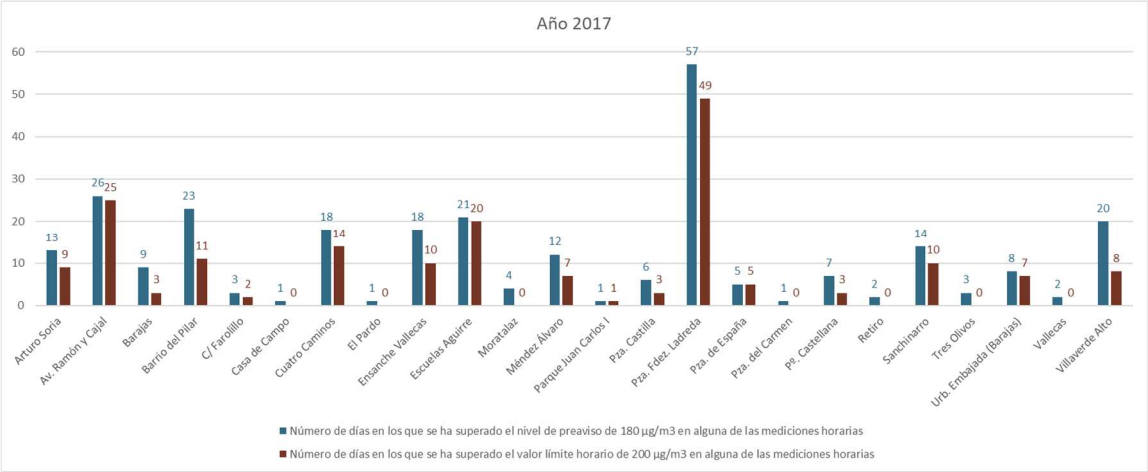


Figura 3. Número de días con NO2 por encima de nivel de preaviso, 2017. (Fuente: Elaboración propia)

Por ello, el Ayuntamiento de Madrid ha establecido una división en zonas de tal manera que las situaciones de alerta puedan declararse en áreas más reducidas con alta densidad de población. Igualmente se definen unos niveles de aviso que permitan, en el caso de registrarse concentraciones elevadas de dióxido de nitrógeno, la puesta en marcha de mecanismos de información adicionales, que sirvan tanto para proteger la salud de los ciudadanos como para sensibilizar a la opinión pública, recabar su colaboración para la reducción de la contaminación y, en función de los niveles alcanzados y la duración del episodio, llevar a cabo medidas de restricción de tráfico en la ciudad y sus accesos para reducir los niveles de contaminación y evitar que se alcance la situación de alerta.

La ciudad de Madrid, a efectos de aplicación del Protocolo de medidas a adoptar durante episodios de alta contaminación, se ha dividido en cinco zonas. Cada una de las estaciones de medición del aire se encuentra enmarcada en alguna de estas zonas siendo la distribución la que se indica a continuación:

Zona	Estaciones
1 (Interior M30)	7 de tráfico (Escuelas Aguirre, Castellana, Plaza de Castilla, Ramón y Cajal, Cuatro Caminos, Plaza de España y Barrio del Pilar) + 3 de fondo (Plaza del Carmen, Méndez Álvaro y Retiro)
2 (Sureste)	1 de tráfico (Moratalaz) + 2 de fondo (Vallecas y Ensanche de Vallecas)
3 (Noreste)	5 de fondo (Arturo Soria, Sanchinarro, Urbanización Embajada, Barajas pueblo y Tres Olivos) + 1 suburbana (Juan Carlos I)
4 (Noroeste)	2 suburbanas (El Pardo y Casa de Campo)
5 (Suroeste)	1 de tráfico (Fernández Ladreda) + 2 de fondo (Farolillo y Villaverde)

Figura 4. Estaciones calidad aire por zona. (Fuente: Ayto. Madrid - Protocolo de actuación para episodios de contaminación por dióxido de nitrógeno)

Las zonas se han definido atendiendo a los siguientes criterios:

- La distribución de la población.
- La tipología y distribución de estaciones del sistema de vigilancia de la calidad del aire.
- El viario de tráfico, para facilitar la implantación de posibles actuaciones de restricción del mismo

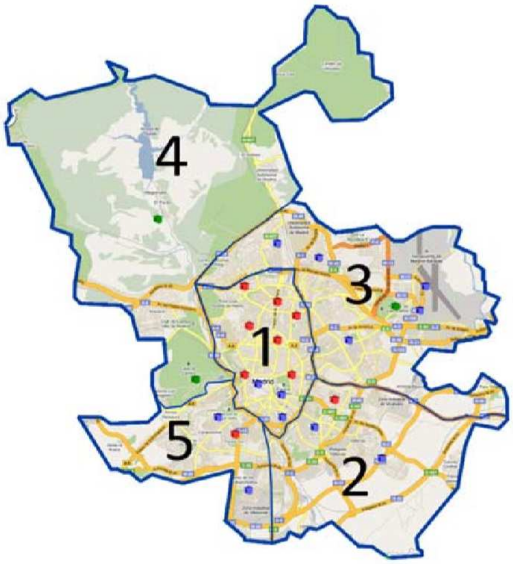


Figura 5. Delimitación de zonas a efectos de aplicación del protocolo de actuación para episodios de contaminación por dióxido de nitrógeno. (Fuente: Ayto. Madrid)

- Zona 1: área comprendida en el interior de la M30.
- Zona 2: área delimitada por la Avda. de Andalucía, Calle 30, la autovía M23 continuando por la R3 y hasta el límite del término municipal de Madrid.
- Zona 3: área delimitada por la autovía M23 y la continuación de la R3, Calle 30 hasta la M40 en la zona oeste y desde allí limita al norte con la M40 hasta el límite del término municipal de Madrid. Esta zona incluye parte del Aeropuerto de Barajas.
- Zona 4: área delimitada por el contorno del límite del municipio de Madrid por el norte, la M40 norte, Calle 30 hasta la A5 y el límite del municipio.
- Zona 5: área delimitada por el contorno sur de la Casa de Campo, Calle 30, Avda. de Andalucía y el término municipal de Madrid.

Se establecen tres niveles de actuación en función de las concentraciones de dióxido de nitrógeno que se registren en las zonas que se han definido:

Nivel de preaviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 180 microgramos/m3 durante dos horas consecutivas.

Nivel de aviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 200 microgramos/m3 durante dos horas consecutivas.

Nivel de alerta: cuando en tres estaciones cualesquiera de una misma zona (o dos si se trata de la zona 4) se superan los 400 microgramos/m3 durante tres horas consecutivas.

En el presente trabajo, se han tomado como referencia estas zonas definidas por el Ayuntamiento de Madrid para la agrupación de estaciones de calidad del aire, climatológicas y de los puntos de medición del tráfico (profundizamos sobre ello en páginas posteriores).

2. Objetivos

Este Trabajo de Fin de Máster (TFM) pretende alcanzar los siguientes objetivos:

I. Una parte muy importante de cualquier proyecto de minería de datos consiste en la recopilación de datos. Se trata de una fase larga y a veces pesada si se hace manualmente. En este sentido, uno de los objetivos del presente TFM consiste en programar una herramienta rápida y eficaz que sea capaz de hacer consultas masivas a una API con el fin de alimentar una base de datos propia sobre la que trabajar posteriormente.

   En nuestro caso, se ha llegado a dicha solución mediante la programación en Python de un script que recopila datos climatológicos históricos de las distintas estaciones de medición de la Comunidad de Madrid. Dichos datos servirán de input para la posterior modelización de cara a predecir el nivel máximo de NO2 de un día.

II. Crear modelos de cara a predecir si un determinado día se superará el nivel de preaviso de NO2 (180 microgramos/m3) a partir de la información climatológica y de tráfico recopilada de los distintos orígenes. Para ello, se ha partido de ficheros proporcionados por el Ayuntamiento de Madrid que recogen el histórico de mediciones de NO2 de los distintos medidores distribuidos por toda la ciudad. Se ha optado por trabajar con una profundidad temporal de 3 años (mediciones

horarias desde 01/07/2015 hasta 30/06/2018). Teniendo datos de 24 estaciones de medición, partimos de una base de datos con 629.760 mediciones horarias.

Actualmente, las acciones a llevar a cabo se ejecutan de una manera reactiva una vez se ha verificado que existe el episodio de alta contaminación. La idea es, si obtenemos un modelo con buena capacidad predictiva, poder llevar a cabo acciones de una manera proactiva, adelantándonos a un escenario de alta contaminación.

III. Realizar una comparativa de bondad de ajuste entre distintos métodos de clasificación utilizados en minería de datos tales como Regresión Logística, Redes Neuronales, Random Forest y Gradient Boosting.

3. Metodología

Para el primero de los objetivos, facilitar o automatizar la recopilación de datos climatológicos se ha programado en Python 3.7.0 un script que realiza esta tarea. En su programación se han utilizado ciertos módulos y funciones que se detallan en el epígrafe correspondiente.

Para alcanzar los objetivos II y III ha sido necesario ejecutar ciertas fases. La primera de ellas, que podríamos denominar fase de preprocesamiento de la información, engloba toda la importación, depuración, homogeneización a las mismas unidades de medida, agrupación y tratamiento de los datos de cara a introducir el menor ruido posible en los modelos que se calculen. Una vez hemos considerado que la calidad de los datos con los que estamos trabajando es buena, hemos utilizado distintas técnicas de clasificación binaria: Regresión Logística, Redes Neuronales, Random Forest y Gradient Boosting.

3.1. Orígenes de datos

Para la realización del presente trabajo, se ha partido de distintos orígenes de información y diferentes métodos para su obtención. La profundidad temporal de los datos con los que hemos trabajado es de 3 años completos, más concretamente, se ha trabajado con datos históricos comprendidos entre el 01/07/2015 hasta el 30/06/2018, ambos inclusive.

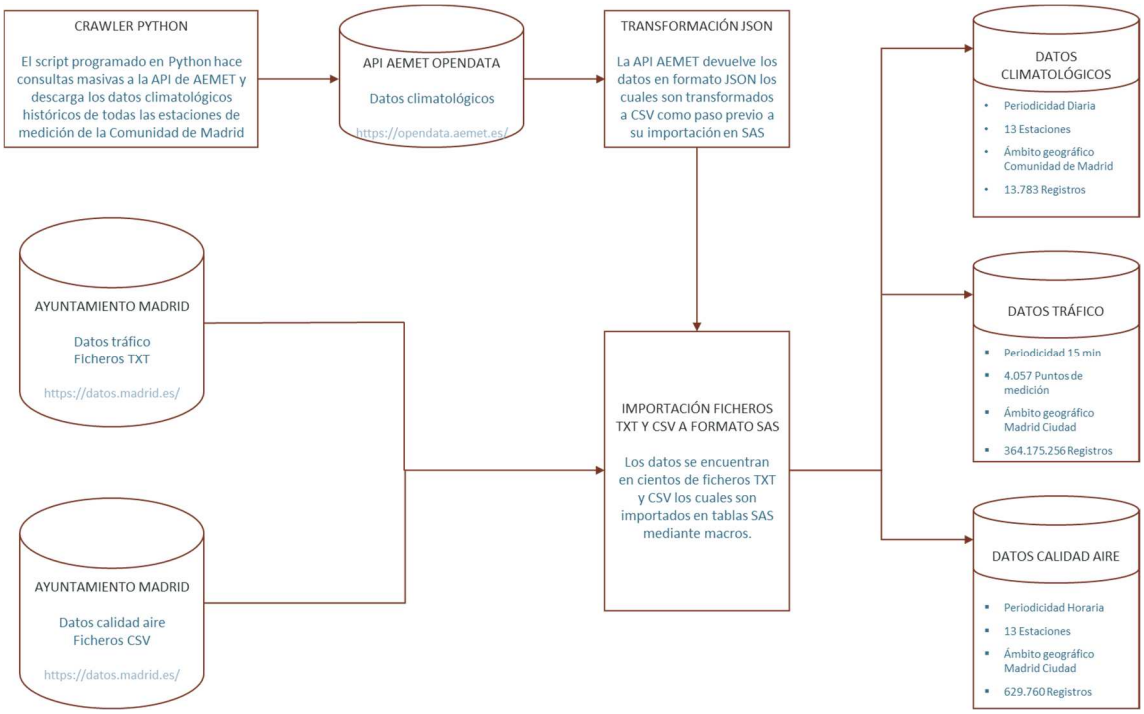


Figura 6. Diagrama de aprovisionamiento de datos

3.2. API Aemet

El primero de los orígenes de los que hemos recuperado datos es la Agencia Estatal de Meteorología (AEMET) de donde hemos obtenido información climatológica histórica de las estaciones de la Comunidad de Madrid. La idea es recuperar datos climatológicos que utilizaremos posteriormente como variables input en nuestra modelización predictiva para determinar si se superará o no el nivel de preaviso de NO2 en un determinado día.

AEMET dispone de una API REST que permite a desarrolladores la posibilidad de hacer consultas periódicas, e incluso programadas, de datos. La abreviatura API procede del inglés (Application Programming Interface) que, traducido al castellano, sería interfaz de programación de aplicaciones. Es decir, las consultas no se realizan a través de una interfaz amigable e intuitiva como puede ser una página web que consulta cualquier usuario, sino que requieren de un desarrollo informático para realizarlas. Una API "REST" es una API apoyada totalmente en el estándar HTTP. Visto de una forma más sencilla, una API REST es un servicio que nos provee de funciones que nos dan la capacidad de hacer uso de un servicio web que no es nuestro, dentro de una aplicación propia, de manera segura.

Para poder acceder a AEMET OpenData, es necesario solicitar una API Key. Una API Key es un identificador, mediante el cual se contabilizan e imputan los accesos que un

usuario realiza al API. De esta forma se limita a los usuarios el número de consultas que se puede hacer por minuto y así no saturar los servidores.

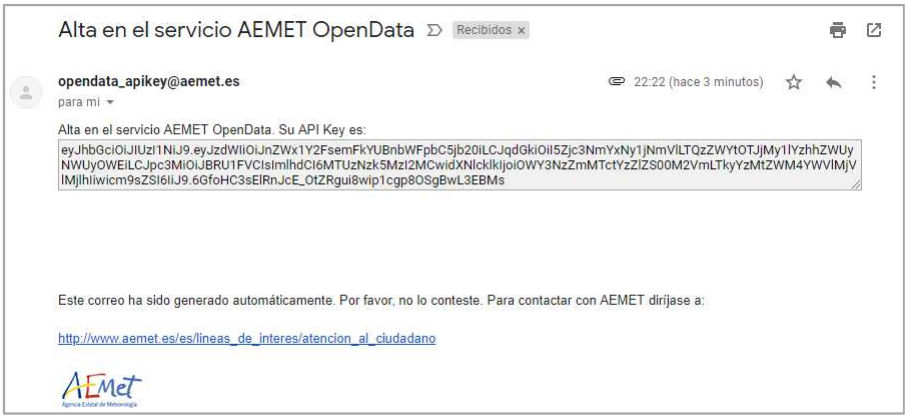
3.3. Recopilación automática de datos climatológicos históricos procedentes de API AEMET utilizando lenguaje Python

Para la realización del presente trabajo y explotar la API REST que ofrece AEMET, se ha desarrollado en Python una solución que hace consultas masivas a dicha API de cara a recuperar los datos climatológicos diarios de aquellas estaciones climatológicas deseadas. Los únicos parámetros que el usuario tendría que definir son el rango de fechas a consultar y el listado de estaciones de las que se quieren recuperar datos.

En nuestro caso buscamos descargar los datos climatológicos históricos de todas las estaciones de la Comunidad de Madrid de una manera automatizada. AEMET únicamente permite consultar datos con una profundidad de un mes y para una única estación. Es decir, no permite aumentar la profundidad y en una única consulta ver datos con una profundidad superior a un mes. Esto supone que, por ejemplo, para un rango de fechas de tres años como el que estamos trabajando, habiendo 13 estaciones climatológicas, tendríamos 468 combinaciones posibles de consultas. Es decir, nuestra aplicación permite descargar en una sola ejecución los datos climatológicos mensuales de todas las estaciones de la comunidad de Madrid en ese periodo de años. De hacerlo manualmente habría que hacer 468 consultas manuales a la web. Simplemente cambiando la fecha de inicio y la fecha de fin en nuestro script se puede ampliar la profundidad histórica de datos a descargar pero, para nuestro estudio, hemos considerado suficiente trabajar con un periodo de tres años. Los datos climatológicos conseguidos por esta vía se tratarán posteriormente en SAS y nos servirán como variables input en la elaboración de modelos de cara a predecir alertas por nivel alto de NO2.

Entrando en la explicación técnica, se indica a continuación qué módulos/paquetes se han utilizado, con qué fin y se aporta y comentan fragmentos clave del código. El código entero se adjunta como “Anexo 1” a este TFM.

- **Paquete 'Requests'**: Paquete que permite el envío de solicitudes HTTP usando Python. Al hacer la solicitud HTTP es necesario incluir en la llamada la API_KEY recibida al registrarnos al servicio:



Como hemos señalado anteriormente, esta API Key es la forma que tiene AEMET de establecer un límite de consultas en un determinado tiempo, por lo que es un parámetro que es obligatorio incluir al enviar la solicitud.

En Python definimos el objeto "api_key" con el código recibido por email y con el comando requests.get, cada vez que hagamos una solicitud HTTP a AEMET, incluiremos como parámetro dicha API Key (más detalle en los siguientes puntos):

```
api_key = {"api_key":"eyJhbGciO (…) cgp8OSgBwL3EBMs"}
respuesta = requests.get(urlnueva, params=api_key, headers=headers,
verify=False)
```

- **Paquete 'Itertools'**: Módulo que facilita la construcción de iteraciones. En nuestro caso se ha utilizado para definir un objeto que recopile todas las posibles combinaciones de Fecha primer día mes – Fecha último día mes – Id estación para un rango de años definido. Sobre estas combinaciones se irá iterando para hacer las sucesivas llamadas a la API AEMET.

```
# Loop para crear la lista de primeros días y últimos días del mes (es un loop
anidado, meses dentro de años)
inicio=2010
fin=2017

# Creamos dos listas iniciales vacías y luego se alimentan: Lista con día de inicio
de mes y lista con último día del mes. Posteriormente las combinaremos creando
tuplas
mesesinicio = []
mesesfinal = []

for i in range(inicio,fin+1):
    for j in range(1,13):
        primerdia = datetime.date(i, j, 1)
        ultimodia = calendar.monthrange(i, j)
        ultimodia2 = datetime.date(i, j, ultimodia[1])
        mesesinicio.append(primerdia)
        mesesfinal.append(ultimodia2)
        print(primerdia.strftime('%m-%d-%Y') + ' ' + ultimodia2.strftime('%m-%d-
%Y'))
        print("\n")

# Utilizamos la funcion ZIP para combinar ambas listas y crear una lista de tuplas
con día inicio - fin mes
listameses=list(zip(mesesinicio,mesesfinal))

# 3100B - Aranjuez
# 3110C - Buitrago Del Lozoya
# 3191E - Colmenar Viejo
# 3200 - Getafe
# 3129 - Madrid Aeropuerto
# 3194U - Madrid, Ciudad Universitaria
# 3196 - Madrid, Cuatro Vientos
# 3195 - Madrid, Retiro
# 3266A - Puerto Alto Del Leon
# 2462 - Puerto de Navacerrada
# 3338 - Robledo De Chavela
# 3111D - Somosierra
# 3175 - Torrejon De Ardoz

estaciones = ['3100B', '3110C', '3191E', '3200', '3129', '3194U', '3196', '3195',
'3266A', '2462', '3338', '3111D', '3175']

# Ahora vamos a crear las posibles combinaciones de Día inicio - Día fin mes -
Estación de medición. Para ello combinamos la lista de tuplas de fechas que tenemos
con la lista de estaciones
combinaciones = list(itertools.product(listameses,estaciones))
```

- **Paquete 'Calendar'**: Módulo con funciones útiles para el tratamiento de fechas. En nuestro caso se ha utilizado para obtener el primer y último día de cada mes en formato YYYY-MM-DD (ver pantallazo código anterior).

- **Paquete 'JSON'**: Módulo que permite el tratamiento de ficheros JSON (JavaScript Object Notation). Los ficheros JSON son un fichero de tipo texto ligero con cierta estructura. Al enviar las consultas a la API AEMET, la respuesta que da ésta es un enlace web del que descargar un fichero JSON. Dicho fichero es el que contiene los datos climatológicos que queremos recuperar. En resumen, el paquete JSON se ha utilizado para el tratamiento de los ficheros que devuelve AEMET en cada consulta.

- **Paquete 'RE'**: Paquete que permite a Python trabajar con expresiones regulares ("Regular Expressions"). Las expresiones regulares son un mini lenguaje en sí mismo, por lo que para poder utilizarlas eficientemente primero debemos entender los componentes de su sintaxis. Importando este módulo podemos encontrar todas las coincidencias de un patrón y sustituirlas por una cadena. En nuestro desarrollo en concreto se define una URL a la que Python va a llamar para recuperar datos:

```
https://opendata.aemet.es/opendata/api/valores/climatologicos/diarios/datos/fechaini/{fechainicio}T00:00:00UTC/fechafin/{fechafin}T23:59:59UTC/estacion/{idestacion}/
```

Dicha URL incluye 3 parámetros (los marcados en color) que vamos a ir sustituyendo en cada iteración gracias a este paquete. De esta forma, en cada llamada, se recuperarán los datos definidos para cada una de las posibles combinaciones de estos 3 parámetros.

```
# Loop de consultas a la web de AEMET

for i in combinaciones:
    url=
"https://opendata.aemet.es/opendata/api/valores/climatologicos/diarios/datos/fechaini/{fechainicio}T00:00:00UTC/fechafin/{fechafin}T23:59:59UTC/estacion/{idestacion}/"
    # Con la funcion lista.index() conseguimos el índice de la lista que sirve para obtener cada una de las fechas en cada iteración
    # [0] Corresponde a primer día de mes
    fechainicio = combinaciones[combinaciones.index(i)][0][0].strftime('%Y-%m-%d')
    # [1] Corresponde a último día de mes
    fechafin = combinaciones[combinaciones.index(i)][0][1].strftime('%Y-%m-%d')
    # Marcamos con [1] para quedarnos con la segunda parte de la tupla
    estacionreemplazo = combinaciones[combinaciones.index(i)][1]
    # Reemplazo fecha inicio
    urlaux = re.sub(r'{fechainicio}', fechainicio, url)
    # Reemplazo fecha fin sobre urlaux anterior
    urlaux2 = re.sub(r'{fechafin}', fechafin, urlaux)
    # Reemplazo idestacion sobre urlaux2 anterior
    urlnueva = re.sub(r'{idestacion}', estacionreemplazo, urlaux2)
    print("Descargando" + " " + urlnueva)
    respuesta = requests.get(urlnueva, params=api_key, headers=cabecera_llamada, verify=False)
    # AEMET devuelve un enlace temporal que contiene los datos en formato JSON
    json_data = json.loads(respuesta.text)
    # Parseamos el fichero JSON extrayendo el link a la web
    urldatos = json_data['datos']
    # Hacemos una nueva llamada a la web temporal que tiene los datos
    respuesta2 = requests.get(urldatos, params=api_key, headers=cabecera_llamada, verify=False)
    json_data2 = json.loads(respuesta2.text)
```

- **Paquete 'CSV'**: Utilizamos dicho paquete para transformar los ficheros JSON recibidos a CSV ya que es un formato más manejable para su posterior importación y utilización en SAS.

```
# Defino la lista de campos en función de los metadatos proporcionados por la web de
AEMET
# Este listado se utilizará como cabecera del CSV que se genere

campos = ['fecha',
'indicativo','nombre','provincia','altitud','tmed','prec','tmin','horatmin','tmax','
horatmax','dir','velmedia','racha','horaracha','sol','presMax','horaPresMax','presMi
n','horaPresMin']

    if len(json_data2)>2:
        with open(estacionreemplazo + '-' + fechainicio + '-' + fechafin + '.csv',
'wb') as fichero_csv:
            dict_writer = csv.DictWriter(fichero_csv, delimiter=";",
fieldnames=campos)
            dict_writer.writeheader()
            dict_writer.writerows(json_data2)
        print("Escribiendo fichero CSV" + " " + estacionreemplazo + '-' +
fechainicio + '-' + fechafin +'.csv')
        print(json_data2)

# En caso de que un día no haya datos grabados no genero fichero
    else:
        print("Sin datos. Pasando a siguiente iteración")

# Retardo de tres segundos entre cada ejecución para no superar límite de llamadas
por minuto
    time.sleep(3)
```

La salida resultante tras ejecutar nuestro script serán cientos de ficheros .CSV. Cada uno de estos ficheros contiene la información climatológica diaria de un periodo de un mes para una estación en concreto.

3.4. Preprocesamiento

Como hemos introducido anteriormente (Figura 6), partimos de una situación en la que la información con la que estamos trabajando proviene de diferentes orígenes, con formatos, ámbitos geográficos y periodicidades distintas. Por ello, es necesario llevar a cabo una tarea de preprocesamiento de dicha información como paso previo a la modelización. Esta tarea de preprocesamiento se compone de las siguientes etapas:

- Fase de importación mediante macros. Por el elevado número de ficheros con el que se ha trabajado, se descartó la importación manual y ha sido necesario buscar una solución más rápida y automatizada.
- Exploración y tratamiento de los datos para depurar valores erróneos, valores ausentes y, en definitiva, mejorar la calidad de la información para introducir el menor ruido posible en la modelización.
- Homogeneización del ámbito geográfico de los diferentes datos con los que se ha trabajado, la cual se ha realizado mediante la agrupación de estaciones de calidad

del aire, estaciones climatológicas y puntos de medición del tráfico en las mismas 5 zonas.

- Homogeneización de la periodicidad de los datos. Puesto que el objetivo del presente trabajo es predecir si un determinado día se superará o no el nivel de preaviso de NO2, nos interesa transformar toda la información a una periodicidad diaria. Cada fuente de información tiene periodicidades distintas lo que ha exigido establecer criterios para la transformación de los datos a diarios.
- Análisis de la información disponible y creación de nuevas variables que puedan ser útiles en la modelización.

En los siguientes puntos profundizaremos sobre lo aquí expuesto.

3.5. Marco teórico sobre las técnicas de clasificación utilizadas

De cara a predecir si un día se superará o no el nivel de preaviso de NO2 hemos aplicado las siguientes técnicas estadísticas:

3.5.1. Regresión Logística

La regresión logística estudia la asociación entre una variable dependiente categórica o binaria y un conjunto de variables explicativas las cuales pueden ser categóricas o continuas. Mediante esta técnica se modela la probabilidad de que la variable independiente tome uno u otro valor en función de las posibles combinaciones de valores de las variables independientes. El resultado final será una función en la que la estimación de los parámetros se realiza por máxima verosimilitud. Dado que las observaciones son independientes, no existe una fórmula explícita para la obtención de los parámetros que maximizan la verosimilitud, por lo que será necesario recurrir a métodos iterativos de optimización. Una de las ventajas de la regresión es que permite cuantificar los efectos de los predictores sobre la respuesta a través de los Odds Ratio. Los Odds Ratio cuantifican el cambio en la probabilidad de que la variable estimada pertenezca a una u otra categoría en base al cambio de categoría de cada variable incluida en el modelo.

3.5.2. Redes Neuronales

Una red neuronal consiste en una serie de algoritmos que tratan de reconocer las relaciones subyacentes en un conjunto de datos a través de un proceso que imita la forma en que funciona el cerebro humano. La analogía con el funcionamiento del cerebro consiste en que cada neurona procesa y combina estímulos de muchas otras neuronas a

través de estructuras de procesos de entrada llamada dendritas. A medida que las neuronas procesan información, se va formando el mecanismo de memoria del cerebro.

En este sentido, la idea subyacente de las redes neuronales es tratar de imitar el funcionamiento del cerebro de cara a elaborar reglas a partir de unos datos de entrada. Por ello, para entrenar una red neuronal es conveniente disponer de grandes cantidades de datos para que el proceso de entrenamiento de la red sea lo más robusto posible. Entrenar la red consiste en proporcionar datos de entrada y decirle a la red qué salida se espera. El objetivo de dicho entrenamiento es minimizar el error entre la predicción y la respuesta esperada.

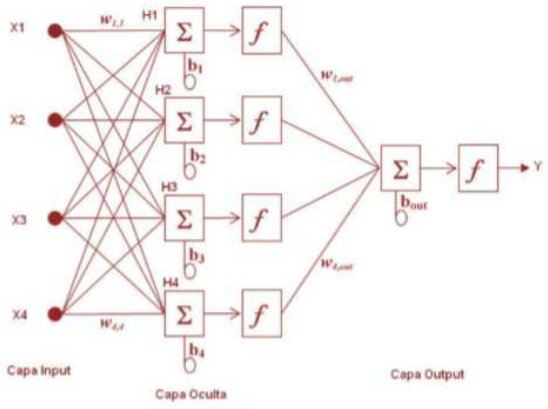


Figura 7. Esquema red neuronal

Una red neuronal está formada por muchos nodos conectados. Los nodos se organizan en grupos llamados "capas". Una red típica consiste de una secuencia de capas con total o aleatorias conexiones entre capas sucesivas. La capa input se conecta a la capa oculta mediante la función de combinación, representada por Σ, donde los pesos W_{ij} hacen el papel de parámetros a estimar. Tras aplicar la función de combinación, se aplica a cada nodo oculto la función de activación, representada por *f* y esta suele ser una función de tipo no lineal. Dicha función es la que utiliza la suma de estímulos para determinar la actividad de salida de la neurona. Es decir, al definir las reglas cada nodo de la red decide qué enviar al siguiente nivel en función de los aportes del nivel anterior.

3.5.3. Árboles de decisión

Un árbol de decisión es un modelo de predicción similar a un diagrama de flujo, en el que se llega a puntos en donde se toman decisiones en base a la capacidad discriminante de una variable. A cada evento se le asignan probabilidades y a cada una de las ramas se le determina un resultado. De cara a mejorar la capacidad predictiva de éstos, se han

desarrollado distintas técnicas que combinan el resultado de varios árboles. En este trabajo utilizaremos únicamente aquellas técnicas que combinan varios árboles:

3.5.4. Bootstrap Averaging

Breiman, en 1996, planteó la posibilidad de combinar la salida de varios árboles con el objetivo de generar uno más robusto. El proceso consistiría en replicar los siguientes pasos varias veces:

- Crear n muestras de los datos originales.
- Se crean m modelos predictivos para cada muestra.
- Se construye un único modelo predictivo a partir del promedio de los anteriores.

A la hora de predecir una variable categórica, como es nuestro caso, la salida del clasificador combinado será aquella clase que resulte ser elegida por la mayoría de los m clasificadores.

3.5.5. Random Forest

La técnica de Random Forest surge como intento de mejorar a Bagging. Este algoritmo mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en las variables utilizadas para segmentar cada nodo del árbol. El proceso sigue los mismos pasos que Bagging, la única diferencia radica en que, en cada modelo predictivo que se ajusta, se seleccionarán en cada nodo p variables de las k originales, y de las p elegidas, se escogerá la mejor variable para la partición del nodo, obteniendo así en cada modelo, diferentes registros y diferentes variables.

3.5.6. Gradient Boosting

Gradient Boosting (Friedman H., 2001) se basa en la idea de entrenar el algoritmo mediante la actualización de los pesos de las observaciones pertenecientes a las clases del suceso de interés a través de la optimización en dirección descendente de una función de pérdida o error determinada, consiguiendo dar mayor relevancia en cada iteración a las observaciones mal clasificadas en pasos anteriores.

3.5.7. Validación Cruzada

Para valorar la bondad de ajuste de todas las técnicas anteriormente comentadas se ha aplicado el método de validación cruzada. Dicha técnica consiste en particionar los datos originales en dos: una de las partes sobre la que se construirá el modelo y la otra sobre la

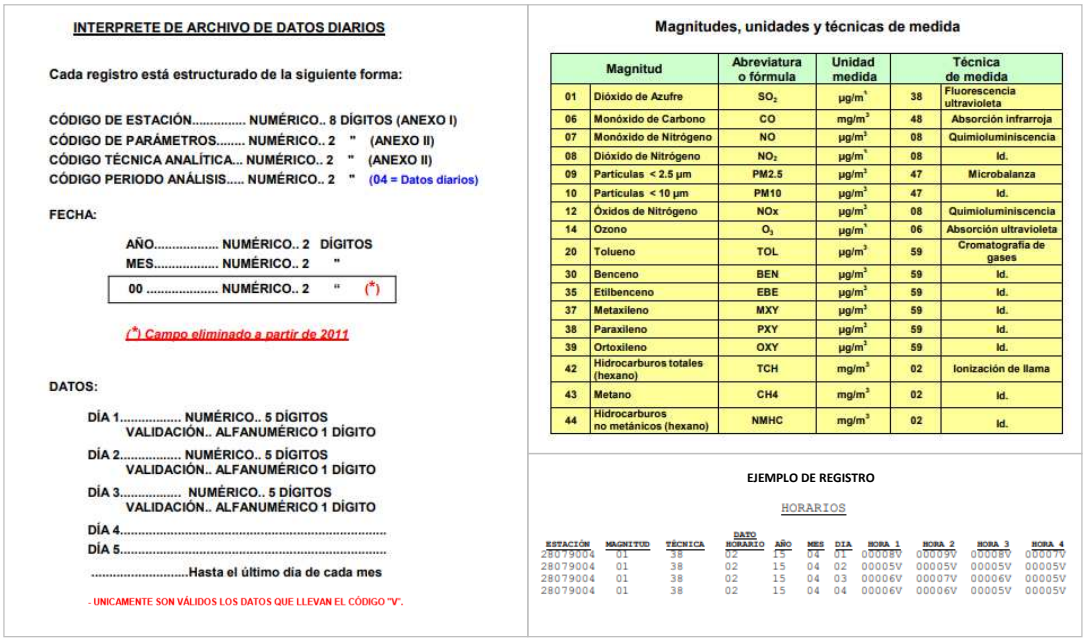
que se validarán los resultados obtenidos. Este proceso se hará n veces de cara a minimizar el factor aleatorio en la generación de las particiones.

4. Fase de preprocesamiento

4.1. Importación de datos en SAS mediante macros

Debido al elevado número de ficheros con el que estamos trabajando, se hace necesario la programación de macros que automaticen el proceso de importación en SAS de dichos ficheros. Se ha programado por lo tanto una serie de macros adaptadas a la estructura de cada origen de información.

A modo de ejemplo, en el proceso de importación de los datos climatológicos, la macro replica la estructura de campos definida por el Ayuntamiento de Madrid. Se tratan de ficheros .TXT con la siguiente estructura de campos:



INTERPRETE DE ARCHIVO DE DATOS DIARIOS

Cada registro está estructurado de la siguiente forma:

CÓDIGO DE ESTACIÓN..... NUMÉRICO.. 8 DÍGITOS (ANEXO I)
CÓDIGO DE PARÁMETROS..... NUMÉRICO.. 2 " (ANEXO II)
CÓDIGO TÉCNICA ANALÍTICA... NUMÉRICO.. 2 " (ANEXO II)
CÓDIGO PERIODO ANÁLISIS..... NUMÉRICO.. 2 " (04 = Datos diarios)

FECHA:

AÑO..... NUMÉRICO.. 2 DÍGITOS
MES..... NUMÉRICO.. 2 "
00 ..... NUMÉRICO.. 2 " (*)

(*) Campo eliminado a partir de 2011

DATOS:

DÍA 1..... NUMÉRICO.. 5 DÍGITOS
VALIDACIÓN.. ALFANUMÉRICO 1 DÍGITO
DÍA 2..... NUMÉRICO.. 5 DÍGITOS
VALIDACIÓN.. ALFANUMÉRICO 1 DÍGITO
DÍA 3..... NUMÉRICO.. 5 DÍGITOS
VALIDACIÓN.. ALFANUMÉRICO 1 DÍGITO
DÍA 4.....
DÍA 5.....
.....Hasta el último dia de cada mes

- UNICAMENTE SON VÁLIDOS LOS DATOS QUE LLEVAN EL CÓDIGO "V".

Magnitudes, unidades y técnicas de medida

Magnitud		Abreviatura o fórmula	Unidad medida	Técnica de medida	
01	Dióxido de Azufre	SO_2	µg/m³	38	Fluorescencia ultravioleta
06	Monóxido de Carbono	CO	mg/m³	48	Absorción infrarroja
07	Monóxido de Nitrógeno	NO	µg/m³	08	Quimioluminiscencia
08	Dióxido de Nitrógeno	NO_2	µg/m³	08	Id.
09	Partículas < 2.5 µm	PM2.5	µg/m³	47	Microbalanza
10	Partículas < 10 µm	PM10	µg/m³	47	Id.
12	Óxidos de Nitrógeno	NOx	µg/m³	08	Quimioluminiscencia
14	Ozono	O_3	µg/m³	06	Absorción ultravioleta
20	Tolueno	TOL	µg/m³	59	Cromatografía de gases
30	Benceno	BEN	µg/m³	59	Id.
35	Etilbenceno	EBE	µg/m³	59	Id.
37	Metaxileno	MXY	µg/m³	59	Id.
38	Paraxileno	PXY	µg/m³	59	Id.
39	Ortoxileno	OXY	µg/m³	59	Id.
42	Hidrocarburos totales (hexano)	TCH	mg/m³	02	Ionización de llama
43	Metano	CH4	mg/m³	02	Id.
44	Hidrocarburos no metánicos (hexano)	NMHC	mg/m³	02	Id.

EJEMPLO DE REGISTRO

HORARIOS

ESTACIÓN	MAGNITUD	TÉCNICA	DATO HORARIO	AÑO	MES	DIA	HORA 1	HORA 2	HORA 3	HORA 4
28079004	01	38	02	15	04	01	00008V	00009V	00008V	00007V
28079004	01	38	02	15	04	02	00005V	00005V	00005V	00005V
28079004	01	38	02	15	04	03	00006V	00007V	00006V	00005V
28079004	01	38	02	15	04	04	00006V	00006V	00005V	00005V

Figura 8. Estructura de campos ficheros calidad aire. (Fuente: Ayto. Madrid)

La macro importa uno a uno cada uno de los ficheros descargados, aplicando la anterior estructura de campos.

```
/*
###############################################################################################

MACRO DE IMPORTACIÓN DE FICHEROS CALIDAD AIRE
FORMATO ANTIGUO - A PARTIR DE OCTUBRE 2017 EL AYTO MODIFICÓ LA ESTRUCTURA DE CAMPOS

###############################################################################################
*/

/* Ruta donde se encuentran los ficheros de calidad del aire (Formato Antiguo) */

%LET RUTA_FICHEROS_AIRE_OLD=C:\Users\Gelu\Desktop\TFM\DATOS\AIRE\FORMATO ANTIGUO\;

FILENAME AIRE_OLD PIPE "dir ""&RUTA_FICHEROS_AIRE_OLD*.txt"" /b";



/* Creación de una tabla auxiliar que contiene el listado de todos los ficheros del
directorio */

DATA LISTA_FICHEROS_AIRE_OLD;
LENGTH NOMBRE_FICHERO $35;
INFILE AIRE_OLD TRUNCOVER;
INPUT NOMBRE_FICHERO $35.;
CALL SYMPUT ('NUMERO_FICHEROS',_N_);
/* Guardo en la variable NUMERO_FICHEROS el número de observaciones para utilizar
después como rango superior del bucle */
RUN;

/* Defino la macro IMPORTA_DATOS_AIRE_OLD con la que voy a importar todos los ficheros
de calidad del aire descargados de la web */

%MACRO IMPORTA_DATOS_AIRE_OLD;
%DO J=1 %TO %EVAL(&NUMERO_FICHEROS.);
/* Variable que he definido en el paso anterior. Habrá tantas iteraciones como ficheros
haya en la carpeta */
DATA _NULL_;
SET LISTA_FICHEROS_AIRE_OLD;
IF _N_=&J;
CALL SYMPUT ('FICHEROENTRADA',NOMBRE_FICHERO);
%PUT &FICHEROENTRADA;
RUN;

DATA WORK.FICHERO_AIRE_OLD&J;

INFILE "&RUTA_FICHEROS_AIRE_OLD\&FICHEROENTRADA" LRECL=164;
INPUT
@1 COD_ESTACION $8.
/* (…) Se omite el listado completo de campos por su extensión (…) */
@164 VALHORA23 $1.
;
RUN;
%END;
%MEND IMPORTA_DATOS_AIRE_OLD;

%IMPORTA_DATOS_AIRE_OLD;

/* Junto todos los ficheros importados en uno único */

PROC SQL NOPRINT;
SELECT COUNT(*) INTO :CONTEO
FROM LISTA_FICHEROS_AIRE_OLD;
QUIT;

DATA _NULL_;
CALL SYMPUT("FICHEROFINAL",CAT('FICHERO_AIRE_OLD',&CONTEO));
RUN;

DATA DATOS_TOTAL_AIRE_OLD;
SET FICHERO_AIRE_OLD1-&FICHEROFINAL;
RUN;
```

La misma solución (adaptando la estructura de campos a cada caso) se ha utilizado para la importación de los ficheros de tráfico y de climatologías diarias.

Datos	Nº de estaciones o puntos de medición	Frecuencia	Número de registros	Ámbito geográfico	Formato coordenadas	Fuente origen
Datos calidad aire (NO2)	24	Horaria	629.760	Madrid ciudad	WGS84	Ayto. Madrid
Datos climatológicos	13	Diaria	13.783	Comunidad de Madrid	WGS84	AEMET
Datos tráfico	4.057	Cada 15 min	364.175.256	Madrid ciudad	UTM Zone 30N (ETRS89)	Ayto. Madrid

Figura 9. Datos iniciales

Como puede verse en el resumen de la Figura 8, cada fuente de información abarca un ámbito geográfico diferente, las mediciones tienen distinta periodicidad y las estaciones/puntos de medición no son las mismas. Puesto que nuestro objetivo es predecir si un determinado día se superara el nivel de preaviso de NO2 a partir de los datos climatológicos y datos de tráfico, ha sido necesario llevar a cabo las siguientes actuaciones (se profundiza en ellas en los siguientes apartados).

- Categorización a binaria de nuestra variable objetivo (Nivel de NO2), siendo 1 el evento de que se supere los 180 microgramos/m3 y 0 en caso contrario.
- Agrupación mediante clusters de las estaciones climatológicas y los puntos de medición de intensidad de tráfico, de cara a poder imputar como variables explicativas los datos registrados en ellos.
- Tratamiento de las distintas bases de datos para transformar la periodicidad a diaria, tomando criterios que vayan en línea con el objetivo perseguido.
- Creación de variables nuevas que puedan ser útiles para la predicción.

4.2. Exploración y tratamiento de los datos de calidad del aire

El intérprete de archivo de datos diarios que ofrece el Ayuntamiento, indica que *"únicamente son válidos los datos que llevan el código V"*. Observando los datos originales vemos que existen algunas mediciones erróneas, marcadas con "N".

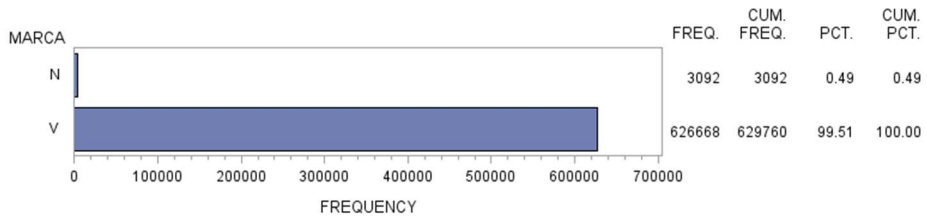


Figura 10. Mediciones NO2 no válidas

Revisamos si los valores no válidos se concentran en alguna de las estaciones, pero no parece ser el caso:

Estación	Mediciones válidas	Mediciones NO válidas
Arturo Soria	26.234	70
Av. Ramón y Cajal	26.217	87
Barajas	26.222	82
Barrio del Pilar	26.079	225
C/ Farolillo	25.620	180
Casa de Campo	26.010	198
Cuatro Caminos	26.239	65
El Pardo	26.143	113
Ensanche Vallecas	26.188	116
Escuelas Aguirre	26.121	183
Moratalaz	25.631	145
Méndez Álvaro	26.195	109
Parque Juan Carlos I	26.133	147
Pza. Castilla	26.087	217
Pza. Fdez. Ladreda	26.160	144
Pza. de España	26.226	78
Pza. del Carmen	25.803	189
Pº. Castellana	26.204	76
Retiro	26.090	214
Sanchinarro	26.234	70
Tres Olivos	26.243	61
Urb. Embajada (Barajas)	26.173	131
Vallecas	26.221	83
Villaverde Alto	26.195	109
TOTAL	**626.668**	**3.092**

Figura 11. Distribución de mediciones no válidas por estación

Por ello, puesto que las mediciones no válidas representan tan sólo el 0,49% y no parece haber una estación que concentre la mayoría de ellas, optamos por imputar el último valor válido registrado de la estación en cuestión ya que no debería haber excesiva variación en el nivel de NO2 de una hora a la siguiente.

Estación	Media NO_2	Máximo NO_2
Arturo Soria	40	356
Av. Ramón y Cajal	45	424
Barajas	38	228
Barrio del Pilar	41	328
C/ Farolillo	40	223
Casa de Campo	23	137
Cuatro Caminos	44	291
El Pardo	18	108
Ensanche Vallecas	39	272
Escuelas Aguirre	59	369
Moratalaz	40	198
Méndez Álvaro	40	235
Parque Juan Carlos I	24	190
Pza. Castilla	43	291
Pza. Fdez. Ladreda	57	347
Pza. de España	47	283
Pza. del Carmen	48	196
Pº. Castellana	39	227
Retiro	32	227
Sanchinarro	34	294
Tres Olivos	35	188
Urb. Embajada (Barajas)	44	239
Vallecas	41	242
Villaverde Alto	45	302

Figura 12. Valores máximos y media NO2 por estación

Tras la depuración de estos valores erróneos revisamos los valores máximos y mínimos, no detectando datos anómalos. El valor máximo registrado en la estación de Ramón y Cajal (424 microgramos/m3) parece algo alto, pero se ha comprobado que la medición es correcta ya que, haciendo una simple búsqueda en Internet, encontramos noticias hablando de ello:

Extraído de una noticia de www.telemadrid.es:

"Ayer miércoles, estuvieron por encima del límite las estaciones de Plaza de España (219), Escuelas Aguirre (369), ***avenida Ramón y Cajal (424),*** *Arturo Soria (280), (…), en todos los casos entre las 19.00 y las 23.00 horas."*

4.3. Clusters de estaciones climatológicas y puntos de medición del tráfico

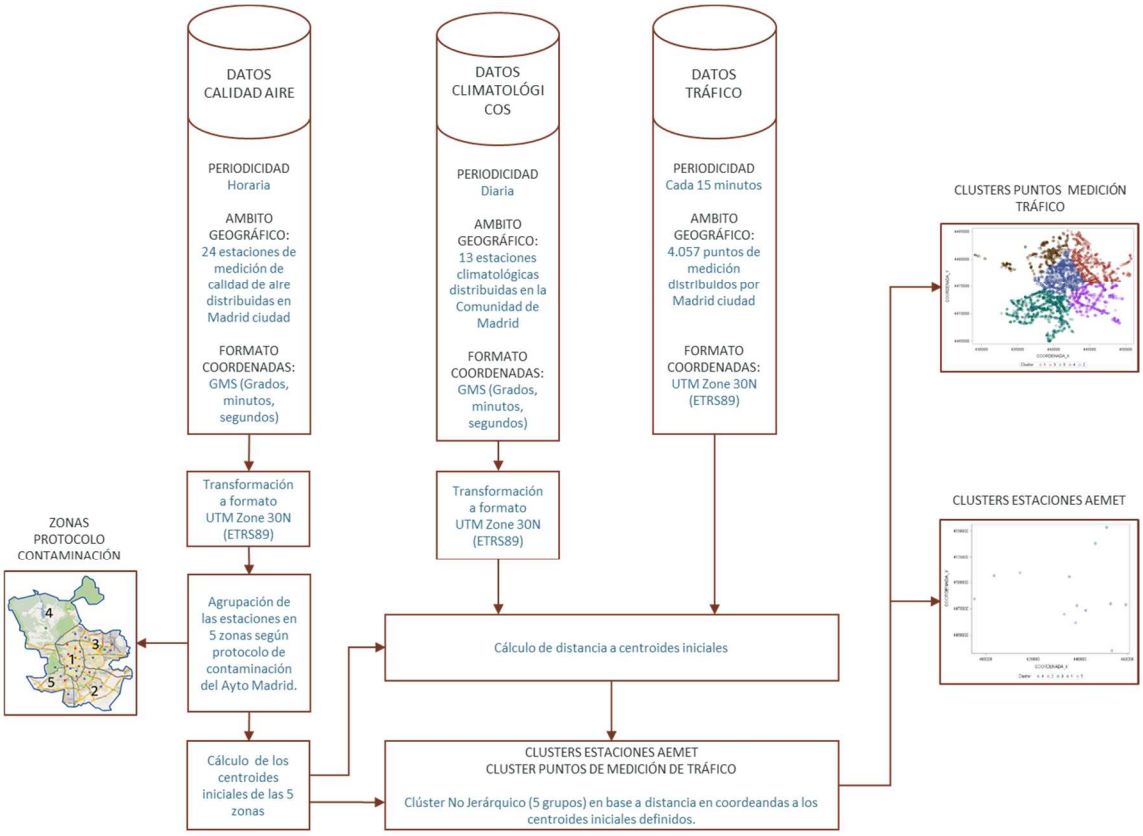


Figura 13. Proceso de clusterización de estaciones y puntos de medición tráfico

Como hemos comentado anteriormente, partimos de una situación en la que las estaciones/puntos de medición no son los mismos para cada fuente de información con las que estamos trabajando. Uno de los pasos necesarios previos a la modelización consistirá en definir un criterio de cara a homogeneizar los diferentes datos, tanto en el ámbito geográfico (los distintos puntos de medición están situados en lugares diferentes) como de periodicidad del dato (partimos de unos datos originales con mediciones de calidad del aire horarias, mientras que los datos climatológicos de los que disponemos son diarios y los datos de intensidad de tráfico son cada quince minutos).

Comenzando por el primer problema, distinto ámbito geográfico, se han agrupado las 24 estaciones de calidad del aire en 5 zonas en línea con lo definido en el Protocolo de contaminación del Ayuntamiento de Madrid (Figura 4).

Puesto que nuestro objetivo es predecir el nivel máximo de NO2 del día y, para ello, vamos a utilizar los datos de intensidad de tráfico y climatológicos, se hace necesario

definir un criterio en base al cual imputar dichos datos a cada una de esas 5 zonas. Para hacerlo, como primer paso hemos recurrido a su agrupación mediante clusters no jerárquicos en base a la distancia en coordenadas entre los distintos puntos. La idea es determinar para cada punto de medición cuál sería la zona de pertenencia.

El proceso ha sido el siguiente:

1. Homogeneización de formatos de coordenadas: Las coordenadas de las estaciones de calidad del aire y de las estaciones climatológicas AEMET figuran en formato WGS84 (grados, minutos, segundos) mientras que las coordenadas de los puntos de medición de intensidad del tráfico se encuentran expresadas en formato UTM Zone 30N (ETRS89). Como paso previo, se han transformado todas las coordenadas WGS84 a UTM Zone 30N (ETRS89) utilizando conversores de coordenadas online.

Figura 14. Localización de la estación de calidad del aire de Plaza de Castilla. (Fuente: Ayto. Madrid)

Tomando como ejemplo la conversión de coordenadas para la estación de medición de Plaza de Castilla, hemos realizado los siguientes pasos:

Partiendo de las coordenadas que indica el ayuntamiento se obtiene la latitud y longitud en grados decimales a partir de los grados, minutos y segundos. La herramienta que hemos utilizado es:

https://www.coordenadas-gps.com/convertidor-de-coordenadas-gps

Figura 15. Conversión de coordenadas: de GMS a GD

Y una vez tenemos la latitud y longitud en grados decimales podemos convertir a coordenadas en formato UTM Zone 30N (ETRS89). Para este paso, la herramienta utilizada es Plexscape:

http://ws.plexscape.com/Services/CoordSysWS/Pages/Transformations.aspx

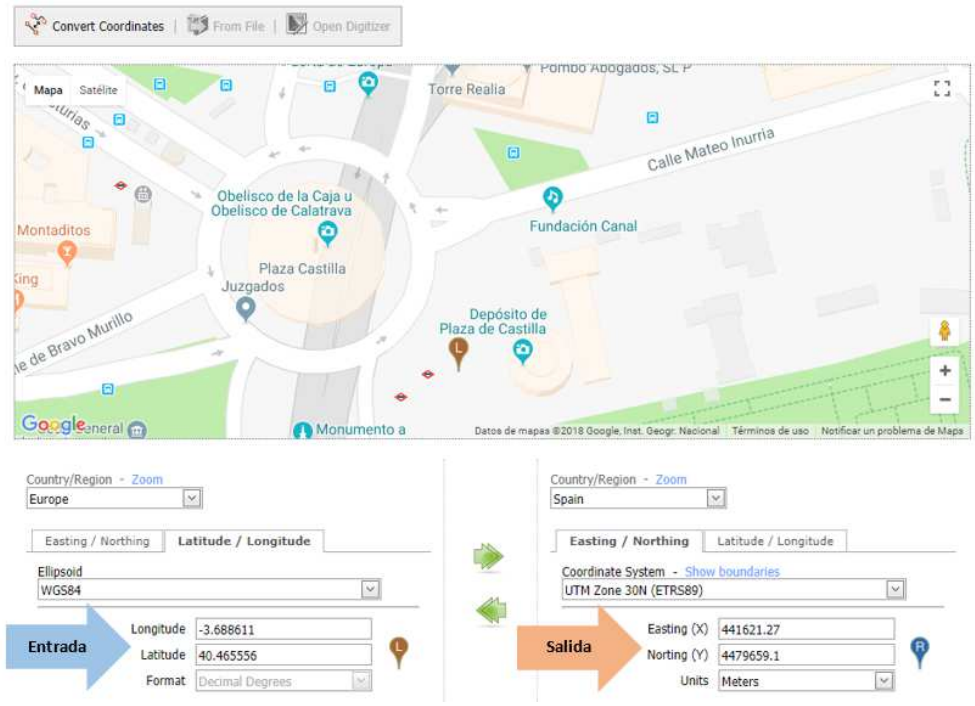


Figura 16. Conversión de coordenadas: de WGS84 a UTM Zone 30N (ETRS89). ws.plexscape.com

Esta conversión de coordenadas se ha hecho para todas las estaciones de calidad del aire y para todas las estaciones AEMET, siendo el resultado el que se indica en la tabla siguiente:

Estaciones Calidad del Aire Madrid			
Estación calidad aire	**Coordenada X**	**Coordenada Y**	**Zona**
Plaza de España	439572.93	4475061.46	**1 - Interior**
Escuelas Aguirre	442116.94	4474771.88	
Ramon y Cajal	442563.72	4478088.43	
Plaza del Carmen	440345.85	4474524.28	
Cuatro Caminos	440033.67	4477449.96	
Barrio del Pilar	439688.6	4481081.14	
Méndez Álvaro	441727.88	4472165.33	
Paseo de la Castellana	441457.56	4476792.69	
Retiro	442095.52	4473981.74	
Plaza de Castilla	441621.27	4479659.1	
Arturo Soria	445786.93	4476796.19	**2 - Noreste**
Barajas Pueblo	450835.04	4480855.27	
Urbanización Embajada (Barajas)	450779.74	4479254.08	
Sanchinarro	444028.27	4482819.82	
Juan Carlos I	448379.54	4479547.76	
Tres Olivos	441557.43	4483544.86	
Casa de Campo	436598.57	4474571.39	**3 - Noroeste**
El Pardo	434381.5	4485547.09	
Moratalaz	445246.87	4473237.77	**4 - Sureste**
Vallecas	444702.23	4471043.25	
Ensanche de Vallecas	448049.6	4469312.95	
Villaverde	439419.72	4466527.24	**5 - Suroeste**
Farolillo	437891.88	4471832.25	
Plaza Fernández Ladreda	439004.95	4470706.88	

Estaciones AEMET		
Estación climatológica	**Coordenada X**	**Coordenada Y**
Navacerrada	414407.23	4509371.75
Buitrago de Lozoya	446473.17	4538212.00
Aranjuez	453429.74	4435361.13
Somosierra	451249.07	4553659.38
Aeropuerto	452902.23	4479702.95
Torrejón	459397.06	4478801.53
Colmenar Viejo	435367.45	4505304.76
Ciudad Universitaria	438594.29	4478141.55
Retiro	442470.47	4473701.34
Cuatro Vientos	433266.51	4469738.14
Getafe	438035.07	4461741.86
Puerto Alto del León	403534.84	4506791.27
Robledo de Chavela	395001.93	4484177.22

Figura 17. Resultado de conversión de coordenadas por estación

En los “Anexos 3 y 4”, adjuntos a este TFM, se incluye el detalle con la localización de cada una de ellas.

2. Una vez tenemos las coordenadas de todas las estaciones y puntos de medición adaptadas a un mismo formato procedemos al cálculo de los centroides iniciales de las 5 zonas definidas. Posteriormente agruparemos en clusters no jerárquicos en función de la distancia de cada punto de medición a dichos centroides. El cálculo de los centroides iniciales se ha hecho definiendo una función sencilla en Python. Se muestra la función y un ejemplo de aplicación a continuación:

```
>>> def centroide(*puntos):
...     x_coords = [p[0] for p in puntos]
...     y_coords = [p[1] for p in puntos]
...     lista = len(puntos)
...     centroide_x = sum(x_coords)/lista
...     centroide_y = sum(y_coords)/lista
...     return [centroide_x, centroide_y]
...
>>> # CENTROIDE ZONA NOROESTE #
... centroide_noroeste = centroide((436598.57,4474571.39),(434381.5,4485547.09))
... print("Centroide Zona Noroeste " + str(centroide_noroeste))
...
Centroide Zona Noroeste [435490.03500000003, 4480059.24]
```

3. Una vez calculados los centroides iniciales se obtienen las distancias en coordenadas de cada punto de medición de tráfico / estación climatológica a dichos centroides.
4. A partir de dichas distancias se realiza la agrupación en los 5 clusters (método no jerárquico). De esta manera conseguimos asignar cada punto de medición de tráfico y cada estación climatológica a una de las 5 zonas en base a la distancia calculada entre el punto en cuestión y los centroides iniciales.

El resultado de la asignación es el siguiente:

Estaciones AEMET			
Estación climatológica	**Coordenada X**	**Coordenada Y**	**Cluster / Zona asignada**
Aeropuerto	452902.23	4479702.95	**1 - Interior**
Torrejón	459397.06	4478801.53	
Ciudad Universitaria	438594.29	4478141.55	
Retiro	442470.47	4473701.34	
Aranjuez	453429.74	4435361.13	**2 - Sureste**
Buitrago de Lozoya	446473.17	4538212.00	**3 - Noreste**
Somosierra	451249.07	4553659.38	
Colmenar Viejo	435367.45	4505304.76	
Navacerrada	414407.23	4509371.75	**4 - Noroeste**
Puerto Alto del León	403534.84	4506791.27	
Robledo de Chavela	395001.93	4484177.22	
Cuatro Vientos	433266.51	4469738.14	**5 - Suroeste**
Getafe	438035.07	4461741.86	

Figura 18. Clusters resultantes de estaciones climatológicas

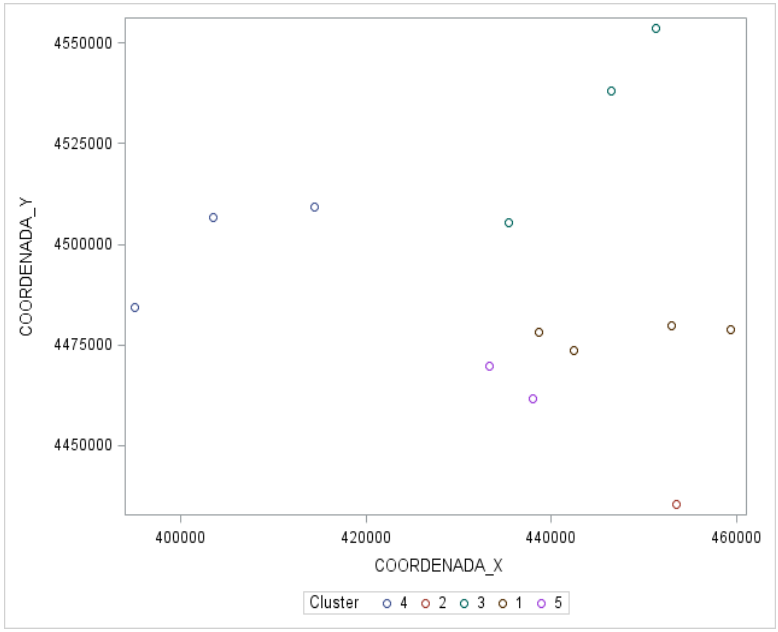


Figura 19. Representación gráfica de los clusters resultantes de estaciones climatológicas

El mismo método se ha utilizado para clusterizar los puntos de medición de tráfico. En este caso, al tratarse de 4.057 puntos omitimos el detalle con la asignación de cada uno de ellos, pero gráficamente se puede ver cómo la agrupación en clusters va en línea con las cinco zonas definidas:

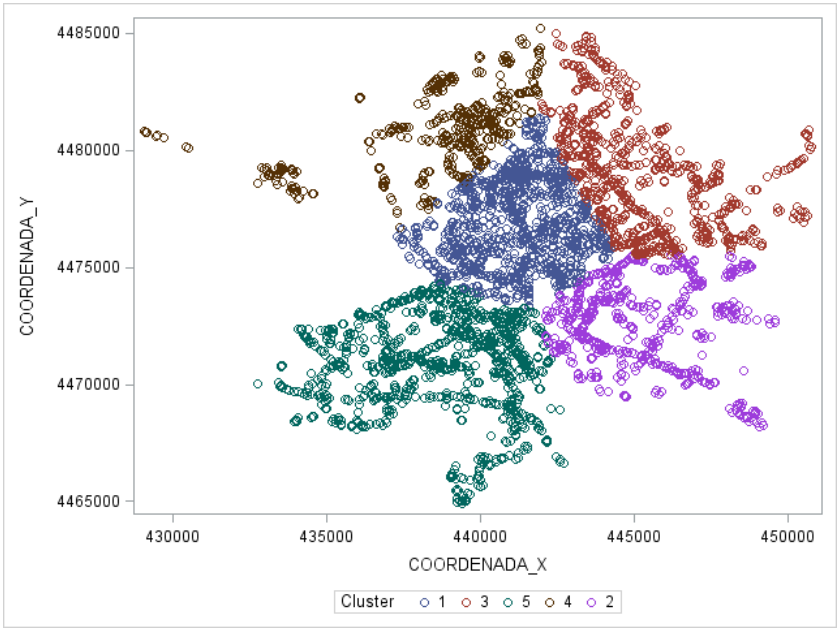


Figura 20. Representación gráfica de los clusters de puntos de medición de tráfico

4.4. Exploración y tratamiento de los datos de tráfico

Disponemos de información de los 4.057 puntos de medición de tráfico repartidos por la ciudad de Madrid. La periodicidad de medición es de cada 15 minutos, resultando en un total de 364.175.256 registros. Cada uno de ellos contiene la siguiente información:

Variable	Descripción	Tipo
INTENSIDAD	Intensidad de número de vehículos por hora.	Continua
VMED	Velocidad media asociada al punto de medición.	Continua
OCUPACIÓN	Porcentaje de ocupación del punto de control por los vehículos.	Continua
CARGA	Parámetro de carga del vial en función de la intensidad, ocupación y características de la infraestructura.	Continua

Figura 21. Variables incluidas en ficheros de tráfico

Explorando la variable intensidad de tráfico observamos que existen valores que podrían ser erróneos (Por ejemplo: 99999, incrementos exageradamente altos respecto a la medición anterior, etc.).

Obs	NAME	NMISS	N	MIN	MAX	MEAN	STD	SKEWNESS	KURTOSIS
1	INTENSIDAD	0	364175256	0	99999	441.301	696.382	8.80845	568.330

Figura 22. Principales estadísticos variable intensidad tráfico

El propio ayuntamiento al generar los ficheros incluye un campo con la marca ERROR que va informada a ‘S’ cuando el dato sea incorrecto. Sin embargo, no parece ser muy fiable dicho campo, ya que, explorando los valores máximos de la variable intensidad, nos hemos encontrado con casos claramente erróneos pese a que no vienen marcados con ERROR=’S’:

IDELEM	FECHA	IDENTIF	TIPO_ELEM	INTENSIDAD	OCUPACION	CARGA	VMED	ERROR
3938	2015-11-05 08:45:00	05010	PUNTOS MEDIDA URBANOS	2968	28	92	0	N
3938	2015-11-05 09:00:00	05010	PUNTOS MEDIDA URBANOS	2257	16	79	0	N
3938	2015-11-05 09:15:00	05010	PUNTOS MEDIDA URBANOS	1393	8	43	0	N
3938	2015-11-05 09:30:00	05010	PUNTOS MEDIDA URBANOS	1849	12	55	0	N
3938	2015-11-05 09:45:00	05010	PUNTOS MEDIDA URBANOS	1813	12	53	0	N
3938	2015-11-05 10:00:00	05010	PUNTOS MEDIDA URBANOS	1864	13	42	0	N
3938	2015-11-05 10:15:00	05010	PUNTOS MEDIDA URBANOS	2917	42	80	0	N
3938	2015-11-05 10:30:00	05010	PUNTOS MEDIDA URBANOS	51330	23	93	0	N
3938	2015-11-05 10:45:00	05010	PUNTOS MEDIDA URBANOS	3055	28	100	0	N
3938	2015-11-05 11:00:00	05010	PUNTOS MEDIDA URBANOS	2723	32	100	0	N
3938	2015-11-05 11:15:00	05010	PUNTOS MEDIDA URBANOS	2620	23	85	0	N
3938	2015-11-05 11:30:00	05010	PUNTOS MEDIDA URBANOS	1875	9	60	0	N
3938	2015-11-05 11:45:00	05010	PUNTOS MEDIDA URBANOS	2303	17	73	0	N
3938	2015-11-05 12:00:00	05010	PUNTOS MEDIDA URBANOS	3168	28	84	0	N
3938	2015-11-05 12:15:00	05010	PUNTOS MEDIDA URBANOS	1857	11	79	0	N

Figura 23. Detección de valores erróneos en variable intensidad

En el ejemplo anterior, vemos que el dato de la medición de las 10:30:00 es claramente erróneo. No tiene sentido que la intensidad de tráfico sea tan alta si comparamos con la medición de los quince minutos anteriores y los quince minutos posteriores.

De cara a minimizar el impacto de estos valores erróneos se ha optado por la siguiente solución:

Para cada punto de medición (de los 4.057 que hay) mantendremos únicamente aquellas observaciones cuyo valor de intensidad de tráfico esté dentro del percentil 95. Es decir, desechamos el 5% de los valores superiores de intensidad para cada punto de medición. El hacer este ajuste para cada punto de medición complica ligeramente el proceso, más aun trabajando con un fichero de más de 360 millones de observaciones, pero no queda otra alternativa ya que, si no tuviéramos en cuenta la agrupación por punto de medición, estaríamos desechando el 5% de los valores superiores del total de nuestros datos afectando principalmente a aquellos puntos de medición con mayor intensidad de tráfico (por ejemplo, la intensidad de un punto de medición situado en la M30 siempre será mayor que la intensidad de una calle poco transitada).

Para llevar a cabo este ajuste, utilizamos el PROC RANK de SAS de cara a calcular el percentil 95 para cada identificador de punto de medición y, en la salida, desechamos aquellas observaciones cuyo valor de intensidad supera dicho percentil.

```
/* DESECHAMOS VALORES DE INTENSIDAD SUPERIORES AL PERCENTIL 95 */
PROC RANK DATA=TRABAJO.DATOS_TOTAL_TRAFICO4
OUT=TRABAJO.DATOS_TOTAL_TRAFICO5 (WHERE=(R_INTENSIDAD<19))
GROUPS=20 TIES=LOW;
VAR INTENSIDAD;
RANKS R_INTENSIDAD;
BY IDELEM;
RUN;
```

Variables de intervalo

Obs	NAME	NMISS	N	MIN	MAX	MEAN	STD	SKEWNESS	KURTOSIS
1	INTENSIDAD	0	346167042	0	7496	408.877	626.561	3.73011	19.1015

Figura 24. Estadísticos intensidad tras ajuste

Observamos que, tras el ajuste, el nivel máximo de intensidad es de 7.496 vehículos por hora. Comprobamos a qué punto de medición corresponde dicho valor y el valor sí parece ser consecuente con el resto de medidas cercanas a esa hora:

IDELEM	FECHA	IDENTIF	TIPO_ELEM	INTENSIDAD	OCUPACION	CARGA	VMED	ERROR
6666	2015-07-15 16:45:00	PM10711	494	6017	12	65	81	N
6666	2015-07-15 17:00:00	PM10711	494	6104	11	64	83	N
6666	2015-07-15 17:15:00	PM10711	494	6279	12	67	83	N
6666	2015-07-15 17:30:00	PM10711	494	6548	13	70	80	N
6666	2015-07-15 17:45:00	PM10711	494	6616	13	70	80	N
6666	2015-07-15 18:00:00	PM10711	494	6936	13	74	79	N
6666	2015-07-15 18:15:00	PM10711	494	6868	13	72	78	N
6666	2015-07-15 18:30:00	PM10711	494	7288	15	77	72	N
6666	2015-07-15 18:45:00	PM10711	494	7496	16	78	73	N
6666	2015-07-15 19:00:00	PM10711	494	6697	14	72	76	N
6666	2015-07-15 19:15:00	PM10711	494	6788	13	71	78	N
6666	2015-07-15 19:30:00	PM10711	494	6664	13	72	77	N
6666	2015-07-15 19:45:00	PM10711	494	6547	12	69	79	N

Figura 25. Comprobación del valor máximo de intensidad

Con este ajuste nuestro dataset ha pasado de 364 Millones de observaciones a 346 Millones. El tener tantos puntos de medición y con una frecuencia tan alta nos permite hacer este ajuste y desechar aquellas erróneas.

Además de ésta, se ha llevado a cabo otras tareas de depuración de datos que resumimos a continuación:

- Se eliminan aquellas observaciones en las que el valor de intensidad sea 0: Según el descriptivo de los datos, el campo de intensidad corresponde al número de vehículos por hora por lo que parece muy extraño que en una hora no pase ningún vehículo en una ciudad tan transitada como Madrid. Tampoco parece deberse a que los puntos de medición tengan un valor mínimo, ya que encontramos observaciones con valores de intensidad muy pequeños (incluso 1). Por todo ello, entendemos que valores 0 equivalen a valores missing y optamos por su eliminación.

- Se eliminan aquellas observaciones en las que el identificador del punto de medición viene vacío.
- Se desecha la información relativa a las variables velocidad Media, carga y ocupación por la baja calidad de los datos: Por ejemplo, muchísimas observaciones con VMED informada a cero teniendo INTENSIDAD registrada. No se trata de observaciones puntuales que podamos imputar u obviar, sino de días enteros sin marcar VMED por lo que decidimos desechar esta variable al completo.

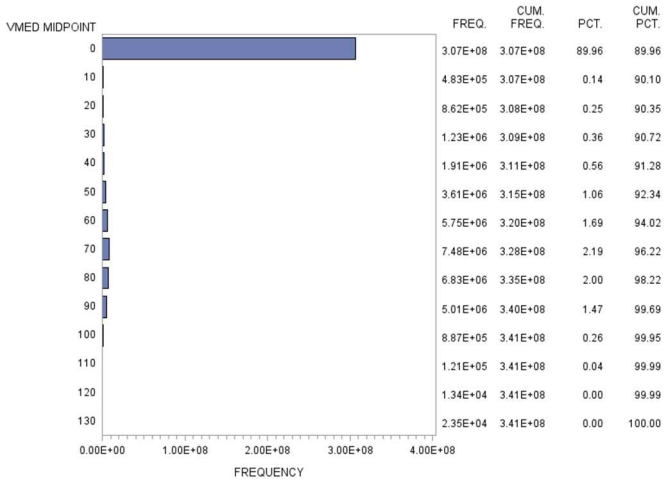


Figura 26. Distribución de la variable VMED

Tras la depuración de datos incorrectos hemos procedido a crear nuevas variables en relación a los datos de intensidad de tráfico. Puede que sea interesante incluir alguna de ellas en la modelización posterior:

	VARIABLES NUEVAS GENERADAS - DESCRIPCIÓN Y DETALLE
MAX INTENSIDAD	Valor máximo de intensidad diaria de la zona
	Se ha calculado como el valor máximo de intensidad marcado por cualquiera de los puntos de medición englobados en la zona de estudio.
SUM MAX INTENSIDAD	Suma de los 30 valores máximos por zona
	Se obtiene el valor máximo diario de intensidad para cada punto de medición. Se suman los valores máximos de las 30 primeras estaciones, agrupados por zona de estudio. Limitamos a 30 los valores a sumar porque hay zonas que tienen muchos más puntos de medición que otras. Pero al menos 30 tienen todas.
MAX MEDIA INTENSIDAD	Media de valores máximos por zona
	Se obtiene valor máximo diario de intensidad para cada punto de medición. Se saca la media de los valores máximos entre todos los puntos de medición englobados en la zona de estudio.
MEDIA INTENSIDAD	Valor medio diario de intensidad de la zona
	Corresponde a la media de los valores medios de todos los puntos de medición de la zona. Es decir: En un primer paso se obtiene el valor medio diario para cada punto de medición de tráfico. 2º Partiendo de los valores medios anteriormente calculados obtendremos el valor medio diario global de la zona.
MEDIA INTENSIDAD PTOMAX	Valor medio diario asociado al punto de control que ha marcado máximo en la zona
	Se obtiene como la media de intensidad de todas las mediciones del día del punto de control que ha marcado máximo en la zona.

Figura 27. Variables de tráfico generadas

4.5. Exploración y tratamiento de los datos climatológicos

Disponemos de información de las 13 estaciones climatológicas repartidas por la Comunidad de Madrid en 13.783 registros. Cada uno de ellos contiene la siguiente información:

Variable	Descripción	Tipo
CLUSTER	Toma valores de 1 a 5 en función de la zona a la que pertenezca la estación.	Categórica
TMED	Temperatura media	Continua
PREC	Nivel de precipitación	Continua
TMIN	Temperatura mínima marcada en el día	Continua
TMAX	Temperatura máxima marcada en el día	Continua
DIR	Dirección del viento asociado a la máxima racha	Continua
VELMEDIA	Velocidad media del viento	Continua
RACHA	Máxima velocidad racha viento	Continua
PRESMAX	Presión máxima marcada en el día	Continua
PRESMIN	Presión mínima marcada en el día	Continua
HORATMIN_CAT	Se categoriza la hora a la que se ha marcado la temperatura mínima (Valores 1/2/3)	Categórica
HORATMAX_CAT	Se categoriza la hora a la que se ha marcado la temperatura máxima (Valores 1/2/3)	Categórica
HORARACHA_CAT	Se categoriza la hora a la que se ha marcado la racha máxima de viento (Valores 1/2/3)	Categórica
HORAPRESMIN_CAT	Se categoriza la hora a la que se ha marcado la presión mínima (Valores 1/2/3)	Categórica
HORAPRESMAX_CAT	Se categoriza la hora a la que se ha marcado la presión máxima (Valores 1/2/3)	Categórica

Figura 28. Variables incluidas en dataset de datos climatológicos

El principal problema detectado al tratar los datos climatológicos es la elevada presencia de valores missing en algunas variables. Para solucionarlo se han aplicado los siguientes criterios de imputación:

1. Ante la aparición de un valor missing, recuperamos el valor medido por otra estación en el mismo día y asociado al mismo cluster.
2. Se imputa (si lo hay) el valor medido por otra estación en el mismo día con independencia de que pertenezca o no a la misma zonificación (cluster).
3. Las variables con % missing superior al 25% son desechadas.

		INFORMADO	MISSING	% MISSING
Datos originales	CLUSTER	13.783	0	0,00%
	TMED	13.164	619	4,49%
	PREC	12.870	913	6,62%
	TMIN	13.164	619	4,49%
	TMAX	13.164	619	4,49%
	DIR	12.173	1.610	11,68%
	VELMEDIA	12.297	1.486	10,78%
	RACHA	12.173	1.610	11,68%
	SOL	6.574	7.209	52,30%
	PRESMAX	10.873	2.910	21,11%
	PRESMIN	10.873	2.910	21,11%
	HORATMIN_CAT	12.464	1.319	9,57%
	HORATMAX_CAT	11.929	1.854	13,45%
	HORARACHA_CAT	10.487	3.296	23,91%
	HORAPRESMIN_CAT	8.557	5.226	37,92%
	HORAPRESMAX_CAT	7.375	6.408	46,49%
	CLUSTER	13.783	0	0,00%
FASE1: Aplicación del primer criterio de imputación	TMED_AJUST	13.770	13	0,09%
	PREC_AJUST	13.736	47	0,34%
	TMIN_AJUST	13.770	13	0,09%
	TMAX_AJUST	13.770	13	0,09%
	DIR_AJUST	13.523	260	1,89%
	VELMEDIA_AJUST	13.525	258	1,87%
	RACHA_AJUST	13.523	260	1,89%
	PRESMAX_AJUST	12.724	1.059	7,68%
	PRESMIN_AJUST	12.724	1.059	7,68%
	HORATMIN_CAT_AJUST	13.696	87	0,63%
	HORATMAX_CAT_AJUST	13.568	215	1,56%
	HORARACHA_CAT_AJUST	13.238	545	3,95%
FASE2: Aplicación del segundo criterio de imputación	CLUSTER	13.783	0	0,00%
	TMED_AJUST2	13.783	0	0,00%
	PREC_AJUST2	13.783	0	0,00%
	TMIN_AJUST2	13.783	0	0,00%
	TMAX_AJUST2	13.783	0	0,00%
	DIR_AJUST2	13.783	0	0,00%
	VELMEDIA_AJUST2	13.783	0	0,00%
	RACHA_AJUST2	13.783	0	0,00%
	PRESMAX_AJUST2	13.783	0	0,00%
	PRESMIN_AJUST2	13.783	0	0,00%
	HORATMIN_CAT_AJUST2	13.783	0	0,00%
	HORATMAX_CAT_AJUST2	13.783	0	0,00%
	HORARACHA_CAT_AJUST2	13.783	0	0,00%

Figura 29. Proceso de imputación valores missing

Se desechan las variables SOL, HORAPRESMIN_CAT Y HORAPRESMAX_CAT por tener un porcetaje de missing superior al 25%. Tras la aplicación de los criterios comentados anteriormente y, una vez comprobado la ausencia de valores missing (resultados en Figura 27), definimos las variables finales que van a ser imputadas a cada zona:

- Para las variables continuas nos quedamos con la media asociada a la zona/cluster.
- Para las variables categóricas nos quedamos con el valor más frecuente (moda) asociado al cluster.

4.6. Estudio de la autocorrelación de los niveles de NO2

Muy posiblemente el nivel de NO2 de un día esté muy relacionado con el del día anterior. Esto es algo que tenemos que tener en cuenta a la hora de modelizar. Por ello, como paso previo, vamos a estudiar en este punto la posible autocorrelación de la serie de NO2 con sus valores anteriores.

Para este estudio vamos a seleccionar la estación que más veces ha superado el nivel de preaviso: Estación de Pza. Fernández Ladreda (ver Figura 3).

El siguiente gráfico muestra los valores máximos de NO2 para esta estación:

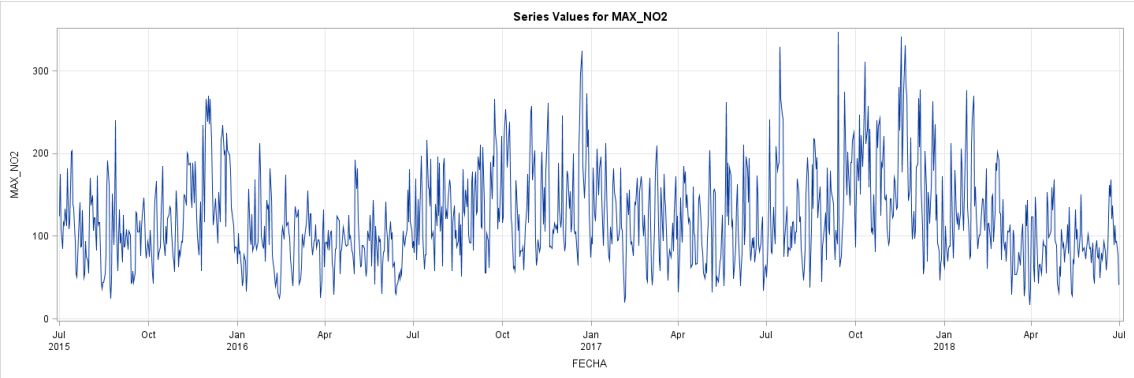


Figura 30. Serie temporal máximo diario NO2

Como primer paso, realizamos el test de Durbin-Watson para determinar si los valores máximos de (NO2Maxd) están relacionados con los del día anterior (NO2Maxd-1). La hipótesis nula es H0: no existe correlación.

Dependent Variable	MAX_NO2

Ordinary Least Squares Estimates			
SSE	3289720	DFE	1094
MSE	3007	Root MSE	54.83664
SBC	11899.8641	AIC	11889.8653
MAE	43.0127034	AICC	11889.8763
MAPE	46.7696162	HQC	11893.6486
		Regress R-Square	0.0043
		Total R-Square	0.0043

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	0.7948	<.0001	1.0000
2	1.2069	<.0001	1.0000
3	1.4484	<.0001	1.0000
4	1.5348	<.0001	1.0000

Observando el estadístico de primer orden, vemos que es significativo, por lo que se rechaza la hipotésis nula de inexistencia de correlación. Es decir, las observaciones de un día sí están relacionadas con las del día anterior.

Figura 31. Test Durbin Watson

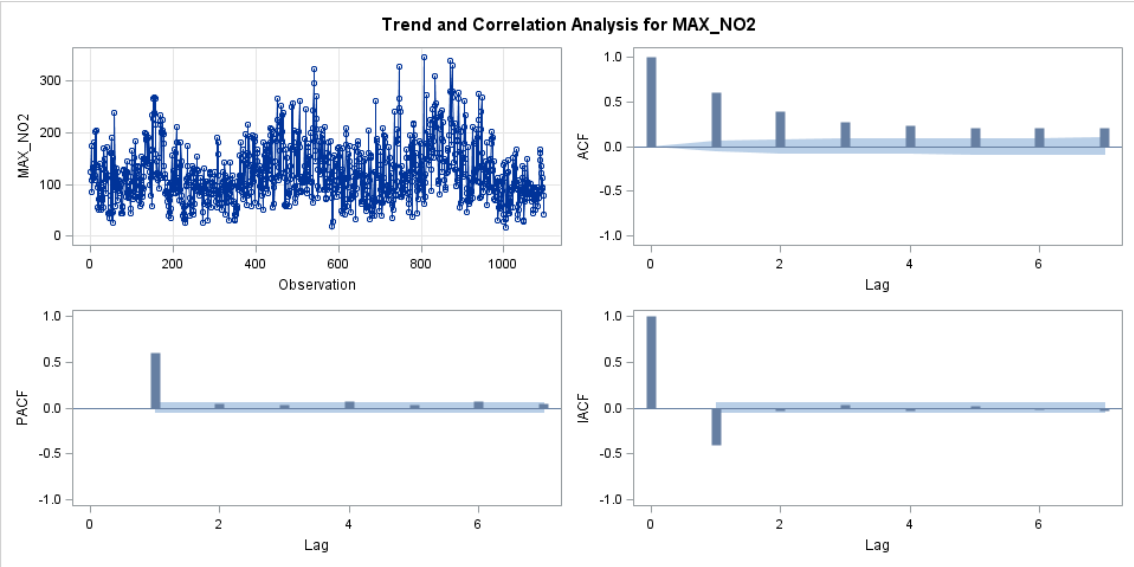


Figura 32. Análisis de tendencia y autocorrelación para la serie del NO2

Si nos fijamos en la función de autocorrelación simple (ACF) vemos que el coeficiente va descendiendo muy poco a poco, lo que indica que la observación en un periodo está muy relacionada con la observación del periodo anterior, es decir, la serie tiene tendencia.

También vamos a estudiar si existe o no correlación entre los residuos. Para ello, realizamos el test de ruido blanco en el que la hipótesis nula es que los residuos son aleatorios o independientes.

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	813.58	6	<.0001	0.603	0.397	0.277	0.234	0.204	0.211

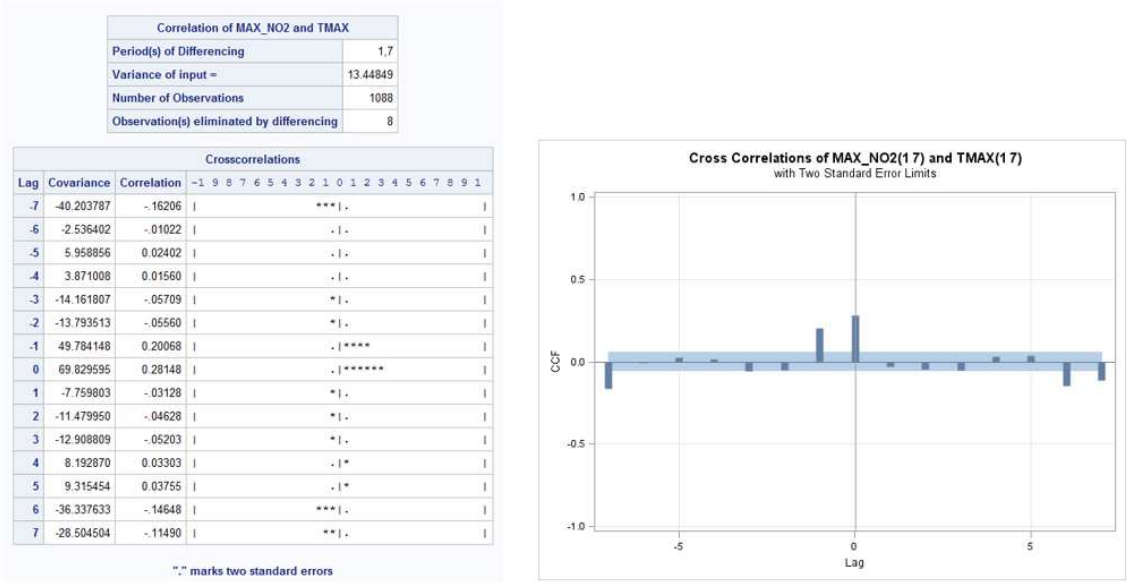
Figura 33. Test de ruido blanco

En cuanto al test de ruido blanco, viendo los coeficientes de autocorrelación se rechaza la hipótesis de independencia para un nivel de confianza del 99%.

Considerando todo lo anterior, concluimos que será conveniente incluir el retardo del NO2 máximo como posible variable en la modelización.

4.7. Estudio de la correlación cruzada: NO2 y variables independientes

En este punto vamos a estudiar si existe correlación entre los distintos retardos de las variables independientes y el nivel máximo de NO2. La idea es mantener fija una variable y la otra que queremos comparar irla retardando día a día. Para cada retardo se calculará la correlación entre ambas series. Para ello nos ayudaremos de los correlogramas con retardos.



Correlation of MAX_NO2 and TMAX

Period(s) of Differencing	1,7
Variance of input =	13.44849
Number of Observations	1088
Observation(s) eliminated by differencing	8

Crosscorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
-7	-40.203787	-.16206	***|. |
-6	-2.536402	-.01022	.|. |
-5	5.958856	0.02402	.|. |
-4	3.871008	0.01560	.|. |
-3	-14.161807	-.05709	*|. |
-2	-13.793513	-.05560	*|. |
-1	49.784148	0.20068	.|**** |
0	69.829595	0.28148	.|****** |
1	-7.759803	-.03128	*|. |
2	-11.479950	-.04628	*|. |
3	-12.908809	-.05203	*|. |
4	8.192870	0.03303	.|* |
5	9.315454	0.03755	.|* |
6	-36.337633	-.14648	***|. |
7	-28.504504	-.11490	**|. |

"." marks two standard errors

Figura 34. Análisis de correlaciones cruzadas entre NO2 y TMAX

Por ejemplo, comparando la temperatura máxima con el nivel máximo de NO2 se observa que existe autocorrelación entre el valor de la temperatura máxima en d-1 con el valor de NO2.

Debido a ello y, de cara a la modelización, hemos creado variables con los valores retardados de nuestras variables explicativas originales.

4.8. Construcción de un tablón sobre el que modelizar

Una vez se han llevado a cabo las actuaciones comentadas en los pasos anteriores, se ha construido un tablón que será la base para nuestra modelización. Dicho tablón contiene la información de los distintos orígenes más las nuevas variables definidas, pero ahora de una manera agrupada, con misma periodicidad (diaria) y con los datos climatológicos y de tráfico imputados a cada zona de estudio. En el "Anexo 2", se adjunta el código con el detalle de la construcción de dicho tablón.

5. Técnicas de clasificación

5.1. Clasificación mediante regresión logística

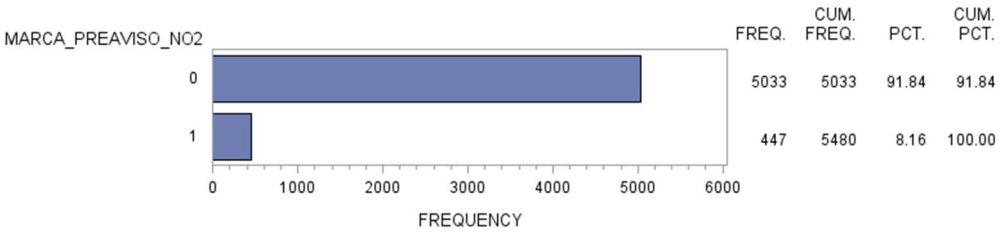


Figura 35. Distribución de evento (Preaviso NO_2=1)

El tablón del que partimos contiene 5.480 observaciones. Como puede observarse, partimos de una situación en la que en el 8,16% de las observaciones se ha superado el nivel de preaviso de NO2. El objetivo de nuestro modelo por lo tanto ha consistido en mejorar dicho porcentaje, pues sería la tasa de error que obtendríamos en un modelo que clasificara a todas las mediciones con MARCA_PREAVISO_NO_2=0.

Se han creado las variables con retardo que comentábamos en el apartado anterior y, observando las matrices de correlaciones, vemos que alguna de ellas puede ser interesante para incluir como efecto en la regresión.

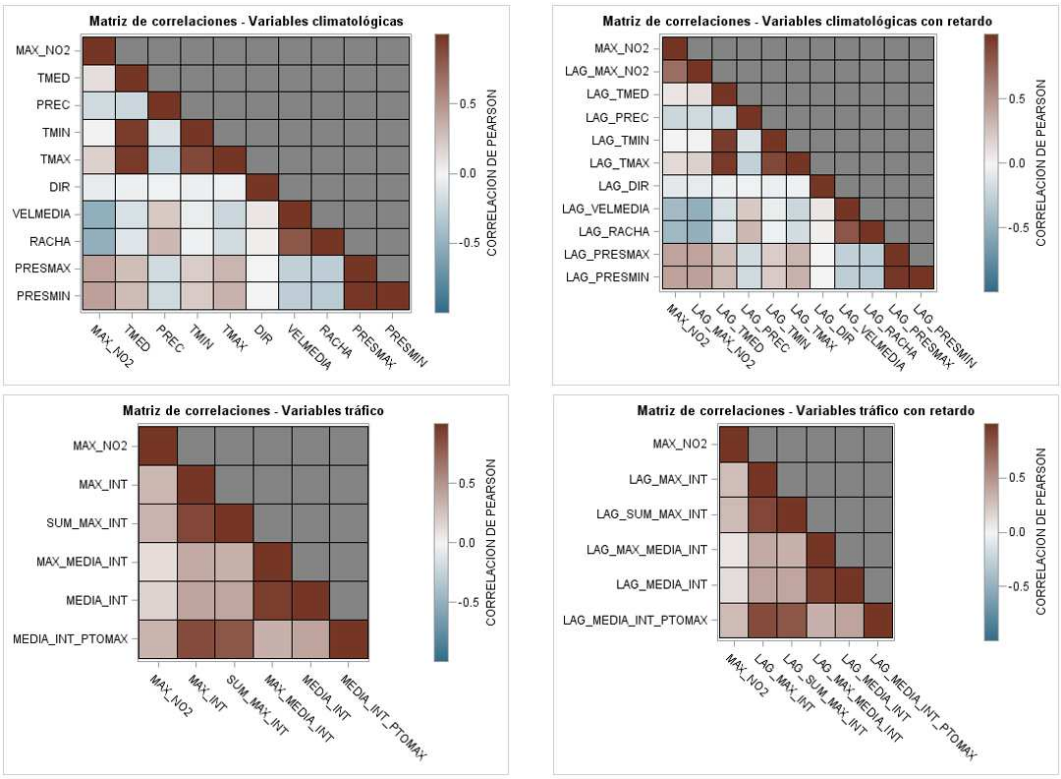


Figura 36. Matriz de correlaciones con y sin retardos

A la hora de modelizar mediante regresión, la primera prueba ha consistido en un modelo sin interacciones. Se han incluido únicamente los efectos principales y se ha calculado el mejor modelo por el método STEPWISE:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-17.2219	2.7970	37.9111	<.0001
PREC	1	-0.2872	0.1109	6.7059	0.0096
TMIN	1	-0.1706	0.0262	42.5019	<.0001
TMAX	1	0.1780	0.0221	64.7420	<.0001
VELMEDIA	1	-0.7367	0.1226	36.1140	<.0001
RACHA	1	-0.1430	0.0443	10.4406	0.0012
PRESMIN	1	0.0580	0.0180	10.3948	0.0013
MAX_INTENSIDAD	1	-0.00112	0.000158	50.0031	<.0001
SUM_MAX_INTENSIDAD	1	0.000070	7.294E-6	93.2588	<.0001
LAG_MAX_NO2	1	0.0220	0.00143	237.6365	<.0001
LAG_VELMEDIA	1	0.1910	0.0689	7.6771	0.0056
LAG_PRESMIN	1	-0.0477	0.0179	7.1288	0.0076

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
PREC	0.750	0.604	0.933
TMIN	0.843	0.801	0.888
TMAX	1.195	1.144	1.248
VELMEDIA	0.479	0.376	0.609
RACHA	0.867	0.795	0.945
PRESMIN	1.060	1.023	1.098
MAX_INTENSIDAD	0.999	0.999	0.999
SUM_MAX_INTENSIDAD	1.000	1.000	1.000
LAG_MAX_NO2	1.022	1.019	1.025
LAG_VELMEDIA	1.210	1.057	1.386
LAG_PRESMIN	0.953	0.921	0.987

Association of Predicted Probabilities and Observed Responses

Percent Concordant	93.2	Somers' D	0.867
Percent Discordant	6.5	Gamma	0.870
Percent Tied	0.3	Tau-a	0.131
Pairs	2160400	c	0.934

Figura 37. Efectos seleccionados Stepwise

A continuación, utilizando la macro *%interacttodolog* (autoría: Javier Portela) se ha construido una lista con todas las posibles interacciones. Hemos pedido que la salida salga ordenada según el estadístico F:

Obs	variable	AIC	FValue	ProbF
1	LAG_MAX_NO2	-9983.74648	1443.29	<.0001
2	MODA_HORATMIN_CAT*LAG_MAX_NO2	-9981.75635	497.70	<.0001
3	MODA_HORARACHA_CAT*LAG_MAX_NO2	-9960.63327	488.75	<.0001
4	MODA_HORATMAX_CAT*LAG_MAX_NO2	-9934.04666	477.53	<.0001
5	LAG_MODA_HORATMIN_CAT*LAG_MAX_NO2	-9944.34148	477.14	<.0001
6	LAG_MODA_HORARACHA_CAT*LAG_MAX_NO2	-9932.78945	472.28	<.0001
7	LAG_MODA_HORATMAX_CAT*LAG_MAX_NO2	-9931.97860	471.94	<.0001
8	RACHA	-9102.59326	441.29	<.0001
9	VELMEDIA	-9056.00084	391.15	<.0001
10	IDZONA*LAG_MAX_NO2	-10039	304.51	<.0001
11	LAG_RACHA	-8969.83347	300.93	<.0001
12	LAG_VELMEDIA	-8921.16890	249.82	<.0001
13	PRESMIN	-8894.95966	221.12	<.0001
14	LAG_PRESMIN	-8889.79878	217.11	<.0001
15	PRESMAX	-8884.16515	209.90	<.0001
16	LAG_PRESMAX	-8880.50087	207.46	<.0001
17	MODA_HORATMAX_CAT*RACHA	-9103.77433	148.91	<.0001
18	MODA_HORARACHA_CAT*RACHA	-9103.44586	148.79	<.0001
19	LAG_MODA_HORATMIN_CAT*RACHA	-9076.24386	148.08	<.0001
20	MODA_HORATMIN_CAT*RACHA	-9100.86770	147.86	<.0001
21	LAG_MODA_HORARACHA_CAT*RACHA	-9073.63545	147.14	<.0001
22	LAG_MODA_HORATMAX_CAT*RACHA	-9073.47639	147.08	<.0001
23	LAG_MODA_HORARACHA_CAT*VELMEDIA	-9031.16130	131.90	<.0001
24	MODA_HORATMAX_CAT*VELMEDIA	-9055.09055	131.44	<.0001
25	LAG_MODA_HORATMAX_CAT*VELMEDIA	-9029.84466	131.43	<.0001
26	LAG_MODA_HORATMIN_CAT*VELMEDIA	-9028.89401	131.09	<.0001
27	MODA_HORARACHA_CAT*VELMEDIA	-9053.48051	130.86	<.0001
28	MODA_HORATMIN_CAT*VELMEDIA	-9053.00714	130.69	<.0001
29	SUM_MAX_INTENSIDAD	-8675.53332	111.63	<.0001
30	MODA_HORARACHA_CAT*LAG_RACHA	-8963.02062	107.70	<.0001
31	IDZONA*RACHA	-9171.32541	104.88	<.0001
32	MODA_HORATMIN_CAT*LAG_RACHA	-8951.86505	103.77	<.0001
33	LAG_MODA_HORATMAX_CAT*LAG_RACHA	-8970.47582	101.91	<.0001
34	LAG_MODA_HORATMIN_CAT*LAG_RACHA	-8967.85839	100.99	<.0001
35	MEDIA_INTENSIDAD_PTOMAX	-8664.68623	100.57	<.0001
36	LAG_MODA_HORARACHA_CAT*LAG_RACHA	-8966.65220	100.56	<.0001
37	MODA_HORATMAX_CAT*LAG_RACHA	-8942.67599	100.53	<.0001
38	IDZONA*VELMEDIA	-9135.23188	97.00	<.0001
39	LAG_SUM_MAX_INTENSIDAD	-8643.67458	91.87	<.0001
40	MODA_HORARACHA_CAT*LAG_VELMEDIA	-8907.79573	88.31	<.0001
41	MODA_HORATMIN_CAT*PRESMIN	-8929.89824	87.23	<.0001
42	MODA_HORATMIN_CAT*LAG_PRESMIN	-8903.35157	86.76	<.0001
43	MODA_HORATMIN_CAT*LAG_VELMEDIA	-8896.89698	84.50	<.0001
44	LAG_MODA_HORATMAX_CAT*LAG_VELMEDIA	-8920.47091	84.39	<.0001
45	MODA_HORATMIN_CAT*PRESMAX	-8921.42444	84.27	<.0001
46	MODA_HORATMIN_CAT*LAG_PRESMAX	-8895.59727	84.05	<.0001
47	MODA_HORATMAX_CAT*LAG_VELMEDIA	-8895.37459	83.97	<.0001
48	LAG_MODA_HORATMIN_CAT*LAG_VELMEDIA	-8917.80175	83.46	<.0001
49	LAG_MODA_HORARACHA_CAT*LAG_VELMEDIA	-8917.52986	83.37	<.0001
50	LAG_MODA_HORATMIN_CAT*PRESMIN	-8893.21173	83.22	<.0001

Figura 38. Interacciones creadas

Tras construir todas las posibles interacciones (260) se han incluido éstas como posibles efectos a seleccionar en nuestro modelo. Incluyendo las interacciones y, ejecutando de nuevo por el método Stepwise, los efectos elegidos han sido los siguientes:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-46.3707	39.7089	1.3637	0.2429
LAG_MAX_NO2*IDZONA	1	1	-0.00333	0.0438	0.0058	0.9394
LAG_MAX_NO2*IDZONA	2	1	0.00171	0.0439	0.0015	0.9689
LAG_MAX_NO2*IDZONA	3	1	0.0140	0.0439	0.1011	0.7506
LAG_MAX_NO2*IDZONA	4	1	-0.0126	0.1751	0.0052	0.9427
SUM_MAX_INTENSIDAD		1	0.000022	4.36E-6	25.4477	<.0001
TMAX*LAG_MODA_HORATM	1	1	-0.1277	0.1276	1.0019	0.3169
TMAX*LAG_MODA_HORATM	2	1	0.0299	0.0643	0.2159	0.6421
TMED		1	-0.5653	0.0635	79.1705	<.0001
LAG_MODA_HORATMAX_CA	1	1	3.2043	1.5105	4.5002	0.0339
LAG_MODA_HORATMAX_CA	2	1	-0.5919	0.7797	0.5762	0.4478
MODA_HORARACHA_CAT	1	1	-0.5423	0.3811	2.0250	0.1547
MODA_HORARACHA_CAT	2	1	0.8221	0.2925	7.8989	0.0049
TMAX		1	0.5224	0.0858	37.0491	<.0001
PREC		1	-0.2157	0.1064	4.1071	0.0427
PRESMIN		1	0.0383	0.0468	0.6707	0.4128
IDZONA	1	1	-18.8595	42.5338	0.1966	0.6575
IDZONA	2	1	-99.8881	45.9309	4.7295	0.0296
IDZONA	3	1	23.4581	41.1060	0.3257	0.5682
IDZONA	4	1	51.6908	153.4	0.1136	0.7361
PRESMIN*IDZONA	1	1	0.0227	0.0496	0.2088	0.6477
PRESMIN*IDZONA	2	1	0.1046	0.0528	3.9209	0.0477
PRESMIN*IDZONA	3	1	-0.0246	0.0484	0.2580	0.6115
PRESMIN*IDZONA	4	1	-0.0595	0.1825	0.1063	0.7444
VELMEDIA		1	-0.6793	0.1321	26.4332	<.0001
RACHA		1	-0.0849	0.0486	3.0567	0.0804
RACHA*MODA_HORARACHA	1	1	0.1530	0.0474	10.4374	0.0012
RACHA*MODA_HORARACHA	2	1	-0.1245	0.0373	11.1390	0.0008
LAG_MAX_NO2		1	0.0166	0.0438	0.1431	0.7052

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
SUM_MAX_INTENSIDAD	1.000	1.000	1.000
TMED	0.568	0.502	0.644
PREC	0.806	0.654	0.993
VELMEDIA	0.507	0.391	0.657

Association of Predicted Probabilities and Observed Responses

Percent Concordant	94.3	Somers' D	0.888
Percent Discordant	5.4	Gamma	0.891
Percent Tied	0.3	Tau-a	0.134
Pairs	2160400	c	0.944

Figura 39. Efectos seleccionados con interacciones

A continuación, hemos dividido las variables categóricas en tantas dummys como categorías tienen, de modo que se puedan escoger efectos concretos que puedan ser significativos y desechar aquéllos otros que no sean relevantes.

Incluir categorías poco representadas en el modelo puede dar lugar a sobreajuste, por lo que hemos definido un parámetro de corte de 400 (aproximadamente 8% sobre total de observaciones, % similar al evento) para no crear dummys con categorías que no lleguen a ese mínimo de observaciones. Para este paso se ha utilizado la macro *%nombresmodbien* la cual facilita la labor (programada por Javier Portela).

A continuación, se ha calculado sucesivas regresiones stepwise para 300 semillas diferentes. Nos interesa comprobar si, para estas 300 semillas existe algún modelo que se repita varias veces. La macro *%randomselectlog* (Javier Portela), nos facilita la salida con los modelos más frecuentes en estas 300 semillas.

Primero se ha realizado con variables originales, sin tener en cuenta interacciones ni dummys:

	efecto	Frequency Count	Percent of Total Frequency
1	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_MAX_INTENSIDAD	17	5.6478405316
2	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_MAX_INTENSIDAD LAG_MEDIA_INTENSIDAD	11	3.6544850498
3	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN LAG_VELMEDIA LAG_MAX_INTENSIDAD	7	2.3255813953
4	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN LAG_VELMEDIA LAG_MAX_INTENSIDAD LAG_MEDIA_INTENSIDAD	6	1.9933554817
5	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_MAX_INTENSIDAD LAG_MEDIA_INTENSIDAD	6	1.9933554817
6	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_PRESMAX LAG_MAX_INTENSIDAD	6	1.9933554817
7	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_MAX_INTENSIDAD LAG_MEDIA_INTENSIDAD	5	1.6611295681
8	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_PREC LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_SUM_MAX_INTENSID	5	1.6611295681
9	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_TMIN LAG_TMAX LAG_VELMEDIA LAG_RACHA LAG_MAX_INTENSIDAD	5	1.6611295681
10	Intercept PREC TMIN TMAX DIR VELMEDIA RACHA PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD MEDIA_INTENSIDAD_PTO LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN LAG_VELMEDIA LAG_PRESMAX LAG_MAX_INTENSIDAD	5	1.6611295681

Figura 40. Extracto de los modelos más frecuentes tras 300 semillas

En el pantallazo anterior vemos los modelos que más veces se han repetido utilizando 300 semillas diferentes. Posteriormente, se ha utilizado la misma técnica pero incluyendo como posibles efectos las dummys creadas anteriormente. Tras ejecutar las 300 regresiones con semillas diferentes no hay ningún modelo que se haya repetido, pero sí que vemos ciertos efectos que son bastante frecuentes:

	efecto	Frequency Count	Percent of Total Frequency
1	LAG_MAX_NO2IDZONA_1	301	3.7206427689
2	LAG_TMINIDZONA_4	293	3.6217552534
3	LAG_VELMEDIAMODA_HORATMIN_C	292	3.609394314
4	MEDIA_INTENSIDADIDZONA_5	266	3.2880098888
5	LAG_PRECIDZONA_5	261	3.2262051916
6	LAG_TMED	256	3.1644004944
7	PRECIDZONA_5	254	3.1396786156
8	LAG_TMINIDZONA_5	246	3.0407911001
9	SUM_MAX_INTENSIDADIDZONA_1	243	3.0037082818
10	IDZONA_1LAG_MODA_HORATMAX_C	235	2.9048207664

Figura 41. Efectos más frecuentes

A continuación, comprobaremos mediante validación cruzada cuál es el mejor modelo calculado de entre las distintas alternativas. Utilizaremos 100 semillas para cada modelo y, en cada semilla, haremos una partición de datos training 70% y test 30%. Nos interesará aquel modelo que ofrezca una menor tasa de fallos.

Modelo	Detalle
REG1	Modelo resultante de la selección de variables originales + variables con retardos pero sin interacciones, por el método Stepwise
REG2	Modelo resultante de la selección de variables originales + variables con retardos + interacciones, por el método Stepwise
REG3	Modelo más frecuente al ejecutar con 300 semillas con variables originales + variables con retardos y sin interacciones
REG4	Segundo modelo más frecuente al ejecutar con 300 semillas con variables originales + variables con retardos y sin interacciones
REG5	Modelo con las 10 interacciones con mayor valor del estadístico F
REG6	Modelo con las 10 interacciones más frecuentes
REG7	Modelo tentativo: Temperatura mínima, máxima presión, velocidad viento, intensidad tráfico y valor retardado del NO2

Figura 42. Modelos de regresión a comparar

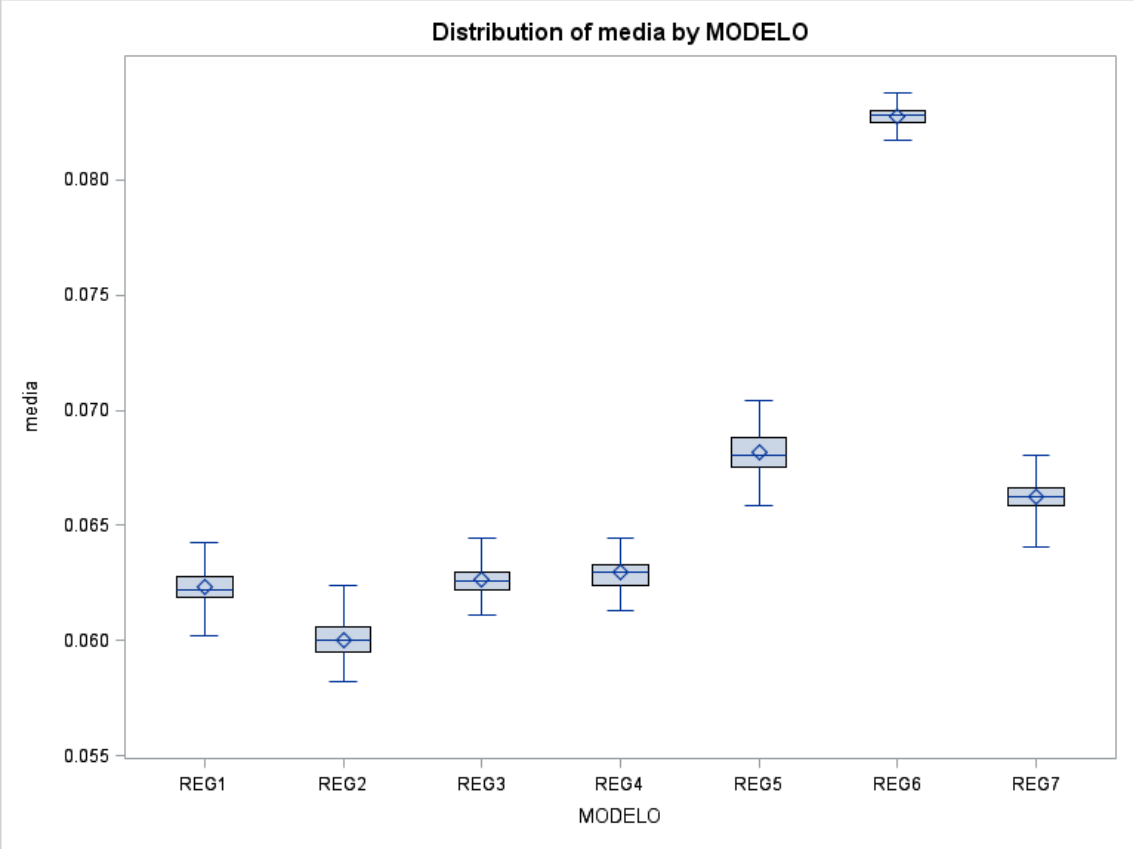


Figura 43. Tasa de error por modelo, tras 100 iteraciones.

Como comentamos en la exploración inicial, se partía de una situación en la que el 8,16% de los sucesos son MARCA_PREAVISO_NO_2=1. Viendo que la media de tasa de fallos se sitúa en el 6% para 100 semillas podemos concluir que mejoramos respecto a la aleatoriedad de un modelo que clasifica a todos los sucesos como MARCA_PREAVISO_NO_2=0.

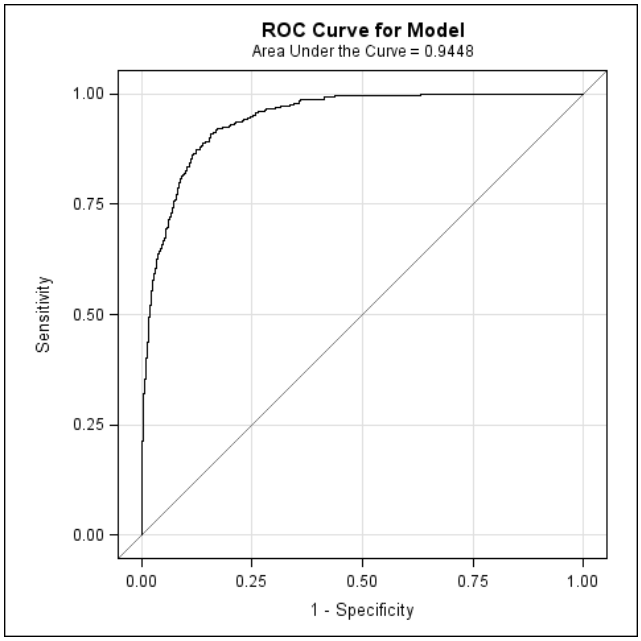


Figura 44. Curva ROC asociada al mejor modelo (Modelo 2)

La curva ROC representa la relación entre la Sensibilidad (probabilidad de que el modelo clasifique correctamente la superación del nivel de preaviso de NO2) y la especificidad (probabilidad de que el modelo no clasifique correctamente) obtenidas para distintos puntos de corte.

La calificación del modelo, se presenta mediante el estadístico 'C', es un valor real situado entre 0 y 1. Este será más exacto cuanto más próximos se encuentre de 1. En el modelo el estadístico 'C' toma un valor 0,9448 por consiguiente se considera buen modelo en términos de poder predictivo.

5.2. Clasificación mediante Redes Neuronales

En una primera aproximación a la modelización mediante redes, se ha estudiado el impacto que tiene en la tasa de error variar el número de nodos y la función de activación.

En esta primera aproximación se han utilizado todas las variables de las que disponemos lo que, seguramente, haga que los modelos calculados sobreajusten y no se comporten del todo bien para datos test.

Modelo	Número de nodos	Algoritmo	Función de activación	Variables
RED01	5	Back propagation	Tanh	Todas
RED02	10	Back propagation	Tanh	Todas
RED03	15	Back propagation	Tanh	Todas
RED04	20	Back propagation	Tanh	Todas
RED05	50	Back propagation	Tanh	Todas
RED06	5	Back propagation	Lin	Todas
RED07	10	Back propagation	Lin	Todas
RED08	15	Back propagation	Lin	Todas
RED09	20	Back propagation	Lin	Todas
RED10	50	Back propagation	Lin	Todas
RED11	5	Levmar	Tanh	Todas
RED12	10	Levmar	Tanh	Todas
RED13	15	Levmar	Tanh	Todas
RED14	20	Levmar	Tanh	Todas
RED15	50	Levmar	Tanh	Todas
RED16	5	Levmar	Lin	Todas
RED17	10	Levmar	Lin	Todas
RED18	15	Levmar	Lin	Todas
RED19	20	Levmar	Lin	Todas
RED20	50	Levmar	Lin	Todas

Figura 45. Parametrización de las redes calculadas

Observando los resultados de los anteriores modelos y basándonos en la tasa de error, en los siguientes gráficos puede verse que la RED16 con función de activación lineal (Lin), algoritmo Levmar y cinco nodos es la que mejor funciona. Las redes con algoritmo Back propagation funcionan todas similar y relativamente bien aunque ninguna supera a los resultados obtenidos mediante regresión. Por otro lado, puede verse que el hecho de aumentar el número de nodos tampoco parece mejorar demasiado la tasa de error, de hecho, al utilizar la función lineal los empeora. Esto seguramente sea debido a que las relaciones entre las distintas variables sean lineales, de ahí que los resultados de la regresión hayan sido tan buenos.

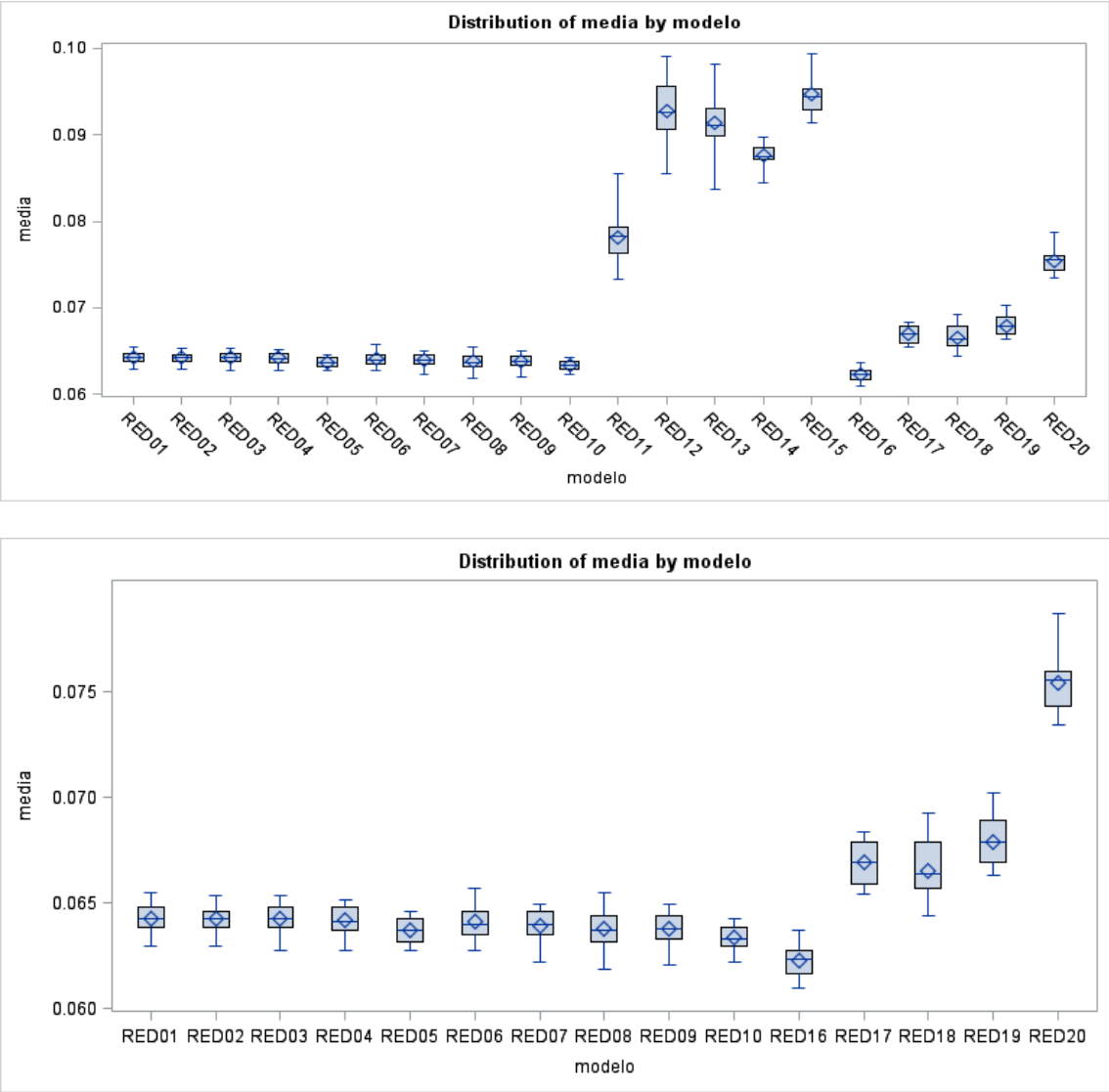
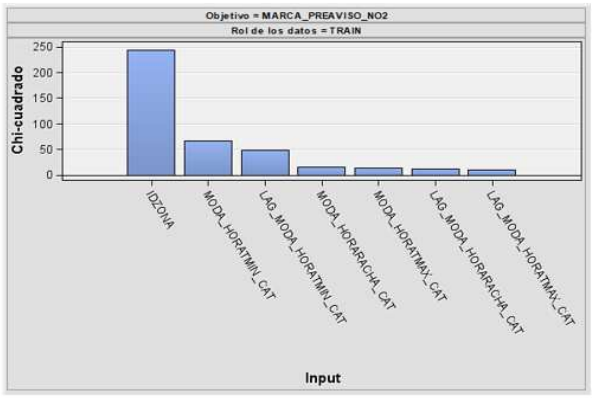


Figura 46. Primera aproximación mediante redes sin selección de variables. Tasa de error por modelo.

También es posible que en esta aproximación se hayan incluido demasiadas variables en el modelo lo que hace que la red sobreajuste y para datos nuevos no se comporte bien. Por ello, de cara a intentar mejorar la capacidad predictiva de nuestra red se ha llevado a cabo una selección de variables. Se han incluido como variables categóricas únicamente aquellas variables más significativas basándonos en el coeficiente Chi-cuadrado. Dicho coeficiente tiene como fin examinar la asociación entre variables categóricas. Existe asociación entre variables cuando los valores de una de ellas dependen de los valores de otra.



Input	Chi-cuadrado	Df	Prob
IDZONA	242.9638	4	<.0001
MODA_HORATMIN_CAT	67.0610	3	<.0001
LAG_MODA_HORATMIN_CAT	48.9383	3	<.0001
MODA_HORARACHA_CAT	16.1635	3	0.0010
MODA_HORATMAX_CAT	13.3384	3	0.0040
LAG_MODA_HORARACHA_CAT	11.5393	3	0.0091
LAG_MODA_HORATMAX_CAT	10.3698	3	0.0157

Figura 47. Chi Cuadrado de variables categóricas

Observando los anteriores estadísticos, se rechaza la hipótesis nula de independencia entre las variables categóricas con la variable independiente. Sin embargo, atendiendo al valor de la Chi-cuadrado puede verse que la variable categórica más significativa con diferencia es la zona de estudio. Por ello, en nuestra segunda modelización únicamente se ha utilizado ésta como variable categórica en la modelización.

En lo que respecta a variables continuas se ha optado por incluir únicamente las diez primeras según el valor de importancia que calcula SAS Miner. Dicho valor es un estadístico calculado en base a un algoritmo interno de SAS Miner que mide la importancia discriminante de cada variable en árboles de decisión.

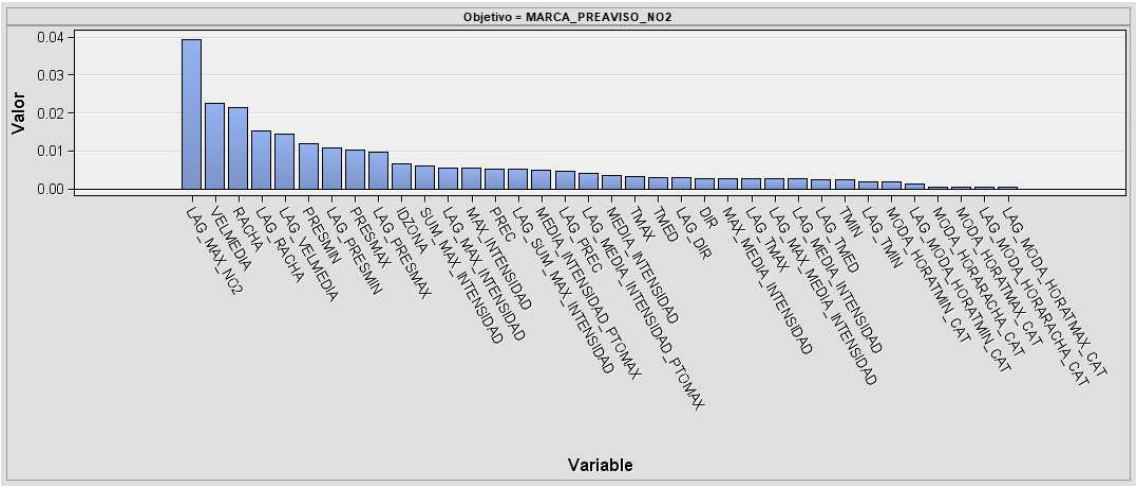


Figura 48. Importancia de variables calculada por SAS Miner según reglas en árboles de decisión

Modelo	Número de nodos	Algoritmo	Función de activación	Variables
RED21	20	Back propagation	Lin	Con selección
RED22	50	Back propagation	Lin	Con selección
RED23	20	Levmar	Lin	Con selección
RED24	50	Levmar	Lin	Con selección

Figura 49. Parametrización de redes con selección de variables

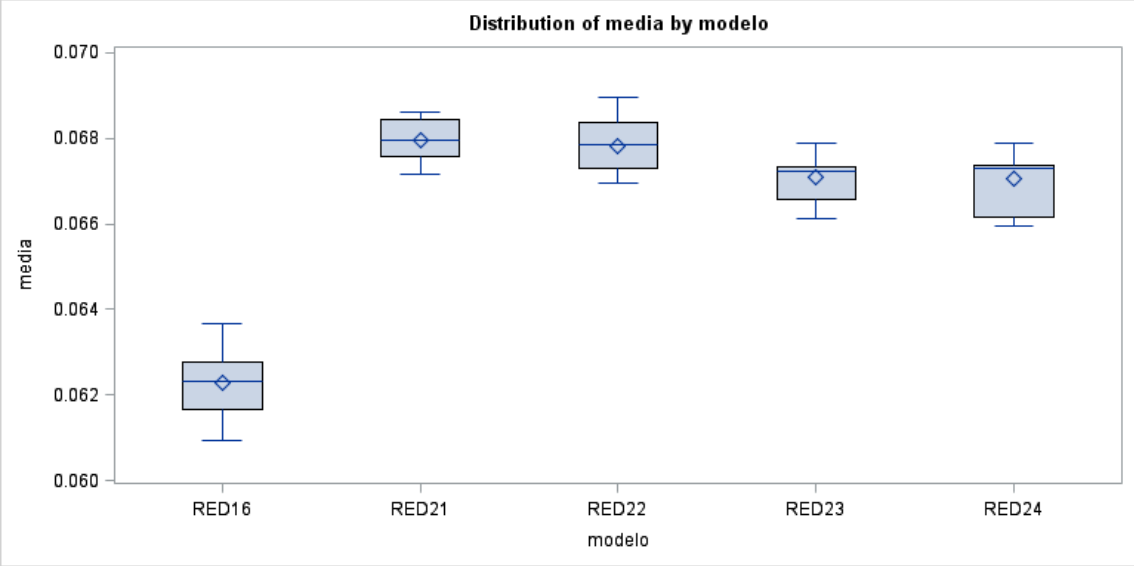


Figura 50. Comparación entre la tasa de error de la mejor red calculada (Red16) vs redes con selección de variables.

Observando los anteriores gráficos vemos que ninguna de las redes con selección de variables consigue mejora la anterior.

5.3. Bootstrap averaging (Bagging)

En este apartado hemos construido varios modelos utilizando la técnica de Bootstrap Averaging, que consiste en generar diferentes árboles sobre muestras aleatorias de nuestro conjunto de datos y, sobre ellas, calcular las predicciones. La predicción final se calculará a partir de la media aritmética de todas las predicciones generadas por los diferentes árboles.

En la elaboración de este modelo intervienen todas las variables y los parámetros que se han monitorizado son: el porcentaje de la muestra a realizar, el número de hojas del árbol, y el número de muestras Bootstrap. Se han hecho pruebas con diferentes valores en estos parámetros de cara a comparar cuál funciona mejor.

Modelo	% de muestra	Nº Hojas	Nº máx árboles
BAG01	70%	10	20
BAG02	70%	10	40
BAG03	70%	10	60
BAG04	70%	20	20
BAG05	70%	20	40
BAG06	70%	20	60
BAG07	70%	30	20
BAG08	70%	30	40
BAG09	70%	30	60
BAG10	90%	10	20
BAG11	90%	10	40
BAG12	90%	10	60
BAG13	90%	20	20
BAG14	90%	20	40
BAG15	90%	20	60
BAG16	90%	30	20
BAG17	90%	30	40
BAG18	90%	30	60

Figura 51. Parametrización de los modelos calculados mediante el algoritmo Bagging

La tasa media de error para cada una de estas pruebas puede verse reflejada en el siguiente diagrama de cajas:

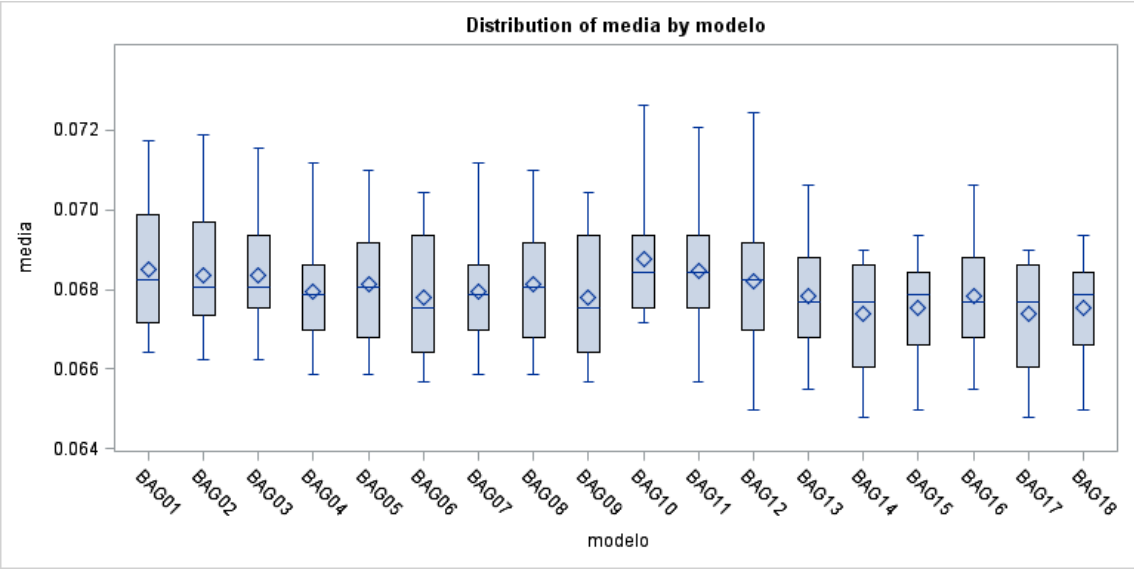


Figura 52. Tasa de error por modelo (Bagging)

5.4. Random Forest

El algoritmo de Random Forest introduce dos fuentes de variación para evitar el sobreajuste: el remuestreo de observaciones y de variables. De esta forma se gana en capacidad de generalización, es decir, puede considerarse como una técnica que difícilmente va a caer en sobreajuste.

Los parámetros a monitorizar en la elaboración de los modelos son: el tamaño de las muestras y si se va a utilizar bootstrap (con reemplazo) o no, el número de árboles a promediar, el número de variables a muestrear en cada nodo y los parámetros básicos de los árboles: número de hojas final, número de divisiones máxima que en nuestro caso es dos por ser una variable binomial, número de observaciones mínimas en cada rama y el p-valor para las distintas divisiones en cada nodo.

Teniendo en cuenta todo ello, se han propuesto diferentes modelos cuya parametrización se resume a continuación:

Modelo	% de muestra	Nº variables	Nº Hojas	Nº máx árboles
RF01	0,7	20	10	20
RF02	0,7	40	10	20
RF03	0,7	20	30	20
RF04	0,7	40	30	20
RF05	0,7	20	10	60
RF06	0,7	40	10	60
RF07	0,7	20	30	60
RF08	0,7	40	30	60
RF09	0,9	20	10	20
RF10	0,9	40	10	20
RF11	0,9	20	30	20
RF12	0,9	40	30	20
RF13	0,9	20	10	60
RF14	0,9	40	10	60
RF15	0,9	20	30	60
RF16	0,9	40	30	60

Figura 53. Parametrización de los modelos calculados mediante el algoritmo Random Forest

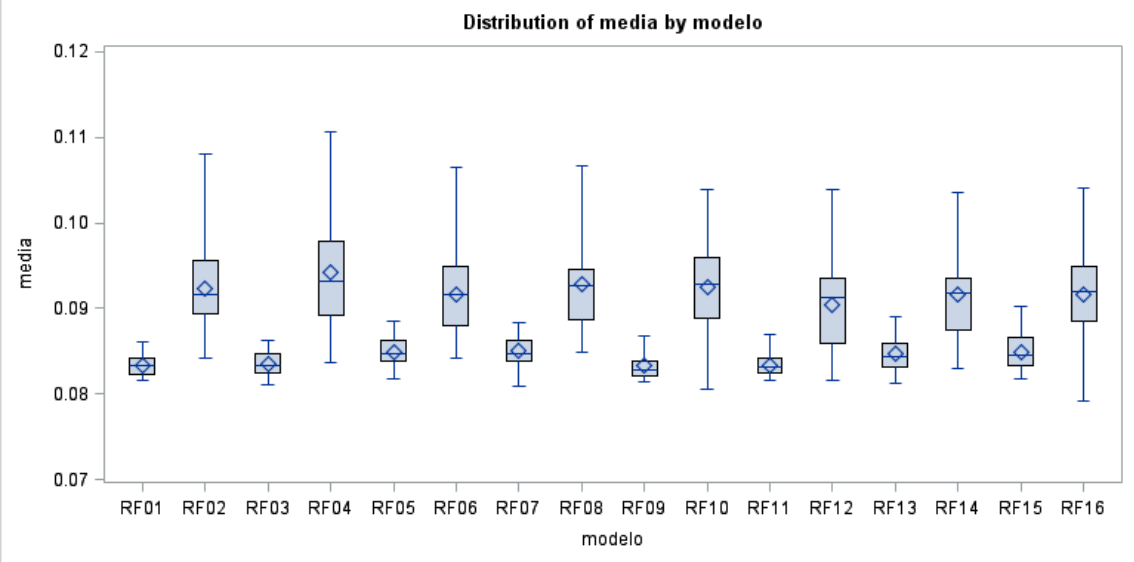


Figura 54. Tasa de error por modelo (Random Forest)

Observando el anterior gráfico se puede apreciar que el hecho de incluir mayor número de variables empeora la capacidad predictora de nuestros modelos. Los modelos parametrizados con 20 variables como máximo son consistentemente mejores que los parametrizados con 40 variables.

5.5. Gradient Boosting

El método de Gradient Boosting consiste en repetir la construcción de árboles de clasificación, modificando ligeramente las predicciones iniciales en cada una de las iteraciones para minimizar los residuos en la dirección de decrecimiento. Debido a ello, si el número de iteraciones aumenta, el residuo tenderá a cero. De cara a evitar el sobreajuste hemos fijado diferentes tasas de aprendizaje, con las que se busca reducir el peso de los modelos sucesivos. Asimismo, se ha monitorizado el funcionamiento de los diferentes modelos variando el número de iteraciones y tamaño final de hojas. El resumen con la parametrización de los diferentes modelos calculados puede verse a continuación:

Modelo	Tasa de aprendizaje	Nº Iteraciones	Profundidad
GB01	0,01	100	5
GB02	0,01	100	15
GB03	0,01	500	5
GB04	0,01	500	15
GB05	0,05	100	5
GB06	0,05	100	15
GB07	0,05	500	5
GB08	0,05	500	15
GB09	0,10	100	5
GB10	0,10	100	15
GB11	0,10	500	5
GB12	0,10	500	15

Figura 55. Parametrización de los modelos calculados mediante el algoritmo Gradient Boosting

La tasa media de error para cada una de estas pruebas puede verse reflejada en el siguiente diagrama de cajas:

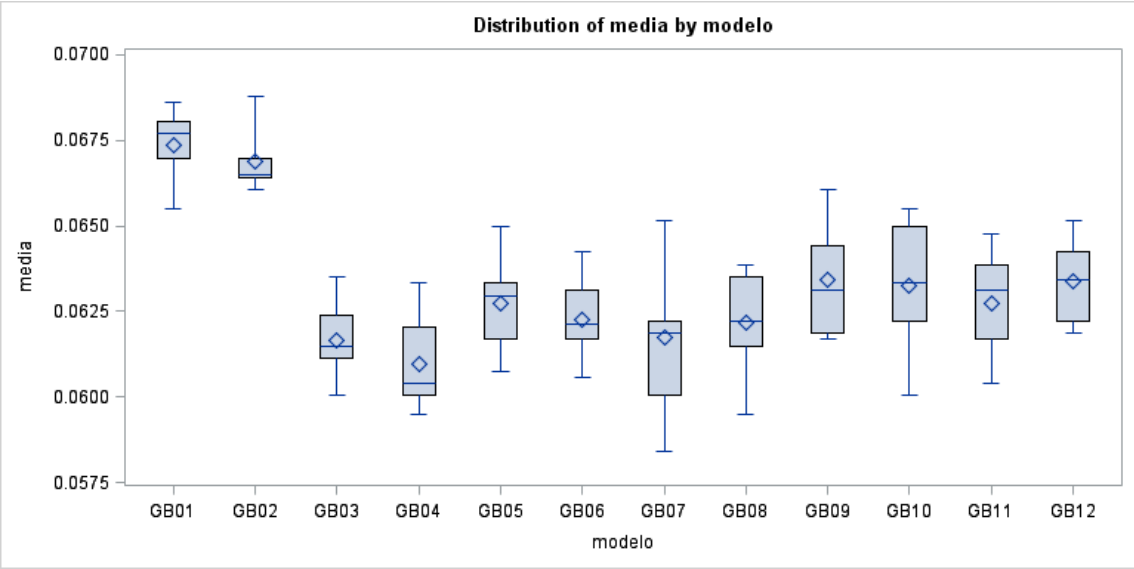


Figura 56. Tasa de error por modelo (Gradient Boosting)

5.6. Comparación de las técnicas de clasificación

Una vez seleccionados los mejores modelos calculados para cada técnica, se procede a su comparación.

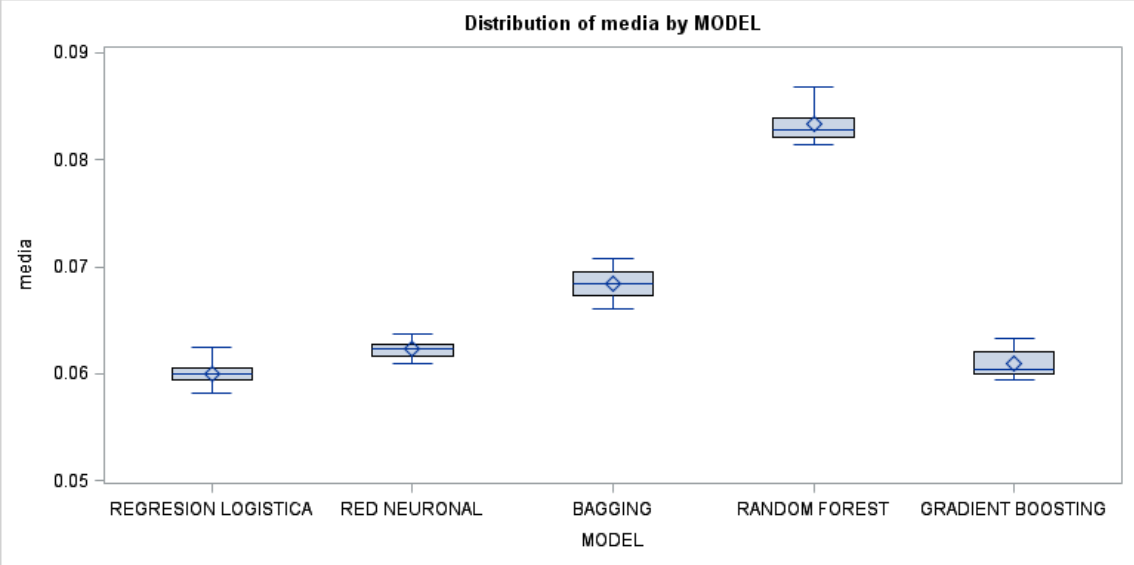


Figura 57. Comparación de la tasa de error en las distintas técnicas de clasificación

- Regresión logística: Modelo resultante de la selección de variables originales + variables con retardos + interacciones, por el método Stepwise.
- Red neuronal: Red con 5 nodos, algoritmo *Levemberg-Marquardt,*función de activación lineal y utilizando todas las variables.
- Bagging: modelo calculado utilizando un porcentaje de muestra del 90%, 30 hojas y 40 muestras Bootstrap.
- Random Forest: modelo calculado utilizando un porcentaje de muestra del 90%, 20 variables, 10 hojas y 20 árboles como máximo.
- Gradient Boosting: modelo calculado con tasa de aprendizaje del 1%, 500 iteraciones y una profundidad de 15 hojas.

En la anterior figura puede verse que la regresión logística es el modelo que mejor se comporta, seguramente debido a la relación lineal entre nuestra variable objetivo y las distintas variables explicativas. La mejor red neuronal y el Gradient Boosting también parecen dar buenos resultados, siendo la tasa de error inferior a nuestro umbral de evento que es el 8,16%. El Random Forest, en cambio, no parece comportarse igual de bien. Posiblemente al hacer el remuestreo de variables haya casos en los que se selecciona alguna variable como discriminante que en observaciones nuevas no se comporta igual de bien, dando lugar a que la tasa de error sea más alta.

6. Principales Conclusiones

En el presente trabajo se ha tratado de abordar distintas fases de un proyecto de minería de datos, desde el aprovisionamiento de los datos hasta la exploración y tratamiento de éstos para su posterior modelización. En nuestro caso, el objetivo marcado buscaba elaborar un modelo con buena capacidad predictiva para poder llevar a cabo acciones de una manera proactiva, adelantándonos a un escenario de alta contaminación por NO_2.

Tras el estudio llevado a cabo en este TFM se pueden extraer las siguientes conclusiones.

- El modelo de regresión logística es el que da un mejor resultado. De cara a la predicción se comporta mejor que otros algoritmos más complejos y ello evidencia que las relaciones entre las distintas variables explicativas y nuestra variable dependiente son de tipo lineal. La tasa de error media se sitúa en torno al 6%, mejorando el umbral de evento del 8,16% por lo que puede ser una buena base sobre la que trabajar.
- El aprovisionamiento automatizado de datos es una parte muy interesante de cualquier proyecto de minería de datos. En nuestro caso, con Python hemos podido construir una herramienta simple pero eficaz para generar una base de datos climatológicos históricos que hemos utilizado posteriormente en la modelización. En el presente TFM se ha utilizado únicamente en la fase inicial de aprovisionamiento de datos pero dicho script podría servir de base para procesos ETL con modelización periódica de cara a predecir los niveles de NO_2 no solamente en Madrid, sino también en otras ciudades.
- Se pone de manifiesto la importancia de una correcta depuración de los datos. Dicha fase nos ha permitido detectar datos de intensidad de tráfico erróneos que, de no haberlos depurado, habrían introducido mucho ruido en la modelización. Al estar utilizando como variable input el valor máximo de intensidad de tráfico por día y zona era necesario eliminar aquellos valores extremos, claramente erróneos. De hecho, en la mejor regresión calculada se elige esta variable en el modelo definido.
- También ha sido muy esclarecedor el estudio de la autocorrelación de los niveles de NO_2 ya que el nivel de NO2 de un día se ve muy influenciado por el nivel de NO2 del día anterior. Generar variables nuevas con retardos para la predicción ha mejorado mucho los resultados.

7. Bibliografía

1. Protocolo de actuación para episodios de contaminación por dióxido de nitrógeno. Disponible para su descarga en el siguiente enlace:

   https://www.madrid.es/UnidadesDescentralizadas/Sostenibilidad/CalidadAire/Ficheros/ProtocoloNO2AprobFinal_201809.pdf

2. Memoria Calidad del Aire 2017 elaborada por el Ayuntamiento de Madrid. Disponible para su descarga en el siguiente enlace: http://www.mambiente.munimadrid.es/opencms/export/sites/default/calaire/Anexos/Memoria2017.pdf
3. Agencia Europea del Medio Ambiente. https://www.eea.europa.eu/themes/air
4. **Alonso Revenga, JM**. (2012) "*Series Temporales. Análisis práctico con SPSS y SAS*." Editorial Académica Española.
5. **D.Michie, D.J.Spiegelhalter, C.C.Taylor** *"Machine Learning, Neural and Statistical Classification",* 1994.
6. **Breiman, L.** (1996) *"Bagging predictors"*. Machine Learning,.
7. **Breiman, L**. (2001) *"Random forests"*. Machine Learning,
8. **Portela J.** (2006). *"Manual de Programación en SAS."* Ediciones Fiec.
9. **Cuadras, C.M.** (2014). *"Nuevos Métodos de Análisis Multivariante."*
10. **Portela, J.** Macros validación cruzada para las distintas técnicas de modelización.

ANEXOS

ANEXO 1: Código Python para explotar API AEMET

```
# Importación de paquetes necesarios
# En caso de no tener algún paquete instalado habrá que instalarlo. Se puede hacer desde
CMD, por ejemplo: python -m pip install unicodecsv

import requests
import json
import re
import csv
import unicodecsv as csv
import itertools
import calendar, datetime
from datetime import date, timedelta
import time

# Obviamos aviso de HTTPS
from requests.packages.urllib3.exceptions import InsecureRequestWarning

requests.packages.urllib3.disable_warnings(InsecureRequestWarning)

# Definimos los datos de cabecera + API_KEY recibida para hacer la llamada HTTP

proxies = {
    'http': 'http://180.185.219.10:8080',
    'https': 'http://180.185.219.10:8080',
}

api_key = {
    "api_key":
"eyJhbGciOiJIUzI1NiJ9.eyJzdWIiOiJnZWx1Y2FsemFkYUBnbWFpbC5jb20iLCJqdGkiOiI5Zjc3NmYxNy1jNm
VlLTQzZWYtOTJjMy1lYzhhZWUyNWUyOWEiLCJpc3MiOiJBRU1FVCIsImlhdCI6MTUzNzk5MzI2MCwidXNlcklkIj
oiOWY3NzZmMTctYzZlZS00M2VmLTkyYzMtZWM4YWVlMjVlMjlhIiwicm9sZSI6IiJ9.6GfoHC3sElRnJcE_OtZRg
ui8wip1cgp8OSgBwL3EBMs"}

cabecera_llamada = {'Accept':
"text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
                    'Accept-Encoding': "gzip, deflate, sdch, br",
                    'Accept-Language': "es-ES,es;q=0.8,en;q=0.6",
                    'Cache-Control': "max-age=0",
                    'Connection': "keep-alive",
                    'Host': "opendata.aemet.es",
                    'Upgrade-Insecure-Requests': "1",
                    'User-Agent': "Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.76 Safari/537.36}"}

# Loop para crear la lista de primeros días y últimos días del mes(es un loop anidado)

inicio = 2011
fin = 2017


# Creamos dos listas iniciales vacías y luego se alimentan
# Lista con día de inicio de mes
# Lista con fin de mes
# Posteriormente las combinaremos creando tuplas

mesesinicio = []
mesesfinal = []

for i in range(inicio, fin + 1):
    for j in range(1, 3):
        primerdia = datetime.date(i, j, 1)
        ultimodia = calendar.monthrange(i, j)
        ultimodia2 = datetime.date(i, j, ultimodia[1])
        mesesinicio.append(primerdia)
        mesesfinal.append(ultimodia2)
        print(primerdia.strftime('%m-%d-%Y') + ' ' + ultimodia2.strftime('%m-%d-%Y'))
        print("\n")

# Utilizamos la funcion ZIP para combinar ambas listas y crear una lista de tuplas con
día inicio - fin mes
listameses = list(zip(mesesinicio,mesesfinal))
```

```
# Listado de estaciones
#
# 3100B - Aranjuez
# 3110C - Buitrago Del Lozoya
# 3191E - Colmenar Viejo
# 3200 - Getafe
# 3129 - Madrid Aeropuerto
# 3194U - Madrid, Ciudad Universitaria
# 3196 - Madrid, Cuatro Vientos
# 3195 - Madrid, Retiro
# 3266A - Puerto Alto Del Leon
# 2462 - Puerto de Navacerrada
# 3338 - Robledo De Chavela
# 3111D - Somosierra
# 3175 - Torrejon De Ardoz

estaciones = ['3100B', '3110C', '3191E', '3200', '3129', '3194U', '3196', '3195',
'3266A', '2462', '3338', '3111D', '3175']



# Ahora vamos a crear las posibles combinaciones de Día inicio - Día fin mes - Estación
de medición
# Para ello combinamos la lista de tuplas de fechas que tenemos con la lista de
estaciones

combinaciones = list(itertools.product(listameses, estaciones))



# Defino la lista de campos en función de los metadatos proporcionados por la web de
AEMET
# Esta listado se utilizará como cabecera del CSV que se genere
campos = ['fecha', 'indicativo', 'nombre', 'provincia', 'altitud', 'tmed', 'prec',
'tmin', 'horatmin', 'tmax', 'horatmax', 'dir', 'velmedia', 'racha', 'horaracha', 'sol',
'presMax', 'horaPresMax', 'presMin', 'horaPresMin']



# Loop de consultas a la web de AEMET

for i in combinaciones:
    url =
"https://opendata.aemet.es/opendata/api/valores/climatologicos/diarios/datos/fechaini/{f
echainicio}T00:00:00UTC/fechafin/{fechafin}T23:59:59UTC/estacion/{idestacion}/"
    # Con la funcion lista.index() conseguimos el índice de la lista que sirve para
obtener cada una de las fechas en cada iteración
    # [0] Corresponde a primer día de mes
    fechainicio = combinaciones[combinaciones.index(i)][0][0].strftime('%Y-%m-%d')
    # [1] Corresponde a último día de mes
    fechafin = combinaciones[combinaciones.index(i)][0][1].strftime('%Y-%m-%d')
    # Marcamos con [1] para quedarnos con la segunda parte de la tupla
    estacionreemplazo = combinaciones[combinaciones.index(i)][1]
    # Reemplazo fecha inicio
    urlaux = re.sub(r'{fechainicio}', fechainicio, url)
    # Remplazo fecha fin sobre urlaux anterior
    urlaux2 = re.sub(r'{fechafin}', fechafin, urlaux)
    # Reemplazo idestacion sobre urlaux2 anterior
    urlnueva = re.sub(r'{idestacion}', estacionreemplazo, urlaux2)
    print("Descargando" + " " + urlnueva)
    respuesta = requests.get(urlnueva, params=api_key, headers=cabecera_llamada,
verify=False)
    # AEMET devuelve un enlace temporal que contiene los datos en formato JSON
    json_data = json.loads(respuesta.text)
    # Parseamos el fichero JSON extrayendo el link a la web
    urldatos = json_data['datos']
    # Hacemos una nueva llamada a la web temporal que tiene los datos
    respuesta2 = requests.get(urldatos, params=api_key, headers=cabecera_llamada,
verify=False)
    json_data2 = json.loads(respuesta2.text)
```

```
    # Exportación a fichero CSV para su posterior tratamiento en SAS
    # En cada iteración vamos a generar un fichero CSV con el formato 'Idestacion -
Fecha_inicio - Fecha_fin.csv'

    if len(json_data2) > 2:
        with open(estacionreemplazo + '-' + fechainicio + '-' + fechafin + '.csv', 'wb')
as fichero_csv:
            dict_writer = csv.DictWriter(fichero_csv, delimiter=";", fieldnames=campos)
            dict_writer.writeheader()
            dict_writer.writerows(json_data2)
        print("Escribiendo fichero CSV" + " " + estacionreemplazo + '-' + fechainicio +
'-' + fechafin + '.csv')
        print(json_data2)

    # En caso de que un día no haya datos grabados no genero fichero
    else:
        print("Sin datos. Pasando a siguiente iteración")

    # Retardo de tres segundos entre cada ejecución para no superar límite de llamadas
por minuto
    time.sleep(3)
```

ANEXO 2: Códigos SAS

2.1 Importación de ficheros

```
/*##############################################################################################################
##############################################################################################################

                                                CÓDIGO 1
                    DEFINICIÓN DE MACROS PARA LA IMPORTACIÓN Y GENERACIÓN DE LOS FICHEROS BASE
                                    IMPORTACIÓN DE FICHEROS CALIDAD DEL AIRE
                                IMPORTACIÓN DE FICHEROS CLIMATOLÓGICOS AEMET
                                  IMPORTACIÓN DE FICHEROS INTENSIDAD TRÁFICO
                    IMPORTACIÓN DE FICHERO DE COORDENADAS DE LOS PUNTOS DE MEDICIÓN DE TRÁFICO

##############################################################################################################
##############################################################################################################



LA LIBRERÍA Y LAS RUTAS EN LA QUE ESTÉN LOS FICHEROS ES LO ÚNICO QUE HAY QUE ACTUALIZAR EN CASO DE
EJECUTAR EN OTRO ORDENADOR  */

LIBNAME TRABAJO 'C:\Users\Gelu\Desktop\TFM\TRABAJO';

/* FORMATO ANTIGUO FICHEROS CALIDAD AIRE AYUNTAMIENTO */
%LET RUTA_FICHEROS_AIRE_OLD=C:\Users\Gelu\Desktop\TFM\DATOS\AIRE\FORMATO ANTIGUO\;
/* A PARTIR DE OCT17 CAMBIA LIGERAMENTE FORMATO DE LOS FICHEROS */
%LET RUTA_FICHEROS_AIRE_NEW=C:\Users\Gelu\Desktop\TFM\DATOS\AIRE\FORMATO NUEVO - DESDE OCT17\;

/* FORMATO ANTIGUO FICHEROS TRAFICO AYUNTAMIENTO */
%LET RUTA_FICHEROS_TRAFICO_F1=C:\Users\Gelu\Desktop\TFM\DATOS\TRAFICO\FORMATO1\;
/* A PARTIR DE SEP15 CAMBIA LIGERAMENTE FORMATO DE LOS FICHEROS, ELIMINANDO EL CAMPO TIPO Y SE
MODIFICA EL FORMATO DEL CAMPO TIPO_ELEM */
%LET RUTA_FICHEROS_TRAFICO_F2=C:\Users\Gelu\Desktop\TFM\DATOS\TRAFICO\FORMATO2\;
/* A PARTIR DE OCT17 CAMBIA LIGERAMENTE FORMATO DE LOS FICHEROS, ELIMINANDO EL CAMPO IDENTIF Y
MODIFICANDO EL NOMBRE DEL ID (IDELEM-ID)*/
%LET RUTA_FICHEROS_TRAFICO_F3=C:\Users\Gelu\Desktop\TFM\DATOS\TRAFICO\FORMATO3\;
/* RUTA DEL FICHERO QUE CONTIENE LAS COORDENADAS DE CADA PUNTO DE MEDICIÓN */
%LET FICHERO_PUNTOSMED_TRAFICO=C:\Users\Gelu\Desktop\TFM\DATOS\PUNTOS MEDIDA
TRAFICO\pmed_ubicacion_06-2018.csv;

%LET RUTA_FICHEROS_AEMET=C:\Users\Gelu\Desktop\TFM\DATOS\AEMET\;



/* ### COMIENZO DE LA RUTINA DE IMPORTACIÓN DE LOS FICHEROS ### */

FILENAME AIRE_OLD PIPE "dir ""&RUTA_FICHEROS_AIRE_OLD*.txt"" /b";
FILENAME AIRE_NEW PIPE "dir ""&RUTA_FICHEROS_AIRE_NEW*.txt"" /b";
FILENAME TRAF_F1 PIPE "dir ""&RUTA_FICHEROS_TRAFICO_F1*.csv"" /b";
FILENAME TRAF_F2 PIPE "dir ""&RUTA_FICHEROS_TRAFICO_F2*.csv"" /b";
FILENAME TRAF_F3 PIPE "dir ""&RUTA_FICHEROS_TRAFICO_F3*.csv"" /b";
FILENAME AEMET PIPE "dir ""&RUTA_FICHEROS_AEMET*.csv"" /b";




/*##############################################################################################################
##############################################################################################################

                                        FICHEROS CALIDAD AIRE FORMATO ANTIGUO
EL AYUNTAMIENTO HA IDO VARIANDO EL FORMATO DE LOS FICHEROS DE CALIDAD DEL AIRE. LOS FICHEROS GENERADOS
A PARTIR DE OCTUBRE 2017 CAMBIARON LIGERAMENTE DE FORMATO: DELIMITADOS POR COMA Y AÑO CON 4 DIGITOS EN
VEZ DE DOS. POR LA EXCESIVA EXTENSIÓN SE ADJUNTA A ESTE TFM ÚNICAMENTE EL MÉTODO DE IMPORTACIÓN DE LOS
FICHEROS DE UNO DE ESTOS FORMATOS YA QUE LA IMPORTACIÓN DE LOS OTROS ES SIMILAR PERO INCLUYENDO LOS
PEQUEÑOS CAMBIOS COMENTADOS

##############################################################################################################
##############################################################################################################

CREO UNA TABLA AUXILIAR QUE CONTIENE UN LISTADO CON TODOS LOS FICHEROS DEL DIRECTORIO */

DATA LISTA_FICHEROS_AIRE_OLD;
 LENGTH NOMBRE_FICHERO $35;
 INFILE AIRE_OLD TRUNCOVER;
 INPUT NOMBRE_FICHERO $35.;
 CALL SYMPUT ('NUMERO_FICHEROS',_N_); /* GUARDO EN LA VARIABLE NUMERO_FICHEROS EL NÚMERO DE
OBSERVACIONES PARA UTILIZAR DESPÚES COMO EL RANGO SUPERIOR DEL BUCLE */
RUN;


/* DEFINO LA MACRO IMPORTA_DATOS_AIRE_FORMATO1 CON LA QUE VOY A IMPORTAR TODOS LOS FICHEROS DE
CALIDAD DEL AIRE DESCARGADOS DE LA WEB */
```

```
%MACRO IMPORTA_DATOS_AIRE_OLD;
%DO J=1 %TO %EVAL(&NUMERO_FICHEROS.); /* VARIABLE QUE HE DEFINIDO EN EL PASO ANTERIOR. HABRÁ TANTAS
ITERACIONES COMO FICHEROS HAYA PARA IMPORTAR */
DATA _NULL_;
SET LISTA_FICHEROS_AIRE_OLD;
IF _N_=&J;
CALL SYMPUT ('FICHEROENTRADA',NOMBRE_FICHERO);
%PUT &FICHEROENTRADA;
RUN;

data WORK.FICHERO_AIRE_OLD&J;

INFILE "&RUTA_FICHEROS_AIRE_OLD\&FICHEROENTRADA" LRECL=164;
INPUT
@1 COD_ESTACION $8.
@9 COD_PARAMETROS $2.
@11 COD_TECNICA $2.
@13 COD_PERIODO $2.
(…) SE OMITE PARTE POR EXCESIVA EXTENSIÓN (…)
@159 HORA23 5.
@164 VALHORA23 $1.
;
RUN;

%end;

%MEND IMPORTA_DATOS_AIRE_OLD;

%IMPORTA_DATOS_AIRE_OLD;


/* JUNTO TODOS LOS FICHEROS IMPORTADOS EN UNO ÚNICO */

PROC SQL NOPRINT;
SELECT COUNT(*) INTO :CONTEO
FROM LISTA_FICHEROS_AIRE_OLD;
QUIT;

DATA _NULL_;
CALL SYMPUT("FICHEROFINAL",CAT('FICHERO_AIRE_OLD',&CONTEO));
RUN;

DATA DATOS_TOTAL_AIRE_OLD;
SET FICHERO_AIRE_OLD1-&FICHEROFINAL;
RUN;

/* ELIMINAMOS FICHEROS TEMPORALES */

PROC DATASETS LIB=WORK;
DELETE FICHERO_AIRE_OLD1-&FICHEROFINAL;
QUIT;




/*##################################################################################################
##################################################################################################

                                        FICHEROS AEMET

##################################################################################################
##################################################################################################

CREO UNA TABLA AUXILIAR QUE CONTIENE UN LISTADO CON TODOS LOS FICHEROS DEL DIRECTORIO */

DATA LISTA_FICHEROS_AEMET;
 LENGTH NOMBRE_FICHERO $35;
 INFILE AEMET TRUNCOVER;
 INPUT NOMBRE_FICHERO $35.;
 CALL SYMPUT ('NUMERO_FICHEROS',_N_); /* GUARDO EN LA VARIABLE NUMERO_FICHEROS EL NÚMERO DE
OBSERVACIONES PARA UTILIZAR DESPÚES COMO LÍMITE SUPERIOR DEL BUCLE */
RUN;

/* DEFINO LA MACRO IMPORTA_DATOS_AEMET CON LA QUE VOY A IMPORTAR TODOS LOS FICHEROS CSV DE AEMET
DESCARGADOS EN PYTHON CON EL CRAWLER */

%MACRO IMPORTA_DATOS_AEMET;
%DO J=1 %TO %EVAL(&NUMERO_FICHEROS.); /* VARIABLE QUE HE DEFINIDO EN EL PASO ANTERIOR. HABRÁ TANTAS
ITERACIONES COMO FICHEROS HAYA PARA IMPORTAR */
DATA _NULL_;
 SET LISTA_FICHEROS_AEMET;
 IF _N_=&J;
 CALL SYMPUT ('FICHEROENTRADA',NOMBRE_FICHERO);
 %PUT &FICHEROENTRADA;
RUN;
```

```
data WORK.FICHERO_AEMET&J;
INFILE "&RUTA_FICHEROS_AEMET\&FICHEROENTRADA" dlm=';' dsd truncover FIRSTOBS=2;
INPUT
fecha            :       $10.
indicativo       :       $5.
nombre           :       $40.
provincia        :       $20.
altitud          :       $40.
tmed             :       $40.
(…) SE OMITE PARTE POR EXCESIVA EXTENSIÓN (…)
presMin          :       $40.
horaPresMin      :       $40.;
RUN;

%end;

%MEND IMPORTA_DATOS_AEMET;

%IMPORTA_DATOS_AEMET;

/* JUNTO TODOS LOS FICHEROS IMPORTADOS EN UNO ÚNICO */

PROC SQL NOPRINT;
SELECT COUNT(*) INTO :CONTEO
FROM LISTA_FICHEROS_AEMET;
QUIT;

DATA _NULL_;
CALL SYMPUT("FICHEROFINAL",CAT('FICHERO_AEMET',&CONTEO));
RUN;

DATA DATOS_TOTAL_AEMET;
SET FICHERO_AEMET1-&FICHEROFINAL;
RUN;

/* NOS QUEDAMOS ÚNICAMENTE CON EL RANGO DE 07/2015 A 06/2018 (PERIODO DE 3 AÑOS) */

DATA TRABAJO.DATOS_TOTAL_AEMET;
SET FICHERO_AEMET1-&FICHEROFINAL;;
WHERE SUBSTR(FECHA,1,7)>='2015-07' AND SUBSTR(FECHA,1,7)<='2018-06';
RUN;

/* ORDENO */

PROC SORT DATA=TRABAJO.DATOS_TOTAL_AEMET;
BY FECHA INDICATIVO;
RUN;

/* ELIMINAMOS FICHEROS TEMPORALES */

PROC DATASETS LIB=WORK;
DELETE FICHERO_AEMET1-&FICHEROFINAL;
QUIT;


/*##########################################################################################################
##########################################################################################################

                                        FICHEROS TRÁFICO – FORMATO 1
 EL AYUNTAMIENTO HA IDO VARIANDO EL FORMATO DE LOS FICHEROS DE TRÁFICO. EN TOTAL ME HE ENCONTRADO CON
    TRES TIPOS DE FORMATO. POR LA EXCESIVA EXTENSIÓN SE ADJUNTA A ESTE TFM ÚNICAMENTE EL MÉTODO DE
  IMPORTACIÓN DE LOS FICHEROS DE UNO DE ESTOS FORMATOS YA QUE EL RESTO ES SIMILAR PERO ADAPTADO A LA
                                        ESTRUCTURA CON LOS CAMBIOS

##########################################################################################################
##########################################################################################################*/

DATA LISTA_FICHEROS_TRAFICO_F1;
 LENGTH NOMBRE_FICHERO $35;
 INFILE TRAF_F1 TRUNCOVER;
 INPUT NOMBRE_FICHERO $35.;
 CALL SYMPUT ('NUMERO_FICHEROS',_N_); /* GUARDO EN LA VARIABLE NUMERO_FICHEROS EL NÚMERO DE
OBSERVACIONES PARA UTILIZAR DESPÚES COMO EL RANGO SUPERIOR DEL BUCLE */
RUN;

/* DEFINO LA MACRO IMPORTA_DATOS_TRAFICO_F1 CON LA QUE VOY A IMPORTAR TODOS LOS FICHEROS DE TRAFICO
(CON EL TIPO DE FORMATO 1) */

%MACRO IMPORTA_DATOS_TRAFICO_F1;
%DO J=1 %TO %EVAL(&NUMERO_FICHEROS.); /* VARIABLE QUE HE DEFINIDO EN EL PASO ANTERIOR. HABRÁ TANTAS
ITERACIONES COMO FICHEROS HAYA PARA IMPORTAR */
DATA _NULL_;
 SET LISTA_FICHEROS_TRAFICO_F1;
 IF _N_=&J;
 CALL SYMPUT ('FICHEROENTRADA',NOMBRE_FICHERO);
 %PUT &FICHEROENTRADA;
RUN;
```

```
DATA FICHERO_TRAFICO_F1_&J;
INFILE "&RUTA_FICHEROS_TRAFICO_F1&FICHEROENTRADA" DELIMITER = ';,' TRUNCOVER DSD LRECL=32767
FIRSTOBS=2 ;
INFORMAT IDELEM $10.;
INFORMAT FECHA $21. ;
INFORMAT IDENTIF $15. ;
(…) SE OMITE PARTE POR EXCESIVA EXTENSIÓN (…)
FORMAT IDELEM $10. ;
FORMAT FECHA $21. ;
FORMAT IDENTIF $15. ;
(…) SE OMITE PARTE POR EXCESIVA EXTENSIÓN (…)
INPUT
IDELEM $
FECHA $
(…) SE OMITE PARTE POR EXCESIVA EXTENSIÓN (…)
;
RUN;

%END;

%MEND IMPORTA_DATOS_TRAFICO_F1;

%IMPORTA_DATOS_TRAFICO_F1;

/* JUNTO TODOS LOS FICHEROS IMPORTADOS EN UNO ÚNICO */

PROC SQL NOPRINT;
SELECT COUNT(*) INTO :CONTEO
FROM LISTA_FICHEROS_TRAFICO_F1;
QUIT;
DATA _NULL_;
CALL SYMPUT("FICHEROFINAL",CAT('FICHERO_TRAFICO_F1_',&CONTEO));
RUN;

DATA DATOS_TOTAL_TRAFICO_F1;
SET FICHERO_TRAFICO_F1_1-&FICHEROFINAL;
DROP TIPO; /* SE ELIMINA EL CAMPO TIPO YA QUE LOS FICHEROS A PARTIR DE SEP15 NO INCLUYEN DICHO DICHO
CAMPO */
RUN;

/* ELIMINAMOS FICHEROS TEMPORALES */

PROC DATASETS LIB=WORK;
DELETE FICHERO_TRAFICO_F1_1-&FICHEROFINAL;
QUIT;

/* ORDENACION DEL FICHERO DE TRÁFICO */

PROC SORT DATA=DATOS_TOTAL_TRAFICO_F1;
BY FECHA IDELEM;
RUN;


/*##############################################################################################################
##############################################################################################################

        FICHERO COORDENADAS TRÁFICO: FICHERO ÚNICO QUE CONTIENE LAS COORDENADAS DE CADA PUNTO

##############################################################################################################
##############################################################################################################*/

DATA TRABAJO.PUNTOS_MEDICION_TRAFICO;
INFILE "&FICHERO_PUNTOSMED_TRAFICO" DELIMITER = ';' TRUNCOVER DSD LRECL=32767 FIRSTOBS=2;
INFORMAT TIPO_ELEM $30. ;
INFORMAT IDELEM $10. ;
INFORMAT COD_CENT $15. ;
INFORMAT NOMBRE $100. ;
INFORMAT X COMMAX. ;
INFORMAT Y COMMAX. ;

FORMAT TIPO_ELEM $30. ;
FORMAT IDELEM $11. ;
FORMAT COD_CENT $15. ;
FORMAT NOMBRE $100. ;
FORMAT X BEST16. ;
FORMAT Y BEST16. ;

INPUT
TIPO_ELEM $
IDELEM $
COD_CENT $
NOMBRE $
X
Y
;
RUN;
```

2.2 Tratamiento ficheros calidad aire

```
/*##################################################################################################
####################################################################################################

                                                CÓDIGO 2
                                EXPLORACIÓN Y TRATAMIENTO DE DATOS DE CALIDAD DEL AIRE
                                        SUTITUCIÓN DE MEDICIONES ERRÓNEAS
                            AGRUPACIÓN EN LAS 5 ZONAS DEFINIDAS POR EL PROTOCOLO DE AYUNTAMIENTO DE MADRID
                                CREACIÓN DE CENTROIDES INICIALES COMO PASO PREVIO AL CÁLCULO DE CLUSTERS
                            TRANSFORMACIÓN DE DATOS HORARIOS A DATOS DIARIOS (MÁXIMO NO2 DIARIO POR ESTACIÓN)

####################################################################################################
####################################################################################################


EXISTEN ALGUNAS MEDICIONES NO VÁLIDAS (MARCA VALIDEZ = 'N')
VAMOS A SUSTITUIR ESTOS VALORES POR EL ÚLTIMO VALOR DE MEDICIÓN VÁLIDO CONOCIDO PARA LA ESTACIÓN */

DATA TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST;
SET TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL;
BY COD_ESTACION;
RETAIN VALOR_AJUSTADO;
IF MARCA^='N' THEN VALOR_AJUSTADO=VALOR;
OUTPUT;
IF LAST.COD_ESTACION THEN VALOR_AJUSTADO = .;
RUN;

/* EL AYUNTAMIENTO DE MADRID, A EFECTOS DE APLICACIÓN DEL PROTOCOLO DE CONTAMINACIÓN, HA DEFINIDO
CINCO ZONAS TENIENDO EN CUENTA:

- LA DISTRIBUCIÓN DE LA POBLACIÓN
- LA TIPOLOGÍA Y DISTRIBUCIÓN DE ESTACIONES DEL SISTEMA DE VIGILANCIA DE LA CALIDAD DEL AIRE
- EL VIARIO DE TRÁFICO, PARA FACILITAR LA IMPLANTACIÓN DE POSIBLES ACTUACIONES DE RESTRICCIÓN
DEL MISMO

SE TRANSCRIBE DICHAS ZONAS A NUESTRO CÓDIGO

ASIMISMO, PUESTO QUE HAY MUCHOS PUNTOS DE MEDICIÓN DE TRÁFICO LOS NECESITAMOS AGRUPAR, LO CUAL LO
HAREMOS MEDIANTE CLUSTERS

COMO PASO PREVIO, HEMOS CALCULADO EN PYTHON EL CENTROIDE DE LAS 5 ZONAS DE MEDICIÓN DE CALIDAD DEL
AIRE, LOS CUALES SE INCORPORAN AL FICHERO */

DATA TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST;

SET TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST;

FORMAT COORDENADA_X BEST20.;
FORMAT COORDENADA_Y BEST20.;

IF COD_ESTACION IN
('28079004','28079008','28079011','28079035','28079038','28079039','28079047','28079048','28079049','
28079050') THEN DO;
        ZONA='INTERIOR M30';
        IDZONA='1';
        /* COORDENADAS CENTROIDE DE LA ZONA */
        COORDENADA_X=441122.394;
        COORDENADA_Y=4476357.601000001;
        END;

IF COD_ESTACION IN ('28079036','28079040','28079054') THEN DO;
        ZONA='SURESTE';
        IDZONA='2';
        /* COORDENADAS CENTROIDE DE LA ZONA */
        COORDENADA_X=445999.56666666665;
        COORDENADA_Y=4471197.989999999;
        END;

IF COD_ESTACION IN ('28079016','28079027','28079055','28079057','28079059','28079060') THEN DO;
        ZONA='NORESTE';
        IDZONA='3';
        /* COORDENADAS CENTROIDE DE LA ZONA */
        COORDENADA_X=446894.4916666667;
        COORDENADA_Y=4479697.062857143;
        END;

IF COD_ESTACION IN ('28079024','28079058') THEN DO;
        ZONA='NOROESTE';
        IDZONA='4';
        /* COORDENADAS CENTROIDE DE LA ZONA */
        COORDENADA_X=435490.03500000003;
        COORDENADA_Y=4480059.24;
        END;
```

```
IF COD_ESTACION IN ('28079017', '28079018', '28079056') THEN DO;
        ZONA='SUROESTE';
        IDZONA='5';
        /* COORDENADAS CENTROIDE DE LA ZONA */
        COORDENADA_X=438772.18333333335;
        COORDENADA_Y=4469688.79;
        END;

/* DISTANCIA A CENTROIDE DE LA M30 */
DIST_CENT_X_M30=ABS(COORDENADA_X-441122.394);
DIST_CENT_Y_M30=ABS(COORDENADA_Y-4476357.601000001);

/* DISTANCIA A CENTROIDE ZONA NORESTE */
DIST_CENT_X_NORESTE=ABS(COORDENADA_X-446894.4916666667);
DIST_CENT_Y_NORESTE=ABS(COORDENADA_Y-4479697.062857143);

/* DISTANCIA A CENTROIDE ZONA NOROESTE */
DIST_CENT_X_NOROESTE=ABS(COORDENADA_X-435490.03500000003);
DIST_CENT_Y_NOROESTE=ABS(COORDENADA_Y-4480059.24);

/* DISTANCIA A CENTROIDE ZONA SURESTE  */
DIST_CENT_X_SURESTE=ABS(COORDENADA_X-445999.56666666665);
DIST_CENT_Y_SURESTE=ABS(COORDENADA_Y-4471197.989999999);

/* DISTANCIA A CENTROIDE ZONA SUROESTE */
DIST_CENT_X_SUROESTE=ABS(COORDENADA_X-438772.18333333335);
DIST_CENT_Y_SUROESTE=ABS(COORDENADA_Y-4469688.79);

RUN;


/* CREACIÓN DEL FICHERO DE CENTROIDE INICIALES */

PROC SQL;
CREATE TABLE TRABAJO.CENTROIDES_INICIALES AS
SELECT DISTINCT
ZONA, COORDENADA_X, COORDENADA_Y,
DIST_CENT_X_M30, DIST_CENT_Y_M30,
DIST_CENT_X_NORESTE, DIST_CENT_Y_NORESTE,
DIST_CENT_X_NOROESTE, DIST_CENT_Y_NOROESTE,
DIST_CENT_X_SURESTE, DIST_CENT_Y_SURESTE,
DIST_CENT_X_SUROESTE, DIST_CENT_Y_SUROESTE,
CASE WHEN ZONA='INTERIOR M30' THEN 1
WHEN ZONA='SURESTE' THEN 2
WHEN ZONA='NORESTE' THEN 3
WHEN ZONA='NOROESTE' THEN 4
WHEN ZONA='SUROESTE' THEN 5
ELSE . END AS CLUSTER
FROM TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST;
QUIT;

PROC SORT DATA=TRABAJO.CENTROIDES_INICIALES;
BY CLUSTER;
RUN;

DATA TRABAJO.CENTROIDES_INICIALES;
RETAIN ZONA CLUSTER;
SET TRABAJO.CENTROIDES_INICIALES;
RUN;

/* A CONTINUACIÓN VAMOS A PASAR LOS DATOS HORARIOS DE NO2 A DIARIOS, QUEDÁNDONOS CON EL MÁXIMO DIARIO
MARCADO POR LA ESTACIÓN */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_NO2_DIARIO AS
SELECT DISTINCT COD_ESTACION, ESTACION, FECHA, IDZONA, ZONA, CAT(IDZONA,'_',PUT(FECHA,YYMMDD10.))AS
AUX_ZONA_DIA, MAX(VALOR_AJUSTADO) AS MAX_NO2
FROM TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST
GROUP BY 1,2,3,4,5;
QUIT;

PROC SORT DATA=TRABAJO.DATOS_TOTAL_NO2_DIARIO;
BY AUX_ZONA_DIA DESCENDING MAX_NO2;
RUN;

/* UNA VEZ OBTENIDO LOS MÁXIMOS DIARIOS POR CADA ESTACIÓN, SELECCIONAMOS EL VALOR MÁXIMO PARA CADA
ZONA */

DATA TRABAJO.NO2_MAXNO2_ZONA;
SET TRABAJO.DATOS_TOTAL_NO2_DIARIO;
IF FIRST.AUX_ZONA_DIA;
BY AUX_ZONA_DIA;
RUN;
```

2.3 Clusterización de puntos de tráfico

```
/*##########################################################################################################
##########################################################################################################

                                                CÓDIGO 3
  CLUSTERIZACIÓN DE LOS PTOS DE MEDICIÓN EN BASE A LA DISTANCIA EN COORD A LOS CENTROIDES INICIALES
                                REPRESENTACIÓN GRÁFICA DE LOS CLUSTER

##########################################################################################################
##########################################################################################################

CRUCE PARA AÑADIR LAS COORDENADAS A LOS PUNTOS DE MEDICIÓN DE TRÁFICO */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_TRAFICO2 (COMPRESS=YES) AS
SELECT DISTINCT A.*, B.X AS COORDENADA_X, B.Y AS COORDENADA_Y
FROM TRABAJO.DATOS_TOTAL_TRAFICO AS A
LEFT JOIN TRABAJO.PUNTOS_MEDICION_TRAFICO AS B
ON A.IDELEM=B.IDELEM
/* AQUELLOS PUNTOS DE MEDICIÓN DE LOS QUE NO TENGA SUS COORDENADAS LOS DESECHAMOS */
WHERE B.X^=.; QUIT;

/* PARA LA TABLA QUE CONTIENE LOS PUNTOS DE MEDICIÓN DE INTENSIDAD DE TRÁFICO, CALCULAMOS LA
DISTANCIA DEL PUNTO A CADA UNO DE LOS 5 CENTROIDES, YA QUE
LA AGRUPACIÓN DE CLUSTERS LA HAREMOS A PARTIR DE DICHAS DISTANCIAS */

DATA TRABAJO.DATOS_TOTAL_TRAFICO2;
SET TRABAJO.DATOS_TOTAL_TRAFICO2;

/* DISTANCIA DEL PUNTO DE MEDICIÓN DE TRÁFICO A CENTROIDE DE LA M30 */
DIST_CENT_X_M30=ABS(COORDENADA_X-441122.394);
DIST_CENT_Y_M30=ABS(COORDENADA_Y-4476357.601000001);

/* DISTANCIA DEL PUNTO DE MEDICIÓN DE TRÁFICO A CENTROIDE ZONA NORESTE */
DIST_CENT_X_NORESTE=ABS(COORDENADA_X-446894.4916666667);
DIST_CENT_Y_NORESTE=ABS(COORDENADA_Y-4479697.062857143);

/* DISTANCIA DEL PUNTO DE MEDICIÓN DE TRÁFICO A CENTROIDE ZONA NOROESTE */
DIST_CENT_X_NOROESTE=ABS(COORDENADA_X-435490.03500000003);
DIST_CENT_Y_NOROESTE=ABS(COORDENADA_Y-4480059.24);

/* DISTANCIA DEL PUNTO DE MEDICIÓN DE TRÁFICO A CENTROIDE ZONA SURESTE */
DIST_CENT_X_SURESTE=ABS(COORDENADA_X-445999.56666666665);
DIST_CENT_Y_SURESTE=ABS(COORDENADA_Y-4471197.989999999);

/* DISTANCIA DEL PUNTO DE MEDICIÓN DE TRÁFICO A CENTROIDE ZONA SUROESTE */
DIST_CENT_X_SUROESTE=ABS(COORDENADA_X-438772.18333333335);
DIST_CENT_Y_SUROESTE=ABS(COORDENADA_Y-4469688.79);
RUN;

/* COMO PASO PREVIO AL CÁLCULO DE LOS CLUSTER CREAMOS UNA TABLA CON LOS DISTINTOS PUNTOS DE MEDICIÓN
Y SUS COORDENADAS */

PROC SQL;
CREATE TABLE TRABAJO.PUNTOS_MEDICION_TRAFICO AS
SELECT DISTINCT IDELEM, COORDENADA_X, COORDENADA_Y, DIST_CENT_X_M30, DIST_CENT_Y_M30,
DIST_CENT_X_NORESTE, DIST_CENT_Y_NORESTE, DIST_CENT_X_NOROESTE, DIST_CENT_Y_NOROESTE,
DIST_CENT_X_SURESTE, DIST_CENT_Y_SURESTE, DIST_CENT_X_SUROESTE, DIST_CENT_Y_SUROESTE
FROM TRABAJO.DATOS_TOTAL_TRAFICO2;
QUIT;

/* CLUSTER NO JERARQUICO - AGRUPAMOS EN 5 CLUSTERS UTILIZANDO LOS CENTROIDES INICIALES CALCULADOS */

PROC FASTCLUS DATA=TRABAJO.PUNTOS_MEDICION_TRAFICO OUT=TRABAJO.PUNTOS_MEDICION_TRAFICO_CLUSTER
SEED=TRABAJO.CENTROIDES_INICIALES LIST DISTANCE MAXCLUSTERS=5 REPLACE=NONE RADIUS=100 MAXITER=1
OUTSEED=TRABAJO.CENTROIDES_FINALES ;
VAR COORDENADA_X COORDENADA_Y DIST_CENT_X_M30 DIST_CENT_Y_M30 DIST_CENT_X_NORESTE DIST_CENT_Y_NORESTE
DIST_CENT_X_NOROESTE DIST_CENT_Y_NOROESTE DIST_CENT_X_SURESTE DIST_CENT_Y_SURESTE
DIST_CENT_X_SUROESTE DIST_CENT_Y_SUROESTE;
RUN;

PROC SORT;BY CLUSTER; RUN; PROC PRINT; BY CLUSTER; RUN;
PROC SGPLOT DATA=TRABAJO.PUNTOS_MEDICION_TRAFICO_CLUSTER;
SCATTER X=COORDENADA_X Y=COORDENADA_Y / GROUP=CLUSTER;
RUN;

/* AÑADIMOS CLUSTER CALCULADO A LOS DATOS ORIGINALES DE TRÁFICO */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_TRAFICO3 (COMPRESS=YES) AS
SELECT DISTINCT A.*, B.CLUSTER, B.DISTANCE AS DISTANCIA_SEMILLA
FROM TRABAJO.DATOS_TOTAL_TRAFICO2 AS A
LEFT JOIN TRABAJO.PUNTOS_MEDICION_TRAFICO_CLUSTER AS B
ON A.IDELEM=B.IDELEM; QUIT;
```

2.4 Exploración y tratamiento de los datos de tráfico

```
/*##############################################################################################################
##############################################################################################################

                                                CÓDIGO 4
                                    EXPLORACIÓN Y TRATAMIENTO DE LOS DATOS DE TRÁFICO
                                            GENERACIÓN DE VARIABLES NUEVAS

##############################################################################################################
##############################################################################################################

DEPURACIÓN DE DATOS ERRÓNEOS */

PROC SQL;
CREATE TABLE DATOS_TRAFICO_ERRONEOS AS
SELECT DISTINCT SUBSTR(FECHA,1,10) AS FECHA, ERROR, COUNT(*) AS CONTEO
FROM TRABAJO.DATOS_TOTAL_TRAFICO3
GROUP BY 1, 2;
QUIT;

/* OBSERVAMOS QUE, PRINCIPALMENTE EN 2015, HAY ALGUNAS MEDICIONES MARCADAS COMO ERRÓNEAS. EL VOLUMEN
DE ERRÓNEOS ES MUY PEQUEÑO COMPARADO CON EL DE MEDICIONES VÁLIDAS Y TAMPOCO HAY NINGÚN DÍA EN EL QUE
TODAS LAS MEDICIONES HAYAN SIDO ERRÓNEAS, POR LO QUE DESECHAMOS DICHAS OBSERVACIONES. */

DATA TRABAJO.DATOS_TOTAL_TRAFICO4 (COMPRESS=YES);
SET TRABAJO.DATOS_TOTAL_TRAFICO3;
WHERE ERROR='N';

/* ASÍMISMO VAMOS A GENERAR VARIABLES DIARIAS A PARTIR DE LOS DATOS DE MEDICIÓN DE TRÁFICO
COMO PASO PREVIO CREO UNA VARIABLE AUXILIAR QUE ME VA A SERVIR POSTERIORMENTE PARA EL TRATAMIENTO DE
TABLAS (AGRUPACIONES, ETC.)*/
AUX_CLUSTER_DIA=CAT(CLUSTER,'_',SUBSTR(FECHA,1,10));
RUN;

/* REVISANDO LOS VALORES DE INTENSIDAD VEMOS QUE HAY VALORES EXTRAÑAMENTE ALTOS QUE PARECEN SER
INCORRECTOS PARA CADA PUNTO DE MEDICIÓN (DE LOS 4.057 QUE HAY) MANTENDREMOS ÚNICAMENTE AQUELLAS
OBSERVACIONES CUYO VALOR DE INTENSIDAD DE TRÁFICO ESTÉ DENTRO DEL PERCENTIL 95. ES DECIR, DESECHAMOS
EL 5% DE LOS VALORES SUPERIORES DE INTENSIDAD PARA CADA PUNTO DE MEDICIÓN */

PROC RANK DATA=TRABAJO.DATOS_TOTAL_TRAFICO4 OUT=TRABAJO.DATOS_TOTAL_TRAFICO5
(WHERE=(R_INTENSIDAD<19)) GROUPS=20 TIES=LOW; _
VAR INTENSIDAD;
RANKS R_INTENSIDAD;
BY IDELEM;
RUN;

/* TAMBIÉN SE ELIMINAN AQUELLAS OBSERVACIONES EN LAS QUE EL VALOR DE INTENSIDAD SEA 0 O IDELEM SIN
INFORMAR */

DATA TRABAJO.DATOS_TOTAL_TRAFICO6 (COMPRESS=YES);
SET TRABAJO.DATOS_TOTAL_TRAFICO5;
WHERE INTENSIDAD^=0 AND IDELEM^='';
/* MANTENEMOS ÚNICAMENTE CAMPOS NECESARIOS */
KEEP IDELEM FECHA IDENTIF TIPO_ELEM INTENSIDAD OCUPACION CARGA VMED ERROR PERIODO_INTEGRACION CLUSTER
AUX_CLUSTER_DIA;
RUN;

PROC SORT DATA=TRABAJO.DATOS_TOTAL_TRAFICO6;
BY FECHA IDELEM;
RUN;

/* REPRESENTACIÓN GRÁFICA DE VARIABLES RELATIVAS A VELOCIDAD MEDIA, OCUPACIÓN Y CARGA */

PROC GCHART DATA=TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX;
HBAR VMED / MIDPOINTS=0 TO 130 BY 10;
RUN;

PROC GCHART DATA=TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX;
HBAR OCUPACION / MIDPOINTS=0 TO 110 BY 10;
RUN;

PROC GCHART DATA=TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX;
HBAR CARGA / MIDPOINTS=0 TO 110 BY 10;
RUN;

/* NOS LLEVA A DESCARTAR ESTAS VARIABLES, ÚNICAMENTE RECUPERAREMOS DATOS DE INTENSIDAD */
```

```
/*##############################################################################################################
##############################################################################################################

                                        GENERACIÓN DE NUEVAS VARIABLES TRÁFICO

##############################################################################################################
##############################################################################################################

PUESTO QUE LA PERIODICIDAD DE MEDICIÓN ES CADA QUINCE MINUTOS NOS INTERESA TRATAR ESTOS DATOS DE CARA
A QUEDARNOS CON VARIABLES DE PERIODICIDAD DIARIA, IMPUTABLES A UNA ÚNICA ZONA PARA QUE PUEDAN SERVIR
DE INPUT EN NUESTRA MODELIZACIÓN. PARA ELLO SE GENERAN LAS SIGUIENTES VARIABLES: */

/* VALOR MÁXIMO DE INTENSIDAD DIARIA DE LA ZONA: SELECCIONAMOS AQUELLOS PUNTOS DE MEDICIÓN QUE HAN
MARCADO EL MÁXIMO DE INTENSIDAD DE TRÁFICO DE LA ZONA. UNA VEZ SELECCIONADO EL PUNTO DE MEDICIÓN,
OBTENDREMOS EL VALOR MÁXIMO DE INTENSIDAD DE TRÁFICO DIARIO ASOCIADO A DICHO PUNTO */
PROC SQL;
CREATE TABLE TRABAJO.MAX_INTENSIDAD_POR_ESTACION AS
SELECT DISTINCT IDELEM, SUBSTR(FECHA,1,10) AS DIA, CLUSTER, MAX(INTENSIDAD) AS MAX_INTENSIDAD,
AUX_CLUSTER_DIA
FROM TRABAJO.DATOS_TOTAL_TRAFICO6
GROUP BY 1, 2;
QUIT;

PROC SORT DATA=TRABAJO.MAX_INTENSIDAD_POR_ESTACION OUT=TRABAJO.MAX_INTENSIDAD_POR_ZONA_AUX;
BY AUX_CLUSTER_DIA DESCENDING MAX_INTENSIDAD;
RUN;

DATA TRABAJO.MAX_INTENSIDAD_POR_ZONA;
SET TRABAJO.MAX_INTENSIDAD_POR_ZONA_AUX;
IF FIRST.AUX_CLUSTER_DIA;
BY AUX_CLUSTER_DIA;
RUN;

/* SUMA DE LOS VALORES MÁXIMOS POR ZONA: COMO PASO PREVIO A CALCULAR ESTA VARIABLE VAMOS A VER CUÁL
ES EL NÚMERO MÍNIMO DE PUNTOS DE MEDICIÓN CON MEDICIONES EN UN DÍA HACEMOS ESTO PORQUE HAY ZONAS QUE
TIENEN MUCHOS MÁS PUNTOS DE MEDICIÓN QUE OTRAS DE MODO QUE ÚNICAMENTE VAMOS A SUMAR LOS N PRIMEROS*/
PROC SQL;
CREATE TABLE TRABAJO.TRAFICO_CONTEO_PUNTOS_POR_DIA AS
SELECT AUX_CLUSTER_DIA, COUNT(DISTINCT(IDELEM)) AS CONTEO
FROM TRABAJO.DATOS_TOTAL_TRAFICO6
GROUP BY 1
ORDER BY CONTEO;
QUIT;

/* EJECUTAMOS Y VEMOS QUE EL VALOR MÍNIMO SON 33, POR LO QUE OPTAMOS POR DELIMITAR A 30 LOS SUMANDOS
PARA EL CÁLCULO DE DICHA VARIABLE */
PROC SORT DATA=TRABAJO.MAX_INTENSIDAD_POR_ESTACION;
BY AUX_CLUSTER_DIA DESCENDING MAX_INTENSIDAD;
RUN;

/* AÑADIMOS VALOR AUXILIAR CON EL RANKING CALCULADO */
PROC RANK DATA=TRABAJO.MAX_INTENSIDAD_POR_ESTACION OUT=TRABAJO.R_MAX_INTENSIDAD_POR_ESTACION
DESCENDING TIES=HIGH;
/* CON LA OPCION TIES=HIGH SUMO +1 EN CASO DE EMPATE */
BY AUX_CLUSTER_DIA;
VAR MAX_INTENSIDAD;
RANKS R_MAX_INTENSIDAD;
RUN;

/* FILTRAMOS PARA QUEDARNOS CON LOS 30 VALORES MÁXIMOS */
DATA TRABAJO.SUMA_MAX_INTENSIDA_POR_ZONA_AUX;
SET TRABAJO.R_MAX_INTENSIDAD_POR_ESTACION;
WHERE R_MAX_INTENSIDAD<=30;
RUN;

/* UNA VEZ GENERADA LA TABLA AUXILIAR, HACEMOS EL CÁLCULO DE LA VARIABLE */
PROC SQL;
CREATE TABLE TRABAJO.SUMA_MAX_INTENSIDAD_POR_ZONA AS
SELECT AUX_CLUSTER_DIA, SUM(MAX_INTENSIDAD) AS SUM_MAX_INTENSIDAD
FROM TRABAJO.SUMA_MAX_INTENSIDA_POR_ZONA_AUX
GROUP BY 1;
QUIT;

/* MEDIA DE VALORES MÁXIMOS POR ZONA: PARA CADA DÍA Y CADA UNA DE LAS CINCO ZONAS, OBTENEMOS LA MEDIA
DE LOS VALORES MÁXIMOS MARCADOS POR CADA UNA DE LAS ESTACIONES A LAS QUE PERTENECE CADA PUNTO DE
MEDICIÓN */
PROC MEANS DATA=TRABAJO.MAX_INTENSIDAD_POR_ESTACION NOPRINT NWAY;
CLASS AUX_CLUSTER_DIA;
VAR MAX_INTENSIDAD;
OUTPUT OUT=TRABAJO.INTENSIDAD_MAX_MEDIA_POR_ZONA(DROP=_TYPE_) mean=MAX_MEDIA_INTENSIDAD;
RUN;
```

```
/* VALOR MEDIO DIARIO DE INTENSIDAD DIARIA DE LA ZONA */
PROC MEANS DATA=TRABAJO.DATOS_TOTAL_TRAFICO6 NOPRINT NWAY;
CLASS AUX_CLUSTER_DIA;
VAR INTENSIDAD;
OUTPUT OUT=TRABAJO.INTENSIDAD_MEDIA_POR_ZONA(DROP=_TYPE_) mean=MEDIA_INTENSIDAD;
RUN;

/* VALOR MEDIO DIARIO DE INTENSIDAD ASOCIADO AL PUNTO DE CONTROL QUE HA MARCADO MÁXIMO EN LA ZONA.
COMO PASO PREVIO PARA EL CÁLCULO CREAMOS UNA TABLA AUXILIAR MANTENIENDO ÚNICAMENTE LAS MEDICIONES DE
LAS ESTACIONES QUE HAN MARCADO MÁXIMO DE ZONA EN CADA UNO DE LOS DÍAS */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX AS
SELECT DISTINCT A.*
FROM TRABAJO.DATOS_TOTAL_TRAFICO6 AS A
INNER JOIN TRABAJO.MAX_INTENSIDAD_POR_ZONA AS B
ON A.IDELEM=B.IDELEM AND A.AUX_CLUSTER_DIA=B.AUX_CLUSTER_DIA;
QUIT;

/* UNA VEZ GENERADA LA TABLA AUXILIAR PODEMOS CALCULAR EL VALOR MEDIO DIARIO DE INTENSIDAD ASOCIADO
AL PUNTO DE CONTROL QUE HA MARCADO MÁXIMO EN LA ZONA */

PROC MEANS DATA=TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX NOPRINT NWAY;
CLASS AUX_CLUSTER_DIA;
VAR INTENSIDAD;
OUTPUT OUT=TRABAJO.VALOR_MEDIO_INTENS_PTOMAX(DROP=_TYPE_) MEAN=MEDIA_INTENSIDAD;
RUN;

/* VALOR MEDIO DIARIO DE INTENSIDAD ASOCIADO AL PUNTO DE CONTROL QUE HA MARCADO MÁXIMO DIARIO */

PROC MEANS DATA=TRABAJO.DATOS_TRAFICO_PUNTO_MAXINT_AUX NOPRINT NWAY;
CLASS AUX_CLUSTER_DIA;
VAR INTENSIDAD;
OUTPUT OUT=TRABAJO.INTENSIDAD_MEDIA_EST_MAXINTENS (DROP=_TYPE_) mean=MEDIA_INTENSIDAD;
RUN;
```

2.5. Clusterización de estaciones AEMET

```
/*#############################################################################################################
#############################################################################################################

                                                CÓDIGO 5
        CLUSTERIZACIÓN DE LAS ESTACIONES CLIMATOLÓGICAS DE AEMET EN BASE A LA DISTANCIA EN
                                    COORDENADAS A LOS CENTROIDES INICIALES
                                    REPRESENTACIÓN GRÁFICA DE LOS CLUSTER

#############################################################################################################
#############################################################################################################

DEL MISMO MODO QUE HEMOS HECHO CON LOS PUNTOS DE MEDICION DE TRÁFICO VAMOS A ASIGNAR CADA ESTACIÓN DE
AEMET A UNA DE LAS 5 ZONAS DEFINIDAS EN EL PROTOCOLO DE CONTAMINACIÓN DEL AYUNTAMIENTO DE MADRID

AÑADIMOS A LA TABLA ORIGEN LAS COORDENADAS DE LAS 23 ESTACIONES Y LA DISTANCIA A LOS CENTROIDES
CALCULADOS */

DATA TRABAJO.DATOS_TOTAL_AEMET2;

SET TRABAJO.DATOS_TOTAL_AEMET;

FORMAT COORDENADA_X BEST20.;
FORMAT COORDENADA_Y BEST20.;

/* PUERTO DE NAVACERRADA */
IF INDICATIVO='2462' THEN DO;
COORDENADA_X=414407.23; COORDENADA_Y=4509371.75;
END;
/* ARANJUEZ */
IF INDICATIVO='3100B' THEN DO;
COORDENADA_X=453429.74; COORDENADA_Y=4435361.13;
END;
/* BUITRAGO DEL LOZOYA */
IF INDICATIVO='3110C' THEN DO;
COORDENADA_X=446473.17; COORDENADA_Y=4538212;
END;
/* SOMOSIERRA */
IF INDICATIVO='3111D' THEN DO;
COORDENADA_X=451249.07;COORDENADA_Y=4553659.38;
END;
/* MADRID AEROPUERTO */
IF INDICATIVO='3129' THEN DO;
COORDENADA_X=452902.23; COORDENADA_Y=4479702.95;
END;
/* TORREJÓN DE ARDOZ */
IF INDICATIVO='3175' THEN DO;
COORDENADA_X=459397.06; COORDENADA_Y=4478801.53;
END;
```

```
/* COLMENAR VIEJO */
IF INDICATIVO='3191E' THEN DO;
COORDENADA_X=435367.45; COORDENADA_Y=4505304.76;
END;

/* MADRID, CIUDAD UNIVERSITARIA */
IF INDICATIVO='3194U' THEN DO;
COORDENADA_X=438594.29; COORDENADA_Y=4478141.55;
END;
/* MADRID, RETIRO */
IF INDICATIVO='3195' THEN DO;
COORDENADA_X=442470.47; COORDENADA_Y=4473701.34;
END;
/* MADRID, CUATRO VIENTOS */
IF INDICATIVO='3196' THEN DO;
COORDENADA_X=433266.51; COORDENADA_Y=4469738.14;
END;
/* GETAFE */
IF INDICATIVO='3200' THEN DO;
COORDENADA_X=438035.07; COORDENADA_Y=4461741.86;
END;
/* PUERTO ALTO DEL LEÓN */
IF INDICATIVO='3266A' THEN DO;
COORDENADA_X=403534.84; COORDENADA_Y=4506791.27;
END;
/* ROBLEDO DE CHAVELA */
IF INDICATIVO='3338' THEN DO;
COORDENADA_X=395001.93; COORDENADA_Y=4484177.22;
END;

/* DISTANCIA DE LA ESTACIÓN AEMET AL CENTROIDE DE LA M30 */
DIST_CENT_X_M30=ABS(COORDENADA_X-441122.394);
DIST_CENT_Y_M30=ABS(COORDENADA_Y-4476357.601000001);

/* DISTANCIA DE LA ESTACIÓN AEMET AL CENTROIDE ZONA NORESTE */
DIST_CENT_X_NORESTE=ABS(COORDENADA_X-446894.4916666667);
DIST_CENT_Y_NORESTE=ABS(COORDENADA_Y-4479697.062857143);

/* DISTANCIA DE LA ESTACIÓN AEMET AL CENTROIDE ZONA NOROESTE */
DIST_CENT_X_NOROESTE=ABS(COORDENADA_X-435490.03500000003);
DIST_CENT_Y_NOROESTE=ABS(COORDENADA_Y-4480059.24);

/* DISTANCIA DE LA ESTACIÓN AEMET AL CENTROIDE ZONA SURESTE */
DIST_CENT_X_SURESTE=ABS(COORDENADA_X-445999.56666666665);
DIST_CENT_Y_SURESTE=ABS(COORDENADA_Y-4471197.989999999);

/* DISTANCIA DE LA ESTACIÓN AEMET AL CENTROIDE ZONA SUROESTE */
DIST_CENT_X_SUROESTE=ABS(COORDENADA_X-438772.18333333335);
DIST_CENT_Y_SUROESTE=ABS(COORDENADA_Y-4469688.79);
RUN;

/* COMO PASO PREVIO CREAMOS UNA TABLA CON LOS DISTINTOS PUNTOS DE MEDICIÓN Y SUS COORDENADAS */

PROC SQL;
CREATE TABLE TRABAJO.ESTACIONES_AEMET AS
SELECT DISTINCT INDICATIVO, NOMBRE, COORDENADA_X, COORDENADA_Y, DIST_CENT_X_M30, DIST_CENT_Y_M30,
DIST_CENT_X_NORESTE, DIST_CENT_Y_NORESTE, DIST_CENT_X_NOROESTE, DIST_CENT_Y_NOROESTE,
DIST_CENT_X_SURESTE, DIST_CENT_Y_SURESTE, DIST_CENT_X_SUROESTE, DIST_CENT_Y_SUROESTE
FROM TRABAJO.DATOS_TOTAL_AEMET2;
QUIT;

/* CLUSTER NO JERARQUICO - AGRUPAMOS EN 5 CLUSTERS */

PROC FASTCLUS DATA=TRABAJO.ESTACIONES_AEMET OUT=TRABAJO.ESTACIONES_AEMET_CLUSTER
SEED=TRABAJO.CENTROIDES_INICIALES LIST DISTANCE MAXCLUSTERS=5 REPLACE=NONE MAXITER=1
OUTSEED=TRABAJO.CENTROIDES_FINALES_AEMET ;
VAR COORDENADA_X COORDENADA_Y DIST_CENT_X_M30 DIST_CENT_Y_M30 DIST_CENT_X_NORESTE DIST_CENT_Y_NORESTE
DIST_CENT_X_NOROESTE DIST_CENT_Y_NOROESTE DIST_CENT_X_SURESTE DIST_CENT_Y_SURESTE
DIST_CENT_X_SUROESTE DIST_CENT_Y_SUROESTE;
RUN;

PROC SORT;BY CLUSTER; RUN; PROC PRINT; BY CLUSTER; RUN;

PROC SGPLOT DATA=TRABAJO.ESTACIONES_AEMET_CLUSTER;
SCATTER X=COORDENADA_X Y=COORDENADA_Y / GROUP=CLUSTER;
RUN;

/* AÑADIMOS CLUSTER A OBSERVACIONES ORIGINALES DE AEMET */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_AEMET3 AS
SELECT DISTINCT A.*, B.CLUSTER, B.DISTANCE AS DISTANCIA_SEMILLA
FROM TRABAJO.DATOS_TOTAL_AEMET2 AS A
LEFT JOIN TRABAJO.ESTACIONES_AEMET_CLUSTER AS B
ON A.INDICATIVO=B.INDICATIVO;
QUIT;
```

2.6. Tratamiento datos AEMET

```
/*##########################################################################################################
##########################################################################################################

                                                CÓDIGO 6
                        EXPLORACIÓN Y TRATAMIENTO DE LOS DATOS CLIMATOLÓGICOS DE AEMET
                            DETECCIÓN DE VALORES MISSING Y CRITERIOS DE IMPUTACIÓN
                         GENERACIÓN DE NUEVAS VARIABLES PARA UTILIZAR EN MODELIZACIÓN

##########################################################################################################
##########################################################################################################

TRATAMIENTO DE DATOS CLIMATOLOGICOS. PUESTO QUE HAY ZONAS CON VARIAS ESTACIONES, TENEMOS QUE
ESTABLECER CRITERIOS PARA IMPUTAR DATOS A UNA ZONA EN VEZ DE A UNA ESTACION, POR EJEMPLO, QUEDARNOS
CON LA MEDIA DIARIA DE CADA VARIABLE ASOCIADA A LA ZONA

COMO PASO PREVIO TRANSFORMO LAS VARIABLES A NÚMERICAS

TAMBIEN VOY A CATEGORIZAR LAS VARIABLES RELATIVAS A HORAS: HORATMAX HORATMIN HORARACHA HORAPMAX
HORAPMIN */

DATA TRABAJO.DATOS_TOTAL_AEMET4;

SET TRABAJO.DATOS_TOTAL_AEMET3;

ALTITUD2=INPUT(ALTITUD,32.);
TMED2=INPUT(TMED,COMMA10.1);
PREC2=INPUT(PREC,COMMA10.1);
TMIN2=INPUT(TMIN,COMMA10.1);
TMAX2=INPUT(TMAX,COMMA10.1);
DIR2=INPUT(DIR,COMMA10.);
VELMEDIA2=INPUT(VELMEDIA,COMMA10.1);
RACHA2=INPUT(RACHA,COMMA10.1);
SOL2=INPUT(SOL,COMMA10.1);
PRESMAX2=INPUT(PRESMAX,COMMA10.1);
PRESMIN2=INPUT(PRESMIN,COMMA10.1);

/* CATEGORIZO VARIABLES RELATIVAS A HORAS MAX/MIN EN 3 VALORES */

/* HORA TEMPERATURA MINIMA */
IF SUBSTR(HORATMIN,1,2)>='00' AND SUBSTR(HORATMIN,1,2)<'08' THEN HORATMIN_CAT='1';
IF SUBSTR(HORATMIN,1,2)>='08' AND SUBSTR(HORATMIN,1,2)<'16' THEN HORATMIN_CAT='2';
IF SUBSTR(HORATMIN,1,2)>='16' AND SUBSTR(HORATMIN,1,2)<'24' THEN HORATMIN_CAT='3';

/* HORA TEMPERATURA MAXIMA */
IF SUBSTR(HORATMAX,1,2)>='00' AND SUBSTR(HORATMAX,1,2)<'08' THEN HORATMAX_CAT='1';
IF SUBSTR(HORATMAX,1,2)>='08' AND SUBSTR(HORATMAX,1,2)<'16' THEN HORATMAX_CAT='2';
IF SUBSTR(HORATMAX,1,2)>='16' AND SUBSTR(HORATMAX,1,2)<'24' THEN HORATMAX_CAT='3';

/* HORA RACHA VIENTO */
IF SUBSTR(HORARACHA,1,2)>='00' AND SUBSTR(HORARACHA,1,2)<'08' THEN HORARACHA_CAT='1';
IF SUBSTR(HORARACHA,1,2)>='08' AND SUBSTR(HORARACHA,1,2)<'16' THEN HORARACHA_CAT='2';
IF SUBSTR(HORARACHA,1,2)>='16' AND SUBSTR(HORARACHA,1,2)<'24' THEN HORARACHA_CAT='3';

/* HORA PRESION MINIMA */
IF SUBSTR(HORAPRESMIN,1,2)>='00' AND SUBSTR(HORAPRESMIN,1,2)<'08' THEN HORAPRESMIN_CAT='1';
IF SUBSTR(HORAPRESMIN,1,2)>='08' AND SUBSTR(HORAPRESMIN,1,2)<'16' THEN HORAPRESMIN_CAT='2';
IF SUBSTR(HORAPRESMIN,1,2)>='16' AND SUBSTR(HORAPRESMIN,1,2)<'24' THEN HORAPRESMIN_CAT='3';

/* HORA PRESION MAXIMA */
IF SUBSTR(HORAPRESMAX,1,2)>='00' AND SUBSTR(HORAPRESMAX,1,2)<'08' THEN HORAPRESMAX_CAT='1';
IF SUBSTR(HORAPRESMAX,1,2)>='08' AND SUBSTR(HORAPRESMAX,1,2)<'16' THEN HORAPRESMAX_CAT='2';
IF SUBSTR(HORAPRESMAX,1,2)>='16' AND SUBSTR(HORAPRESMAX,1,2)<'24' THEN HORAPRESMAX_CAT='3';

KEEP FECHA INDICATIVO NOMBRE COORDENADA_X COORDENADA_Y CLUSTER DISTANCIA_SEMILLA ALTITUD2 TMED2 PREC2
TMIN2 TMAX2 DIR2 VELMEDIA2 RACHA2 SOL2 PRESMAX2 PRESMIN2 HORATMIN_CAT HORATMAX_CAT HORARACHA_CAT
HORAPRESMIN_CAT HORAPRESMAX_CAT;

RUN;

DATA TRABAJO.DATOS_TOTAL_AEMET4;
SET TRABAJO.DATOS_TOTAL_AEMET4;
RENAME ALTITUD2=ALTITUD;
RENAME TMED2=TMED;
RENAME PREC2=PREC;
RENAME TMIN2=TMIN;
RENAME TMAX2=TMAX;
RENAME DIR2=DIR;
RENAME VELMEDIA2=VELMEDIA;
RENAME RACHA2=RACHA;
RENAME SOL2=SOL;
RENAME PRESMAX2=PRESMAX;
RENAME PRESMIN2=PRESMIN;
RUN;
```

```
/* EXPLORACIÓN DE VARIABLES AEMET */

/* PRINCIPALES ESTADÍSTICOS */

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET4;
OUTPUT OUT=AEMET_MEANS; RUN;

/* MISSINGS */

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET4 NMISS;
OUTPUT OUT=AEMET_MISSINGS;RUN;


/* IMPUTACIÓN DE MISSINGS */

/* PRIMER CRITERIO: ANTE LA APARICIÓN DE UN VALOR MISSING, RECUPERAMOS EL VALOR MEDIDO POR ESTACIÓN
EN EL MISMO DÍA Y ASOCIADO AL MISMO CLUSTER
LAS VARIABLES
- HORA PRESIÓN MÍNIMA
- HORA PRESIÓN MÁXIMA
- SOL
LAS DESECHAMOS POR EL ELEVADO % DE OBSERVACIONES CON MISSING. LA IMPUTACIÓN SE HARÁ SOBRE EL RESTO DE
VARIABLES. COMO PASO PREVIO AL CRUCE DE TABLAS PARA IMPUTAR LOS VALORES, CREO TABLAS AUXILIARES CON
AQUELLAS OBSERVACIONES NO-MISSING */

DATA
TMED_NOMISSING
PREC_NOMISSING
TMIN_NOMISSING
TMAX_NOMISSING
DIR_NOMISSING
VELMEDIA_NOMISSING
RACHA_NOMISSING
PRESMAX_NOMISSING
PRESMIN_NOMISSING
HORATMIN_CAT_NOMISSING
HORATMAX_CAT_NOMISSING
HORARACHA_CAT_NOMISSING;

SET TRABAJO.DATOS_TOTAL_AEMET4;

IF TMED^=. THEN OUTPUT TMED_NOMISSING;
IF PREC^=. THEN OUTPUT PREC_NOMISSING;
IF TMIN^=. THEN OUTPUT TMIN_NOMISSING;
IF TMAX^=. THEN OUTPUT TMAX_NOMISSING;
IF DIR^=. THEN OUTPUT DIR_NOMISSING;
IF VELMEDIA^=. THEN OUTPUT VELMEDIA_NOMISSING;
IF RACHA^=. THEN OUTPUT RACHA_NOMISSING;
IF PRESMAX^=. THEN OUTPUT PRESMAX_NOMISSING;
IF PRESMIN^=. THEN OUTPUT PRESMIN_NOMISSING;
IF HORATMIN_CAT^='' THEN OUTPUT HORATMIN_CAT_NOMISSING;
IF HORATMAX_CAT^='' THEN OUTPUT HORATMAX_CAT_NOMISSING;
IF HORARACHA_CAT^='' THEN OUTPUT HORARACHA_CAT_NOMISSING;

RUN;

/* QUITAMOS DUPLICADOS EN TODAS LAS TABLAS AUXILIARES CREADAS EN EL PASO ANTERIOR PORQUE DE CARA A LA
IMPUTACIÓN ÚNICAMENTE NECESITAMOS QUEDARNOS CON UN ÚNICO DATO PARA CADA RELACIÓN FECHA-CLUSTER-
VARIABLE */

PROC SORT DATA=TMED_NOMISSING NODUPKEY OUT=TMED_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=PREC_NOMISSING NODUPKEY OUT=PREC_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=TMIN_NOMISSING NODUPKEY OUT=TMIN_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=TMAX_NOMISSING NODUPKEY OUT=TMAX_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=DIR_NOMISSING NODUPKEY OUT=DIR_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=VELMEDIA_NOMISSING NODUPKEY OUT=VELMEDIA_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=RACHA_NOMISSING NODUPKEY OUT=RACHA_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=PRESMAX_NOMISSING NODUPKEY OUT=PRESMAX_NOMISSING_SD;
BY FECHA CLUSTER; RUN;
PROC SORT DATA=PRESMIN_NOMISSING NODUPKEY OUT=PRESMIN_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=HORATMIN_CAT_NOMISSING NODUPKEY OUT=HORATMIN_CAT_NOMISSING_SD;
BY FECHA CLUSTER; RUN;
```

```
PROC SORT DATA=HORATMAX_CAT_NOMISSING NODUPKEY OUT=HORATMAX_CAT_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

PROC SORT DATA=HORARACHA_CAT_NOMISSING NODUPKEY OUT=HORARACHA_CAT_NOMISSING_SD;
BY FECHA CLUSTER; RUN;

/* UNA VEZ CREADAS LAS TABLAS AUXILIARES SE PROCEDE A LA IMPUTACIÓN SEGÚN EL PRIMER CRITERIO:
SE IMPUTA (SI LO HAY) EL VALOR MEDIDO POR OTRA ESTACIÓN EN EL MISMO DÍA Y ASOCIADO AL MISMO CLUSTER
*/

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_AEMET5 AS
SELECT DISTINCT A.*,
CASE WHEN (A.TMED^=.) THEN A.TMED ELSE B.TMED END AS TMED_AJUST,
CASE WHEN (A.PREC^=.) THEN A.PREC ELSE C.PREC END AS PREC_AJUST,
CASE WHEN (A.TMIN^=.) THEN A.TMIN ELSE D.TMIN END AS TMIN_AJUST,
CASE WHEN (A.TMAX^=.) THEN A.TMAX ELSE E.TMAX END AS TMAX_AJUST,
CASE WHEN (A.DIR^=.) THEN A.DIR ELSE F.DIR END AS DIR_AJUST,
CASE WHEN (A.VELMEDIA^=.) THEN A.VELMEDIA ELSE G.VELMEDIA END AS VELMEDIA_AJUST,
CASE WHEN (A.RACHA^=.) THEN A.RACHA ELSE H.RACHA END AS RACHA_AJUST,
CASE WHEN (A.PRESMAX^=.) THEN A.PRESMAX ELSE I.PRESMAX END AS PRESMAX_AJUST,
CASE WHEN (A.PRESMIN^=.) THEN A.PRESMIN ELSE J.PRESMIN END AS PRESMIN_AJUST,
CASE WHEN (A.HORATMIN_CAT^='') THEN A.HORATMIN_CAT ELSE K.HORATMIN_CAT END AS HORATMIN_CAT_AJUST,
CASE WHEN (A.HORATMAX_CAT^='') THEN A.HORATMAX_CAT ELSE L.HORATMAX_CAT END AS HORATMAX_CAT_AJUST,
CASE WHEN (A.HORARACHA_CAT^='') THEN A.HORARACHA_CAT ELSE M.HORARACHA_CAT END AS HORARACHA_CAT_AJUST

FROM TRABAJO.DATOS_TOTAL_AEMET4 AS A

/* CRUCE CON CADA UNA DE LAS TABLAS AUXILIARES */
LEFT JOIN TMED_NOMISSING_SD AS B
ON A.FECHA=B.FECHA AND A.CLUSTER=B.CLUSTER
LEFT JOIN PREC_NOMISSING_SD AS C
ON A.FECHA=C.FECHA AND A.CLUSTER=C.CLUSTER
LEFT JOIN TMIN_NOMISSING_SD AS D
ON A.FECHA=D.FECHA AND A.CLUSTER=D.CLUSTER
LEFT JOIN TMAX_NOMISSING_SD AS E
ON A.FECHA=E.FECHA AND A.CLUSTER=E.CLUSTER
LEFT JOIN DIR_NOMISSING_SD AS F
ON A.FECHA=F.FECHA AND A.CLUSTER=F.CLUSTER
LEFT JOIN VELMEDIA_NOMISSING_SD AS G
ON A.FECHA=G.FECHA AND A.CLUSTER=G.CLUSTER
LEFT JOIN RACHA_NOMISSING_SD AS H
ON A.FECHA=H.FECHA AND A.CLUSTER=H.CLUSTER
LEFT JOIN PRESMAX_NOMISSING_SD AS I
ON A.FECHA=I.FECHA AND A.CLUSTER=I.CLUSTER
LEFT JOIN PRESMIN_NOMISSING_SD AS J
ON A.FECHA=J.FECHA AND A.CLUSTER=J.CLUSTER
LEFT JOIN HORATMIN_CAT_NOMISSING_SD AS K
ON A.FECHA=K.FECHA AND A.CLUSTER=K.CLUSTER
LEFT JOIN HORATMAX_CAT_NOMISSING_SD AS L
ON A.FECHA=L.FECHA AND A.CLUSTER=L.CLUSTER
LEFT JOIN HORARACHA_CAT_NOMISSING_SD AS M
ON A.FECHA=M.FECHA AND A.CLUSTER=M.CLUSTER;
QUIT;

/* COMPROBACIÓN DE MISSINGS TRAS LA PRIMERA IMPUTACIÓN */

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET5 NMISS;
OUTPUT OUT=AEMET_MISSINGS2; RUN;

/* SEGUNDO CRITERIO DE IMPUTACIÓN DE MISSING:
SE IMPUTA (SI LO HAY) EL VALOR MEDIDO POR OTRA ESTACIÓN EN EL MISMO DÍA CON INDEPENDENCIA DE QUE
PERTENEZCA O NO A LA MISMA ZONIFICACIÓN(CLUSTER) */

/* QUITAMOS DUPLICADOS EN TODAS LAS TABLAS AUXILIARES CREADAS EN EL PASO ANTERIOR PORQUE DE CARA A LA
IMPUTACIÓN ÚNICAMENTE NECESITAMOS QUEDARNOS CON UN ÚNICO DATO PARA CADA RELACIÓN FECHA-VARIABLE */

PROC SORT DATA=TMED_NOMISSING_SD NODUPKEY OUT=TMED_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=PREC_NOMISSING_SD NODUPKEY OUT=PREC_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=TMIN_NOMISSING_SD NODUPKEY OUT=TMIN_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=TMAX_NOMISSING_SD NODUPKEY OUT=TMAX_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=DIR_NOMISSING_SD NODUPKEY OUT=DIR_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=VELMEDIA_NOMISSING_SD NODUPKEY OUT=VELMEDIA_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=RACHA_NOMISSING_SD NODUPKEY OUT=RACHA_NOMISSING_SD2;
BY FECHA; RUN;
```

```
PROC SORT DATA=PRESMAX_NOMISSING_SD NODUPKEY OUT=PRESMAX_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=PRESMIN_NOMISSING_SD NODUPKEY OUT=PRESMIN_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=HORATMIN_CAT_NOMISSING_SD NODUPKEY OUT=HORATMIN_CAT_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=HORATMAX_CAT_NOMISSING_SD NODUPKEY OUT=HORATMAX_CAT_NOMISSING_SD2;
BY FECHA; RUN;

PROC SORT DATA=HORARACHA_CAT_NOMISSING_SD NODUPKEY OUT=HORARACHA_CAT_NOMISSING_SD2;
BY FECHA; RUN;

/* UNA VEZ CREADAS LAS TABLAS AUXILIARES SE PROCEDE A LA IMPUTACIÓN SEGÚN EL SEGUNDO CRITERIO:
SE IMPUTA (SI LO HAY) EL VALOR MEDIDO POR OTRA ESTACIÓN EN EL MISMO DÍA Y ASOCIADO AL MISMO CLUSTER
*/

PROC SQL;
CREATE TABLE TRABAJO.DATOS_TOTAL_AEMET6 AS
SELECT DISTINCT A.*,
CASE WHEN (A.TMED_AJUST^=.) THEN A.TMED_AJUST ELSE B.TMED END AS TMED_AJUST2,
CASE WHEN (A.PREC_AJUST^=.) THEN A.PREC_AJUST ELSE C.PREC END AS PREC_AJUST2,
CASE WHEN (A.TMIN_AJUST^=.) THEN A.TMIN_AJUST ELSE D.TMIN END AS TMIN_AJUST2,
CASE WHEN (A.TMAX_AJUST^=.) THEN A.TMAX_AJUST ELSE E.TMAX END AS TMAX_AJUST2,
CASE WHEN (A.DIR_AJUST^=.) THEN A.DIR_AJUST ELSE F.DIR END AS DIR_AJUST2,
CASE WHEN (A.VELMEDIA_AJUST^=.) THEN A.VELMEDIA_AJUST ELSE G.VELMEDIA END AS VELMEDIA_AJUST2,
CASE WHEN (A.RACHA_AJUST^=.) THEN A.RACHA_AJUST ELSE H.RACHA END AS RACHA_AJUST2,
CASE WHEN (A.PRESMAX_AJUST^=.) THEN A.PRESMAX_AJUST ELSE I.PRESMAX END AS PRESMAX_AJUST2,
CASE WHEN (A.PRESMIN_AJUST^=.) THEN A.PRESMIN_AJUST ELSE J.PRESMIN END AS PRESMIN_AJUST2,
CASE WHEN (A.HORATMIN_CAT_AJUST^='') THEN A.HORATMIN_CAT_AJUST ELSE K.HORATMIN_CAT END AS
HORATMIN_CAT_AJUST2,
CASE WHEN (A.HORATMAX_CAT_AJUST^='') THEN A.HORATMAX_CAT_AJUST ELSE L.HORATMAX_CAT END AS
HORATMAX_CAT_AJUST2,
CASE WHEN (A.HORARACHA_CAT_AJUST^='') THEN A.HORARACHA_CAT_AJUST ELSE M.HORARACHA_CAT END AS
HORARACHA_CAT_AJUST2

FROM TRABAJO.DATOS_TOTAL_AEMET5 AS A

/* CRUCE CON CADA UNA DE LAS TABLAS AUXILIARES
A DIFERENCIA DEL CRUCE ANTERIOR, NOS DA IGUAL QUE EL DATO PERTENEZCA A UNA ESTACIÓN DE LA MISMA
ZONIFICACIÓN POR ELLO ELIMINAMOS COMO CONDICIÓN QUE CRUCE POR CLUSTER */

LEFT JOIN TMED_NOMISSING_SD2 AS B
ON A.FECHA=B.FECHA
LEFT JOIN PREC_NOMISSING_SD2 AS C
ON A.FECHA=C.FECHA
LEFT JOIN TMIN_NOMISSING_SD2 AS D
ON A.FECHA=D.FECHA
LEFT JOIN TMAX_NOMISSING_SD2 AS E
ON A.FECHA=E.FECHA
LEFT JOIN DIR_NOMISSING_SD2 AS F
ON A.FECHA=F.FECHA
LEFT JOIN VELMEDIA_NOMISSING_SD2 AS G
ON A.FECHA=G.FECHA
LEFT JOIN RACHA_NOMISSING_SD2 AS H
ON A.FECHA=H.FECHA
LEFT JOIN PRESMAX_NOMISSING_SD2 AS I
ON A.FECHA=I.FECHA
LEFT JOIN PRESMIN_NOMISSING_SD2 AS J
ON A.FECHA=J.FECHA
LEFT JOIN HORATMIN_CAT_NOMISSING_SD2 AS K
ON A.FECHA=K.FECHA
LEFT JOIN HORATMAX_CAT_NOMISSING_SD2 AS L
ON A.FECHA=L.FECHA
LEFT JOIN HORARACHA_CAT_NOMISSING_SD2 AS M
ON A.FECHA=M.FECHA;

QUIT;

/* COMPROBACIÓN DE MISSINGS TRAS LA SEGUNDA IMPUTACIÓN */

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET6 NMISS;
OUTPUT OUT=AEMET_MISSINGS3;
RUN;

/* UNA VEZ COMPROBADO LA AUSENCIA DE VALORES MISSING, NOS QUEDAMOS SOLO CON LOS CAMPOS QUE NOS
INTERESAN

TAMBIEN SE CREA UNA VARIABLE AUXILIAR QUE ME VA A SERVIR POSTERIORMENTE PARA EL TRATAMIENTO DE TABLAS
(AGRUPACIONES, ETC.) */

DATA TRABAJO.DATOS_TOTAL_AEMET_FINAL;
SET TRABAJO.DATOS_TOTAL_AEMET6;
KEEP FECHA INDICATIVO NOMBRE COORDENADA_X COORDENADA_Y CLUSTER DISTANCIA_SEMILLA ALTITUD TMED_AJUST2
PREC_AJUST2 TMIN_AJUST2 TMAX_AJUST2 DIR_AJUST2 VELMEDIA_AJUST2 RACHA_AJUST2 PRESMAX_AJUST2
PRESMIN_AJUST2 HORATMIN_CAT_AJUST2 HORATMAX_CAT_AJUST2 HORARACHA_CAT_AJUST2;RUN;
```

```
DATA TRABAJO.DATOS_TOTAL_AEMET_FINAL;
SET TRABAJO.DATOS_TOTAL_AEMET_FINAL;
RENAME TMED_AJUST2=TMED;
RENAME PREC_AJUST2=PREC;
RENAME TMIN_AJUST2=TMIN;
RENAME TMAX_AJUST2=TMAX;
RENAME DIR_AJUST2=DIR;
RENAME VELMEDIA_AJUST2=VELMEDIA;
RENAME RACHA_AJUST2=RACHA;
RENAME PRESMAX_AJUST2=PRESMAX;
RENAME PRESMIN_AJUST2=PRESMIN;
RENAME HORATMIN_CAT_AJUST2=HORATMIN_CAT;
RENAME HORATMAX_CAT_AJUST2=HORATMAX_CAT;
RENAME HORARACHA_CAT_AJUST2=HORARACHA_CAT;

AUX_CLUSTER_DIA=CAT(CLUSTER,'_',FECHA);
RUN;

PROC SORT DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL;
BY AUX_CLUSTER_DIA;

RUN;

/* PARA LAS VARIABLES CONTINUAS NOS VAMOS A QUEDAR CON LA MEDIA DE CADA VARIABLE */

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR TMED;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_TMED (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR PREC;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_PREC (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR TMIN;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_TMIN (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR TMAX;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_TMAX (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR DIR;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_DIR (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR VELMEDIA;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_VELMEDIA (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR RACHA;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_RACHA (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR PRESMAX;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_PRESMAX (WHERE=(_STAT_='MEAN')); RUN;

PROC MEANS DATA=TRABAJO.DATOS_TOTAL_AEMET_FINAL NOPRINT;
BY AUX_CLUSTER_DIA;
VAR PRESMIN;
OUTPUT OUT=TRABAJO.AEMET_MEDIA_PRESMIN (WHERE=(_STAT_='MEAN')); RUN;
/* PARA LAS VARIABLES CATEGORICAS NOS QUEDAMOS CON LA MODA */

/* MODA HORA TMIN CATEGORICA */

PROC SQL;
CREATE TABLE AEMET_MODA_HORATMIN_CAT_AUX AS
SELECT AUX_CLUSTER_DIA, HORATMIN_CAT, COUNT(*) AS CONTEO
FROM TRABAJO.DATOS_TOTAL_AEMET_FINAL
GROUP BY 1, 2
ORDER BY AUX_CLUSTER_DIA, CONTEO DESC;
QUIT;

DATA TRABAJO.AEMET_MODA_HORATMIN_CAT;
SET AEMET_MODA_HORATMIN_CAT_AUX;
IF FIRST.AUX_CLUSTER_DIA;
BY AUX_CLUSTER_DIA;
RENAME HORATMIN_CAT=MODA_HORATMIN_CAT;
DROP CONTEO;
RUN;
```

```
/* MODA HORA TMAX CATEGORICA */

PROC SQL;
CREATE TABLE AEMET_MODA_HORATMAX_CAT_AUX AS
SELECT AUX_CLUSTER_DIA, HORATMAX_CAT, COUNT(*) AS CONTEO
FROM TRABAJO.DATOS_TOTAL_AEMET_FINAL
GROUP BY 1, 2
ORDER BY AUX_CLUSTER_DIA, CONTEO DESC;
QUIT;


DATA TRABAJO.AEMET_MODA_HORATMAX_CAT;
SET AEMET_MODA_HORATMAX_CAT_AUX;
IF FIRST.AUX_CLUSTER_DIA;
BY AUX_CLUSTER_DIA;
RENAME HORATMAX_CAT=MODA_HORATMAX_CAT;
DROP CONTEO;
RUN;

/* MODA HORA RACHA CATEGORICA */

PROC SQL;
CREATE TABLE AEMET_MODA_HORARACHA_CAT_AUX AS
SELECT AUX_CLUSTER_DIA, HORARACHA_CAT, COUNT(*) AS CONTEO
FROM TRABAJO.DATOS_TOTAL_AEMET_FINAL
GROUP BY 1, 2
ORDER BY AUX_CLUSTER_DIA, CONTEO DESC;
QUIT;

DATA TRABAJO.AEMET_MODA_HORARACHA_CAT;
SET AEMET_MODA_HORARACHA_CAT_AUX;
IF FIRST.AUX_CLUSTER_DIA;
BY AUX_CLUSTER_DIA;
RENAME HORARACHA_CAT=MODA_HORARACHA_CAT;
DROP CONTEO;
RUN;
```

2.7. Series temporales y estudio de autocorrelación y autocorrelación cruzada

```
/*##################################################################################################
####################################################################################################

                                        CÓDIGO 7
                ESTUDIO DE SERIES TEMPORALES: AUTOCORRELACIÓN Y CORRELACIÓN CRUZADA

####################################################################################################
####################################################################################################


PARA EL ANALISIS DE LA SERIE TEMPORAL NOS VAMOS A CENTRAR EN AQUELLA ESTACIÓN QUE MÁS DÍAS HAYA
SUPERADO EL UMBRAL DE ALARMA */

PROC SQL;
CREATE TABLE TRABAJO.NUM_ALARMA_POR_ESTACION AS
SELECT COD_ESTACION, ESTACION, COUNT(*) AS CONTEO_DIAS
FROM
(SELECT DISTINCT COD_ESTACION, ESTACION, ANO, MES, DIA
FROM TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST
WHERE MARCA_ALERTA_NO2=1)
GROUP BY 1, 2;
QUIT;

/* LA ESTACIÓN DE 28079056 - PZA. FDEZ. LADREDA ES LA QUE MÁS VECES HA SUPERADO EL NIVEL DE ALERTA DE
NO2 */

DATA TRABAJO.DATOS_NO2_28079056;
SET TRABAJO.DATOS_TOTAL_INFO_NO2_FINAL_AJUST;
WHERE COD_ESTACION='28079056';
RUN;

/* NOS QUEDAMOS CON EL VALOR MÁXIMO DIARIO PARA DICHA ESTACIÓN */

PROC SQL;
CREATE TABLE TRABAJO.DATOS_MAXNO2_28079056 AS
SELECT FECHA, MAX(VALOR_AJUSTADO) AS MAX_NO2
FROM TRABAJO.DATOS_NO2_28079056
GROUP BY 1;
QUIT;

PROC TIMESERIES DATA=TRABAJO.DATOS_MAXNO2_28079056 PLOTS =(DECOMP PERIODOGRAM SERIES) PRINT=(SEASONS
DECOMP);
ID FECHA INTERVAL=DAY;
VAR MAX_NO2; RUN;

/* TEST DE DURBIN WATSON PARA VER AUTOCORRELACIÓN */
```

```
PROC AUTOREG DATA=TRABAJO.DATOS_MAXNO2_28079056 OUTEST=TRABAJO.RESIDUOS PLOTS =(RESIDUAL);
MODEL MAX_NO2 = FECHA / DW=4 DWPROB; RUN;

/* VEMOS POSIBLE AUTOCORRELACIÓN EN RESIDUOS */

PROC ARIMA DATA=TRABAJO.DATOS_MAXNO2_28079056;
IDENTIFY VAR=MAX_NO2 NLAG=7; RUN;

/* CORRELACIÓN CRUZADA: NO2 Y VARIABLES INDEPENDIENTES */

/* POR EJEMPLO: TMAX - ZONA 5 (LA ZONA DE PLAZA FERNANDEZ LADREDA) */

PROC SQL;
CREATE TABLE SERIE_TMAX_ZONA5_AUX AS
SELECT INPUT(FECHA,YYMMDD10.) AS FECHA FORMAT DDMMYY10., TMAX
FROM TRABAJO.DATOS_TOTAL_AEMET_FINAL
WHERE CLUSTER=5;
QUIT;

PROC SQL;
CREATE TABLE TRABAJO.SERIE_TMAX_ZONA5 AS
SELECT FECHA, MAX(TMAX) AS TMAX
FROM SERIE_TMAX_ZONA5_AUX
GROUP BY 1;
QUIT;

PROC SQL;
CREATE TABLE TMAX_VS_NO2 AS
SELECT DISTINCT A.*, B.TMAX
FROM TRABAJO.DATOS_MAXNO2_28079056 AS A
LEFT JOIN TRABAJO.SERIE_TMAX_ZONA5 AS B
ON A.FECHA=B.FECHA;
QUIT;

PROC ARIMA DATA=TMAX_VS_NO2 PLOTS=ALL;
IDENTIFY VAR=MAX_NO2(1 7) CROSS=(TMAX(1 7)) NLAGS=7;
RUN;
```

2.8. Generación de tablón base para la modelización

```
/*#############################################################################################################
#############################################################################################################

                                                CÓDIGO 8
                        GENERACIÓN DEL TABLÓN BASE PARA LA MODELIZACIÓN MULTIVARIANTE

#############################################################################################################
#############################################################################################################

VAMOS A GENERAR EL TABLON QUE RECOPILE LA INFORMACIÓN DIARIA DE MAX(NO2) + VARIABLES CLIMATOLOGICAS Y
VARIBLES DE TRÁFICO + VARIABLES GENERADAS DE CARA A COMENZAR CON LA MODELIZACIÓN */

PROC SQL;
CREATE TABLE TRABAJO.TABLON AS
SELECT DISTINCT A.*,

/* CATEGORIZAMOS NUESTRA VARAIABLE OBJETIVO: 1 = PREAVISO (NIVEL NO2 IGUAL O SUPERIOR A UMBRAL DE
PREAVISO 180) 0 = OK (NIVEL NO2 INFERIOR A UMBRAL DE PREAVISO) */

CASE WHEN A.MAX_NO2>=180 THEN 1 ELSE 0 END AS MARCA_PREAVISO_NO2,

/* SE AÑADE INFO CLIMATOLOGICA AEMET AGRUPADA POR DIA Y CLUSTER/ZONA */

B.TMED,
C.PREC,
D.TMIN,
E.TMAX,
F.DIR,
G.VELMEDIA,
H.RACHA,
I.PRESMAX,
J.PRESMIN,
K.MODA_HORATMIN_CAT,
L.MODA_HORATMAX_CAT,
M.MODA_HORARACHA_CAT,

/* SE AÑADE INFO TRÁFICO AGRUPADA POR DIA Y CLUSTER/ZONA */

N.MAX_INTENSIDAD,
O.SUM_MAX_INTENSIDAD,
P.MAX_MEDIA_INTENSIDAD,
Q.MEDIA_INTENSIDAD,
R.MEDIA_INTENSIDAD AS MEDIA_INTENSIDAD_PTOMAX,
INPUT(SUBSTR(AUX_ZONA_DIA,3,10),YYMMDD10.) AS DIA FORMAT=DATE9.

FROM TRABAJO.NO2_MAXNO2_ZONA AS A
```

```
/* CRUCE CON INFO DE AEMET */
LEFT JOIN TRABAJO.AEMET_MEDIA_TMED AS B
ON A.AUX_ZONA_DIA=B.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_PREC AS C
ON A.AUX_ZONA_DIA=C.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_TMIN AS D
ON A.AUX_ZONA_DIA=D.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_TMAX AS E
ON A.AUX_ZONA_DIA=E.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_DIR AS F
ON A.AUX_ZONA_DIA=F.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_VELMEDIA AS G
ON A.AUX_ZONA_DIA=G.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_RACHA AS H
ON A.AUX_ZONA_DIA=H.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_PRESMAX AS I
ON A.AUX_ZONA_DIA=I.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MEDIA_PRESMIN AS J
ON A.AUX_ZONA_DIA=J.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MODA_HORATMIN_CAT AS K
ON A.AUX_ZONA_DIA=K.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MODA_HORATMAX_CAT AS L
ON A.AUX_ZONA_DIA=L.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.AEMET_MODA_HORARACHA_CAT AS M
ON A.AUX_ZONA_DIA=M.AUX_CLUSTER_DIA

/* CRUCE CON INFO DE TRÁFICO */
LEFT JOIN TRABAJO.MAX_INTENSIDAD_POR_ZONA AS N
ON A.AUX_ZONA_DIA=N.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.SUMA_MAX_INTENSIDAD_POR_ZONA AS O
ON A.AUX_ZONA_DIA=O.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.INTENSIDAD_MAX_MEDIA_POR_ZONA AS P
ON A.AUX_ZONA_DIA=P.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.INTENSIDAD_MEDIA_POR_ZONA AS Q
ON A.AUX_ZONA_DIA=Q.AUX_CLUSTER_DIA
LEFT JOIN TRABAJO.VALOR_MEDIO_INTENS_PTOMAX AS R
ON A.AUX_ZONA_DIA=R.AUX_CLUSTER_DIA

ORDER BY AUX_ZONA_DIA; QUIT;

/* DE CARA A LA PREDICCIÓN VAMOS A CREAR TAMBIÉN VARIABLES CON DECALAJE. COMO HEMOS VISTO EN EL
APARTADO DE SERIES TEMPORALES, EXISTE CORRELACIÓN ENTRE NUESTRA VARIABLE OBJETIVO Y EL VALOR DE
NUESTRAS VARIABLES EXPLICATIVAS EN D-1, YA SEA PORQUE EL PROCESO DE GENERACIÓN DE NO2 REQUIERE
TIEMPO, O PORQUE LAS PARTÍCULAS SE MANTIENEN EN EL AIRE UN TIEMPO, ETC.
COMO PASO PREVIO ORDENAMOS POR ZONA Y DÍA */

PROC SORT DATA=TRABAJO.TABLON OUT=TABLON_ORDENADO;
BY AUX_ZONA_DIA; RUN;

/* A CONTINUACIÓN DIVIDIMOS TABLON EN LAS 5 ZONAS, YA QUE CUANDO UTILICEMOS LA FUNCION LAG NO
QUEREMOS IMPUTAR A UN DÍA LA OBSERVACIÓN DE OTRA ZONA */

DATA AUX_TABLON_ZONA1 AUX_TABLON_ZONA2 AUX_TABLON_ZONA3 AUX_TABLON_ZONA4 AUX_TABLON_ZONA5;
SET TABLON_ORDENADO;
IF IDZONA='1' THEN OUTPUT AUX_TABLON_ZONA1;
IF IDZONA='2' THEN OUTPUT AUX_TABLON_ZONA2;
IF IDZONA='3' THEN OUTPUT AUX_TABLON_ZONA3;
IF IDZONA='4' THEN OUTPUT AUX_TABLON_ZONA4;
IF IDZONA='5' THEN OUTPUT AUX_TABLON_ZONA5; RUN;

/* AHORA EN CADA UNA DE LAS 5 TABLAS, CREAMOS LAS VARIABLES RETARDADAS
COMO SON LAS MISMAS VARIABLES PARA CADA ZONA CREO UNA MACRO SENCILLA */

%MACRO GENERA_RETARDOS(TABLAORIGEN=);

DATA &TABLAORIGEN._LAG;
SET &TABLAORIGEN.;
LAG_MAX_NO2=LAG1(MAX_NO2);
LAG_TMED=LAG1(TMED);
LAG_PREC=LAG1(PREC);
LAG_TMIN=LAG1(TMIN);
LAG_TMAX=LAG1(TMAX);
LAG_DIR=LAG1(DIR);
LAG_VELMEDIA=LAG1(VELMEDIA);
LAG_RACHA=LAG1(RACHA);
LAG_PRESMAX=LAG1(PRESMAX);
LAG_PRESMIN=LAG1(PRESMIN);
LAG_MODA_HORATMIN_CAT=LAG1(MODA_HORATMIN_CAT);
LAG_MODA_HORATMAX_CAT=LAG1(MODA_HORATMAX_CAT);
LAG_MODA_HORARACHA_CAT=LAG1(MODA_HORARACHA_CAT);
LAG_MAX_INTENSIDAD=LAG1(MAX_INTENSIDAD);
LAG_SUM_MAX_INTENSIDAD=LAG1(SUM_MAX_INTENSIDAD);
LAG_MAX_MEDIA_INTENSIDAD=LAG1(MAX_MEDIA_INTENSIDAD);
LAG_MEDIA_INTENSIDAD=LAG1(MEDIA_INTENSIDAD);
LAG_MEDIA_INTENSIDAD_PTOMAX=LAG1(MEDIA_INTENSIDAD_PTOMAX);
RUN;

%MEND;
```

```
/* LLAMADAS A LA MACRO PARA GENERAR TABLAS CON RETARDOS */

%GENERA_RETARDOS(TABLAORIGEN=AUX_TABLON_ZONA1);
%GENERA_RETARDOS(TABLAORIGEN=AUX_TABLON_ZONA2);
%GENERA_RETARDOS(TABLAORIGEN=AUX_TABLON_ZONA3);
%GENERA_RETARDOS(TABLAORIGEN=AUX_TABLON_ZONA4);
%GENERA_RETARDOS(TABLAORIGEN=AUX_TABLON_ZONA5);

/* JUNTAMOS TODO Y CREAMOS UN NUEVO TABLON QUE INCLUYE TAMBIÉN LAS VARIABLES CON RETARDOS */

DATA TRABAJO.TABLON2;
SET AUX_TABLON_ZONA1_LAG
AUX_TABLON_ZONA2_LAG
AUX_TABLON_ZONA3_LAG
AUX_TABLON_ZONA4_LAG
AUX_TABLON_ZONA5_LAG;
RUN;

/* MAPA DE CALOR DE LAS CORRELACIONES DE LAS VARIABLES ORIGINALES Y LAS DE RETARDO CREADAS */
/* LA SIGUIENTE MACRO LA HE SACADO DE INTERNET */

%MACRO PREPCORRDATA(IN=,OUT=);
PROC CORR DATA=&IN. NOPRINT PEARSON OUTP=WORK._TMPCORR VARDEF=DF ;
RUN;

DATA &OUT.;
KEEP X Y R;
SET WORK._TMPCORR(WHERE=(_TYPE_="CORR"));
ARRAY V{*} _NUMERIC_;
X = _NAME_;
DO I = DIM(V) TO 1 BY -1;
Y = VNAME(V(I));
R = V(I);
IF (I<_N_) THEN
R=.;
OUTPUT;
END; RUN;

PROC DATASETS LIB=WORK NOLIST NOWARN;
DELETE _TMPCORR;
QUIT;
%MEND;

ODS PATH WORK.MYSTORE(UPDATE) SASHELP.TMPLMST(READ);

PROC TEMPLATE;
DEFINE STATGRAPH CORRHEATMAP;
DYNAMIC _TITLE;
BEGINGRAPH;
ENTRYTITLE _TITLE;
RANGEATTRMAP NAME='MAP';
RANGE -1 - 1 / RANGECOLORMODEL=(CX336B87 CXF5F5F5 CX763626);
ENDRANGEATTRMAP;
RANGEATTRVAR VAR=R ATTRVAR=R ATTRMAP='MAP';
LAYOUT OVERLAY /
XAXISOPTS=(DISPLAY=(LINE TICKS TICKVALUES))
YAXISOPTS=(DISPLAY=(LINE TICKS TICKVALUES));
HEATMAPPARM X = X Y = Y COLORRESPONSE = R /
XBINAXIS=FALSE YBINAXIS=FALSE
NAME = "HEATMAP" DISPLAY=ALL;
CONTINUOUSLEGEND "HEATMAP" /
ORIENT = VERTICAL LOCATION = OUTSIDE TITLE="CORRELACION DE PEARSON";
ENDLAYOUT;
ENDGRAPH;
END;
RUN;

DATA VARIABLES_CLIMA;
SET TRABAJO.TABLON2;
KEEP MAX_NO2 TMED PREC TMIN TMAX DIR VELMEDIA RACHA PRESMAX PRESMIN; RUN;

DATA VARIABLES_CLIMA_LAG;
SET TRABAJO.TABLON2;
KEEP MAX_NO2 MAX_NO2_LAG LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN LAG_TMAX LAG_DIR LAG_VELMEDIA
LAG_RACHA LAG_PRESMAX LAG_PRESMIN; RUN;

/* RENOMBRO PARA QUE SE VEA MEJOR LOS GRÁFICOS */
DATA VARIABLES_TRAFICO;
SET TRABAJO.TABLON2;
KEEP MAX_NO2 MAX_INTENSIDAD SUM_MAX_INTENSIDAD MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD
MEDIA_INTENSIDAD_PTOMAX;
RENAME MAX_INTENSIDAD=MAX_INT;
RENAME SUM_MAX_INTENSIDAD=SUM_MAX_INT;
RENAME MAX_MEDIA_INTENSIDAD=MAX_MEDIA_INT;
RENAME MEDIA_INTENSIDAD=MEDIA_INT;
RENAME MEDIA_INTENSIDAD_PTOMAX=MEDIA_INT_PTOMAX; RUN;
```

```
DATA VARIABLES_TRAFICO_LAG;
SET TRABAJO.TABLON2;
KEEP MAX_NO2 LAG_MAX_INTENSIDAD LAG_SUM_MAX_INTENSIDAD LAG_MAX_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD
LAG_MEDIA_INTENSIDAD_PTOMAX;
RENAME LAG_MAX_INTENSIDAD=LAG_MAX_INT;
RENAME LAG_SUM_MAX_INTENSIDAD=LAG_SUM_MAX_INT;
RENAME LAG_MAX_MEDIA_INTENSIDAD=LAG_MAX_MEDIA_INT;
RENAME LAG_MEDIA_INTENSIDAD=LAG_MEDIA_INT;
RENAME LAG_MEDIA_INTENSIDAD_PTOMAX=LAG_MEDIA_INT_PTOMAX; RUN;

ods graphics /height=600 width=800 imagemap;
%prepCorrData(in=VARIABLES_CLIMA,out=CORR_CLIMA);
proc sgrender data=CORR_CLIMA template=corrHeatmap;
dynamic _title="Matriz de correlaciones - Variables climatológicas"; run;

ods graphics /height=600 width=800 imagemap;
%prepCorrData(in=VARIABLES_CLIMA_LAG,out=CORR_CLIMA_LAG);
proc sgrender data=CORR_CLIMA_LAG template=corrHeatmap;
dynamic _title="Matriz de correlaciones - Variables climatológicas con retardo"; run;

ods graphics /height=600 width=800 imagemap;
%prepCorrData(in=VARIABLES_TRAFICO,out=CORR_TRAFICO);
proc sgrender data=CORR_TRAFICO template=corrHeatmap;
dynamic _title="Matriz de correlaciones - Variables tráfico";run;

ods graphics /height=600 width=800 imagemap;
%prepCorrData(in=VARIABLES_TRAFICO_LAG,out=CORR_TRAFICO_LAG);
proc sgrender data=CORR_TRAFICO_LAG template=corrHeatmap;
dynamic _title="Matriz de correlaciones - Variables tráfico con retardo";
run;
```

2.9 Modelización y resultados

```
/*##########################################################################################################
##########################################################################################################

                                                CÓDIGO 9
                                            MODELIZACIÓN
                                             REGRESIÓN
                                               REDES
                                             BAGGING
                                          RANDOM FOREST
                                       GRADIENT BOOSTING

EN EL PRESENTE TFM SE HA PROBADO CON DIFERENTES TÉCNICAS Y, EN CADA UNA DE ELLAS, CON DIFERENTES
PARAMETRIZACIONES. POR SU EXCESIVA EXTENSIÓN SE ADJUNTA ÚNICAMENTE EL CÓDIGO DE LA VALIDACIÓN CRUZADA
DE LOS MEJORES MODELOS CALCULADOS. TODOS LOS MODELOS HAN SIDO VALIDADOS MEDIANTE VALIDACIÓN CRUZADA
UTILIZANDO MACROS FACILITADAS Y PROGRAMADAS POR EL PROFESOR D.JAVIER PORTELA LAS CUALES NO SE ADJUNTAN
A ESTE TFM.

##########################################################################################################
##########################################################################################################



/*##########################################################################################################
                                        REGRESIÓN LOGÍSTICA BINARIA
##########################################################################################################*/

/* MODELO2 STEPWISE CON INTERACCIONES */

%CRUZADALOGISTICA(ARCHIVO=TRABAJO.TABLON2,
VARDEPEN=MARCA_PREAVISO_NO2,
CONTI=LAG_MAX_NO2*IDZONA SUM_MAX_INTENSIDAD TMAX*LAG_MODA_HORATMAX_CAT TMED LAG_MODA_HORATMAX_CAT
MODA_HORARACHA_CAT TMAX PREC PRESMIN IDZONA PRESMIN*IDZONA VELMEDIA RACHA RACHA*MODA_HORARACHA_CAT
LAG_MAX_NO2,
CATEGOR=IDZONA LAG_MODA_HORATMAX_CAT MODA_HORARACHA_CAT,
NGRUPOS=4,
SINICIO=12375,
SFINAL=12575,
OBJETIVO=TASAFALLOS);
DATA TRABAJO.REG2;
SET FINAL;
MODELO='REG2'; MODEL='REGRESION LOGISTICA';
```

```
/*##########################################################################################################
                                                RED NEURONAL
##########################################################################################################*/

%CRUZADABINARIANEURAL(ARCHIVO=TRABAJO.TABLON2,
VARDEPEN=MARCA_PREAVISO_NO2,
CONTI=TMED PREC TMIN TMAX DIR VELMEDIA RACHA PRESMAX PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD
MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTOMAX LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN
LAG_TMAX LAG_DIR LAG_VELMEDIA LAG_RACHA LAG_PRESMAX LAG_PRESMIN LAG_MAX_INTENSIDAD
LAG_SUM_MAX_INTENSIDAD LAG_MAX_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD_PTOMAX,
CATEGOR=IDZONA MODA_HORATMIN_CAT MODA_HORATMAX_CAT MODA_HORARACHA_CAT LAG_MODA_HORATMIN_CAT
LAG_MODA_HORATMAX_CAT LAG_MODA_HORARACHA_CAT,
NGRUPOS=4,
SINICIO=12345,
SFINAL=12374,
NODOS=5,
FUNCACT=LIN, METO=LEVMAR,
OBJETIVO=TASAFALLOS);
DATA TRABAJO.RED16; SET FINAL; MODELO='RED16'; MODEL='RED NEURONAL'; RUN;


/*##########################################################################################################
                                                BAGGING
##########################################################################################################*/

%CRUZADABAGGINGBIN(ARCHIVO=TRABAJO.TABLON2,
VARDEPEN=MARCA_PREAVISO_NO2,
LISTCONTI=TMED PREC TMIN TMAX DIR VELMEDIA RACHA PRESMAX PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD
MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTOMAX LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN
LAG_TMAX LAG_DIR LAG_VELMEDIA LAG_RACHA LAG_PRESMAX LAG_PRESMIN LAG_MAX_INTENSIDAD
LAG_SUM_MAX_INTENSIDAD LAG_MAX_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD_PTOMAX,
LISTCATEGOR=IDZONA MODA_HORATMIN_CAT MODA_HORATMAX_CAT MODA_HORARACHA_CAT LAG_MODA_HORATMIN_CAT
LAG_MODA_HORATMAX_CAT LAG_MODA_HORARACHA_CAT,
NGRUPOS=4,SINICIO=1000,SFINAL=1010,
SINICIOBAG=1000,SFINALBAG=1039,
PORCENBAG=0.9,MAXBRANCH=2,
NLEAVES=30,TAMHOJA=40,
REEMPLAZO=1,OBJETIVO=TASAFALLOS);
DATA TRABAJO.BAG17; SET FINAL; MODELO="BAG17"; MODEL='BAGGING'; RUN;


/*##########################################################################################################
                                                RANDOM FOREST
##########################################################################################################*/

%CRUZADARANDOMFORESTBIN(ARCHIVO=TRABAJO.TABLON2,
VARDEP=MARCA_PREAVISO_NO2,
LISTCONTI=TMED PREC TMIN TMAX DIR VELMEDIA RACHA PRESMAX PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD
MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTOMAX LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN
LAG_TMAX LAG_DIR LAG_VELMEDIA LAG_RACHA LAG_PRESMAX LAG_PRESMIN LAG_MAX_INTENSIDAD
LAG_SUM_MAX_INTENSIDAD LAG_MAX_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD_PTOMAX,
LISTCATEGOR=IDZONA MODA_HORATMIN_CAT MODA_HORATMAX_CAT MODA_HORARACHA_CAT LAG_MODA_HORATMIN_CAT
LAG_MODA_HORATMAX_CAT LAG_MODA_HORARACHA_CAT,
PORCENBAG=0.9,VARIABLES=20,MAXTREES=20,MAXDEPTH=10,MAXBRANCH=2,TAMHOJA=40,PVALOR=0.1,
NGRUPOS=4,SINICIO=12345,SFINAL=12374,OBJETIVO=TASAFALLOS);
DATA TRABAJO.RF9;SET FINAL;MODELO="RF09"; MODEL='RANDOM FOREST'; RUN;

/*##########################################################################################################
                                                GRADIENT BOOSTING
##########################################################################################################*/

%CRUZADATREEBOOSTBIN(ARCHIVO=TRABAJO.TABLON2,
VARDEPEN=MARCA_PREAVISO_NO2,
CONTI=TMED PREC TMIN TMAX DIR VELMEDIA RACHA PRESMAX PRESMIN MAX_INTENSIDAD SUM_MAX_INTENSIDAD
MAX_MEDIA_INTENSIDAD MEDIA_INTENSIDAD MEDIA_INTENSIDAD_PTOMAX LAG_MAX_NO2 LAG_TMED LAG_PREC LAG_TMIN
LAG_TMAX LAG_DIR LAG_VELMEDIA LAG_RACHA LAG_PRESMAX LAG_PRESMIN LAG_MAX_INTENSIDAD
LAG_SUM_MAX_INTENSIDAD LAG_MAX_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD LAG_MEDIA_INTENSIDAD_PTOMAX,
CATEGOR=IDZONA MODA_HORATMIN_CAT MODA_HORATMAX_CAT MODA_HORARACHA_CAT LAG_MODA_HORATMIN_CAT
LAG_MODA_HORATMAX_CAT LAG_MODA_HORARACHA_CAT,
NGRUPOS=4,SINICIO=12345,SFINAL=12354,OBJETIVO=TASAFALLOS,MAXBRANCH=2,LEAFSIZE=10,MINCATSIZE=20,MINOBS
=30,
SHRINK=0.01,ITERACIONES=500,MAXDEPTH=15);
DATA TRABAJO.GB4;SET FINAL;MODELO="GB04"; MODEL='GRADIENT BOOSTING';RUN;

/*##########################################################################################################
                                                COMPARACIÓN FINAL
##########################################################################################################*/

DATA TRABAJO.COMPARACION_FINAL;
SET TRABAJO.REG2 TRABAJO.RED16 TRABAJO.BAG17 TRABAJO.RF9 TRABAJO.GB4; RUN;

ods graphics /height=600 width=1200 imagemap;

PROC BOXPLOT DATA=COMPARACION_FINAL;
PLOT MEDIA*MODEL;
RUN;
```

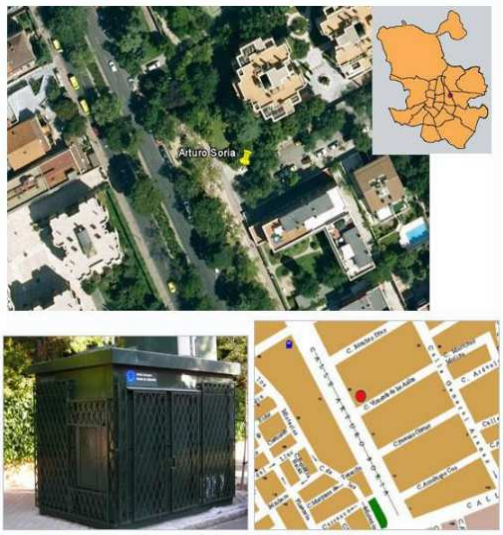
ANEXO 3: Estaciones de medición de calidad del aire en Madrid

ESTACIÓN: Plaza de España	CÓDIGO: 28079004
Dirección: C/ Princesa esq. Plaza de España	
Longitud: 3º 42' 44.40"W	Latitud: 40º 25' 26.37"N — Altitud: 637 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Dióxido de azufre, Monóxido de carbono, Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Humedad relativa, Niveles sonoros	Parámetros meteorológicos: Velocidad del viento, Dirección del viento, Temperatura media, Humedad relativa, Precipitación

ESTACIÓN: Escuelas Aguirre	CÓDIGO: 28079008
Dirección: Entre c/ Alcalá y c/ O'Donell	
Longitud: 3º 40' 56.35" W	Latitud: 40º 25' 17.63" N — Altitud: 672 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Dióxido de azufre, Monóxido de carbono, Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Partículas PM10, Partículas PM2.5, Ozono, Benceno, Hidrocarburos totales, Niveles sonoros	

ESTACIÓN: Ramón y Cajal	CÓDIGO: 28079011
Dirección: Avda. Ramón y Cajal esq. c/ Príncipe de Vergara	
Longitud: 3º 40' 38.47" W	Latitud: 40º 27' 05.30" N — Altitud: 708 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Benceno	

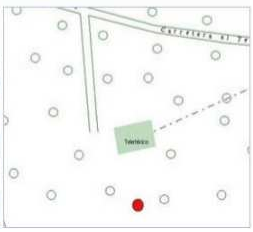
ESTACIÓN: Arturo Soria	CÓDIGO: 28079016
Dirección: C/ Arturo Soria esq. C/ Vizconde de los Asilos	
Longitud: 3º 38' 21.24" W	Latitud: 40º 26' 24.17" N — Altitud: 698 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Monóxido de carbono, Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Ozono, Niveles sonoros	Parámetros meteorológicos: Precipitación



ESTACIÓN: Villaverde	CÓDIGO: 28079017
Dirección: C/ Juan Peñalver.	
Longitud: 3º 42' 47,98" W	Latitud: 40º 20' 49,56" N — Altitud: 601 mts.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Dióxido de azufre; Óxidos de nitrógeno totales; Dióxido de nitrógeno; Monóxido de nitrógeno; Ozono	

ESTACIÓN: Farolillo	CÓDIGO: 28079018
Dirección: C/ Farolillo esq. C/ Ervigio	
Longitud: 3º 43' 54,60" W	Latitud: 40º 23' 41,20" N — Altitud: 581 m.
Tipo de estación: Urbana de fondo	
Contaminantes medidos: Dióxido de azufre; Monóxido de carbono; Óxidos de nitrógeno totales; Dióxido de nitrógeno; Monóxido de nitrógeno; Partículas PM10; Partículas PM2.5; Ozono; Benceno; Niveles sonoros	Parámetros meteorológicos: Temperatura media; Precipitación

ESTACIÓN: Casa de Campo	CÓDIGO: 28079024
Dirección: Casa de Campo. Terminal del Teleférico	
Longitud: 3º 44' 50,44" W	Latitud: 40º 25' 09,68" N — Altitud: 645 m.
Tipo de estación: Suburbana	
Contaminantes medidos: Dióxido de azufre; Monóxido de carbono; Óxidos de nitrógeno totales; Dióxido de nitrógeno; Monóxido de nitrógeno; Partículas PM10; Partículas PM2.5; Ozono; Benceno	Parámetros meteorológicos: Velocidad del viento; Dirección del viento; Precipitación; Presión barométrica; Humedad relativa; Radiación solar; Índice de radiación UVA; Temperatura media



ESTACIÓN: Barajas Pueblo	CÓDIGO: 28079027
Dirección: C/ Júpiter, 21	
Longitud: 3º 34' 48,10" W	Latitud: 40º 28' 36,94" N — Altitud: 631 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Ozono; Óxidos de nitrógeno totales; Dióxido de nitrógeno; Monóxido de nitrógeno; Hidrocarburos totales; Niveles sonoros	

ESTACIÓN: Plaza del Carmen		CÓDIGO: 28079035
Dirección: Plaza del Carmen esq. Tres Cruces		
Longitud	Latitud	Altitud
3º 42' 11,42'' W	40º 25' 09,15'' N	657 m.
Tipo de estación		
Urbana de Fondo		
Contaminantes medidos		
Dióxido de azufre Monóxido de carbono Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Ozono Niveles sonoros		

ESTACIÓN: Moratalaz		CÓDIGO: 28079036
Dirección: Avda. Moratalaz esq. Camino de Vinateros		
Longitud	Latitud	Altitud
3º 38' 43,06'' W	40º 24' 28,64'' N	671 m.
Tipo de estación		
Urbana de tráfico		
Contaminantes medidos	Parámetros meteorológicos	
Dióxido de azufre Monóxido de carbono Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Partículas PM10 Niveles sonoros	Precipitación	

ESTACIÓN: Cuatro Caminos		CÓDIGO: 28079038
Dirección: Avda. Pablo Iglesias esq. c/ Marqués de Lema		
Longitud	Latitud	Altitud
3º 42' 25,66'' W	40º 26' 43,95'' N	699 m.
Tipo de estación		
Urbana de tráfico		
Contaminantes medidos	Parámetros meteorológicos	
Dióxido de azufre Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Partículas PM10 Partículas PM2.5 Benceno Niveles sonoros	Temperatura media Precipitación	

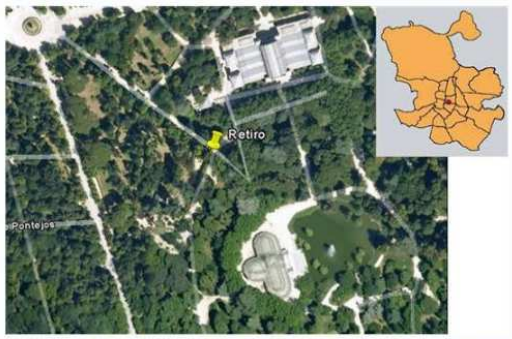
ESTACIÓN: Barrio del Pilar		CÓDIGO: 28079039
Dirección: Avd. Betanzos esq. C/ Monforte de Lemos		
Longitud	Latitud	Altitud
3º 42' 41,55'' W	40º 28' 41,62'' N	673 m.
Tipo de estación		
Urbana de tráfico		
Contaminantes medidos	Parámetros meteorológicos	
Monóxido de carbono Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Ozono Niveles sonoros	Precipitación	

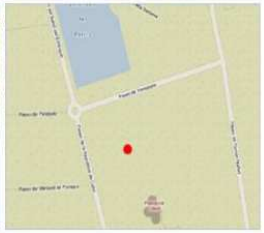
ESTACIÓN: Vallecas	CÓDIGO: 28079040
Dirección: C/ Arroyo del Olivar esq. C/ Río Grande	
Longitud: 3º 39' 05,48'' W	Latitud: 40º 23' 17,34'' N — Altitud: 677 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Dióxido de azufre, Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Partículas PM10, Niveles sonoros	Parámetros meteorológicos: Precipitación

ESTACIÓN: Méndez Álvaro	CÓDIGO: 28079047
Dirección: C/Juan de Mariana - Pza. Amanecer Mendez Alvaro	
Longitud: 3º 41' 12'' O	Latitud: 40º 23' 53'' N — Altitud: 609 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Partículas PM10, Partículas PM2.5	

ESTACIÓN: Castellana	CÓDIGO: 28079048
Dirección: C/ Jose Gutierrez Abascal	
Longitud: 3º 41' 25'' O	Latitud: 40º 26' 23'' N — Altitud: 676 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Óxidos de nitrógeno totales, Monóxido de nitrógeno, Dióxido de nitrógeno, Partículas PM10, Partículas PM2.5	

ESTACIÓN: Parque del Retiro	CÓDIGO: 28079049
Dirección: Paseo Venezuela - Casa de Vacas	
Longitud: 3º 40' 57'' O	Latitud: 40º 24' 52'' N — Altitud: 662 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Ozono	

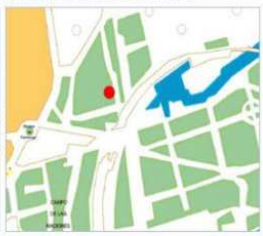




ESTACIÓN: Sanchinarro		CÓDIGO: 28079057
Dirección: C/ Princesa de Éboli - C/ María Tudor		
Longitud	Latitud	Altitud
3º 39' 37,8" O	40º 29' 39,1" N	700 m.
Tipo de estación		
Urbana de Fondo		
Contaminantes medidos		Parámetros meteorológicos
Dióxido de azufre Monóxido de carbono Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Partículas PM10		Temperatura media

ESTACIÓN: El Pardo		CÓDIGO: 28079058
Dirección: Avda. de la Guardia		
Longitud	Latitud	Altitud
3º 46' 28,6" O	40º 31' 5" N	616 m.
Tipo de estación		
Suburbana		
Contaminantes medidos		
Monóxido de nitrógeno Dióxido de nitrógeno Óxidos de nitrógeno totales Ozono Hidrocarburos totales Niveles sonoros		

ESTACIÓN: Juan Carlos I		CÓDIGO: 28079059
Dirección: Parque Juan Carlos I (frente oficinas mantenimiento)		
Longitud	Latitud	Altitud
3º 36' 32" W	40º 27' 54" N	669 m.
Tipo de estación		
Suburbana		
Contaminantes medidos		Parámetros meteorológicos
Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Ozono		Velocidad del viento Dirección del viento Precipitación Presión barométrica Humedad relativa Índice de radiación UVA Radiación solar Temperatura media

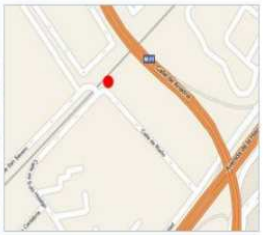


ESTACIÓN: Tres Olivos		CÓDIGO: 28079060
Dirección: Plaza de Tres Olivos		
Longitud	Latitud	Altitud
3º 41' 23' O	40º 30' 02" N	715 m.
Tipo de estación		
Urbana de Fondo		
Contaminantes medidos		Parámetros meteorológicos
Óxidos de nitrógeno totales Dióxido de nitrógeno Monóxido de nitrógeno Partículas PM10 Ozono		Precipitación Presión barométrica

ESTACIÓN: Plaza Castilla	CÓDIGO: 28079050
Dirección: Plaza Castilla - Canal	
Longitud: 3º 41' 19" O	Latitud: 40º 27' 56" N — Altitud: 728 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Partículas PM10, Partículas PM2.5	Parámetros meteorológicos: Temperatura media

ESTACIÓN: Ensanche de Vallecas	CÓDIGO: 28079054
Dirección: Avda. La Gavia - Avda.Las Suertes	
Longitud: 3º 36' 43" O	Latitud: 40º 22' 22" N — Altitud: 630 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Ozono	Parámetros meteorológicos: Dirección del viento, Velocidad del viento, Precipitación, Radiación solar, Temperatura media

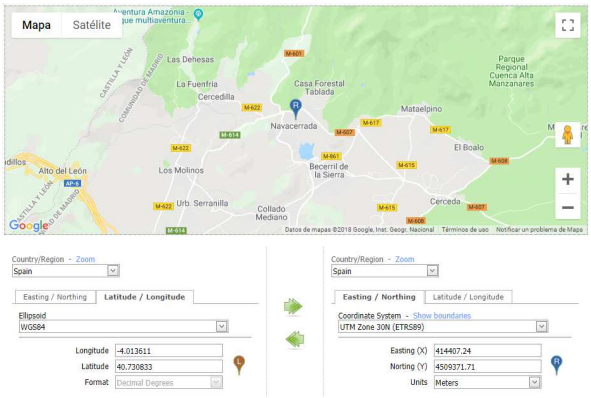
ESTACIÓN: Urbanización Embajada (Barajas)	CÓDIGO: 28079055
Dirección: C/ Riaño	
Longitud: 3º 34' 50" W	Latitud: 40º 27' 45" N — Altitud: 619 m.
Tipo de estación: Urbana de Fondo	
Contaminantes medidos: Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Partículas PM10, Benceno, Hidrocarburos totales	



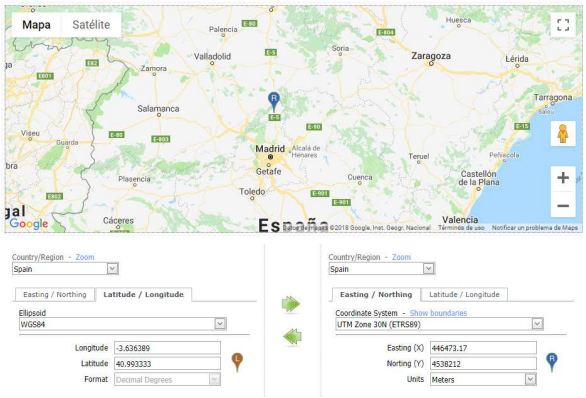
ESTACIÓN: Plaza Fernández Ladreda	CÓDIGO: 28079056
Dirección: Pza. Fernandez Ladreda - Avda.Oporto	
Longitud: 3º 43' 7" W	Latitud: 40º 23' 05" N — Altitud: 605 m.
Tipo de estación: Urbana de Tráfico	
Contaminantes medidos: Monóxido de carbono, Óxidos de nitrógeno totales, Dióxido de nitrógeno, Monóxido de nitrógeno, Ozono	Parámetros meteorológicos: Velocidad del viento, Dirección del viento, Temperatura media, Humedad relativa, Precipitación

ANEXO 4: Estaciones AEMET

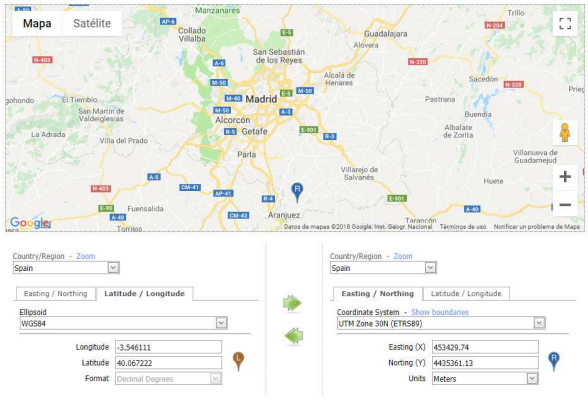
Navacerrada	
Latitud	40º 43' 51" N
Longitud	4º 0' 49" O
Altitud	1207m



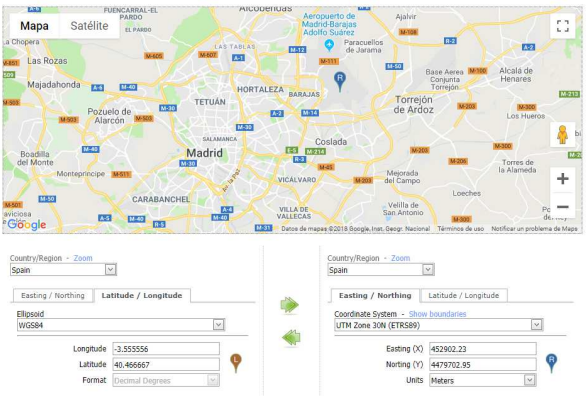
Buitrago de Lozoya	
Latitud	40º 59' 36" N
Longitud	3º 38' 11" O
Altitud	975m



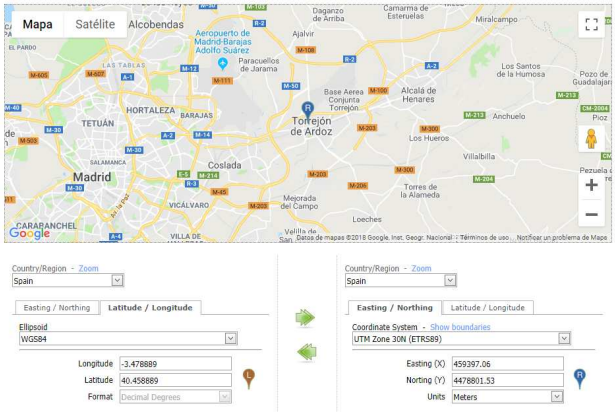
Aranjuez	
Latitud	40º 4' 42" N
Longitud	3º 32' 46" O
Altitud	540m



Madrid Aeropuerto	
Latitud	40º 28' 0" N
Longitud	3º 33' 20" O
Altitud	609m



Torrejón de Ardoz	
Latitud	40º 27' 32" N
Longitud	3º 28' 44" O
Altitud	586m

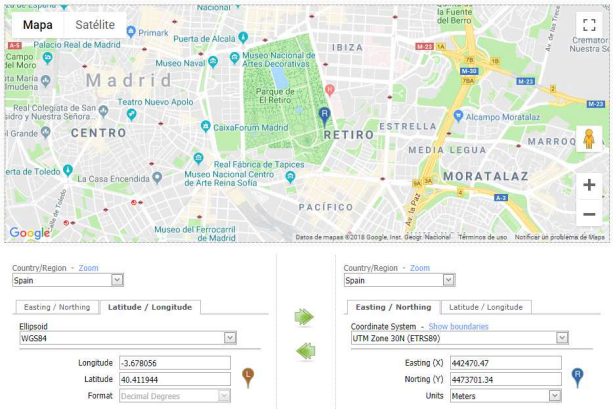


Colmenar Viejo	
Latitud	40º 41' 46" N
Longitud	3º 45' 52" O
Altitud	1004m

Somosierra	
Latitud	41º 7' 58" N
Longitud	3º 34' 51" O
Altitud	1433m

Ciudad Universitaria	
Latitud	40º 27' 6" N
Longitud	3º 43' 27" O
Altitud	664m

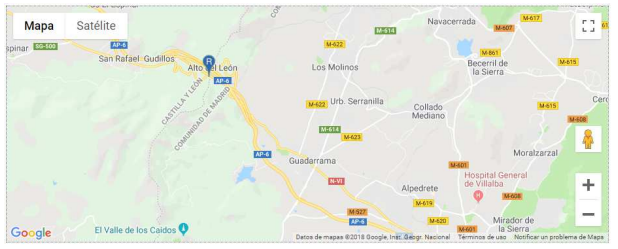
Retiro	
Latitud	40º 24' 43" N
Longitud	3º 40' 41" O
Altitud	667m



Cuatro Vientos	
Latitud	40º 22' 32" N
Longitud	3º 47' 10" O
Altitud	690m

Getafe	
Latitud	40º 18' 14" N
Longitud	3º 43' 45" O
Altitud	622m

Puerto Alto del León	
Latitud	40º 42' 23" N
Longitud	4º 8' 31" O
Altitud	1532m



Robledo de Chavela	
Latitud	40º 30' 6" N
Longitud	4º 14' 21" O
Altitud	901m