TFM
Calidad del Aire en Madrid

XI Máster Data Science - KSchool
Miguel Ángel Fernández González

Tabla de contenido

1 Introducción 3
2 Descripción de los datos 6
3 Estudio de variables 10
3.1 PM10 10
3.2 SO2 11
3.3 NO2 12
3.4 CO 14
3.5 O3 15
3.6 Correlación entre las variables 16
4 Metodología 17
5 Algoritmos empleados 19
5.1 Modelo VAR (Vector Auto Regression) 19
5.2 Modelo Machine Learning Random Forest 19
6 Conclusiones y próximos pasos 21

1 Introducción

El aumento de los niveles de contaminación en las grandes ciudades ha creado una gran preocupación acerca de cómo puede afectar a la salud de la población, de manera que en algunas de ellas se han tomado medidas drásticas como limitar la circulación de vehículos para reducir la concentración de contaminantes.
Distintos estudios han demostrado una relación entre el aumento de los niveles de contaminación y el aumento de la mortalidad en las ciudades con una concentración de contaminantes elevada.
Se entiende por contaminante como aquella sustancia que es ajena a la composición normal de la atmósfera y que permanece en ella durante un tiempo.
Para este estudio nos vamos a centrar en la ciudad de Madrid y en las sustancias contaminantes más dañinas para la salud, descritas según el Portal Web de Calidad del Aire del Ayuntamiento de Madrid:

- SO2
- PM10
- NO2
- CO
- O3

DIÓXIDO DE AZUFRE (SO2)

Es un gas incoloro, no inflamable. Posee un olor fuerte e irritante en altas concentraciones. Se origina por la combustión de carburantes con cierto contenido en azufre (carbón, fuel) y la fundición de minerales ricos en sulfatos. Se genera principalmente por la industria (incluyendo las termoeléctricas), seguido de los vehículos a motor.

Se ha comprobado que la presencia de partículas en suspensión y SO2 aumentan la frecuencia de bronquitis crónicas, de catarros y de dificultades respiratorias en adultos. Cuando las concentraciones de SO2 superan los 500 microgramos por metro cúbico de aire, como promedio de 24 horas, aumenta la mortalidad en la población, especialmente en individuos con procesos cardíacos o pulmonares. Con promedios diarios de 250 microgramos por metro cúbico de SO2 y de humos se produce el empeoramiento en los enfermos con afecciones pulmonares. Sin embargo, las concentraciones de partículas en suspensión y de SO2 que afectan a la salud pueden variar de un lugar a otro según las características físicas y químicas de las partículas y en función de la presencia en el aire de otros contaminantes con los que se puedan combinar.

PARTÍCULAS EN SUSPENSIÓN (PM10)

El material particulado es una mezcla compleja de componentes con características químicas y físicas diversas, formadas a partir de otros contaminantes primarios e, incluso, a partir de elementos naturales. En las ciudades europeas, este material se genera en procesos de combustión provenientes tanto de los sistemas de calefacción de edificios como de las emisiones generadas por el tráfico rodado, con una especial importancia en los motores de ciclo diésel con tecnologías de motor anteriores al año 2000. Además, en el caso de España, por su situación geográfica, se pueden encontrar aportes de origen natural como pueden ser las procedentes del desierto del Sáhara. El término PM10 se refiere a partículas en suspensión

con un diámetro aerodinámico de hasta 10 µm, comprendiendo las fracciones fina y gruesa, y PM2.5 se refiere a partículas en suspensión con un diámetro aerodinámico de hasta 2.5 µm.

La composición de las partículas en suspensión puede ser una mezcla muy variada, sin embargo, se suelen clasificar según su tamaño, y no tanto por su origen o composición. Las partículas de diámetro aerodinámico igual o inferior a 10 µm (PM10) suelen quedarse en la garganta y sus proximidades. Las que tienen un diámetro igual o inferior a 2,5 µm (PM2,5) pueden llegar hasta los pulmones. Finalmente, las partículas ultrafinas, con un diámetro igual o inferior a 0,1 µm, pueden llegar a pasar del alvéolo pulmonar a la sangre. Sus efectos sobre la salud afectan exclusivamente al sistema respiratorio, provocando normalmente tos y cierta dificultad para respirar. En casos extremos y en población de riesgo, pueden agravar el asma, causar daños al pulmón (incluyendo la disminución de la función del pulmón y enfermedades respiratorias de por vida) e, incluso, la muerte prematura en individuos con patologías previas del corazón y del pulmón.

DIÓXIDO DE NITRÓGENO (NO2)

El dióxido de nitrógeno (NO2) es un contaminante indicador de actividades de transporte, especialmente el tráfico rodado. Lo emiten directamente los vehículos, especialmente los diésel (emisiones directas o «primarias»), pero se produce también en la atmósfera a partir de las emisiones de monóxido de nitrógeno (NO) de los vehículos; por un proceso químico, dicho gas se transforma en NO2 (contaminante «secundario»).

La mayor parte de los estudios relativos a los efectos de los óxidos de nitrógeno se han ocupado del NO2 ya que es el más tóxico. Sus efectos se producen casi enteramente en el tracto respiratorio: una concentración media de 190 microgramos de NO2 por metro cúbico de aire, superada el 40% de los días, aumenta la frecuencia de infecciones de las vías respiratorias en la población expuesta.

MONÓXIDO DE CARBONO (CO)

El monóxido de carbono es un contaminante primario indicador del tráfico rodado. Es un gas incoloro, inodoro e insípido. Su presencia se ha reducido de manera continua en los últimos años debido fundamentalmente a los cambios tecnológicos en los vehículos de motor que son los principales emisores de este contaminante.

Una concentración elevada de CO en el aire es peligrosa: al inhalarlo se combina con la hemoglobina de la sangre formando carbooxihemoglobina, lo que reduce la capacidad de la sangre para transportar oxígeno desde los pulmones hasta los tejidos. Se ha comprobado que una saturación de carbooxihemoglobina por encima del 10% afecta a la función psicomotora, causando cansancio, cefaleas y alteraciones de la coordinación. Por encima del 5% de saturación se producen cambios funcionales cardíacos y pulmonares y se aumenta el umbral visual, pero no se han encontrado efectos significativos con una concentración inferior al 2%.

EL OZONO (O3)

El ozono es un contaminante secundario que se forma a partir de una serie de contaminantes precursores cuando encuentran un nivel de insolación suficiente. Las moléculas de este gas azulado y picante están formadas por tres átomos de oxígeno.

Presenta dos propiedades que marcan sus interacciones con la vida de nuestro planeta: su fuerte absorción de la radiación ultravioleta y su gran poder oxidante.

La primera hace que su presencia en la estratosfera sea imprescindible como filtro para evitar que lleguen a la superficie del planeta altos niveles de radiación ultravioleta que resultarían catastróficos para todos los seres vivos. Por eso existen tantas campañas y esfuerzos para evitar el deterioro de la conocida «capa de ozono».

Sin embargo, la segunda propiedad (su alto poder oxidante), lo hace muy peligroso cuando aparece en la troposfera porque, en determinadas concentraciones, puede producir daños en nuestra salud, en la vegetación y en los materiales.

2 Descripción de los datos

Para la realización de este estudio se han tomado diferentes fuentes de datos públicas desde el año 2014 hasta el 2019. Se han obtenido por su naturaleza 4 grupos distintos de datos:

- **Calidad del Aire**
  Estos datos se han obtenido del Portal de datos abiertos del Ayuntamiento de Madrid. En estos ficheros tenemos datos horarios con las mediciones en cada una de las estaciones sobre las sustancias contaminantes en la atmósfera de la ciudad de Madrid. Se han encontrado 2 tipos de formato de archivos .txt y .csv, cada uno de los cuales con una estructura de registro diferente que se muestra en la siguiente tabla.

Columna	Descripción	Tipo Fichero
PROVINCIA	Código de la provincia	.csv y .txt
MUNICIPIO	Código del municipio	.csv y .txt
ESTACION	Código de la estación de medición	.csv y .txt
MAGNITUD	Código de la sustancia contaminante medida	.csv y .txt
PUNTO_MUESTREO	Código del muestreo compuesto por la concatenación del código de la Estación, Magnitud y la Técnica	.csv
TECNICA	Tecnica de la medición	.txt
DATOHORARIO	Indica si el dato medido es horario o no	.txt
ANO	Año en formato YY	.csv y .txt
MES	Mes en formato MM	.csv y .txt
DIA	Dia en formato DD	.csv y .txt
H01 .. H24	Hora en formato numérico	.csv y .txt
V01 .. V24	Campo que sirve para validar el dato obtenido en la medición. Si es V es válido.	.csv y .txt

Ambos formatos de archivos se han unificado en uno solo para el óptimo procesamiento de los datos.
En la preparación de los datos, para los valores inválidos se ha buscado para los valores medidos por la misma estación el valor válido más próximo, primero en horas anteriores y si no existiera en posteriores.

Los códigos de las magnitudes se detallan en la siguiente tabla.

Código	Magnitud	Fórmula	Unidad Medida
1	Dióxido de Azufre	SO2	µg/m3
6	Monóxido de Carbono	CO	mg/m3
8	Dióxido de Nitrogeno	NO2	µg/m3
10	Partículas < 10 µm	PM10	µg/m3
14	Ozono	O3	µg/m3

La red de calidad del aire del ayuntamiento de Madrid tiene las siguientes estaciones de mediciones sobre las cuales se ha realizado el estudio.

Número	Estación	Dirección	Longitud	Latitud
4	Pza. de España	Plaza de España	-3,710726	40,423282
8	Escuelas Aguirre	Entre C/ Alcalá y C/ O' Donell	-3,682319	40,421564
11	Avda. Ramón y Cajal	Avda. Ramón y Cajal esq. C/ Príncipe de Vergara	-3,677144	40,451740
16	Arturo Soria	C/ Arturo Soria esq. C/ Vizconde de los Asilos	-3,639146	40,440369
17	Villaverde	C/. Juan Peñalver	-3,713378	40,348311
18	Farolillo	Calle Farolillo - C/Ervigio	-3,731853	40,394781
24	Casa de Campo	Casa de Campo (Terminal del Teleférico)	-3,747347	40,419356
27	Barajas Pueblo	C/. Júpiter, 21 (Barajas)	-3,577770	40,473660
35	Pza. del Carmen	Plaza del Carmen esq. Tres Cruces.	-3,703717	40,418700
36	Moratalaz	Avd. Moratalaz esq. Camino de los Vinateros	-3,645780	40,407821
38	Cuatro Caminos	Avda. Pablo Iglesias esq. C/ Marqués de Lema	-3,707562	40,445749
39	Barrio del Pilar	Avd. Betanzos esq. C/ Monforte de Lemos	-3,711967	40,478064
40	Vallecas	C/ Arroyo del Olivar esq. C/ Río Grande.	-3,651603	40,388301
47	Mendez Alvaro	C/ Juan de Mariana / Pza. Amanecer Mendez Alvaro	-3,686804	40,398102
48	Castellana	C/ Jose Gutierrez Abascal	-3,690286	40,440217
49	Parque del Retiro	Paseo Venezuela- Casa de Vacas	-3,682065	40,413598
50	Plaza Castilla	Plaza Castilla (Canal)	-3,688740	40,465583
54	Ensanche de Vallecas	Avda La Gavia / Avda. Las Suertes	-3,613995	40,370673
55	Urb. Embajada	C/ Riaño (Barajas)	-3,580747	40,462531
56	Pza. Fernández Ladreda	Pza. Fernández Ladreda - Avda. Oporto	-3,718728	40,384964
57	Sanchinarro	C/ Princesa de Eboli esq C/ Maria Tudor	-3,660503	40,494208
58	El Pardo	Avda. La Guardia	-3,774611	40,518058
59	Juan Carlos I	Parque Juan Carlos I (frente oficinas mantenimiento)	-3,609072	40,465250
60	Tres Olivos	Plaza Tres Olivos	-3,689761	40,500589

Su ubicación en la ciudad puede verse en el siguiente mapa.

No todas las estaciones miden todos los contaminantes. Los contaminantes que miden cada una de las estaciones se representan en la siguiente tabla.

Número	Estación	NO2	SO2	CO	PM10	O3
4	Pza. de España	x	x	x		
8	Escuelas Aguirre	x	x	x	x	x
11	Avda. Ramón y Cajal	x				
16	Arturo Soria	x		x		x
17	Villaverde	x	x			x
18	Farolillo	x	x	x	x	x
24	Casa de Campo	x	x	x	x	x
27	Barajas Pueblo	x				x
35	Pza. del Carmen	x	x	x		x
36	Moratalaz	x	x	x	x	
38	Cuatro Caminos	x	x		x	
39	Barrio del Pilar	x		x		x
40	Vallecas	x	x		x	
47	Mendez Alvaro	x			x	
48	Castellana	x			x	
49	Parque del Retiro	x				x
50	Plaza Castilla	x			x	
54	Ensanche de Vallecas	x				x
55	Urb. Embajada	x			x	
56	Pza. Fernández Ladreda	x		x		x
57	Sanchinarro	x	x	x	x	
58	El Pardo	x				x
59	Juan Carlos I	x				x
60	Tres Olivos	x			x	x

- **Tráfico**

  Estos datos se han obtenido del Portal de datos abiertos del Ayuntamiento de Madrid. Tenemos datos de las mediciones acerca del tráfico de todas las estaciones de medida de la ciudad de Madrid, además de la localización en coordenadas de cada una de las estaciones.

  El formato de cada registro de los ficheros con las mediciones sobre el tráfico se describe en la siguiente tabla.

Columna	Descripción
idelem	Identificación única del Punto de Medida en los sistemas de control del tráfico del Ayuntamiento de Madrid.
fecha	Fecha y hora oficiales de Madrid con formato yyyy-mm-dd hh:mi:ss
identif	Identificador del Punto de Medida en los Sistemas de Tráfico (se proporciona por compatibilidad hacia atrás).
tipo_elem	Nombre del Tipo de Punto de Medida: Urbano o M30.
Intensidad	Intensidad del Punto de Medida en el periodo de 15 minutos (vehículos/hora). Un valor negativo implica la ausencia de datos.
ocupacion	Tiempo de Ocupación del Punto de Medida en el periodo de 15 minutos (%). Un valor negativo implica la ausencia de datos.
carga	Carga de vehículos en el periodo de 15 minutos. Parámetro que tiene en cuenta intensidad, ocupación y capacidad de la vía y establece el grado de uso de la vía de 0 a 100. Un valor negativo implica la ausencia de datos.
vmed	Velocidad media de los vehículos en el periodo de 15 minutos (Km./h). Sólo para puntos de medida interurbanos M30. Un valor negativo implica la ausencia de datos.
error	Indicación de si ha habido al menos una muestra errónea o sustituida en el periodo de 15 minutos. N: no ha habido errores ni sustituciones E: los parámetros de calidad de alguna de las muestras integradas no son óptimos. S: alguna de las muestras recibidas era totalmente errónea y no se ha integrado.
periodo_integracion	Número de muestras recibidas y consideradas para el periodo de integración.

  En la preparación de los datos, para todas las estaciones de medición de tráfico se ha calculado la distancia en metros respecto a las estaciones de Calidad del Aire de las

que hacemos el estudio. Para el cálculo de esta distancia hemos utilizado la Vicenty's formulae. (https://en.wikipedia.org/wiki/Vincenty%27s_formulae)
De esta forma, hemos decidido quedarnos con la media de las mediciones de tráfico de las 4 estaciones más próximas a cada estación de medición de la Calidad del Aire. Se hace así por considerar que el tráfico más próximo es el más influyente a cada medición de la Calidad del Aire.
Para cada uno de los valores inválidos en estos datos se ha decidido cargarlos con el valor más próximo anterior en su misma estación de medición.

- **Calendario Laboral**
  Estos datos se han obtenido del Portal de datos abiertos del Ayuntamiento de Madrid.

3 Estudio de variables

Para realizar el estudio de variables se presentan el estadístico descriptivo de la media agrupado por meses y por días de la semana, además se va a realizar un estudio acerca de en qué días se han vulnerado los límites legales que marca la legislación durante los años 2014 a marzo del 2019.

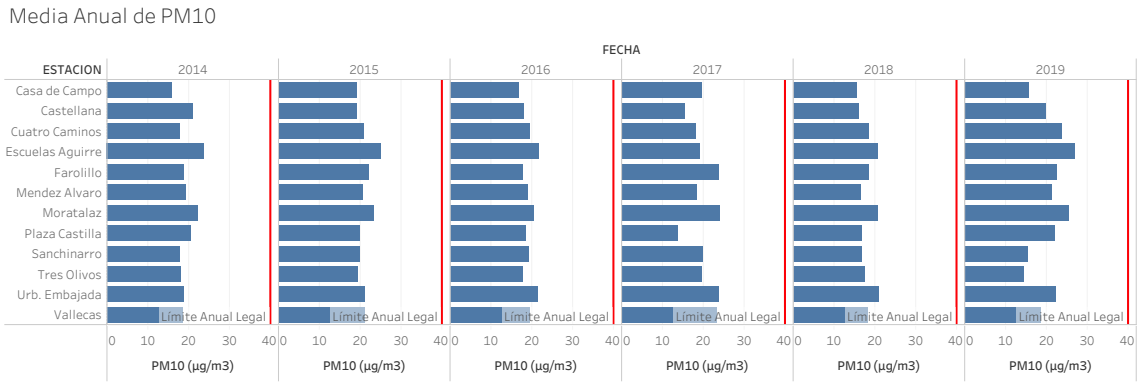
El objetivo de estos límites es el preservar la calidad del aire para evitar, prevenir o reducir los efectos nocivos que puedan ocasionar a la salud de las personas y el medio ambiente.
Se tienen en cuenta dependiendo del contaminante los valores límite en distintos periodos de tiempo: octohorario, horario, diario, anual.

Para los contaminantes sobre los que tratara este estudio los límites legales para la protección de la salud de las personas son:

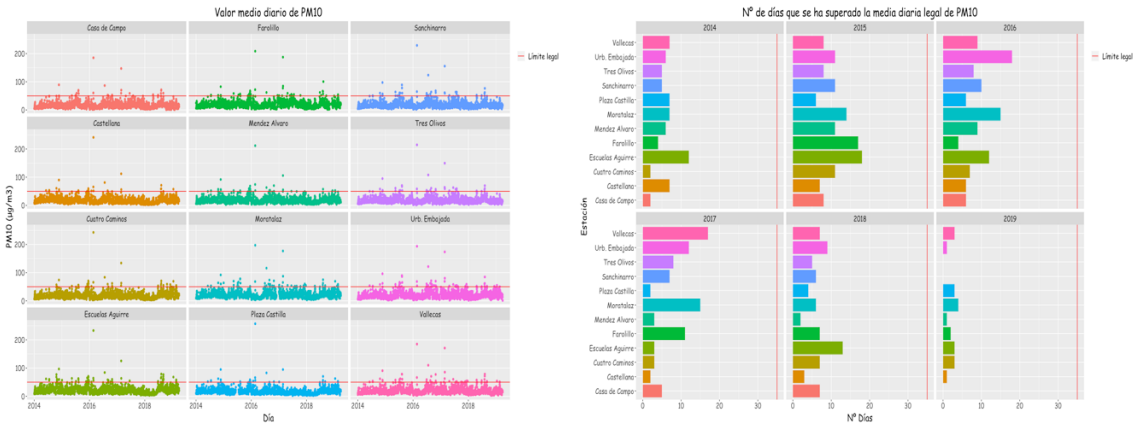
Compuesto	Período	Valor límite	Nº máximo de superaciones
PM10	Media anual	40 µg/m3	
	Media diaria	50 µg/m3	35 días/año
SO2	Media diaria	125 µg/m3	3 días/año
	Media horaria	350 µg/m3	24 horas/año
NO2	Media anual	40 µg/m3	
	Media horaria	200 µg/m3	18 horas/año
CO	Máxima diaria de las medias móviles octohorarias	10 mg/m3	
O3	Máxima diaria de las medias móviles octohorarias	120 µg/m3	25 días/año

3.1 PM10

Para el estudio de esta variable hemos calculado la media anual de este contaminante en cada una de las estaciones de medición.



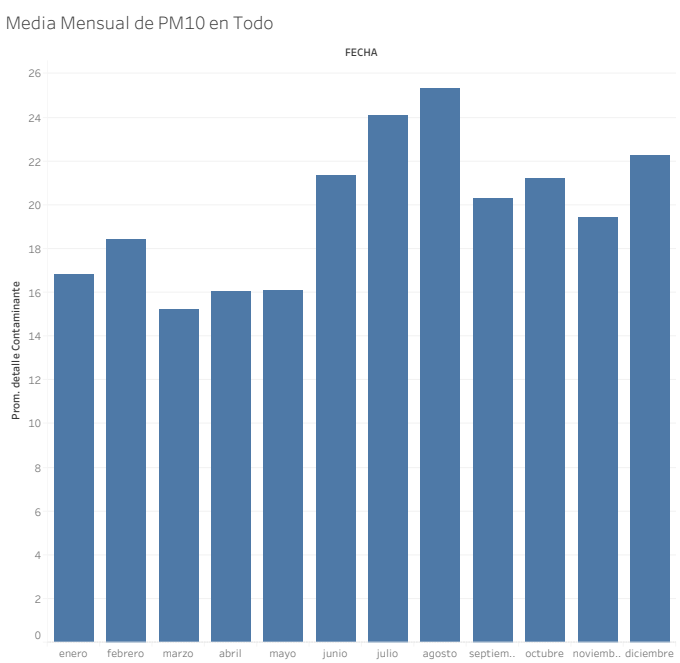
Como puede verse en estas gráficas, no se ha superado el límite anual legal (40 µg/m3) para las medias anuales en ningún año y para ninguna estación.
Para las medias diarias en cada una de las estaciones hemos obtenido las siguientes gráficas.

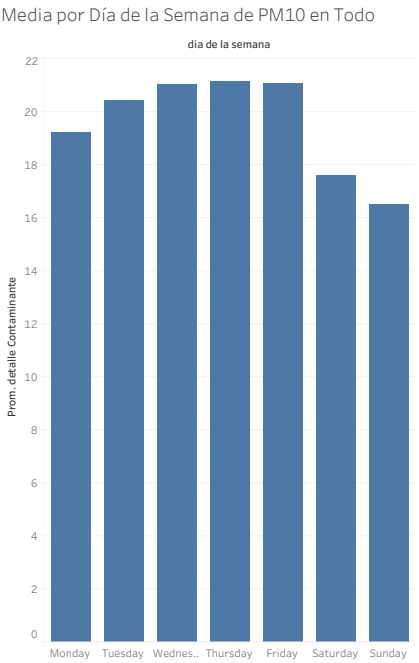


Puede verse que en todas las estaciones se ha superado en algún momento el límite diario, pero se ha cumplido la legislación durante todos estos años ya que no se ha superado el nº de días máximo.

Para la media mensual de PM10 puede apreciarse que son mayores en los meses de verano, lo que podría verse influido por la intrusión de partículas procedentes del desierto del Sahara que en esos meses suele ser habitual.

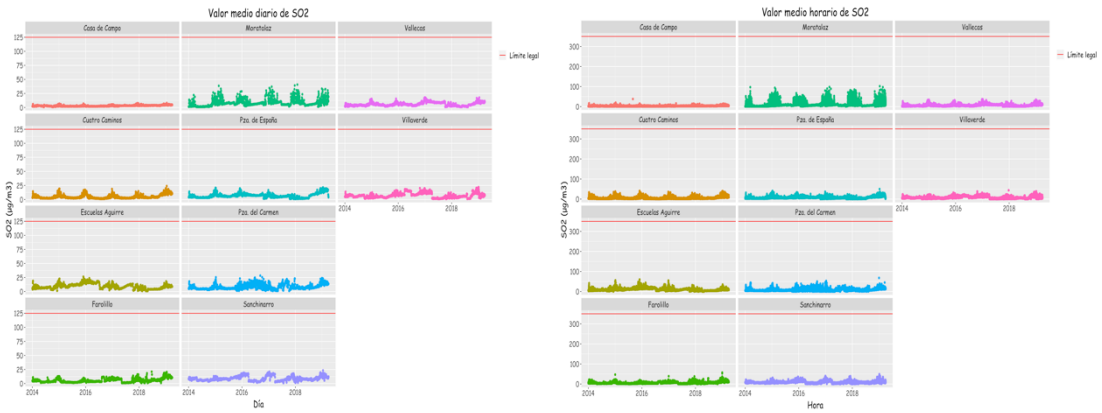
En las medias diarias se aprecia una clara disminución en el fin de semana que podría ser explicado por una disminución del tráfico en esos días.





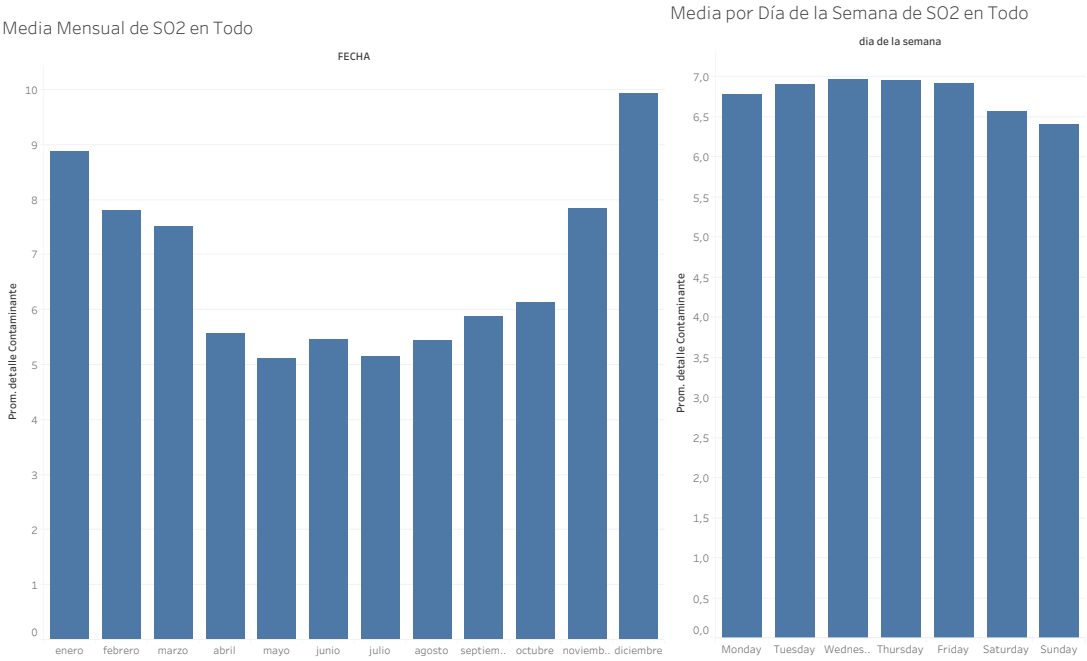
3.2 SO2

Para el estudio de esta variable hemos calculado la media diaria y horaria, y el nº de días al año que supera ambas medias en cada una de las estaciones.
Tanto para las medias diarias como horarias, puede verse que de una manera muy holgada no se ha superado el límite legal en ninguna de las estaciones y por lo tanto puede decirse que este contaminante no supone una amenaza para la salud de las personas.



Los meses con las medias mensuales más altas son los meses de invierno, algo que teniendo en cuenta que este contaminante se origina principalmente por la combustión de carburantes con cierto contenido en azufre (carbón, fuel), podría ser explicado por el uso de las calefacciones durante estos meses.

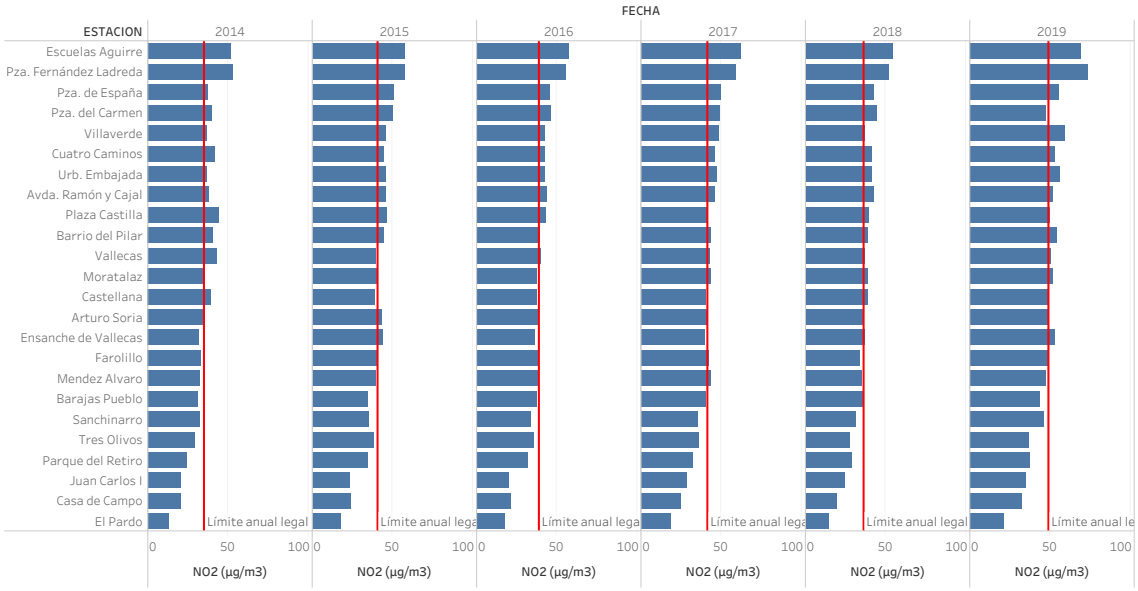
En las medias diarias, aunque se aprecia una disminución en el fin de semana, no es algo altamente reseñable.



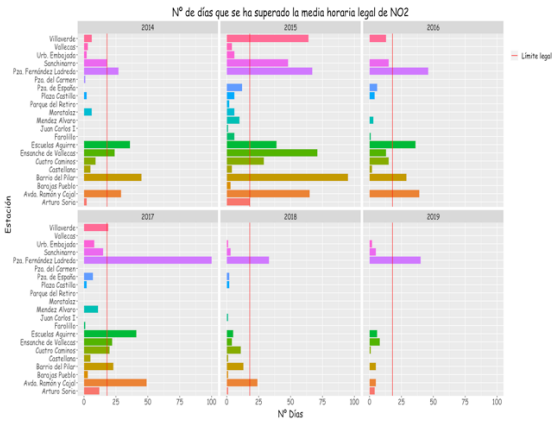
3.3 NO2

Para el estudio de esta variable hemos calculado la media anual y horaria, y el nº de días al año que supera la media horaria en cada una de las estaciones.

Media Anual en NO2



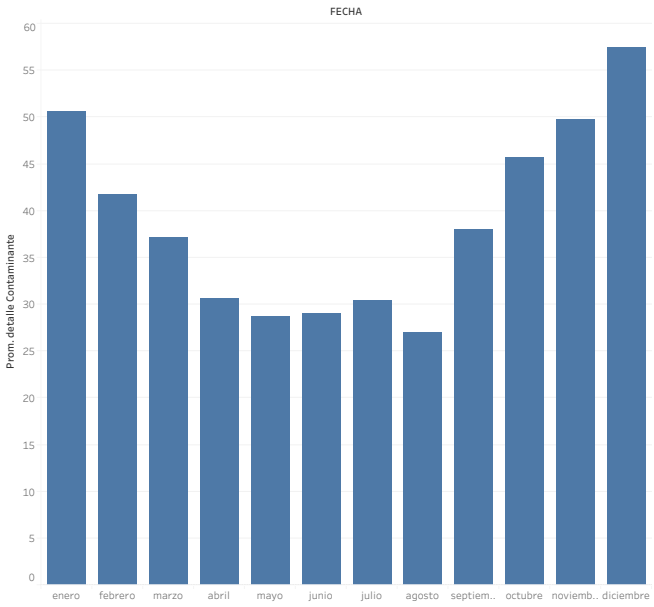
Como puede verse son varias las estaciones que superan la media anual. Esto puede ser debido al tráfico rodado y en especial a los vehículos que consumen diésel. Las estaciones que se mantienen con las medias más bajas son precisamente aquellas que se encuentras en las zonas verdes de la ciudad y con menos tráfico. (Retiro, Juan Carlos I, Casa de Campo y El Pardo)



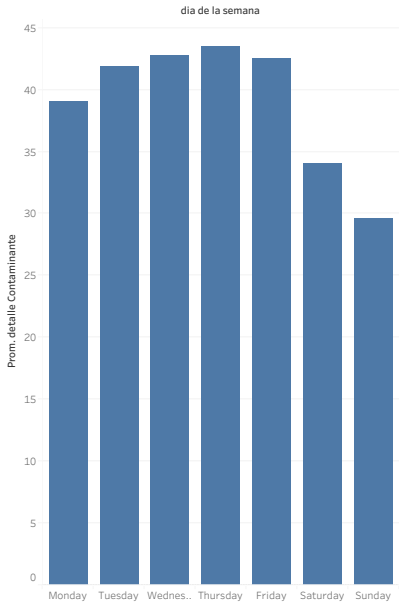
Ha habido un claro descenso en el nº de días que se superan la media horaria en el último año 2018 y en lo que llevamos del 2019 hasta el mes de marzo.

Para este contaminante se puede apreciar un bajón considerable en los meses de primavera y verano, además de en los fines de semana

Media Mensual de NO2 en Todo

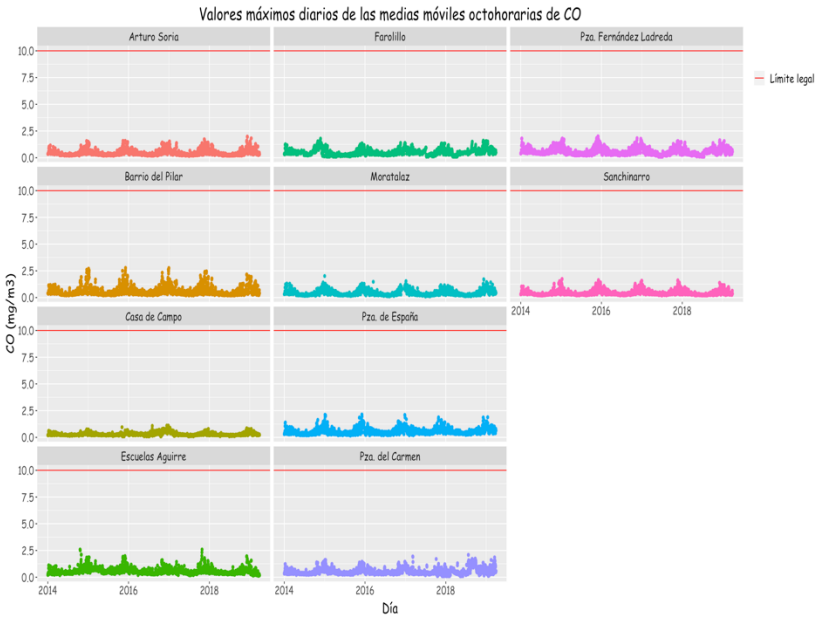


Media por Día de la Semana de NO2 en Todo



3.4 CO

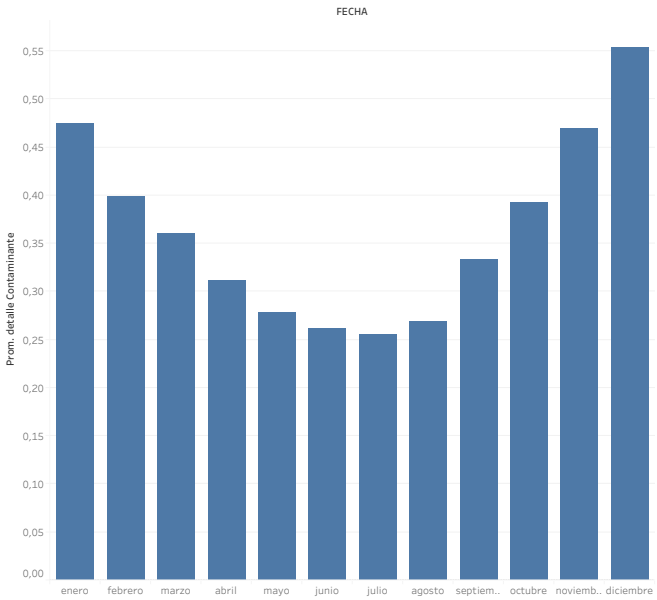
Para el estudio de esta variable hemos calculado la máxima diaria de las medias móviles octohorarias en cada una de las estaciones.



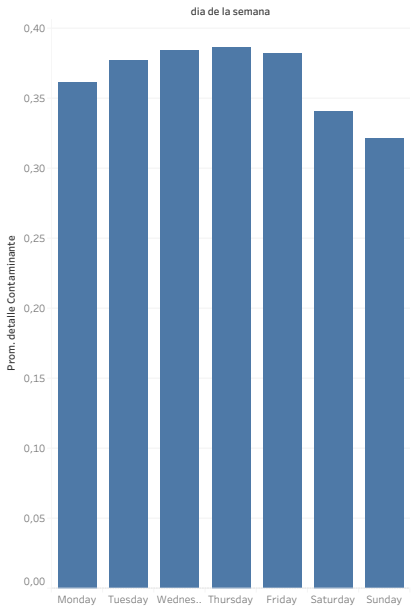
Como podemos ver este contaminante tampoco representa un peligro para la salud de las personas ya que se mantiene lejos de superar los límites legales en todas las estaciones.

Para este contaminante, al igual que para el NO2, se parecía una disminución clara tanto en los meses de primavera y verano como en el fin de semana.

Media Mensual de CO en Todo

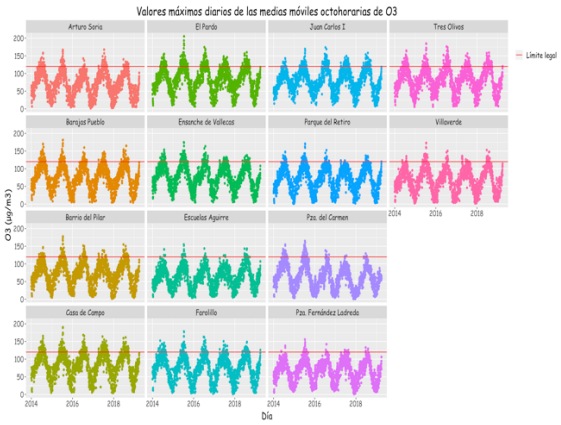


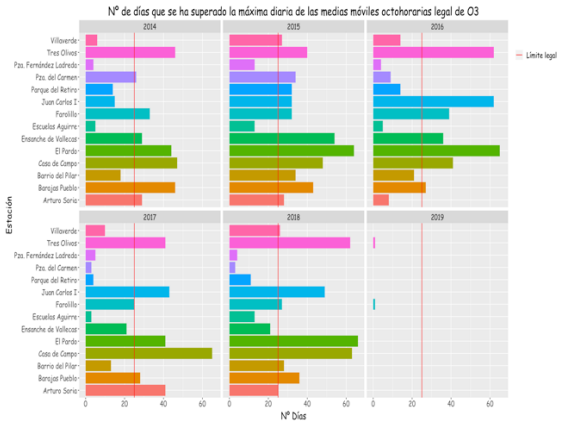
Media por Día de la Semana de CO en Todo



3.5 O3

Para el estudio de esta variable hemos calculado la máxima diaria de las medias móviles octohorarias, y el nº de días que supera la máxima diaria en cada una de las estaciones.

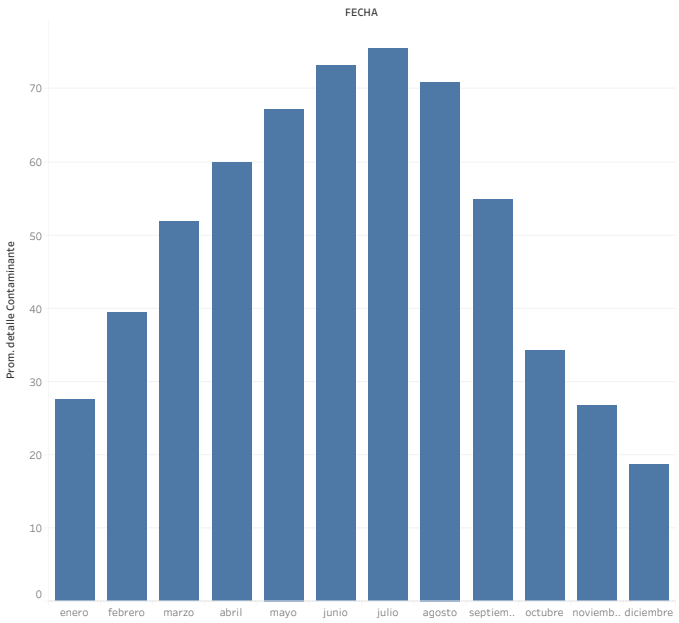




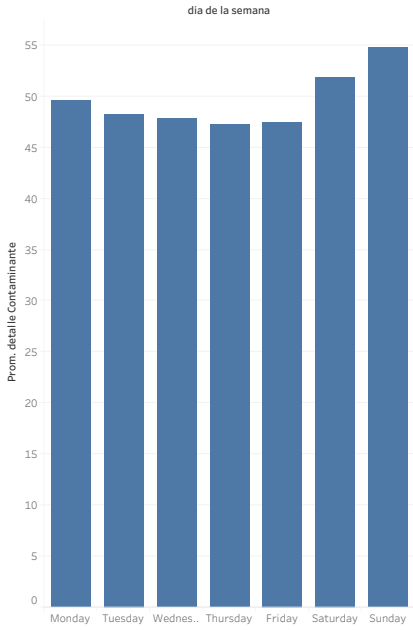
Como puede verse para este contaminante, en todas las estaciones se supera el máximo diario permitido, y en muchas incluso también el número de días en que puede superarse el límite legal. Al contrario que para el NO2, este contaminante tiene niveles más altos en las estaciones que se encuentran en las zonas verdes de la ciudad y además presenta los picos en los meses de temperaturas más altas.

Debido a que este gas es un contaminante que se forma a partir de otros contaminantes cuando encuentran un nivel de insolación suficiente, se puede apreciar claramente como en los meses con más altas temperaturas este contaminante es más abundante, dándose la media más alta en el mes de julio.

Media Mensual de O3 en Todo

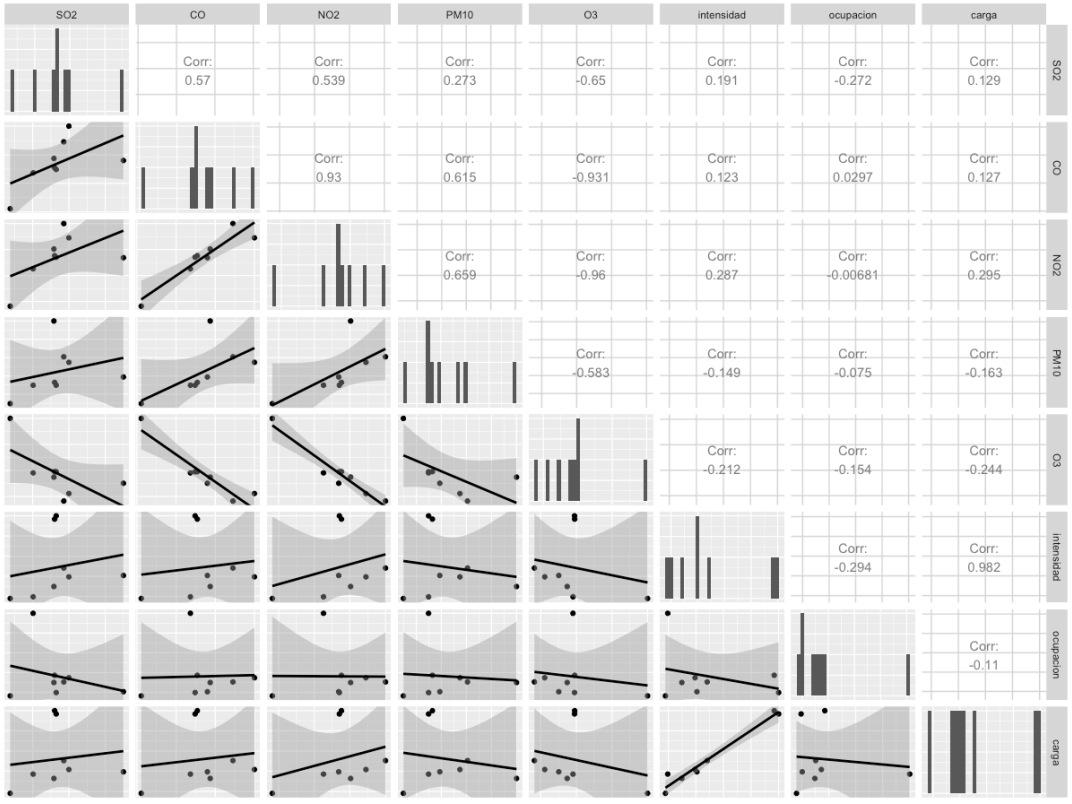


Media por Día de la Semana de O3 en Todo



3.6 Correlación entre las variables

Se ha implementado una matriz de correlación de variables a través de la función ggpairs en R para ver la relación entre las variables de la Calidad del Aire y las de Tráfico, y como puede verse las variables que mantienen una mayor relación son NO2 con el CO, O3 con el CO, O3 con NO2 y la carga con la intensidad.



4 Metodología

Para el desarrollo de este trabajo se ha utilizado el lenguaje de programación de R en RStudio y de Python en Jupyter Notebook en el entorno de trabajo de Anaconda que proporciona multitud de paquetes para la ciencia de datos.

Para la visualización he utilizado la librería ggplot2 en R y además he implementado una interfaz en Tableau con el que poder interactuar como usuario con diferentes Dashboards tales que permitan filtrar y analizar los resultados obtenidos en este estudio.

El proyecto se ha publicado en github que nos ha proporcionado un sistema de versiones adecuado para el desarrollo del mismo.
El repositorio público donde he alojado el proyecto es:

https://github.com/mafdezglez/Air_Quality_Madrid

Para ejecutar el proyecto de inicio a fin habría que ejecutar los siguientes scripts en el orden indicado:

0a_Cargar_Librerias.R
Con este script se cargan todas las librerías necesarias para ejecutar el resto de scripts R.

0b_DescargarDatos_Trafico.R
Los datos necesarios se encuentran en la carpeta /rawdata, todos excepto los de Tráfico que no se han podido alojar en github por exceder el tamaño permitido.

Para poder descargar estos archivos necesarios será suficiente con ejecutar este script que los descarga de Google Drive.

Para ello, cuando se ejecute el script le hará la siguiente pregunta:
Use a local file ('.httr-oauth'), to cache OAuth access credentials between R sessions?
Seleccionar 1: 'Yes'
A continuación, se abrirá una ventana de navegador solicitando hacer logon en una cuenta de google, donde una vez se haya realizado logon se pedirá permiso para que 'tidyverse api packages' acceda a su cuenta de google para poder descargar archivos de Google Drive.

1a_PreparacionDatos_CalidadAire.R
1b_PreparacionDatos_Trafico.R
1c_PreparacionDatos_CalendarioLaboral.R
Con estos scripts se realiza la preparación de los datos de la calidad del aire, el tráfico y el Calendario Laboral respectivamente.

2a_Estudio_Variables_Graficas_CalidadAire.R
Gráficas del estudio de superación de los límites establecidos por la legislación con la librería ggplot2.

2b_AnalisisCorrelacionVariables
Matriz de correlación de las variables con la función ggpairs.

3a_ModeloVAR.ipynb
Script con el desarrollo del modelo VAR de la serie temporal multivariable.

3b_ModeloML_Random Forest.ipynb
Script con el desarrollo del modelo de machine learning con Random Forest.

4_Visualizacion.twb
Este archivo está compuesto por un conjunto de Dashboards en Tableau.
Para ver correctamente la aplicación de visualización diseñada hay que ejecutar el 'Modo de presentación' desde el Dashboard 'Menu'.

5 Algoritmos empleados

Como cada variable en el estudio no solo depende de valores pasados sino también de otras variables, vamos a tratar la predicción como una Serie Temporal Multivariable.
Por tanto, para la predicción de las variables hemos utilizado el modelo estadístico VAR y el algoritmo de Machine Learning Random Forest.

5.1 Modelo VAR (Vector Auto Regression)

VAR es un método capaz de entender y usar la relación entre las distintas variables.
El modelo VAR se crea a partir de variables linealmente independientes y estas han de ser estacionarias, sino lo fueran deberíamos diferenciarlas hasta que lo fueran.
Por tanto para poder aplicar este modelo, previamente debo asegurarme que el sistema está cointegrado comprobando la estacionariedad de todo el sistema. Para comprobar esto, el módulo de los autovalores debe ser menor que 1, y esto lo verificamos usando el test de Johansen.
Una vez hemos comprobado que nuestro sistema está cointegrado, creo los conjuntos de train (80%) y validation (20%). La selección de estos conjuntos es dependiente del tiempo, no podría usar un cross-validation para seleccionar estos conjuntos ya que se trata de un problema de series temporales.
Posteriormente entrenamos el modelo y así poder realizar predicciones en el mismo intervalo de tiempo que el conjunto de validation. Basándonos en estas predicciones y los valores del conjunto de validation podemos comprobar como de óptimo es el modelo.
La métrica utilizada para este modelo es el RMSE (Raíz del Error Cuadrático Medio) y los valores obtenidos no han sido los suficientemente óptimos.

```
RMSE values
SO2 : 5.105466860213757
CO : 0.24448301203082343
NO2 : 30.941074903069413
PM10 : 18.61386343623298
O3 : 31.809517664876644
El error medio es:  6.361903532975329
```

Para la predicción final utilizo el fichero completo (train+validation)

El código de este modelo está implementado en el notebook 3a_ModeloVAR.ipynb.

5.2 Modelo Machine Learning Random Forest

Las series temporales pueden ser planteadas también como un problema de aprendizaje supervisado, y así obtener un modelo mediante algoritmos de Machine Learning.

Previamente a obtener el modelo vamos a preparar los datos adecuadamente. Para ello, inicialmente creamos los conjuntos de train (80%) y test (20%).
Posteriormente, sobre este conjunto de train genero lags y normalizo dichos datos mediante un RobustScaler. Estos lags se calculan de forma que podamos tener como inputs las

observaciones en tiempos anteriores (por ej: t-1) y como valor a predecir la observación actual (t).

Después y mediante un GridSearch, cálculo los parámetros óptimos para el algoritmo RandomForestRegressor.

Además, realizo un feature_importances_ para quedarme con las columnas más importantes en el modelo, y con éstas genero de nuevo el conjunto de datos de train.
Con este conjunto de datos, entreno el modelo para así predecir sobre los datos de test. Con esta predicción hecha y los datos de test puedo medir lo óptimo del modelo, utilizando la misma métrica que para el modelo anterior (RMSE).

RMSE values
SO2_0 : 4.137435612035518
CO_0 : 0.24259836294490342
NO2_0 : 34.71339886683454
PM10_0 : 34.18085476978372
O3_0 : 78.4589061090397
El error medio es: 15.69178122180794

Para la predicción final utilizo el fichero completo inicial (train+test)

El código de este modelo está implementado en el notebook 3b_ModelosML.ipynb.

6 Conclusiones y próximos pasos

Se ha realizado un estudio de los 5 contaminantes más peligrosos para la salud de los humanos, y de ello podemos concluir que tanto el SO2 como el CO no representan una amenaza para la salud con las mediciones realizadas desde el 2014.

Para el PM10, aún no superando en ningún momento el nº máximo de días en que se superan los límites legales, se deberían de tomar medidas sobre todo para los meses de verano, lo que hace que se lleguen a límites peligrosos para la salud.

El NO2 y el O3 representan la mayor amenaza para la salud de los contaminantes en el estudio por su alto índice de presencia durante todo el año, con lo que habría que tomar las medidas necesarias para reducir la emisión de ambos contaminantes.

Tanto el PM10 como el NO2 son contaminantes generados principalmente por el tráfico rodado, por lo que una medida importante podría ser reducirlo, lo que indirectamente también haría disminuir las cantidades de O3 en los meses con altas temperaturas.
Otra medida importante que podría tomarse para realizar un estudio más eficiente con más datos y más precisos sería el de habilitar a todas las estaciones de medición de calidad del aire con la facultad de medir todos los contaminantes. Actualmente, tan solo 3 estaciones miden los 5 contaminantes sobre los que trata el estudio, siendo estos los más peligrosos para la salud de las personas.

Respecto a las predicciones sobre los contaminantes, se ha implementado un modelo estocástico mediante el VAR y un modelo de Machine Learning mediante un Random Forest no resultando ser muy precisos ninguno de ellos.

Como próximos pasos se añadirán datos acerca de la climatología para ver si mejoran las predicciones con los modelos comentados. Además se desarrollara algún modelo con redes neuronales tales como CNN (Red Neuronal Convolucional) o RNN (Red Neuronal Recurrente).