



GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

ANÁLISIS DEL NIVEL DE CONTAMINACIÓN ATMOSFÉRICA DE MADRID UTILIZANDO TÉCNICAS DE MACHINE LEARNING

Autor: Manuel Sainz de Baranda Portela

Director: Eugenio Francisco Sánchez Úbeda

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
**ANÁLISIS DEL NIVEL DE CONTAMINACIÓN ATMOSFÉRICA EN MADRID
UTILIZANDO TÉCNICAS DE MACHINE LEARNING en la ETS de Ingeniería -
ICAI de la Universidad Pontificia Comillas en el**

curso académico 2019/2020 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

**El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido
tomada de otros documentos está debidamente referenciada.**

Fdo.:



Fecha: 13/ 07/ 2020

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.:



Fecha: 13/ 07/ 2020



GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

ANÁLISIS DEL NIVEL DE CONTAMINACIÓN ATMOSFÉRICA DE MADRID UTILIZANDO TÉCNICAS DE MACHINE LEARNING

Autor: Manuel Sainz de Baranda Portela

Director: Eugenio Francisco Sánchez Úbeda

Madrid

ANÁLISIS DEL NIVEL DE CONTAMINACIÓN ATMOSFÉRICA EN MADRID UTILIZANDO TÉCNICAS DE MACHINE LEARNING

Autor: Sainz de Baranda Portela, Manuel.

Director: Sánchez Úbeda, Eugenio Francisco.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Mediante modelos obtenidos utilizando regresión, se ha buscado explicar ocho contaminantes diferentes en Madrid. Para ello, la obtención de variables como el tráfico o la meteorología han sido fundamentales. A partir de las variables obtenidas se ha ajustado un modelo de regresión lineal múltiple para cada contaminante. Los resultados obtenidos son razonables para todos los contaminantes, con la excepción del Dióxido de Azufre y las partículas en los que se ha determinado que faltan variables explicativas adicionales.

Palabras clave: Contaminación, Tráfico, Meteorología, Regresión

1. Introducción

La contaminación, especialmente de las grandes ciudades, es un factor de riesgo para la salud. La Organización Mundial de la salud (OMS) informa que en 2012 un total de siete millones de personas perdieron la vida a causa de la contaminación. Lo cual significa que una de cada ocho personas murió ese año por la baja calidad del aire [1].

Mediante variables como el tráfico, las condiciones meteorológicas y la festividad se pretende explicar las variaciones de ocho contaminantes en Madrid. Los contaminantes que se buscan explicar son los siguientes:

- Dióxido de Azufre (SO_2), es un gas incoloro y nocivo para la salud a altas concentraciones. Se genera principalmente por la combustión de carbón.
- Monóxido de Carbono (CO), es un gas insípido e incoloro emitido como combustión incompleta de carburantes fósiles y biocombustibles.
- Monóxido de Nitrógeno (NO), Dióxido de Nitrógeno (NO_2) y el Óxido de Nitrógeno (NO_x). Son gases que se emiten principalmente en procesos de combustión que ocurren en relación con el tráfico.
- Partículas menores de 10 micras (PM_{10}) y Partículas menores de 2.5 micras ($PM_{2.5}$), son partículas que pueden producirse de manera natural, a causa del polvo, partículas en el suelo o el polen, o como consecuencia de la actividad humana, especialmente debido al tráfico rodado.
- El Ozono (O_3) es un gas incoloro, invisible y de olor agradable formado debido a reacciones químicas con los Óxidos de Nitrógeno o con los compuestos orgánicos volátiles en presencia de luz solar.

2. Definición del proyecto

La información de contaminación se ha obtenido del portal de datos abiertos del ayuntamiento de Madrid [2]. Esta información contiene 17 contaminantes de los que únicamente se seleccionan los 8 contaminantes mencionados en la introducción al

considerarse de mayor importancia. Los datos obtenidos son diarios y se descargaron desde el 1 de enero de 2020 hasta el 31 de mayo de 2020.

La obtención de las variables es una parte fundamental del trabajo. Las variables meteorológicas se obtienen de la Agencia Estatal de Meteorología (AEMET) [3] cuyos datos también se descargaron desde el año 2001 hasta mayo del 2020. Las variables de tráfico se descargaron del portal de datos abiertos del ayuntamiento de Madrid [4]. El tráfico presenta alrededor de 4500 puntos de medida de los que una gran parte contienen varios huecos o falta de información. Cada punto de medida recoge un dato cada 15 minutos por lo que fue necesario calcular la media diaria. La información descargada corresponde con la intensidad del tráfico durante esos 15 minutos y únicamente se tienen datos desde julio del 2013. Además, se creó un calendario donde se recogen otro tipo de variables como la festividad.

Finalmente, antes de ajustar los modelos de regresión para estimar los contaminantes, se ha realizado un estudio analítico de todos los contaminantes de interés. Para llevar a cabo este análisis se han utilizado técnicas de machine learning como ANOVA, clustering o PCA. Con el fin de comprender el comportamiento de estos contaminantes en Madrid y aplicarlo a los modelos.

3. Descripción del modelo

Mediante modelos de regresión lineal múltiple se busca obtener un conjunto de modelos que a partir de las variables obtenidas anteriormente sean capaces de estimar los niveles de concentración de los diferentes contaminantes. Para ello, tomando como datos de aprendizaje aquellos comprendidos entre el 2015 y el 2018 (conjunto de aprendizaje) y, utilizando técnicas de *machine learning*, se han ajustado los modelos por mínimos cuadrados. A continuación, se muestra los datos diarios reales de Dióxido de Nitrógeno y la estimación obtenida con el modelo ajustado.

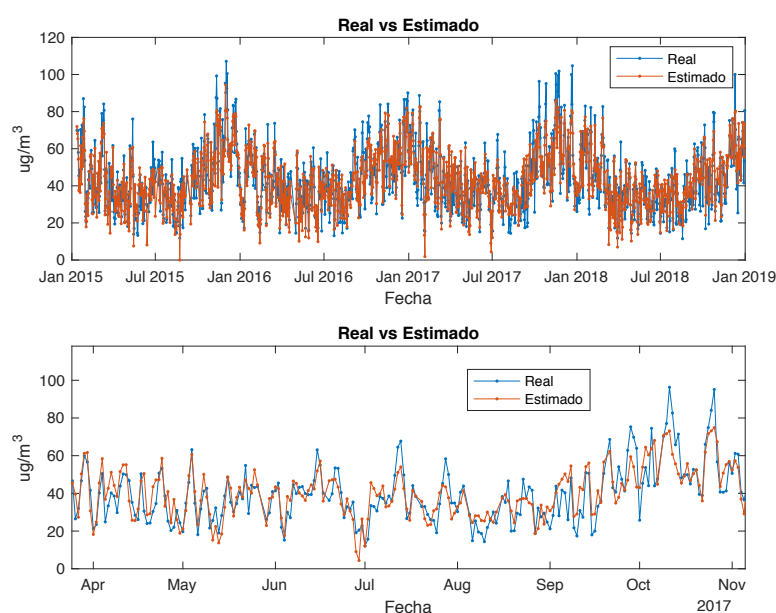


Figura 1 Modelo del Dióxido de Nitrógeno

4. Resultados

Una vez obtenido el modelo se ha validado su capacidad de generalización con los datos de 2019 y 2020 (conjunto de test) para todos los contaminantes. Para ello se ha calculado el coeficiente de determinación ajustado (R^2) y la raíz del error cuadrático medio (RMSE), recogidos en la tabla adjunta.

Tabla 1. Resultados de los Modelos

Contaminante	R^2 (Aprendizaje)	R^2 (Test)	RMSE (Aprendizaje)	RMSE (Test)
Dióxido de Azufre	0.3233	0.2852	2.2396	3.2943
Monóxido de Carbono	0.7361	0.6901	0.0792	0.0915
Monóxido de Nitrógeno	0.7359	0.7213	15.2968	16.5054
Dióxido de Nitrógeno	0.7967	0.8139	7.6904	9.3864
Óxido de Nitrógeno	0.7672	0.7843	29.2123	32.7542
Partículas < 10 micras	0.7011	0.5083	6.586	7.1384
Partículas < 2.5 micras	0.625	0.6	3.4762	3.4756
Ozono	0.801	0.7348	9.8247	11.4501

5. Conclusiones

Una obtención adecuada de variables es fundamental para poder explicar cualquiera de los distintos contaminantes. La preparación de esta información, especialmente con el tráfico, ha sido complicada debido a la falta de homogeneidad y presencia de huecos en los datos.

Se han obtenido ocho modelos utilizando regresión lineal múltiple. Estos modelos son razonables para la mayoría de los contaminantes. Cabe destacar una posible mejora en el caso de las partículas si se incluyese una variable de intervención para recoger las intrusiones de masas de aire sahariano. El Dióxido de Azufre no se consiguió explicar razonablemente bien con las variables planteadas, posiblemente debido a la baja influencia del tráfico en este contaminante.

En definitiva, se ha logrado comprender las variables influyentes en los contaminantes en los que se obtuvieron modelos de calidad. El tráfico, las temperaturas, el viento o el mes del año han sido capaces de explicar razonablemente la calidad del aire en Madrid.

6. Referencias

- [1] World Health Organization, Marzo 2014. OMS | 7 millones de muertes cada año debidas a la contaminación atmosférica. Recuperado 14 de mayo de 2020, de <https://www.who.int/mediacentre/news/releases/2014/air-pollution/es/>

- [2] Ayuntamiento de Madrid. Calidad del aire. Datos diarios años 2001 a 2020 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 10 de febrero de 2020, de <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>
- [3] A.E. Meteorolog. AEMET OpenData. Recuperado 1 de marzo de 2020, de http://www.aemet.es/es/datos_abiertos/AEMET_OpenData
- [4] Ayuntamiento de Madrid. Tráfico. Histórico de datos del tráfico desde 2013 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 15 de marzo de 2020, de <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

ANALYSIS OF THE LEVEL OF ATMOSPHERIC POLLUTION IN MADRID USING MACHINE LEARNING TECHNIQUES

Author: Sainz de Baranda Portela, Manuel

Supervisor: Sánchez Úbeda, Eugenio Francisco

Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

Through models obtained using regression, we have sought to explain eight different contaminants in Madrid. For this, obtaining variables such as traffic or meteorology is essential. From the obtained variables, a multiple linear regression model is adjusted for each pollutant. The results obtained are relatively good (adjusted determination coefficient higher than 0.7) for all pollutants with the exception of Sulfur Dioxide and particles where additional explanatory variables have been found to be missing.

Keywords: Pollution, Traffic, Meteorology, Regression.

1. Introduction

Pollution, especially in large cities, is a risk factor for people and vegetation. The World Health Organization (WHO) reports that in 2012 a total of seven million people lost their lives due to contamination. Which means that one in eight people died that year from poor air quality [1].

By means of variables such as traffic, weather conditions and holidays, the aim is to explain the variations of eight pollutants in Madrid. The pollutants studied are the following:

- Sulfur Dioxide (SO_2), is a colorless gas and harmful to people at high concentrations. It is mainly generated by the combustion of coal.
- Carbon monoxide (CO) it is a tasteless and colorless gas emitted as incomplete combustion of fossil fuels and biofuels.
- Nitrogen Monoxide (NO), Nitrogen dioxide (NO_2) and Nitrogen Oxide (NO_x). They are gases that are emitted mainly in combustion processes that occur in relation to traffic.
- Particles smaller than 10 microns (PM_{10}) and Particles smaller than 2.5 microns ($PM_{2.5}$) are particles that can occur naturally, due to dust, soil particles or pollen, or as a consequence of human activity, especially due to traffic.
- Ozone (O_3) It is a colorless, invisible gas with a pleasant odor formed due to chemical reactions with Nitrogen Oxides or with volatile organic compounds in the presence of sunlight.

2. Definition of the Project

Pollution data has been obtained from the open data portal of the Madrid City Council [2]. This information contains 17 contaminants from which only the 8 contaminants mentioned in the introduction are selected as they are considered to be of major

importance. The data obtained is daily and was downloaded from January 1, 2020 to May 31, 2020.

Obtaining the variables is a fundamental part of this project. The meteorological variables are obtained from the State Meteorological Agency (AEMET) [3], data which was also downloaded from 2001 to May 2020. Traffic variables were downloaded from the open data portal of the Madrid city council [4]. Traffic presents around 4500 measurement points of which a large part contains various gaps or lack of information. Each measurement point collects a data every 15 minutes, so it was necessary to calculate the daily average. The downloaded information corresponds to the intensity of the traffic during those 15 minutes and there are only data downloaded from July 2013 onwards. In addition, a calendar was created where other variables such as the holidays are gathered.

Finally, before adjusting the regression models to estimate the pollutants, an analytical study of all the pollutants of interest is carried out. Machine analysis techniques such as ANOVA, clustering or PCA have been used to carry out this analysis. In order to understand the behavior of these pollutants in Madrid and apply it to the models.

3. Model Description

Using multiple linear regression models, we seek to obtain a set of models that, based on the variables obtained previously, are capable of estimating the concentration levels of the different pollutants. To do this, taking as learning data those between 2015 and 2018 (learning set) and, using machine learning techniques, the models have been adjusted by the least squares method. The actual daily Nitrogen Dioxide data and the estimate obtained with the fitted model are shown below.

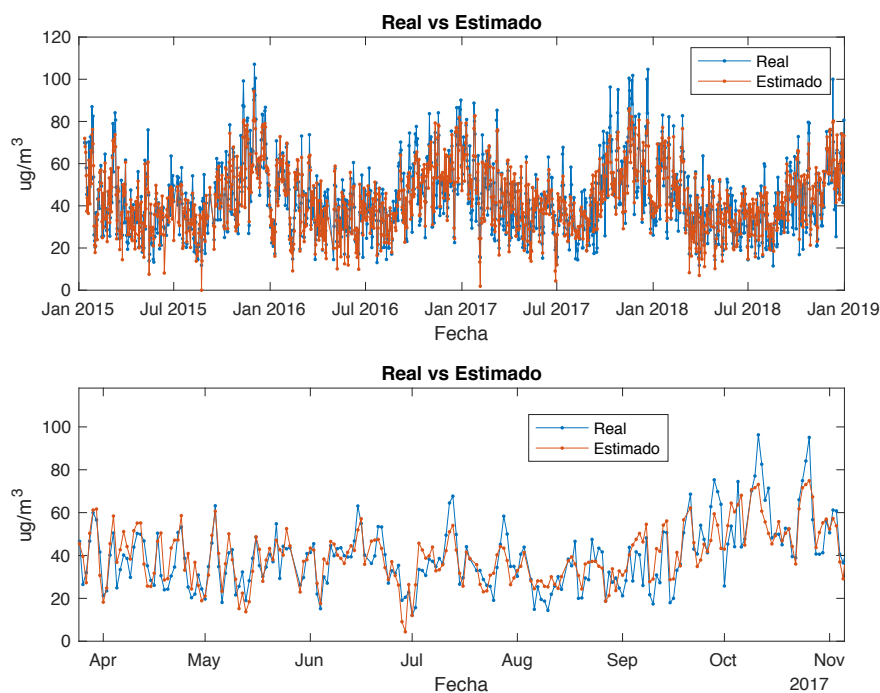


Figure 1 Model of Nitrogen Dioxide

4. Results

Once the model is obtained, its generalization capacity is tested with data from 2019 and 2020 (test set) for all contaminants. Using its adjusted coefficient of determination (R^2) and the root mean square error (RMSE), the quality of the results is evaluated. The results are shown in the attached table.

Table 1. Model Results

Pollutant	R^2 (Learning set)	R^2 (Test set)	RMSE (Learning set)	RMSE (Test set)
Sulfur Dioxide	0.3233	0.2852	2.2396	3.2943
Carbon Monoxide	0.7361	0.6901	0.0792	0.0915
Nitrogen Monoxide	0.7359	0.7213	15.2968	16.5054
Nitrogen Dioxide	0.7967	0.8139	7.6904	9.3864
Nitrogen Oxide	0.7672	0.7843	29.2123	32.7542
Particles smaller than 10 micros	0.7011	0.5083	6.586	7.1384
Particles smaller than 2.5 micros	0.625	0.6	3.4762	3.4756
Ozone	0.801	0.7348	9.8247	11.4501

5. Conclusions

An adequate obtaining of variables is essential to be able to explain any of the different contaminants. The preparation of this information, especially with traffic, has been complicated due to the lack of homogeneity and the presence of gaps in the data.

Eight models have been obtained using multiple linear regression. These models are reasonable for most pollutants. It is worth noting a possible improvement in the case of particles if an intervention variable were included to collect intrusions from Saharan air masses. The Sulfur Dioxide could not be explained reasonably well with the variables proposed, possibly due to the low influence of traffic on this pollutant.

In conclusion, it has been possible to understand the influential variables in the pollutants in which quality models were obtained. Traffic, temperatures, precipitations, wind or the month of the year have been able to reasonably explain the air quality in Madrid.

6. References

- [1] World Health Organization, Marzo 2014. OMS | 7 millones de muertes cada año debidas a la contaminación atmosférica. Recuperado 14 de mayo de 2020, de <https://www.who.int/mediacentre/news/releases/2014/air-pollution/es/>
- [2] Ayuntamiento de Madrid. Calidad del aire. Datos diarios años 2001 a 2020 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 10 de febrero de 2020, de

<https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>

- [3] A.E. Meteorolog. AEMET OpenData. Recuperado 1 de marzo de 2020, de http://www.aemet.es/es/datos_abiertos/AEMET_OpenData
- [4] Ayuntamiento de Madrid. Tráfico. Histórico de datos del tráfico desde 2013 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 15 de marzo de 2020, de <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

Índice de la memoria

Capítulo 1. Introducción.....	8
Capítulo 2. Estado de la Cuestión.....	11
Capítulo 3. Definición del Trabajo.....	13
3.1 Justificación	13
3.2 Metodología	13
3.3 Planificación	14
Capítulo 4. Datos de Contaminación.....	15
Capítulo 5. Variables Explicativas.....	20
5.1 Meteorología	20
5.1.1 Preparación de la Información.....	20
5.1.2 Análisis de la Meteorología	21
5.2 Calendario	25
5.3 Tráfico	25
5.3.1 Preprocesado de la Información.....	26
5.3.2 Análisis de Tráfico	31
5.3.3 Obtención de variables explicativas	35
Capítulo 6. Dióxido de Azufre.....	41
6.1 Límites, Máximos y Mínimos.....	41
6.1.1 Límites Legales	41
6.1.2 Umbrales de Información	41
6.2 Análisis	42
6.3 Modelo explicativo del nivel de concentración	47
6.4 Análisis de Resultados	50
Capítulo 7. Óxidos, Dióxidos y Monóxidos de Nitrógeno.....	53
7.1 Límites, Máximos y Mínimos.....	53
7.1.1 Límites Legales	53
7.1.2 Umbrales de Información	54

7.2	Análisis	55
7.3	Modelo explicativo del nivel de concentración	59
7.4	Análisis de Resultados	67
Capítulo 8. Partículas menores de 10 y 2,5 micras		71
8.1	Límites, Máximos y Mínimos	71
8.1.1	Limites Legales	71
8.1.2	Umbral de Información	72
8.2	Análisis	73
8.3	Modelo explicativo del nivel de concentración	77
8.4	Análisis de Resultados	83
Capítulo 9. Monóxido de Carbono		86
9.1	Máximos y Mínimos	86
9.2	Análisis	86
9.3	Modelo explicativo del nivel de concentración	91
9.4	Análisis de Resultados	93
Capítulo 10. Ozono		95
10.1	Máximos y Mínimos	95
10.2	Análisis	95
10.3	Modelo explicativo del nivel de concentración	99
10.4	Análisis de Resultados	102
Capítulo 11. Conclusiones y Trabajos Futuros		104
11.1	Conclusiones	104
11.2	Trabajos Futuros	106
Bibliografía 107		
ANEXO I – Objetivos de Desarrollo Sostenible		110
ANEXO II – Fuentes de Información		112
11.3	Información de Contaminación	112
11.4	Información meteorológica	114
11.5	Información del tráfico	115

<i>ANEXO III – Límites Calidad del Aire</i>	<i>117</i>
---	------------

Índice de figuras

Figura 1. Distribución por zonas de la ciudad de Madrid [21]	15
Figura 2. Dendograma Contaminantes.....	18
Figura 3. Centroides de los Cluster (Contaminantes)	19
Figura 4. Ubicación de las estaciones de meteorología	21
Figura 5. Variable Temperatura Media y valores suavizados con tres tamaños de ventana.....	22
Figura 6. Variable Velocidad Media del Viento y valores suavizados con tres tamaños de ventana	22
Figura 7. Variable de Precipitación y valores suavizados con tres tamaños de ventana.....	23
Figura 8. Dendrograma de las temperaturas (reales y suavizadas)	24
Figura 9. Dendrograma de la lluvia y del viento (reales y suavizadas)	24
Figura 10. Localización de las estaciones de contaminación y los puntos de medida de tráfico	27
Figura 11. Intensidad de tráfico recogido en los puntos de medida cercanos a Plaza de España	28
Figura 12. Discontinuidades en las series de tráfico (atípicos y huecos).....	28
Figura 13. Calidad de los datos recogidos en los puntos de medida de tráfico.....	29
Figura 14. Calidad de los datos recogidos en los puntos de medida de tráfico después del filtro.....	30
Figura 15. Mapa de las estaciones eliminadas/mantenidas	31
Figura 16. Tráfico Mensual.....	32
Figura 17. Tráfico Anual.....	32
Figura 18. Tráfico Semanal.....	33
Figura 19. Dendrograma de Tráfico (Plaza de España)	34
Figura 20. Variables de Tráfico preseleccionadas en el método 1	36
Figura 21. Variables de Tráfico construidas con el método 1	37
Figura 22. Gráfico de Sedimentación para las variables de tráfico.....	38

Figura 23. Evolución Temporal de las dos primeras componentes principales del tráfico.	39
Figura 24. Peso de las Variables (Primera y Segunda Componente Principal)	40
Figura 25. Serie Temporal del Dióxido de Azufre en las diferentes estaciones	43
Figura 26. Obtención del número de conglomerados (SO_2).....	44
Figura 27. Centroides de los conglomerados (SO_2)	45
Figura 28. Test Anova (SO_2)	46
Figura 29. Prueba de Comparación Múltiple para el mes (SO_2)	46
Figura 30. Análisis de Residuos (Dióxido de Azufre)	48
Figura 31. Estimación vs Real (Dióxido de Azufre).....	49
Figura 32. Predicción vs Real (Dióxido de Azufre).....	51
Figura 33. Estimación y predicción vs Real del Dióxido de Azufre (entrenamiento y test)	52
Figura 34. Serie Temporal del Dióxido, Monóxido y Óxido de Nitrógeno en las diferentes estaciones	55
Figura 35. Obtención del número de conglomerados (NO , NO_x , NO_2)	56
Figura 36. Centroides de los conglomerados (NO , NO_x , NO_2)	57
Figura 37. Prueba de Comparación Múltiple del día de la semana (NO_2)	59
Figura 38. Análisis de los Residuos (Monóxido de Nitrógeno).....	61
Figura 39. Estimado vs Real (Monóxido de Nitrógeno)	61
Figura 40. Análisis de los Residuos (Dióxido de Nitrógeno)	63
Figura 41. Estimado vs Real (Dióxido de Nitrógeno).....	63
Figura 42. Análisis de los Residuos (Óxido de Nitrógeno)	65
Figura 43. Estimación vs Real (Óxido de Nitrógeno).....	66
Figura 44. Predicción vs Real (Monóxido de Nitrógeno)	68
Figura 45. Predicción vs Real (Dióxido de Nitrógeno).....	69
Figura 46. Predicción vs Real (Óxido de Nitrógeno).....	70
Figura 47. Serie Temporal de las partículas menores de 10 y 2.5 micras en las diferentes estaciones	74
Figura 48. Obtención del número de conglomerados ($PM_{2.5}$, PM_{10})	75
Figura 49. Centroides de los conglomerados ($PM_{2.5}$, PM_{10})	76
Figura 50. Análisis de los Residuos (Partículas menores de 2.5 micras).....	79

Figura 51. Estimado vs Real (Partículas menores a 2.5 micras)	79
Figura 52. Análisis de los Residuos (Partículas menores de 10 micras).....	81
Figura 53. Estimado vs Real (Partículas menores de 10 micras)	82
Figura 54. Predicción vs Real (partículas menores de 2.5 micras)	84
Figura 55. Predicción vs Real (partículas menores de 10 micras)	85
Figura 56. Serie Temporal del Monóxido de Carbono en las diferentes estaciones	87
Figura 57. Obtención del número de conglomerados (Monóxido de Carbono).....	88
Figura 58. Centroides de los conglomerados (Monóxido de Carbono)	89
Figura 59. Prueba de comparación múltiple de las estaciones (<i>Monóxido de Carbono</i>).	90
Figura 60. Test Anova con interacción.	90
Figura 61 Análisis de Residuos (Monóxido de Carbono).....	92
Figura 62. Estimación vs Real (Monóxido de Carbono).....	92
Figura 63. Predicción vs Real (Monóxido de Carbono).....	94
Figura 64. Serie Temporal del Ozono en las diferentes estaciones.....	96
Figura 65. Obtención número de conglomerados (Ozono).....	97
Figura 66. Centroides de los conglomerados (Ozono).....	98
Figura 67. Prueba de Comparación Múltiple de la festividad (Ozono)	99
Figura 68 Análisis de Residuos (Ozono)	101
Figura 69. Estimación vs Real (Ozono)	101
Figura 70. Predicción vs Real (Ozono)	103
Figura 71. Datos Contaminantes.	113
Figura 72. Aemet Open Data [19].....	114
Figura 73. Éxito en la obtención de datos	115
Figura 74. Representación datos Tráfico.....	116

Índice de tablas

Tabla 1. Planificación de trabajo (Cronograma)	14
Tabla 2. Estaciones de Contaminación	16
Tabla 3. Datos de Contaminación (1).....	17
Tabla 4. Datos de Contaminación (2).....	17
Tabla 5. Diagnóstico e interpretación del modelo (Dióxido de Azufre).....	49
Tabla 6. Media Anual Dióxido de Nitrógeno.....	54
Tabla 7. Diagnóstico e interpretación del modelo (Monóxido de Nitrógeno)	62
Tabla 8. Diagnóstico e interpretación del modelo (Dióxido de Nitrógeno).....	64
Tabla 9. Diagnóstico e interpretación del modelo (Óxido de Nitrógeno).....	66
Tabla 10. Media Anual Partículas menores de 10 y 2.5 micras	72
Tabla 11. Nº de Superaciones de los límites de partículas menores de 10 micras.....	73
Tabla 12. Diagnóstico e interpretación del modelo (Partículas menores de 2.5 micras)	80
Tabla 13. Diagnóstico e interpretación del modelo (partículas menores de 10 micras)	82
Tabla 14. Diagnóstico e interpretación del modelo (Monóxido de Carbono).....	93
Tabla 15. Diagnóstico e interpretación del modelo (Ozono)	102
Tabla 16. Resumen de los resultados obtenidos.....	105
Tabla 17. Límites Legales [8]	117
Tabla 18. Umbrales de Información [8]	118

Capítulo 1. INTRODUCCIÓN

La contaminación, especialmente de las grandes ciudades, es un factor de riesgo para la salud. La Organización Mundial de la salud (OMS) informa que en 2012 un total de siete millones de personas perdieron la vida a causa de la contaminación. Lo cual significa que una de cada ocho personas murió ese año por la baja calidad del aire [1]. Mejorar la calidad del aire, por tanto, es uno de los grandes retos que las ciudades tienen por delante. Para poder hacerlo, se deben de entender qué variables explican los niveles de concentración con el fin de saber cuál es la mejor forma de reducir las emisiones generadas por los seres humanos y, en definitiva, mejorar la salud de las personas.

Madrid ha ido empleando una serie de medidas como no permitir que los coches que más contaminan puedan circular por el centro (Madrid Central), no permitir que algunos coches puedan aparcar en la zona interior de la M-30 y ha impuesto regulaciones de velocidad en algunas zonas. Aparentemente, todos los esfuerzos de la ciudad se centran en reducir el tráfico rodado.

El objetivo del presente proyecto es encontrar un grupo de variables capaces de explicar la evolución temporal de los distintos contaminantes de Madrid. Concretamente, se parte de variables meteorológicas como el viento o las temperaturas, variables de intensidad de tráfico, y de variables derivadas del calendario como los festivos, el día de la semana, el mes o el año. Cada contaminante será explicado mejor por unas variables y peor por otras. Los contaminantes estudiados en este trabajo son los siguientes:

- Dióxido de Azufre (SO_2) es un gas incoloro que resulta irritante y nocivo para la salud a concentraciones elevadas. Su origen se encuentra en la combustión de combustibles fósiles, principalmente carbón y derivados del petróleo debido a la presencia de Azufre en su composición y, mediante el proceso de combustión, combina con el oxígeno pasando de esta forma a la atmósfera [2]. Su unidad de

medida es microgramo por metro cúbico y la técnica de medida es la fluorescencia ultravioleta.

- Monóxido de Carbono (CO) es un gas insípido e incoloro emitido como combustión incompleta de carburantes fósiles y biocombustibles. El efecto del tráfico en las emisiones de este contaminante se ha visto reducido por el uso de conversores catalíticos [3]. Su unidad de medida es el miligramo por metro cúbico y su técnica de medida es la absorción infrarroja.
- Monóxido de Nitrógeno (NO), Dióxido de Nitrógeno (NO_2) y el Óxido de Nitrógeno (NO_x) son gases que se emiten principalmente en procesos de combustión que ocurren en relación con el tráfico, especialmente a causa de motores diésel. Aunque también se producen debido a la generación de energía eléctrica. Estos focos de emisión producen generalmente Monóxido y Dióxido de Carbono que al oxidarse pueden producir el Óxido de Nitrógeno [4]. La unidad de medida de los tres contaminantes es el microgramo por metro cúbico y la técnica de medida es la quimioluminiscencia.
- Partículas menores de 10 micras (PM_{10}) y Partículas menores de 2.5 micras ($PM_{2.5}$) son partículas que pueden producirse de manera natural, a causa del polvo, partículas en el suelo o el polen, o como consecuencia de la actividad humana, especialmente debido al tráfico rodado [5]. La unidad de medida de ambos contaminantes es el microgramo por metro cúbico y la técnica de medida es la microbalanza.
- El Ozono (O_3) es un gas incoloro, invisible y de olor agradable que se encuentra de forma natural en la estratosfera. También se encuentra en la troposfera y a concentraciones elevadas resulta ser un riesgo para la salud. El Ozono se forma en esta capa debido a reacciones químicas con los Óxidos de Nitrógeno o con los compuestos orgánicos volátiles en presencia de luz solar. Predomina en las zonas rurales y suele aumentar su concentración los domingos, como consecuencia de la disminución de otros contaminantes al reducirse el tráfico [7]. La unidad de medida del Ozono es el microgramo por metro cúbico y la técnica de medida es la absorción ultravioleta.

Todos los contaminantes son distintos entre sí, aunque algunos dependen claramente del tráfico mientras que otros contaminantes apenas se ven influenciados por esta variable. En definitiva, comprender que variables influyen directamente en los contaminantes puede dar una idea de la manera de evitar que se superen unos límites a partir de los cuales el contaminante puede ser nocivo para la salud. Existen unos límites legales que ningún municipio o comunidad deberá de superar, que se encuentran publicados en el Boletín Oficial del Estado y que fueron decretados el 28 de enero de 2011 mediante real decreto [8].

Además, debido a la situación excepcional vivida en España, especialmente en Madrid, causada por la pandemia de COVID-19, se ha vivido una situación de confinamiento y parón económico. Esta situación ha llevado a Madrid a presentar un descenso muy elevado en el tráfico y que la industria se haya tenido que detener durante varias semanas o incluso meses. La contaminación, por tanto, al ser principalmente causada por el ser humano debería presentar un claro descenso durante los meses de marzo, abril y mayo de 2020.

Finalmente, para la realización del presente trabajo se utilizarán técnicas de aprendizaje automático utilizando Matlab. Estas técnicas incluyen ANOVA (por sus siglas en inglés, *ANalysis Of VAriance*), PCA (por sus siglas en inglés, *Principal Component Analysis*), *Clustering* y Regresión. En ANOVA se analiza la varianza para contrastar que las medias de las distintas poblaciones coinciden [9]. En PCA se calculan unas componentes principales capaces de explicar la mayor cantidad de información (varianza) con un número menor de variables [10]. En *Clustering* se agrupan observaciones o variables en *clusters* o conglomerados según alguna medida de semejanza [11]. Para terminar, con la técnica de regresión se trata de explicar una variable dependiente, que en este caso sería cualquiera de los contaminantes, a partir de un conjunto de variables explicativas, como las temperaturas, la lluvia o el tráfico [12].

Capítulo 2. ESTADO DE LA CUESTIÓN

Durante los últimos años la preocupación por la calidad del aire ha ido en aumento. Como consecuencia existen varios trabajos que intentan explicar la evolución de las series temporales de algunos contaminantes en función de otras variables. La mayoría de ellos se centran en explicar alguno de los contaminantes siendo de especial interés el estudio de las partículas menores de 10 y 2.5 micras debido a los problemas de salud que pueden acarrear.

Existen estudios estadísticos que estudian la correlación entre los contaminantes y variables meteorológicas. Se han realizado análisis de regresión lineal múltiple para explicar contaminantes como el Dióxido de Azufre o las partículas menores de diez micras utilizando variables explicativas como las temperaturas o la humedad relativa obteniendo buenos resultados para medias anuales y mensuales [13]. A diferencia de dicho trabajo, el objetivo de este proyecto es realizar un análisis de todos los contaminantes influyentes en la zona urbana de Madrid utilizando diferentes técnicas de aprendizaje automático entre las que se incluye regresión.

Además, debido a los crecientes controles de calidad del aire en el mundo, muchos gobiernos buscan la forma de poder predecir los niveles de contaminación, especialmente en invierno, debido a la mayor concentración de contaminación en la atmósfera. Por ello, a partir de valores anteriores del contaminante, así como de otras variables explicativas como la velocidad del viento o las temperaturas, se pretende predecir el comportamiento de la variable utilizando técnicas como redes neuronales o un análisis no-paramétrico como se explica en la revista científica *Ingeniare* [14]. A diferencia del trabajo mencionado, en este proyecto, el objetivo no es predecir el nivel de contaminación futuro, sino estudiar una cantidad superior de contaminantes mediante otros métodos como ANOVA, PCA, clustering y, sobre todo, regresión para obtener modelos que expliquen la alta variabilidad en los niveles de calidad del aire de Madrid.

Existe un trabajo sobre el análisis de la calidad del aire en Madrid utilizando técnicas de aprendizaje automático. En este trabajo se realiza un análisis y un modelo predictivo utilizando técnicas de regresión y clasificación para el Dióxido de Nitrógeno [15]. El trabajo mencionado incluye variables como el tráfico, el viento y los niveles de concentración del contaminante a días pasados. A diferencia de dicho trabajo, en el presente proyecto se estudia una cantidad mayor de contaminantes (incluido el Dióxido de Nitrógeno), con el objetivo no tanto de predecir los niveles de contaminación, sino de explicar cómo evolucionan los ocho contaminantes de interés con variables como el tráfico, las temperaturas, la lluvia, el viento, el mes o la festividad. En ambos proyectos se utiliza la información de la misma base de datos, con la diferencia de que el trabajo de fin de grado que aquí se desarrolla utiliza los datos diarios mientras que ese trabajo se centra en los datos horarios del Dióxido de Azufre.

En el trabajo presentado en [16] se analiza la calidad del aire en la zona metropolitana del Valle de México con datos desde el año 1995 hasta el 2002. Los contaminantes analizados son el Monóxido de Carbono, el Dióxido y Óxido de Nitrógeno, el Ozono, el Dióxido de Azufre y las partículas menores de 10 micras. Una de las técnicas utilizadas en este trabajo es el método *K-means*, el cual también se utiliza en este trabajo de fin de grado.

Finalmente, existen análisis del comportamiento de distintos contaminantes durante un periodo de tiempo determinado. Por ejemplo, el Ayuntamiento de Madrid presenta cada año una memoria en la que se estudian los distintos contaminantes [17]. Se analizan los promedios diarios, mensuales, anuales, así como si los umbrales de alerta o si los límites establecidos por la legislación vigente se han visto superados en algún momento. En este trabajo también se han analizado los límites de todos los años y, además, se ha realizado un análisis y un modelo explicativo de las series temporales de ocho contaminantes.

Capítulo 3. DEFINICIÓN DEL TRABAJO

3.1 JUSTIFICACIÓN

Este trabajo busca encontrar una serie de variables capaces de explicar la evolución de ocho contaminantes distintos en Madrid. A diferencia de otros proyectos, se consideran variables meteorológicas, variables de tráfico y la festividad, con el fin de lograr comprender que variables explicativas describen mejor el comportamiento de cada contaminante utilizando técnicas de aprendizaje automático. Además, se hace un estudio del comportamiento de cada uno de los contaminantes desde el año 2001 hasta el 2020 incluyendo máximos, estaciones con mayores medidas, influencia de distintas variables como el mes o el año, y finalmente se ajustan un conjunto de modelos cuyo objetivo es explicar las series temporales de todos los contaminantes. Es decir, se realiza un estudio más completo y de una cantidad mayor de contaminantes.

Tal y como se explica en el Capítulo 2. los gobiernos de distintos países deben cumplir unos índices mínimos de calidad del aire y, en algunos casos, de no cumplirse pueden llevar consigo una sanción. Por ello, la mejor forma de evitar dichas sanciones y garantizar la salud de las personas es entendiendo qué variables influyen en la contaminación para poder corregir y evitar sobrepasar los límites establecidos.

En el caso concreto de Madrid, comprendiendo cómo influyen las variables en los contaminantes se pueden tomar decisiones desde el Gobierno municipal de Madrid para evitar, en la medida de lo posible, concentraciones perjudiciales para la salud.

3.2 METODOLOGÍA

Para realizar el trabajo se deben de seguir los siguientes pasos:

1. Descargar los datos de contaminación del portal de datos abiertos del Ayuntamiento de Madrid [18].
2. Descargar los datos meteorológicos de la Agencia Estatal de Meteorología (AEMET) [19].
3. Descargar los datos de tráfico del portal de datos abiertos del Ayuntamiento de Madrid [20].
4. Elaboración de la programación en Matlab.
5. Análisis de los datos utilizando técnicas de aprendizaje automático con Matlab como ANOVA, PCA o *clustering*.
6. Obtención de modelos que estimen la contaminación a partir de las variables consideradas utilizando un análisis de regresión lineal múltiple.
7. Análisis crítico de los resultados.

3.3 PLANIFICACIÓN

En el siguiente cronograma se aprecia el plan de trabajo establecido:

Tabla 1. Planificación de trabajo (Cronograma)

	<i>Enero</i>	<i>Febrero</i>	<i>Marzo</i>	<i>Abril</i>	<i>Mayo</i>	<i>Junio</i>	<i>Julio</i>
Descarga y organización de los datos							
Elaboración de la programación en Matlab							
Análisis y comparación de resultados							
Elaboración de la memoria							

Capítulo 4. DATOS DE CONTAMINACIÓN

Los datos de contaminación se obtienen siguiendo los pasos mencionados en ANEXO I I – Fuentes de Información. Estos datos de contaminación se dividen en múltiples estaciones de las cuales no todas ellas miden todos los contaminantes. Se tienen datos desde el 1 de enero de 2001 hasta el 31 de mayo de 2020 de todos los contaminantes de interés.

Toda la información relativa a las estaciones de contaminación se encuentra detallada en la Tabla 2. Madrid se divide en cinco zonas de la que se ha tenido especial interés en la zona 1 debido a que representa la zona interior a la M-30 y contiene el mayor número de estaciones. La distribución de las estaciones por zonas se muestra en la Figura 1. Nótese que solo aparecen las estaciones que siguen en activo hoy en día.

Distribución por zonas de la ciudad de Madrid:

ZONA	ESTACIONES
1 (interior M-30)	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 (Plaza Elíptica) + 2 de fondo (Farolillo y Villaverde)



Figura 1. Distribución por zonas de la ciudad de Madrid [21]

Tabla 2. Estaciones de Contaminación

Estación	Nombre de Estación	Zona	Alta	Baja
E1	Pº. Recoletos	1	01-Jan-2001	04-May-2009
E2	Glta. De Carlos V	1	01-Jan-2001	04-Dec-2006
E35	Pza. del Carmen	1	01-Jan-2001	
E4	Pza. de España	1	01-Jan-2001	
E39	Barrio del Pilar	1	01-Jan-2001	
E6	Pza. Dr. Marañón	1	01-Jan-2001	27-Nov-2009
E7	Pza. M. de Salamanca	1	01-Jan-2001	30-Dec-2009
E8	Escuelas Aguirre	1	01-Jan-2001	
E9	Pza. Luca de Tena	1	01-Jan-2001	07-Dec-2009
E38	Cuatro Caminos	1	01-Jan-2001	
E11	Av. Ramón y Cajal	1	01-Jan-2001	
E12	Pza. Manuel Becerra	1	01-Jan-2001	30-Dec-2009
E40	Vallecas	2	01-Jan-2001	
E14	Pza. Fdez. Ladreda	1	01-Jan-2001	02-Dec-2009
E15	Pza. Castilla	1	01-Jan-2001	14-Oct-2008
E16	Arturo Soria	3	01-Jan-2001	
E17	Villaverde Alto	5	01-Jan-2001	
E18	C/ Farolillo	5	01-Jan-2001	
E19	Huerta Castañeda	5	01-Jan-2001	30-Dec-2009
E36	Moratalaz	2	01-Jan-2001	
E21	Pza. Cristo Rey	1	01-Jan-2001	04-Dec-2009
E22	Pº. Pontones	1	01-Jan-2001	20-Nov-2009
E23	Final C/ Alcalá	1	01-Jan-2001	30-Dec-2009
E24	Casa de Campo	4	01-Jan-2001	
E25	Santa Eugenia	2	01-Jan-2001	16-Nov-2009
E26	Urb. Embajada (Barajas)	3	01-Jan-2001	11-Jan-2010
E27	Barajas	3	01-Jan-2001	
E47	Méndez Álvaro	1	21-Dec-2009	
E48	Pº. Castellana	1	01-Jun-2010	
E49	Retiro	1	01-Jan-2010	
E50	Pza. Castilla	1	08-Feb-2010	
E54	Ensanche Vallecas	2	11-Dec-2009	
E55	Urb. Embajada (Barajas)	3	20-Jan-2010	
E56	Pza. Elíptica	5	18-Jan-2001	
E57	Sanchinarro	3	24-Nov-2009	
E58	El Pardo	4	30-Nov-2009	
E59	Parque Juan Carlos I	3	14-Dec-2009	
E60	Tres Olivos	3	14-Jan-2010	

Los datos recopilados desde el año 2001 se organizan para facilitar su uso con Matlab en dos tablas distintas. En la primera tabla se organizan los datos de tal manera que existen tantas columnas para cada contaminante como estaciones midan ese contaminante (ver Tabla 3). Mientras que la segunda tabla únicamente tiene una columna para cada contaminante, pero una cantidad mucho mayor de filas debido a que cada día se repite 39 veces (número de estaciones existentes) (ver Tabla 4).

Tabla 3. Datos de Contaminación (1)

Fecha	E1_C1	E1_C6	E1_C7	E1_C8	E1_C10	E1_C12	E1_C14	...
01-Jan-2001	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
02-Jan-2001	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
03-Jan-2001	16	1.3	94	80	NaN	37	225	...
04-Jan-2001	17	1.2	82	76	NaN	36	202	...
05-Jan-2001	16	1.5	130	76	NaN	33	276	...
06-Jan-2001	15	1	65	65	NaN	17	165	...
07-Jan-2001	19	1	75	67	NaN	21	181	...
08-Jan-2001	27	1.3	147	93	NaN	46	319	...
09-Jan-2001	26	1.1	123	85	NaN	36	274	...
10-Jan-2001	31	1.5	186	83	NaN	47	368	...
11-Jan-2001	17	1.2	113	80	NaN	30	253	...
:	:	:	:	:	:	:	:	⋮

Tabla 4. Datos de Contaminación (2)

Fecha	Estación	C1	C6	C7	...
21-nov-10	E59	NaN	NaN	4	...
21-nov-10	E60	NaN	NaN	4	...
22-nov-10	E1	NaN	NaN	NaN	...
22-nov-10	E2	NaN	NaN	NaN	...
22-nov-10	E35	10	400	18	...
22-nov-10	E4	6	400	25	...
22-nov-10	E39	NaN	400	32	...
22-nov-10	E6	NaN	NaN	NaN	...
22-nov-10	E7	NaN	NaN	NaN	...
22-nov-10	E8	9	500	46	...
22-nov-10	E9	NaN	NaN	NaN	...
22-nov-10	E38	13	NaN	35	...
22-nov-10	E11	NaN	NaN	54	...
22-nov-10	E12	NaN	NaN	NaN	...
22-nov-10	E40	9	NaN	23	...
:	:	:	:	:	⋮

Es de esperar que algunos de los contaminantes sean muy similares entre ellos. En el dendrograma de la Figura 2 se observa claramente una correlación de, al menos, un 80% entre el Monóxido, el Dióxido y el Óxido de Nitrógeno, así como una relación similar entre las partículas menores de 10 y 2.5 micras. En cambio, el Ozono no tiene relación directa con el resto de los contaminantes.

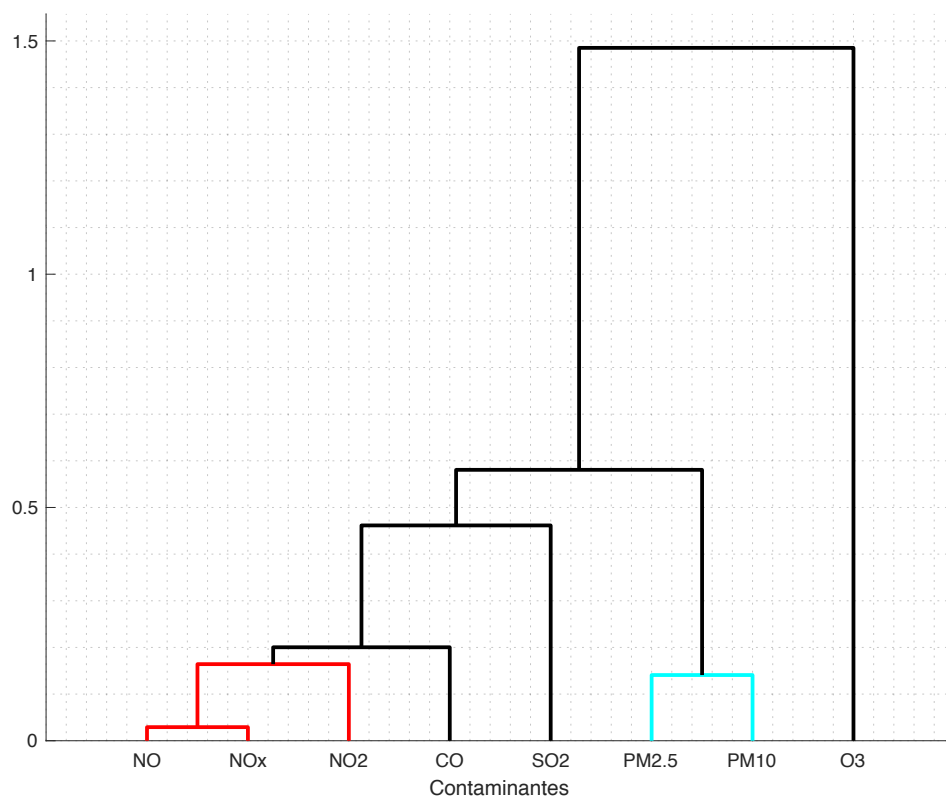


Figura 2. Dendrograma Contaminantes

Finalmente, en la Figura 3 se muestran los centroides de los *clusters* para los contaminantes. Se aprecia claramente como todos los contaminantes, exceptuando el Ozono, al aumentar la concentración de cualquier contaminante aumenta el resto. En cambio, el Ozono reduce su concentración a medida que el resto de los contaminantes aumentan confirmando la existencia de una correlación inversa.

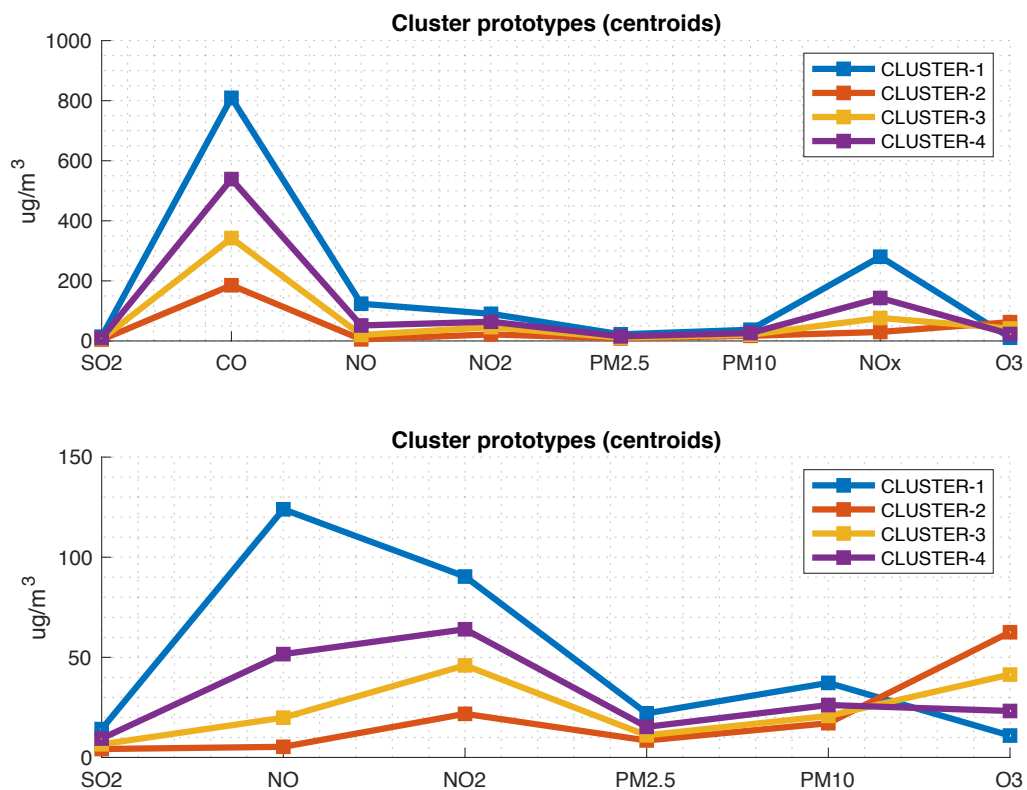


Figura 3. Centroides de los Cluster (Contaminantes)

Capítulo 5. VARIABLES EXPLICATIVAS

5.1 METEOROLOGÍA

La meteorología se considera una variable importante a la hora de explicar los niveles de calidad del aire en Madrid. La lluvia y el viento son fenómenos naturales capaces de limpiar la atmósfera. Las temperaturas y la radiación solar incidente pueden hacer que aumente o disminuya la calidad del aire de una manera indirecta ya que influye significativamente en el comportamiento de las personas. Por ejemplo, un día de temperaturas bajas las personas son más propensas a utilizar la calefacción o a moverse en algún tipo de transporte rodado en vez de ir a pie. Mediante variables como el viento, temperatura, lluvia y la radiación solar se busca comprender el comportamiento de los distintos contaminantes en Madrid. La manera de obtener estos datos está disponible en ANEXO I I – Fuentes de Información. Finalmente, las variables recogidas en las estaciones de AEMET son las temperaturas mínimas, máximas y medias, las precipitaciones, el viento medio diario y la racha máxima de viento y la duración efectiva de la insolación.

5.1.1 PREPARACIÓN DE LA INFORMACIÓN

Las cuatro estaciones de meteorología en Madrid de las que se tomaron datos otorgan una idea precisa de las temperaturas, precipitaciones, y del viento. Las estaciones se encuentran distribuidas por Madrid. Sus ubicaciones son: Retiro, Ciudad Universitaria, Cuatro Vientos y el Aeropuerto de Barajas. En la Figura 4 se muestra una representación gráfica de sus ubicaciones con respecto a las estaciones de tráfico y a las estaciones de medida de calidad del aire. Dado que las variables medidas por las estaciones de meteorología son suficientemente similares, se ha decidido trabajar con la media diaria de las cuatro, reduciendo así las cuatro variables a únicamente una.

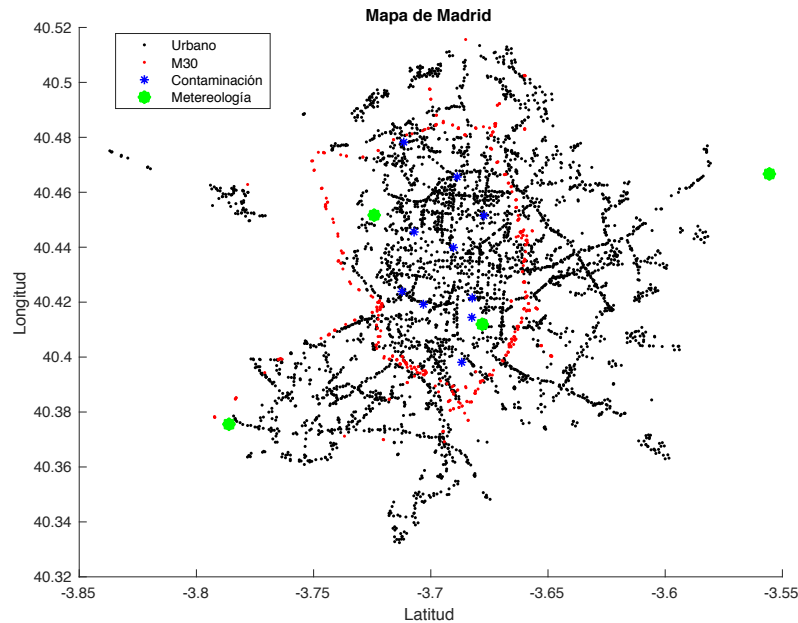


Figura 4. Ubicación de las estaciones de meteorología

5.1.2 ANÁLISIS DE LA METEOROLOGÍA

Las medidas diarias obtenidas tomando la media de las distintas estaciones da una primera visión del comportamiento de estas variables. Evidentemente, en el caso de las temperaturas se espera un comportamiento periódico (subidas en verano y bajadas en invierno). En cambio, variables como las precipitaciones tienen un comportamiento menos predecible.

Ciertas variables pueden ser más relevantes teniendo en cuenta la acumulación de los últimos días. Por ejemplo, es posible que una semana de temperaturas muy bajas sea responsable del aumento de la contaminación y no únicamente la temperatura de ese día. Por ello, además de la media diaria se añaden nuevas variables derivadas tomando medias móviles con diferentes ventanas de tiempo. En este caso, se tomaron medias móviles con ventanas de tiempo de tres, cinco y siete días lo cual quiere decir que se toma el dato de ese día y tantos días atrás como la ventana de tiempo asignada y se realiza la media. A continuación, se muestran las series temporales obtenidas para alguna de las variables.

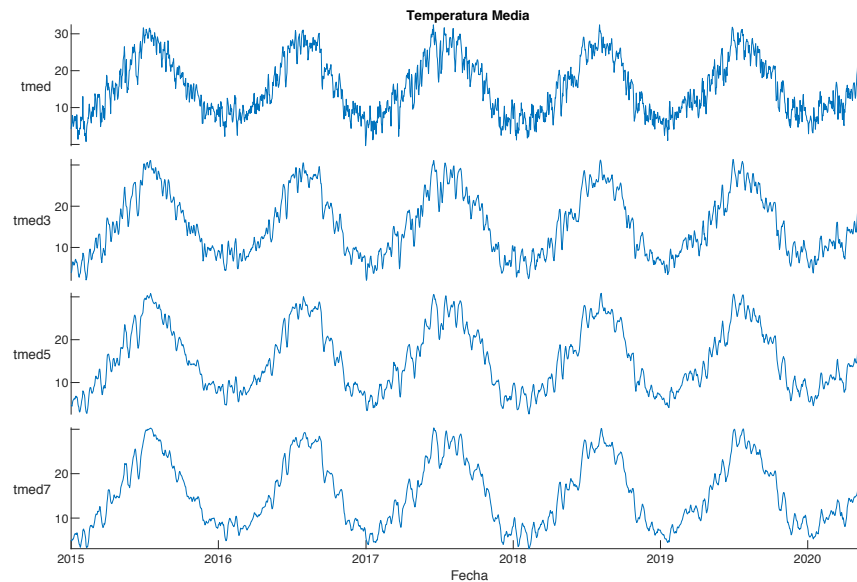


Figura 5. Variable Temperatura Media y valores suavizados con tres tamaños de ventana

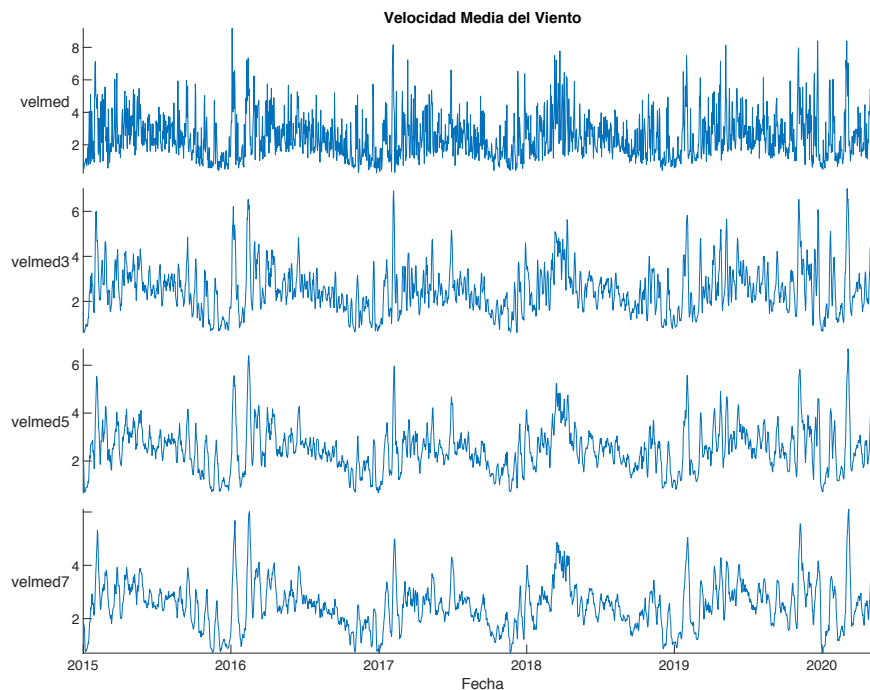


Figura 6. Variable Velocidad Media del Viento y valores suavizados con tres tamaños de ventana

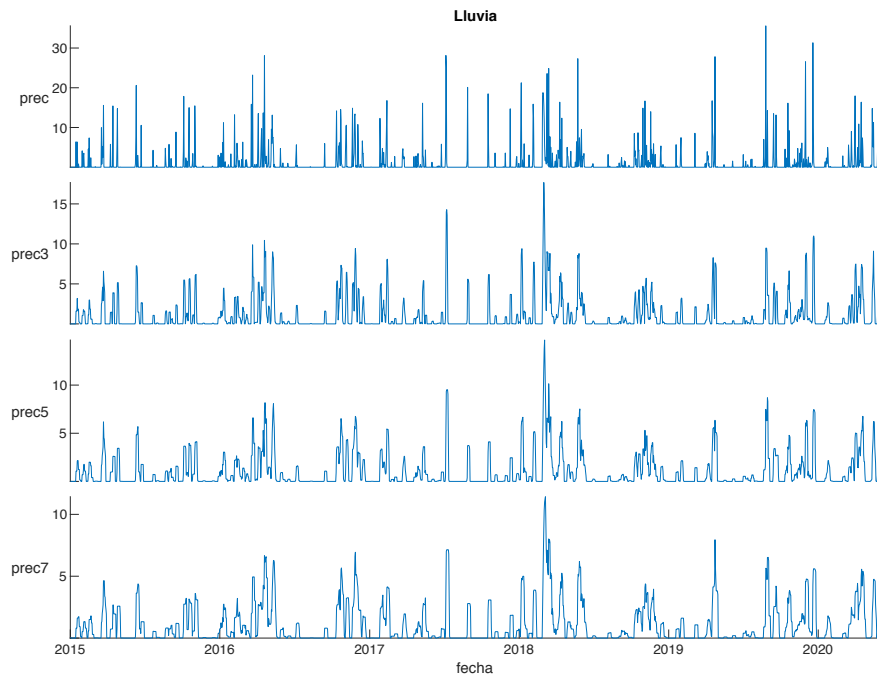


Figura 7. Variable de Precipitación y valores suavizados con tres tamaños de ventana

Para poder comprender la correlación existente entre las distintas variables se ha utilizado *clustering jerárquico*. A partir del dendrograma obtenido se aprecia una clara relación entre las variables originales y la versión suavizada de la misma variable con la media móvil según distintas ventanas de tiempo. También, se aprecia como las temperaturas máximas, mínimas y medias tienen una fuerte relación entre ellas. Más concretamente, las temperaturas tienen entre sí una correlación lineal de aproximadamente un 94% (ver Figura 8). Por norma general, la misma variable suavizada con ventanas de tiempo de cinco y siete días tienen una correlación siempre mayor al 90%. Unas variables tan correlacionadas pueden generar problemas de colinealidad a la hora de utilizar regresión para intentar explicar el comportamiento de los contaminantes.

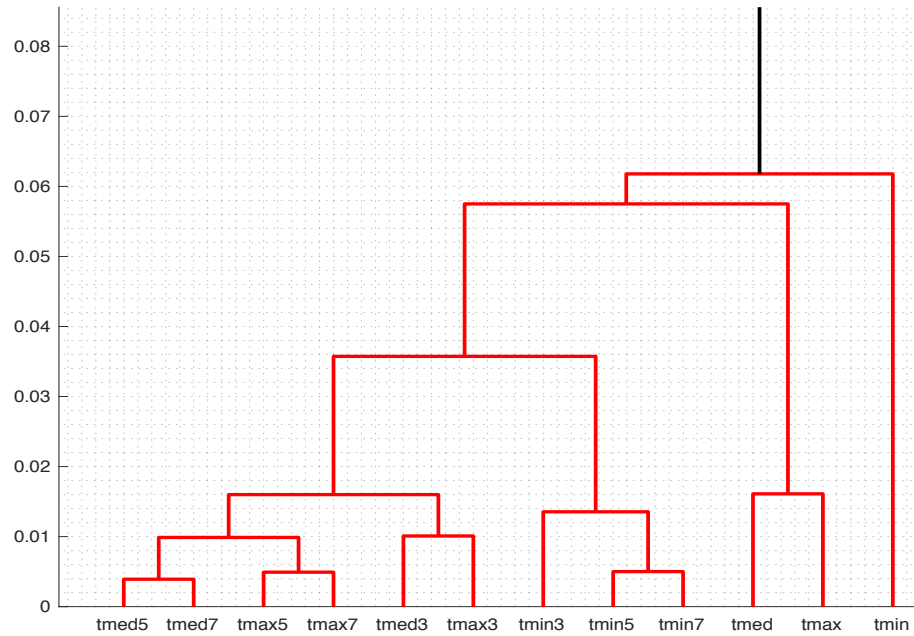


Figura 8. Dendrograma de las temperaturas (reales y suavizadas)

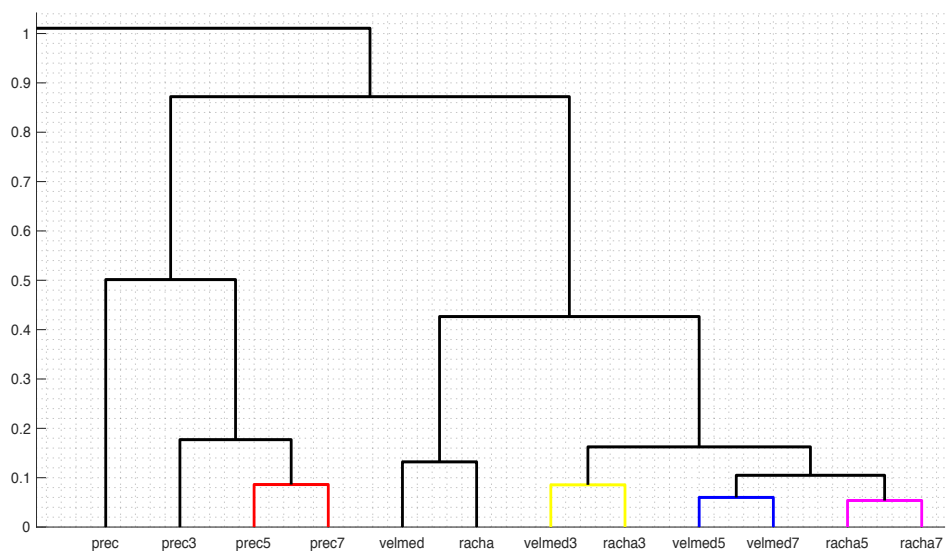


Figura 9. Dendrograma de la lluvia y del viento (reales y suavizadas)

5.2 CALENDARIO

El calendario influye directamente en el comportamiento de las personas. La festividad, por ejemplo, implica una reducción del uso de vehículos o implica una salida de Madrid hacia otras ciudades españolas. Los fines de semana se espera un descenso de los niveles de contaminación a causa del descenso de la actividad laboral. En invierno se espera un aumento de la contaminación debido al uso de, por ejemplo, la calefacción.

En 2020, debido a la pandemia sufrida con especial dureza en España y, especialmente, en Madrid, se tomaron una serie de medidas excepcionales. Esta situación de excepcionalidad es relevante debido al parón económico y social que se produjo durante los meses de marzo, abril y mayo. El estado de alarma se decreta el 15 de marzo y se mantuvo un confinamiento estricto hasta que Madrid entró en la fase 0 el 4 de mayo. En esta fase únicamente se permitió a los ciudadanos pasear o hacer deporte dentro de unas ciertas franjas horarias. El 25 de mayo se conseguía el pase a la fase 1 reactivando parte de la vida habitual de las personas. Evidentemente, durante este periodo de tiempo la calidad del aire en Madrid mejoró considerablemente, como se verá más adelante.

Por todo lo anterior, se creó un conjunto de variables asociadas al calendario para recoger todas las festividades a nivel nacional y de la Comunidad de Madrid, el día de la semana, el mes, el año y el estado de alarma con sus diferentes fases. En concreto se asigna un 0 si no es festivo y un 1 si lo es. También se han generado variables de intervención realizando lo mismo con la fase 0, fase 1, y estado de alarma.

5.3 TRÁFICO

El tráfico es, a priori, una de las variables fundamentales para el análisis de la contaminación en las grandes ciudades. Tal y como se mencionó en la descripción de los contaminantes en el Capítulo 1. , la mayoría de los contaminantes que se estudian en el presente trabajo se producen de alguna forma u otra debido al tráfico rodado.

Los datos se obtienen siguiendo el ANEXO I I – Fuentes de Información. Existen en torno a 4500 puntos de medida de tráfico repartidos en Madrid de los que se tienen datos desde el año 2013 hasta abril del 2020. Evidentemente, es necesario procesar y sintetizar toda esta información con el fin de construir alguna variable o variables capaces de resumir el comportamiento del tráfico en Madrid.

5.3.1 PREPROCESADO DE LA INFORMACIÓN

En este apartado se describe el proceso seguido para construir un conjunto razonable de variables de calidad que recojan el comportamiento diario del tráfico a partir de las variables disponibles para los 4500 puntos de medida existentes. Lograr este objetivo no es sencillo ya que existen huecos aleatorios en los datos, hay medidas erróneas y además los puntos de medida cambian con el tiempo, ya que durante los años se han ido añadiendo nuevos y eliminando antiguos.

Para poder ligar cada estación de contaminación con un grupo de las 4500 estaciones de tráfico próximas, se debe obtener la información de la latitud y longitud de todas las estaciones involucradas. Con esta información se pueden relacionar las estaciones siguiendo el criterio de distancia euclídea mínima. Al existir 24 estaciones de contaminación, cada estación tendrá varios puntos de medida o estaciones de tráfico asignadas. A partir de la latitud y longitud de las estaciones se representado el gráfico de Figura 10 para comprender mejor la relación existente entre las estaciones de contaminación y los puntos de medida de tráfico. Cómo se puede observar, el elevado número de puntos de medida permiten reconocer las principales calles de Madrid, así como el anillo de la M30.

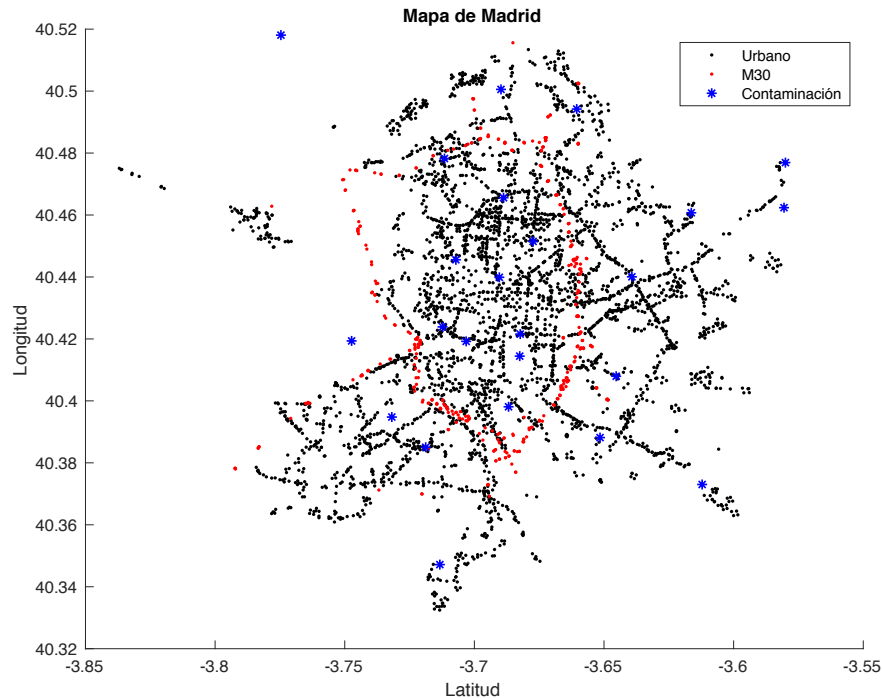


Figura 10. Localización de las estaciones de contaminación y los puntos de medida de tráfico

Con el fin de comprender la calidad de los datos de una manera sencilla, se presenta en la Figura 11 los datos recogidos de las estaciones de medida de tráfico más próximas a la estación de contaminación situada en Plaza de España. En la figura se observa que los datos en los años 2013 y 2014 parecen de una calidad menor debido al pequeño número de estaciones activas entonces. Por ello, los datos de tráfico solo serán considerados desde el 1 de enero de 2015 en adelante. Además, la existencia de datos anómalos (picos repentinos y de gran magnitud) complican la manipulación de los datos.

Una forma de reducir el número de variables es tomando la media de todas las estaciones. Con este método sencillo se simplifica el análisis de la información, aunque existen unos inconvenientes que se muestran en la Figura 12. Un alto número de estaciones dejan de recopilar datos durante un corto periodo de tiempo lo cual lleva a que la media aumente de manera repentina o, cuando todas dejan de medir al mismo tiempo, no exista información

alguna (posiblemente debido a un error en los servidores del Ayuntamiento de Madrid). Por ello, la media no representa correctamente lo que realmente está ocurriendo en el tráfico.

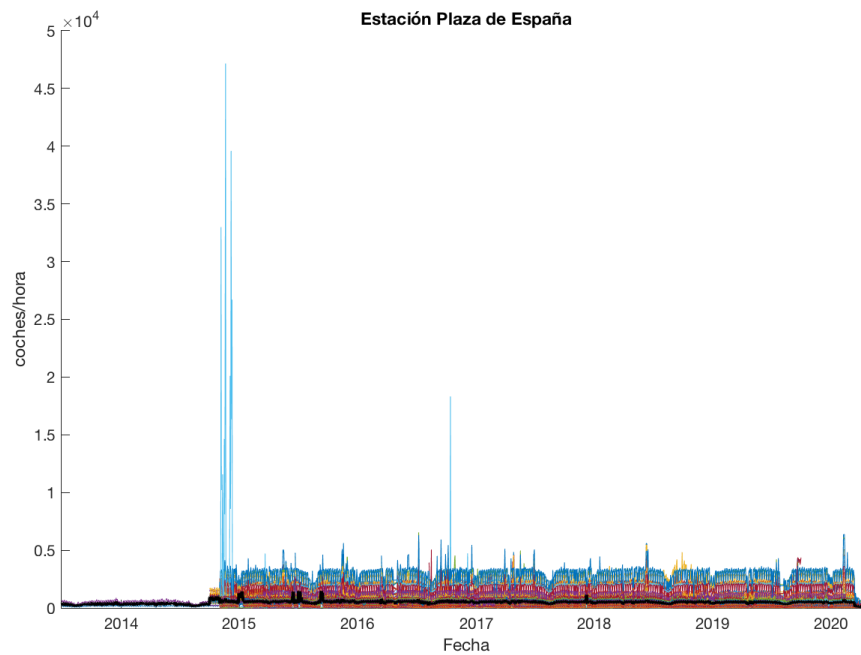


Figura 11. Intensidad de tráfico recogido en los puntos de medida cercanos a Plaza de España

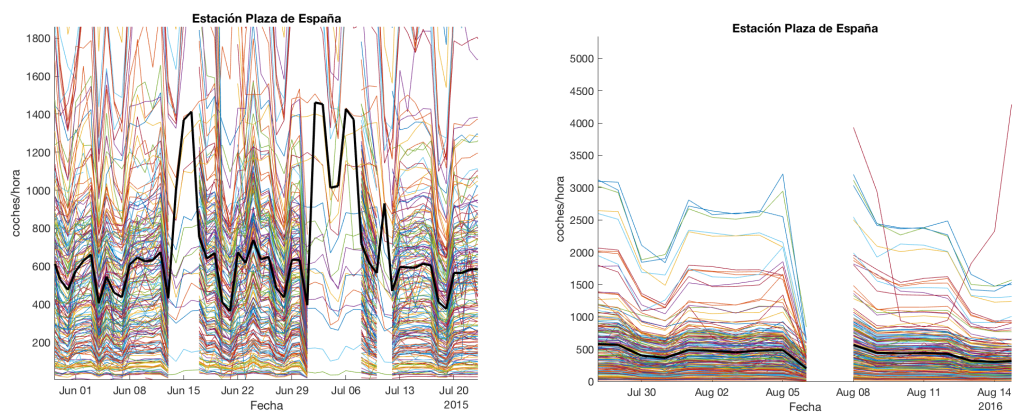


Figura 12. Discontinuidades en las series de tráfico (atípicos y huecos)

Además, las estaciones pueden fallar y dejar de medir durante un tiempo. Esto ocurre por fallos en las estaciones, este siendo reparada, mantenimiento de la estación o errores en los servidores o aparatos de medida. La falta de información o NaN (*not a number*) se puede representar gráficamente utilizando la función de *Matlab pcolor*. Esta función muestra sobre un plano una representación gráfica de la calidad de los datos. Los huecos en blanco son falta de información o NaNs. Esta gráfica se ha utilizado para cada una de las estaciones de contaminación para su análisis. A continuación, se muestra en la Figura 13 un ejemplo para la estación de Plaza de España, con 210 estaciones de tráfico asignadas. En el eje horizontal están los días del periodo de datos considerado y en el eje vertical cada una de las estaciones de tráfico.

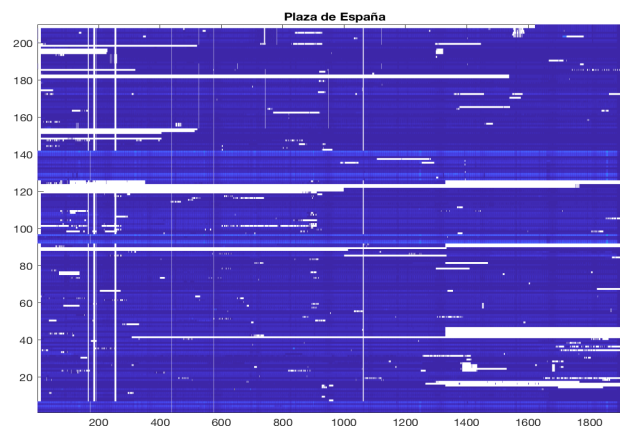


Figura 13. Calidad de los datos recogidos en los puntos de medida de tráfico

A partir de la Figura 13, se aprecia que hay muchas estaciones que presentan varios periodos en los que sus medidas fallan. Con el objetivo de tener un conjunto de variables de calidad del tráfico que tengan medidas para el periodo considerado, se ha decidido mantener los datos de aquellas estaciones con un alto número de lecturas correctas, eliminando las demás estaciones. En la Figura 14 se observa el resultado obtenido una vez realizado dicho filtrado, con una clara mejoría de la información, aunque el número de estaciones se ve reducido en 184.

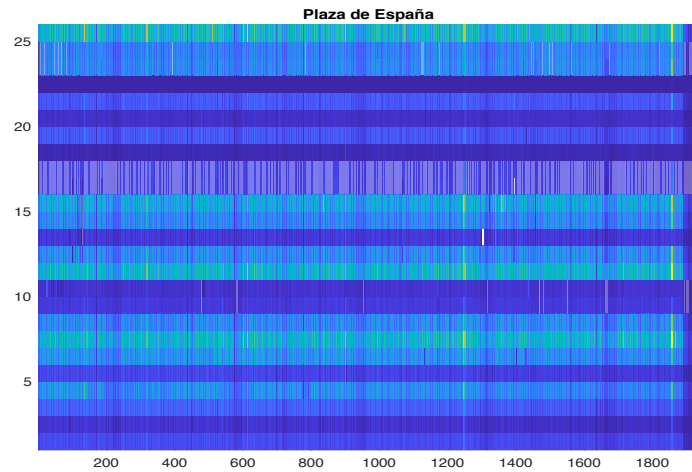


Figura 14. Calidad de los datos recogidos en los puntos de medida de tráfico después del filtro

Este proceso se repite para las 24 estaciones de contaminación. La calidad de los datos mejora enormemente, pero se deben sacrificar una alta cantidad de puntos de medida. Visualmente, en la Figura 15 se muestra un mapa con aquellas estaciones que han sido eliminadas en negro y aquellas que se mantienen en verde.

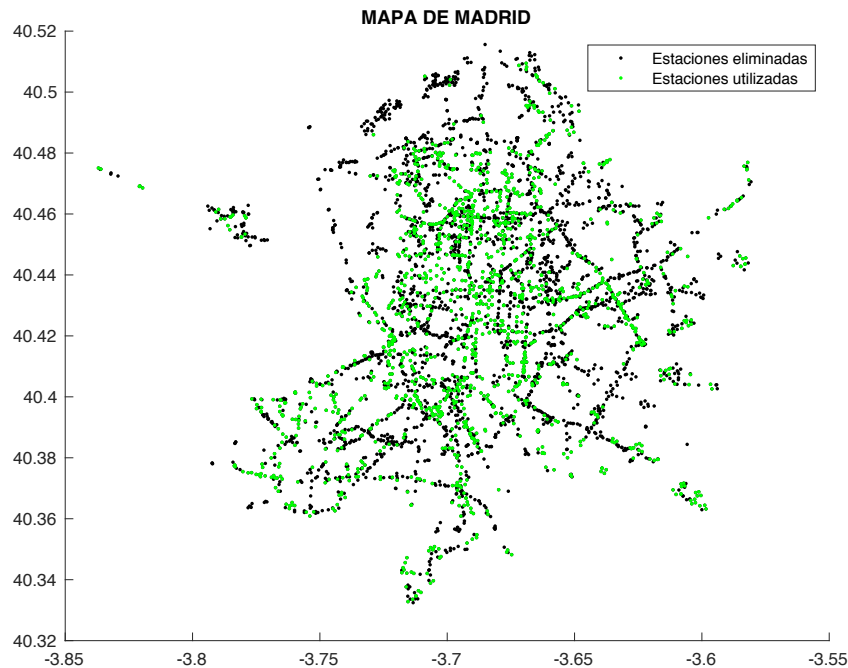


Figura 15. Mapa de las estaciones eliminadas/mantenidas

5.3.2 ANÁLISIS DE TRÁFICO

Una vez reducido el número de errores de los datos se procede al análisis de la información. Principalmente se destacan dos cosas: la primera es que cada año se ve una caída fuerte del intensidad de tráfico diario en agosto (ver Figura 16), y la segunda, debido a la situación de excepcionalidad que se ha dado en España durante los meses de marzo y abril por causa del confinamiento se aprecia una clara bajada del tráfico (Figura 17). Además, se aprecia un aumento progresivo de lunes a viernes mientras que los fines de semana el tráfico tiende a reducirse (Figura 18). Nótese que algunas estaciones de medida registran valores mucho más elevados que otras. Esto ocurre debido a que la intensidad de tráfico mide la media del número de coches por hora en un día por lo que es de esperar que calles como la M-30, la Castellana o Joaquín Abascal tengan una densidad de tráfico mucho mayor a otras.

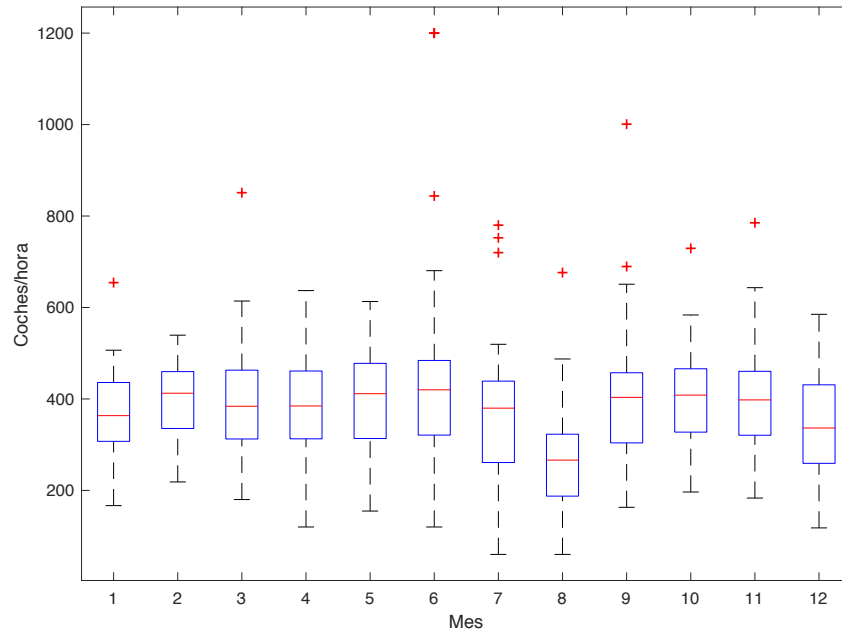


Figura 16. Tráfico Mensual

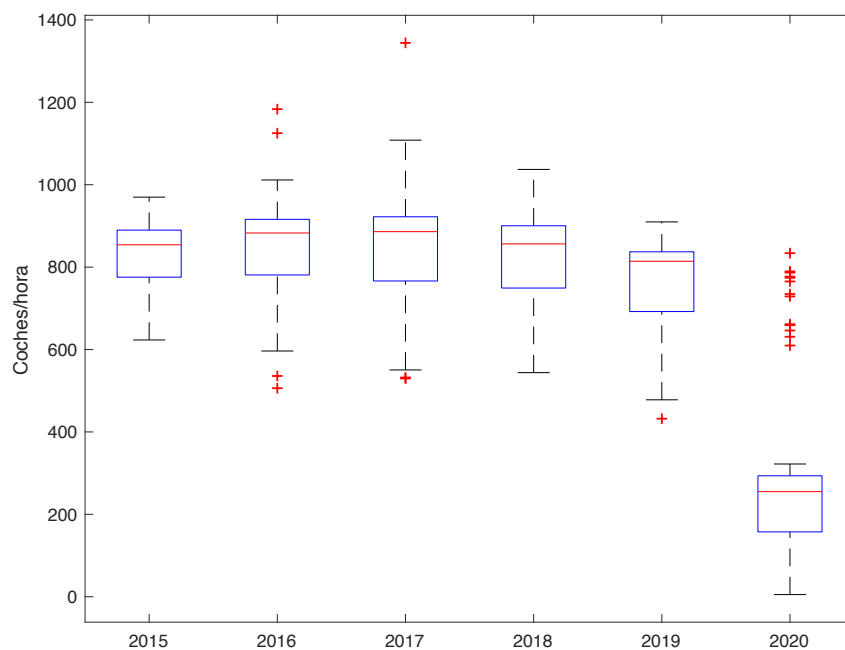


Figura 17. Tráfico Anual

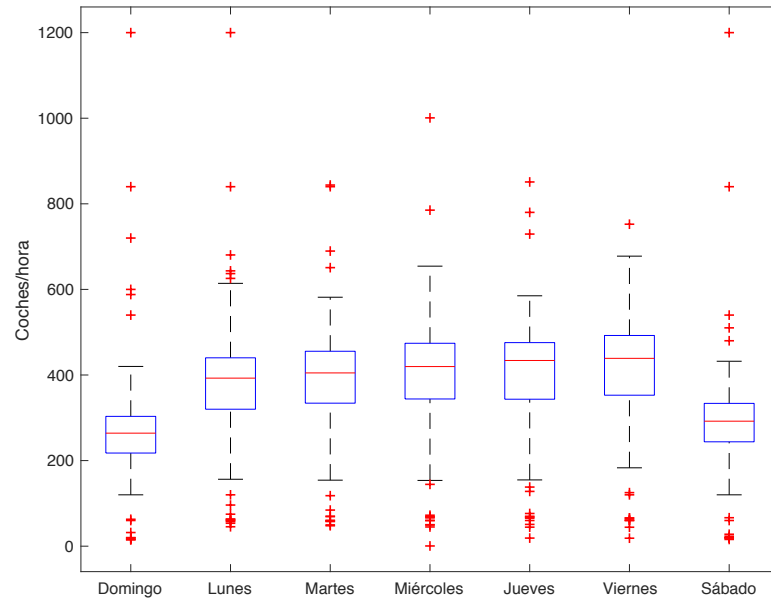


Figura 18. Tráfico Semanal

La Figura 17 se realizó únicamente considerando los meses de marzo y abril.

Aunque la depuración realizada para seleccionar un conjunto de puntos de medida de tráfico de calidad ha obligado a rechazar muchas variables, el número de estaciones de tráfico sigue siendo demasiado alto como para utilizarlas directamente para explicar la contaminación. Por ello, es necesario ordenar y simplificar la información restante. Para lo que se ha utilizado *clustering jerárquico*. Mediante esta técnica de aprendizaje automático se agrupan las estaciones que sean similares entre sí en grupos o clústeres. Por ejemplo, para la estación de contaminación situada en Plaza de España en la Figura 19 se muestra el dendrograma con el criterio de agrupar aquellas estaciones de tráfico que tengan una correlación mayor o igual a 0.7. Esto quiere decir que la altura del dendrograma donde se lleva a cabo el corte es de 0.3.

Debido a la situación especial causado por el estado de alarma en España solamente se utiliza la información existente desde el 2015 hasta el 2019 (inclusive) a la hora de realizar este análisis.

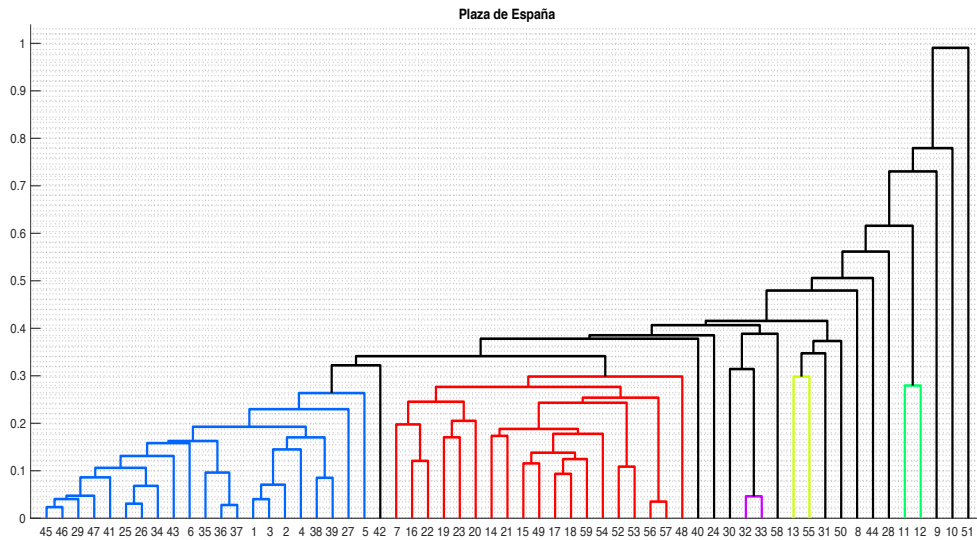


Figura 19. Dendrograma de Tráfico (Plaza de España)

En la Figura 19 se observa que existen dos *clusters* con un alto número de puntos de medida de tráfico y, bajo el criterio de una correlación mínima de 0.7, hay varias estaciones que son independientes y forman sus propios grupos. Para reducir el nº de variables, se ha optado por simplificar los grupos que contienen varios puntos de medida, convirtiendolos en una única medida tomando una media ponderada de los miembros del grupo. En el caso anterior antes de llevar a cabo la agrupación había un total de 59 puntos de medida distintos. Después de realizar el método propuesto, quedan solamente 18 puntos de medida. Al igual que se hizo en casos anteriores, este proceso se repite para las 23 estaciones de contaminación restantes para simplificar con el mismo mecanismo la información de tráfico asociado a cada estación de contaminación.

5.3.3 OBTENCIÓN DE VARIABLES EXPLICATIVAS

A partir de lo anterior, se ha optado por obtener, mediante dos vías distintas, varias variables razonables capaces de recoger de manera robusta lo que ocurre con el tráfico. Dos de las vías incluyen, además del *clustering* usado anteriormente, la reutilización de este método para seguir agrupando el mayor número de variables que compartan una cierta correlación. La otra alternativa utilizada se basa en el método del análisis de las componentes principales o PCA, por sus siglas en inglés (*principal component analysis*).

5.3.3.1 Método 1

El primer método consiste en únicamente considerar los *clusters* obtenidos en la zona interior de la M-30 de Madrid y unir todos los *clusters* de las distintas estaciones pertenecientes a dicha zona.

Concretamente, la zona interior de la M-30 presenta un total de 130 variables de tráfico. Evidentemente, es un número demasiado elevado de variables explicativas para intentar explicar la dinámica de los diferentes contaminantes, por lo que deben de reducirse. Para ello, se eliminan todas los *clusters* que tengan más de 5 días de error comprendidos entre el 1 de enero de 2015 y el 31 de diciembre de 2019 reduciéndose así a un total de 13 variables que se muestran en la Figura 20. Estas variables están formadas por cinco *clusters* de la estación de Plaza de España, cinco de la estación de Méndez Álvaro y dos de la estación de Retiro.

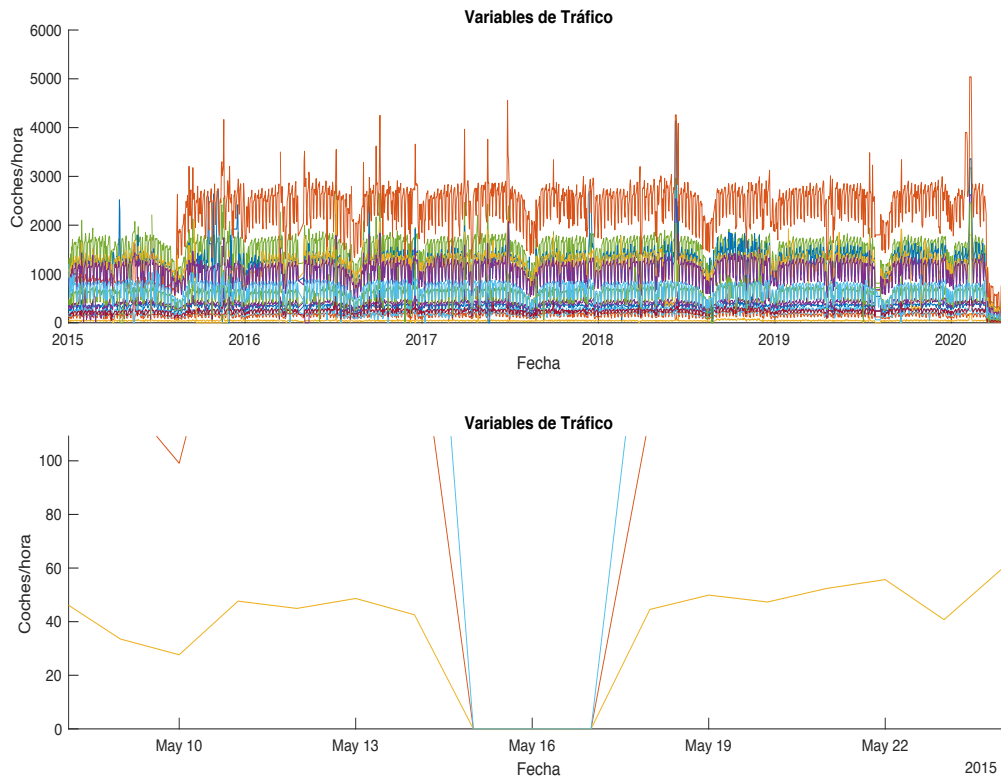


Figura 20. Variables de Tráfico preseleccionadas en el método 1

En la Figura 20 se aprecia que hay varias variables que presentan intervalos temporales en los que no se detecta ningún coche. Esto se produce como consecuencia de la función de Matlab *nanmean* que, en caso de que todos los puntos sobre los que se realiza la media son falta de información o NaN (*not a number*), se contabiliza como un cero. Al tener un número elevado de variables se eliminan aquellas que presenten algún cero durante su serie temporal para evitar el mayor número de errores posibles en las variables. Finalmente, el número de variables se ve reducido a un total de tres, cuyas evoluciones temporales se muestran en la Figura 21. Estas variables son las que se han utilizado como variables de entrada candidatas en el modelo de regresión para estimar el nivel de los contaminantes.

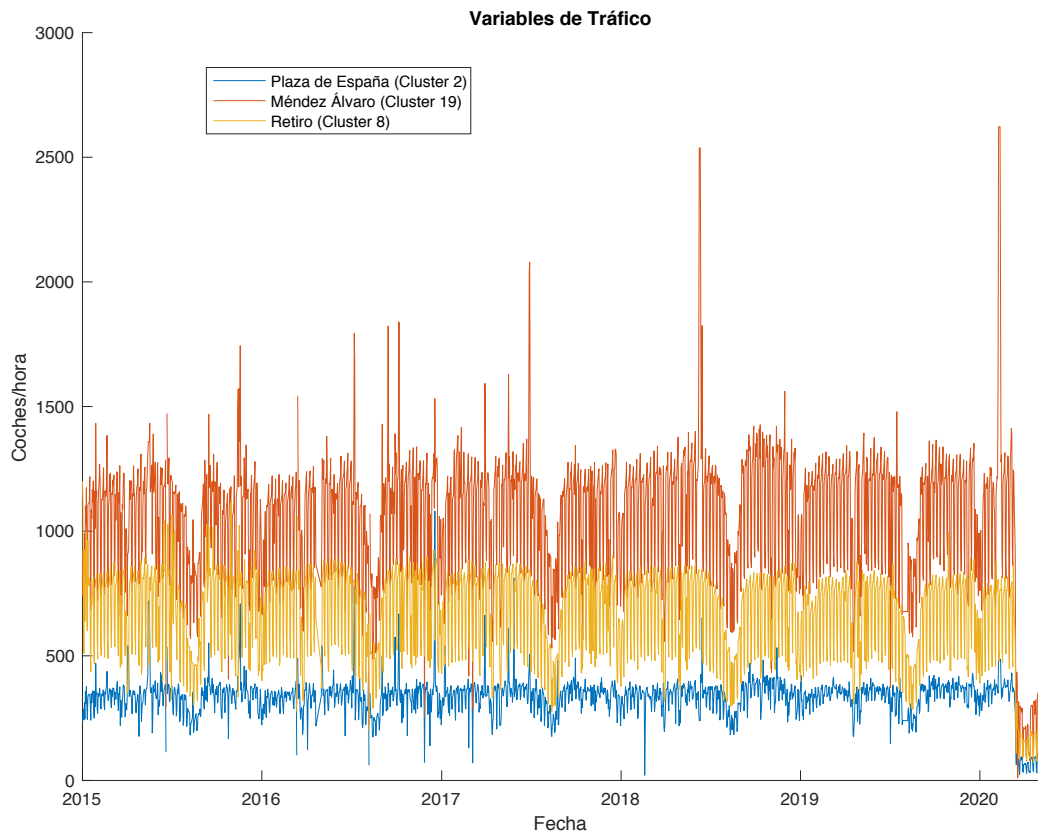


Figura 21. Variables de Tráfico construidas con el método 1

5.3.3.2 Método 2

Finalmente, mediante el análisis de las componentes principales y únicamente teniendo en cuenta los clusters pertenecientes a la zona interior a la M-30, se pretende encontrar un número reducido de componentes que expliquen una cantidad razonable de la variabilidad de los datos. Este método requiere que todas las filas de la tabla que no dispongan de información se eliminen antes de realizar el método. Por ello, primero se eliminan aquellas estaciones, al igual que en los métodos anteriores, que contengan alguna medida igual a cero en su evolución temporal.

Originalmente, todos los grupos encontrados en la zona interior a la M-30 de Madrid presentan un total de 130 variables que, tras eliminar los grupos con medidas igual a cero,

se reducen a un total de 93. A continuación, se eliminan aquellas filas de la tabla que en alguna columna presenten falta de información o NaN perdiendo así 77 días desde el 1 de enero de 2015 hasta el 30 de abril de 2020. Teniendo en cuenta que ese intervalo de fechas considerado supone un total de 1937 días, es razonable perder 77 días de datos. Por tanto, se eliminan otras cuatro variables que presentan unos picos en sus series sin justificación aparente y que únicamente aportan ruido al método. Por tanto, el análisis de componentes principales se realiza sobre las 89 variables restantes.

La Figura 22 muestra el gráfico de sedimentación utilizado para seleccionar el número de componentes principales. Según dicho gráfico, es razonable seleccionar únicamente dos componentes principales. En este caso, una única componente principal es capaz de explicar el 97.27% y dos componentes principales un total de 97.75%. Añadir más componentes apenas mejora la varianza explicada.

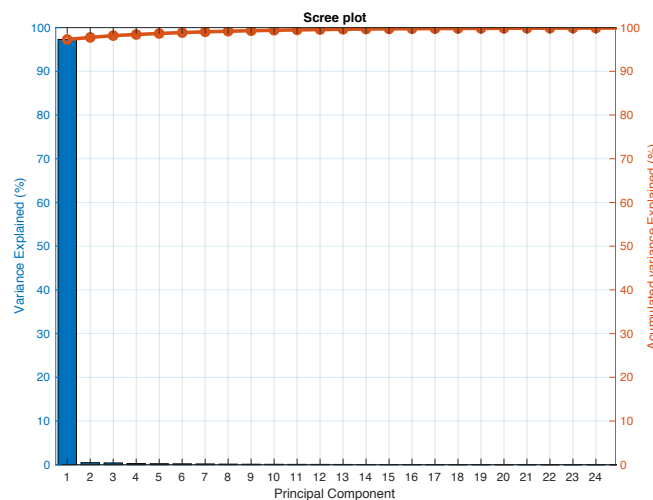


Figura 22. Gráfico de Sedimentación para las variables de tráfico

Como era de esperar, los valores de ambas componentes principales del tráfico oscilan alrededor del cero, siendo la primera componente mucho más importante que la segunda, según se observa en la Figura 23. Ambas componentes muestran un claro descenso en el mes de agosto y otro muy evidente en el año 2020 durante el confinamiento.

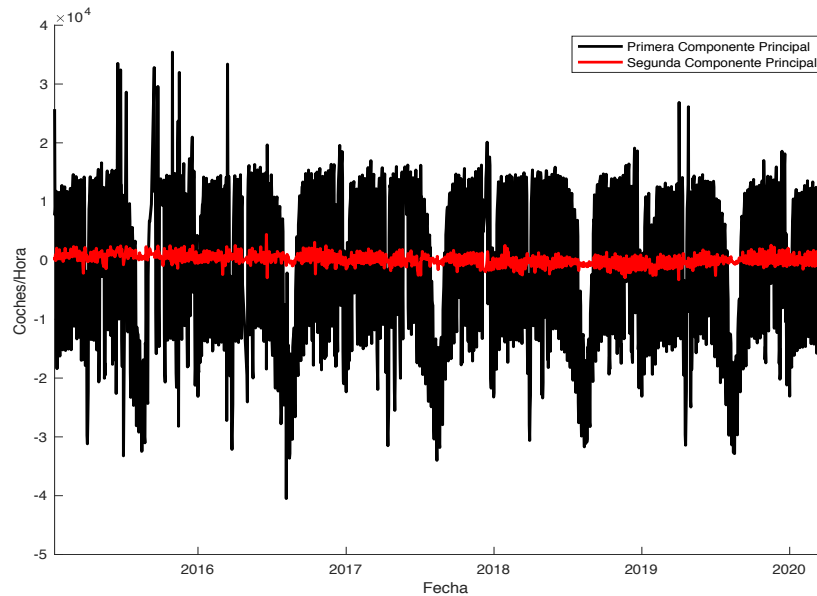


Figura 23. Evolución Temporal de las dos primeras componentes principales del tráfico

A partir de los autovectores de la matriz de covarianza formada por las 89 variables consideradas se pueden conocer qué variables de tráfico son las más influyentes en la primera y segunda componente principal que han sido seleccionadas. Únicamente se considerarán influyentes aquellas con un valor asociado mayor que 0.1 y solo se analiza la primera componente principal ya que esta explica más del 95% de la variabilidad. Existen 8 variables con un peso superior a 0.1 como se puede apreciar en la Figura 24.

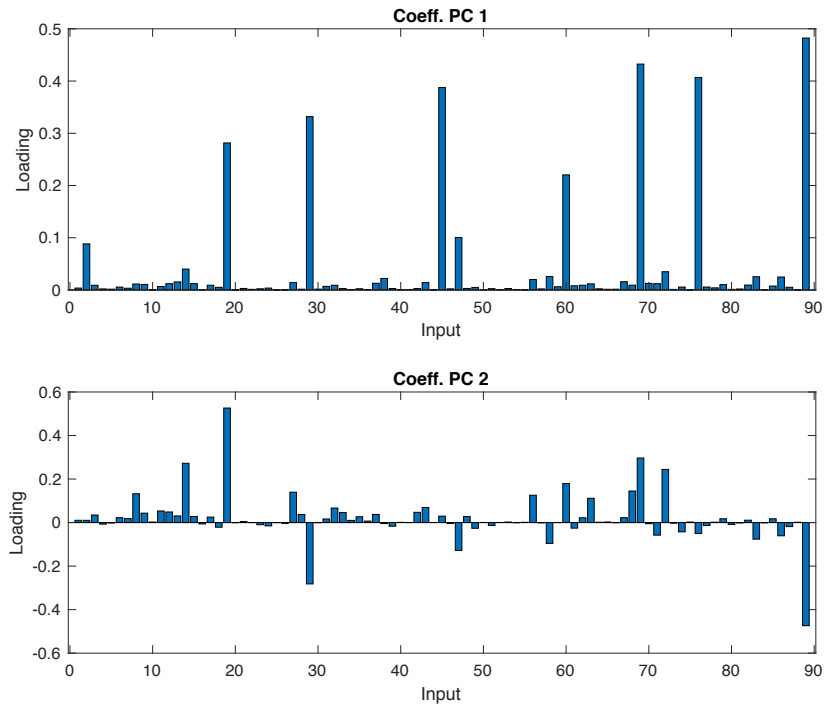


Figura 24. Peso de las Variables (Primera y Segunda Componente Principal)

Las variables con un peso importante en la primera componente principal se corresponden con calles amplias con varios carriles que permiten un enorme flujo de vehículos por hora. Concretamente, se trata de calles como Paseo de la Castellana, Goya, Velázquez, José Abascal, Bravo Murillo, Miguel Ángel, Serrano, Cea Bermúdez o la Avenida de Asturias entre otras muchas calles del estilo.

Capítulo 6. DIÓXIDO DE AZUFRE

En este capítulo se analiza en detalle los niveles de contaminación del dióxido de azufre en Madrid. Para ello se estudia la evolución temporal de la serie de dicho contaminante, ajustando un modelo de regresión lineal múltiple para estimar el nivel de contaminación en función de las variables explicativas seleccionadas.

6.1 LÍMITES, MÁXIMOS Y MÍNIMOS

Los límites legales y los umbrales de información se fijaron mediante el Real Decreto del 28 de enero de 2011 [8]. Los umbrales de información superior e inferior se deben de respetar para proteger de la salud de las personas.

El valor más alto registrado es de $93 \mu\text{g} / \text{m}^3$ y se dio el 22 de febrero de 2001 en la Plaza del Dr. Marañón. Únicamente en cuatro ocasiones desde el 2001 se ha medido una concentración igual a 0.

6.1.1 LÍMITES LEGALES

El valor límite diario del dióxido de azufre es de $125 \mu\text{g} / \text{m}^3$ (ver Tabla 17). Este valor no se ha visto superado en ninguna ocasión desde el 1 de enero de 2001 hasta el 31 de mayo de 2020 en ninguna de las estaciones de Madrid.

6.1.2 UMBRALES DE INFORMACIÓN

Los umbrales superior e inferior de evaluación diarios son de $75 \mu\text{g} / \text{m}^3$ y $50 \mu\text{g} / \text{m}^3$ (ver Tabla 18) que no deberán superarse en más de 3 ocasiones por año.

El umbral superior se sobrepasó un total de 11 veces, siendo todas ellas en el año 2001. Mientras que el umbral inferior se superó en 134 ocasiones. De las 134 ocasiones en las que se superó el umbral inferior, 76 veces fueron en 2001, 18 en 2002, 11 en 2003, 6 en 2004,

12 en 2005, 3 en 2006 y 2007, y 5 en 2008. A partir del 2008 no se superó en ninguna ocasión.

6.2 ANÁLISIS

Con el fin de comprender el comportamiento del Dióxido de Azufre (SO_2) se debe de analizar su comportamiento e intentar comprender que variables influyen en dicho contaminante. Para ello, es imprescindible analizar si todas las estaciones de Madrid tienen datos similares, se debe de analizar la serie temporal del contaminante y, finalmente, identificar las variables que pueden influir en la variabilidad del contaminante.

Analizando la serie temporal (Figura 25) se observa un descenso progresivo en los niveles de contaminación desde el año 2001 hasta el 2013 en todas las estaciones de medida. Desde ese momento, se aprecia una cierta estabilidad hasta alcanzar 2019 donde hubo un repunte de contaminación. Además, en verano el SO_2 disminuye claramente con respecto a los meses de enero y diciembre. Nótese que la media de todas las estaciones se representa en negro en la Figura 25.

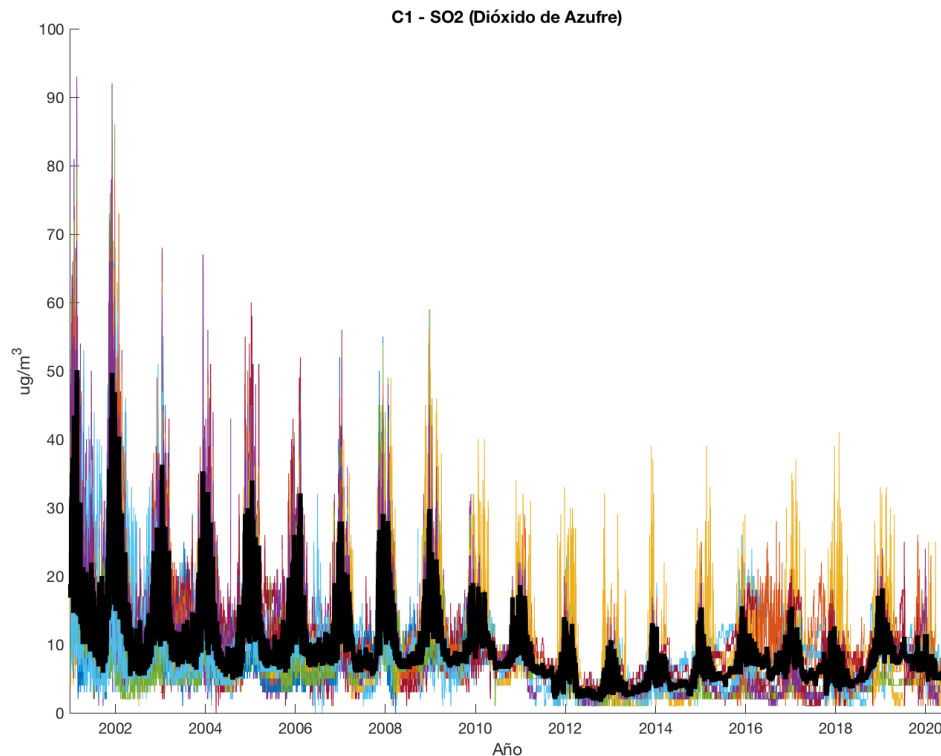


Figura 25. Serie Temporal del Dióxido de Azufre en las diferentes estaciones

A continuación, mediante el método de agrupación de *K-means* se han determinado conglomerados de días caracterizados por tener unos niveles de contaminación similares en todas las estaciones de medida. Para poder utilizar este método es necesario elegir previamente el valor adecuado de *K*, de tal forma que haya una buena relación entre complejidad y error de cuantización (QE). En la Figura 26 se muestra el criterio utilizado para seleccionar el valor de *K*. En este caso particular no existe un codo evidente, aunque un valor de *K* razonable podría ser 3.

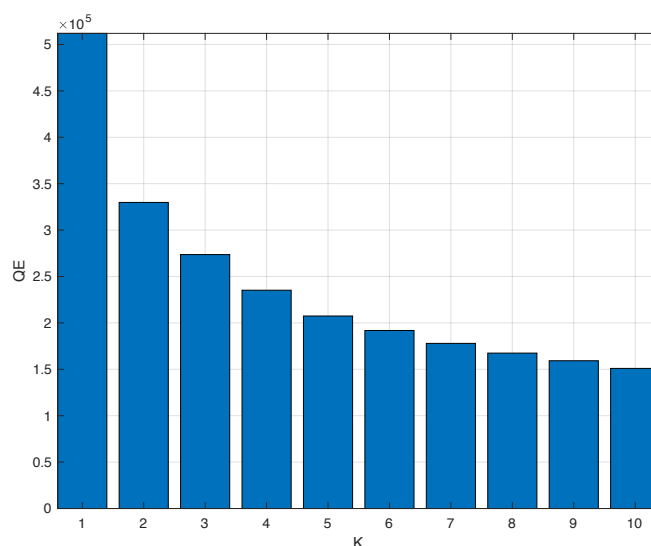


Figura 26. Obtención del número de conglomerados (SO_2)

Una vez seleccionado el número de conglomerados (K) se procede a representar gráficamente los distintos grupos obtenidos, en donde cada clúster representa un grupo de días similares en cuanto a niveles de concentración en las diferentes estaciones de medida. En la Figura 27 se muestran los centros de los conglomerados (centroides) por simplicidad, observándose que el clúster 1 representa los días de mayores niveles de contaminación, mientras que en el clúster 3 se encuentran los días de menor nivel de concentración del contaminante. Además, también se aprecia claramente que la estación con medidas más bajas en los tres clústeres de este contaminante es la estación de la Casa de Campo, lo cual es de esperar ya que se encuentra en una zona con bajo tráfico y alejado de la zona urbana de Madrid. En cambio, es llamativo que en la estación de Moratalaz, con unos niveles relativamente bajos de contaminación en el clúster 3, en el clúster 1 es la estación donde la contaminación alcanza los niveles más elevados, superiores a los que aparecen en la estación de Escuelas Aguirre o Sanchinarro, que son las que presentan valores más elevados en comparación con el resto de las estaciones en los clústeres 2 y 3.

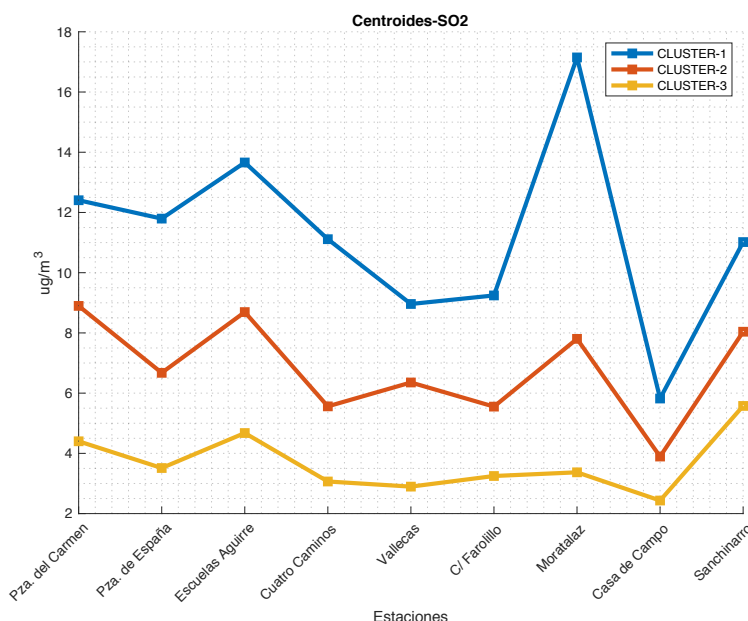


Figura 27. Centroides de los conglomerados (SO_2)

Para comprender mejor el comportamiento del Dióxido de Azufre, se procede a analizar si los niveles de concentración de dicho contaminante cambia y cómo con variables discretas como el mes, el día de la semana, la festividad o la estación de medida y las posibles interacciones entre ellas. Para ello, se ha realizado un análisis de la varianza (ANOVA). El resultado de la prueba se encuentra en la Figura 28. Todas las variables influyen en la concentración debido a que presentan un *p-valor* ($Prob > F$) igual a cero.

Este resultado quiere decir que hay al menos un grupo de cada variable para el que el contaminante es claramente distinto los valores que toma para el resto de grupos. Por ejemplo, si se realiza una prueba de comparación múltiple para el mes, se puede observar que hay muchos meses que son diferentes del resto, observándose una bajada importante de los niveles de contaminación entre los meses de abril y septiembre (ver Figura 29). Realizando la misma comparación para el resto de las variables consideradas se aprecia que la contaminación es menor durante el fin de semana y los festivos. Además, la estación de Casa de Campo mide una concentración considerablemente menor al resto de estaciones, pero la estación de Moratalaz no es claramente diferente al resto.

Finalmente, las interacciones entre las variables muestran que únicamente la interacción entre las estaciones y el mes tienen un *p-valor* suficientemente bajo (0.0008) como para considerar que influye en la concentración del contaminante.

Analysis of Variance					
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
month	76941.4	11	6994.67	492.06	0
Día de la semana	1631.8	6	271.96	19.13	0
Estacion	69040.5	9	7671.17	539.65	0
Festivo	379.8	1	379.81	26.72	0
Error	532748.2	37478	14.21		
Total	680035	37505			

Constrained (Type III) sums of squares.

Figura 28. Test Anova (SO_2)

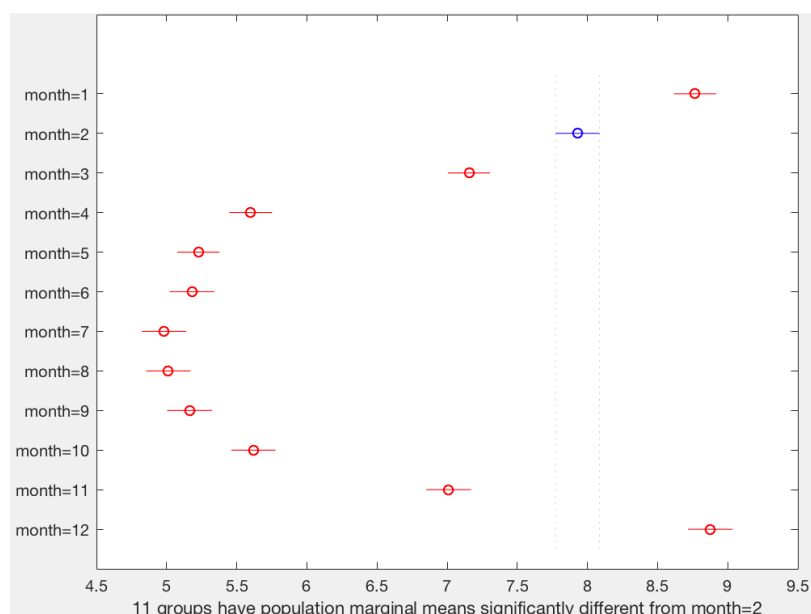


Figura 29. Prueba de Comparación Múltiple para el mes (SO_2)

6.3 MODELO EXPLICATIVO DEL NIVEL DE CONCENTRACIÓN

Utilizando las variables creadas durante el Capítulo 5. y añadiendo como variable la diferencia de concentración de Dióxido de Azufre entre los dos días anteriores se ha planteado un modelo de regresión lineal múltiple para explicar el nivel de contaminación en función de las mismas. El modelo se ha ajustado con los datos de 2015 a 2018.

Dado que se dispone del nivel de concentración del Dióxido de Azufre en diferentes estaciones, se ha estudiado como variable de salida la media de todas las estaciones pertenecientes a la zona interior a la M-30. Además, en el caso del Dióxido de Azufre se utilizaron las variables de tráfico obtenidas mediante el método del análisis de las componentes principales. La variable Trafico1 representa la primera componente principal y la variable Trafico2 la segunda.

Inicialmente, se plantea un modelo con todas las variables disponibles y se van eliminando aquellas con un *p-valor* mayor. Posteriormente, se prueban distintas interacciones o relaciones cuadráticas con el fin reducir la raíz del error cuadrático medio (RMSE) y aumentar el coeficiente ajustado de determinación (R^2).

Finalmente, una vez obtenido el modelo con un número razonable de variables explicativas se impone la condición de que en ningún caso sea menor que cero el valor estimado por el modelo. El modelo finalmente obtenido consta de ocho términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 2.2396$$

$$R^2 = 0.3233$$

Un coeficiente de determinación tan bajo implica que no se trata de un buen ajuste. El análisis de residuos (ver Figura 30) muestra que los residuos claramente no son independientes. Se está estimando por debajo y por arriba periódicamente confirmando que no se trata de un buen modelo. Este contaminante ha sido muy cambiante de un año a otro y no se tiene una variable capaz de explicar estas variaciones. En la Figura 31 se puede

observar la verdadera concentración y la estimación realizada por el modelo. Claramente, se aprecia la gran variación de este contaminante y la dificultad del modelo para explicar dichas variaciones.

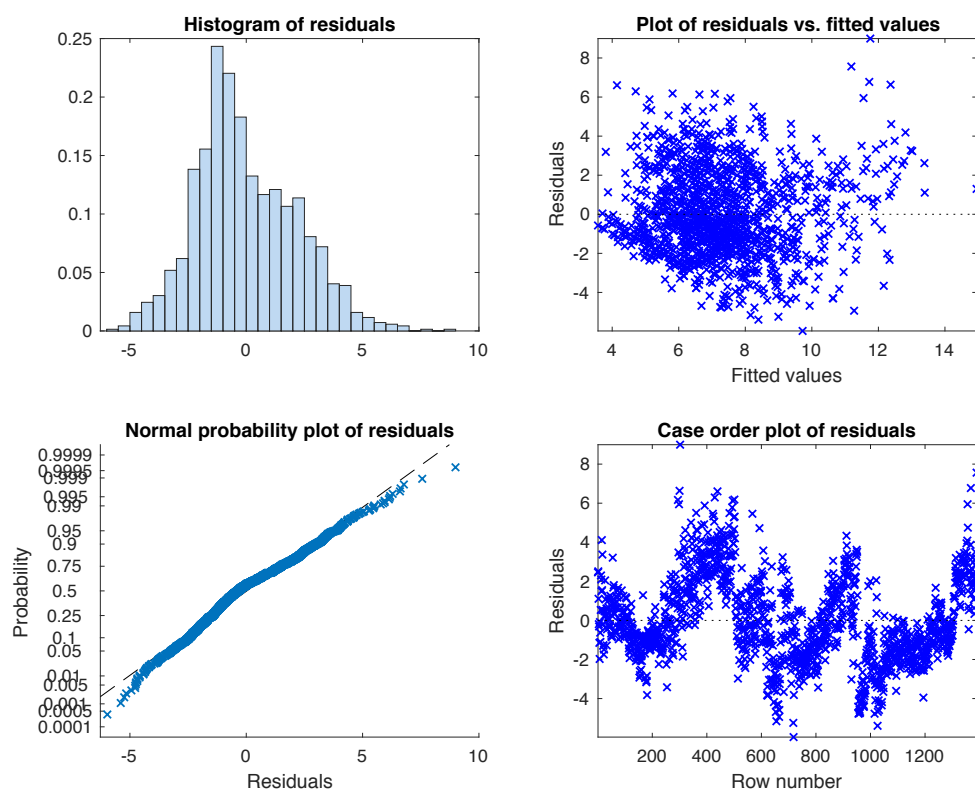


Figura 30 Análisis de Residuos (Dióxido de Azufre)

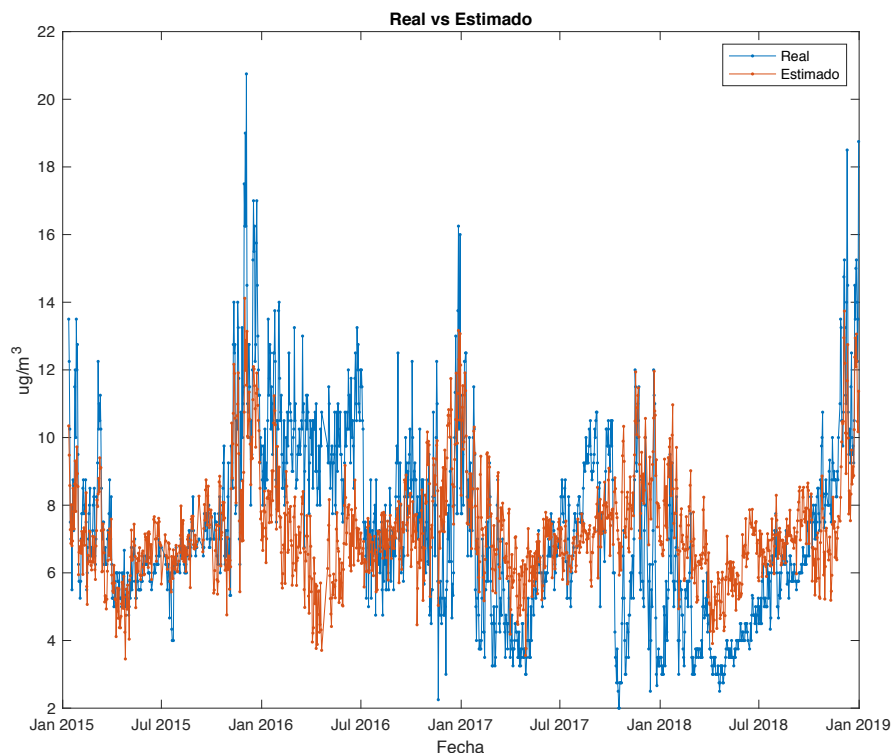


Figura 31. Estimación vs Real (Dióxido de Azufre)

El modelo obtenido consta de 8 términos y 18 coeficientes. En la Tabla 5 se muestran todas las variables explicativas utilizadas en el modelo del Dióxido de Azufre.

Tabla 5. Diagnóstico e interpretación del modelo (Dióxido de Azufre)

Variable	Estimación	SE	tStat	p_valor
Constante	14.86116084	0.843784465	17.61250823	9.46E-63
tmed	0.287373736	0.040766828	7.049205117	2.84E-12
tmin	-0.355888986	0.04099531	-8.681212272	1.09E-17
racha3	-1.435872149	0.150503768	-9.540439881	6.24E-21
incContaminante	0.264991986	0.047806178	5.543048992	3.56E-08
Trafico1	1.95E-05	4.98E-06	3.903392236	9.95E-05
mesc 2	-1.157987448	0.306325849	-3.780247253	1.63E-04
mesc 3	-1.16696997	3.19E-01	-3.662717639	2.59E-04
mesc 4	-2.576458173	0.349374797	-7.374482058	2.84E-13
mesc 5	-1.639460526	0.38902267	-4.214305886	2.67E-05
mesc 6	-0.472839774	0.482350793	-0.980281947	3.27E-01

mesc_7	-0.693195171	0.531492352	-1.304242984	1.92E-01
mesc_8	0.066847224	0.526284029	0.12701739	8.99E-01
mesc_9	-0.200663894	0.440815365	-0.455210753	6.49E-01
mesc_10	-1.11420025	0.355249058	-3.136391853	1.75E-03
mesc_11	-0.40734564	0.309034117	-1.318125145	1.88E-01
mesc_12	0.574464432	0.313382433	1.833109871	6.70E-02
racha3^2	0.060079478	0.007016068	8.563126704	2.91E-17

6.4 ANÁLISIS DE RESULTADOS

Una vez obtenido el modelo se debe de comprobar su funcionamiento en los años 2019 y 2020 para chequear si generaliza correctamente. Para ello, mediante los coeficientes estimados se realizan predicciones de cómo sería la contaminación en los años 2019 y 2020 considerando que se conocen los valores de todas las variables de entrada del modelo. Evidentemente, en este caso el modelo tuvo un mal ajuste durante el aprendizaje por lo que no se espera que mejore durante estos dos años.

La raíz del error cuadrático medio y el coeficiente de determinación ajustado calculados para estos dos años valen:

$$RMSE = 3.2943$$

$$R^2 = 0.2852$$

El RMSE aumenta y el coeficiente de determinación ajustado disminuye, lógico ya que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Representando la contaminación real y la estimada por el modelo (ver Figura 32) se aprecian claramente las carencias del modelo. Durante prácticamente todo el 2019 el modelo estima una concentración muy inferior a la realidad. Cierto es, que este contaminante se produce principalmente por la combustión del carbón en calefacciones centrales, variable no disponible en este trabajo, aunque está relacionada con la temperatura. Además, se aprecia unos niveles bajos de este contaminante durante el periodo de estado de alarma (15 de marzo

en adelante). La estimación durante este periodo cae bruscamente debido a la fuerte caída de la variable de tráfico cuya importancia no es la estimada por el modelo de regresión. La predicción mejora considerablemente durante el invierno de 2020.

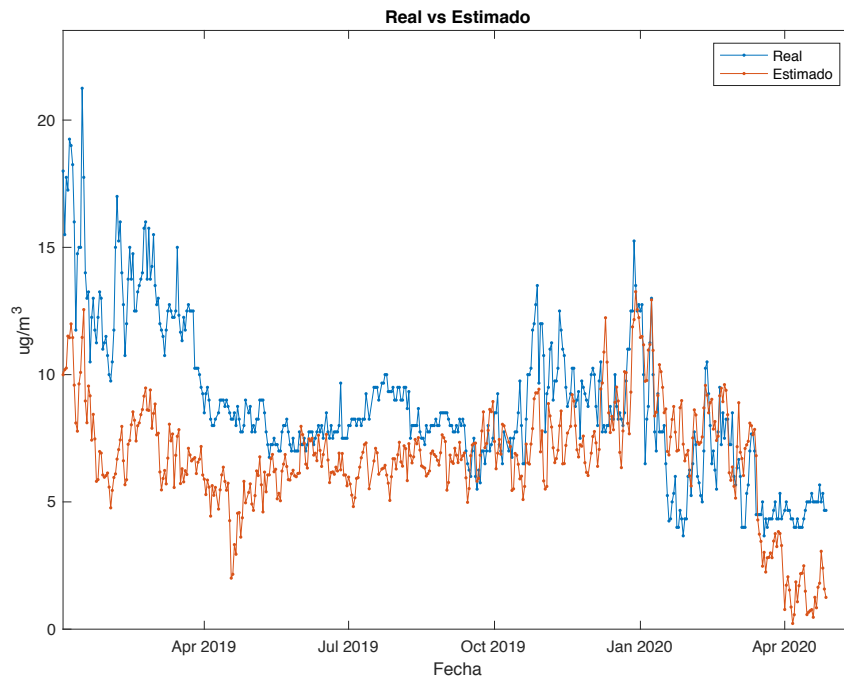


Figura 32. Predicción vs Real (Dióxido de Azufre)

En la Figura 33, se han representado las estimaciones obtenidas para todo el periodo estudiado (entrenamiento y test), para comprender mejor las razones de los errores de estimación de este modelo. Tal y como se mencionó, el Dióxido de Azufre tiene una fuerte concentración en Madrid durante los años 2016 y 2019 y es muy propenso a tener variaciones bruscas. En conclusión, las variables utilizadas para explicar el Dióxido de Azufre son insuficientes y, por tanto, el modelo no estima de manera razonable los niveles de concentración en Madrid.

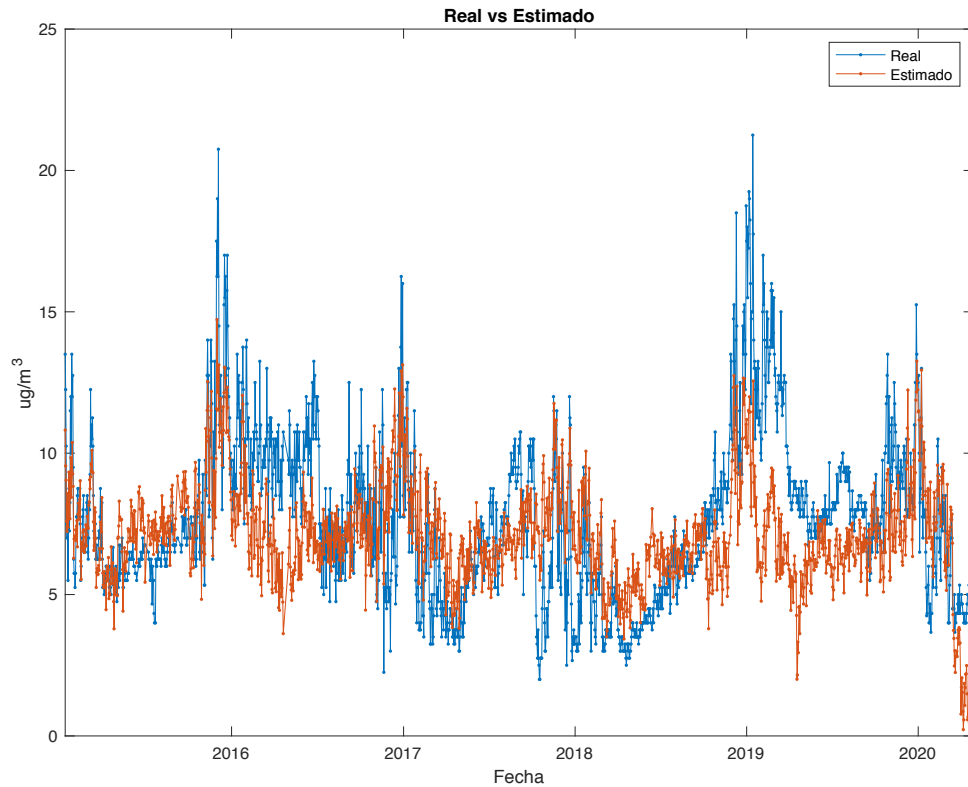


Figura 33. Estimación y predicción vs Real del Dióxido de Azufre (entrenamiento y test)

Capítulo 7. ÓXIDOS, DIÓXIDOS Y MONÓXIDOS DE NITRÓGENO

En este capítulo se analiza en detalle los niveles de contaminación del Óxido, Dióxido y Monóxido de Nitrógeno en Madrid. Para ello se estudian las evoluciones temporales de la serie de dichos contaminantes, ajustando un modelo de regresión lineal múltiple para estimar el nivel de contaminación en función de las variables explicativas seleccionadas.

7.1 LÍMITES, MÁXIMOS Y MÍNIMOS

Los límites legales y los umbrales de información se fijaron mediante Real Decreto el 28 de enero de 2011 [8]. Los umbrales de información superior e inferior se deben de respetar para proteger de la salud de las personas.

El valor más alto registrado de óxido de nitrógeno es de $1376 \mu\text{g} / \text{m}^3$ y se dio el 5 de diciembre de 2001 en la estación de Paseo Recoletos. El mínimo registrado de este contaminante es de $2 \mu\text{g} / \text{m}^3$ ocurriendo hasta en tres ocasiones en la estación de Casa de Campo.

El valor más alto registrado de dióxido de nitrógeno es de $219 \mu\text{g} / \text{m}^3$ y se dio el 12 de enero de 2007 en la estación de la Plaza Luca de Tena. El mínimo registrado de este contaminante es de $1 \mu\text{g} / \text{m}^3$.

El valor más alto registrado de monóxido de nitrógeno es de $771 \mu\text{g} / \text{m}^3$ y se dio el 5 de diciembre de 2001 en la estación de Paseo Recoletos. El mínimo registrado de este contaminante es de 0. Nótese que el máximo de obtenido se dio en la misma estación y el mismo día que el máximo del óxido de nitrógeno.

7.1.1 LÍMITES LEGALES

El valor límite anual del dióxido de nitrógeno y del óxido de nitrógeno es de $40 \mu\text{g} / \text{m}^3$ (ver Tabla 17). Desde el año 2001 únicamente se ha respetado este límite en los años 2012,

2013, 2014, 2016, 2018, 2019, 2020 (ver Tabla 6). Siendo el 2001 el peor año con una media de $62.9 \mu\text{g} / \text{m}^3$.

7.1.2 UMBRALES DE INFORMACIÓN

Los umbrales superior e inferior de evaluación anual son de $32 \mu\text{g} / \text{m}^3$ y $26 \mu\text{g} / \text{m}^3$ (ver Tabla 18).

El umbral superior se sobrepasó todos los años con la excepción del 2020 del que solo se tienen datos hasta el mes de mayo inclusive. El umbral inferior, en cambio, se supera todos los años incluyendo el 2020 (ver Tabla 6).

Tabla 6. Media Anual Dióxido de Nitrógeno

Año	Concentración ($\mu\text{g}/\text{m}^3$)
2001	62.9165603
2002	61.6477391
2003	58.4087073
2004	60.7659574
2005	61.8556195
2006	60.5938115
2007	60.0156285
2008	55.3591964
2009	54.3198456
2010	44.2713786
2011	44.9046196
2012	38.6468839
2013	34.7044985
2014	35.0529574
2015	40.9798073
2016	38.5578403
2017	41.6073845
2018	36.5471222
2019	34.5760268
2020	28.0893258

7.2 ANÁLISIS

Con el fin de comprender el comportamiento del Óxido, Dióxido y Monóxido de Nitrógeno se debe de analizar su comportamiento e intentar comprender que variables influyen en dicho contaminante. Para ello, es imprescindible analizar si todas las estaciones de Madrid tienen datos similares, se deben de analizar las series temporales de los contaminantes y, finalmente, identificar las variables que pueden influir en la variabilidad de los contaminantes.

Analizando las tres series temporales (Figura 34) se aprecia que el Monóxido y el Óxido de Nitrógeno tienen una evolución temporal muy similar mientras que el Dióxido de Nitrógeno es algo distinta en todas las estaciones de medida. Todas las series han ido disminuyendo desde el 2001 aunque los meses de invierno presentan, en algunas ocasiones, valores más altos a los años anteriores. Todos los contaminantes presentan mayores concentraciones durante el invierno y caen a valores muy bajos durante el verano. Nótese que la media de todas las estaciones se representa en negro en la Figura 34

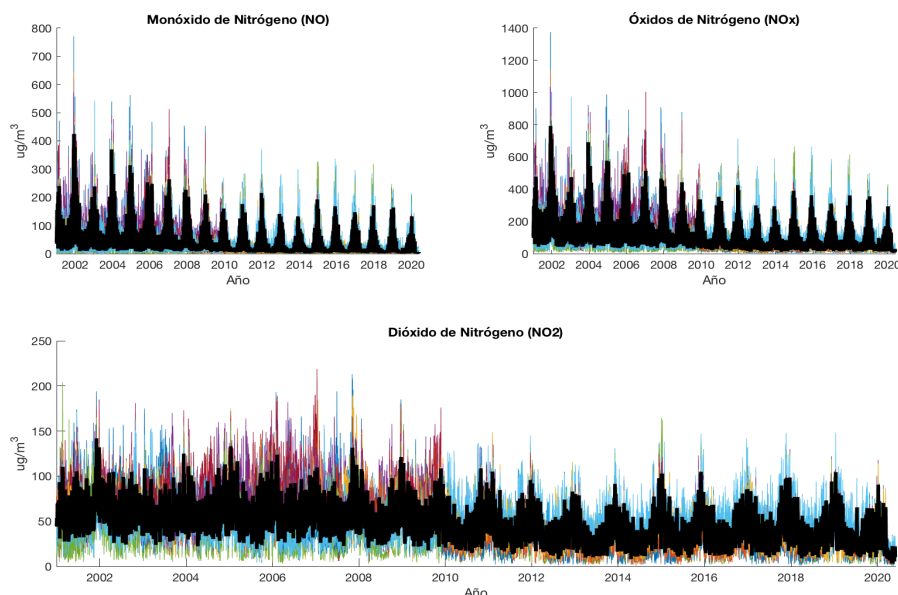


Figura 34. Serie Temporal del Dióxido, Monóxido y Oxido de Nitrógeno en las diferentes estaciones

A continuación, mediante el método de agrupación de *K-means* se han determinado conglomerados de días caracterizados por tener unos niveles de concentración similares en todas las estaciones de medida. Para poder utilizar este método es necesario elegir previamente el valor adecuado de *K*, de tal forma que haya una buena relación entre complejidad y error de cuantización (QE). En la Figura 35 se muestra el criterio utilizado para asignar el valor de *K*. Para los tres contaminantes se decide tomar *K* igual a tres.

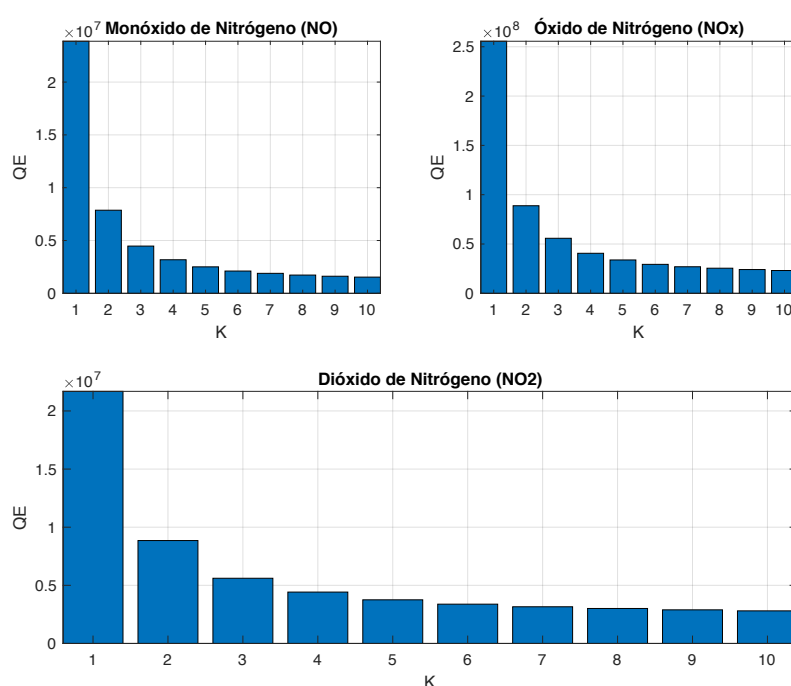


Figura 35. Obtención del número de conglomerados (NO , NO_x , NO_2)

Una vez seleccionado el número de conglomerados (*K*) se procede a representar gráficamente los distintos grupos obtenidos, en donde cada clúster representa un grupo de días similares en cuanto a niveles de concentración en las diferentes estaciones de medida. En la Figura 36 se muestran los centros de los conglomerados (centroides) por simplicidad. Nuevamente, se observa la similitud entre el *NO* y el *NO_x*, destacando que las medidas más altas en los tres clústeres se produjeron en las estaciones de Plaza de España y Barrio del Pilar mientras que las más bajas se produjeron en estaciones alejadas del centro de Madrid

como el Pardo, la Casa de Campo y el Parque Juan Carlos I. En cambio, el NO_2 presenta medidas más bajas en Arturo Soria y Méndez Álvaro para los tres clústeres. También es interesante resaltar que en los días de mayor concentración de estos contaminantes (recogidos en el cluster 1), se observa un aumento mucho mayor en las cuatro primeras estaciones para los contaminantes NO y NOx , algo que no ocurre para el Dióxido de Nitrógeno.

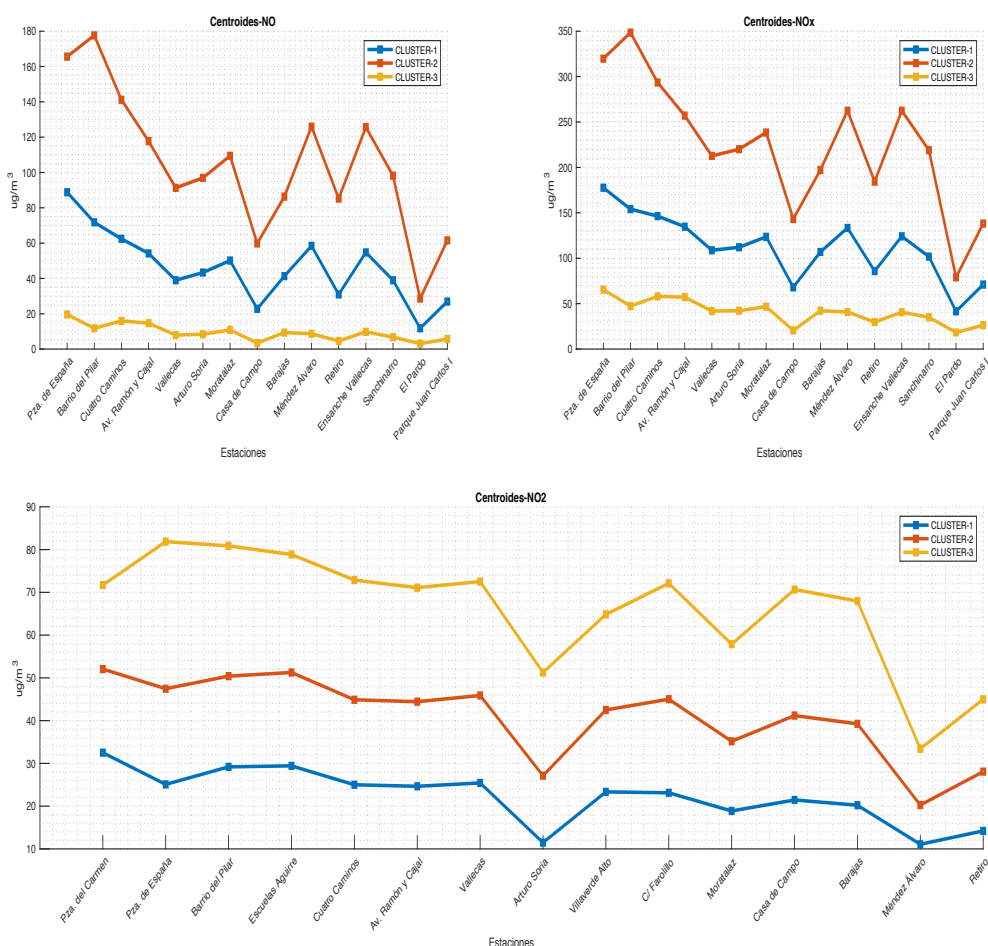


Figura 36. Centroides de los conglomerados (NO , NOx , NO_2)

Para comprender mejor el comportamiento de los contaminantes se procede a analizar el si los niveles de concentración de dichos contaminantes cambian y cómo con variables

discretas como el mes, el día de la semana, la festividad o la estación de medida y las posibles interacciones entre ellas. Para ello, se ha realizado un análisis de la varianza (ANOVA). Para todos los contaminantes estudiados en este capítulo todas las variables influyen, sin tener en cuenta interacciones entre ellas, en la concentración debido a que presentan un *p-valor* ($Prob > F$) igual a cero.

Este resultado quiere decir que hay al menos un grupo de cada variable para el que el contaminante es claramente distinto a los valores que toma para el resto de grupos. Por ejemplo, si se realiza distintas pruebas de comparación múltiple al contaminante NO_2 se aprecia que el sábado y el domingo son claramente diferentes del resto, observándose una subida importante de los niveles de contaminación entre el lunes y el viernes (ver Figura 37). Realizando la misma comparación para el resto de las variables consideradas se aprecia que las estaciones que menos concentración miden son, efectivamente, las estaciones de la Casa de Campo, el Pardo y el Parque Juan Carlos I, y que existe una concentración mayor en los meses de invierno y que los días festivos la contaminación es menor. Realizando las mismas pruebas con los otros dos contaminantes se aprecia que la estación que menos concentración es la estación del Pardo que no aparece en la Figura 36 debido a que contiene grandes periodos de tiempo en los que no presenta información y, por tanto, se descarta a la hora de usar el método *K-means*.

En cuanto a las interacciones, todos los contaminantes presentan un *p-valor* de 0 en la interacción entre el día de la semana y el mes. Así mismo, todos presentan un *p-valor* muy cercano a 0 en la interacción entre el día de la semana y la festividad.

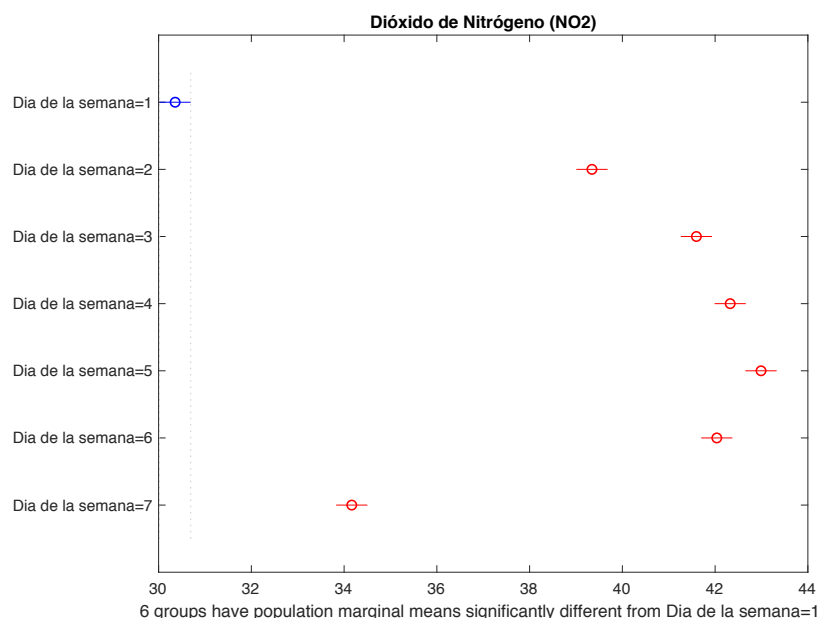


Figura 37. Prueba de Comparación Múltiple del día de la semana (NO₂)

En la Figura 37 se muestra la prueba de comparación múltiple realizada a la variable discreta día de la semana. Se observa claramente, como el día de menor contaminación es el domingo (día 1 en la figura) mientras que el jueves es el día de mayor contaminación. Esto es una tendencia que siguen los demás contaminantes analizados en este capítulo y que está claramente relacionado con la actividad humana, muy similar a la que se puede observar en la serie de consumo de energía eléctrica en donde los lunes y los viernes actúan como días de transición entre el fin de semana y los laborables de máxima actividad.

7.3 MODELO EXPLICATIVO DEL NIVEL DE CONCENTRACIÓN

Utilizando las variables creadas durante el Capítulo 5. y añadiendo como variable la diferencia de concentración de los contaminantes entre los dos días anteriores se ha planteado un modelo de regresión lineal múltiple para explicar el nivel de contaminación en función de las mismas. El modelo se ha ajustado con los datos de 2015 a 2018.

Dado que se dispone del nivel de concentración del Monóxido, Dióxido y Óxido de Nitrógeno en diferentes estaciones, se ha estudiado como variable de salida la media de todas las estaciones pertenecientes a la zona interior a la M-30. Además, en el caso del Óxido, Dióxido y Monóxido de Nitrógeno se utilizaron las variables de tráfico obtenidas mediante el método del análisis de las componentes principales. La variable Trafico1 representa la primera componente principal y la variable Trafico2 la segunda componente.

Inicialmente, se plantea un modelo con todas las variables disponibles y se van eliminando aquellas con un *p-valor* mayor. Posteriormente, se prueban distintas interacciones o relaciones cuadráticas con el fin reducir la raíz del error cuadrático medio (RMSE) y aumentar el coeficiente ajustado de determinación (R^2).

Finalmente, una vez obtenido el modelo con un número razonable de variables explicativas se impone la condición de que en ningún caso sea menor que cero el valor estimado por el modelo.

El modelo finalmente obtenido del Monóxido de Nitrógeno consta de 13 términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 15.2968$$

$$R^2 = 0.7359$$

Según el análisis de residuos mostrado en la Figura 38 se puede cuestionar la falta de normalidad de los mismos, o que su varianza no es constante, además de la independencia de los residuos debido a que a mayor valor estimado se dan residuos mayores. Esto ocurre en los meses de invierno. En la Figura 39 se aprecia la calidad de la estimación del modelo concluyendo dos cosas: el modelo no es capaz de estimar los picos de concentración repentinos que se dan principalmente en el mes de enero y el modelo estima razonablemente bien el resto de las situaciones.

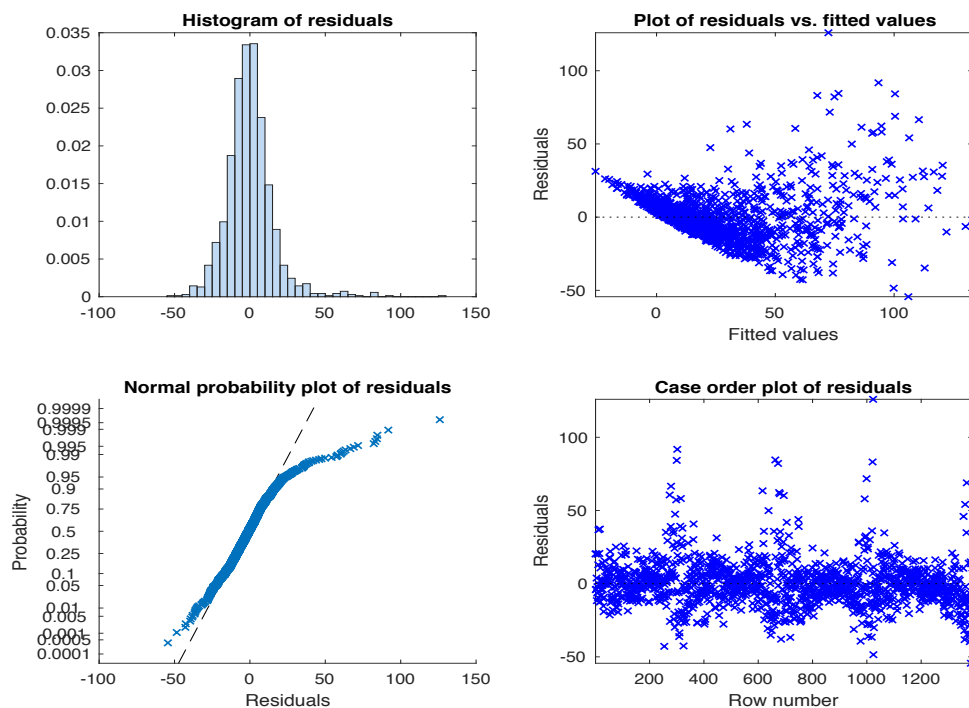


Figura 38. Análisis de los Residuos (Monóxido de Nitrógeno)

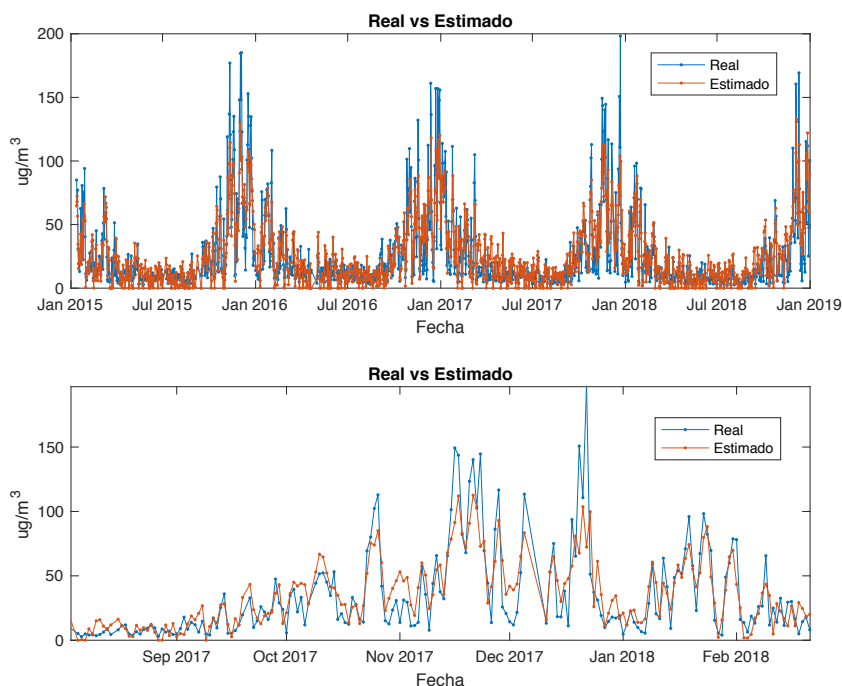


Figura 39. Estimado vs Real (Monóxido de Nitrógeno)

El modelo obtenido para le Monóxido de Nitrógeno se muestra en la Tabla 7:

Tabla 7. Diagnóstico e interpretación del modelo (Monóxido de Nitrógeno)

Variable	Estimación	SE	tStat	p-valor
Constante	125.4221678	6.349878479	19.75190048	1.36E-76
tmed	-5.00128702	0.539404827	-9.271861813	6.90E-20
tmax	3.695209727	0.417101907	8.859249188	2.45E-18
racha	-1.415966474	0.195727797	-7.234365765	7.77E-13
tmed3	5.173135379	0.835239411	6.193595885	7.76E-10
tmin3	-4.90E+00	8.05E-01	-6.087775631	1.48E-09
racha3	-17.71205532	1.09439931	-16.18427127	4.92E-54
sol5	-1.769652895	4.18E-01	-4.229789901	2.49E-05
incContaminante	0.284113472	0.021835059	13.01180223	1.41E-36
Trafico1	0.000431057	3.53E-05	12.20772004	1.32E-32
mesc_2	-14.20171899	2.149962076	-6.605567209	5.66E-11
mesc_3	-15.13461032	2.252462038	-6.719141129	2.68E-11
mesc_4	-21.93473871	2.532431586	-8.661532588	1.29E-17
mesc_5	-21.22808862	2.8577067	-7.428365066	1.93E-13
mesc_6	-19.06073227	3.678289381	-5.181955604	2.53E-07
mesc_7	-12.90634829	4.116242734	-3.135468223	1.75E-03
mesc_8	-9.262135559	4.066116763	-2.277882338	2.29E-02
mesc_9	-20.05939824	3.373175538	-5.946740101	3.47E-09
mesc_10	-10.69846426	2.675250767	-3.999051001	6.70E-05
mesc_11	3.46975097	2.218647608	1.563903595	1.18E-01
mesc_12	6.958024071	2.210943288	3.147083921	0.001684688
racha3^2	0.756119635	0.049755687	15.19664747	2.61E-48

El modelo finalmente obtenido del Dióxido de Nitrógeno consta de 16 términos, con las siguientes medidas de bondad del ajuste obteniendo:

$$RMSE = 7.6904$$

$$R^2 = 0.7967$$

Según el análisis de los residuos de la Figura 40. se observa que los residuos obtenidos son razonables.

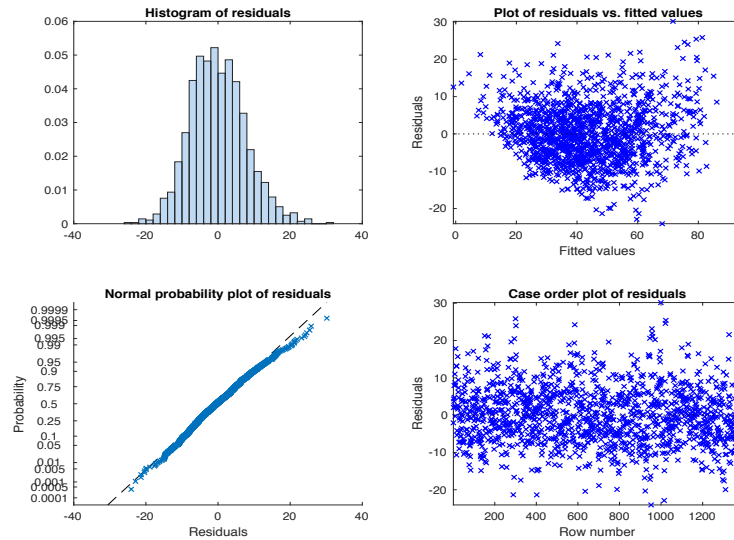


Figura 40. Análisis de los Residuos (Dióxico de Nitrógeno)

En la Figura 41 se puede observar que efectivamente el modelo genera estimaciones que siguen razonablemente bien la dinámica de la serie real.

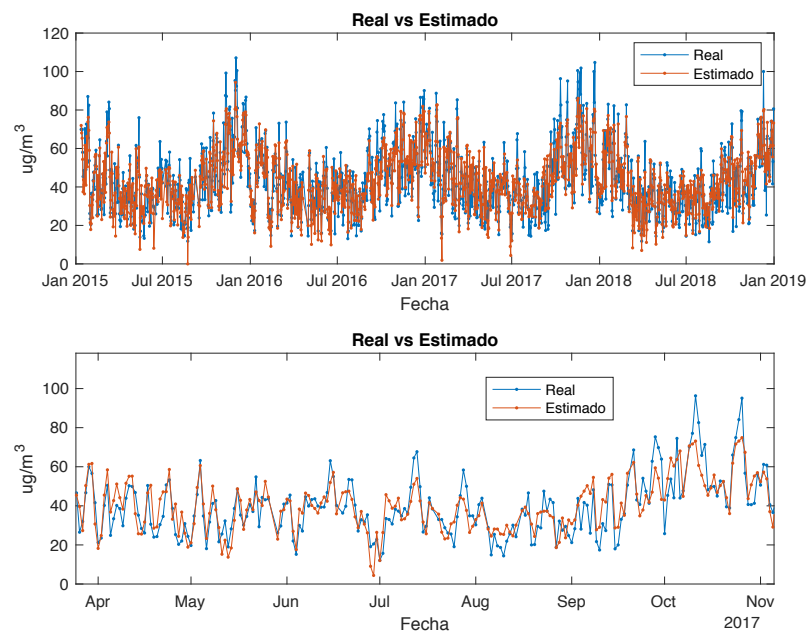


Figura 41. Estimado vs Real (Dióxico de Nitrógeno)

El modelo obtenido para el Dióxido de Nitrógeno se muestra en la Tabla 8:

Tabla 8. Diagnóstico e interpretación del modelo (Dióxido de Nitrógeno)

Variable	Estimación	SE	tStat	p-valor
Constante	48.81394106	1.97295092	24.74158914	6.26E-112
tmed	-4.283707296	0.323041574	-13.26054488	7.75E-38
tmax	4.471339447	0.275406031	16.23544494	2.52E-54
velmed	-3.101145049	0.22966894	-13.50267499	4.35E-39
sol	-1.263968256	0.104464205	-12.09953453	4.39E-32
tmed3	2.388	3.74E-01	6.385416404	2.34E-10
prec3	0.863495937	0.128946959	6.696520355	3.11E-11
tmin3	-4.036671609	4.96E-01	-8.144662837	8.51E-16
racha3	-0.854391554	0.137808237	-6.199858394	7.48E-10
tmin5	1.433626773	4.30E-01	3.332785984	8.83E-04
tmed7	-1.324488817	0.291217236	-4.54811272	5.89E-06
incContaminante	0.185061589	0.017438501	10.61224206	2.47E-25
Trafico1	0.000562703	3.48E-05	16.17014574	6.14E-54
mesc_2	-5.083124347	1.062597563	-4.783677775	1.91E-06
mesc_3	-8.880292694	1.175680946	-7.553318546	7.74E-14
mesc_4	-14.73319427	1.40401365	-10.49362609	7.93E-25
mesc_5	-17.85743832	1.595070845	-11.19538883	6.79E-28
mesc_6	-19.88699348	1.954835395	-10.17323174	1.76E-23
mesc_7	-18.86163273	2.185717044	-8.629494281	1.69E-17
mesc_8	-17.20067111	2.179439951	-7.892243646	6.05E-15
mesc_9	-13.91057018	1.87526387	-7.417926829	2.08E-13
mesc_10	-4.485076056	1.572217691	-2.852706773	4.40E-03
mesc_11	0.752079179	1.202450015	0.62545567	0.531776739
mesc_12	3.00916564	1.080994436	2.783701321	0.005448501
velmed:Trafico1	-6.01E-05	1.23E-05	-4.869554035	1.25E-06
tmin3^2	0.056921192	0.007318776	7.777419659	1.45E-14

El modelo finalmente obtenido del Óxido de Nitrógeno consta de 14 términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 29.2123$$

$$R^2 = 0.7672$$

El análisis de residuos mostrado en la Figura 42 es muy similar a lo obtenido en el análisis de residuos del Monóxido de Nitrógeno. Al igual que en el caso del Monóxido de Nitrógeno, tanto la independencia como la normalidad de los residuos es cuestionable. De la misma manera, en los meses de invierno el modelo está subestimando y es incapaz de estimar los días de mayores concentraciones.

En la Figura 43, se muestra el valor real de los niveles de concentración de este contaminante y la estimación. Se aprecia una buena estimación por norma general exceptuando las subestimaciones explicadas anteriormente.

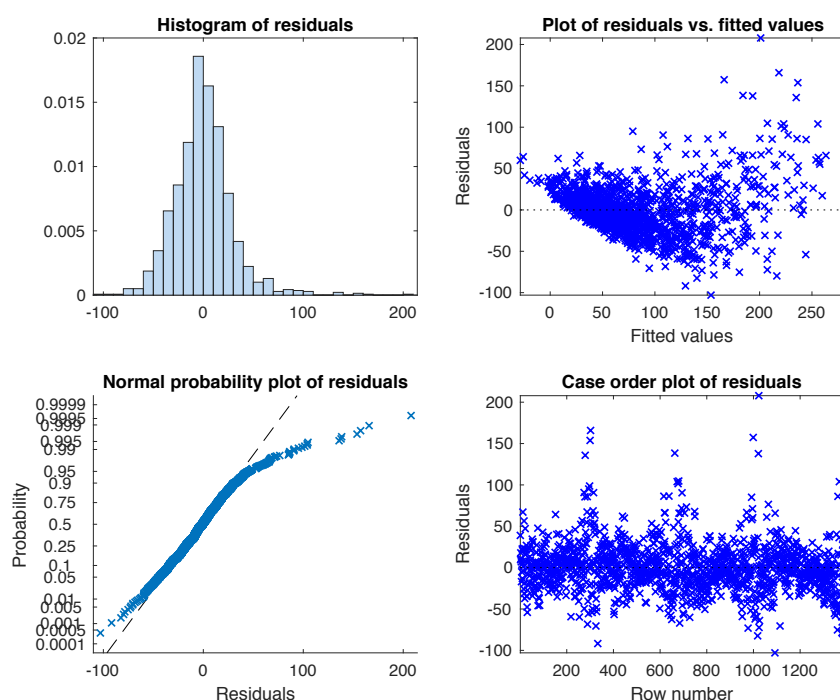


Figura 42. Análisis de los Residuos (Óxido de Nitrógeno)

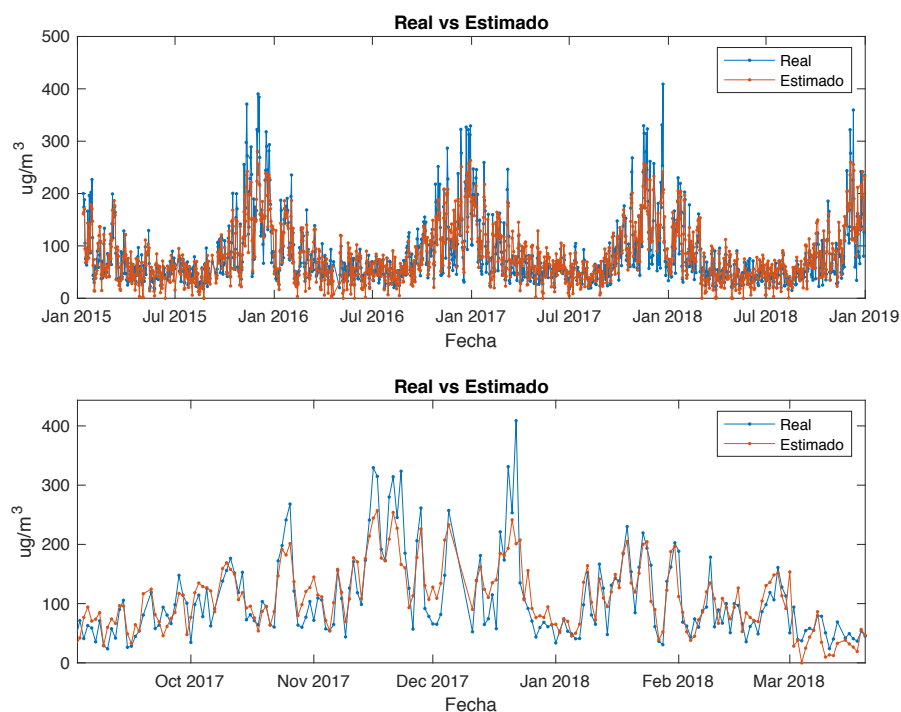


Figura 43. Estimación vs Real (Óxido de Nitrógeno)

El modelo obtenido para el Óxido de Nitrógeno se muestra en la Tabla 9:

Tabla 9. Diagnóstico e interpretación del modelo (Óxido de Nitrógeno)

Variable	Estimación	SE	tStat	p-valor
Constante	211.3742868	9.319420028	22.68105592	6.61E-97
tmed	-9.924514934	1.029294235	-9.642058217	2.50E-21
tmax	8.241242499	0.795681941	10.35745827	2.98E-24
racha	-20.98724432	1.257586797	-16.68850561	4.81E-57
tmed3	16.74569766	1.828794307	9.156687329	1.90E-19
tmin3	-1.74E+01	1.66E+00	-10.43411491	1.41E-24
racha3	-2.617222492	0.560935655	-4.665815892	3.38E-06
sol3	-5.359113419	7.24E-01	-7.404378725	2.30E-13
tmax7	-2.389684562	0.625250455	-3.821963732	1.38E-04
incContaminante	0.231554677	2.02E-02	11.45147472	4.68E-29
Trafico1	0.001044207	6.67E-05	15.66221771	5.68E-51
mesc_2	-23.16933798	4.045786783	-5.726781767	1.26E-08
mesc_3	-29.16302198	4.38E+00	-6.662724649	3.89E-11
mesc_4	-41.51864764	5.226000331	-7.944631652	4.04E-15

mesc_5	-42.53977029	5.913171501	-7.1940701	1.03E-12
mesc_6	-42.57339886	7.229400329	-5.88892535	4.89E-09
mesc_7	-35.07476771	8.057741773	-4.352927743	1.44E-05
mesc_8	-30.3888028	8.015654243	-3.79118184	1.56E-04
mesc_9	-41.06818478	6.805723911	-6.034359506	2.05E-09
mesc_10	-13.35229895	5.676290689	-2.352293018	1.88E-02
mesc_11	10.63259884	4.481226038	2.372698622	1.78E-02
mesc_12	17.60511555	4.145704276	4.246592225	2.32E-05
racha^2	0.759730472	0.052878219	14.36755038	1.07E-43
tmin5^2	0.149699809	0.028442876	5.263174193	1.64E-07

7.4 ANÁLISIS DE RESULTADOS

Una vez obtenido el modelo se debe de comprobar su funcionamiento en los años 2019 y 2020 para chequear si generaliza correctamente. Para ello, mediante los coeficientes estimados con los datos de 2015-2018 se realizan predicciones de cómo sería la contaminación en los años 2019 y 2020 considerando que se conocen los valores de todas las variables de entrada del modelo.

Para el Monóxido de Nitrógeno el RMSE y el R^2 calculados para estos dos años valen:

$$RMSE = 16.5054$$

$$R^2 = 0.7213$$

La raíz del error cuadrático medio aumenta ligeramente y el coeficiente de determinación ajustado también disminuye ligeramente, lógico ya que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Representando la contaminación real y la estimada por el modelo (ver Figura 44) se aprecia una leve sobreestimación en la mayoría de la serie. Especialmente interesante es el nivel de concentración de este contaminante durante el estado de alarma (15 de marzo de 2020 en adelante). Durante el estado de alarma el modelo en la gran mayoría de los días estima una concentración de prácticamente cero. Además, la variabilidad de la estimación es

considerablemente mayor que la real durante los meses de verano (posiblemente debido a la gran variabilidad de los datos durante el invierno). En definitiva, los resultados obtenidos para el Monóxido de Carbono son relativamente buenos.

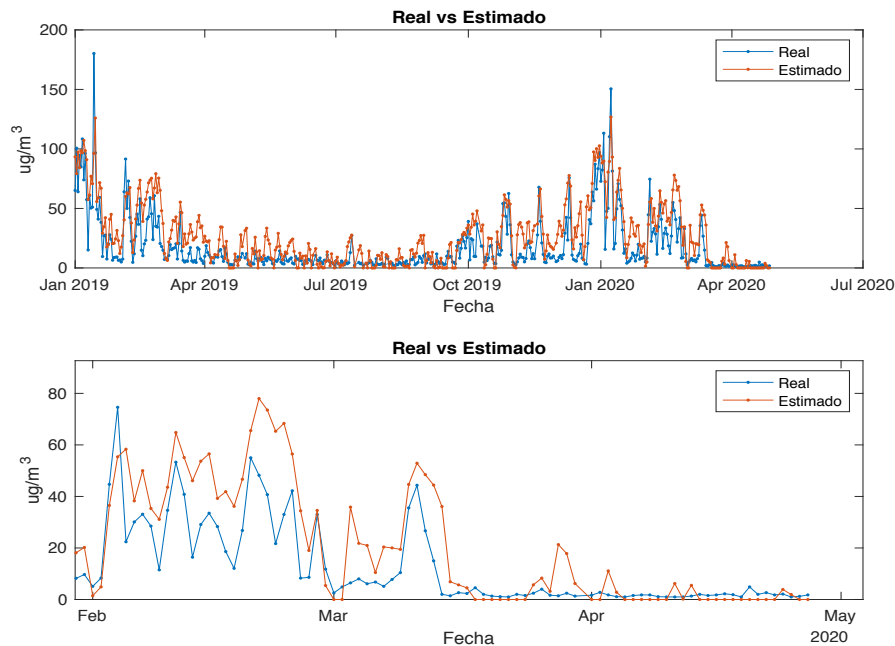


Figura 44. Predicción vs Real (Monóxido de Nitrógeno)

Para el Dióxido de Nitrógeno el RMSE y el R^2 calculados para estos dos años valen:

$$RMSE = 9.3864$$

$$R^2 = 0.8139$$

El RMSE aumenta y el coeficiente de determinación ajustado aumenta. Es lógico que el RMSE aumente debido a que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Representando la contaminación real y la estimada por el modelo (ver Figura 45) se aprecia un buen seguimiento de los niveles de concentración del contaminante. Especialmente

interesante es el nivel de concentración de este contaminante durante el estado de alarma (15 de marzo de 2020 en adelante). En este caso particular durante el estado de alarma se estima razonablemente bien la bajada de los niveles de concentración del Dióxido de Nitrógeno. Por lo que la importancia del tráfico está bien estimada por el modelo de regresión. En definitiva, los resultados obtenidos son de buena calidad para este contaminante.

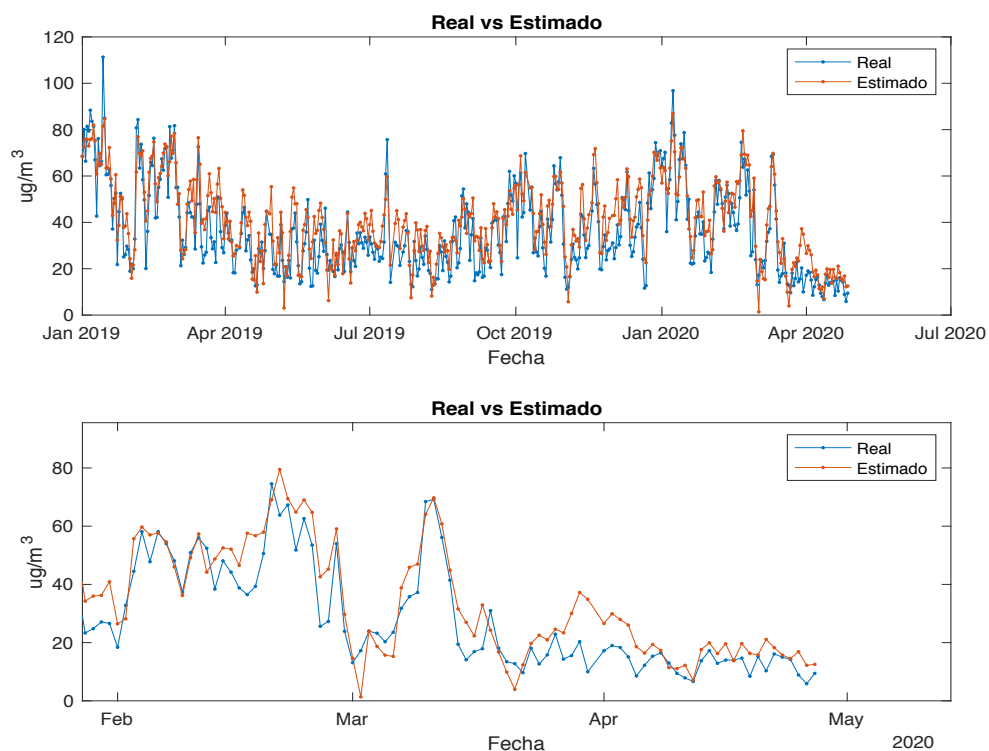


Figura 45. Predicción vs Real (Dióxido de Nitrógeno)

Para el Óxido de Nitrógeno el RMSE y el R^2 calculados para estos dos años valen:

$$RMSE = 32.7542$$

$$R^2 = 0.7843$$

La raíz del error cuadrático medio aumenta y el coeficiente de determinación ajustado aumenta. Es lógico que el RMSE aumente debido a que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Representando la contaminación real y la estimada por el modelo (ver Figura 46) se aprecia un buen seguimiento de los niveles de concentración del contaminante exceptuando alguna subestimación durante los meses de invierno. Durante el confinamiento, los niveles de concentración se mantuvieron bajos cosa que la estimación recoge. De nuevo, la bajada del tráfico en Madrid influye, prácticamente al instante, en la mejora de los niveles de calidad del aire de este contaminante. El modelo obtenido para el Óxido de Nitrógeno realiza estimaciones muy razonables.

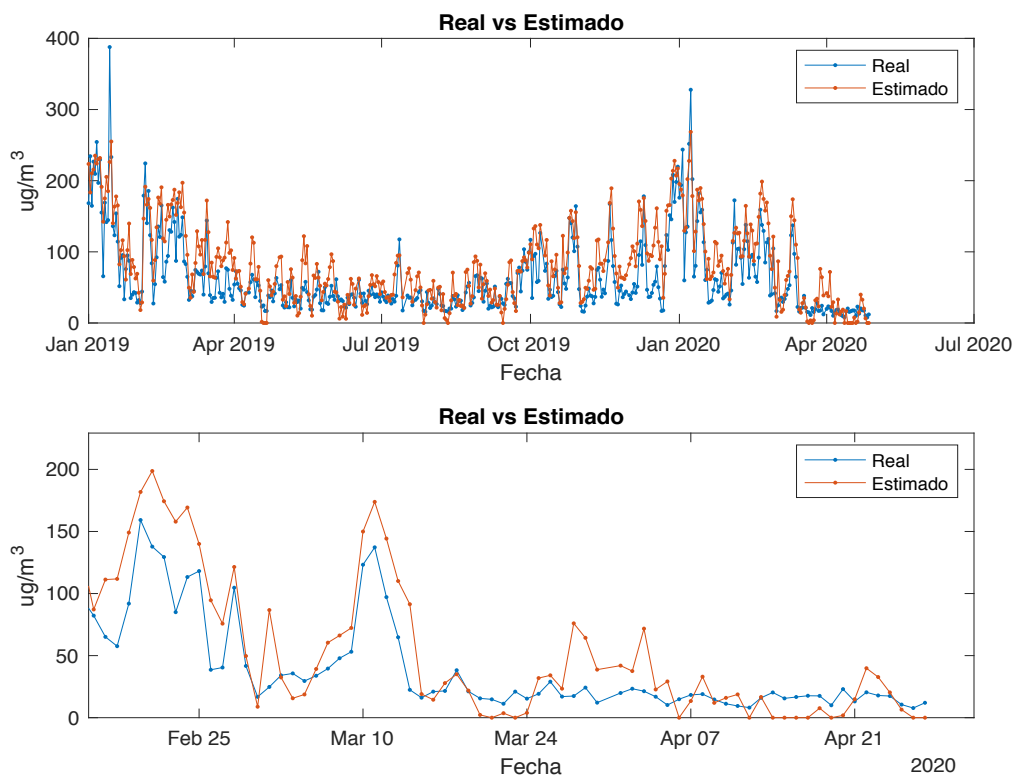


Figura 46. Predicción vs Real (Óxido de Nitrógeno)

Capítulo 8. PARTÍCULAS MENORES DE 10 Y 2,5 MICRAS

En este capítulo se analiza en detalle los niveles de contaminación de las partículas menores de 10 y 2.5 micras en Madrid. Para ello se estudian las evoluciones temporales de la serie de dichos contaminantes, ajustando un modelo de regresión lineal múltiple para estimar el nivel de contaminación en función de las variables explicativas seleccionadas.

8.1 LÍMITES, MÁXIMOS Y MÍNIMOS

Los límites legales y los umbrales de información se fijaron mediante Real Decreto el 28 de enero de 2011 [8]. Los umbrales de información superior e inferior se deben de respetar para proteger de la salud de las personas.

El valor más alto registrado de partículas menores de 10 micras es de $257 \mu\text{g} / \text{m}^3$ y se dio el 22 de febrero de 2016 en la estación de Plaza de Castilla. El 22 de febrero de 2016 hubo una intrusión sahariana [22] (masa de aire africano) en Madrid lo cual explicaría un valor tan alto de estas partículas. El mínimo registrado de este contaminante es de $1 \mu\text{g} / \text{m}^3$.

El valor más alto registrado de partículas menores de 2.5 micras es de $126 \mu\text{g} / \text{m}^3$ y se dio el 24 de julio de 2004 en la estación de Cuatro Caminos. El mínimo registrado de este contaminante es de $1 \mu\text{g} / \text{m}^3$.

8.1.1 LÍMITES LEGALES

El valor límite anual de partículas menores de 10 micras y de partículas menores de 2.5 micras es de $40 \mu\text{g} / \text{m}^3$ y de $25 \mu\text{g} / \text{m}^3$ (ver Tabla 17). El límite legal establecido no se supera para ninguno de los dos contaminantes ningún año. El valor más elevado de las partículas menores de 10 micras se da en 2006 mientras que el valor más elevado para las partículas menores de 2.5 micras se da en 2004.

El valor límite diario establecidos para las partículas menores de 10 micras es de $50 \mu g / m^3$ con un tope de superaciones de 35 veces al año (ver Tabla 11). Este requisito lo cumplen los años 2010 y del 2012 en adelante. La Tabla 11 contabiliza la cantidad de días en la que al menos 1 estación ha medido más del límite o umbral establecido.

8.1.2 UMBRALES DE INFORMACIÓN

El umbral superior anual de partículas menores de 10 micras y de partículas menores de 2.5 micras es de $28 \mu g / m^3$ y de $17 \mu g / m^3$ (ver Tabla 18). El umbral inferior anual de partículas menores de 10 micras y de partículas menores de 2.5 micras es de $20 \mu g / m^3$ y de $12 \mu g / m^3$. Los años en los que superan estos límites se aprecian en la Tabla 10.

En la Tabla 11 se observa el número de veces que se ha sobrepasado ese año los umbrales diarios para las partículas menores de 10 micras (no se debe de superar más de 35 veces al año).

Tabla 10. Media Anual Partículas menores de 10 y 2.5 micras

Año	$PM_{2.5}$	PM_{10}
2001	-	34.5839478
2002	-	33.2455972
2003	18.4383228	35.552092
2004	20.078355	33.435348
2005	18.7574081	34.1980572
2006	18.6199729	36.9278735
2007	17.4031809	33.1611167
2008	14.4288618	26.6226136
2009	12.8	24.5310539
2010	11.9794133	21.8140741
2011	12.2993421	23.3019435
2012	11.8377757	23.6119881
2013	9.93441109	18.3836782
2014	10.6334716	19.3098462
2015	11.4854033	21.0439047
2016	10.3879706	19.1690989
2017	9.93876611	19.9443422
2018	10.4698331	18.4033884

2019	9.8551532	17.4595238
2020	9.87364921	16.4950752

Tabla 11. N° de Superaciones de los límites de partículas menores de 10 micras

Año	Umbral Inferior N° de Superaciones	Umbral Superior N° de Superaciones	Límite legal N° de Superaciones
2001	342	294	185
2002	339	279	161
2003	342	292	197
2004	334	268	177
2005	340	285	194
2006	352	312	224
2007	351	312	171
2008	315	218	83
2009	289	201	61
2010	203	106	22
2011	229	132	49
2012	220	116	32
2013	159	67	11
2014	169	62	12
2015	204	92	27
2016	162	79	22
2017	207	97	29
2018	179	91	24
2019	194	85	18
2020	65	34	11

8.2 ANÁLISIS

Con el fin de comprender el comportamiento de las Partículas menores de 10 y 2.5 micras se debe de analizar su comportamiento e intentar comprender que variables influyen en dicho contaminante. Para ello, es imprescindible analizar si todas las estaciones de Madrid tienen datos similares, se deben de analizar las series temporales de los contaminantes y, finalmente, identificar las variables que pueden influir en la variabilidad de los contaminantes.

Analizando la serie temporal (Figura 47) se aprecia que las partículas menores de 10 y 2.5 micras presentan una evolución temporal muy similar en todas las estaciones de medida. Los picos tan relevantes observados en la serie temporal de ambos contaminantes se deben a intrusiones saharianas (masas de aire africano) debido a que estos episodios elevan enormemente los niveles de concentración del contaminante. Nótese que la media de todas las estaciones se representa en negro en la Figura 47.

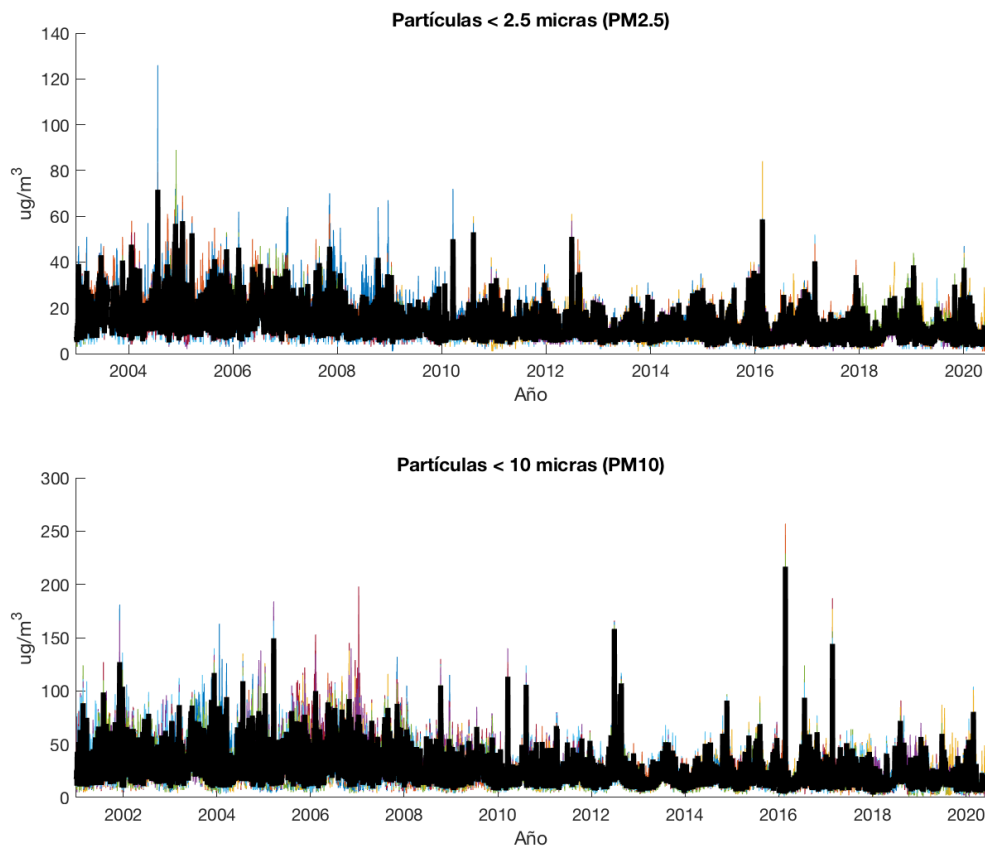


Figura 47. Serie Temporal de las partículas menores de 10 y 2.5 micras en las diferentes estaciones

A continuación, mediante el método de agrupación de *K-means*, se han determinado conglomerados de días caracterizados por tener unos niveles de contaminación similares en todas las estaciones de medida. Para poder utilizar este método es necesario elegir

previamente el valor adecuado de K , de tal forma que haya una buena relación entre complejidad y error de cuantización (QE). En la Figura 48 se muestra el criterio utilizado para seleccionar el valor de K . Se toma K igual a tres y K igual a cuatro para las $PM_{2.5}$ y PM_{10} respectivamente. Únicamente se han tenido en cuenta los datos desde el año 2010 hasta el 2019 inclusive debido a la mejor calidad de los datos durante este tiempo.

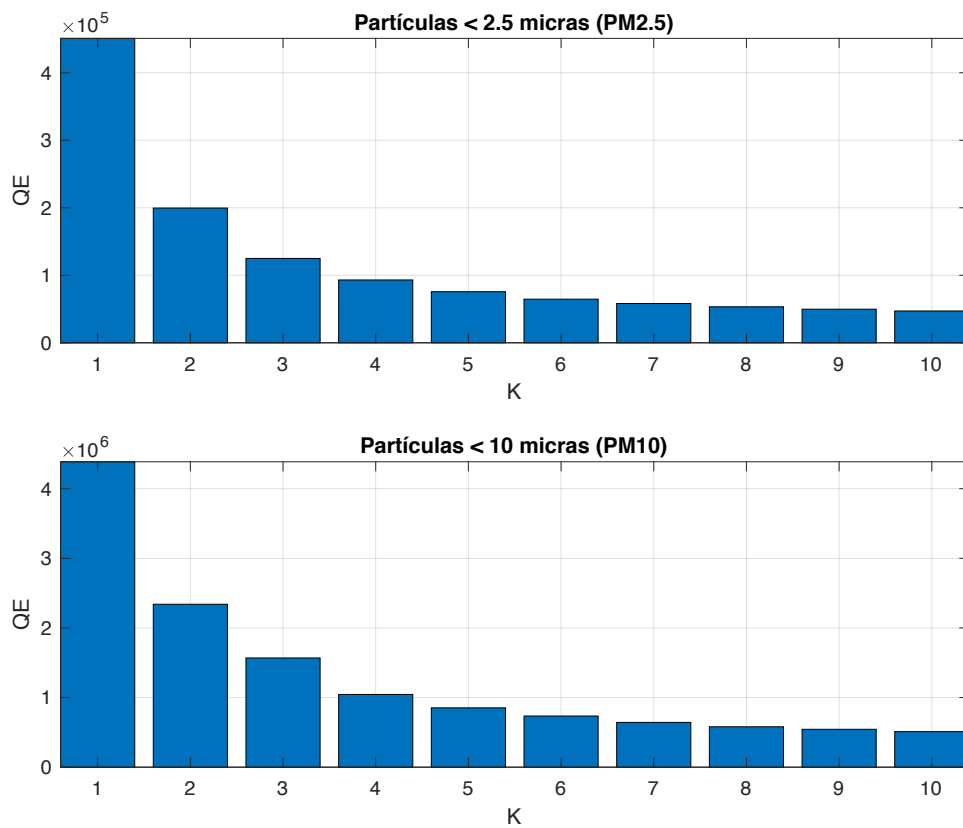


Figura 48. Obtención del número de conglomerados ($PM_{2.5}$, PM_{10})

Una vez seleccionado el número de conglomerados (K) se procede a representar gráficamente los distintos grupos obtenidos, en donde cada clúster representa un grupo de días similares en cuanto a niveles de concentración en las diferentes estaciones de medida. En la Figura 49 se muestran los centros de los conglomerados (centroides) por simplicidad. Las partículas menores de 2.5 micras únicamente se miden durante el periodo de tiempo

seleccionado (2010-2019) sin apenas errores en la información en las cuatro estaciones mostradas en la Figura 49. Se aprecia un comportamiento muy parecido entre todas las estaciones, exceptuando la estación de la Casa de Campo para ambos contaminantes. De nuevo, esta estación recoge mejor calidad del aire para los tres clústeres que las demás.

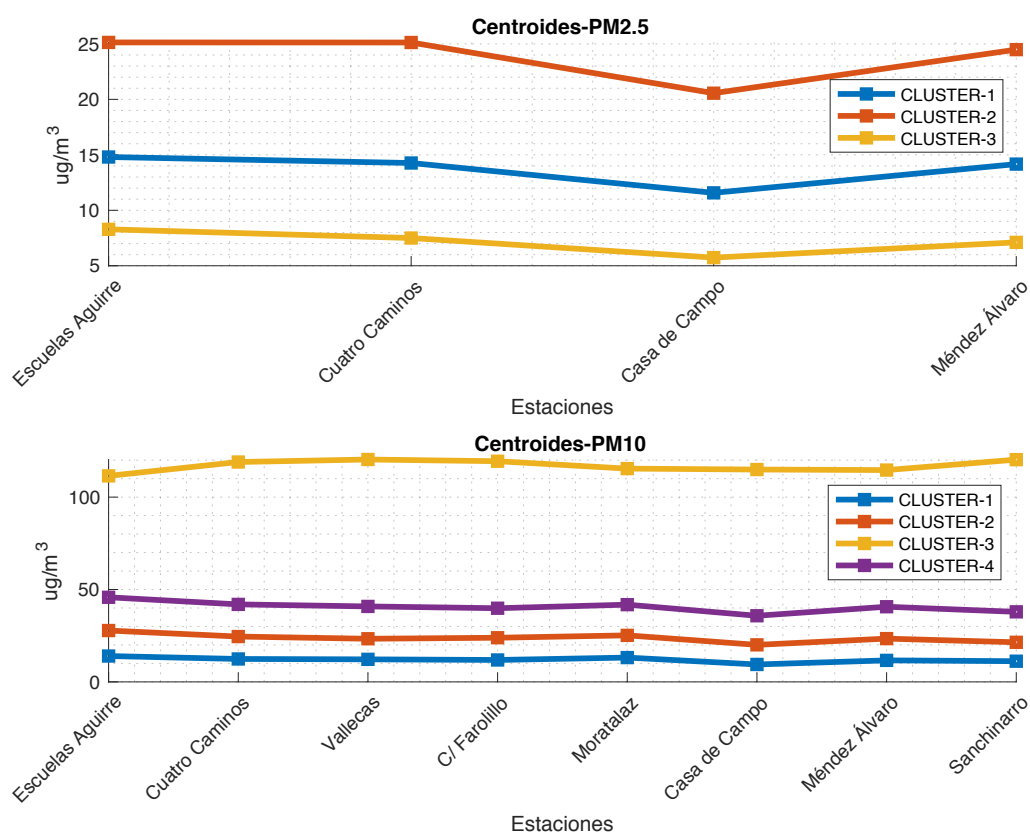


Figura 49. Centroides de los conglomerados ($PM_{2.5}$, PM_{10})

Para comprender mejor el comportamiento de los contaminantes se procede a analizar si los niveles de concentración de dichos contaminantes cambian y cómo con variables discretas como el mes, el día de la semana, la festividad o la estación y las posibles interacciones entre ellas. Para ello, se ha realizado un análisis de la varianza (ANOVA). Para todos los contaminantes estudiados en este capítulo todas las variables influyen, sin tener en cuenta

interacciones entre ellas, en la concentración debido a que presentan un *p-valor* ($Prob > F$) igual a cero o muy próximo a dicho valor.

Ambos contaminantes muestran un comportamiento similar al realizar la prueba de comparación múltiple a las distintas variables. Los contaminantes muestran un aumento de la contaminación durante el invierno, aumenta los días entre semana y los no festivos. Para ambos tipos de partículas las interacciones relevantes son el mes con la estación, el mes con la festividad y el mes con el día de la semana. En cambio, presenta un *p-valor* cercanos a la unidad en la interacción estación con el día de la semana.

8.3 MODELO EXPLICATIVO DEL NIVEL DE CONCENTRACIÓN

Utilizando las variables creadas durante el Capítulo 5. , añadiendo como variable la diferencia de concentración de los contaminantes entre los dos días anteriores y añadiendo variables para explicar las intrusiones saharianas se ha planteado un modelo de regresión lineal múltiple para explicar el nivel de contaminación en función de las mismas. El modelo se ha ajustado con los datos de 2015 a 2018.

Dado que se dispone del nivel de concentración de las partículas en diferentes estaciones, se ha estudiado como variable de salida la media de todas las estaciones pertenecientes a la zona interior a la M-30. Además, en el caso de las partículas menores de 10 y 2.5 micras se utilizaron las variables de tráfico obtenidas mediante el método del análisis de las componentes principales. La variable Trafico1 representa la primera componente principal y la variable Trafico2 la segunda componente.

Inicialmente, se plantea un modelo con todas las variables disponibles y se van eliminando aquellas con un *p-valor* mayor. Posteriormente, se prueban distintas interacciones o relaciones cuadráticas con el fin reducir la raíz del error cuadrático medio (RMSE) y aumentar el coeficiente ajustado de determinación (R^2).

Finalmente, una vez obtenido el modelo con un número razonable de variables explicativas se impone la condición de que en ningún caso sea menor que cero el valor estimado por el modelo.

El modelo finalmente obtenido de las partículas menores de 2.5 micras consta de 15 términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 3.4762$$

$$R^2 = 0.6250$$

Su análisis de residuos mostrado en la Figura 50 no cuestiona la independencia de los residuos mientras que la normalidad de los residuos no parece aceptable. Únicamente se han introducido dos episodios de intrusiones saharianas (las de mayor impacto en el contaminante) por lo que existen varios días sin considerar. Esto conlleva residuos elevados ya que el modelo es incapaz de explicarlos correctamente con el resto de las variables. En la Figura 51 se aprecia la calidad de la estimación y se observa claramente como el modelo subestima los niveles de concentración de partículas durante varios episodios. Únicamente se consideraron los episodios de febrero de 2016 y febrero de 2017 como episodios de intrusiones saharianas.

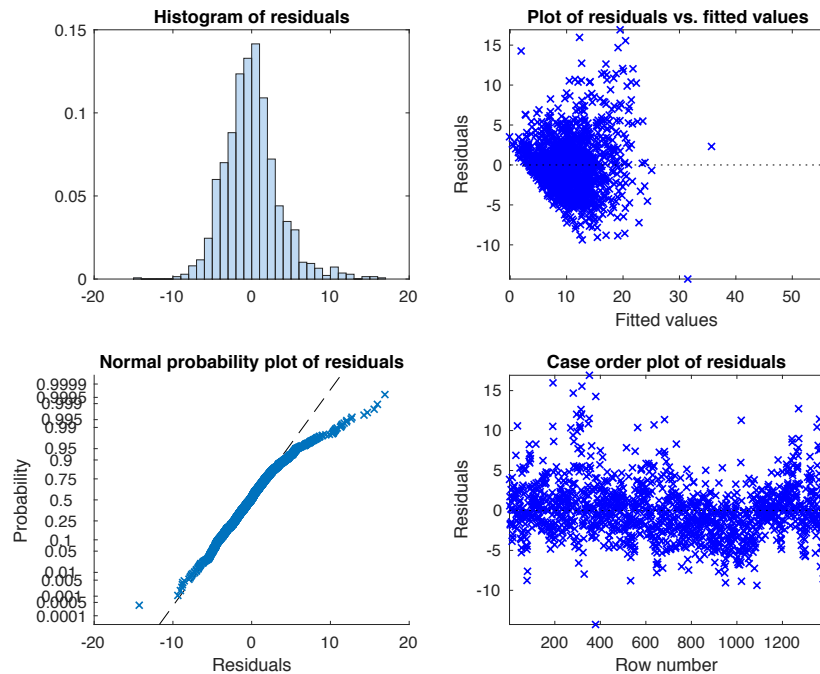


Figura 50. Análisis de los Residuos (Partículas menores de 2.5 micras)

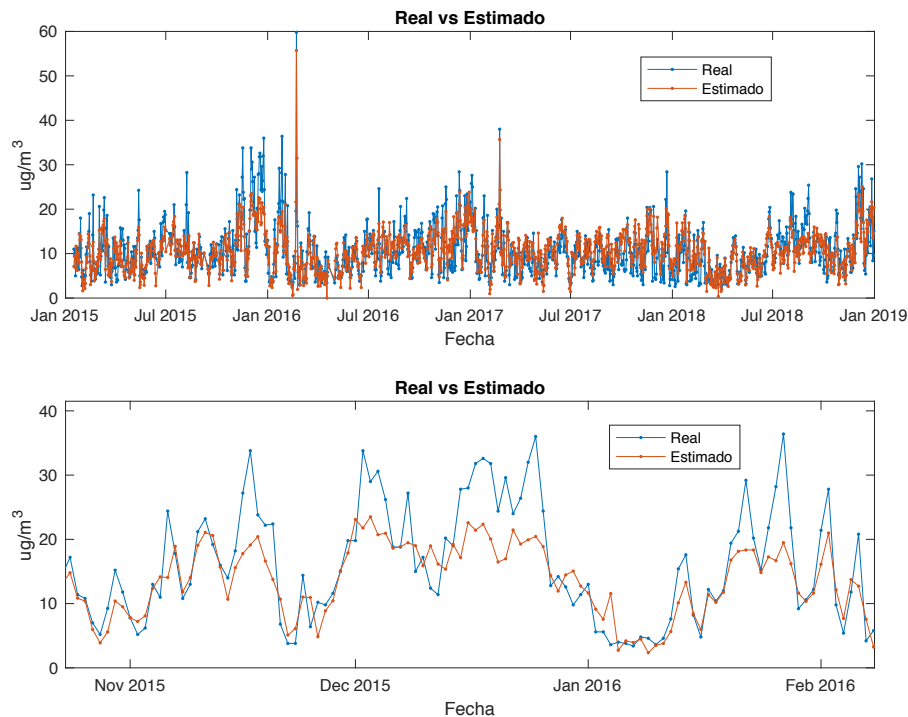


Figura 51. Estimado vs Real (Partículas menores a 2.5 micras)

El modelo obtenido para las partículas menores de 2.5 micras se muestra en la Tabla 12:

Tabla 12. Diagnóstico e interpretación del modelo (Partículas menores de 2.5 micras)

Variable	Estimación	SE	tStat	p-valor
Constante	23.65953771	1.058901622	22.34347103	1.90E-94
tmed	-1.127284642	0.142733435	-7.897831643	5.81E-15
tmax	0.969023447	0.107912862	8.979684389	8.82E-19
racha	-1.524612609	0.148020278	-10.30002528	5.23E-24
sol	-0.392692608	0.04643531	-8.456767302	6.99E-17
racha3	-5.83E-01	7.80E-02	-7.482128761	1.31E-13
tmax5	0.191304841	0.046483691	4.115526081	4.10E-05
velmed7	-0.770570405	2.15E-01	-3.582275662	3.53E-04
incContaminante	0.229748285	0.02178023	10.54847836	4.65E-25
Trafico1	5.41E-05	7.78E-06	6.956269388	5.41E-12
mesc_2	-0.529493445	4.93E-01	-1.074688269	2.83E-01
mesc_3	-1.117603106	0.522868899	-2.13744422	3.27E-02
mesc_4	-3.045775364	6.20E-01	-4.91083621	1.02E-06
mesc_5	-4.725323042	0.685764449	-6.89059202	8.47E-12
mesc_6	-5.301441383	0.837374666	-6.331026718	3.30E-10
mesc_7	-5.915129118	0.925335453	-6.392415961	2.24E-10
mesc_8	-5.353435261	0.915207605	-5.849421736	6.17E-09
mesc_9	-6.465520885	0.785431435	-8.231808147	4.28E-16
mesc_10	-5.287916317	0.639614918	-8.267343629	3.22E-16
mesc_11	-2.430452165	0.515066427	-4.71871595	2.62E-06
mesc_12	0.892823569	0.492833064	1.811614588	7.03E-02
sahara2016feb	40.6719354	3.373985507	12.05456731	7.25E-32
sahara2017feb	17.96376413	3.377793006	5.318195667	1.22E-07
tmin^2	0.01821894	0.002745053	6.637008109	4.61E-11
racha^2	5.51E-02	6.24E-03	8.822022231	3.38E-18

El modelo finalmente obtenido de las partículas menores de 10 micras consta de 13 términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 6.5860$$

$$R^2 = 0.7011$$

La Figura 52 muestra un comportamiento muy similar a las partículas menores de 2.5 micras. Únicamente se han considerado dos episodios de intrusión sahariana en la obtención del modelo. Por ello, existen residuos de mayores dimensiones al no tener ninguna variable capaz de explicar esos aumentos en los niveles de concentración tan repentinos. Finalmente, se muestra en la Figura 53 la estimación obtenida con el modelo. Se aprecia que se trata de una buena estimación siempre y cuando no exista una intrusión sahariana.

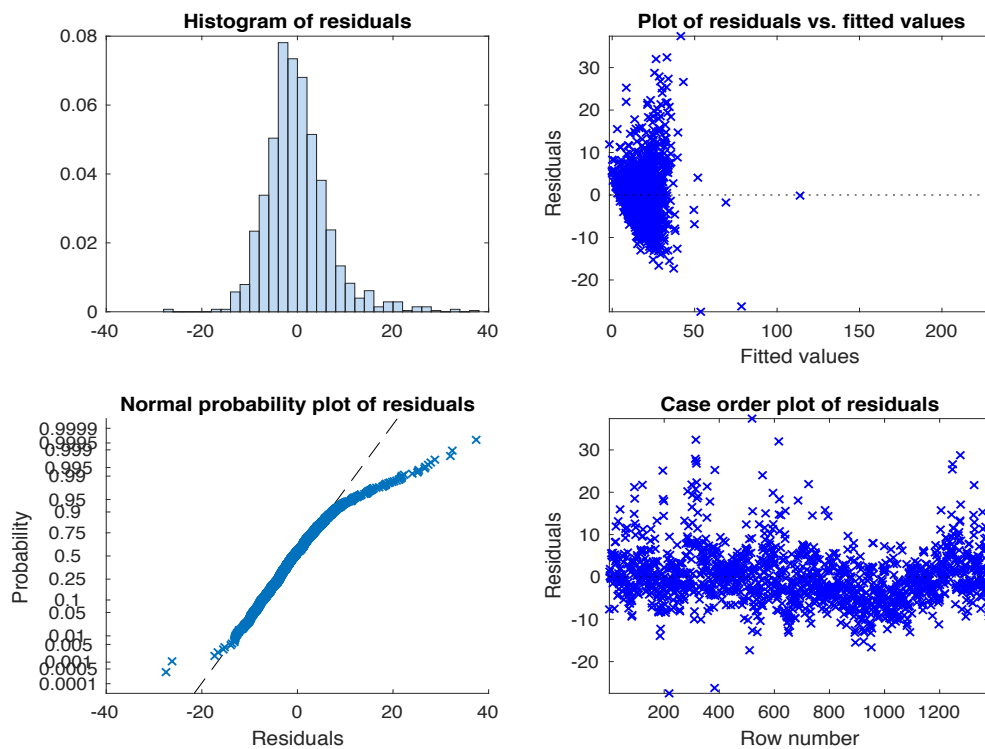


Figura 52. Análisis de los Residuos (Partículas menores de 10 micras)

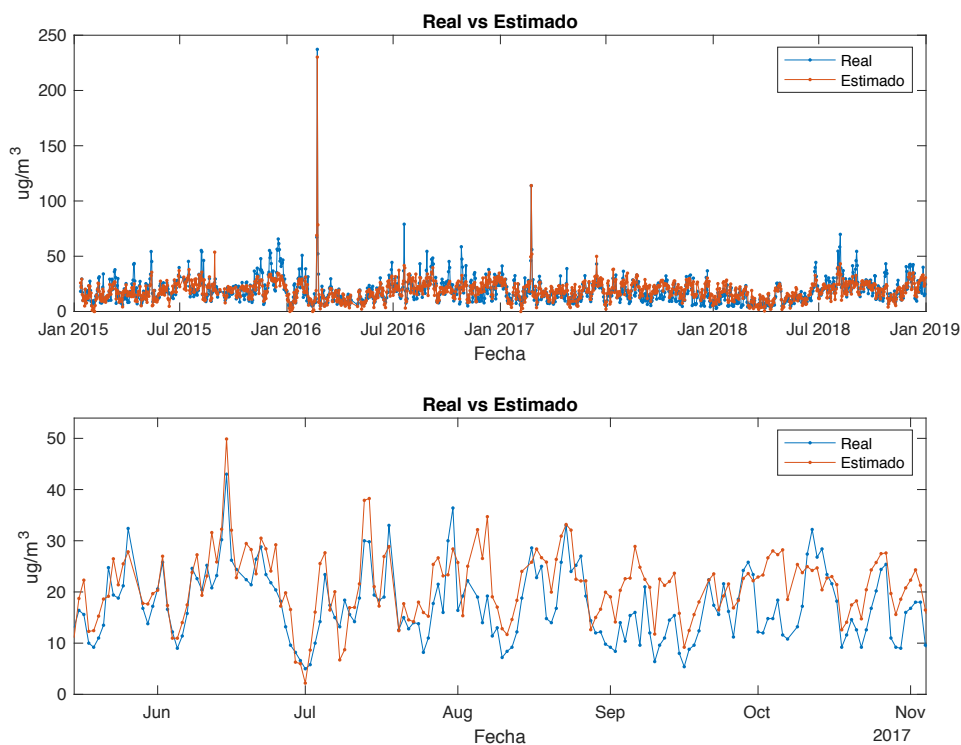


Figura 53. Estimado vs Real (Partículas menores de 10 micras)

El modelo obtenido para las partículas menores de 10 micras se muestra en la Tabla 13:

Tabla 13. Diagnóstico e interpretación del modelo (partículas menores de 10 micras)

Variable	Estimación	SE	tStat	p-valor
Constante	48.92804442	1.70721585	28.65955376	7.80E-142
tmin	-0.643814791	0.119401249	-5.392027292	8.20E-08
racha	-4.028346867	0.283600362	-14.20430794	8.20E-43
sol	-0.37062861	0.067432166	-5.49631772	4.62E-08
racha5	-1.127797335	0.119473434	-9.439733107	1.55E-20
Trafico1	1.34E-04	1.47E-05	9.094765523	3.25E-19
mesc_2	0.874144445	0.916358468	0.95393285	3.40E-01
mesc_3	-0.120888381	9.50E-01	-0.127195934	8.99E-01
mesc_4	-3.096604869	1.069029232	-2.896651257	3.83E-03
mesc_5	-7.98E+00	1.19E+00	-6.717060099	2.71E-11
mesc_6	-10.65980913	1.43E+00	-7.462956489	1.50E-13

mesc_7	-12.17657689	1.585276102	-7.681044882	2.99E-14
mesc_8	-7.314592487	1.55E+00	-4.716950714	2.64E-06
mesc_9	-7.753769141	1.291957592	-6.001566295	2.50E-09
mesc_10	-4.658739498	1.092415318	-4.264623007	2.14E-05
mesc_11	-2.24600612	0.939264472	-2.3912393	1.69E-02
mesc_12	1.621470774	0.925181873	1.752596783	7.99E-02
sahara2016feb	207.9608147	6.341527566	32.79348904	1.84E-174
sahara2017feb	85.72082149	6.388482477	13.41802561	1.18E-38
tmax:racha	0.088180143	0.005609177	15.72069172	2.59E-51
tmin3:incContaminante	0.017463681	0.001874621	9.315844481	4.68E-20
tmin^2	0.03569791	0.005212737	6.848207953	1.13E-11
racha^2	0.091368454	0.011732761	7.78746419	1.34E-14

8.4 ANÁLISIS DE RESULTADOS

Una vez obtenido el modelo se debe de comprobar su funcionamiento en los años 2019 y 2020 para chequear si generaliza correctamente. Para ello, mediante los coeficientes estimados con los datos de 2015-2018 se realizan predicciones de cómo sería la contaminación en los años 2019 y 2020 considerando que se conocen los valores de todas las variables de entrada del modelo.

Para las partículas menores de 2.5 micras la raíz del error cuadrático medio y el coeficiente de determinación ajustado calculados para estos dos años valen:

$$RMSE = 3.4756$$

$$R^2 = 0.6$$

El RMSE, sorprendentemente, disminuye ligeramente mientras el coeficiente de determinación también disminuye.

Representando la contaminación real y la estimada por el modelo (ver Figura 54) se aprecia una buena estimación durante la mayoría de la serie temporal. Es destacable que el RMSE sea menor en esta ocasión ya que no existe ninguna variable capaz de explicar las intrusiones

saharianas. A diferencia de otros contaminantes, no se aprecia una caída de la concentración del contaminante tan relevante debido al estado de alarma. El modelo predice los niveles de contaminación con bastante precisión durante este tiempo.

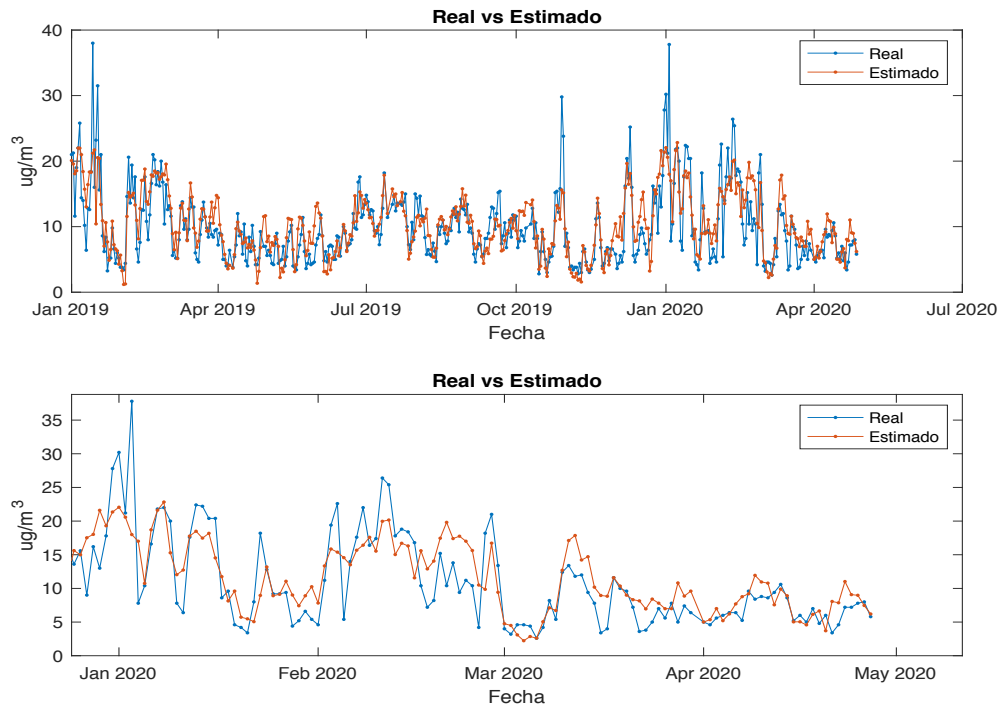


Figura 54. Predicción vs Real (partículas menores de 2.5 micras)

Para las partículas menores de 10 micras la raíz del error cuadrático medio y el coeficiente de determinación ajustado calculados para estos dos años valen:

$$RMSE = 7.1384$$

$$R^2 = 0.5083$$

El RMSE aumenta y el coeficiente ajustado de determinación disminuye mucho. Esto ocurre debido a que ninguna de las variables del modelo es capaz de explicar las intrusiones saharianas puesto que el modelo subestima estos episodios.

Representando la contaminación real y la estimada por el modelo (ver Figura 55) se observan las limitaciones debido a la ausencia de una variable explicativa capaz de explicar las intrusiones saharianas. Por ejemplo, se aprecia un episodio fuerte a finales de febrero de 2020. De todos modos, cuando estos episodios no se dan, la estimación de los niveles de concentración es buena. Durante el confinamiento las concentraciones no son especialmente bajas y el modelo estima la concentración de manera bastante precisa.

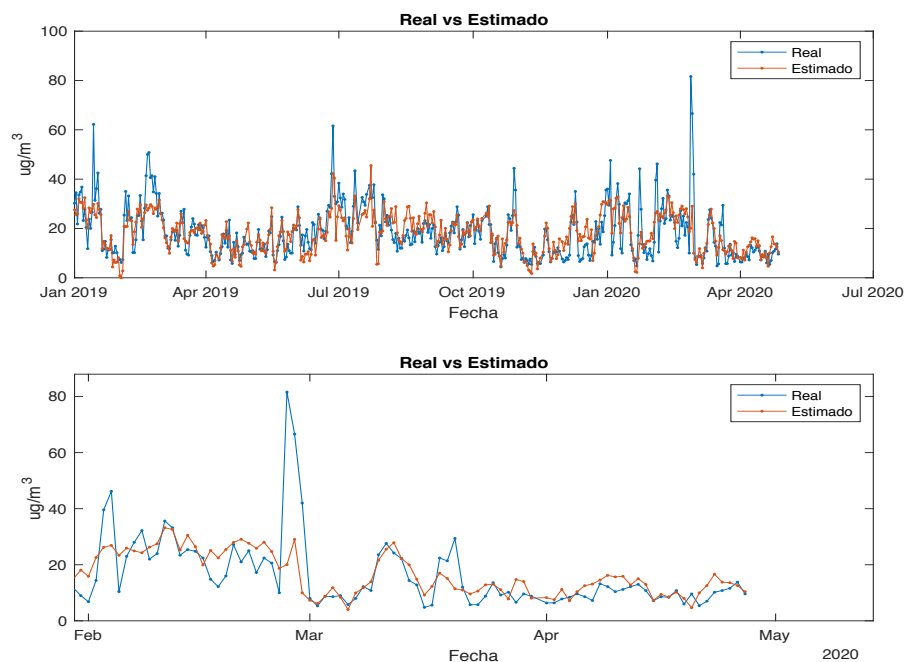


Figura 55. Predicción vs Real (partículas menores de 10 micras)

En definitiva, para las partículas se debe de añadir una nueva variable capaz de explicar las intrusiones saharianas lo cual mejoraría mucho el modelo obtenido. Aún así, mientras estos episodios no se den las estimaciones de ambos contaminantes son relativamente buenas.

Capítulo 9. MONÓXIDO DE CARBONO

En este capítulo se analiza en detalle los niveles de contaminación del Monóxido de Carbono en Madrid. Para ello se estudia la evolución temporal de la serie de dicho contaminante, ajustando un modelo de regresión lineal múltiple para estimar el nivel de contaminación en función de las variables explicativas seleccionadas.

9.1 MÁXIMOS Y MÍNIMOS

El Monóxido de Carbono al estar regulado a partir de las medias móviles octohorarias no se pueden estudiar si dichos límites se han ido superando a lo largo de los años debido a que los datos de estudio de la contaminación son diarios.

El valor más alto registrado es de $7.6 \text{ mg} / \text{m}^3$ y se dio el 5 de diciembre de 2001 en la estación de Cuatro Caminos. La mínima concentración detectada de este contaminante es 0.

9.2 ANÁLISIS

Con el fin de comprender el comportamiento del Monóxido de Carbono (CO) se debe de analizar su comportamiento e intentar comprender que variables influyen en dicho contaminante. Para ello, es imprescindible analizar si todas las estaciones de Madrid tienen datos similares, se debe de analizar la serie temporal del contaminante y, finalmente, identificar las variables que pueden influir en la variabilidad del contaminante.

Analizando la serie temporal (Figura 56) se observa un descenso progresivo en los niveles de contaminación desde el año 2001 hasta, aproximadamente, el 2012 en todas las estaciones de medida. Desde ese momento, la concentración de este contaminante se estabiliza hasta el día de hoy. Además, en verano el CO disminuye claramente con respecto a los meses de invierno. Nótese que la media de todas las estaciones se representa en negro en la Figura 56.

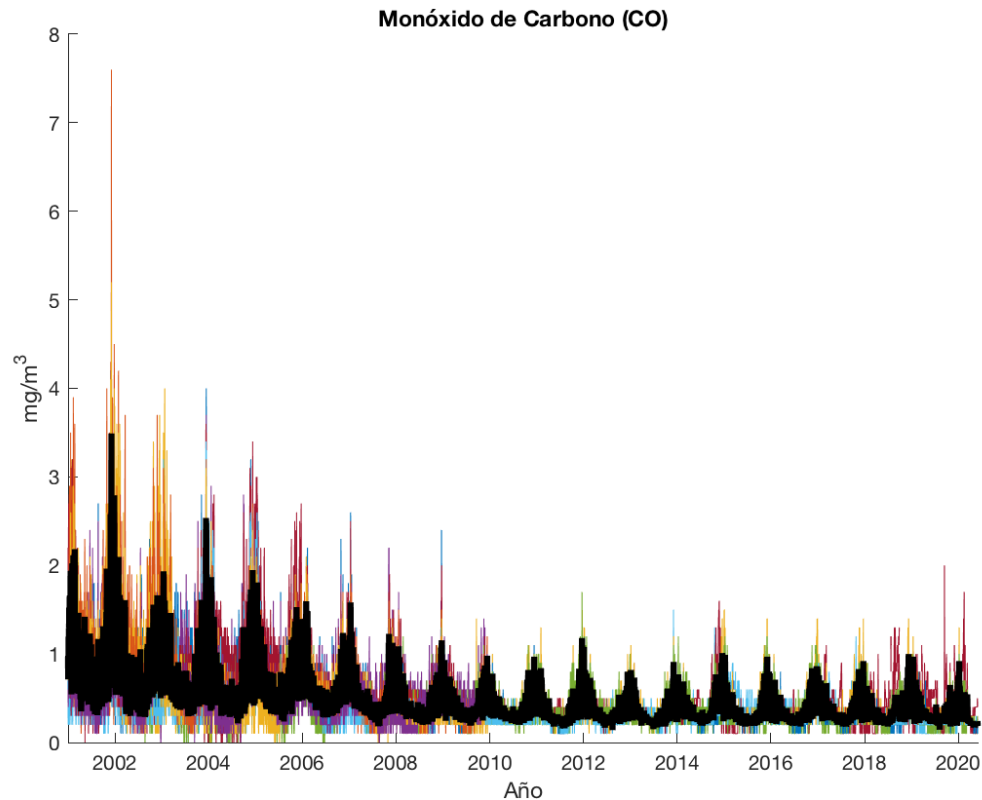


Figura 56. Serie Temporal del Monóxido de Carbono en las diferentes estaciones

A continuación, mediante el método de agrupación de *K-means* se han determinado conglomerados de días caracterizados por tener unos niveles de contaminación similares en todas las estaciones de medida. Para poder utilizar este método es necesario elegir previamente el valor adecuado de *K*, de tal forma que haya una buena relación entre complejidad y error de cuantización (QE). En la Figura 57 se muestra el criterio utilizado para asignar el valor de *K*. En este caso particular no existe un codo evidente, aunque un *K* razonable podría ser 3.

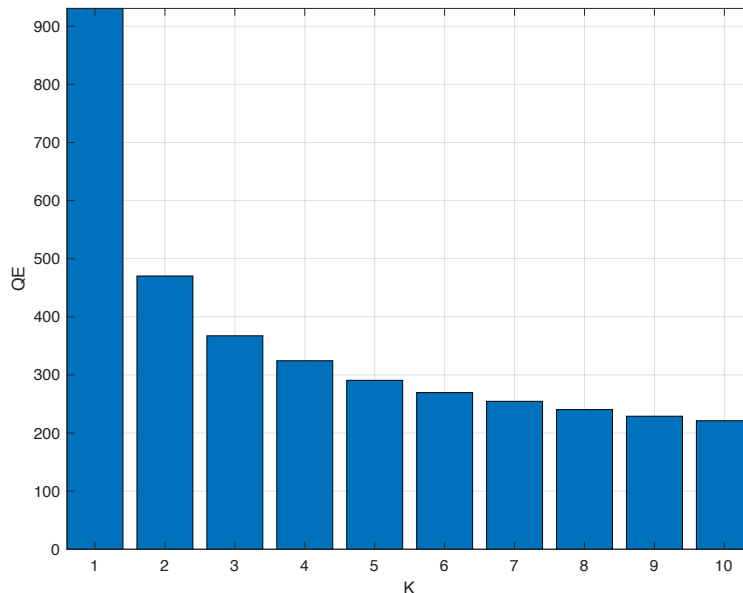


Figura 57. Obtención del número de conglomerados (Monóxido de Carbono)

Una vez seleccionado el número de conglomerados (K) se procede a representar gráficamente los distintos grupos obtenidos, en donde cada clúster representa un grupo de días similares en cuanto a niveles de concentración en las diferentes estaciones de medida. En la Figura 58 se muestran los centros de los conglomerados (centroides) por simplicidad, observándose que el clúster 2 representa los días de mayores niveles de contaminación, mientras que en el clúster 3 se encuentran los días de menor nivel de concentración del contaminante. Además, también se aprecia claramente que la estación con medidas más bajas de este contaminante para los tres clústeres es la estación de la Casa de Campo, lo cual es de esperar ya que se encuentra en una zona con bajo tráfico y alejado de Madrid Centro. En cambio, las que más concentración miden es la estación de la Plaza Elíptica para los tres clústeres.

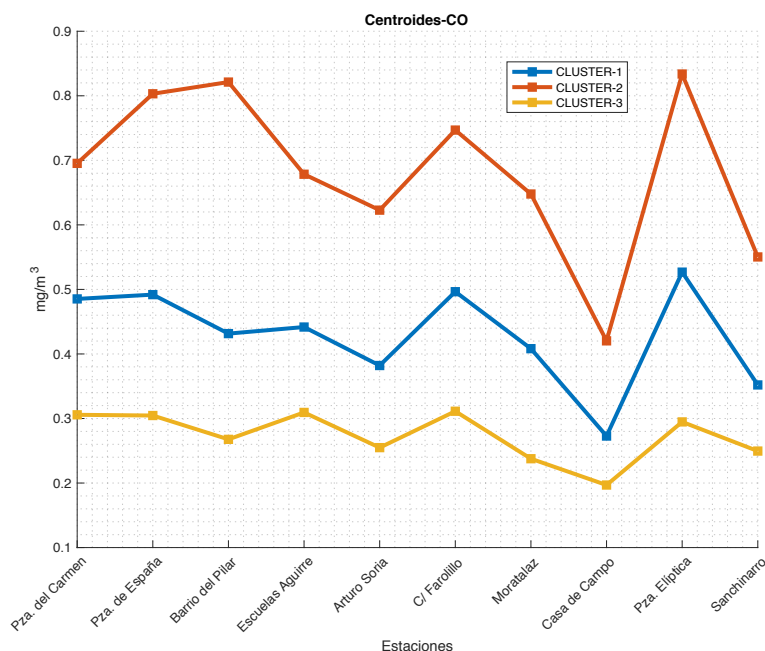


Figura 58. Centroides de los conglomerados (Monóxido de Carbono)

Para comprender mejor el comportamiento del Monóxido de Carbono, se procede a analizar si los niveles de concentración de dicho contaminante cambian y cómo con variables discretas como el mes, el día de la semana, la festividad o la estación de medida y las posibles interacciones entre ellas. Para ello, se ha realizado un análisis de la varianza (ANOVA). Los *p-valores*, sin tener en cuenta las interacciones, fueron de 0 para las cuatro variables mencionadas anteriormente por lo que todas ellas afectan al nivel de concentración presente de este contaminante en Madrid.

Este resultado quiere decir que hay al menos un grupo de cada variable para el que el contaminante es claramente distinto a los valores que toma para el resto de grupos. Por ejemplo, si se realiza una prueba de comparación múltiple a la estación de medida se observa que la estación de medida de la Casa de Campo es mucho menor a las demás (ver Figura 59). Realizando la misma comparación para el resto de las variables consideradas se aprecia que la contaminación es menor durante el fin de semana y los festivos y, finalmente, la concentración es mucho mayor durante el invierno que durante el verano.

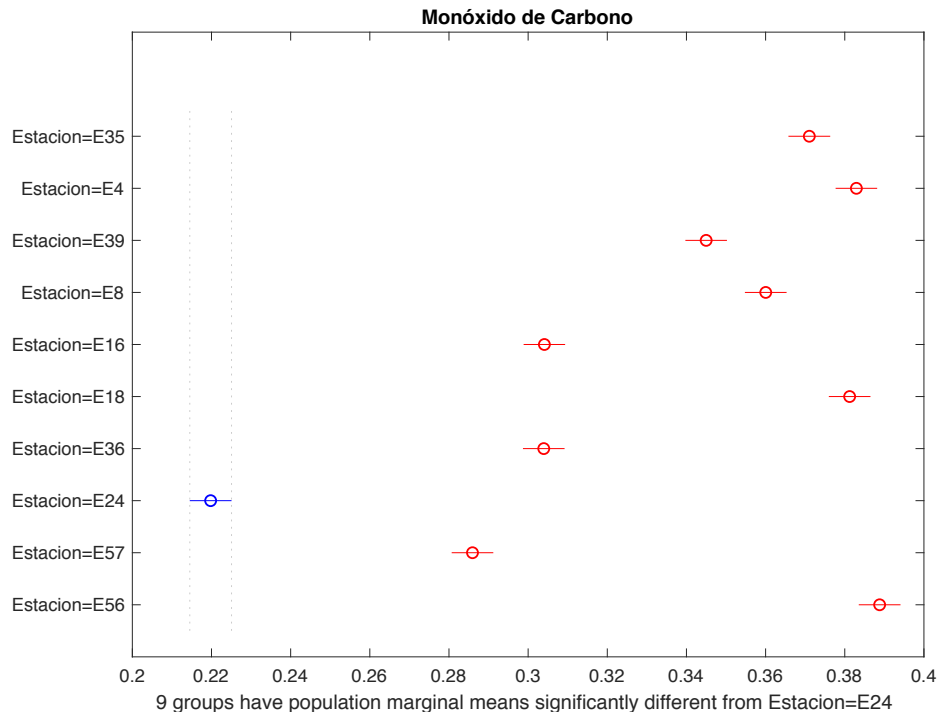


Figura 59. Prueba de comparación múltiple de las estaciones (Monóxido de Carbono)

Finalmente, en la Figura 60 se observa que hay 5 interacciones muy relevantes ya que tienen un *p-valor* de 0 o prácticamente 0. Además, la interacción entre la estación y la festividad tiene un *p-valor* de 0.0136 por lo que también parece ser relevante.

Analysis of Variance					
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
# Dia de la semana	1.45	5	0.28908	15.12	0
# Mes	45.7	9	5.07734	265.57	0
Estacion	9.89	9	1.09891	57.48	0
# Festivo	0	0	0	0	NaN
Dia de la semana*Mes	5.99	66	0.09072	4.75	0
Dia de la semana*Estacion	1.93	54	0.03568	1.87	0.0001
# Dia de la semana*Festivo	0.75	5	0.15033	7.86	0
Mes*Estacion	25.81	99	0.26075	13.64	0
# Mes*Festivo	0.9	9	0.10007	5.23	0
Estacion*Festivo	0.4	9	0.04418	2.31	0.0136
Error	685.38	35849	0.01912		
Total	1122.32	36118			

Constrained (Type III) sums of squares. Terms marked with # are not full rank.

Figura 60. Test Anova con interacción.

9.3 MODELO EXPLICATIVO DEL NIVEL DE CONCENTRACIÓN

Utilizando las variables creadas durante el Capítulo 5. y añadiendo como variable la diferencia de concentración de Monóxido de Carbono entre los dos días anteriores se ha planteado un modelo de regresión lineal múltiple para explicar el nivel de contaminación en función de las mismas. El modelo se ha ajustado con los datos de 2015 a 2018.

Dado que se dispone del nivel de concentración del Monóxido de Carbono en diferentes estaciones, se ha estudiado como variable de salida la media de todas las estaciones pertenecientes a la zona interior a la M-30. Además, en el caso del Monóxido de Carbono se utilizaron las variables de tráfico obtenidas mediante el método del análisis de las componentes principales. La variable Trafico1 representa la primera componente principal y la variable Trafico2 la segunda componente.

Inicialmente, se plantea un modelo con todas las variables disponibles y se van eliminando aquellas con un *p-valor* mayor. Posteriormente, se prueban distintas interacciones o relaciones cuadráticas con el fin reducir la raíz del error cuadrático medio (RMSE) y aumentar el coeficiente ajustado de determinación (R^2).

Finalmente, una vez obtenido el modelo con un número razonable de variables explicativas se impone la condición de que en ningún caso sea menor que cero el valor estimado por el modelo. El modelo finalmente obtenido consta de doce términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 0.0792$$

$$R^2 = 0.7361$$

El análisis de residuos (ver Figura 61) muestra una independencia en los residuos razonable, mientras que la normalidad de los mismos está cuestionada debido a la subestimación del modelo de los picos de invierno. En la Figura 62 se puede observar la verdadera concentración y la estimación realizada por el modelo. El modelo del Monóxido de Carbono

explica razonablemente bien la concentración de este contaminante en Madrid. En la figura se aprecia claramente la subestimación realiza por el modelo durante los picos de invierno.

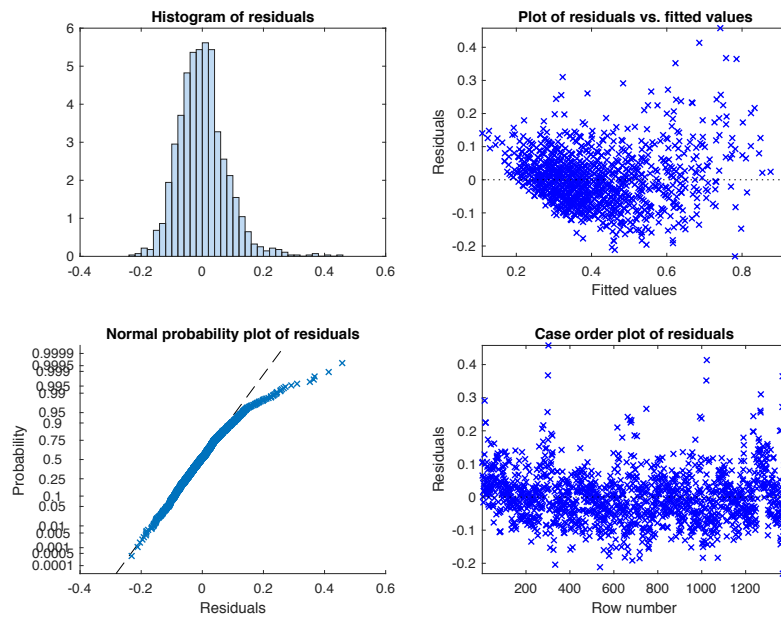


Figura 61 Análisis de Residuos (Monóxido de Carbono)

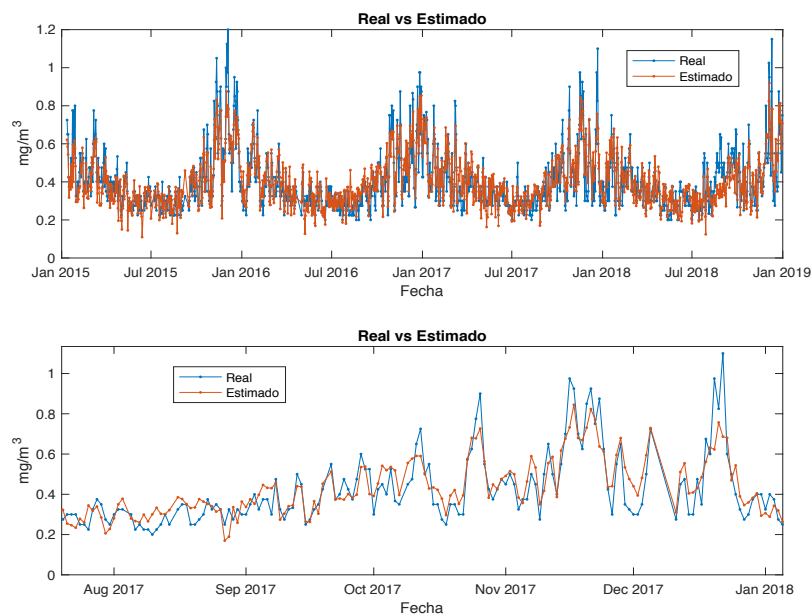


Figura 62. Estimación vs Real (Monóxido de Carbono)

En la Tabla 14 se muestran todas las variables explicativas utilizadas en el modelo del Monóxido de Carbono.

Tabla 14. Diagnóstico e interpretación del modelo (Monóxido de Carbono)

Variable	Estimación	SE	tStat	p-valor
Constante	0.978690091	0.03179147	30.78467594	1.56E-158
tmin	-0.0138255	0.001048396	-13.18729099	1.81E-37
tmax	0.012911004	0.000820369	15.73804165	2.04E-51
velmed	-0.065514107	0.006564569	-9.979956101	1.09E-22
racha3	-0.08223659	0.005795723	-14.18918573	9.85E-43
tmed7	-9.42E-03	1.80E-03	-5.226884635	1.99E-07
incContaminante	0.247271668	0.021793587	11.34607486	1.41E-28
Trafico1	2.12E-06	1.79E-07	11.88461501	4.55E-31
mesc 2	-0.017057158	0.010908027	-1.563725287	1.18E-01
mesc 3	-4.09E-03	1.19E-02	-0.34335636	7.31E-01
mesc 4	-0.031461731	1.44E-02	-2.180010803	2.94E-02
mesc 5	-0.089413027	0.01645554	-5.433612492	6.53E-08
mesc 6	-0.108288391	2.00E-02	-5.418807166	7.08E-08
mesc 7	-0.113815598	0.022223702	-5.12136084	3.47E-07
mesc 8	-0.065342513	0.022309837	-2.928865594	3.46E-03
mesc 9	-0.04822827	0.019236289	-2.5071504	1.23E-02
mesc 10	-0.031051492	0.015807997	-1.964290175	4.97E-02
mesc 11	0.026680397	0.012188917	2.188906295	2.88E-02
mesc 12	0.023335605	0.011238777	2.07634738	3.80E-02
velmed^2	0.006885773	0.000939104	7.332282372	3.86E-13
racha3^2	0.003172516	0.00027008	11.74658711	2.02E-30
tmin5^2	0.00037809	7.37E-05	5.130495038	3.31E-07

9.4 ANÁLISIS DE RESULTADOS

Una vez obtenido el modelo se debe de comprobar su funcionamiento en los años 2019 y 2020 para chequear si generaliza correctamente. Para ello, mediante los coeficientes estimados se realizan predicciones de cómo sería la contaminación en los años 2019 y 2020 considerando que se conocen los valores de todas las variables de entrada del modelo.

La raíz del error cuadrático medio y el coeficiente de determinación ajustado calculados para estos dos años valen:

$$RMSE = 0.0915$$

$$R^2 = 0.6901$$

El RMSE aumenta y el coeficiente de determinación ajustado disminuye, lógico ya que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Representando la contaminación real y la estimada por el modelo (ver Figura 63) se aprecia una buena estimación del modelo para estos años. Durante el confinamiento que empezaba el 15 de marzo de 2020 se observa un claro descenso repentino de la contaminación por Monóxido de Carbono. El modelo también estima dicha caída, pero después está constantemente sobreestimando. Por ello, es razonable pensar que el tráfico explica parte de la caída de contaminación, pero no toda. Esto podría ocurrir debido al parón de otros agentes contaminantes como los restaurantes, los hoteles o las pequeñas industrias.

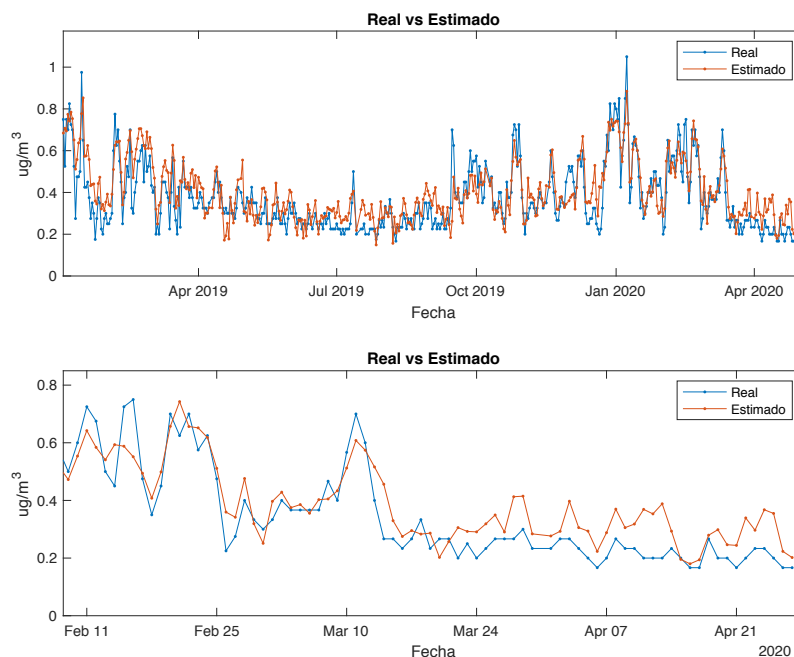


Figura 63. Predicción vs Real (Monóxido de Carbono)

Capítulo 10. OZONO

En este capítulo se analiza en detalle los niveles de contaminación del Ozono en Madrid. Para ello se estudia la evolución temporal de la serie de dicho contaminante, ajustando un modelo de regresión lineal múltiple para estimar el nivel de contaminación en función de las variables explicativas seleccionadas.

10.1 MÁXIMOS Y MÍNIMOS

El ozono al estar regulado a partir de las medias móviles octohorarias y no medias diarias no se pueden estudiar si dichos límites se han ido superando a lo largo de los años.

El valor más alto registrado es de $148 \mu\text{g} / \text{m}^3$ y se dio el 16 de julio de 2006 en la estación de la Casa de Campo. La mínima concentración detectada de este contaminante es $1 \mu\text{g} / \text{m}^3$.

10.2 ANÁLISIS

Con el fin de comprender el comportamiento del Ozono (O_3) se debe de analizar su comportamiento e intentar comprender que variables influyen en dicho contaminante. Para ello, es imprescindible analizar si todas las estaciones de Madrid tienen datos similares, se debe de analizar la serie temporal del contaminante y, finalmente, identificar las variables que pueden influir en la variabilidad del contaminante.

Analizando la serie temporal (Figura 64) se observa un aumento progresivo en los niveles de contaminación desde el año 2001 hasta, aproximadamente, el 2013 en todas las estaciones de medida. Desde ese momento, la concentración de este contaminante se estabiliza hasta el día de hoy. Esto llama la atención debido a que el resto de los contaminantes han experimentado un claro descenso desde el año 2001. Además, en verano el O_3 aumenta claramente con respecto a los meses de invierno. Esto, aunque sea un comportamiento

distinto al resto de los contaminantes, es de esperar ya que el Ozono requiere de la presencia de la luz solar para poder formarse al nivel del suelo. Nótese que la media de todas las estaciones se representa en negro en la Figura 64.

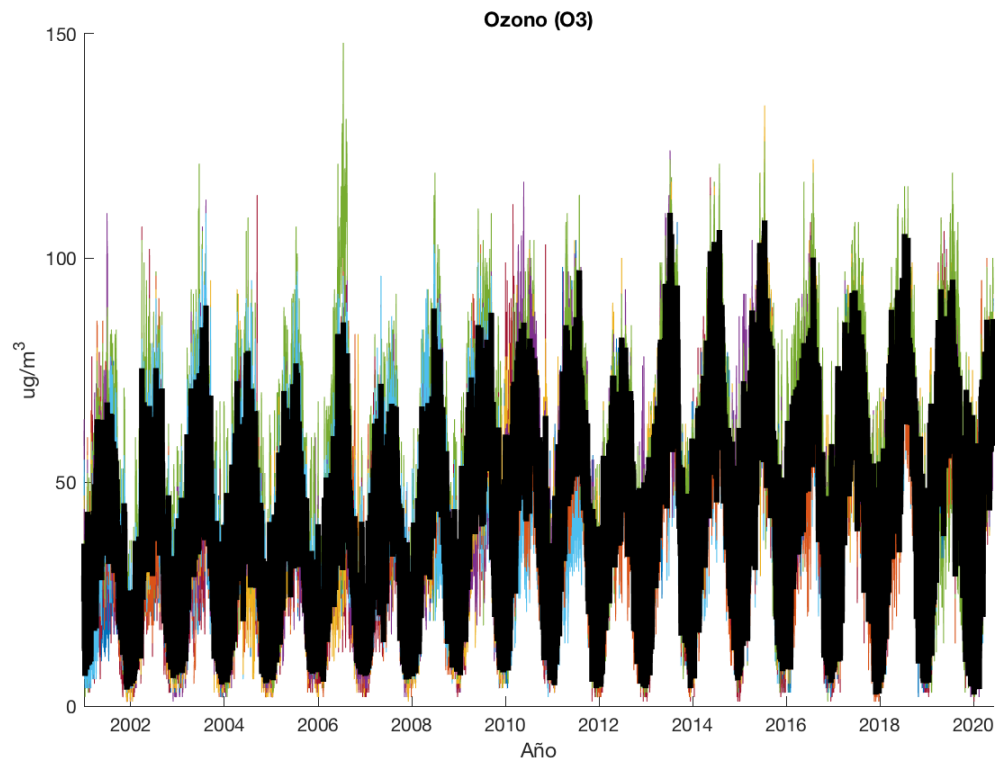


Figura 64. Serie Temporal del Ozono en las diferentes estaciones

A continuación, mediante el método de agrupación de *K-means*, se han determinado conglomerados de días caracterizados por tener unos niveles de contaminación similares en todas las estaciones de medida. Para poder utilizar este método es necesario elegir previamente el valor adecuado de *K*, de tal forma que haya una buena relación entre complejidad y error de cuantización (QE). En la Figura 65 se muestra el criterio utilizado para seleccionar el valor de *K*. En este caso particular no existe un codo evidente, aunque un valor de *K* razonable podría ser 4.

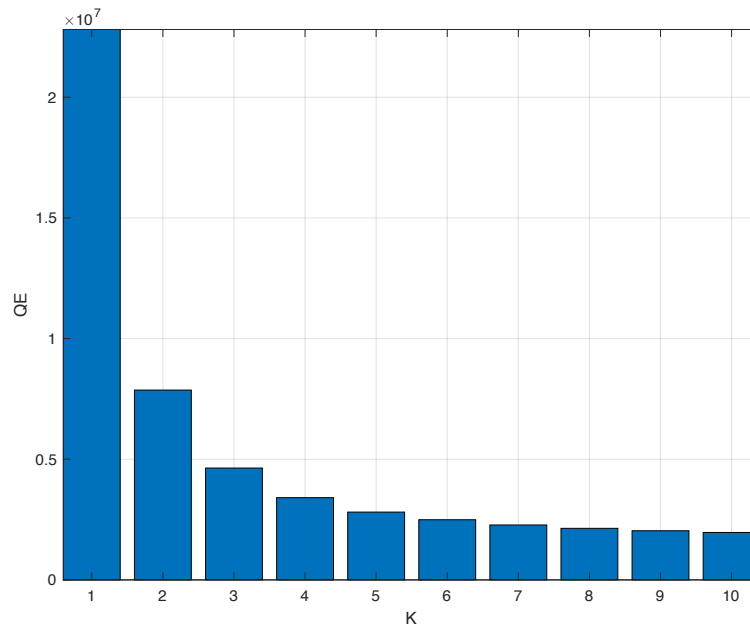


Figura 65. Obtención número de conglomerados (Ozono)

Una vez seleccionado el número de conglomerados (K) se procede a representar gráficamente los distintos grupos obtenidos, en donde cada clúster representa un grupo de días similares en cuanto a niveles de concentración en las diferentes estaciones de medida. En la Figura 66 se muestran los centros de los conglomerados (centroides) por simplicidad, observándose que el clúster de color rojo representa los días de mayores niveles de contaminación, mientras que en el clúster de color morado se encuentran los días de menor nivel de concentración del contaminante. Además, se aprecia claramente que la estación con medidas más altas de este contaminante son las estaciones de la Casa de Campo, el Pardo y el Parque Juan Carlos I. De nuevo, aún siendo un resultado distinto al resto de contaminantes, en el Capítulo 1. se advirtió que este contaminante se encuentra en mayores cantidades en las zonas rurales alejadas de zonas urbanas. A su vez, las estaciones que presentan menores medidas para todos los clústeres son estaciones como Escuelas Aguirre, Villaverde Alto y Plaza del Carmen.

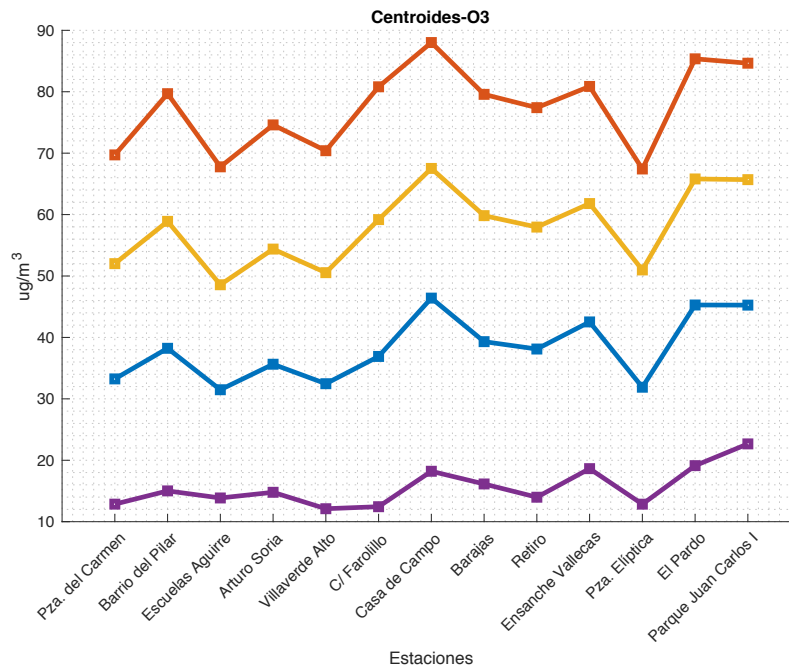


Figura 66. Centroides de los conglomerados (Ozono)

Para comprender mejor el comportamiento del Ozono, se procede a analizar si los niveles de concentración de dicho contaminante cambian y cómo con variables discretas como el mes, el día de la semana, la festividad o la estación y las posibles interacciones entre ellas. Para ello, se ha realizado un análisis de la varianza (ANOVA). Mediante esta prueba se confirma que el Ozono es un contaminante muy distinto a los demás debido a las razones que se observan mediante las pruebas de comparación múltiple: los festivos y los días entre semana la contaminación aumenta, y durante los meses cálidos también aumentan las concentraciones (ver Figura 64). Todo esto es exactamente lo contrario a lo que ocurre con el resto de los contaminantes. Aún así, igual que el resto, el nivel de Ozono depende de los meses, de la festividad, de la estación de medida y del día de la semana debido a que presentan un *p-valor* nulo.

Finalmente, las interacciones obtenidas fueron tres:

- Interacción entre el día de la semana y el mes (*p-valor* de 0)

- Interacción entre el día de la semana y la festividad (p -valor de 0)
- Interacción entre el mes y estación (p -valor de 0)

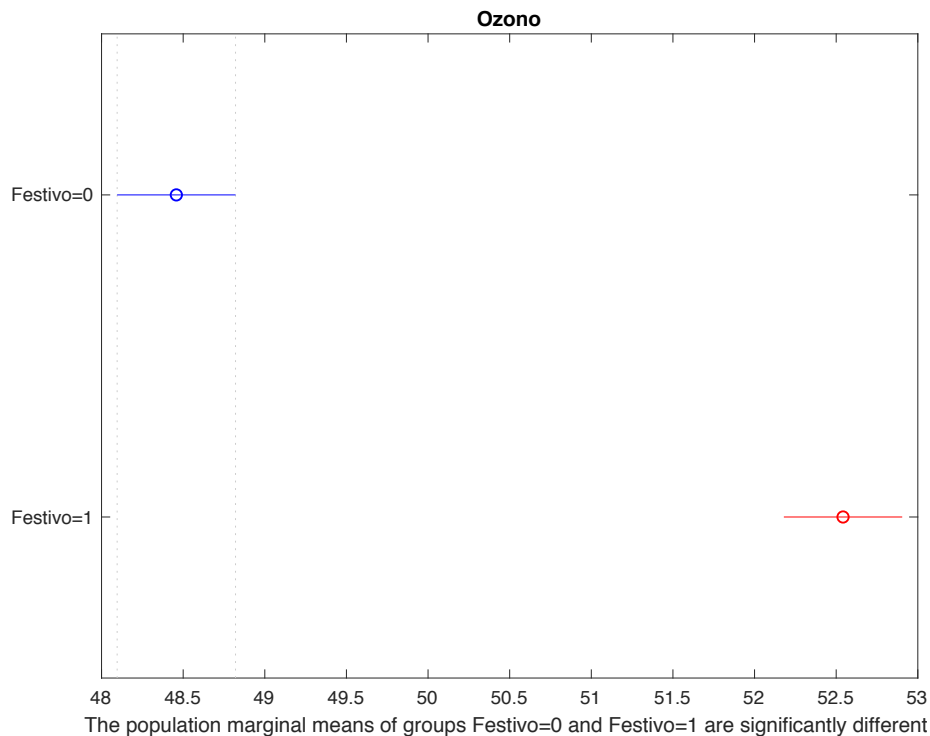


Figura 67. Prueba de Comparación Múltiple de la festividad (Ozono)

Se observa claramente en la Figura 67 que los días festivos la concentración es mayor que los días no festivos.

10.3 MODELO EXPLICATIVO DEL NIVEL DE CONCENTRACIÓN

Utilizando las variables creadas durante el Capítulo 5. y añadiendo como variable la diferencia de concentración de Ozono entre los dos días anteriores se ha planteado un modelo de regresión lineal múltiple para explicar el nivel de contaminación en función de las mismas. El modelo se ha ajustado con los datos de 2015 a 2018.

Dado que se dispone del nivel de concentración del Ozono en diferentes estaciones, se ha estudiado como variable de salida la media de todas las estaciones pertenecientes a la zona interior a la M-30. Además, en el caso del Ozono se utilizaron las variables de tráfico obtenidas mediante el método del análisis de las componentes principales. La variable Trafico1 representa la primera componente principal y la variable Trafico2 la segunda componente.

Inicialmente, se plantea un modelo con todas las variables disponibles y se van eliminando aquellas con un *p-valor* mayor. Posteriormente, se prueban distintas interacciones o relaciones cuadráticas con el fin reducir la raíz del error cuadrático medio (RMSE) y aumentar el coeficiente de correlación ajustado de determinación (R^2).

Finalmente, una vez obtenido el modelo con un número razonable de variables explicativas se impone la condición de que en ningún caso sea menor que cero el valor estimado por el modelo. El modelo finalmente obtenido consta de diez términos, con las siguientes medidas de bondad del ajuste:

$$RMSE = 9.8247$$

$$R^2 = 0.801$$

El análisis de residuos (ver Figura 68) muestra diferencias en las colas de la distribución de probabilidad normal. Además, los residuos muestran algo de periodicidad debido a la alta variación del Ozono durante los meses de verano. Aún así, los residuos parecen independientes y normales. En la Figura 69 se puede observar la verdadera concentración y la estimación realizada por el modelo.

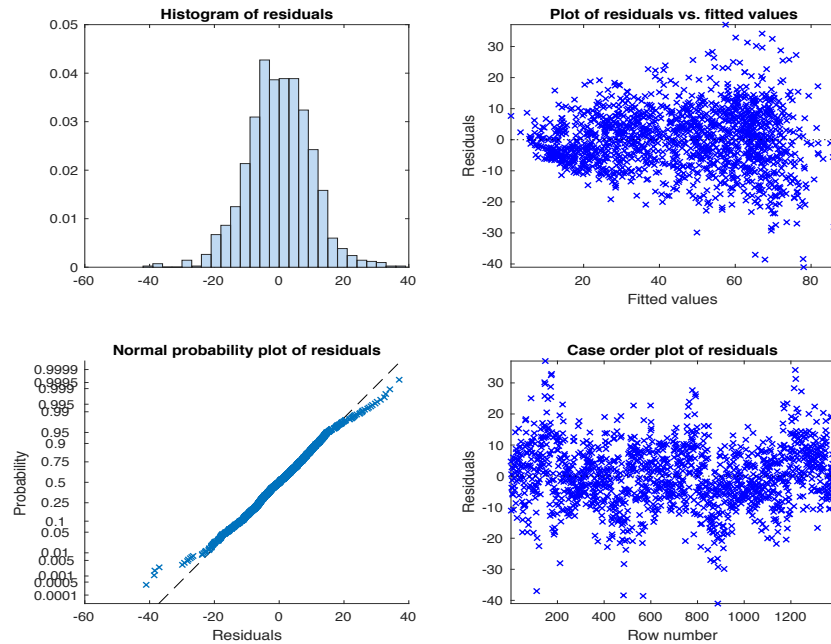


Figura 68 Análisis de Residuos (Ozono)

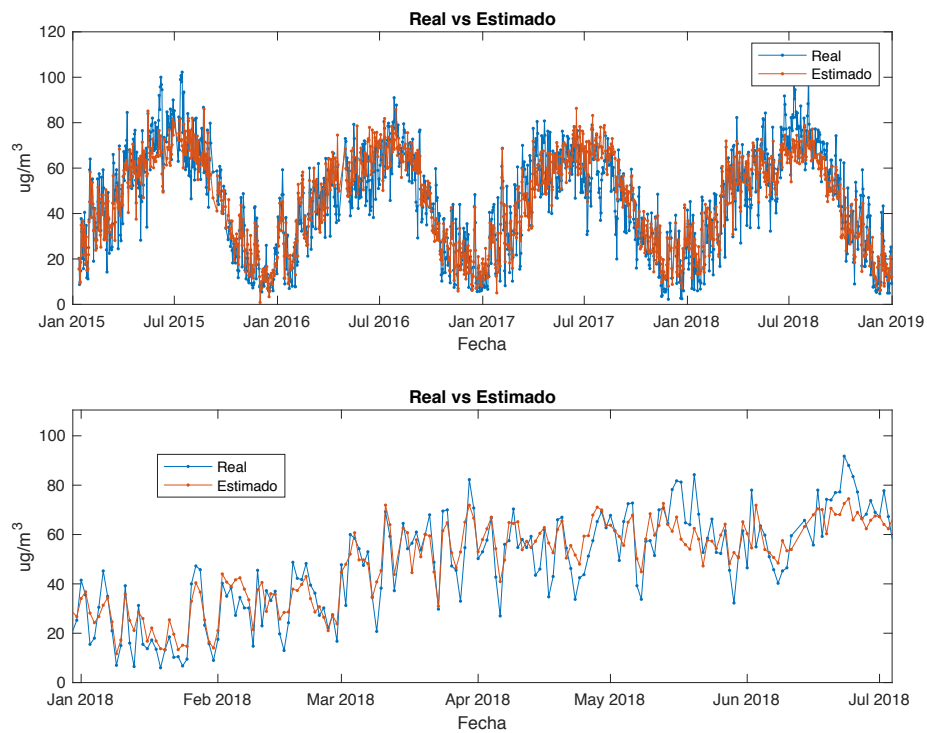


Figura 69. Estimación vs Real (Ozono)

En la Tabla 15 se muestran todas las variables explicativas utilizadas en el modelo del Ozono.

Tabla 15. Diagnóstico e interpretación del modelo (Ozono)

Variable	Estimación	SE	tStat	p-valor
Constante	22.29570587	2.673247953	8.340305971	1.79E-16
tmed	3.435020913	0.310377732	11.06722733	2.51E-27
tmax	-4.440321942	0.337184739	-13.16881052	2.24E-37
velmed	2.333100509	0.283433882	8.231551178	4.26E-16
sol	1.899916564	0.130593679	14.54830421	1.08E-44
racha3	0.870882062	0.16188676	5.379575599	8.77E-08
incContaminante	0.197040521	0.022568341	8.730837638	7.23E-18
Trafico1	-0.000203451	2.19E-05	-9.273059147	6.82E-20
mesc 2	9.080801202	1.347618681	6.73840555	2.35E-11
mesc 3	18.70894679	1.432686957	13.0586425	8.13E-37
mesc 4	29.49168052	1.599048048	18.44327352	5.16E-68
mesc 5	30.34844904	1.775930275	17.08876157	1.70E-59
mesc 6	27.59127402	2.122219847	13.00113844	1.59E-36
mesc 7	24.25327227	2.332161337	10.39948304	1.97E-24
mesc 8	21.08481716	2.309083803	9.131248132	2.36E-19
mesc 9	21.99838768	1.978761823	11.11724889	1.50E-27
mesc 10	8.854649169	1.666905092	5.312029588	1.26E-07
mesc 11	2.715426643	1.405137435	1.932498967	0.053504027
mesc 12	-2.426052997	1.353416725	-1.792539543	0.073267303
tmax^2	0.036386282	0.004603314	7.90436669	5.50E-15

10.4 ANÁLISIS DE RESULTADOS

Una vez obtenido el modelo se debe de comprobar su funcionamiento en los años 2019 y 2020 para chequear si generaliza correctamente. Para ello, mediante los coeficientes estimados se realizan predicciones de cómo sería la contaminación en los años 2019 y 2020 considerando que se conocen los valores de todas las variables de entrada del modelo.

La raíz del error cuadrático medio y el coeficiente de determinación ajustado calculados para estos dos años valen:

$$RMSE = 11.4501$$

$$R^2 = 0.7348$$

La raíz del error cuadrático medio aumenta y el coeficiente ajustado de determinación disminuye, lógico ya que estos datos no han sido utilizados para ajustar los coeficientes del modelo.

Evaluando la serie temporal (ver Figura 69) se aprecia una buena estimación del modelo para estos años con la excepción de dos caídas repentinas de concentración de Ozono en 2019. Durante el confinamiento que comenzaba el 15 de marzo los niveles de Ozono se han mantenido similares a los de otros años. El modelo obtenido para el Ozono es una buena estimación de la realidad.

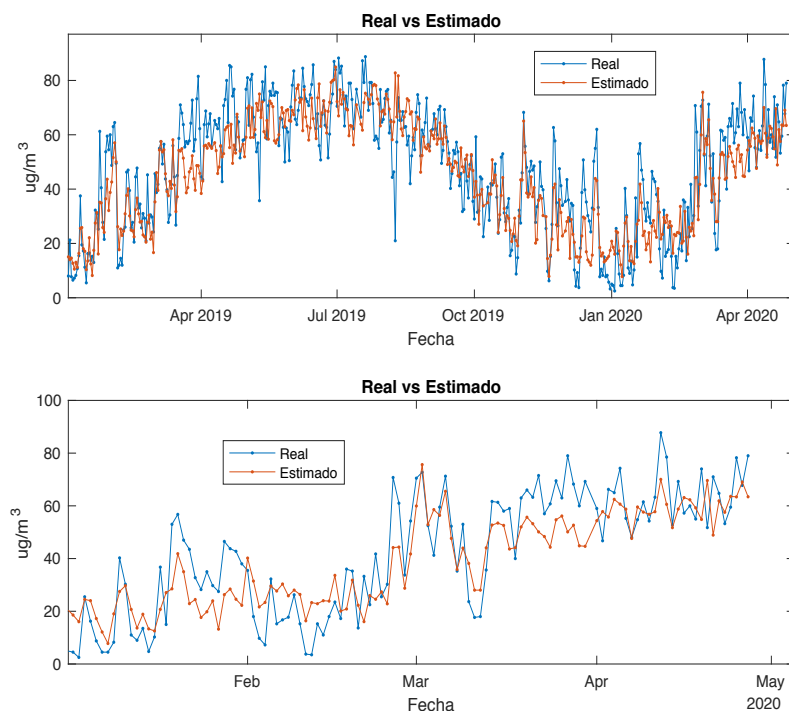


Figura 70. Predicción vs Real (Ozono)

Capítulo 11. CONCLUSIONES Y TRABAJOS FUTUROS

11.1 CONCLUSIONES

Una obtención adecuada de variables es fundamental para poder explicar cualquiera de los distintos contaminantes. La preparación de esta información, especialmente con el tráfico, ha sido complicada debido a la falta de homogeneidad y presencia de huecos en los datos.

El análisis realizado para cada uno de los contaminantes previa implementación del modelo muestra una serie de comportamientos que todos los contaminantes comparten, con la excepción del Ozono. Concretamente, se aprecia un aumento progresivo de los niveles de contaminación de lunes a jueves y el día de mejor calidad del aire es el domingo. El verano presenta valores mucho menores a los obtenidos durante el invierno y los días festivos la contaminación es menor. También, se puede concluir que la calidad del aire en Madrid ha ido mejorando a medida que pasan los años. Esto puede ser debido a la renovación del parque automovilístico, con un aumento progresivo de coches eléctricos o híbridos. El Ozono, en cambio, se comporta de una manera completamente opuesta al resto de contaminantes.

Se han obtenido ocho modelos utilizando regresión lineal múltiple. Estos modelos son razonables para la mayoría de los contaminantes. Cabe destacar una posible mejora en el caso de las partículas si se incluyese una variable de intervención para recoger las intrusiones de masas de aire sahariano. El Dióxido de Azufre no se consiguió explicar razonablemente bien con las variables planteadas, posiblemente debido a la baja influencia del tráfico en este contaminante. Un resumen de los resultados obtenidos se muestra en la Tabla 16, en donde se aprecian unos resultados relativamente buenos para el Monóxido de Carbono, el Monóxido, Dióxido y Óxido de Nitrógeno y el Ozono.

Tabla 16. Resumen de los resultados obtenidos

Contaminante	R^2 (Aprendizaje)	R^2 (Test)	RMSE(Aprendizaje)	RMSE(Test)
Dióxido de Azufre	0.3233	0.2852	2.2396	3.2943
Monóxido de Carbono	0.7361	0.6901	0.0792	0.0915
Monóxido de Nitrógeno	0.7359	0.7213	15.2968	16.5054
Dióxido de Nitrógeno	0.7967	0.8139	7.6904	9.3864
Óxido de Nitrógeno	0.7672	0.7843	29.2123	32.7542
Partículas < 10 micras	0.7011	0.5083	6.586	7.1384
Partículas < 2.5 micras	0.625	0.6	3.4762	3.4756
Ozono	0.801	0.7348	9.8247	11.4501

En definitiva, se ha logrado analizar cuantitativamente la importancia de las variables influyentes en los contaminantes mediante los modelos de regresión ajustados. De dichos modelos, el tráfico, las temperaturas, el viento o el mes del año han sido las variables que más afectan a la contaminación. La lluvia, sorprendentemente, apenas ha tenido relevancia para ninguno de los modelos obtenidos, con la excepción del modelo de Dióxido de Nitrógeno. Tampoco se han utilizado para ninguno de los modelos las variables día de la semana ni la festividad debido a que el comportamiento de la variable de tráfico explica las bajadas o subidas de concentraciones en esos días.

El estado de alarma producido por la pandemia también ha hecho que Madrid mejore sustancialmente el nivel de calidad del aire. Las variables consideradas en este trabajo, a excepción del tráfico, no deberían de cambiar durante este tiempo. Por tanto, la única variable capaz de predecir este descenso es el tráfico que se redujo de manera drástica de un día para otro. Por ejemplo, en el caso del Monóxido de Carbono, el modelo ajustado estima una fuerte caída el 15 de marzo (día de inicio del confinamiento) pero sobreestima ligeramente los niveles de concentración. Esto quiere decir que posiblemente existan otros elementos contaminantes que no se han tenido en cuenta en los modelos y que también cerraron sus puertas por la pandemia (restaurantes, hoteles, fábricas, etc). En cambio, el modelo del Óxido de Nitrógeno estima, a partir del 15 de

marzo, valores iguales a cero lo cual indica que el peso otorgado a la variable de tráfico para ese modelo es algo superior al real.

11.2 TRABAJOS FUTUROS

Los modelos realizados se han llevado a cabo tomando la media de todas las estaciones que pertenecen a la zona interior a la M-30 en Madrid. Por lo tanto, una posible mejora es intentar ajustar un modelo para cada contaminante y cada estación. Es cierto que los resultados serían muy similares para estaciones próximas, pero daría una idea mucho más detallada de la importancia de cada una de las variables para cada estación de contaminación. Por ejemplo, en la estación de Casa de Campo o el Pardo los niveles de tráfico son mucho menores a las zonas urbanas de Madrid y, por tanto, sea posible entender de mejor manera la influencia de las demás variables.

En segundo lugar, la obtención de una mayor cantidad de variables para aumentar la calidad de los modelos. Esto es especialmente importante para el Dióxido de Azufre puesto que el resultado obtenido es mucho peor que para el resto de los contaminantes. Las variables que podrían mejorar estos resultados serían aquellas relacionadas con la quema de carbón, como por ejemplo el uso de las calefacciones centrales de este tipo. Por otro lado, se identificó la importancia de la variable de intrusiones saharianas en el estudio de las partículas menores de 10 y 2.5 micras. Esta variable es difícil de introducir en los modelos debido a que existen varios episodios cada año y su intensidad varía constantemente.

Por último, se podrían construir modelos de carácter predictivo a partir de los desarrollados. Para ello, sería necesario disponer de previsiones para las variables de entrada que mejor expliquen las series temporales de los distintos contaminantes, como las proporcionadas por la Agencia Estatal de Meteorología (AEMET).

BIBLIOGRAFÍA

- [1] World Health Organization, Marzo 2014. OMS | 7 millones de muertes cada año debidas a la contaminación atmosférica. Recuperado 14 de mayo de 2020, de <https://www.who.int/mediacentre/news/releases/2014/air-pollution/es/>
- [2] CMA, Dióxido de azufre. Consejería de Medio Ambiente y Ordenación del Territorio, Junta de Andalucía. Recuperado 1 de julio de 2020, de <http://www.juntadeandalucia.es/medioambiente/site/porta/web/menuitem.7e1cf46ddf59bb227a9ebe205510e1ca/?vgnextoid=815af7ffa3828310VgnVCM1000001325e50aRCRD&vgnextchannel=33fc99e6e6828310VgnVCM2000000624e50aRCRD>
- [3] Ministerio para la Transición Ecológica y el Reto Demográfico. Monóxido de carbono. Recuperado 3 de junio de 2020, de <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/monoxido-carbono.aspx>
- [4] Ministerio para la Transición Ecológica y el Reto Demográfico. Óxidos de Nitrógeno. Recuperado 3 de junio de 2020, de <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/oxidos-nitrogeno.aspx>
- [5] Consejería de Salud de la Región de Murcia. Murciasalud. Recuperado 7 de junio de 2020, de <http://www.murciasalud.es/pagina.php?id=244308&idsec=1573>
- [6] Ministerio para la Transición Ecológica y el Reto Demográfico. Ozono. Recuperado 7 de junio de 2020, de <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/ozono.aspx>
- [7] Ministerio para la Transición Ecológica y el Reto Demográfico. (s. f.-b). Óxidos de Nitrógeno. Recuperado 3 de junio de 2020, de <https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/oxidos-nitrogeno.aspx>
- [8] España. Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire. *Boletín Oficial del Estado*, 29 de enero de 2011 núm. 25.

- [9] Sánchez Úbeda, Eugenio Francisco. Febrero 2019. ANOVA (Estadística II). Universidad Pontificia de Comillas.
- [10] Sánchez Úbeda, Eugenio Francisco. Febrero 2019. Principal Component Analysis (Estadística II). Universidad Pontificia de Comillas.
- [11] Sánchez Úbeda, Eugenio Francisco. Febrero 2019. Clustering (Estadística II). Universidad Pontificia de Comillas.
- [12] Sánchez Úbeda, Eugenio Francisco. Febrero 2019. Linear Regression (Estadística II). Universidad Pontificia de Comillas.
- [13] Ramos-Herrera, S., Bautista-Margulis, R. y Valdez-Manzanilla, A. (2010). *Estudio estadístico de la correlación entre contaminantes atmosféricos y variables meteorológicas en la zona norte de Chiapas, México. Universidad y Ciencia*, vol. 26(1), 65-80.
- [14] Salini Calderón, G., y Pérez Jara, P. (2006). Estudio de las series temporales de contaminación ambiental mediante técnicas de redes neuronales artificiales. *Ingeniare. Revista chilena de ingeniería*, vol. 14(3), 284-290.
- [15] Ruiz Pastor, C. (2017). *Técnicas de aprendizaje automático para el análisis de datos de calidad del aire* (Trabajo de fin de grado). Universidad Autónoma de Madrid, España.
- [16] Caramillo-Ramírez, P., Sánchez López, A., Calva-Rosales, L. y Pérez-Vázquez, I. (2014). *Análisis de datos de calidad del aire de la Zona Metropolitana del Valle de México mediante técnicas de agrupamiento. Research in Computing Science* 72, 137-150.
- [17] Dirección General de Sostenibilidad y Control Ambiental. (2020). *Calidad del Aire 2019*.
- [18] Ayuntamiento de Madrid. Calidad del aire. Datos diarios años 2001 a 2020 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 10 de febrero de 2020, de <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=aecb88a7e2b73410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>
- [19] A.E. Meteorolog. AEMET OpenData. Recuperado 1 de marzo de 2020, de http://www.aemet.es/es/datos_abiertos/AEMET_OpenData
- [20] Ayuntamiento de Madrid. Tráfico. Histórico de datos del tráfico desde 2013 - Portal de datos abiertos del Ayuntamiento de Madrid. Recuperado 15 de marzo de 2020, de <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=33cb30c367e78410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfint=default>

- [21] Dirección General de Sostenibilidad y Control Ambiental. (2020). *Boletín diario de calidad del aire del Ayuntamiento de Madrid 07/07/2020*.
- [22] Dirección General de Sostenibilidad y Control Ambiental. (2017). *Calidad del Aire 2016*.

ANEXO I – OBJETIVOS DE DESARROLLO SOSTENIBLE

De cara al 2030, se deben de cumplir unos Objetivos de Desarrollo Sostenible (ODS). Existen 17 objetivos de desarrollo sostenible de los que este Trabajo de Fin de Grado aporta a las metas del objetivo 11 y 13. El objetivo 11 es lograr que las ciudades sean más inclusivas, seguras, resilientes y sostenibles. La meta que este trabajo puede ayudar a lograr es el siguiente:

- De aquí a 2030, reducir el impacto ambiental negativo per cápita de las ciudades, incluso prestando especial atención a la calidad del aire y la gestión de los desechos municipales y de otro tipo

Concretamente, el estudio realizado de los principales contaminantes de Madrid. Ayuda a buscar maneras de mejorar la calidad del aire en las grandes ciudades. Por ejemplo, según lo visto en este Trabajo de Fin de Grado, las medidas a tomar para evitar altas concentraciones de contaminantes en la atmósfera deberían de ser implementadas durante el invierno debido a la enorme concentración de contaminantes durante ese tiempo. Según se ha visto, los contaminantes van reduciendo sus concentraciones en Madrid de manera progresiva, posiblemente debido al aumento de coches eléctricos o híbridos, por lo que es de esperar que la calidad del aire mejore sustancialmente de aquí a 2030 cumpliendo el objetivo.

El siguiente objetivo es el 13, adoptar medidas urgentes para combatir el cambio climático y sus efectos. Las metas que este trabajo puede ayudar a lograr son las siguientes:

- Incorporar medidas relativas al cambio climático en las políticas, estrategias y planes nacionales.
- Mejorar la educación, la sensibilización y la capacidad humana e institucional respecto de la mitigación del cambio climático, la adaptación a él, la reducción de sus efectos y la alerta temprana.

La contaminación es una causa fundamental del cambio climático. Por ello, el estudio realizado en el Trabajo de Fin de Grado puede ayudar a comprender las variables fundamentales en la mayoría de los contaminantes y así, implementar políticas útiles que se centren en mejorar la calidad del aire. Conociendo las variables significativas para cada contaminante es posible alertar de situaciones de alta contaminación como, por ejemplo, días secos sin ventilación de la atmósfera y de temperaturas muy bajas. Tomar medidas este tipo de días ayudaría a reducir los niveles de contaminación y, en definitiva, a cumplir las metas señaladas anteriormente.

ANEXO II – FUENTES DE INFORMACIÓN

11.3 INFORMACIÓN DE CONTAMINACIÓN

Los datos de contaminación requeridos para poder llevar a cabo el proyecto se han descargado en su totalidad de la información aportada por el Ayuntamiento de Madrid. Para acceder a dicha información se deben seguir los siguientes pasos:

1. Acceder al siguiente *link*: <https://datos.madrid.es/portal/site/egob>
2. En el buscador se debe buscar: Calidad del aire. Datos diarios años 2001 a 2020.
3. Otra opción: Catálogo de datos > Conjuntos de datos > Calidad del aire. Datos diarios años 2001 a 2020
4. En descargas se muestra la opción de obtener la información de diferentes maneras. Ficheros en formato TXT, CSV O XML. En este caso, se accede al archivo TXT. Se debe copiar todo el contenido y pegarlo en cualquier tipo de documento que EXCEL sea capaz de abrir más adelante. Se optó por guardar la información en un Word pero se debe cambiar el formato del archivo a TXT.
5. Mediante Excel se debe acceder a la pestaña Archivo > Abrir. Se debe abrir el archivo en el que se pegaron los datos y seleccionar la opción delimitados por comas.
6. Seleccionar la opción Delimitados y pulsar en siguiente. Se debe elegir la opción delimitado por comas que aparece a continuación. Después se finaliza y los datos aparecen en una tabla de Excel.

7. Lo que se debe de obtener se muestra en la Figura 71, una tabla de una gran cantidad de filas y 70 columnas:

	A	B	C	D	E	F	G
1	28	79	4	1	38	4	2020
2	28	79	4	1	38	4	2020
3	28	79	4	1	38	4	2020
4	28	79	4	1	38	4	2020
5	28	79	4	1	38	4	2020
6	28	79	4	6	48	4	2020
7	28	79	4	6	48	4	2020
8	28	79	4	6	48	4	2020
9	28	79	4	6	48	4	2020
10	28	79	4	6	48	4	2020
11	28	79	4	7	8	4	2020
12	28	79	4	7	8	4	2020
13	28	79	4	7	8	4	2020
14	28	79	4	7	8	4	2020
15	28	79	4	7	8	4	2020
16	28	79	4	8	8	4	2020
17	28	79	4	8	8	4	2020
18	28	79	4	8	8	4	2020
19	28	79	4	8	8	4	2020
20	28	79	4	8	8	4	2020
21	28	79	4	12	8	4	2020
22	28	79	4	12	8	4	2020
23	28	79	4	12	8	4	2020
24	28	79	4	12	8	4	2020

Figura 71. Datos Contaminantes.

8. Finalmente, para comprender la información recogida en la tabla de Excel se debe acceder al siguiente PDF: [Intérprete de ficheros de calidad del aire.](#)

Este proceso debe de repetirse tantas veces como sea necesario para poder descargar todos los años de interés.

11.4 INFORMACIÓN METEOROLÓGICA

Los datos meteorológicos requeridos para poder llevar a cabo el proyecto se han descargado en su totalidad de la Agencia Estatal de Meteorología (AEMET). Para acceder a dicha información se deben seguir los siguientes pasos:

1. Acceder a los datos abiertos de AEMET, en la Figura 72 se muestra el contenido de la página: [AEMET/DatosAbiertos](#)

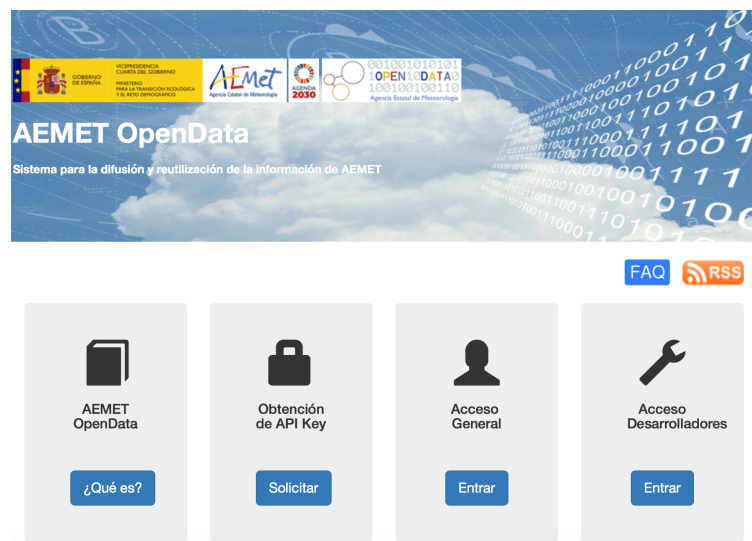


Figura 72. Aemet Open Data [19]

2. Para poder acceder a los datos se debe de tener una API key. Para obtenerla, se debe pinchar en solicitar obtención de API key. Simplemente se requiere introducir su correo electrónico para conseguir la llave que da el acceso a los datos.
3. Una vez obtenida la API key, se acceden a los datos pinchando en acceso general. Los datos requeridos para este proyecto se encuentran en Valores Climatológicos > Climatologías Diarias. El rango máximo de datos que se pueden obtener en una búsqueda es de 5 años por lo que para obtener datos de periodos de tiempo mayores se deben descargar varias tandas de datos.

4. Los datos se encuentran en formato JSON. Debe aparecer el mensaje que se observa en la Figura 73. Cualquier otro mensaje querrá decir que los datos no se han obtenido correctamente.

```
{
  "descripcion" : "exito",
  "estado" : 200,
  "datos" : "https://opendata.aemet.es/opendata/sh/6aac4297",
  "metadatos" : "https://opendata.aemet.es/opendata/sh/b3aa9d28"
}
```

Figura 73. Éxito en la obtención de datos

Se debe seleccionar y copiar el enlace correspondiente a datos y acceder a dicho enlace. Al acceder al enlace correspondiente se le mostrarán los datos en formato JSON.

5. Finalmente, se debe cambiar el formato a Excel. Una forma cómoda de hacerlo es accediendo al siguiente enlace: www.convertcsv.com

Para más información: opendata.aemet.es

11.5 INFORMACIÓN DEL TRÁFICO

Los datos de tráfico se han descargado en su totalidad de la información aportada por el Ayuntamiento de Madrid. Para acceder a dicha información se deben seguir los siguientes pasos:

1. Acceder al siguiente *link*: <https://datos.madrid.es/portal/site/egob>
2. En el buscador se debe buscar: Tráfico. Histórico de datos de tráfico desde 2013
3. Otra opción: Catálogo de datos > Conjuntos de datos > Tráfico. Histórico de datos de tráfico desde 2013
4. En descargas se puede descargar un fichero que contiene los datos en formato CSV. Se tienen los datos por meses desde julio de 2013.

Estos ficheros contienen varios puntos de medida, alrededor de 4500, que recogen varias variables cada 15 minutos (ver Figura 74). En este caso, solo se selecciona la variable

intensidad de tráfico y se debe de realizar la media de los datos de un punto de medida para poder tener un único dato diario por cada punto de medida.

id	fecha	tipo_elem	intensidad	ocupacion	carga	vmed	error	periodo_integracion
1001	1/5/20 0:00	M30	72	NaN	0	59	N	5
1001	1/5/20 0:15	M30	96	NaN	0	58	N	5
1001	1/5/20 0:30	M30	264	1	0	53	N	5
1001	1/5/20 0:45	M30	84	0	0	31	N	5
1001	1/5/20 1:00	M30	12	NaN	0	6	N	5
1001	1/5/20 1:15	M30	36	0	0	34	N	5
1001	1/5/20 1:30	M30	120	0	0	52	N	5
1001	1/5/20 1:45	M30	84	1	0	38	N	5
1001	1/5/20 2:00	M30	0	0	0	0	N	5
1001	1/5/20 2:15	M30	0	0	0	0	N	5
1001	1/5/20 2:30	M30	72	NaN	0	44	N	5
1001	1/5/20 2:45	M30	0	0	0	0	N	5
1001	1/5/20 3:00	M30	24	0	0	26	N	5
1001	1/5/20 3:15	M30	60	NaN	0	60	N	5
1001	1/5/20 3:30	M30	12	NaN	0	14	N	5
1001	1/5/20 3:45	M30	48	NaN	0	54	N	5
1001	1/5/20 4:00	M30	36	0	0	35	N	5
1001	1/5/20 4:15	M30	12	NaN	0	11	N	5
1001	1/5/20 4:30	M30	0	0	0	0	N	5
1001	1/5/20 4:45	M30	24	0	0	26	N	5
1001	1/5/20 5:00	M30	12	NaN	0	10	N	5
1001	1/5/20 5:15	M30	48	0	0	29	N	5

Figura 74. Representación datos Tráfico

Para más información acerca del resto de variables: <https://datos.madrid.es/Trafico>

ANEXO III – LÍMITES CALIDAD DEL AIRE

Existen unos límites legales que no se deben superar y estos se recogen a continuación.

Tabla 17. Límites Legales [8]

<i>Contaminante</i>	<i>Concentración</i>	<i>Período de Promedio</i>	<i>Nº de superaciones toleradas</i>
SO_2	$125 \mu g/m^3$	<i>Diario</i>	3 veces/año
NO_2	$40 \mu g/m^3$	<i>Anual</i>	
PM_{10}	$50 \mu g/m^3$	<i>Diario</i>	35 veces/año
	$40 \mu g/m^3$	<i>Anual</i>	
$PM_{2.5}$	$25 \mu g/m^3$	<i>Anual</i>	
CO	$10 \text{ mg}/m^3$	<i>Medias móviles octohorarias</i>	
O_3	$120 \mu g/m^3$	<i>Medias móviles octohorarias</i>	25 veces/año

Existen unos umbrales de evaluación para la protección de la salud y se recogen en la siguiente tabla.

Tabla 18. Umbrales de Información [8]

<i>Contaminante</i>	<i>Concentración Superior</i>	<i>Concentración Inferior</i>	<i>Período de Promedio</i>	<i>Nº de superaciones toleradas</i>
SO_2	$75 \mu g/m^3$	$50 \mu g/m^3$	Diario	3 veces/año
NO_2	$32 \mu g/m^3$	$26 \mu g/m^3$	Anual	
NO_x	$24 \mu g/m^3$	$19.5 \mu g/m^3$	Anual	
PM_{10}	$35 \mu g/m^3$	$25 \mu g/m^3$	Diario	35 veces/año
	$28 \mu g/m^3$	$20 \mu g/m^3$	Anual	
$PM_{2.5}$	$17 \mu g/m^3$	$12 \mu g/m^3$	Anual	
CO	$7 mg/m^3$	$5 \mu g/m^3$	Medias móviles octohorarias	