Universitat Oberta de Catalunya

uoc.edu

Estudio espacio-temporal de la calidad del aire en la ciudad de Madrid y búsqueda de modelos de predicción

Sergio Romero Córdoba
Máster Universitario de Ciencia de Datos
Área 1

Consultora: Anna Muñoz Bollas
Profesor responsable de la asignatura: Albert Solé Ribalta

05 de junio de 2022

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio espacio-temporal de la calidad del aire en la ciudad de Madrid y búsqueda de modelos de predicción*
Nombre del autor:	*Sergio Romero Córdoba*
Nombre del consultor/a:	*Anna Muñoz Bollas*
Nombre del PRA:	*Albert Solé Ribalta*
Fecha de entrega (mm/aaaa):	06/2022
Titulación::	*Máster Universitario de Ciencia de Datos*
Área del Trabajo Final:	*Área 1*
Idioma del trabajo:	*Español*
Palabras clave	*Calidad del aire, Madrid, SIG*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

La contaminación del aire que respiramos supone un riesgo para la salud humana y para el medio ambiente. Según la Organización Mundial de la Salud, el 99% de la población mundial vive en lugares donde no se respetan sus directrices sobre calidad del aire. Madrid no es una excepción. Es una gran ciudad que soporta un intenso tráfico y con un gran número de hogares cuyo consumo de energía para calefacción, cocinar o iluminación supone una importante fuente de emisión de agentes contaminantes.

En los últimos años se ha tomado más conciencia del problema que esto supone y se han implantado algunas medidas como el aumento de la red de vías ciclistas o la implantación de una Zona de Bajas Emisiones en el centro de la ciudad. Sin embargo, los episodios de alta contaminación siguen siendo habituales.

El portal de datos abiertos del Ayuntamiento de Madrid proporciona datos sobre las mediciones de los agentes contaminantes más importantes desde el año 2001 tomados por las distintas estaciones de control distribuidas por la ciudad. Con estos datos, el siguiente trabajo pretende analizar la evolución de la calidad del aire en la ciudad y estudiar el impacto que determinadas medidas o episodios sucedidos durante estos años han podido tener. Para realizar este estudio se creará un Sistema de Información Geográfica que facilite la comprensión y el análisis de los datos.

Abstract (in English, 250 words or less):

Air pollution is a great risk for human health and the environment. According to the World Health Organization (WHO), 99% of the world population lives in places where the air quality guidelines levels are not met. Madrid is no exception. As a large city, it has heavy road traffic and a big number of inhabitants whose household energy consumption for heating, cooking and lighting is an important source of pollution.

During the last years, people are more aware of this problem and competent institutions have implemented some measures like the creation of increasing the number of bike lanes or the creation of a Low Emission Zone in the center of the city. Nevertheless, atmospheric pollution episodes are still usual.

The City of Madrid open data portal provides information of the measurements of the main pollutants since 2001 taken by control stations distributed along the city. With this data, this work will try to analyze the evolution of air quality in the city and study the impact of certain measures and episodes that may have happened in the last years. To carry out this task, a Geographical Information System will be created to help in the comprehension and analysis of the data.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

La calidad del aire que respiramos en nuestro día a día tiene un impacto fundamental en nuestra salud y en la del medio ambiente. Según estimaciones de 2016 de la Organización Mundial de la Salud (OMS), la contaminación atmosférica provoca cada año 4,2 millones de muertes prematuras en el mundo [1].

La contaminación atmosférica no es un fenómeno nuevo, pero su impacto se disparó a partir de la Revolución Industrial. Aunque hay algunos procesos naturales que generan emisiones gaseosas contaminantes de la atmósfera, la actividad humana es en gran medida la principal causante del problema. En las últimas décadas en los países desarrollados se ha avanzado en una legislación que permita controlarla y reducirla, pero, lo cierto es que se siguen observando episodios de alta contaminación de forma constante cada año.

El problema afecta tanto a áreas rurales como a núcleos urbanos. Pese a no ser exclusivo de las grandes ciudades, las características de éstas, con un gran número de habitantes y un intenso tráfico que las recorre, hacen que sea una de las principales preocupaciones de sus ciudadanos.

El siguiente trabajo se centra en la ciudad de Madrid y pretende realizar un análisis de la calidad del aire en esta ciudad en los últimos años. El Ayuntamiento de esta ciudad ha llevado a cabo actividades de control y vigilancia de la calidad del aire desde el año 1968 aunque es a partir del año 1998 cuando se crea un Sistema Integral de Vigilancia, Predicción e Información [2]. Este Sistema Integral cuenta con 2 unidades móviles, 24 estaciones fijas de vigilancia de la calidad del aire, 83 sensores y equipos meteorológicos y 95 analizadores de gases y partículas. La ciudad cuenta con protocolos de actuación para episodios de contaminación [3] y desde noviembre de 2018 tiene una Zona de Bajas Emisiones (ZBE) con una extensión de 472 hectáreas en donde la circulación de tráfico está restringida. Sin embargo, estos esfuerzos son insuficientes. La última memoria disponible de calidad del aire de Madrid, correspondiente al año 2020 [4] indica que, pese a la mejoría sustancial respecto a años anteriores, los niveles límite y objetivo establecidos para el Dióxido de Nitrógeno (NO_2) y el Ozono (O_3), se han continuado excediendo. Esta mejoría además tuvo mucho que ver con la situación de pandemia de la COVID-19 que limitó la movilidad y las actividades en la ciudad desde el 15 de marzo hasta el 21 de junio.

1.1.1 Entendiendo el problema. ¿Qué es la contaminación atmosférica?

La contaminación atmosférica es la presencia que existe en el aire de pequeñas partículas o productos secundarios gaseosos que pueden implicar riesgo, daño o molestia para las personas, plantas y animales que se encuentran expuestas a dicho ambiente [5].

El primer paso en este proceso de contaminación es la emisión de gases contaminantes a la atmósfera [6]. Estas emisiones proceden de dos tipos de fuentes:

- Fuentes difusas. Son aquellas que no tienen ningún foco ni localización geográfica concreta. Ejemplos de estas fuentes serían el transporte o la agricultura.
- Fuentes fijas. Se encuentran bien localizadas y es fácil medir sus niveles de emisión. Una fábrica o una central eléctrica serían fuentes de este tipo.

Las emisiones liberadas en una localización específica pueden causar la contaminación del aire en dicha localización. Sin embargo, también es posible que estos contaminantes viajen grandes distancias, sean dispersados o transformados en otras especies químicas.

Finalmente, el depósito atmosférico es el proceso por el cual estos contaminantes son transportados hacia la superficie, provocando efectos nocivos en seres humanos, ecosistema y materiales. Existen dos modalidades de depósito:

- Depósito húmedo. Los contaminantes son transportados a la superficie a través de fenómenos meteorológicos como la lluvia, la nieve o la niebla.
- Depósito seco. La superficie captura directamente los contaminantes mediante impacto, sedimentación o difusión.

En la siguiente figura se muestra gráficamente el proceso completo de contaminación.

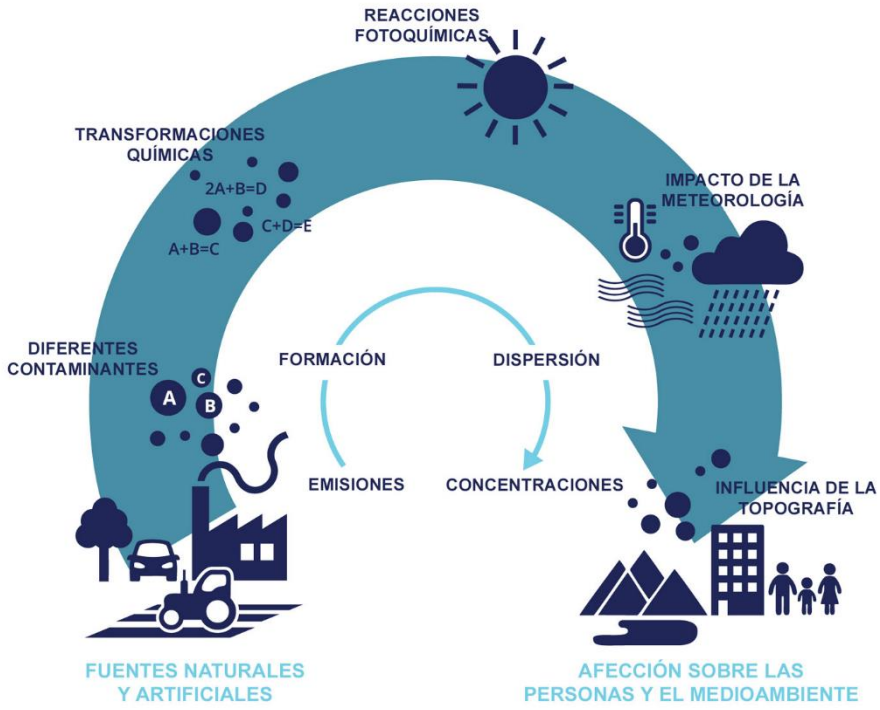


Figura 1. Contaminación: De emisiones a exposición
Fuente: Agencia Europea de Medio Ambiente

1.1.2 Principales actividades contaminantes

La Agencia Europea del Medio Ambiente señala [7] que Las partículas contaminantes (PC), el dióxido de nitrógeno (NO_2) y el ozono troposférico (O_3) son los contaminantes que provocan mayores daños a la salud humana y el medio ambiente en Europa y que las principales fuentes de estos contaminantes, como muestra la figura 2, son el transporte por carretera, las calefacciones domésticas, la agricultura y la industria.

Figura 2. Fuentes de contaminación atmosférica en Europa
Fuente. Agencia Europea de Medio Ambiente

1. Agricultura.

Las prácticas agrícolas no sostenibles son un problema para la pérdida de biodiversidad y la degradación del ecosistema.

Más del 90% de las emisiones de amoniaco de Europa, cerca del 80% de las emisiones de metano y cerca del 20% de las emisiones de compuestos orgánicos volátiles distintos de metano (COVDM), como benceno y etanol, provienen de actividades agrícolas.

2. Industria.

La producción y distribución de energía son la principal fuente de emisiones de óxidos de azufre (SO_x) y una de las principales de óxidos de nitrógeno (NO_x). Alrededor del 60% de las emisiones de óxidos de azufre provienen de la producción de energía eléctrica.

3. Fuentes naturales.

Los fenómenos naturales también contribuyen a la contaminación atmosférica. Entre ellos están las emisiones volcánicas y marinas, tormentas de arena, incendios forestales o la actividad de los seres vivos (como por ejemplo la digestión de los rumiantes).

4. Tratamiento de residuos.

Una mala gestión de los residuos que producimos puede contribuir a la contaminación del aire. Estos residuos son una importante fuente de metano.

5. Transporte por carretera.

El transporte por carretera es responsable de alrededor del 45% de las emisiones de óxidos de nitrógeno (NO_x) y es una importante fuente de emisión de otros contaminantes fundamentales.

Además, es la fuente de ruido ambiental más habitual que afecta a más de 100 millones de personas en Europa.

6. Calefacción.

La calefacción, ya sea en hogares, negocios o edificios públicos, es una importante fuente de contaminación. Es responsables del 53 % de emisiones de partículas finas ($PM_{2.5}$) y contribuye fundamentalmente en las emisiones de monóxido de carbono (CO)

1.1.3 Principales agentes contaminantes

A continuación, se describen los contaminantes con mayor presencia en la atmósfera.

Material particulado (PM)

Son sustancias en estado sólido o líquido que se encuentran en suspensión en la atmósfera o que se depositan de forma muy lenta en la superficie. Son un conjunto de contaminantes muy variados con características físicas y químicas muy diversas. Las fuentes de emisión son también múltiples, siendo de hecho un agente contaminante que se puede considerar tanto primario (se emiten directamente a la atmósfera) como secundario (se genera a partir de reacciones químicas en la atmósfera).

En general se suelen clasificar en función de su diámetro en material particulado fino y grueso.

El material particulado fino ($PM_{2.5}$) es aquel cuyo diámetro es inferior a 2,5 μm. Estas partículas son las más nocivas para la salud humana ya que son capaces de

atravesar la barrera pulmonar y entrar en el sistema sanguíneo. Pueden producir enfermedades cardiovasculares y respiratorias, así como cáncer de pulmón.

Debido a su reducido tamaño pueden permanecer en la atmósfera suspendidas durante mucho tiempo y pueden viajar transportadas grandes distancias, hasta miles de kilómetros de su origen.

El material particulado grueso (PM_{10}) son las partículas con diámetro inferior a 10μm. La mayoría de estas partículas se originan por erosión de la superficie terrestre o forman parte de las emisiones marinas.

Ozono (O_3)

La troposfera es la capa más baja de la atmósfera. El ozono presente en esta capa supone un riesgo para la salud pudiendo causar problemas respiratorios, provocar asma, reducir la función pulmonar y dar lugar a enfermedades pulmonares. Es un contaminante secundario, es decir, generado por reacciones químicas a partir de otros contaminantes.

Dióxido de nitrógeno (NO_2)

El dióxido de nitrógeno es un contaminante primario que a su vez puede dar lugar a otros contaminantes secundarios como el ozono o el ácido nítrico. El transporte por carretera es responsable del 45% del dióxido de nitrógeno, siendo otras fuentes de emisión la calefacción o la generación de electricidad.

Entre los efectos adversos para la salud se encuentran el aumento de los síntomas de bronquitis en niños asmáticos o la disminución del desarrollo de la función pulmonar.

Dióxido de azufre (SO_2)

El dióxido de azufre es un gas de olor desagradable que se genera principalmente por la quema de combustibles fósiles con contenido de azufre que se usan para la generación de energía eléctrica, la calefacción doméstica y los vehículos a motor. Otras fuentes naturales de dióxido de azufre son las erupciones volcánicas y las emisiones procedentes de los océanos.

Puede afectar al sistema respiratorio y a las funciones pulmonares, así como causar irritación pulmonar.

Otros agentes contaminantes que se pueden encontrar en el aire son otros compuestos nitrogenados (NO, N_2O y NH_3), monóxido de carbono (CO), material particulado ultrafino (diámetro inferior a 1nm), metano (CH_4), compuestos orgánicos volátiles como el benceno y compuestos orgánicos persistentes como las dioxinas.

1.1.4 Recomendaciones de la OMS e Índice de Calidad del Aire (ICA)

Las directrices de la OMS [1] sobre la calidad del aire ofrecen orientaciones a escala mundial sobre los umbrales y límites de los contaminantes atmosféricos clave que entrañan riesgos para la salud (AGQ level, de sus siglas en inglés, Air Quality Guideline). El objetivo de las directrices es que todos los países alcancen los niveles de calidad del aire recomendados. Sin embargo, también proponen metas intermedias (Interim target) para aquellos países o regiones en los que estos objetivos son complicados de alcanzar a día de hoy. Estas metas intermedias sirven como objetivos más realistas que se pueden plantear para ir consiguiendo logros graduales que signifiquen un beneficio para la salud de sus ciudadanos.

Las directrices actuales corresponden al año 2021 y suponen una actualización de las anteriores del año 2005.

La siguiente tabla muestra los valores recomendados (AGQ level), así como los límites intermedios (Interim target) para los cinco contaminantes principales descritos en el punto anterior añadiendo además del Monóxido de Carbono (CO).

Pollutant	Averaging time	Interim target				AQG level
		1	2	3	4	
$PM_{2.5}$, µg/m3	Annual	35	25	15	10	5
	24-hour[a]	75	50	37.5	25	15
PM_{10}, µg/m3	Annual	70	50	30	20	15
	24-hour [a]	150	100	75	50	45
O_3, µg/m3	Peak season[b]	100	70	-	-	60
	8-hour	160	120	-	-	100
NO_2, µg/m3	Annual	40	30	20	-	10
	24-hour [a]	120	50	-	-	25
SO_2, µg/m3	24-hour [a]	125	50	-	-	40
CO, mg/m3	24-hour [a]	7	-	-	-	4

a 99th percentile (i.e. 3-4 exceedance days per year). Por ejemplo, registrar valores medios diarios (24 horas) por encima de 25 de µg/m^3 de NO_2 durante más de 4 días al año, supondría incumplir estas directrices (por encima de 50 si el país tiene como objetivo el límite intermedio 2 o por encima de 120 si tiene por objetivo el límite intermedio 1)
b Average of daily maximum 8-hour mean O_3 concentration in the six consecutive months with the highest six-month running-average O_3 concentration. Es decir, la ventana de los 6 meses con mayor media de concentración de O_3 será la peak-season. La media del máximo valor medio diario de un periodo de 8 horas durante esos 6 meses no debería superar los 60 µg/m^3.

Tabla 1. Resumen de valores recomendados a largo y corto plazo y límites intermedios
Fuente. Directrices de la OMS sobre la calidad del aire 2021.

Con motivo de esta revisión por parte de la OMS, la Comisión Europea lanzó una consulta pública [8] para recabar las opiniones de los ciudadanos y de las partes interesadas sobre la revisión de las Directivas sobre la calidad del aire ambiente (2008/50/CE y 2004/107/CE). La iniciativa tiene por objeto presentar una

propuesta legislativa de revisión de estas Directivas cuyos principales elementos serían: 1) adaptar más las normas de calidad del aire de la UE a las recomendaciones de la Organización Mundial de la Salud (actualizadas en 2021); 2) seguir mejorando la seguridad jurídica y la aplicabilidad del marco legislativo, sobre todo las disposiciones sobre información pública, sanciones y acceso a vías de resarcimiento efectivas; y 3) reforzar el seguimiento, la modelización y la planificación de la calidad del aire.

Además de estos niveles de calidad, es común que los organismos competentes tengan un Índice de Calidad del Aire (ICA, también conocido como AQI por sus siglas en inglés). Este índice es un indicador que sirve para informar de la calidad del aire a la población de una manera clara y sencilla. Indica el grado de pureza o contaminación atmosférica del aire, y los efectos para la salud asociados [9]. Para el proyecto se cogerá como referencia el estándar CAQI (Índice Común de Calidad del Aire) desarrollado en el marco del proyecto CITEAIR cofinanciado por la Unión Europea. Para cada uno de los contaminantes se establece un índice parcial, de forma que el peor valor de los cinco define el índice global. Existe una versión horaria y una versión diaria. El índice horario se calcula a partir de la concentración horaria registrada ($\mu g/m^3$) mientras que el índice diario se calcula a partir del valor horario más alto alcanzado durante las 24 horas de cada jornada. Los niveles establecidos se muestran en la siguiente tabla:

Contaminantes	Muy bueno	Bueno	Regular	Malo	Muy malo
Partículas PM2,5	0-15	16-30	31-55	56-110	>110
Partículas PM10	0-25	26-50	51-90	91-180	>180
Dióxido de Nitrógeno (NO_2)	0-50	51-100	101-200	201-400	>400
Ozono (O_3)	0-60	61-120	121-180	181-240	>240
Dióxido de Azufre (SO_2)	0-50	51-100	101-350	351-500	>500

Tabla 2. Índice de Calidad del Aire CAQI
Fuente. Unión Europea [9]

1.1.5 Protocolos alta contaminación de Madrid

Actualmente existen dos protocolos de actuación por episodios de alta contaminación en Madrid dependiendo del contaminante que lo cause: dióxido de nitrógeno y ozono troposférico.

Protocolo de actuación para episodios de contaminación por NO_2 [10]

El actual protocolo entró en vigor el 10 de diciembre de 2018 y realiza una división de la ciudad en 5 zonas estableciendo 3 niveles de actuación en función de las concentraciones de dióxido de nitrógeno que se registren o se prevean registrar en las zonas que se han definido. Estos niveles son:

PREAVISO: cuando dos estaciones cualesquiera de una misma zona superan, simultáneamente, 180 µg/m3 durante dos horas consecutivas, o tres estaciones cualesquiera de la red de vigilancia superan, simultáneamente, 180 µg/m3 durante tres horas consecutivas.
AVISO: cuando dos estaciones cualesquiera de una misma zona superan, simultáneamente, 200 µg/m3 durante dos horas consecutivas, o tres estaciones cualesquiera de la red de vigilancia superan, simultáneamente, 200 µg/m3 durante tres horas consecutivas.
ALERTA: cuando tres estaciones cualesquiera de una misma zona (o dos si se trata de la zona 4) superan, simultáneamente, 400 µg/m3 durante tres horas consecutivas.

Hay 5 escenarios previstos que se activarán cuando se alcancen alguno de los niveles anteriormente citados:

ESCENARIO 1: 1 día con superación del nivel de preaviso.
ESCENARIO 2: 2 días consecutivos con superación del nivel de preaviso o 1 día con superación del nivel de aviso.
ESCENARIO 3: 3 días consecutivos con superación del nivel de preaviso o 2 días consecutivos con superación del nivel de aviso.
ESCENARIO 4: 4 días consecutivos con superación del nivel de aviso.
ESCENARIO 5: 1 día de nivel de alerta.

Las medidas se van acumulando en los distintos escenarios siendo el escenario 1 el menos restrictivo y el 5 el más restrictivo.

La contaminación por NO_2 en Madrid es debido sobre todo al tráfico, por lo que las medidas van encaminadas a reducir el número de vehículos a motor en circulación.

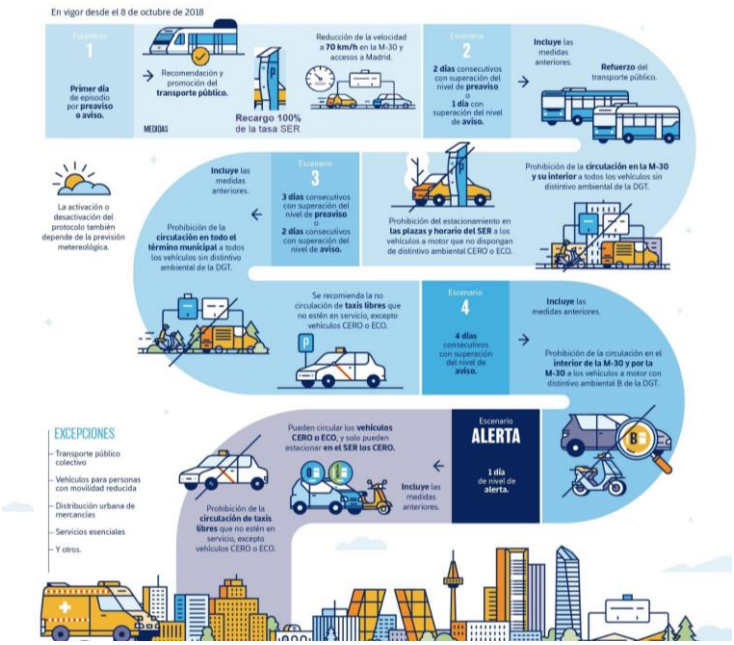


Figura 3. Protocolo por alta contaminación
Fuente. Ayuntamiento de Madrid

La figura 3 muestra el diagrama de actuación de este protocolo.

Protocolo de actuación para episodios de contaminación por Ozono [11]

Este protocolo, a diferencia del anterior, no define de forma tan clara escenarios con restricciones, sino que es más bien informativo, aunque sí que establece la coordinación de actuaciones mediante la activación del Sistema de Alertas en Salud Ambiental.

Se establecen tres niveles de actuación en función de las concentraciones registradas por las estaciones de la red de vigilancia de la calidad del aire:

PREAVISO (de carácter interno): 160 µg/m3. Contempla la verificación de mecanismos en caso de llegar a un nivel superior.
UMBRAL DE INFORMACIÓN: 180 µg/m3. Activa todos los canales de información disponibles: página web del Ayuntamiento, mensajes SMS a los usuarios dados de alta en el servicio, notificación y mensajes en la aplicación para móviles "Aire de Madrid" y publicación en la cuenta de Twitter @airedemadrid.
UMBRAL DE ALERTA: 240 µg/m3. Se convocará a la Comisión de Calidad del aire de la ciudad de Madrid.

En todos los casos se trata de valores medios horarios por estación.

Los informes de los episodios de alta contaminación en los que se ha activado alguno de estos protocolos se pueden consultar en el portal web de calidad del aire del Ayuntamiento de Madrid: https://www.mambiente.madrid.es/opencms/calaire/Episodios/listados_informes_episodios/

1.2 Motivación

Madrid es una ciudad en la que se repiten episodios de alta contaminación desde hace muchos años. El tráfico intenso que padece la ciudad es el principal causante de emisiones y, aunque en los últimos años se está tomando mayor conciencia del problema, tanto a nivel social como a nivel institucional, no es raro encontrarnos con imágenes como la siguiente que se muestra a la izquierda.

Figura 5. Contaminación en Madrid
Fuente. Telemadrid.es

Figura 4. Madrid sin contaminación
Fuente. Instagram. Usuario josel.fotografia

Nos hemos acostumbrado a convivir con ello cuando la realidad es que otras imágenes como la de la derecha, nos muestran como se ve la ciudad los días en los que no hay tanta contaminación atmosférica.

Como habitante de esta ciudad durante muchos años, me resulta interesante analizar los diferentes factores que influyen en la calidad del aire que respiramos a diario. A lo largo del Máster de Ciencia de Datos he podido ver la gran cantidad de información recogida en relación a agentes contaminantes desde hace ya bastantes años. Ha resultado ser una sorpresa agradable el ver que las instituciones tienen mecanismos para la medición y análisis de emisiones y contaminantes que afectan a la calidad del aire, pese a que llevar a cabo medidas para la reducción de la contaminación sea complicado, en muchos casos por la oposición de los propios ciudadanos.

Además, la situación de pandemia de la COVID-19 produjo en el año 2020 durante unos meses una situación que nos dio la oportunidad de comprobar la mejora que teníamos en temas de contaminación atmosférica al reducir drásticamente la movilidad y la actividad en la ciudad.

Mi interés es poder llegar a comprender mediante el análisis y visualización de estos datos un problema que es complejo ya que, obviamente, parar toda actividad de la ciudad no es una solución.

1.3 Objetivos del Trabajo

El trabajo pretende realizar un análisis y visualización de los datos de calidad del aire en la ciudad de Madrid entre los años 2001, primer año para el que se dispone de datos, hasta la actualidad. Se analizará la evolución de los principales agentes contaminantes, así como la posible relación que puedan tener con factores climatológicos y con la densidad del tráfico.

Un aspecto a analizar en detalle es el periodo entre el 15 de marzo y el 21 de junio de 2020. En este periodo, debido a la situación sanitaria causada por la pandemia de la COVID-19, el Gobierno de España decretó el estado de alarma limitando al mínimo la movilidad. Este hecho debería verse reflejado en la calidad del aire de una ciudad como Madrid, en la que una de las principales causas de contaminación es el tráfico rodado.

También se analizarán las medidas que se están tomando en los últimos años para intentar reducir las emisiones y mejorar la calidad del aire. En concreto, se estudiará el impacto que ha tenido la creación de una Zona de Bajas Emisiones en el centro de la ciudad.

Por último, se intentarán buscar patrones que causen episodios de alta contaminación en la ciudad.

1.3.1 Hipótesis (Objetivo principal)

Comprobar la relación que existe entre el tráfico y la calidad del aire y la mejoría que se obtiene con la aplicación de medidas restrictivas como la creación de una zona de bajas emisiones.

1.3.2 Objetivos parciales (Preguntas de investigación)

Para llegar a confirmar la hipótesis planteada anteriormente, se definen a continuación una serie de objetivos o preguntas de investigación. Se van a distinguir los objetivos principales, derivados de la realización del Trabajo de Fin de Máster y los objetivos secundarios, más específicos para el proyecto que se pretende abordar.

Objetivos principales

1) Analizar el estado del arte en relación a investigaciones sobre calidad del aire, no sólo en la ciudad de Madrid sino para cualquier región.
2) Adquisición y compresión de los datos necesarios para la elaboración del proyecto. Realizar la limpieza y preparación de los datos necesaria para su explotación.
3) Redacción de la memoria que documente el TFM.
4) Realizar una presentación creativa e innovadora orientada a la difusión del proyecto.
5) Defender el TFM ante un tribunal evaluador.

Objetivos secundarios

1) Obtención y comprensión de fuentes de datos necesarios para el análisis de calidad del aire en Madrid.
2) Limpieza y preparación del juego de datos que se utiliza para el desarrollo del proyecto utilizando el lenguaje de programación R.
3) Análisis de los datos utilizando el lenguaje de programación R.
4) Crear una visualización interactiva que permita el análisis temporal de los datos utilizando herramientas SIG.
5) Realizar de nuevo un análisis de los datos utilizando la visualización creada.

1.4 Enfoque y método seguido

El producto final que se quiere obtener es una visualización interactiva que nos facilite la comprensión y el análisis de los datos de calidad del aire de la ciudad de Madrid durante los últimos años. Además, como se ha comentado en los objetivos del proyecto, esta visualización incorporará datos meteorológicos y de tráfico para ver su relación con los distintos componentes contaminantes de la atmósfera.

Muchas instituciones públicas han desarrollado mapas que permiten ver los datos de calidad del aire en una determinada ciudad. Muchos de ellos permiten la visualización en tiempo real, de las últimas 24 horas e incluso en algunos casos permiten visualizar una previsión de las próximas horas. No existe sin embargo un visualizador que permita ver estos datos desde el origen de los registros y que

además los ponga en relación con los otros datos que pretende analizar este proyecto.

El portal de datos abiertos del Ayuntamiento de Madrid va a ser el principal origen de datos. Este portal cuenta con los siguientes datos en relación a la calidad del aire:

- Datos por franjas horarias. Años 2001 a 2021.
- Datos diarios. Años 2001 a 2021.
- Estaciones de control.

Y los siguientes datos meteorológicos:

- Datos por franjas horarias desde 2019.
- Datos diarios desde 2019.
- Estaciones de control.

Para completar los datos meteorológicos anteriores a 2019 se podrán consultar otras fuentes como la Agencia Española de Meteorología (AEMET).

Se va a utilizar la siguiente propuesta metodológica de desarrollo de un GIS [12] que consta de 5 pasos (figura 6).

- Planificación inicial.
- El estudio de viabilidad.
- Análisis y diseño detallados.
- Implementación.
- Mantenimiento y revisión.

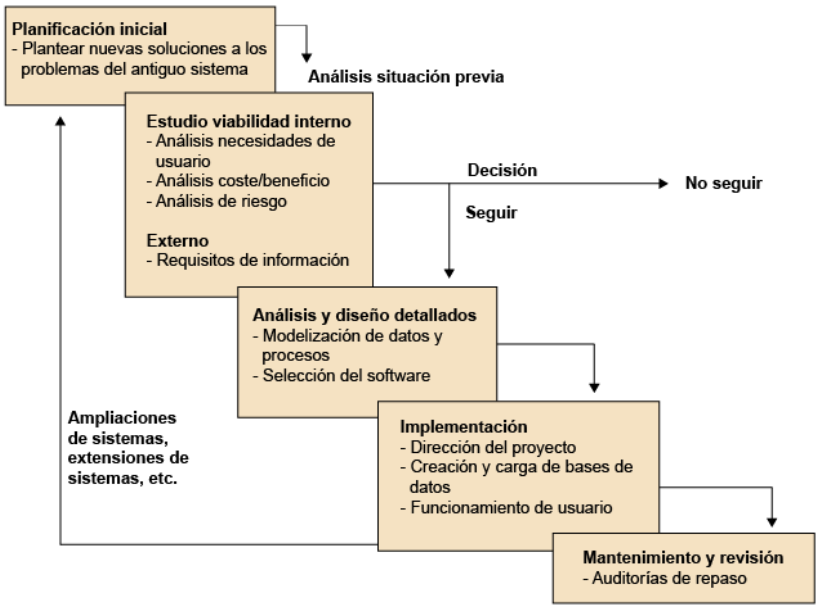


Figura 6. Metodología SIG
Fuente. J. Rodríguez Lloret; R. Olivella; A. Muñoz Bollas; V. Velarde Gutiérrez. Introducción a los sistemas de información geográfica

1.5 Planificación del Trabajo

Durante el proyecto se utilizarán los siguientes recursos:

Software

- R. Lenguaje de programación. Se utilizará en distintas fases: limpieza y preparación de los datos, análisis estadístico y modelos predictivos.
- R Studio. Entorno de desarrollo integrado para R.
- PostgreSQL v14. Almacenamiento de datos del proyecto. Se crearán las vistas necesarias para la creación de capas en QGIS.
- PostGIS. Módulo para bases de datos espaciales en PostgreSQL.
- pgAdmin 4. Entorno para la gestión y administración en PostgreSQL.
- QGIS v3. Sistema de Información Geográfica de código abierto. Análisis espacial.
- Leaflet y D3. Librerías JavaScript utilizadas para la creación de un mapa web para consulta de datos de calidad del aire.
- GitHub. Almacenamiento y gestión de versiones.
- Word. Redacción de memoria.
- PowerPoint. Creación de presentación

Hardware

- Portátil Intel(R) Core(TM) i7-3517U CPU @ 1.90GHz 1.90 GHz y 8.00 GB de RAM con Sistema Operativo Windows 10.

A continuación, se detallan las 6 fases en las que se divide el proyecto con las actividades e hitos de cada una de ellas.

Fase 1. Definición y planificación del trabajo final

- Investigación del proyecto
- Poner en contexto la temática del trabajo
- Justificar la elección y describir la motivación personal
- Definir el objetivo principal
- Definir los objetivos parciales
- Definir la metodología de trabajo que se va a utilizar
- Planificar el proyecto mediante un diagrama de Gantt

Fase 2. Estado del arte

- Buscar trabajos relacionados con la visualización y análisis de la calidad del aire
- Describir los éxitos conseguidos hasta la fecha
- Investigar visualizaciones similares

Fase 3. Diseño e implementación del trabajo

- Obtener datos principales del proyecto (datos de calidad del aire de Madrid)
- Buscar otros datos que puedan ser útiles en el análisis (meteorológicos, de tráfico…)
- Obtener datos cartográficos de Madrid por distritos
- Preparar los datos: transformación de valores, discretización, reducción de dimensionalidad y tratamiento de datos vacíos
- Realizar un análisis estadístico de datos y buscar patrones de contaminación

- Crear del modelo de datos en PostgreSQL
- Cargar los datos en la base de datos
- Crear un mapa en QGIS con las capas necesarias para realizar el análisis espacial
- Analizar los datos a través de la visualización
- Crear una visualización interactiva con Leaflet
- Validar el objetivo principal del trabajo
- Creación de modelos de predicción de ICA basado en datos meteorológicos

Fase 4. Redacción de la memoria
- Explicar los resultados obtenidos
- Redactar las conclusiones
- Explicar líneas de trabajo futuras

Fase 5. Preparación de la defensa
- Identificar puntos clave del proyecto
- Crear presentación que resuma los aspectos fundamentales
- Crear un vídeo para la presentación elaborada

Fase 6. Defensa del proyecto
- Preparar la defensa.
- Defender el proyecto.

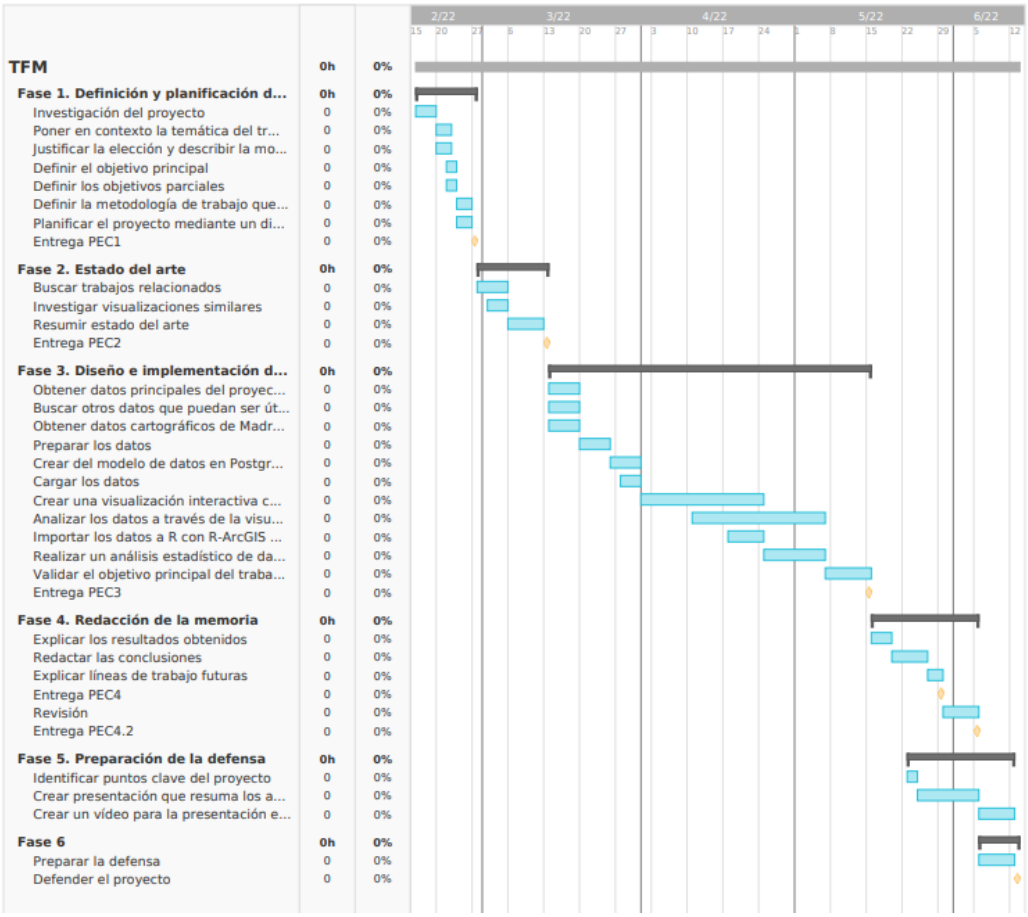


Figura 7. Diagrama de Gantt
Fuente. Elaboración propia

La figura 7 muestra el diagrama de Gantt de la planificación.

1.6 Breve sumario de productos obtenidos

Se listan a continuación los principales productos obtenidos durante el proyecto:

- Tres ficheros de R Markdown. El primer fichero, Aire.rmd, contiene la carga, limpieza, preparación y análisis estadístico de los datos horarios de calidad del aire de Madrid desde el año 2001. El segundo fichero, Meteo.rmd, contiene la carga, limpieza, preparación y análisis estadístico de los datos horarios meteorológicos de Madrid desde el año 2019. Por último, el fichero Modelos.rmd contiene los modelos de predicción con distintos hiperparámetros para la predicción del índice de calidad del aire en función de datos meteorológicos.

- Un proyecto QGIS con el mapa que contiene distintas capas para la consulta de datos de calidad del aire de Madrid desde el año 2001. En este proyecto también se han generado una serie de capas por interpolación IDW de los últimos cuatro años ara distintos agentes contaminantes.

- Un mapa web accesible a través de Internet desarrollado con Leaflet. Este mapa permite consultar datos de calidad del aire así como las distintas capas de interpolación creadas en QGIS.

- Un conjunto de gráficas con el análisis de los datos de calidad del aire y datos meteorológicos.

1.7 Descripción del resto de capítulos de la memoria

En los próximos capítulos de esta memoria se hablará sobre el estado del arte de tres líneas de investigación:

- Impacto de la calidad del aire en la salud
- Sistemas de información geográfica para la visualización de la calidad del aire
- Modelos de calidad de aire usando técnicas de aprendizaje automático

A continuación, se hablará del diseño e implementación del proyecto, detallando el proceso de búsqueda de datos, su preparación y análisis y la generación de modelos predictivos así como de una visualización interactiva.

Por último, se comentará las conclusiones a las que se ha llegado tras la realización del proyecto.

2. Estado del arte

Como se comentó en el capítulo anterior, este trabajo se va a centrar en el estudio y análisis de la calidad del aire en la ciudad de Madrid utilizando herramientas SIG para la visualización de los distintos agentes contaminantes y técnicas de aprendizaje automático para la creación de modelos predictivos. Las líneas de investigación que cubren el proyecto son las siguientes:

- Impacto de la calidad del aire en la salud
- Sistemas de información geográfica para la visualización de la calidad del aire
- Modelos de calidad de aire usando técnicas de aprendizaje automático

A continuación se muestra en detalle el estado del arte de cada uno de estos puntos que se utilizará como punto de partida para el trabajo a realizar.

2.1 Impacto de la calidad del aire en la salud

Aunque no es objeto de desarrollo de este trabajo investigar sobre el impacto que la contaminación atmosférica tiene sobre la salud humana, sí que es una parte importante del mismo pues es un aspecto que ha contribuido en la elección del proyecto. Por este motivo, a continuación se exponen brevemente las investigaciones realizadas a este respecto así como la legislación vigente que afecta a la ciudad de Madrid.

En la década de los años 1980, la asociación de problemas de salud con la contaminación atmosférica ya era un hecho que contaba con suficiente respaldo científico y es a finales de esta década, en 1987, cuando la OMS publica por primera vez unas directrices con unos valores límite para 28 partículas contaminantes [13]. Estas directrices, únicamente publicadas para la región de Europa, no tardaron en ser actualizadas. En los siguientes años, se realizan un gran número de estudios epidemiológicos y se acumulan las pruebas de que los efectos adversos en la salud se producen a concentraciones por debajo de las establecidas por la OMS [14], lo que provoca una revisión de estos valores y una actualización publicada en el año 2000.

A principios del siglo XXI continúan las investigaciones sobre las implicaciones que tienen los agentes contaminantes que respiramos en nuestra salud y estudios como los de Brunekreef y Holgate [15] en el año 2002 o Pope [16] en el 2000 de nuevo vuelven a sugerir que los valores límite establecidos siguen siendo altos y concluyen que estas investigaciones necesitaban recibir un mayor interés tanto regulatorio como científico. Ante estas nuevas evidencias, la OMS vuelve a actualizar sus directrices en el año 2005, ya a nivel mundial, y en esta ocasión las recomendaciones se centran en los cinco agentes contaminantes más perjudiciales: material particulado (fino y grueso), ozono, dióxido de nitrógeno y dióxido de azufre.

Las directrices que están en vigor actualmente corresponden al año 2001, última actualización que se ha llevado a cabo. Estas nuevas directrices suponen de nuevo una revisión a la baja de los valores límite y están desarrolladas según los siguientes pasos principales:

- Formulación del alcance y preguntas clave de la guía
- Revisión de la evidencia relevante
- Evaluación del grado de veracidad de la evidencia resultante del paso anterior
- Formulación de niveles de calidad del aire
- Formulación de otro tipo de guía auxiliar

La tabla 1, mostrada en el capítulo 1.1.4, resume los valores establecidos en estas directrices.

En la actualidad, es un hecho aceptado que la contaminación atmosférica conlleva problemas importantes para la salud. La siguiente imagen resume los problemas que se asocian a las distintas partículas contaminantes que respiramos.

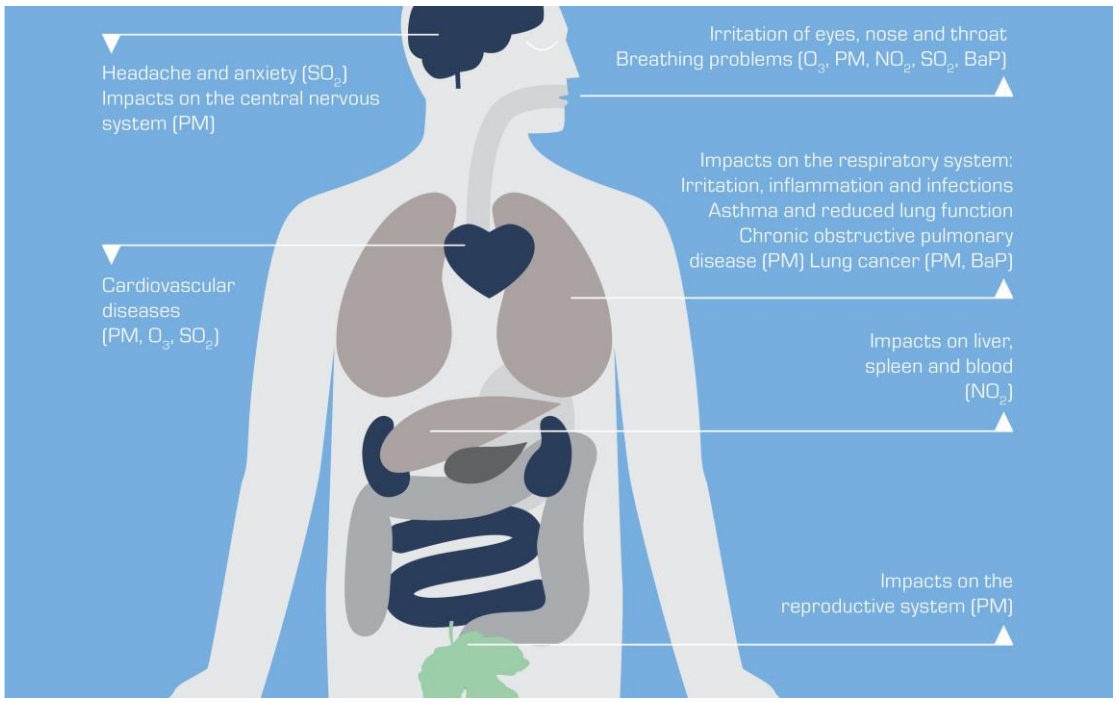


Figura 8. Principales fuentes de contaminación atmosférica y problemas potenciales en la salud humana
Fuente. Agencia Europea de Medio Ambiente

A nivel internacional existen una serie de convenios en relación a la contaminación atmosférica [17]:

- Convenio de Ginebra de 1979 sobre contaminación atmosférica transfronteriza a gran distancia
- Convenio de Viena de 1985 para la Protección de la Capa de Ozono
- Convención Marco de Naciones Unidas sobre el Cambio Climático de 1992

- Convenio de Estocolmo de 2001 sobre Contaminantes Orgánicos Persistentes (COP, POP en inglés)
- Convenio de Minamata de 2013 sobre Mercurio

A nivel europeo, la legislación actual existente se basa en tres directivas [18]:

- Directiva 2008/50/EC sobre calidad del aire ambiente y un aire limpio para Europa.
- Directiva 2004/107/EC del Parlamento Europeo y el Consejo relativa a arsénico, cadmio, mercurio e hidrocarburos aromáticos policlínicos en el medio ambiente.
- Directiva 2015/1480/EC que modifica algunos anexos de las directivas anteriores.

Por último, en España la legislación en vigor viene representada por las siguientes normas [17]:

- Ley 34/2007, de 15 de noviembre, de calidad del aire y protección de la atmósfera
- Real Decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire

2.2 Sistemas de Información geográfica para la visualización de la calidad del aire

Desde su creación en el año 1962 por el geógrafo británico Roger Tomlinson, los Sistemas de Información Geográfica (SIG) han evolucionado enormemente de la mano de los avances tecnológicos que permitían el desarrollo de herramientas abiertas a un gran número de usuarios. Los sistemas, que en un primer momento estaban compuestos por aplicaciones de escritorio complejas únicamente utilizadas por unos pocos usuarios, pasaron a popularizarse con el desarrollo de Internet. Los mapas podían ser distribuidos fácilmente y ser accesibles para cualquier persona que tuviera acceso a la web. Así, teníamos mapas en muchos casos estáticos o que permitían interacciones limitadas. Pero a partir de mediados de la década de los 2000 se lleva a cabo un paso más en esta evolución; los usuarios ya no sólo son consumidores de la información, sino que también aportan datos geoespaciales. Surgen proyectos como OpenStreetMap o Google Maps.

Hoy en día, miles de organizaciones en prácticamente cualquier campo utilizan SIG para identificar problemas, monitorizar cambios, gestionar y responder a eventos, hacer predicciones, establecer prioridades o entender tendencias [19].

El análisis de la calidad del aire es uno de los campos que se han beneficiado de estas herramientas. Durante muchos años la evaluación de calidad del aire se hacía en base a puntos de medición y todavía en la actualidad la mayoría de Organismos Públicos con algún tipo de competencias ofrecen mapas interactivos a los ciudadanos desarrollados con algún tipo de herramienta SIG que permiten consultar la calidad del aire en tiempo real en determinados puntos. La información de las

estaciones de control de Madrid, por ejemplo, la podemos encontrar en la Agencia Europea de Medio Ambiente [20], el Ministerio para la Transición Ecológica y el Reto Demográfico [21], la Comunidad de Madrid [22] o el Ayuntamiento de Madrid [23]. Todos estos mapas ofrecen una única capa de visualización con la información, con mayor o menor detalle, de la calidad del aire recogida en las distintas estaciones de control. Estas visualizaciones son útiles a nivel informativo para los ciudadanos que quieran conocer el estado atmosférico en el momento, sin embargo, no permiten un análisis con otras capas de información como podría ser datos meteorológicos o de tráfico. La siguiente figura muestra el mapa que ofrece la Agencia Europea de Medio Ambiente.

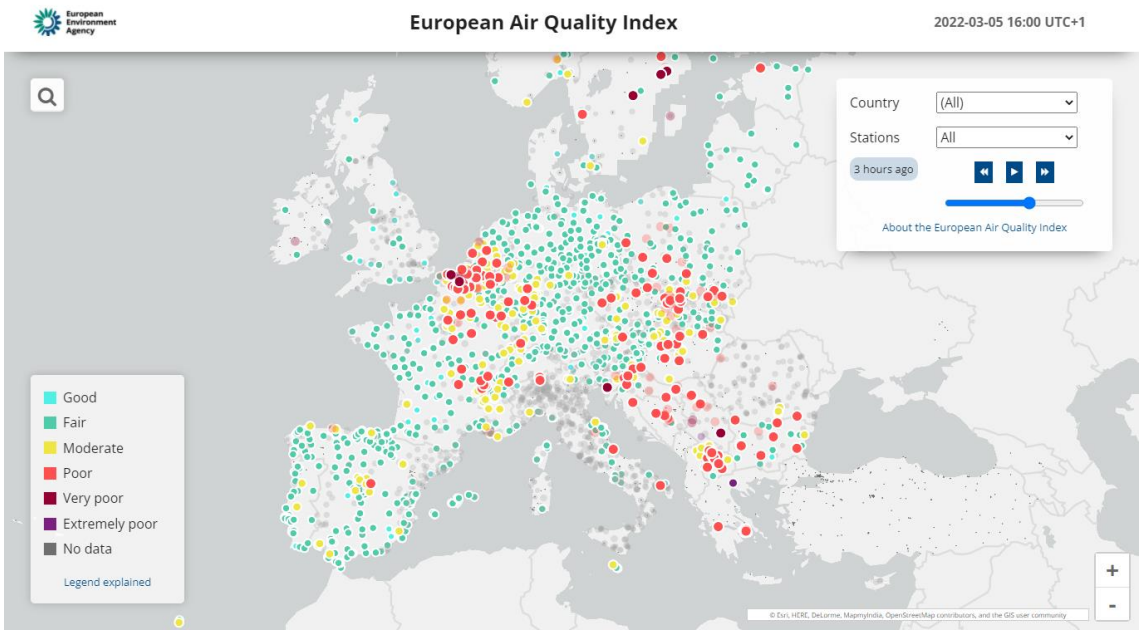


Figura 9. European Air Quality Index Viewer
Fuente. Agencia Europea de Medio Ambiente - https://airindex.eea.europa.eu/Map/AQI/Viewer/

Sin embargo, el uso de las herramientas SIG va más allá de la visualización de puntos de medición. Así, podemos encontrar numerosos trabajos de análisis espacio-temporales y de caracterización de calidad del aire. La publicación de S.S. Jensen et al. [24] ya describía en el año 2001 AirGIS, un sistema diseñado para ayudar a las autoridades locales en la gestión de la contaminación atmosférica en las ciudades de Dinamarca.

También podemos encontrar mapas para la visualización de modelos de predicción de la calidad del aire como CALIOPE-Urban [25], un proyecto del centro de supercomputación de Barcelona que acopla un modelo de predicción a mesoescala (escala nacional o urbana) con un modelo de dispersión sobre calles. A continuación, se muestra una imagen de esta aplicación con los nivele de NOx de Barcelona a las 7:00 AM. Las zonas rojas indican el flujo de vehículos de entrada a la ciudad.

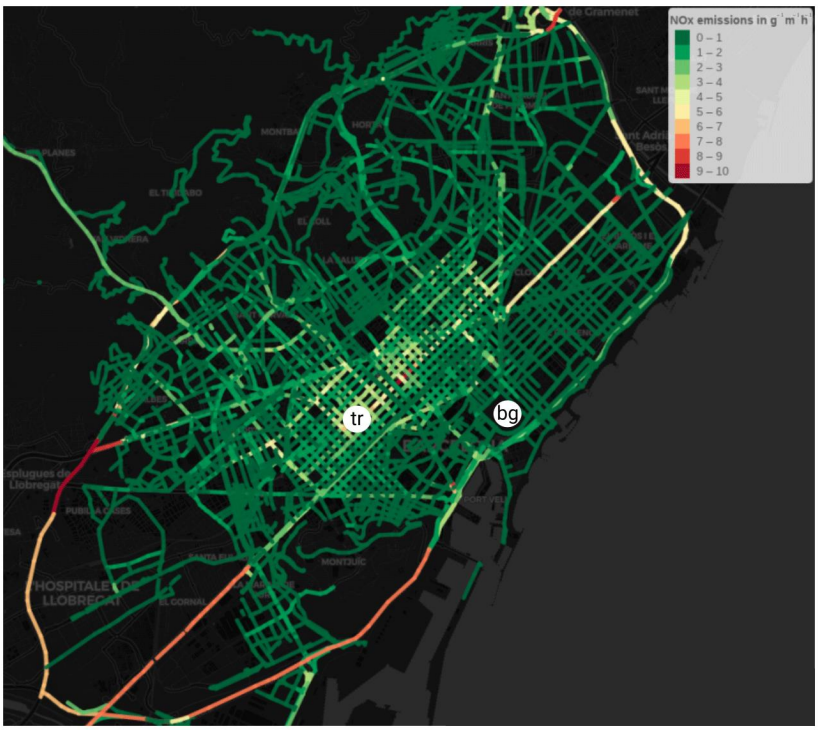


Figura 10. Emisiones de NOx en la ciudad de Barcelona a las 7:00 AM
Fuente. European Goesciences Union

En definitiva, la visualización, tanto de valores monitorizados como de modelos de concentraciones de emisiones, ha demostrado ser una herramienta efectiva para la ayuda en el análisis y toma de decisiones en relación a la contaminación en las ciudades. Numerosos estudios se llevan presentando en las dos últimas décadas. Por citar algunos de los más relevantes encontrados, tenemos las publicaciones de DJ Briggs et al. [26] del año 1997 o de M. Jerret et al. [27] del año 2001.

2.3 Modelos de calidad de aire usando técnicas de aprendizaje automático

La tendencia actual para el análisis la calidad del aire no es tanto la visualización en determinados puntos de medición sino el uso de modelos que nos permitan definir la cantidad de partículas contaminantes en determinadas zonas, así como predecir sus valores a futuro.

La directiva europea 2008/50/EC sobre calidad de aire ambiente y aire limpio ponía ya énfasis en el uso de modelos que complementen la información monitorizada. Aunque en los años previos se habían estado utilizando modelos en los distintos Estados miembro, tanto a nivel nacional como a nivel local, estos eran de muy diversas formas y con métodos de evaluación que en muchos casos ni siquiera eran comparables [30]. Por este motivo surge el FAIRMODE (Forum for Air Quality Modelling in Europe) en el año 2008, que pretende armonizar el uso de modelos de calidad de aire para su comparación documentación y fiabilidad. Estos modelos son usados para:

- Evaluar problemas de calidad del aire existentes
- Predicción de calidad del aire

- Planear medidas que reduzcan la contaminación del aire

En los últimos años se han realizado múltiples trabajos intentando buscar modelos que permitan estimar y predecir la concentración de agentes contaminantes en el aire. Muchos de estos trabajos se centran en la búsqueda de modelos de dispersión del aire que investigan fuentes de emisión en base a datos medidos de contaminantes, datos meteorológicos y datos geográficos. Estos modelos requieren de conocimientos de las propiedades físicas y químicas tanto de contaminantes como de la atmósfera y son adoptados principalmente por organizaciones de protección medioambiental como es el caso de la EPA, la Agencia de Protección Medioambiental de los Estados Unidos con su modelo AERMOD [28].

Para este trabajo, sin embargo, nos centraremos en una segunda corriente que utiliza métodos de aprendizaje automático confiando en datos de medición históricos. El análisis de A. Masih [29] recoge un amplio número de estos estudios y los divide en dos grandes grupos: estimación y predicción. Para el primer grupo, se concluye que el aprendizaje por conjuntos y la regresión lineal son los métodos más adecuados, mientras que para el segundo grupo las redes neuronales y SVM son las aproximaciones preferidas. La siguiente imagen muestra la evolución del número de publicaciones para la predicción de calidad del aire entre los años 2013 y 2018.

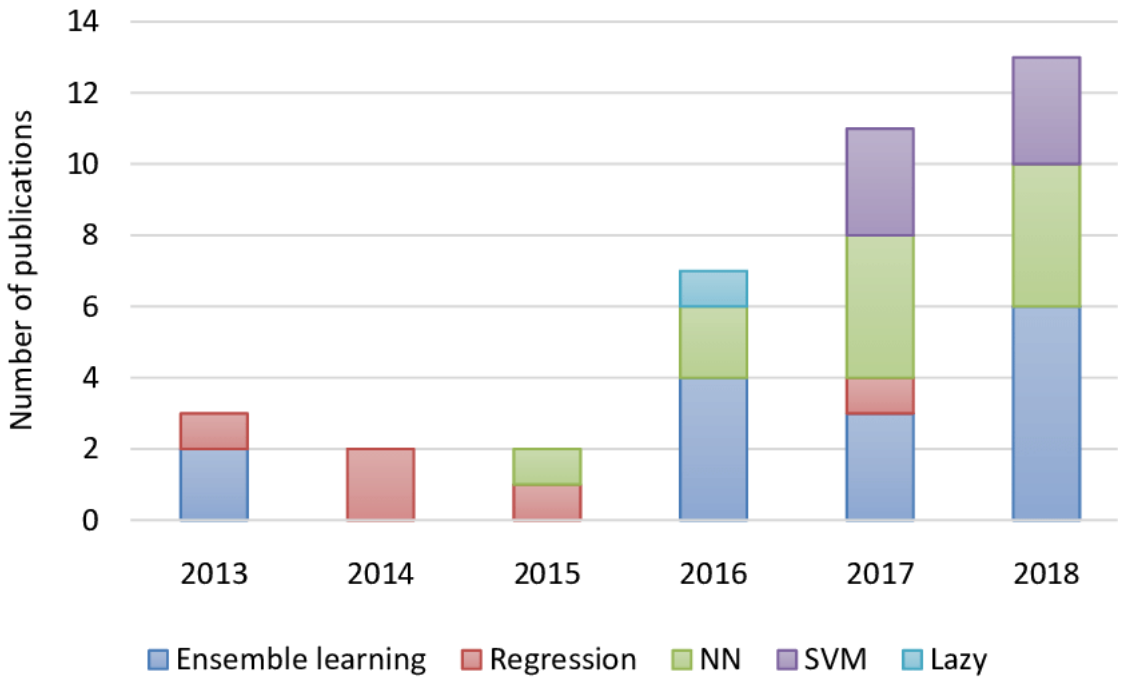


Figura 11. Número de estudios basados en diferentes algoritmos de aprendizaje automático
Fuente. A. Masih - Machine learning algorithms in air quality modeling

El último año de referencia muestra la preferencia de utilización de métodos de aprendizaje por conjuntos, esto es, la utilización de múltiples modelos de aprendizaje automático para combinar sus resultados, y entre estos, se sugiere Random Forest como la técnica más popular. Un ejemplo de esto es el trabajo de Joanna A. Kaminska [30] quien a través de este tipo de algoritmo crea un modelo para la predicción de O_2, NO_x y $PM_{2.5}$ para la ciudad polaca de Wrocław. Un aspecto

interesante de este trabajo es la división del conjunto de datos en cuatro subconjuntos, uno por cada estación, debido a las condiciones meteorológicas tan diferentes. Según la autora, esta decisión se mostró acertada ya que los valores de concentraciones varían mucho en los cuatros subconjuntos creados.

Para la evaluación de los modelos, se utilizan principalmente cuatro métodos:

- Coeficiente de correlación R^2
- Error medio absoluto (MAE)
- Raíz del error cuadrático medio (RECM)
- Error relativo absoluto (ERA).

En cuanto a los parámetros empleados para la predicción, los datos más utilizados son datos meteorológicos, siendo los más relevantes los siguientes:

- Temperatura
- Velocidad del viento
- Dirección del viento
- Presión atmosférica
- Humedad relativa.

Otros datos empleados son los referentes al tráfico, que se han revelado como los más importantes en relación a la predicción de óxidos de nitrógeno (NO_x).

Un último grupo de datos, menos frecuente su utilización, son los relativos a la superficie (altura, utilización del terreno, tipo de terreno…).

También en el campo del aprendizaje profundo se han realizado estudios para la predicción de la calidad del aire. Bekkar, A. et al. [31] proponen un modelo hibrido basado en redes neuronales convolucionales (CNN) y Long Short-Term Memory (LSTM) para la predicción de la concentración de partículas $PM_{2.5}$ en el área urbana de Beijing con resultados muy estables.

En España, el CIEMAT (Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas), dentro del Departamento de Medio Ambiente, cuenta con el Grupo de Modelización de la Contaminación Atmosférica que lleva más de quince años simulando la contaminación urbana, modelizándola tanto a nivel meso-escalar (la ciudad en su conjunto) como micro-escalar (barrios y distritos urbanos a muy alta resolución) [33]. La concentración se estima mediante la resolución numérica de un balance de masas que considera la emisión de contaminantes a la atmósfera, su transporte por el viento, la mezcla vertical como consecuencia del estado turbulento de la atmósfera, su depósito por vía húmeda (bien por arrastre por la lluvia o por su incorporación a las gotas de agua dentro de las nubes), o seca y la química atmosférica [34]. La siguiente imagen muestra de forma visual los distintos componentes que participan en la modelización de la concentración de partículas contaminantes en el aire.

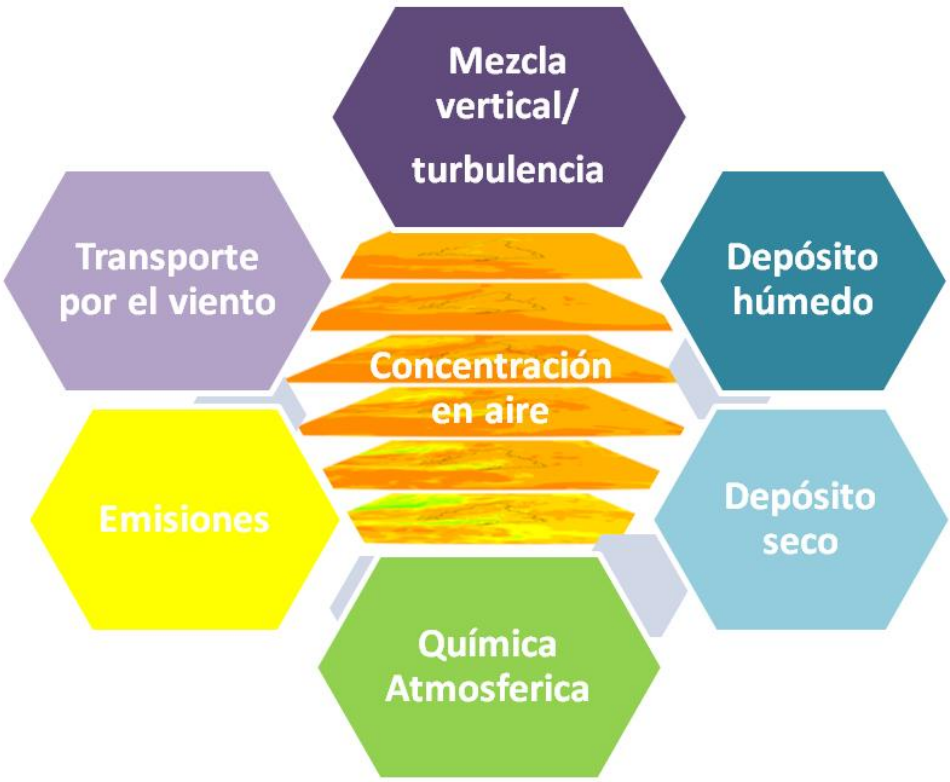


Figura 12. Modelización calidad del aire
Fuente. Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

2.4 Decisiones tecnológicas

Las herramientas que se utilizarán para el desarrollo del proyecto serán aquellas con las que se ha trabajado a lo largo del máster.

La preparación de los datos, el análisis estadístico y la aplicación de algoritmos de aprendizaje automático para la creación de un modelo de regresión de la calidad del aire de la ciudad de Madrid se realizarán con el leguaje de programación R utilizando el entorno de desarrollo integrado R Studio. Todo el código necesario se creará en un único fichero R Markdown que permitirá crear un fichero HTML una vez terminado. Una de las librerías que se utilizarán es openair, que proporciona herramientas para el análisis de datos de calidad del aire.

Según lo visto por los últimos trabajos realizados sobre predicción de calidad del aire, los algoritmos que se utilizarán serán Random Forest y SVM, que han probado ser los que mejores resultados ofrecen.

En cuanto a la herramienta GIS, finalmente se ha optado por la utilización de QGIS, que, al tratarse de software libre, nos permite hacer uso de todas sus funcionalidades. Se va a trabajar con la última versión estable disponible actualmente que es la 3.24.0. QGIS dispone de múltiples complementos, entre ellos la posibilidad de crear capas de tipo ráster mediante interpolación espacial, lo cual se utilizará para las visualizaciones que se crearán. Se descartó el uso de ArcGIS Pro

o Carto al ser software de pago y con licencias gratuitas únicamente por un periodo de 21 días el primero y 14 el segundo.

Para la creación de mapas temáticos en la web se utilizará la librería de JavaScript Leaflet. El mapa se podrá consultar a través de Internet en una página alojada en GitHub Pages.

Los datos necesarios para la realización del proyecto se guardarán en una base de datos Postgre SQL que dispone de la extensión PostGIS que permite convertir el sistema en una base de datos espacial mediante la adición de tres características: tipos de datos espaciales, índices espaciales y funciones que operan sobre ellos. Además, se usará la versión 4 de pgAdmin para todo el trabajo con la base de datos.

Tanto el código como la documentación generada durante el proyecto se guardará en GitHub para poder gestionar el control de versiones.

3. Diseño e implementación del trabajo

En este apartado se va a detallar el desarrollo del proyecto. En la primera parte se describen las fuentes de datos utilizadas, así como el proceso de carga, limpieza y preparación de los datos, proceso que ha sido realizado en el entorno de desarrollo R Studio. La segunda parte del proyecto consiste en un estudio de los datos de calidad del aire en la ciudad de Madrid, analizando los principales agentes contaminantes con los datos disponibles de los últimos 20 años. Por último, se han desarrollado varios modelos de predicción de calidad del aire a partir de datos meteorológicos.

Todo el código desarrollado para el análisis se ha subido a un repositorio de GitHub público: https://github.com/SergioRC70/UOC_Ciencia_de_Datos_TFM Este repositorio contiene los scripts de R en formato Markdown, el proyecto de QGIS con todas las capas necesarias para su ejecución y el código de visor web de datos de calidad del aire desarrollado con Leaflet. El fichero README.rm describe cada uno de los ficheros incluidos en el repositorio.

3.1 Carga, limpieza y preparación de los datos con R

3.1.1 Datos de calidad del aire

El portal de datos abiertos del Ayuntamiento de Madrid proporciona varios juegos de datos con información recogida por el Sistema Integral de Calidad del Aire. Para el desarrollo del proyecto se han utilizado los siguientes:

- Calidad del aire. Datos horarios desde 2001.
- Calidad del aire. Estaciones de control.

Estos datos se pueden encontrar en varios formatos (txt, xml y csv) siendo el formato csv el utilizado para este proyecto, encontrando la información dividida en ficheros mensuales.

Cada registro está estructurado de la siguiente manera:

PROVINCIA	MUNICIPIO	ESTACIÓN	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	...	H24	V24

Donde H01 corresponde al dato de la 1 de la mañana de ese día, con su código de validación correspondiente, V01; H02 es el dato de las 2 de la mañana, con código de validación V02 y así sucesivamente hasta H24/V24. Los datos horarios de las magnitudes corresponden a la media aritmética de los valores diezminutales que se registran cada hora.

Se ha trabajado únicamente con valores que han sido validados por lo que se han eliminado del dataset todos aquellos valores cuyo código de validación correspondiente (Vxx) sea 'N'.

Además, para poder trabajar con mayor comodidad, se ha transformado la estructura del dataset de manera que cada registro muestre el valor de un contaminante concreto para una única hora y para una estación determinada. En total disponemos de 3.464.568 observaciones.

Los ficheros contienen los datos recogidos para los siguientes agentes contaminantes:

ID	*Nombre contaminante*	*Código*
01	Dióxido de Azufre	SO2
06	Monóxido de Carbono	CO
07	Monóxido de Nitrógeno	NO
08	Dióxido de Nitrógeno	NO2
09	Partículas < 2.5 µm	PM2.5
10	Partículas < 10 µm	PM10
12	Óxidos de Nitrógeno	NOx
14	Ozono	O3
20	Tolueno	TOL
30	Benceno	BEN
35	Etilbenceno	EBE
37	Metaxileno	MXY
38	Paraxileno	PXY
39	Ortoxileno	OXY
42	Hidrocarburos totales (hexano)	TCH
43	Metano	CH4
44	Hidrocarburos no metánicos (hexano)	NMHC

Tabla 3. Lista de agentes contaminantes recogidos por las estaciones de control de Madrid
Fuente. Portal de datos abierto del Ayuntamiento de Madrid.

En el punto 1.1.3 se detallaron los agentes contaminantes con mayor presencia en la atmósfera y que además son utilizados para el cálculo del AQI (índice de calidad del aire). El proyecto se centrará en el análisis de estos cinco agentes por lo que la información relativa al resto de agentes se obviará.

Por su parte, el juego de datos de estaciones de control contiene la información geométrica además del tipo de estación. Existen tres tipos de estación:

- Urbanas de fondo: son representativas de la exposición de la población urbana en general.
- De tráfico: situada de tal manera que su nivel de contaminación está influido principalmente por las emisiones procedentes de una calle o carretera próxima, pero se ha de evitar que se midan microambientes muy pequeños en sus proximidades.
- Suburbanas: están situadas a las afueras de la ciudad, en los lugares donde se encuentran los mayores niveles de ozono.

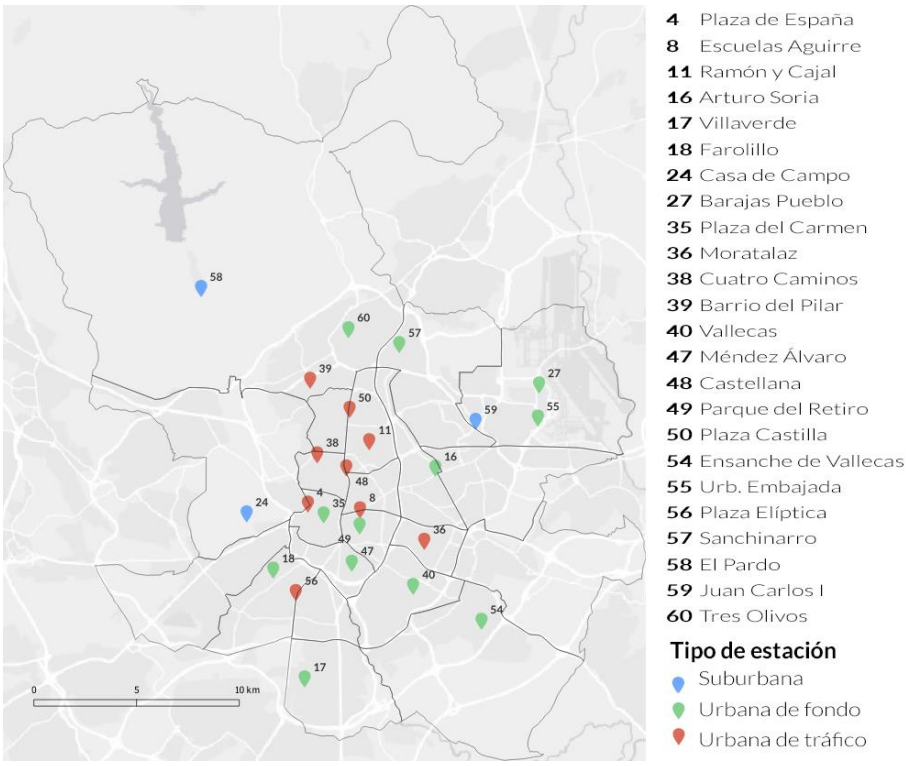


Figura 13. Mapa de Estaciones de Control – Red de Vigilancia Madrid
Fuente. Elaboración propia

3.1.2 Datos meteorológicos

Los datos meteorológicos han sido tomados también del portal de datos abiertos del Ayuntamiento de Madrid. El Sistema Integral de la Calidad del Aire del Ayuntamiento de Madrid incluye la red meteorológica municipal que proporciona distintos juegos de datos a este respecto con distinta granularidad. Los dataset con los que se ha trabajado en el proyecto son los siguientes:

- Datos meteorológicos. Datos horarios desde 2019.
- Datos meteorológicos. Estaciones de control.

Como vemos, el número de datos meteorológicos de los que disponemos es bastante menor ya que sólo contamos con datos desde el año 2019.

Cada registro está estructurado de la misma manera que los datos de calidad del aire:

PROVINCIA	MUNICIPIO	ESTACIÓN	MAGNITUD	PUNTO_MUESTREO	ANO	MES	DIA	H01	V01	...	H24	V24

Donde H01 corresponde al dato de la 1 de la mañana de ese día, con su código de validación correspondiente, V01; H02 es el dato de las 2 de la mañana, con código de validación V02 y así sucesivamente hasta H24/V24.

Al igual que con los datos de calidad de aire, se ha trabajado únicamente con valores que han sido validados por lo que se han eliminado del dataset todos aquellos

valores cuyo código de validación correspondiente (Vxx) sea 'N' y se ha transformado la estructura del dataset de manera que cada registro muestre el valor de un parámetro específico para una única hora y una estación determinada. En total contamos con 685.896 observaciones.

Los parámetros recogidos por las estaciones meteorológicas son los siguientes:

ID	*Nombre contaminante*	*Unidad de Medida*
80	Radiación Ultravioleta	Mw/m^2
81	Velocidad Viento	m/s
82	Dirección de Viento	-
83	Temperatura	^oC
86	Humedad Relativa	%
87	Presión Bariométrica	mb
88	Radiación Solar	W/m^2
89	Precipitación	l/m^2

Tabla 4. Lista de variables recogidas por las estaciones meteorológicas de Madrid
Fuente. Portal de datos abiertos del Ayuntamiento de Madrid

Desafortunadamente, las estaciones meteorológicas no coinciden en todos los casos con las estaciones de control de calidad del aire. Además, muchas de ellas no disponen de todas las variables meteorológicas.

Con el objetivo de disponer del mayor número de datos posibles, se han identificado tres estaciones meteorológicas que sí recogen información para todos los datos y que se encuentran muy próximas a otras tres estaciones que disponen de menos datos (las parejas de estaciones son 102-36, 103-17 y 108-39). La localización de todas las estaciones se puede ver en la siguiente imagen.

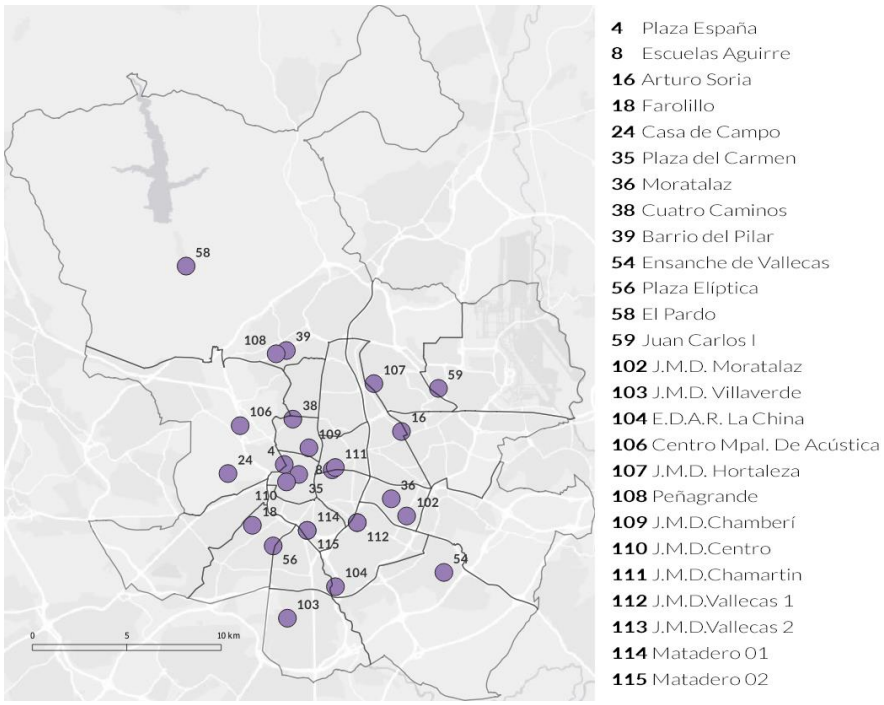


Figura 14. Mapa de Estaciones Meteorológicas de Madrid
Fuente. Elaboración propia

Además, se ha eliminado la información de radiación ultravioleta (80) ya que no hay datos en 675860 observaciones de las 685896 de las que disponemos.

3.1.3 QGIS

Con el objetivo de añadir un análisis geoespacial al estudio de los datos de calidad del aire, se ha utilizado QGIS. En primer lugar, los datos de los que disponemos (datos de calidad del aire, información meteorológica, estaciones de control y estaciones meteorológicas) se han añadido a una base de datos PostgreSQL a la que previamente se había añadido el módulo PostGIS para dotarla de soporte de objetos geográficos.

En base a esta información, se han ido creando las vistas necesarias en PostgreSQL para añadir las distintas capas en QGIS. Además, se ha añadido una capa obtenida del portal de datos abiertos del Ayuntamiento de Madrid con la información geográfica de los distintos distritos de la ciudad y la de la zona de bajas emisiones del distrito centro, ambas en formato shapefile.

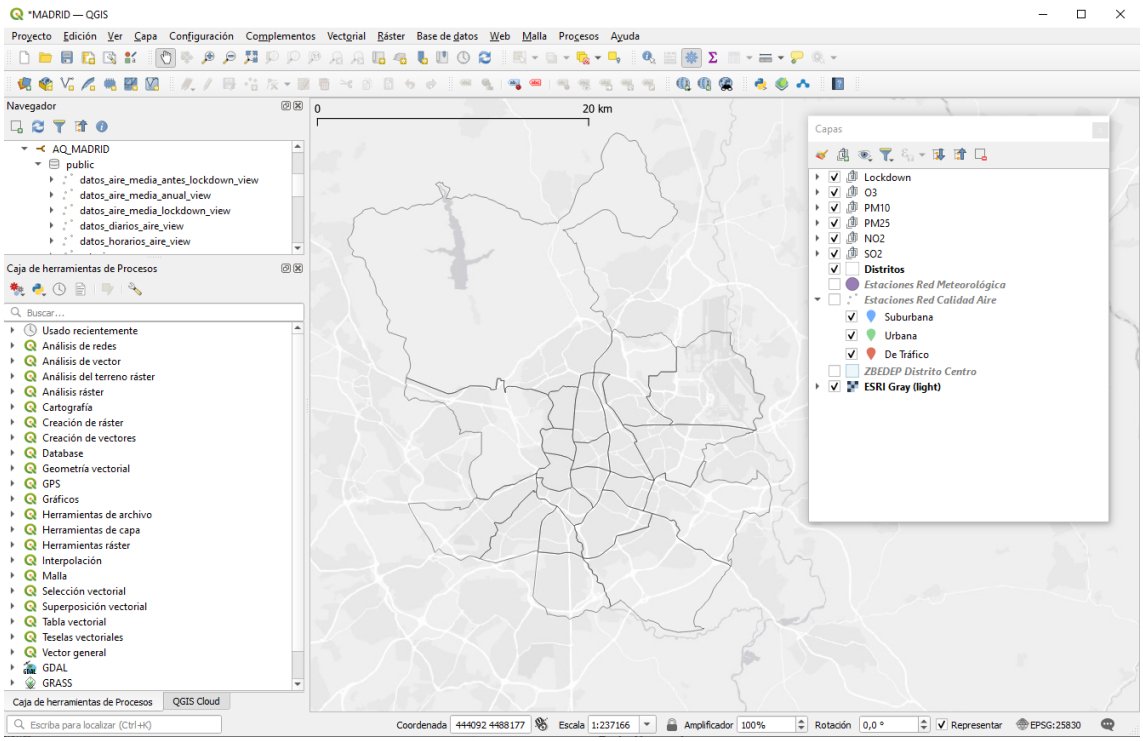


Figura 15. Mapa QGIS creado para el proyecto
Fuente. Elaboración propia

Las distintas interpolaciones se han realizado utilizando esta herramienta.

3.1.4 Visor Leaflet

Uno de los objetivos planteados en el apartado 1.3.2 para este proyecto era la creación de una visualización interactiva que permita analizar los datos de calidad de aire a lo largo de todo el periodo para el que se disponen de datos (desde el año 2001). El proyecto de QGIS podría cubrir este objetivo, sin embargo, no es accesible desde la web, por lo que adicionalmente se ha desarrollado una visualización utilizando la librería de JavaScript de código abierto Leaflet.

El mapa cuenta con un selector de fecha que permite determinar un día entre el 1 de enero de 2001 y el 31 de diciembre de 2021 para el que se muestran los datos de calidad del aire de las distintas estaciones de control. El color mostrado para la estación de control indica el valor del Índice de Calidad del Aire para esa estación y pulsando sobre la misma se pueden ver los valores diarios de los distintos agentes contaminantes. Para cargar los datos de las estaciones de control y de calidad del aire se utiliza la librería de JavaScript D3. Se muestra a continuación el fragmento de código que carga estos datos incluido en la función renderStations, dentro del fichero renderStations.js.

```
// Cargamos los datos de los ficheros csv. Lo primero es cargar la estaciones de control
d3.dsv(";", "./data/estaciones_control_aire.csv").then(function(data) {
        data.forEach(function(d) {
                var latitud = d.LATITUD;
                var longitud = d.LONGITUD;
                var estacion = d.CODIGO_CORTO;
                var nom_estacion = d.ESTACION

                // Cargarmos el fichero con los datos para el año seleccionado
                var file_name = "./data/datos" + fecha.getFullYear() + "12.csv";
                d3.dsv(";", file_name).then(function(data) {
                        var filteredData = data.filter(function(row, i) {
                                var myMonth = ("0" + (fecha.getMonth() + 1)).slice(-2);
                                return row.MES == myMonth && row.ESTACION == estacion &&
(row.MAGNITUD == '1' || row.MAGNITUD == '8' || row.MAGNITUD == '9' || row.MAGNITUD == '10'
|| row.MAGNITUD == '14');
                        });
```

La siguiente imagen muestra el mapa web al que se puede acceder en la siguiente URL: https://sergiorc70.github.io/UOC_Ciencia_de_Datos_TFM/Visor/mapa.html. Este mapa ha sido probado con la última versión de los navegadores Google Chrome, Mozilla Firefox y Microsoft Edge.

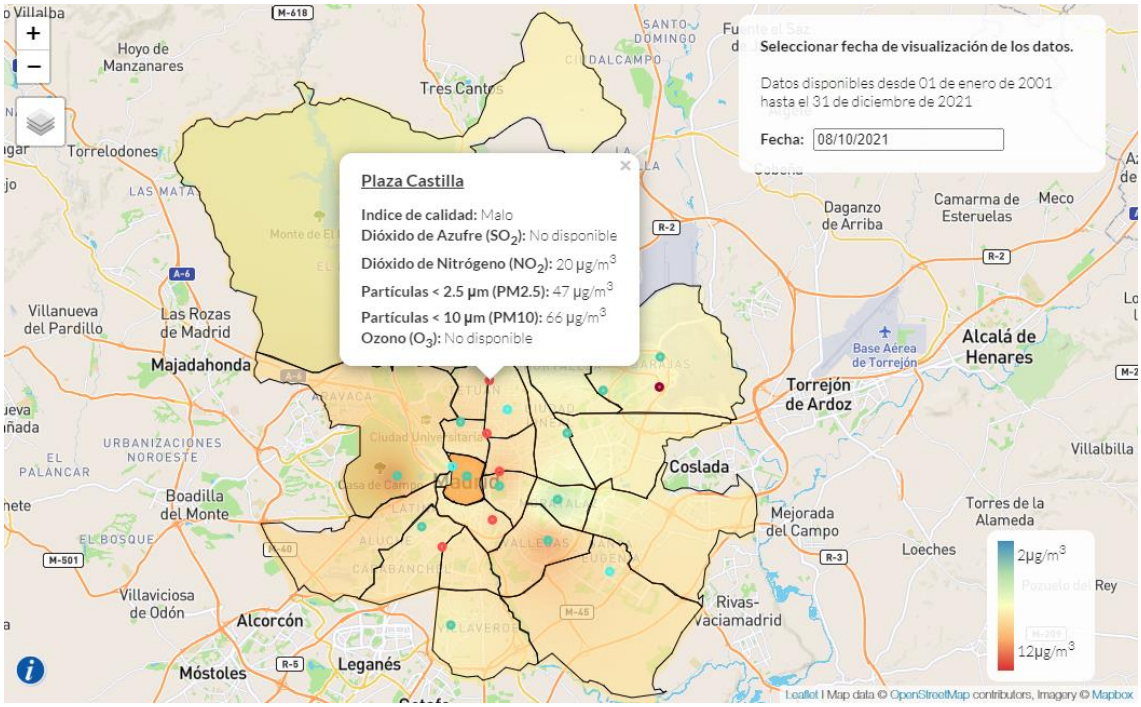


Figura 16. Mapa web basado en Leaflet
Fuente. Elaboración propia

Se han añadido al mapa la capa de distritos de Madrid y la capa de la zona de bajas emisiones, ambas como capas de tipo overlay, así como capas base de las distintas interpolaciones realizadas en QGIS correspondientes a los cuatro últimos años para los cinco agentes contaminantes con los que se ha trabajado en el proyecto.

3.2 Análisis de datos de calidad del aire

3.2.1 SO_2

El dióxido de azufre se origina principalmente por la combustión de carburantes fósiles y la fundición de minerales ricos en sulfatos. En la ciudad de Madrid se origina principalmente por el sector residencial, comercial e institucional [35].

Los valores observados durante los últimos 20 años se sitúan muy por debajo de la recomendación de la OMS (40 µg/m^3 en un periodo de 24horas). Además, se observa una tendencia a la baja, aunque es el año 2012 en el que se recogieron los valores más bajos en media. Esto se puede observar en la siguiente gráfica.

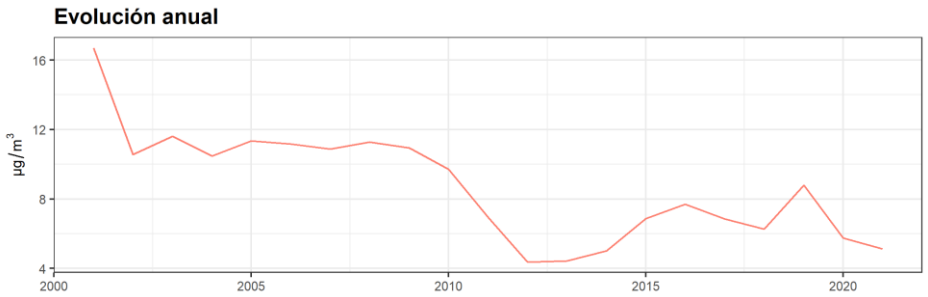


Figura 17. Evolución anual SO_2 en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

Respecto a las distintas estaciones de control, no se aprecian grandes diferencias si bien es cierto que los valores más elevados se recogen en las estaciones de tipo urbano de tráfico mientras que las estaciones suburbanas son las que recogen mejores valores. A continuación, se muestra el diagrama de cajas con los valores recogidos en las distintas estaciones y en las estaciones agrupadas por tipo.

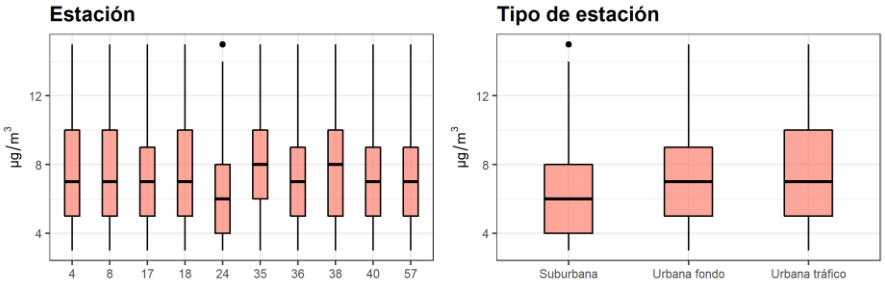


Figura 18. Valores medios de SO_2 por estación de control y tipo de estación en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

A nivel horario, vemos que el valor medio se mantiene constante durante todo el día, sin embargo, si observamos los valores a nivel mensual sí que vemos diferencias. Los meses con valores más altos corresponden a los valores más fríos y en los que por lo tanto se utiliza más la calefacción. Esto tiene sentido ya que como hemos dicho, ésta es la principal fuente de contaminación de SO_2 que tenemos en la ciudad. Se muestra a continuación el diagrama de cajas con los valores recogidos por mes y por hora.

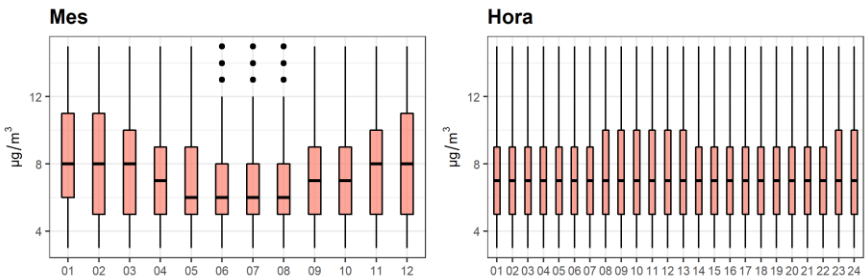


Figura 19. Valores medios de SO_2 en Madrid por mes y hora (periodo 2001-2021)
Fuente. Elaboración propia

Las siguientes imágenes muestran los niveles de SO_2 en los últimos 4 años obtenidos por interpolación en base a los valores de las estaciones de control. La interpolación se ha realizado con el método de distancia inversa ponderada (IDW) utilizando un valor de potencia igual a 2.

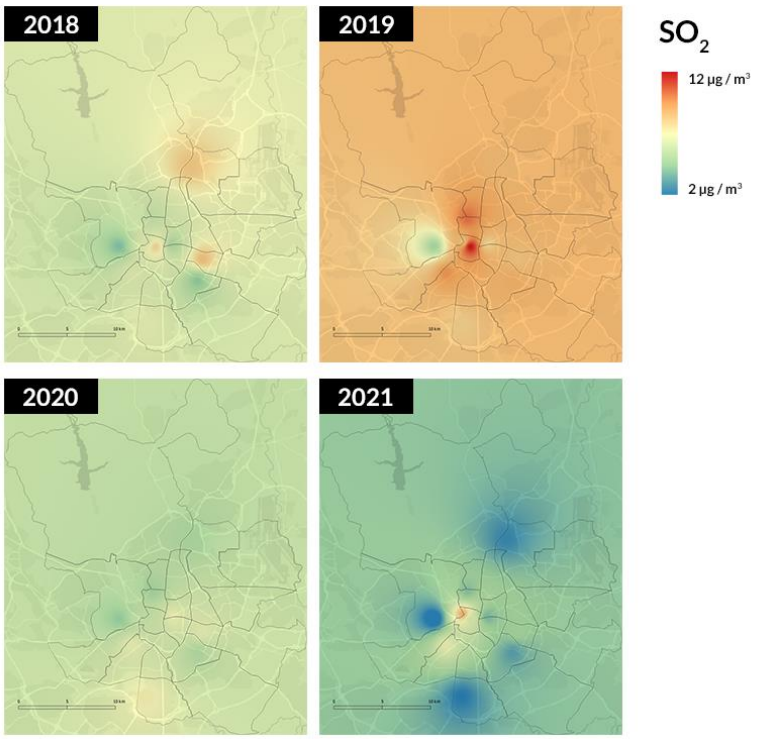


Figura 20. Mapas de niveles de SO_2 en Madrid
Fuente. Elaboración propia

El patrón que presentan los datos es algo extraño, pero si observamos la evolución anual podemos ver el aumento en los valores de SO_2 durante el año 2019. En cualquier caso, como se comentaba anteriormente, los valores se sitúan muy por debajo de las recomendaciones de las OMS.

3.2.2 NO_2

El dióxido de nitrógeno es un contaminante indicador principalmente de tráfico rodado. La principal fuente de emisión de NO_2 en la ciudad de Madrid son los vehículos, especialmente los diésel.

La recomendación de la OMS es que el valor medio anual no debe superar los 10 µg/m^3 en media anual. Como se puede observar en la siguiente gráfica que muestra el valor medio anual de los últimos 20 años, este valor ha sido siempre superado.

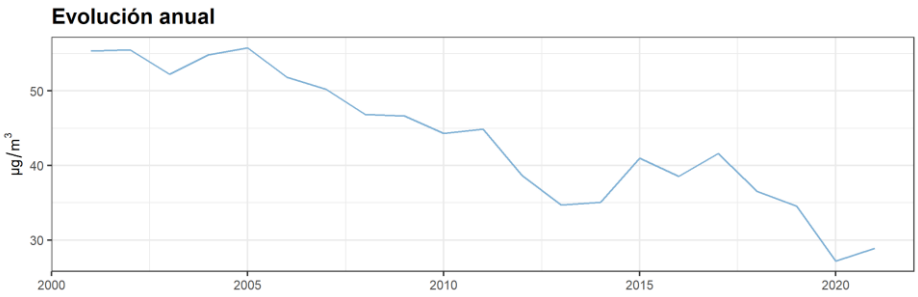


Figura 21. Evolución anual NO_2 en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

Lógicamente, los peores valores se observan en las estaciones de tipo urbano de tráfico, que son aquellas cuyo nivel de contaminación está influido principalmente por una carretera próxima. Las siguientes gráficas muestran los valores medios por estación y por tipo de estación.

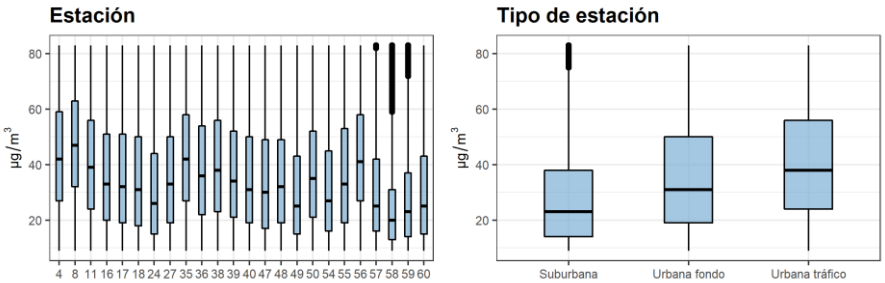


Figura 22. Valores medios de NO_2 por estación de control y tipo de estación en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

Las horas del día con mayores niveles de NO_2 son las horas punta de tráfico, apreciándose el aumento de este contaminante casi de forma inmediata. Los meses del año con menores niveles de contaminación son los de verano. Una posibilidad es que coincida con que estos meses son los que menos tráfico hay. Sin embargo, el descenso en otros meses con más actividad, como pueden ser mayo o junio podrían sugerir una relación con la temperatura. A continuación, se muestran las gráficas con los valores medios de NO_2 por mes y por hora.

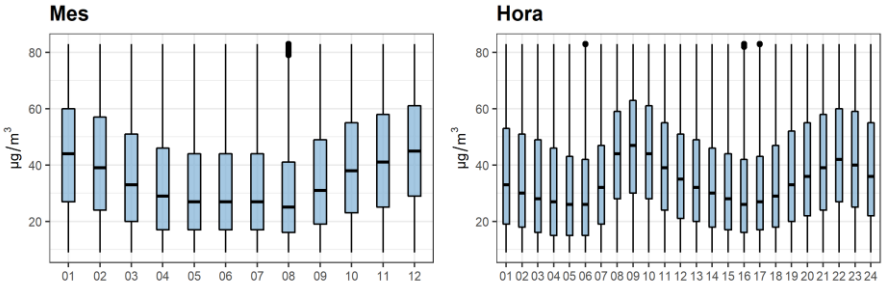


Figura 23. Valores medios de NO_2 en Madrid por mes y hora (periodo 2001-2021)
Fuente. Elaboración propia

Las siguientes imágenes muestran los niveles de NO_2 en los últimos 4 años obtenidos por interpolación en base a los valores de las estaciones de control. La

interpolación se ha realizado con el método de distancia inversa ponderada (IDW) utilizando un valor de potencia igual a 2.

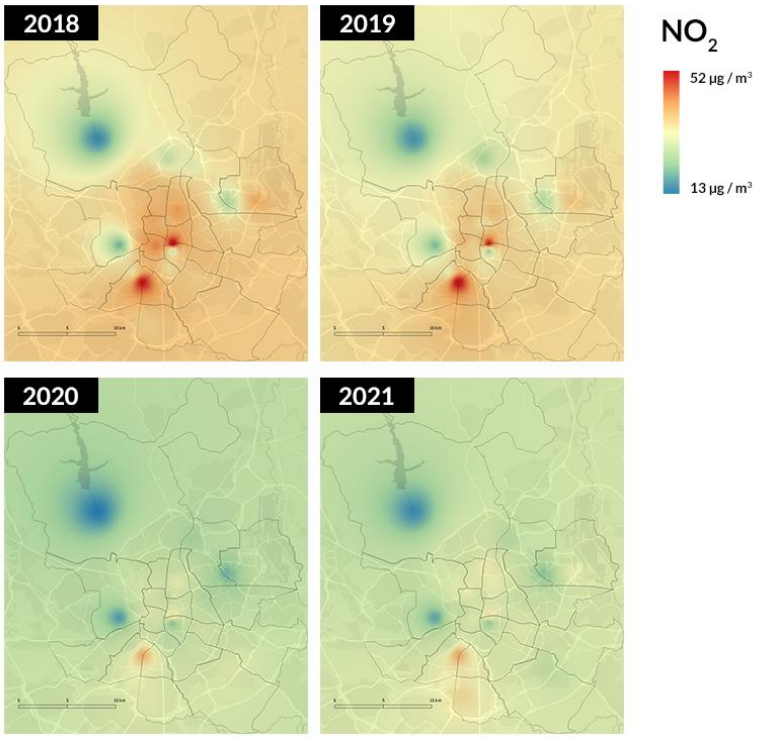


Figura 24. Mapas de niveles de NO_2 en Madrid
Fuente. Elaboración propia

Se observa un cambio drástico en el año 2020 que coincide con el año de mayor reducción de la movilidad y actividad en la ciudad debido a la situación sanitaria provocada por la pandemia de COVID-19. Esto es algo que se verá para varios agentes contaminantes y que se estudiará con algo más de detalle en el apartado 3.2.7.

3.2.3 PM_{10}

Tal y como describen las guías de la OMS, el material particulado es una compleja mezcla de características físicas y químicas muy diversas. Las partículas con diámetro inferior a 10µm constituyen el material particulado grueso, que en Madrid se origina principalmente por el tráfico rodado. La contribución del tráfico engloba tanto las emisiones directas de partículas primarias desde el tubo de escape de los vehículos motorizados, como la suspensión de materiales que se acumulan en el pavimento (productos de abrasión mecánica de vehículos, frenos, ruedas, emisiones derivadas de obras de construcción o demolición, etc.) [35].

La siguiente gráfica muestra la evolución anual de los niveles medios de PM_{10} recogidos en Madrid durante los últimos 20 años. Se observa la tendencia a la baja, sin embargo, únicamente en el año 2020 se presenta una media anual inferior al valor recomendado por la OMS en sus últimas directrices sobre la calidad del aire

(15 µg/m^3 en media anual), aunque sí que se mantiene por debajo del valor límite establecido en la legislación española (40 µg/m^3 en media anual)

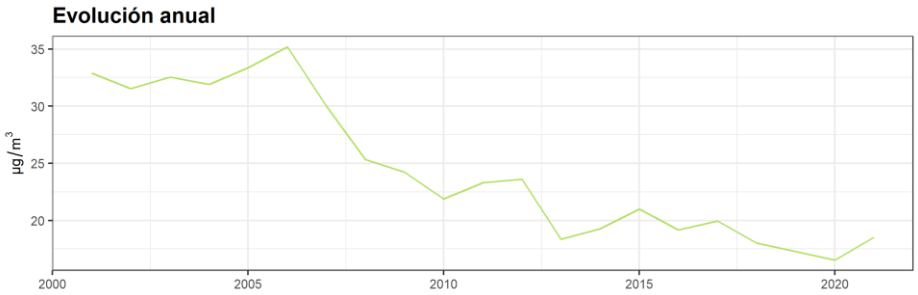


Figura 25. Evolución anual PM_{10} en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

No se aprecian grandes diferencias de valores entre las distintas estaciones de control ni por tipo de estación. Aunque el material particulado PM_{10} se origina en Madrid principalmente por el tráfico rodado, no se observan valores significativamente mayores en las estaciones de control de tipo urbanas de tráfico. Esto quizás se deba a que, dado su reducido tamaño, el material particulado puede viajar transportado grandes distancias y por lo tanto es fácil que se disperse. Las siguientes dos gráficas muestran los diagramas de cajas para las distintas estaciones de control, así como por tipo de estación.

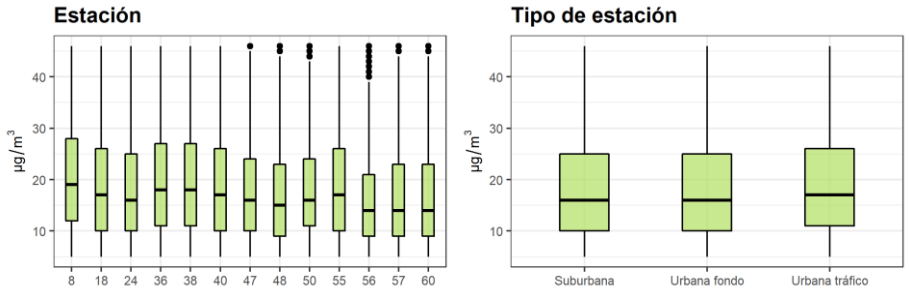


Figura 26. Valores medios de PM_{10} por estación de control y tipo de estación en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

A nivel horario, vemos que los valores más altos coinciden con las horas de mayor tráfico, mientras que a nivel mensual, los valores más altos coinciden con los meses más calurosos, lo que podría indicar una relación con aspectos meteorológicos. Esta apreciación se estudiará más en detalle en el capítulo 3.3. Se muestra a continuación el diagrama de cajas con los valores recogidos por mes y por hora.

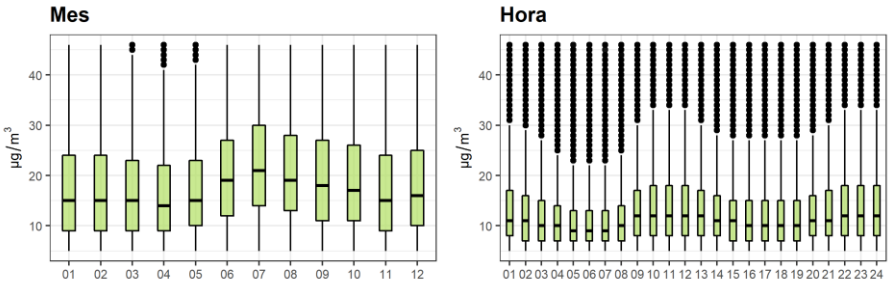


Figura 27. Valores medios de PM_{10} en Madrid por mes y hora (periodo 2001-2021)
Fuente. Elaboración propia

Las siguientes imágenes muestran los niveles de PM_{10} en los últimos 4 años obtenidos por interpolación en base a los valores de las estaciones de control. La interpolación se ha realizado con el método de distancia inversa ponderada (IDW) utilizando un valor de potencia igual a 2.

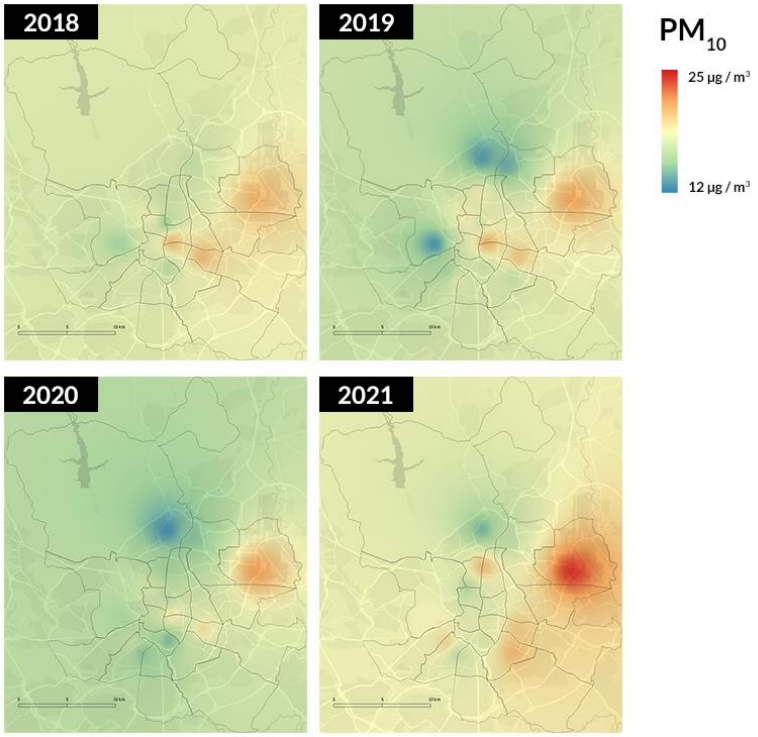


Figura 28. Mapas de niveles de PM_{10} en Madrid
Fuente. Elaboración propia

De nuevo, vemos como el mejor año es el 2020. Además, vemos como el punto de mayor contaminación por PM_{10} se encuentra cerca del aeropuerto internacional de Madrid.

3.2.4 $PM_{2.5}$

El material particulado fino es aquel cuyo diámetro es inferior a 2,5 µm y constituyen las partículas más nocivas para la salud humana.

También se observa una tendencia muy a la baja, especialmente marcada en los diez primeros años con datos (desde el año 2003, primer año con datos para PM_{25}). El valor medio anual supera en todos los años el valor recomendado por la OMS en sus últimas directrices sobre la calidad del aire (5 μg/m^3 en media anual). La siguiente gráfica muestra esta evolución anual.

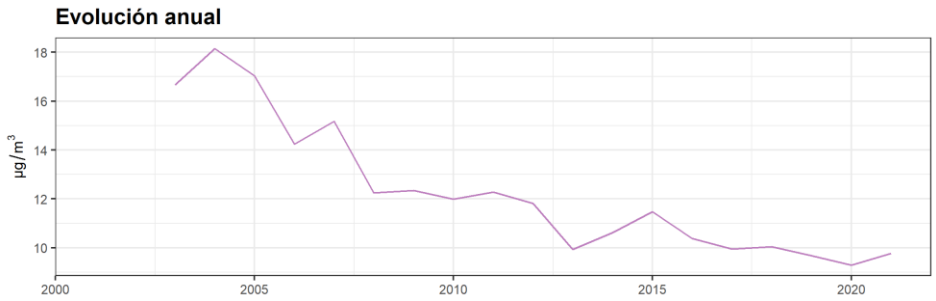


Figura 29. Evolución anual $PM_{2.5}$ en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

Respecto a los valores recogidos por las distintas estaciones, no se aprecian diferencias ni por estación ni por tipo. Al igual que pasaba con las partículas de PM_{10}, el material particulado fino se dispersa fácilmente debido a su reducido tamaño y es quizás por esto por lo que no se aprecian diferencias significativas. Las siguientes dos gráficas muestran los diagramas de cajas para las distintas estaciones de control, así como por tipo de estación.

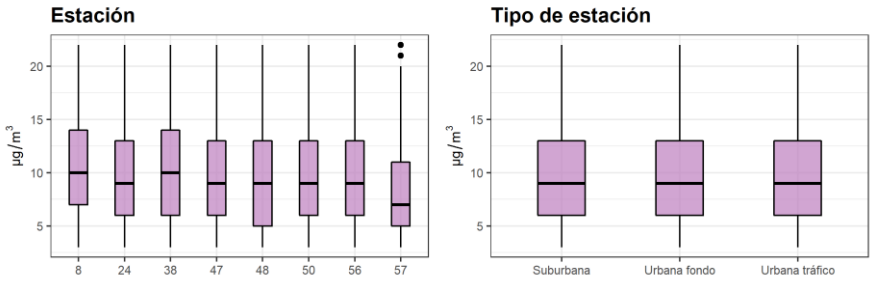


Figura 30. Valores medios de $PM_{2.5}$ por estación de control y tipo de estación en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

A nivel horario sí que se observan diferencias, viéndose los valores más altos durante las horas de mayor actividad del día. A nivel mensual sí que se aprecian diferencias, sin embargo, no parece existir un patrón claro relacionado con la meteorología ya que los valores más altos coinciden con los meses de agosto y diciembre. Una posible explicación puede estar en que una de las principales fuentes de emisión de PM_{25} son las fuentes naturales, por lo que sería necesario analizarlas, principalmente las masas de aire sahariano (https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/fuentes-naturales/). Se muestra a continuación el diagrama de cajas con los valores recogidos por mes y por hora.

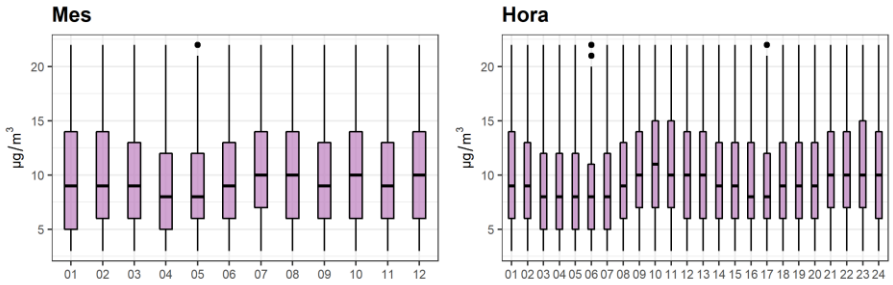


Figura 31. Valores medios de $PM_{2.5}$ en Madrid por mes y hora (periodo 2001-2021)
Fuente. Elaboración propia

Las siguientes imágenes muestran los niveles de PM_{25} en los últimos 4 años obtenidos por interpolación en base a los valores de las estaciones de control. La interpolación se ha realizado con el método de distancia inversa ponderada (IDW) utilizando un valor de potencia igual a 2.

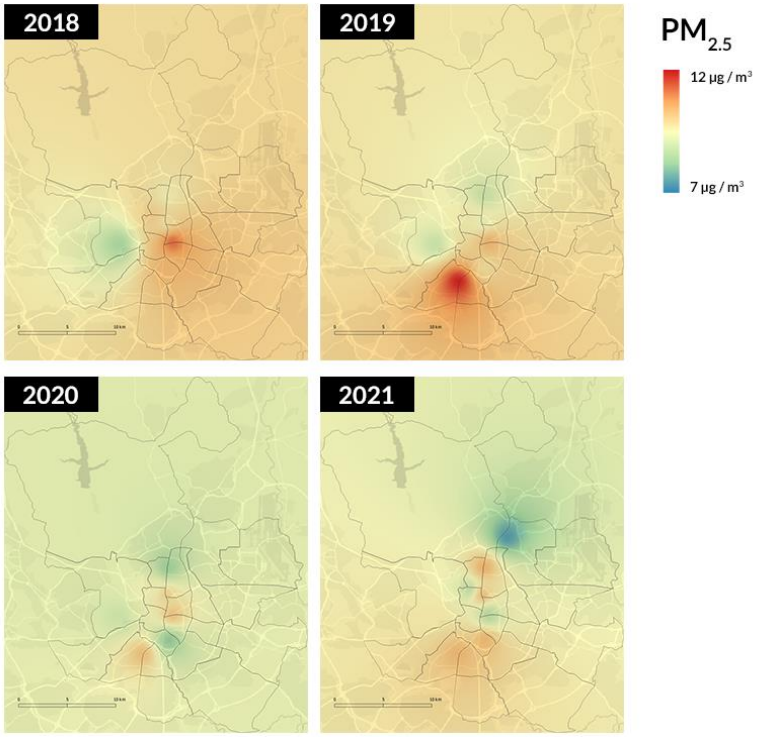


Figura 32. Mapas de niveles de $PM_{2.5}$ en Madrid
Fuente. Elaboración propia

Vemos de nuevo como el mejor año es el 2020.

3.2.5 O_3

El Ozono es un contaminante secundario, formado por una serie de reacciones complejas en la atmósfera a partir de otros contaminantes primarios, principalmente el dióxido de nitrógeno y los compuestos orgánicos volátiles.

Mientras que para el resto de agentes contaminantes veíamos la tendencia a la baja durante los últimos 20 años, la siguiente gráfica muestra la evolución durante este periodo del O_3 y en ella se puede apreciar que como los valores registrados van en aumento. Todavía no se supera el valor máximo establecido en las guías de la OMS (60 µg/m^3 en media anual), pero esta tendencia hace que se vaya aproximando cada vez más.

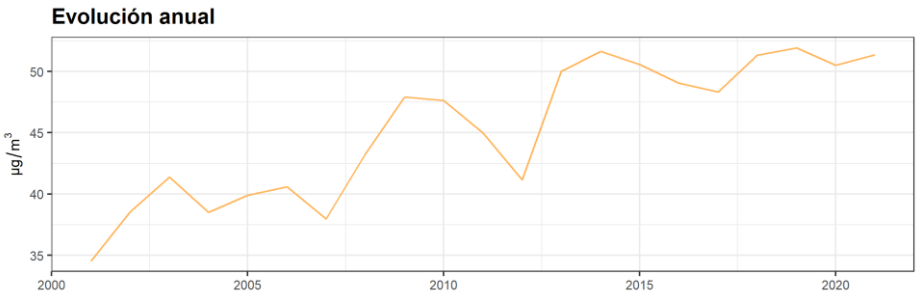


Figura 33. Evolución anual O_3 en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

Los mayores valores de O_3 se registran en las zonas periféricas de la ciudad. Se puede apreciar en las siguientes gráficas que muestran los valores medios por estación y por tipo de estación. Las estaciones suburbanas son las que registran los valores más altos.

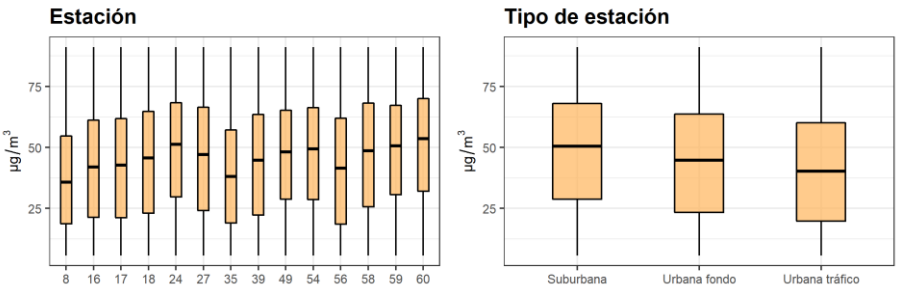


Figura 34. Valores medios de O_3 por estación de control y tipo de estación en Madrid (periodo 2001-2021)
Fuente. Elaboración propia

En cuanto a los meses del año con mayor cantidad de O_3, se observa que son los meses con temperaturas más altas los que tienen valores más elevados. Los rayos UV intensos del sol son el catalizador de las reacciones entre las emisiones de NO_2 y los COV y es por esto que las condiciones que conducen a días de alto ozono son típicas de los días calurosos de verano. En relación a las horas del día con mayor cantidad de ozono, son típicamente las primeras horas de la tarde las que registran peores valores. Pese a que el ozono troposférico se genera a partir de NO_2, que hemos visto que se producía en mayor cantidad en horas punta de tráfico, el tiempo que tardan las reacciones químicas en generar ozono hace que los valores máximos de O_3 se retrasen hasta estas horas de la tarde. Las siguientes gráficas muestran los valores medios registrados por mes y por hora.

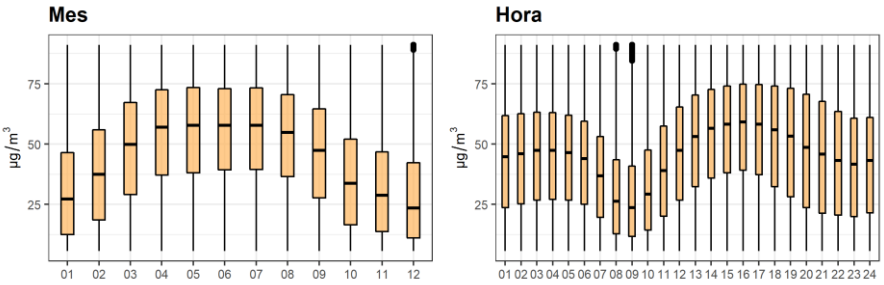


Figura 35. Valores medios de O_3 en Madrid por mes y hora (periodo 2001-2021)
Fuente. Elaboración propia

Las siguientes imágenes muestran los niveles de O_3 en los últimos 4 años obtenidos por interpolación en base a los valores de las estaciones de control. La interpolación se ha realizado con el método de distancia inversa ponderada (IDW) utilizando un valor de potencia igual a 2.

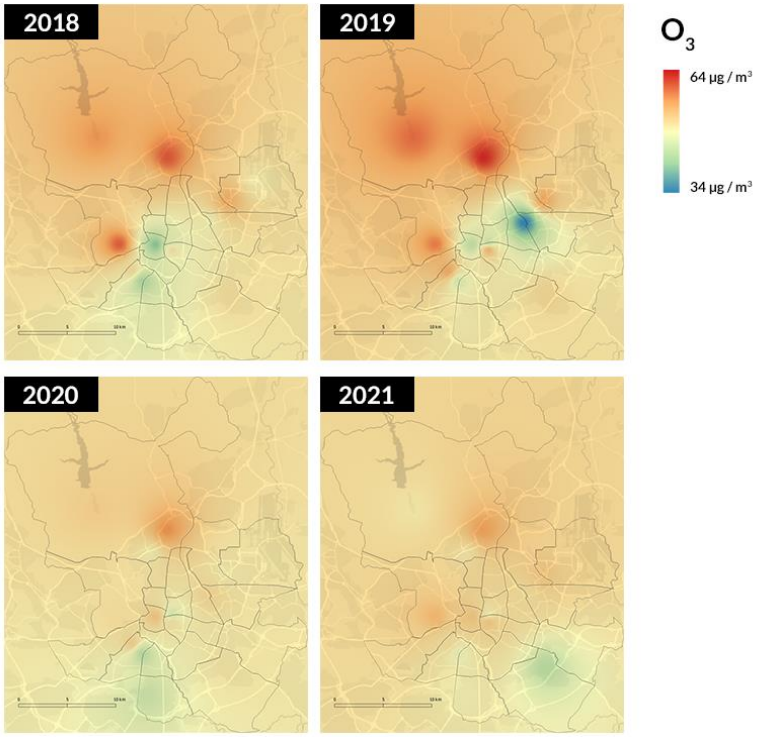


Figura 36. Mapas de niveles de O_3 en Madrid
Fuente. Elaboración propia

Vemos que los niveles de ozono se han mantenido más o menos en los mismos valores durante estos cuatro años incluso parece que mejorando algo en las estaciones periféricas, pero empeorando en el centro de la ciudad.

3.2.6 Estacionariedad de las series

La estacionariedad es una propiedad importante en el análisis de series temporales. Una serie temporal es estacionaria si sus propiedades (media, varianza, autocorrelación…) no cambian con el tiempo.

El test de Dickey-Fuller aumentado nos siver para indicar si una serie temporal univariante es estacionaria. Este test define una hipótesis nula y una hipótesis alternativa:

Hipótesis nula (H_0): La serie temporal tiene una raíz unitaria, lo que significa que no es estacionaria. Tiene alguna estructura dependiente del tiempo.

Hipótesis alternativa (H_1): Se rechaza la hipótesis nula, lo que significa que la serie es estacionaria.

Se ha ejecutado el test para las series de valor medio horario de cada uno de los contaminantes obteniéndose en todos los casos un valor p muy bajo que rechaza la hipótesis nula, es decir, todas las series son estacionarias.

En el apéndice dos se puede ver la gráfica de descomposición de estas series para sus distintas componentes.

3.2.6 Zona de Bajas Emisiones

Un evento que merece la pena analizar con más detalle es la creación de una zona de bajas emisiones en el centro de la ciudad.

La zona de bajas emisiones de Madrid fue inaugurada el 30 de noviembre de 2018 con el nombre de Madrid Central. El periodo que se ha analizado en este trabajo es el periodo de un año desde la entrada en vigor de Madrid Central y el mismo periodo para el año anterior.

Como se muestra en el siguiente mapa, únicamente la estación de control 35 se encuentra dentro del ámbito de la zona de bajas emisiones. Esta estación de control proporciona datos para dióxido de nitrógeno, dióxido de azufre y ozono.

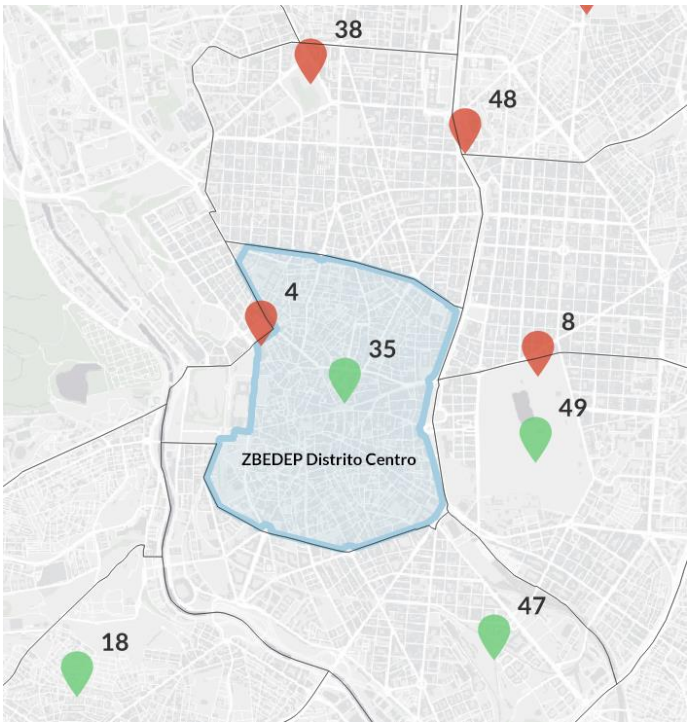


Figura 37. Mapa de Madrid Central con estaciones de control
Fuente. Elaboración propia

La siguiente figura muestra los valores medios de los tres agentes contaminantes en el periodo en el que no había ninguna restricción en comparación con el periodo en el que ya había entrado en vigor la zona de bajas emisiones.

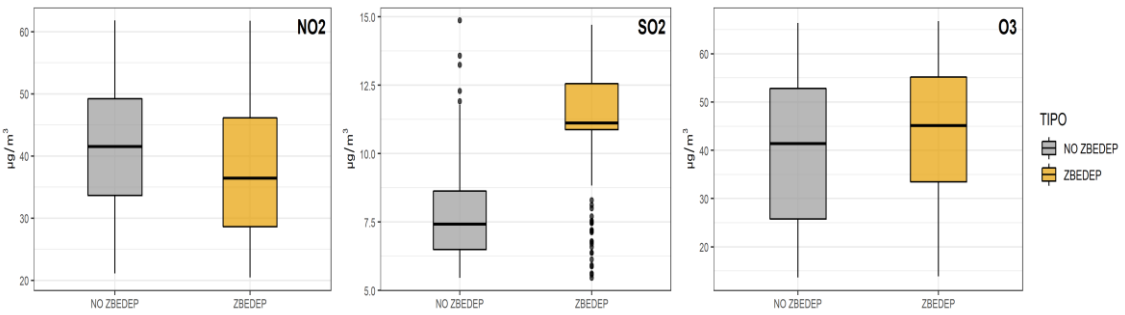


Figura 38. Comparación NO_2, SO_2 y O_3 con y sin zona de bajas emisiones
Fuente. Elaboración propia

En las gráficas se puede ver la esperada bajada de los niveles de dióxido de nitrógeno. Los valores de dióxido de azufre y de ozono, sin embargo, son peores con la zona de bajas emisiones que sin ella, siendo especialmente llamativa el importante aumento de dióxido de azufre.

Si comparamos estos valores con los que obtenemos para estos mismos periodos pero tomando todas las estaciones de control de la ciudad, podemos ver que no sólo en la zona centro se vio un fuerte incremento de dióxido de azufre, sino que fue algo que se experimento en todas las estaciones. El dióxido de nitrógeno también bajo a nivel general, pero no de forma tan marcada como lo hizo en la nueva zona de bajas

emisiones, por lo que en este caso sí que podemos concluir que las restricciones impuestas fueron responsables de la baja de NO_2. Sin embargo, para el ozono tenemos el comportamiento contrario; el aumento que se recoge en la estación de control de la zona centro no se corresponde con la media de toda la ciudad. Se muestra a continuación las gráficas con los valores para todas las estaciones de control en los dos periodos de estudio.

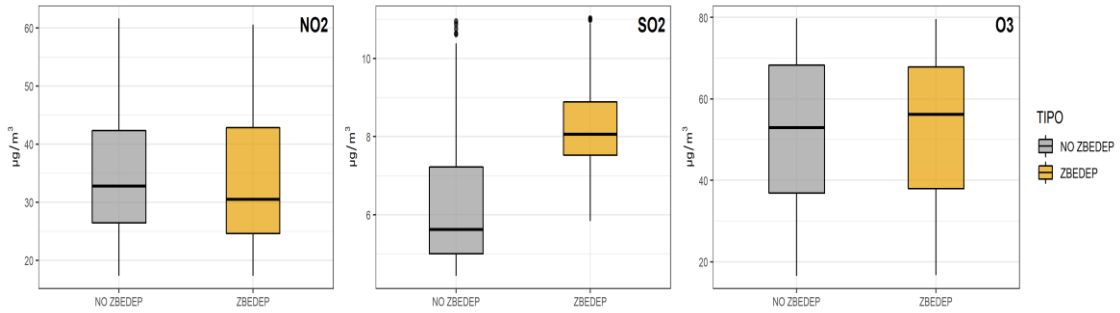


Figura 39. Comparación NO_2, SO_2 y O_3 con y sin zona de bajas emisiones para toda la ciudad
Fuente. Elaboración propia

La siguiente tabla muestra el porcentaje de incremento/decremento de los valores medios de los contaminantes en el periodo anterior a la entrada en vigor de la zona de bajas emisiones con respecto al periodo después de su entrada en vigor. Los valores que se muestran son por un lado únicamente para la estación de control de la zona centro y por el otro para todas las estaciones de control

Contaminante	*Madrid Central*	*Total Madrid*
NO_2	-19,49%	1,86%
SO_2	40,39%	35,82%
O_3	10,04%	-2,21%

Tabla 5. Porcentajes incremento/decremento NO_2, SO_2 y O_3 Zona Bajas Emisiones
Fuente. Elaboración propia.

El descenso en los niveles de NO_2 es el esperado teniendo en cuenta las restricciones de circulación de tráfico rodado, sin embargo, habría que estudiar cuál es el motivo del incremento de O_3. Una posible explicación sería el inicio de la remodelación de Plaza de España que supuso unas obras importantes. Sería necesario poder estudiar la evolución de este agente contaminante en el futuro para tener datos concluyentes.

3.2.7 Confinamiento

El segundo evento interesante surge en el año 2020. Debido a la situación sanitaria origina por la pandemia de COVID-19, el 14 de marzo el gobierno de España declaro el estado de alarma que se implementaría al día siguiente como medida de contención de los contagios. Esto hizo que la actividad y movilidad en la ciudad se redujera drásticamente, lo que dio lugar a un escenario para la investigación sobre el impacto que estas actividades tienen en la calidad del aire.

La siguiente gráfica muestra la evolución de los distintos agentes contaminantes en el periodo comprendido entre el 1 de enero de 2020 y el 25 de mayo de 2020, día en el que comenzaron las primeras medidas de desescalada establecidas por el plan de desconfinamiento en la Comunidad de Madrid. La línea vertical roja marca el 15 de marzo, primer día de confinamiento.

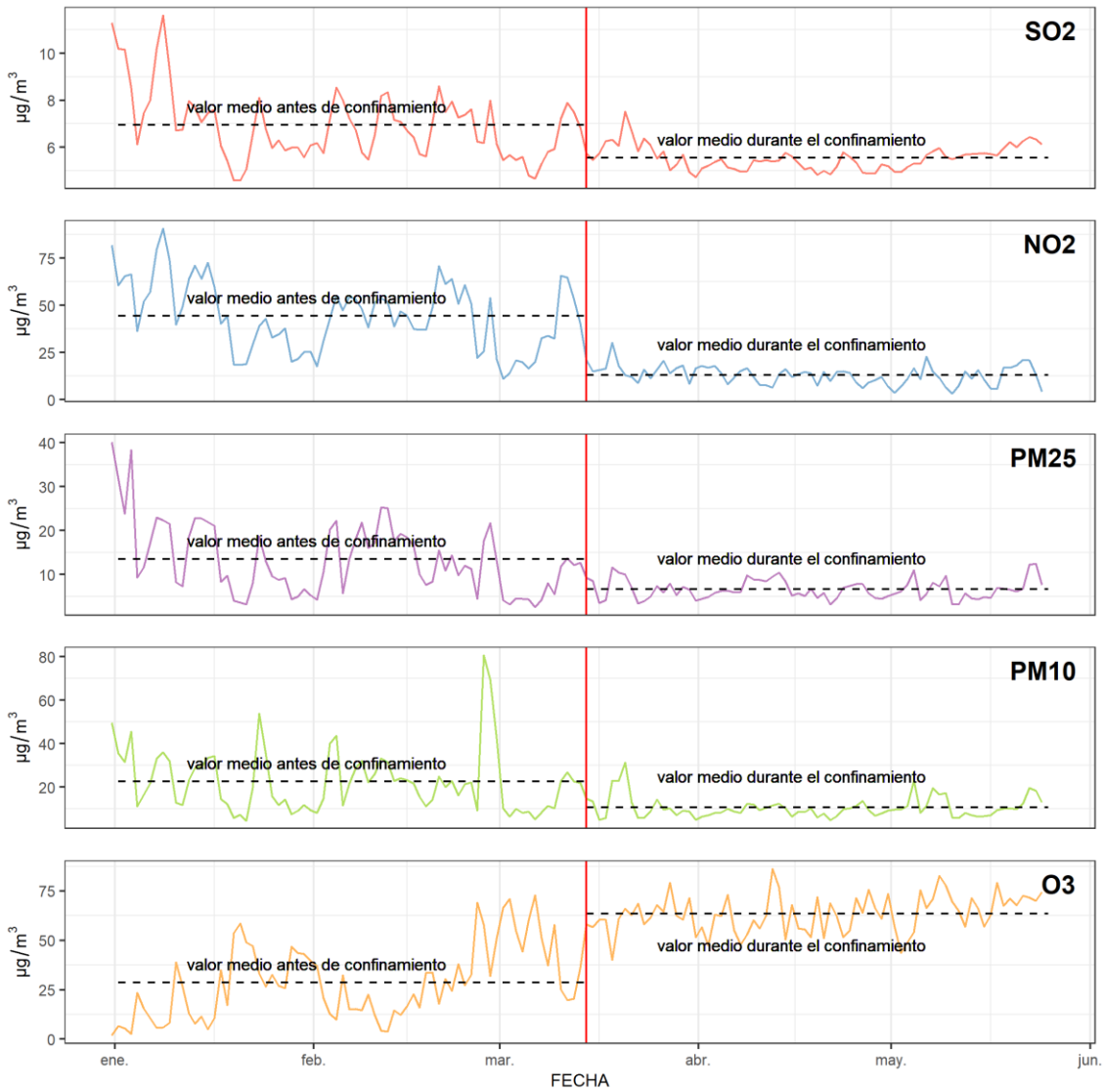


Figura 40. Evolución de valores medios diarios de agentes contaminantes en Madrid (periodo 01 enero 2020 – 25 mayo 2021)
Fuente. Elaboración propia

El descenso en el valor medio de SO_2, NO_2, $PM_{2.5}$ y PM_{10} es notable durante los meses de confinamiento respecto al valor medio de los meses anteriores. Los contaminantes en donde se observa un mayor descenso son SO_2 y especialmente NO_2. Además, se observa que los valores máximos alcanzados durante el confinamiento son también bastante más bajos que los alcanzados en los meses previos.

La gráfica del Ozono, sin embargo, muestra cómo se produjo un aumento de los valores medios registrados, así como de los valores máximos.

Como hemos visto anteriormente, muchos de estos contaminantes se ven afectados por las condiciones climatológicas mostrando un comportamiento estacional, con valores más altos o bajos dependiendo de mes del año en que nos encontramos. Por esto es quizás más interesante comprar los valores durante el confinamiento con los valores medios de esos mismos meses en años anteriores. A continuación, se muestran las gráficas en las que se puede ver esta comparación.

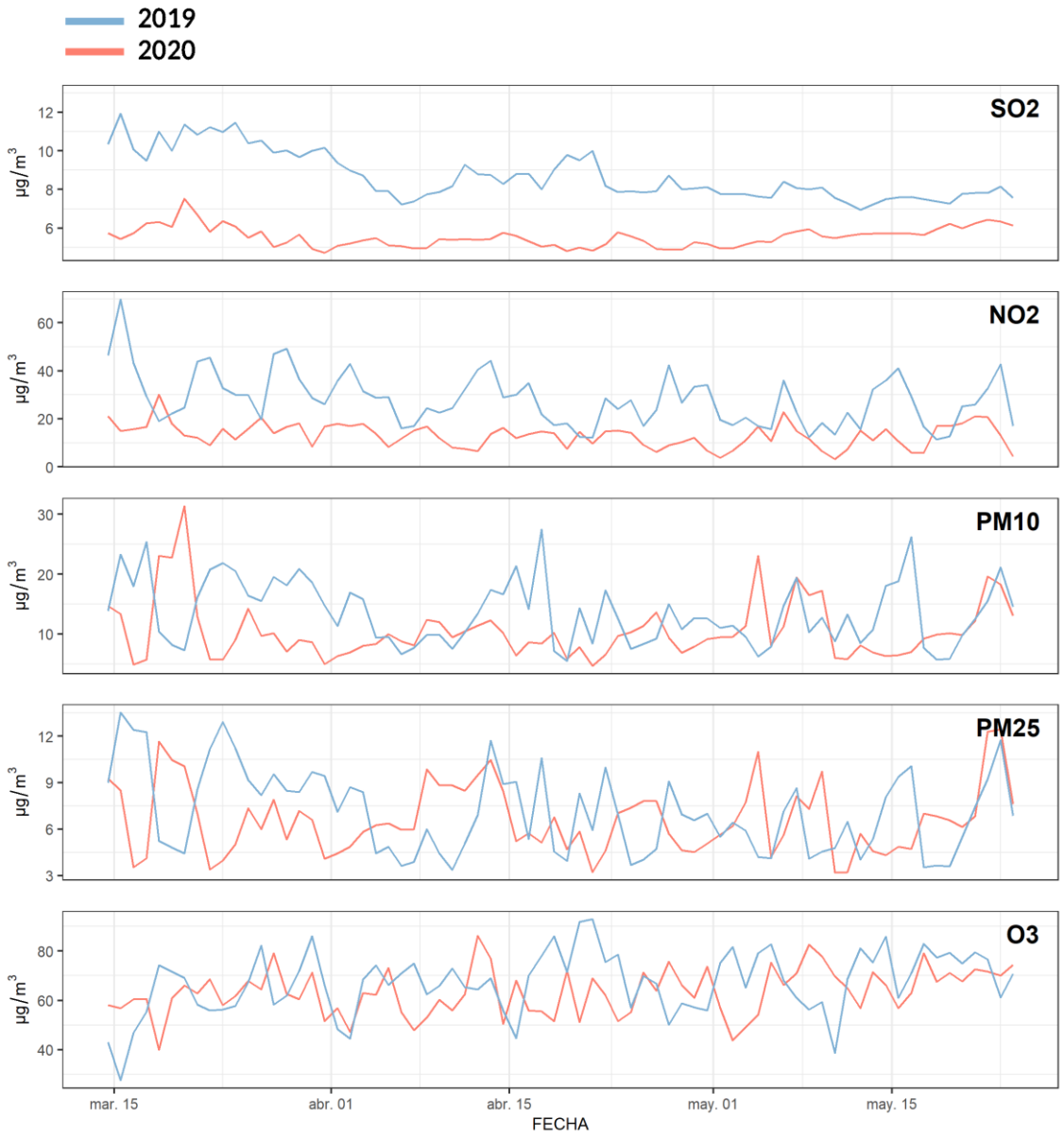


Figura 41. Evolución de valores medios diarios de agentes contaminantes en Madrid (periodo 04 enero 2020 – 25 mayo 2021 en los años 2019 y 2020)
Fuente. Elaboración propia

Vemos que la comparación con el mismo periodo del año anterior sigue dando valores muy inferiores para SO_2 y NO_2, pero el resto de contaminantes son menos claros.

Por último, mostramos los mapas de interpolación de los valores previos al confinamiento en comparación con los valores durante el confinamiento donde se pueden observar estos cambios comentados anteriormente.

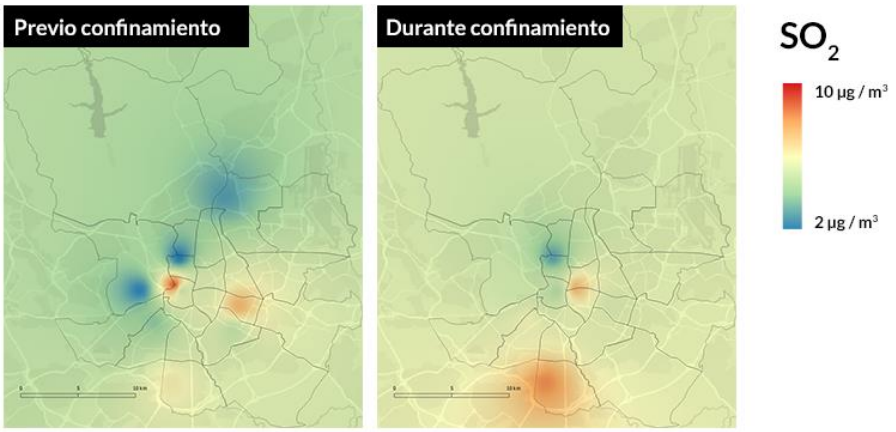


Figura 42. Mapa de niveles de SO_2 previos y durante el confinamiento en Madrid
Fuente. Elaboración propia

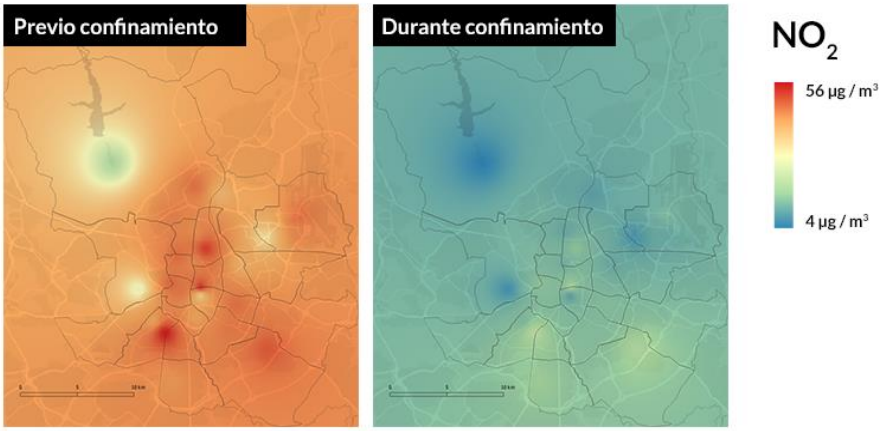


Figura 43. Mapa de niveles de NO_2 previos y durante el confinamiento en Madrid
Fuente. Elaboración propia

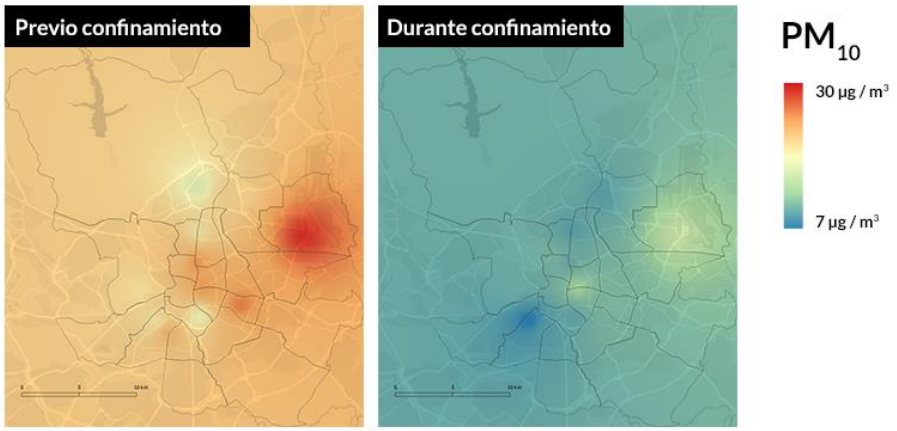


Figura 44. Mapa de niveles de PM_{10} previos y durante el confinamiento en Madrid
Fuente. Elaboración propia

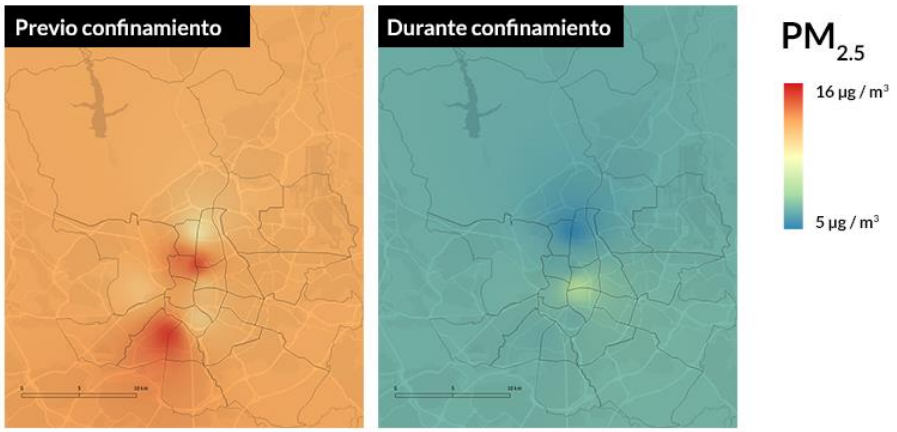


Figura 45. Mapa de niveles de $PM_{2.5}$ previos y durante el confinamiento en Madrid
Fuente. Elaboración propia

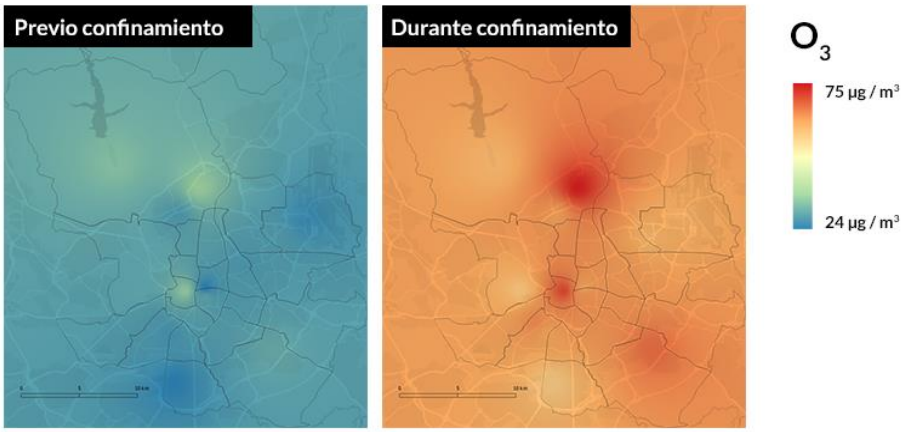


Figura 46. Mapa de niveles de O_3 previos y durante el confinamiento en Madrid
Fuente. Elaboración propia

3.3 Modelos de predicción de calidad del aire

Esta parte del trabajo se centra en la obtención de un modelo de clasificación que permita predecir el valor del Índice de Calidad del Aire en función de datos meteorológicos. Tal y como se vio en el capítulo 2, actualmente las técnicas más utilizadas a este respecto son la combinación de clasificadores y SVM, y dentro de la combinación de clasificadores, random forest es el algoritmo preferido. Por tanto, estas son las dos técnicas que se han estudiado.

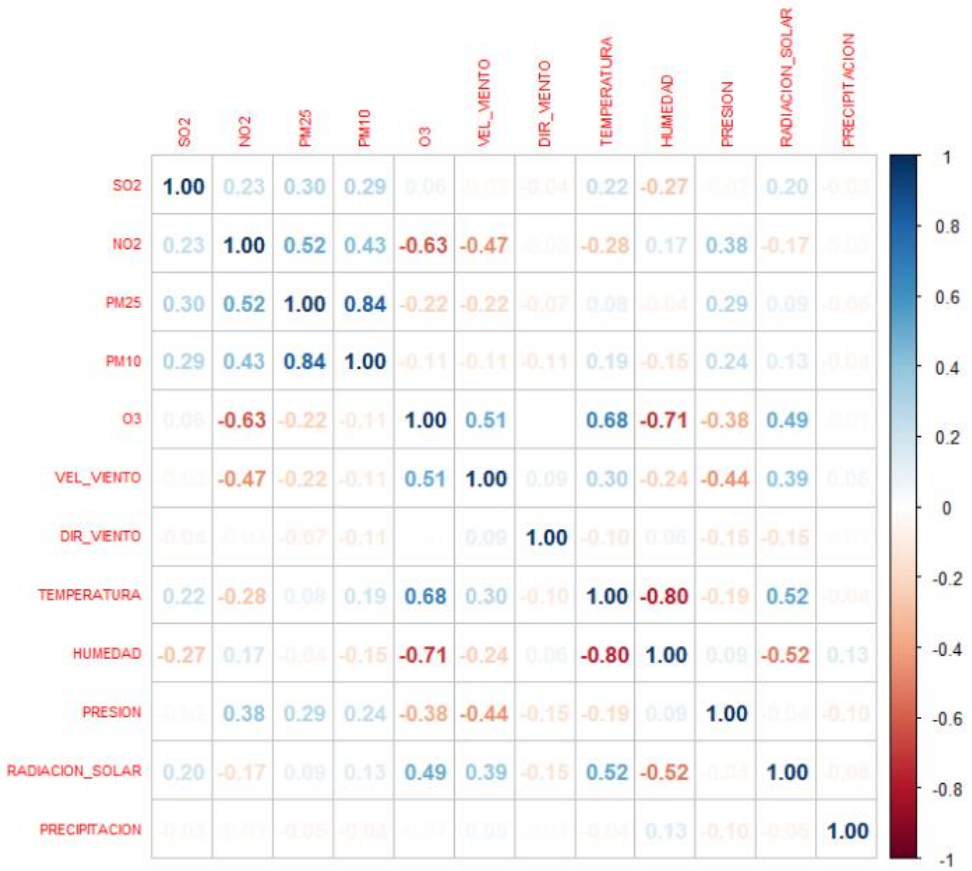
Para calcular el índice de calidad del aire se ha usado el estándar CAQI (Índice Común de Calidad del Aire) desarrollado en el marco del proyecto CITEAIR de la Unión Europea. Para cada uno de los contaminantes que se han analizado durante el trabajo, se establece un índice parcial de forma que el peor valor es el que determina el índice global.

Contaminante	*Muy bueno*	*Bueno*	*Regular*	*Malo*	*Muy malo*
Partículas $PM_{2.5}$	0-15	16-30	31-55	56-110	> 110
Partículas PM_{10}	0-25	26-50	51-90	91-180	> 180
Dióxido de Nitrógeno (NO_2)	0-50	51-100	101-200	201-400	> 400
Ozono (O_3)	0-60	61-120	151-180	181-240	> 240
Dióxido de Azufre (SO_2)	0-50	51-100	101-350	351-500	> 500

Tabla 6. Valores Índice Común de Calidad del Aire CAQI
Fuente. CAQI Air Quality Index [39].

3.3.1 Datos meteorológicos

Para la creación los distintos modelos de clasificación se van a usar los datos meteorológicos disponibles. Antes de generar los modelos, vamos a ver la influencia que los distintos factores tienen sobre cada uno de los agentes contaminante. A continuación, se muestra la matriz de correlación entre las distintas variables calculada con el coeficiente de correlación de Pearson.



De esta tabla vemos que entre las variables meteorológicas existe una correlación inversa entre la temperatura y la humedad. Entre los agentes contaminantes, la correlación más alta es entre las partículas PM_{10} y las $PM_{2.5}$, pero también existe correlación, aunque menor, entre estas dos partículas y el dióxido de nitrógeno. El ozono en la troposfera, como se ha comentado en otros capítulos del trabajo, se forma por las reacciones de otros contaminantes primarios como el dióxido de nitrógeno y los compuestos orgánicos volátiles. Resulta por lo tanto sorprendente que veamos una correlación inversa entre O_3 y NO_2 en esta matriz. Por último, vemos que el O_3 es el agente que más correlación muestra con variables meteorológicas (velocidad del viento, temperatura y radiación solar e inversa con la humedad) y que el NO_2 tiene una correlación inversa con la temperatura.

Vemos a continuación gráficamente algunas de las correlaciones más altas. Las siguientes gráficas muestran los datos de NO_2 respecto a la velocidad del viento y el O_3 con la velocidad del viento, temperatura, radiación solar y humedad.

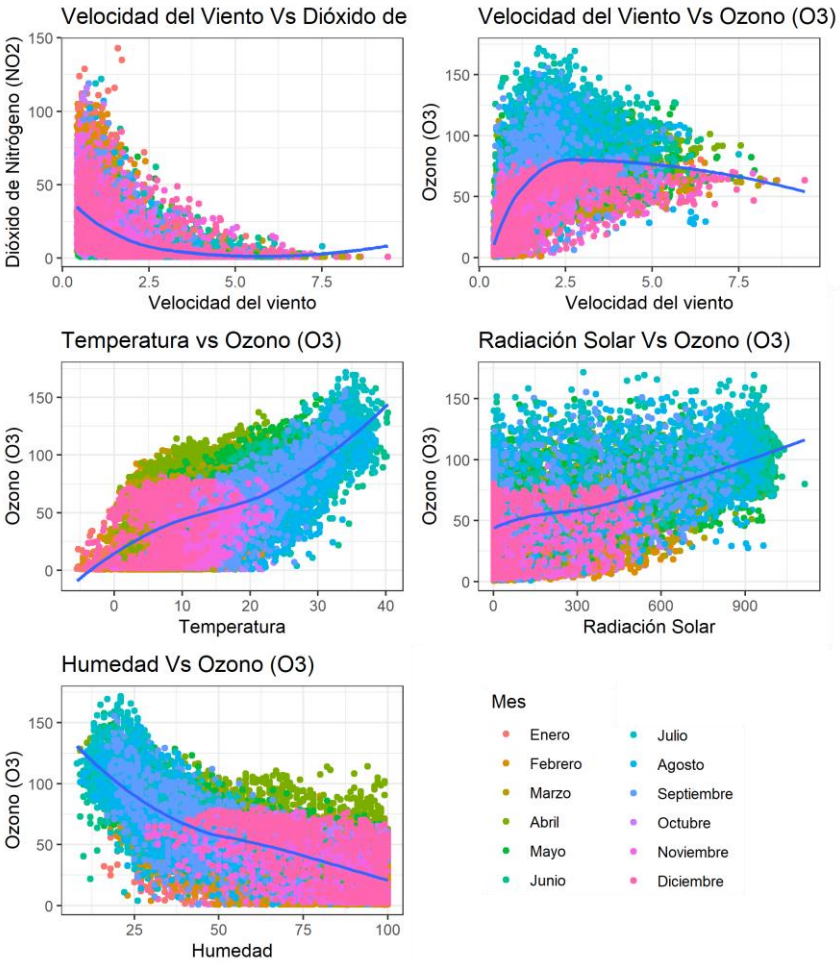


Figura 47. Relación agentes contaminantes – variables meteorológicas
Fuente. Elaboración propia

En relación a la velocidad y dirección del viento, la librería openair de R nos permite hacer un análisis relacionando estas variables con los distintos agentes contaminantes. En primer lugar, vemos que la dirección del viento en Madrid está dominada por la componente suroeste que es además la que tiene más porcentaje de vientos fuertes. Se puede observar en la gráfica de tipo rosa de los vientos.

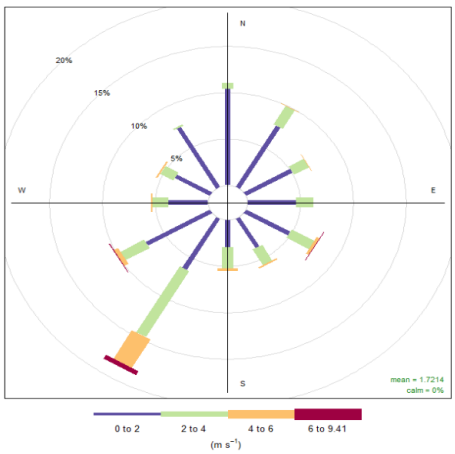


Figura 48. Rosa de los vientos Madrid

Fuente. Elaboración propia

Sin embargo, para las partículas PM_{10} y $PM_{2.5}$ vemos que, con vientos de dirección sudeste, el porcentaje de altas concentraciones de ambas partículas es mayor. Se puede ver en las siguientes gráficas.

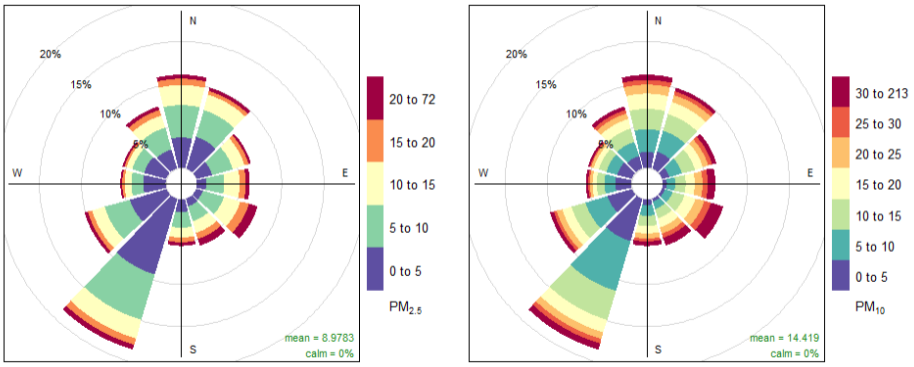


Figura 49. Rosa de contaminación $PM_{2.5}$ y PM_{10} Madrid
Fuente. Elaboración propia

Otra forma de verlo es mostrando los percentiles en lugar de porcentajes. Se muestra en las gráficas siguientes.

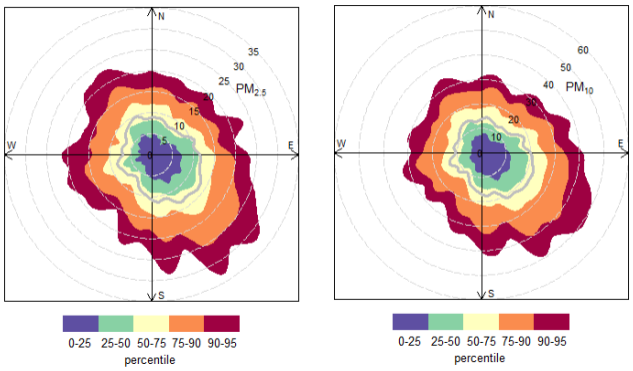


Figura 50. Rosa de percentiles de contaminación $PM_{2.5}$ y PM_{10} Madrid
Fuente. Elaboración propia

3.3.2 Preparación de los datos. Ventana deslizante

La organización de nuestro dataset ahora mismo nos permite plantear un problema de aprendizaje supervisado por el cual tratemos de predecir del ICA en base a los datos meteorológicos en un momento determinado, el tipo de estación y la hora. Sin embargo, por la naturaleza de nuestro problema, tiene más sentido tener en cuenta los datos meteorológicos de las últimas horas, ya que el efecto de los mismo se nota a partir de un determinado tiempo y no de forma inmediata. Por ello, se han creado nuevas variables para recoger el valor medio de las últimas 6, 24 y 48 horas para cada una de las variables meteorológicas.

Además de los datos meteorológicos, también se ha introducido la estación, el tipo de estación y la hora como variables para la predicción, ya que se vio que pueden influir en el valor de los distintos agentes contaminantes.

Por último, en lugar de crear un modelo de clasificación para intentar predecir el ICA del momento, lo realmente interesante sería ver si podemos predecir la calidad del aire a futuro. Para ello, se han creado dos variables nuevas que recogen el valor del ICA en el momento de tiempo t + 6 horas y otra para t + 24 horas, es decir, queremos ver si nuestro modelo puede predecir el ICA que tendremos dentro de 6 horas y dentro de un día.

Cada observación quedará por tanto de la siguiente manera:

DATOS METEO T -48 HORAS	DATOS METEO T – 24 HORAS	DATOS METEO T -6 HORAS	DATOS METEO T	TIPO ESTACIÓN	HORA	ICA T	ICA T + 6	ICA T + 24

3.3.2 Random Forest

Random forest es una técnica de minería de datos que consiste en la combinación de árboles de decisión utilizando la técnica de bagging. La idea básica del bagging es utilizar el conjunto de entrenamiento original para generar un gran número de conjuntos similares usando muestreo con remplazo, es decir, de un conjunto de N elementos se escogen N' elementos al azar, siendo N' <= N y con la posibilidad de escoger un mismo elemento más de una vez. Durante el proceso de generación del modelo se construye cada árbol de decisión con un subconjunto de las variables del conjunto de datos de manera que se tienen en cuenta variables que normalmente quedarían eclipsadas por otra más relevantes.

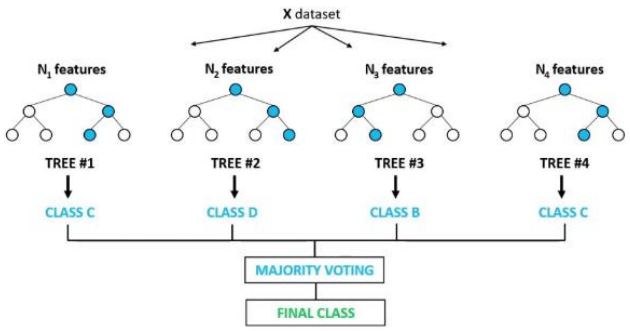


Figura 51. Esquema algoritmo Random Forest
Fuente. freeCodeCamp.org

Una forma de medir el error del modelo random forest es mediante el error out-of-bag, que consiste en estimar el error como el promedio de todos los errores parciales cometidos por cada árbol de decisión usando como conjunto de pruebas todos aquellos elementos del dataset original que no han sido elegidos en el conjunto de entrenamientos durante el muestreo con remplazamiento.

El paquete ranger de R ofrece una implementación rápida de random forest [36].

Ajuste del modelo y optimización de hiperparámetros.

Los siguientes hiperparámetros han sido los que se han modificado para el ajuste del modelo:

- Num_trees: máximo número de árboles que se van a generar.
- mtry: número de variables que se escogen aleatoriamente cada vez que se divide un nodo.
- Max_depth: profundidad máxima que tienen los árboles.

Se ha creado una matriz con los siguientes valores para cada uno de estos parámetros.

Parámetro			
num_trees	50	100	200
mtry	3	5	7
max_depth	10	20	0 (sin límite)

Tabla 7. Valores hiperparámetros random forest
Fuente. Elaboración propia.

Con estas combinaciones se ha entrenado los 4 juegos de datos que hemos creado (datos meteorológicos del momento, media 6 horas anterior, media 24 horas anterior y media 48 horas anterior). A continuación, se muestran las tablas con los 5 mejores resultados para cada dataset de entrenamiento.

Datos meteorológicos actuales			
num_trees	***mtry***	***max_depth***	***oob_error***
200	3	20	0.1389557
200	3	0	0.1390045
200	5	20	0.1399148
200	5	0	0.1399555
200	7	20	0.1402724

Ventana deslizante de 6 horas			
num_trees	***mtry***	***max_depth***	***oob_error***
200	3	0	0.1114100
200	5	0	0.1129056
100	3	0	0.1133526
200	7	0	0.1136534
200	3	20	0.1141979

Ventana deslizante de 24 horas			
num_trees	***mtry***	***max_depth***	***oob_error***
200	3	0	0.09091944
100	3	0	0.09223617
200	5	0	0.09237434
100	5	0	0.09390240
200	7	0	0.09391053

Ventana deslizante de 48 horas			
num_trees	***mtry***	***max_depth***	***oob_error***
200	3	0	0.09011477
200	5	0	0.09090318
100	3	0	0.09191105
200	7	0	0.09265069
100	5	0	0.09272384

Tabla 8. Mejores resultados random forest por dataset
Fuente. Elaboración propia.

Vemos que la mejora realmente importante la obtenemos cuando pasamos de usar datos únicamente del momento y usamos datos de las últimas 6 horas o datos de las últimas 24. Con datos de las últimas 48 horas se mejora aún más el modelo, pero la diferencia no es apenas apreciable.

Los mejores hiperparámetros obtenidos son para el dataset de 48 horas con num_trees = 200, mtry = 3 y max_depth = 0. Entrenando un modelo con estos hiperparámetros podemos ver la importancia que tienen las distintas variables:

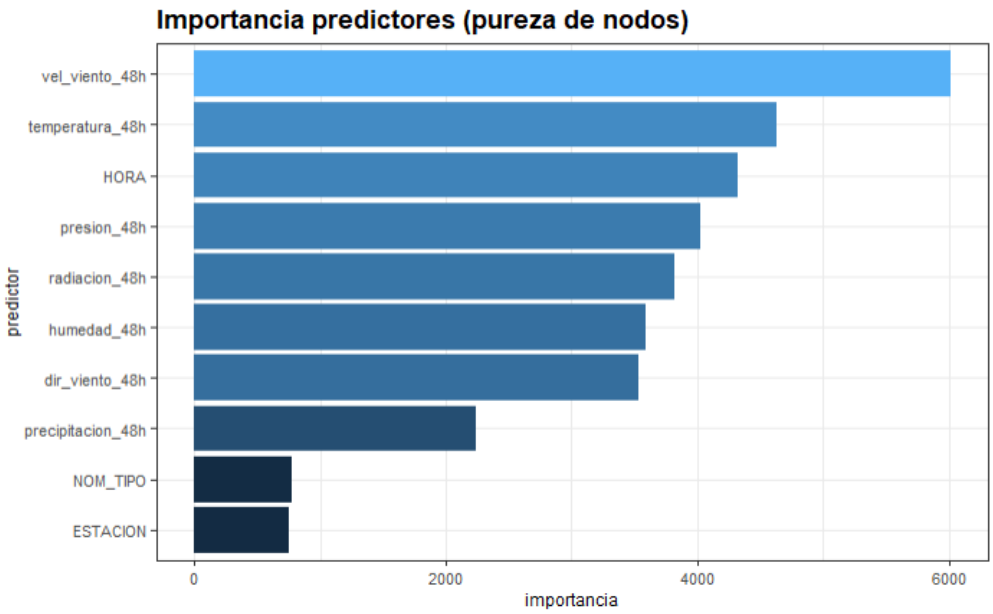


Figura 52. Importancia de las variables en el modelo random forest
Fuente. Elaboración propia

Vemos que la velocidad del viento y la temperatura son las variables más importantes mientras que la estación o el tipo de estación no tiene apenas importancia.

Forecast 6 horas.

Como comentamos anteriormente, lo realmente interesante sería ver si podemos predecir el valor de ICA para un tiempo futuro. En un primer lugar se ha buscado un modelo capaz de predecir el valor del índice a las 6 horas. Para ello, hemos usado de nuevo los cuatro juegos de datos del apartado anterior. Los resultados se muestran en las siguientes tablas.

Datos meteorológicos actuales			
num_trees	***mtry***	***max_depth***	***oob_error***
200	5	0	0.2735519
200	3	0	0.2743729
200	5	20	0.2756247
200	7	0	0.2756490
200	7	20	0.2769658

Ventana deslizante de 6 horas			
num_trees	***mtry***	***max_depth***	***oob_error***
200	5	0	0.2034367
200	7	0	0.2036480
200	3	0	0.2046478
100	5	0	0.2084275
100	7	0	0.2091021

Ventana deslizante de 24 horas			
num_trees	*mtry*	*max_depth*	*oob_error*
200	5	0	0.1235836
200	7	0	0.1239494
200	3	0	0.1241282
100	7	0	0.1258027
100	5	0	0.1259165

Ventana deslizante de 48 horas			
num_trees	*mtry*	*max_depth*	*oob_error*
200	5	0	0.1212670
200	7	0	0.1221937
200	3	0	0.1225188
100	7	0	0.1233804
100	5	0	0.1234617

Tabla 9. Mejores resultados random forest forecast 6 horas por dataset
Fuente. Elaboración propia.

Los resultados obtenidos son algo peores que el anterior caso, sin embargo, utilizando los datos meteorológicos de las últimas 24 y 48 horas obtenemos resultados bastante buenos, con el mejor out-of-bag error de 0.123 y 0.121 respectivamente.

Forecast 24 horas.

Por último se muestran los resultados obtenidos al intentar predecir el índice de calidad del aire del día siguiente.

Datos meteorológicos actuales			
num_trees	*mtry*	*max_depth*	*oob_error*
200	5	0	0.2865830
200	7	0	0.2871683
200	3	0	0.2889893
200	5	20	0.2899649
200	7	20	0.2899811

Ventana deslizante de 6 horas			
num_trees	*mtry*	*max_depth*	*oob_error*
200	7	0	0.2109619
200	5	0	0.2125715
100	7	0	0.2159453
200	3	0	0.2164737
100	5	0	0.2175387

Ventana deslizante de 24 horas			
num_trees	*mtry*	*max_depth*	*oob_error*
200	5	0	0.1223660
200	3	0	0.1233822
200	7	0	0.1243984
100	5	0	0.1256504
100	3	0	0.1258292

Ventana deslizante de 48 horas			
num_trees	*mtry*	*max_depth*	*oob_error*
200	5	0	0.1205287
200	3	0	0.1207970
200	7	0	0.1212604
100	5	0	0.1220652
100	3	0	0.1230652

Tabla 10. Mejores resultados random forest forecast 6 horas por dataset
Fuente. Elaboración propia.

De nuevo, obtenemos resultados algo peores pero bastante buenos usando los datasets de datos meteorológicos de 24 y 48 horas con un out-of-bag error en el mejor de los casos de 0.120.

3.3.3 SVM

Las máquinas de soporte vectorial son un algoritmo de aprendizaje supervisado utilizado en problemas de clasificación tanto lineales como no lineales. Se fundamenta en el margen y el hiperplano de separación. Un *hiperplano* es un subespacio cuya dimensión es uno menos que su espacio ambiental. Por lo tanto, en un espacio p-dimensional, un hiperplano será un subespacio plano de dimensiones p – 1. El *margen* es la zona de separación entre los distintos grupos a clasificar delimitada por hiperplanos. El objetivo de las SVM es encontrar el hiperplano

separador óptimo que maximice el margen del juego de datos de entrenamiento [38].

Ajuste del modelo y optimización de hiperparámetros.

Debido a los largos tiempos de ejecución, únicamente se ha utilizado el dataset con datos meteorológicos de las últimas 48 horas. Los hiperparámetros que se han intentado optimizar son:

- C: el coste indica la penalización que tienen los errores en la clasificación.
- gamma: coeficiente que multiplica la distancia entre dos puntos en el kernel radial. Cuanto más pequeño es, más influencia tienen dos puntos cercanos.

Los valores que se han probado son los siguientes:

Parámetro			
C	0,01	1	5
gamma	0,5	2	5

Tabla 11. Hiperparámetros SVM
Fuente. Elaboración propia.

Además, debido a los requisitos de capacidad computacional para la ejecución del algoritmo, únicamente se ha probado con un dataset de 24583 observaciones. El kernel utilizado es el radial.

Los resultados obtenidos se muestran en la siguiente gráfica:

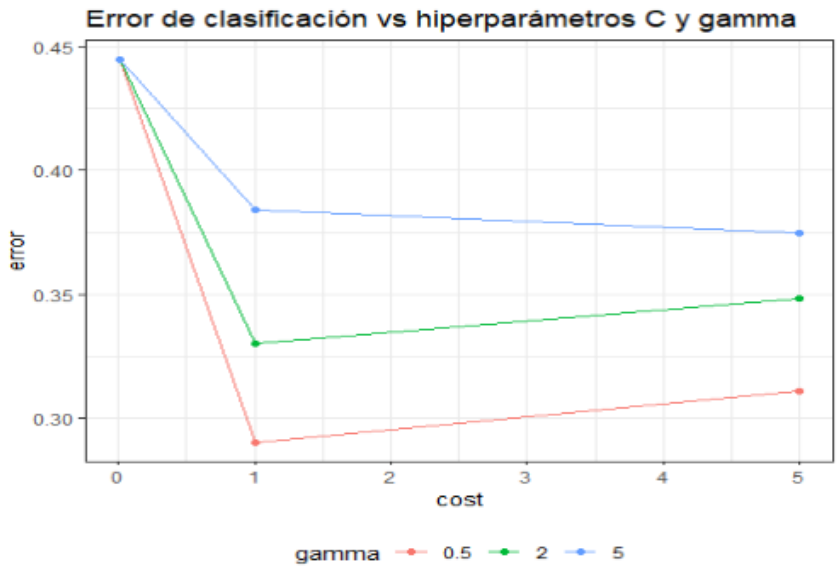


Figura 53. Error de clasificación vs hiperparámetros C y gamma – entrenamiento SVM
Fuente. Elaboración propia

Y en la tabla podemos ver los valores de los 5 mejores resultados obtenidos:

C	*gamma*	*error*	*dispersion*
1	0,5	0,2900791	0.010385546
5	0,5	0.3109467	0.007787320
1	2	0.3299436	0.007882211
5	2	0.3481264	0.008171977
5	5	0.3746493	0.005769755

Tabla 12. Mejores resultados SVM
Fuente. Elaboración propia.

Los resultados obtenidos son mucho peores que los de random forest. El ajuste con otros hiperparámetros no devuelve resultados mejores, por lo que sería necesario probar a encontrar otro kernel o intentar ejecución con un mayor número de observaciones lo cual requeriría mucho tiempo de computación.

Ejecutando el modelo sobre el conjunto de datos de prueba obtenemos un error del 27,34%, algo mejor que el error de entrenamiento, pero de nuevo mucho peor que los obtenidos con random forest.

4. Conclusiones

El proyecto se centra en el análisis de los datos de calidad del aire en la ciudad de Madrid en los últimos 20 años y en concreto, en los cinco contaminantes más relevantes y que son los utilizados para el cálculo de índice de calidad del aire: dióxido de azufre, dióxido de nitrógeno, partículas PM_{10} y $PM_{2.5}$ y el ozono troposférico. Durante el desarrollo del trabajo se ha podido comprobar la tendencia a la baja de la mayoría de los contaminantes. El ozono es el único contaminante que muestra una tendencia al alza lo que hace que, aunque todavía no llegue a superar los límites máximos sugeridos por la Organización Mundial de la Salud en su guía sobre calidad del aire, se esté aproximando.

Se han estudiado dos episodios interesantes que han influido en la calidad del aire. El primero es la creación de una zona de bajas emisiones en el distrito centro de la ciudad. Hemos visto que las restricciones al tráfico rodado han supuesto un descenso de los niveles de dióxido de nitrógeno, sin embargo, los otros dos contaminantes recogidos por la estación de control situada en esta zona, que son el dióxido de azufre y el ozono, registran valores superiores a los recogidos antes de la creación de la zona de bajas emisiones.

El segundo evento es el periodo de confinamiento vivido durante varios meses en el año 2020 debido a la situación sanitaria originada por la pandemia de COVID-19. Esta situación supuso un descenso de la movilidad en toda la ciudad que permiten estudiar la influencia del tráfico rodado en la calidad del aire. Hemos visto que, efectivamente, los niveles de NO_2, SO_2, PM_{10} y $PM_{2.5}$ bajaron, siendo especialmente significativo los descensos de NO_2.

Además del análisis de los datos haciendo uso de R, se ha realizado un análisis geográfico. Para ello, se han generado distintos mapas a través de la interpolación de los datos recogidos por las distintas estaciones de control, lo que ha permitido ver cómo se distribuyen los distintos agentes contaminantes en las distintas regiones de la ciudad. En general, se observa como las áreas con mayor tráfico son las que cuentan con mayores niveles de NO_2 mientras que las zonas suburbanas son las que tienen mayores niveles de O_3.

Por último, se han generado varios modelos de predicción de calidad del aire haciendo uso de técnicas de aprendizaje automático. En concreto, se han utilizado dos algoritmos que son los que se vieron en el capítulo 2 que eran los más empleados en los últimos estudios realizados: random forest y SVM. Los resultados con random forest han sido bastante buenos, sin embargo, para SVM, los mejores modelos obtenidos tienen una tasa de error del 27%.

Debido al tiempo limitado para la realización del proyecto, hay una serie de líneas que se podrían plantear como trabajo futuro:

- Una de las cosas que me ha sorprendido durante el análisis de los datos de calidad del aire es que tras la creación de la zona de bajas emisiones y durante el confinamiento, los niveles de ozono subieron, siendo el único agente contaminante que mostro este comportamiento (todos los demás contaminantes estudiados bajaron significativamente). El ozono se genera principalmente por las reacciones química de óxidos de nitrógeno (NO_x) y por los Compuestos Orgánicos Volátiles (COV). Por esto, me parece sorprendente que bajando los niveles de NO_2 los niveles de O_3 sin embargo suban. Este es un punto en el que me hubiera gustado profundizar más.

- QGIS es un Sistema de Información Geográfica de software libre y código abierto que por lo que he podido experimentar cuenta con una gran capacidad y un gran número de funcionalidades. Sin embargo, no he encontrado una forma sencilla de generar un cuadro de mando online a través del mapa generado en QGIS. Creo que esto hubiera sido más sencillo de haber utilizado otras herramientas de pago como ArcGIS o CARTO, sin embargo, las limitaciones de las cuentas de prueba de estas dos herramientas has hecho que finalmente haya desarrollado el trabajo únicamente con QGIS.

- Gran parte del tiempo empleado en el trabajo se ha empleado en la limpieza y análisis de los datos, dejándome finalmente poco tiempo para profundizar en los modelos de predicción. Los resultados obtenidos con SVM han sido bastante malos y creo que se podrían mejorar utilizando un conjunto de datos de entrenamiento más grande. Sin embargo, el tiempo que tarda el entrenamiento de los modelos de SVM ha hecho que finalmente el modelo presentado en este trabajo tenga una tasa de error bastante elevada. Además, me hubiera gustado haber utilizado alguna otra técnica, como redes neuronales.

- Durante el proyecto he querido utilizar también datos de tráfico ya que es uno de los factores principales de contaminación. Sin embargo, no he encontrado ninguna fuente de datos buena que me permitiera tener estos datos horarios por tramos.

- Por último, creo que hubiera sido interesante realizar un estudio similar comparando distintas ciudades con condiciones climatológicas con comportamientos de movilidad distintos.

A nivel personal, el trabajo me ha servido para profundizar en un tema que me resulta muy interesante como es el de la calidad del aire. Viéndolo en perspectiva, como decía antes, creo que hubiera sido incluso más interesante si no me hubiera centrado únicamente en mi ciudad y en lugar de eso hubiera sido un estudio a nivel más global. Además he podido poner en práctica algunos de los conocimientos adquiridos durante el máster, principalmente la limpieza y preparación de grandes conjuntos de datos, el análisis estadístico y geográfico de estos datos y la utilización de algoritmos de aprendizaje automático supervisados para la generación de modelos de clasificación.

5. Glosario

OMS	Organización Mundial de la Salud
SO_2	Dióxido de azufre
NO_2	Dióxido de nitrógeno
PM_{10}	Partículas de diámetro menor a 10μg
$PM_{2.5}$	Partículas de diámetro menor a 2.5μg
O_3	Ozono
COV	Compuestos Orgánicos Volátiles
COVDM	Compuestos Orgánicos Volátiles de Metano
ICA (AQI)	Índice de Calidad del Aire (Air Quality Index)
CAQI	Índice de Calidad del Aire Común (Common Air Quality Index)
SIG	Sistema de Información Geográfica
IDW	Distancia Inversa Ponderada (Inverse Distance Weighting)
AEMET	Agencia Española de Meteorología
SVM	Support Vector Machines

6. Bibliografía

[1] Contaminación del aire ambiente (exterior) – Organización Mundial de la Salud (OMS). Visitado el 25/02/2022. https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] Carta de servicios de la calidad del aire - Portal Web del Ayuntamiento de Madrid. Visitado el 25/02/2022. https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Calidad-y-Evaluacion/Modelos-de-Calidad-y-Excelencia/Cartas-de-Servicios/Cartas-de-Servicios-vigentes/Carta-de-Servicios-de-Calidad-del-Aire/?vgnextfmt=default&vgnextoid=1a698ae7017bb610VgnVCM2000001f4a900aRCRD&vgnextchannel=5ef34225faf07410VgnVCM2000000c205a0aRCRD

[3] Protocolos de Actuación para Episodios de Contaminación – Portal Web de Calidad del Aire del Ayuntamiento de Madrid. Visitado el 25/02/2022. http://www.mambiente.munimadrid.es/opencms/calaire/Episodios/ProtocolosActuacion/

[4] Memoria 2020 de calidad del aire - Ayuntamiento de Madrid. http://www.mambiente.munimadrid.es/opencms/export/sites/default/calaire/Anexos/Memorias/MEMORIA_2020.pdf

[5] Contaminación atmosférica - IDEAM - Instituto de Hidrología, Meteorología y Estudios Ambientales del Gobierno de Colombia. Visitado el 25/02/2022. http://www.ideam.gov.co/web/contaminacion-y-calidad-ambiental/contaminacion-atmosferica#:~:text=La%20contaminaci%C3%B3n%20atmosf%C3%A9rica%20es%20la,encuentran%20expuestas%20a%20dicho%20ambiente.

[6] Gallego Picó, Alejandrina. "La contaminación atmosférica". *Contaminación atmosférica,* editado por UNED - Universidad Nacional de Educación a Distancia

[7] SEÑALES DE LA AEMA 2020 - Hacia una contaminación cero en Europa – Agencia Europea de Medio Ambiente (AEMA). https://www.eea.europa.eu/es/publications/senales-de-la-aema-2020

[8] Consulta pública de calidad del aire – Comisión Europea. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12677-Calidad-del-aire-revision-de-las-normas-de-la-UE/public-consultation_es

[9] CAQI Air Quality Index – European Union. https://www.airqualitynow.eu/download/CITEAIR-Comparing_Urban_Air_Quality_across_Borders.pdf

[10] Protocolo de actuación para episodios de contaminación por Dióxido de Nitrógeno en la ciudad de Madrid - Ayuntamiento de Madrid.

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

[11] Protocolo de actuación para episodios de contaminación por Ozono en la ciudad de Madrid - Ayuntamiento de Madrid. https://www.mambiente.madrid.es/opencms/export/sites/default/calaire/Anexos/Procedimiento_ozono.pdf

[12] J. Rodríguez Lloret; R. Olivella; A. Muñoz Bollas; V. Velarde Gutiérrez. Introducción a los sistemas de información geográfica

[13] Aire Quality Guidelines for Europe – World Health Organization. https://apps.who.int/iris/handle/10665/107364

[14] Bert Brunekreef, Douglas W. Dockery, and Michal Krzyzanowski, 1995 - Epidemiologic Studies on Short-Term Effects of Low Levels of Major Ambient Air Pollution Components

[15] Bert Brunekreef, Stephen T Holgate, 2002 - Air pollution and health

[16] C. Arden Pope III, 2000 - Epidemiology of Fine Particulate Air Pollution and Human Health: Biologic Mechanisms and Who's at Risk?

[17] Normativa Calidad del Aire – Ministerio para la Transición Ecológica y el Medio Ambiente. Visitado el 03/03/2022. https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/normativa/

[18] Air Quality – Existing Legislation, Comisión Europea. Visitado el 03/03/2022. https://ec.europa.eu/environment/air/quality/existing_leg.htm

[19] GIS Overview – ESRI. Visitado el 05/03/2022. https://www.esri.com/en-us/what-is-gis/overview

[20] European Air Quality Index – Agencia Europea de Medio Ambiente. Visitado el 04/03/2022. https://airindex.eea.europa.eu/Map/AQI/Viewer/

[21] Visor de Calidad del Aire – Ministerio para la Transición Ecológica y el Reto Demográfico. Visitado el 04/03/2022. https://sig.mapama.gob.es/calidad-aire/

[22] Calidad del Aire – Comunidad de Madrid. Visitado el 04/03/2022. https://idem.madrid.org/visor/?v=calidadaire&ZONE=430000,4485000,8

[23] Consulta de datos del mapa de red de vigilancia – Ayuntamiento de Madrid. Visitado el 04/03/2022. http://www.mambiente.munimadrid.es/sica/scripts/index.php

[24] Steen Solvang Jensen, Ruwim Berkowicz, Henning Sten Hansen, Ole Hertel. A Danish decision-support GIS tool for management of urban
air quality and human exposures (2001).

[25] CALIOPE-Urban v1.0: coupling R-LINE with a mesoscale air quality modelling system for urban air quality forecasts over Barcelona city (Spain) - Jaime Benavides, Michelle Snyder, Marc Guevara, Albert Soret, Carlos Pérez García-Pando, Fulvio Amato, Xavier Querol, and Oriol Jorba. 2019.

[26] David J. Briggs, Susan Collins, Paul Elliott, Paul Fischer, Simon Kingham, Erik Lebret, Karel Pryl, Hans Van Reeuwijk, Kirsty Smallbone and Andre Van Der Veen. Mapping urban air pollution using GIS: a regression-based approach. (1997)

[27] Michael Jerrett, Richard T Burnett, Pavlos Kanaroglou, John Eyles, Norm Finkelstein, Chris Giovis, Jeffrey R Brook. A GIS ^ environmental justice analysis of particulate air pollution in Hamilton, Canada. (2001).

[28] Air Quality Dispersion Modelling – United States Environmental Protection Agency. Visitado el 05/03/2022. https://www.epa.gov/scram/air-quality-dispersion-modeling

[29] A. Masih - Machine learning algorithms in air quality modelling. (2019)

[30] Joanna A. Kaminska - The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław (2018)

[31] Bekkar, A., Hssina, B., Douzi, S. et al. - Air-pollution prediction in smart city, deep learning approach. J Big Data 8, 161 (2021)

[32] Using models for air quality assessment and planning: a guide – Agencia Europea de Medio Ambiente. Visitado el 05/03/2022.
https://www.eea.europa.eu/highlights/using-models-for-air-quality

[33] La modelización de la contaminación atmosférica urban – CIEMAT. Visitado el 05/03/2022.
https://www.ciemat.es/cargarAplicacionNoticias.do;jsessionid=08E7E81F3331B4EEB8F17934DC2EDD3A?idArea=-1&identificador=2020

[34] Calidad del aire/Modelización – CIEMAT. Visitado e 05/03/2022. http://retos-aire.ciemat.es/-impactos-en-calidad-del-aire

[35] Memoria anual de calidad del aire 2021.
https://airedemadrid.madrid.es/UnidadesDescentralizadas/Sostenibilidad/CalidadAire/Publicaciones/Memorias_anuales/Ficheros/MEMORIA_2021.pdf

[36] Breiman L. - Random forest (2001).

[37] CAQI Air quality index. https://www.airqualitynow.eu/download/CITEAIR-Comparing_Urban_Air_Quality_across_Borders.pdf

[38] Jordi Gironés, Jordi Casas, Julià Minguillón, Ramón Caihuelas. Minería de datos. Modelos y Algoritmos (2017).

7. Anexos

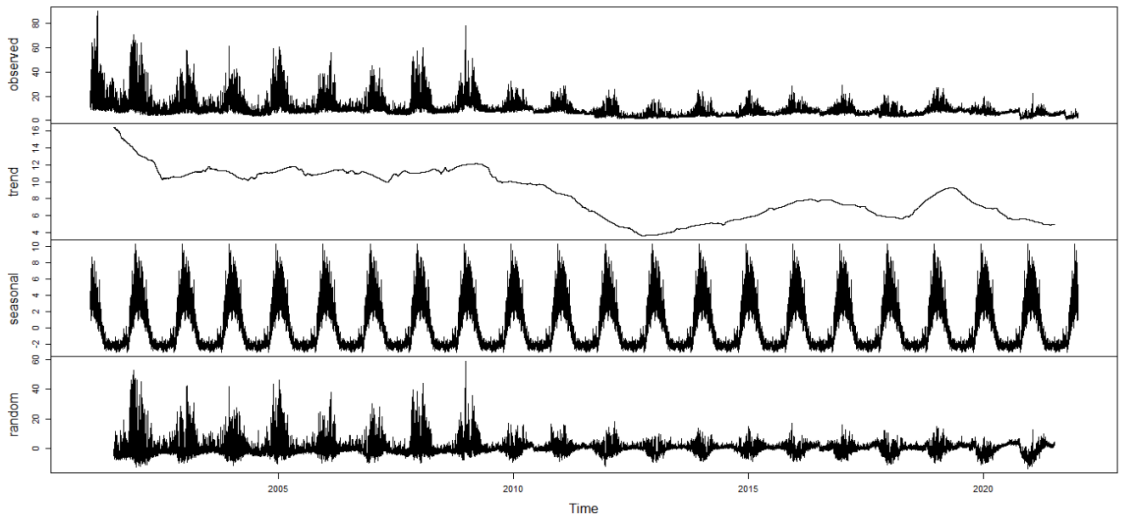
Anexo 1. Código del proyecto

El código desarrollado para el análisis se ha subido a un repositorio de GitHub público: https://github.com/SergioRC70/UOC_Ciencia_de_Datos_TFM

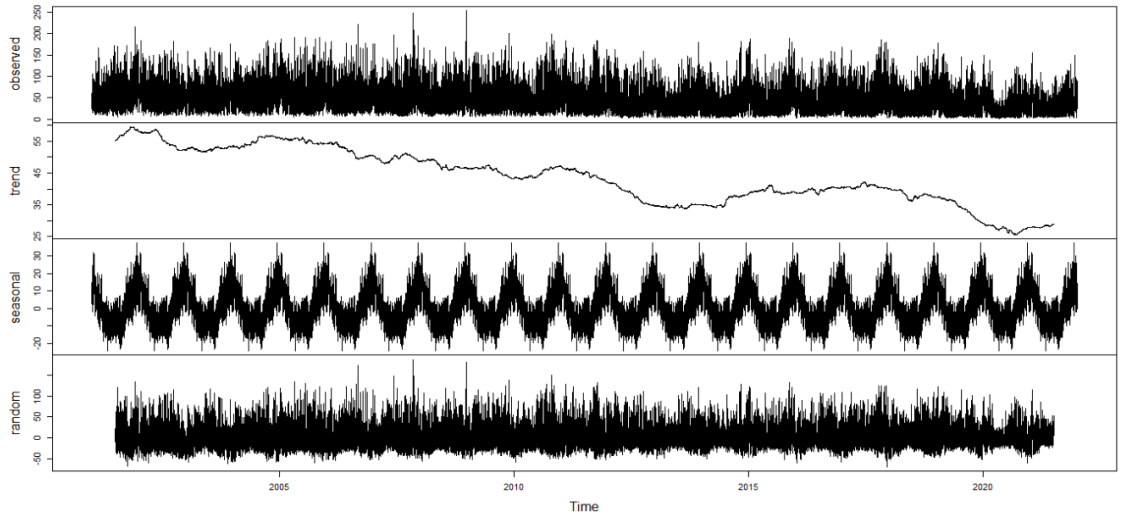
El repositorio contiene tanto el código R como el proyecto de QGIS con todos los archivos de datos necesarios para su ejecución. El fichero README.rm describe cada uno de los ficheros incluidos en el repositorio.

Anexo 2. Descomposición series temporales agentes contaminantes

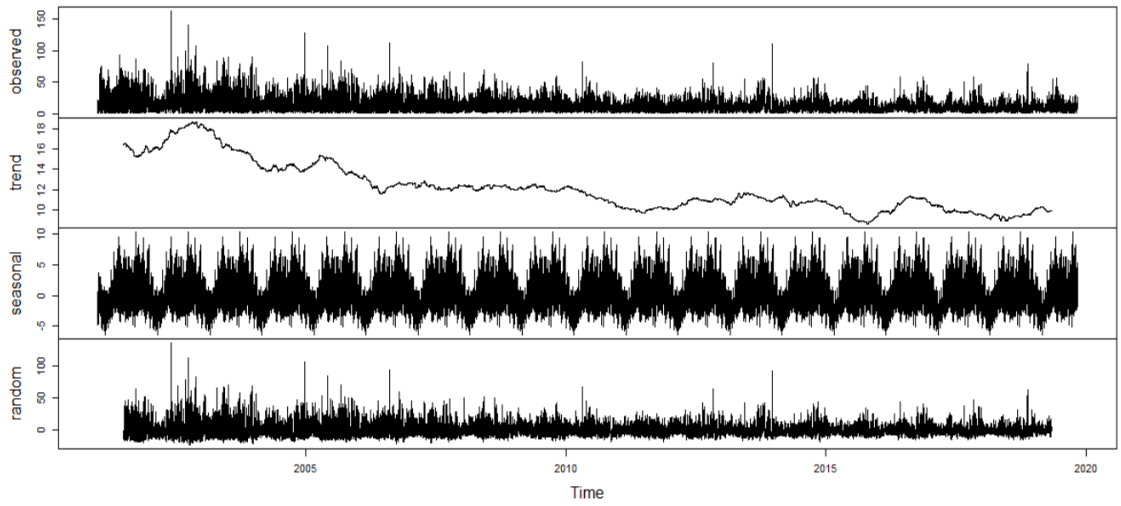
SO_2



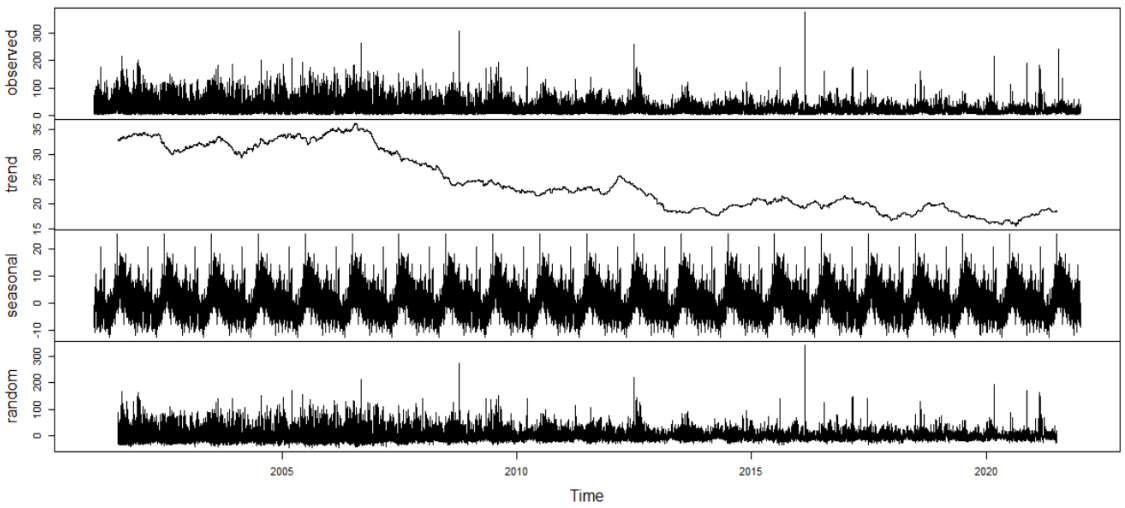
NO_2



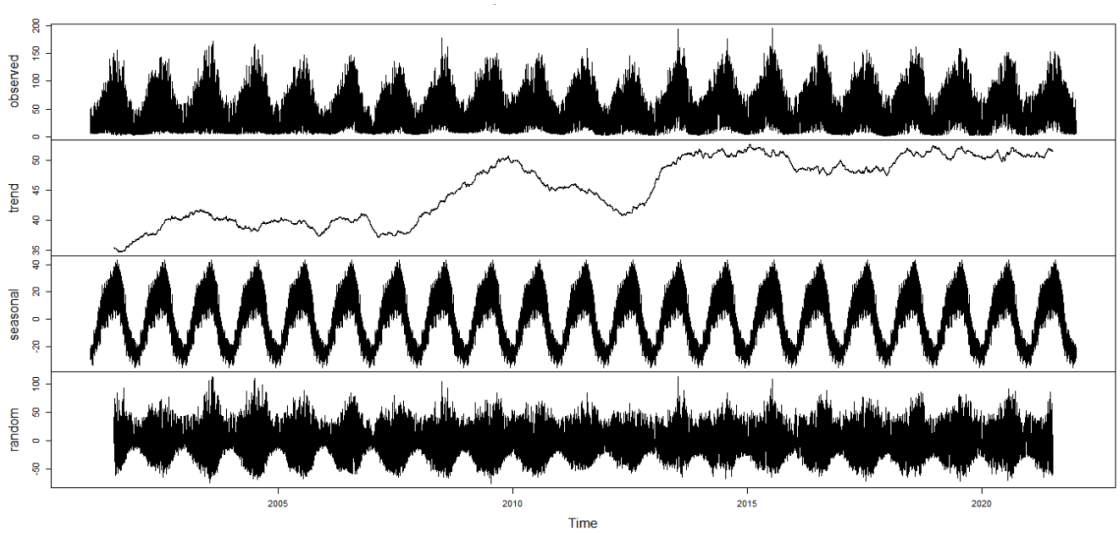
$PM_{2.5}$


observed
trend
seasonal
random
2005
2010
2015
2020
Time

PM_{10}


observed
trend
seasonal
random
2005
2010
2015
2020
Time

O_3


observed
trend
seasonal
random
2005
2010
2015
2020
Time