

GRADO EN CIENCIA DE DATOS



VNIVERSITAT
DE VALÈNCIA

TRABAJO FIN DE GRADO

DETECCIÓN DE ANOMALÍAS EN LA
CONTAMINACIÓN DEL AIRE EN VALENCIA

AUTOR:
LUCÍA CARRO ARCAYA

TUTOR:
FERNANDO MATEO JIMÉNEZ



VNIVERSITAT
DE VALÈNCIA



Escola Tècnica Superior
d'Enginyeria **ETSE-UV**

TRABAJO FIN DE GRADO

DETECCIÓN DE ANOMALÍAS EN LA CONTAMINACIÓN DEL AIRE EN VALENCIA

AUTOR: LUCÍA CARRO ARCAYA

TUTOR: FERNANDO MATEO JIMÉNEZ

Declaración de autoría:

Yo, Lucía Carro Arcaya, declaro la autoría del Trabajo Fin de Grado titulado “Detec-
ción de anomalías en la contaminación del aire en Valencia” y que el citado trabajo no
infringe las leyes en vigor sobre propiedad intelectual. El material no original que figura
en este trabajo ha sido atribuido a sus legítimos autores.

Valencia, 17 de julio de 2025

Fdo: Lucía Carro Arcaya

Resumen:

El objetivo principal de este TFG es detectar anomalías en datos de contaminación atmosférica de la ciudad de Valencia, utilizando registros horarios obtenidos entre 2022 y 2024 a través de estaciones oficiales de la Red Valenciana de Vigilancia y Control de la Contaminación Atmosférica. El estudio se centra en contaminantes clave como NO_2 , $\text{PM}_{2.5}$, O_3 y SO_2 , por su impacto en la salud pública y relevancia normativa.

La metodología aplicada abarca varias fases. En primer lugar, se realizó un proceso de preprocesado para unificar, limpiar y estructurar los datos. Seguido de un análisis exploratorio detallado que permitió identificar patrones temporales y estacionales, así como relaciones entre variables y comportamientos por estación de medida. Posteriormente, se aplicaron diferentes técnicas para la detección de anomalías, incluyendo tanto métodos estadísticos clásicos como modelos avanzados.

Los resultados obtenidos permiten valorar la utilidad de estos enfoques para la monitorización automática de la calidad del aire.

Abstract:

The main objective of this Final Degree Project is to detect anomalies in air pollution data from the city of Valencia, using hourly records collected between 2022 and 2024 through official monitoring stations of the Valencian Network for the Monitoring and Control of Atmospheric Pollution. The study focuses on key pollutants such as NO₂, PM_{2.5}, O₃, and SO₂, due to their impact on public health and regulatory relevance.

The methodology applied covers several phases. First, a preprocessing stage was carried out to unify, clean, and structure the data. This was followed by a detailed exploratory analysis that enabled the identification of temporal and seasonal patterns, as well as relationships between variables and behaviors across different stations. Subsequently, various techniques for anomaly detection were applied, including both classical statistical methods and advanced models.

The results obtained allow for an assessment of the usefulness of these approaches for the automatic monitoring of air quality.

Resum:

L'objectiu principal d'aquest Treball de Fi de Grau és detectar anomalies en dades de contaminació atmosfèrica de la ciutat de València, utilitzant registres horaris obtinguts entre 2022 i 2024 a través d'estacions oficials de la Xarxa Valenciana de Vigilància i Control de la Contaminació Atmosfèrica (RVVCCA). L'estudi se centra en contaminants clau com el NO₂, PM_{2.5}, O₃ i SO₂, pel seu impacte en la salut pública i la seua rellevància normativa.

La metodologia aplicada abasta diverses fases. En primer lloc, es va dur a terme un procés de preprocessament per unificar, netejar i estructurar les dades. A continuació, es va realitzar una anàlisi exploratòria detallada que va permetre identificar patrons temporals i estacionals, així com relacions entre variables i comportaments segons l'estació de mesura. Posteriorment, es van aplicar diferents tècniques per a la detecció d'anomalies, incloent tant mètodes estadístics clàssics com models avançats.

Els resultats obtinguts permeten valorar la utilitat d'aquests enfocaments per a la monitorització automàtica de la qualitat de l'aire.

Agradecimientos:

A Valencia, por todo lo que me ha hecho crecer y aprender en estos cuatro años. Por toda la gente que he conocido, que ha hecho que esta etapa de mi vida siempre la vaya a recordar con una sonrisa.

A mi familia y amigos, gracias por darme vuestro apoyo y cariño en todo momento, por confiar en mí más que nadie. Porque sé que puedo contar con vosotros para todo, y aun estando lejos, os tengo siempre presentes.

A Naiara, Irune y Amaia, hacéis que volver siempre sea bonito.

A mis padres y a mi hermano, gracias por estar y por cuidar de mí. Siempre estaré agradecida de teneros.

Y en especial, a mi madre, por ser mi modelo a seguir. Gracias por todo, te admiro.

Índice general

1. Introducción	17
1.1. Introducción	17
1.2. Motivación	17
1.3. Objetivos	18
1.4. Organización de la memoria	18
2. Estado del arte	21
2.1. Definición y tipología de anomalías en series temporales	21
2.2. Métodos estadísticos clásicos	22
2.3. Algoritmos	23
2.3.1. ARIMA	23
2.3.2. Isolation Forest	24
2.3.3. Autoencoder LSTM	26
3. Metodología	29
3.1. Datos del estudio	29
3.1.1. Fuentes y variables medidas	29
3.1.2. Estructura de los datos utilizados	30
3.2. Preprocesamiento de datos	30
3.3. Análisis exploratorio	33
3.3.1. Análisis inicial	33
3.3.2. Correlaciones	35
3.3.3. Análisis temporal	38
3.3.4. Análisis STL	39
3.4. Umbral de anomalías	43
3.4.1. Umbrales estadísticos	43
3.4.2. Comité de anomalías	48
3.4.3. Umbral legal	51
3.5. Análisis no supervisado	53

3.6. Selección de estaciones	56
4. Resultados	57
4.1. Introducción	57
4.2. ARIMA	59
4.3. Isolation Forest	63
4.4. Autoencoder LSTM	71
4.5. Comparación	76
5. Conclusiones	81
5.1. Conclusiones	81
5.2. Trabajo futuro	84
A. Anexos	85
Bibliografía	89

Capítulo 1

Introducción

1.1. Introducción

La contaminación del aire en entornos urbanos constituye un grave problema ambiental y de salud pública. Las emisiones procedentes del transporte, la industria y la actividad energética liberan contaminantes como dióxido de nitrógeno (NO_2), partículas en suspensión (PM_{10} y $\text{PM}_{2.5}$) y ozono (O_3). Estos compuestos se han vinculado a un aumento de enfermedades respiratorias, cardiovasculares y a efectos adversos en la población infantil.

La ciudad de Valencia, con su área metropolitana altamente poblada, presenta episodios críticos de contaminación que en ocasiones superan los límites establecidos por la normativa europea y estatal. Analizar estos episodios y detectar anomalías en las series temporales de contaminantes puede facilitar la implementación de sistemas de alerta temprana y mejorar la toma de decisiones de las autoridades ambientales.

1.2. Motivación

La motivación principal de este trabajo es demostrar que la detección automática de anomalías en las series temporales de contaminación puede mejorar de forma significativa la gestión de la calidad del aire en Valencia. Se parte de la hipótesis de que un sistema que capte y aproveche patrones atípicos —más allá de los simples umbrales— proporcionará información más rica y accionable para las autoridades ambientales.

Importancia del problema

- **Salud pública:** Detectar de forma temprana episodios contaminantes críticos permite reducir la exposición de grupos vulnerables (niños, ancianos, enfermos crónicos) y mitigar efectos adversos en la salud.
- **Cumplimiento normativo:** Un sistema continuo y fiable de alerta ayuda a garantizar que no se superen límites legales, facilitando la adopción inmediata de medidas correctivas.
- **Eficiencia en la gestión:** Contar con criterios robustos de detección de anomalías optimiza el uso de recursos (personal, mediciones extraordinarias) y prioriza intervenciones donde realmente sean necesarias.

- **Base para sistemas avanzados:** Los resultados y metodologías obtenidos sentarán las bases para el desarrollo de herramientas proactivas de gestión ambiental, integrando modelos predictivos y análisis espacial.

1.3. Objetivos

Este Trabajo de Fin de Grado se centra en la detección automática de anomalías en series temporales de contaminación del aire en la ciudad de Valencia. El propósito es aprovechar los datos de concentración de distintos contaminantes, con registros cada diez minutos desde enero de 2022 hasta mayo de 2024, para diseñar y evaluar métodos que permitan identificar de forma fiable episodios atípicos.

Para cumplir con este objetivo, se plantean las siguientes metas:

- **Analizar el rendimiento de los algoritmos de detección.** Evaluar la precisión de diferentes técnicas para diferenciar episodios contaminantes significativos de la variabilidad de fondo habitual.
- **Caracterizar las anomalías en el dominio temporal y estacional.** Clasificar los eventos detectados según su duración, picos breves frente a elevaciones prolongadas, y estudiar su distribución a lo largo de las distintas estaciones del año.
- **Comparar el comportamiento entre estaciones de monitorización.** Analizar la frecuencia y magnitud de las anomalías en cada estación de la red de vigilancia de Valencia, con el fin de identificar zonas con mayor incidencia de episodios atípicos.
- **Definir criterios de detección operativos.** Establecer umbrales estadísticos y reglas prácticas que permitan automatizar la identificación de anomalías en tiempo real.

Como resultado, se espera que los métodos seleccionados identifiquen anomalías en las series temporales de contaminantes con alta precisión y un bajo ratio de falsas alarmas. Además, permitan caracterizar de manera detallada el comportamiento temporal y espacial de los eventos atípicos, y faciliten la definición de criterios operativos para automatizar la detección en tiempo real, sirviendo así como base sólida para el desarrollo futuro de un sistema de alerta temprana, al aportar evidencia cuantitativa de su viabilidad y fiabilidad.

1.4. Organización de la memoria

La memoria comienza introduciendo el contexto general del estudio, incluyendo la motivación que lo impulsa, los objetivos planteados y la relevancia del problema abordado. A continuación, se presentan los fundamentos teóricos y metodológicos, describiendo los principales enfoques utilizados en la detección de anomalías en series temporales, desde métodos estadísticos tradicionales hasta modelos más avanzados de aprendizaje automático.

Seguidamente, se detalla la metodología empleada, incluyendo las características del conjunto de datos utilizado, el proceso de preprocesamiento aplicado y el análisis explora-

torio que permitió identificar patrones temporales y estacionales relevantes. Posteriormente, se describen las técnicas implementadas para la detección de anomalías, justificando su elección y aplicación en función de las propiedades de los datos.

Concluye con la exposición de los resultados obtenidos y su interpretación, seguida de un apartado de conclusiones donde se valoran los principales hallazgos del estudio. Finalmente, se plantean posibles líneas de trabajo futuro y se incluye un anexo con material complementario de interés.

Capítulo 2

Estado del arte

2.1. Definición y tipología de anomalías en series temporales

Una anomalía en el contexto de una serie temporal es un valor o un conjunto de valores que se desvían significativamente del comportamiento esperado de la serie. Estas desviaciones pueden indicar eventos excepcionales o errores de medición. [1]

Se distinguen tres tipologías principales [1]:

- Anomalías puntuales: un único valor que se separa claramente de la tendencia y la variabilidad habitual de la serie. Por ejemplo, un pico aislado de concentración que no guarda relación con los instantes anteriores ni posteriores.

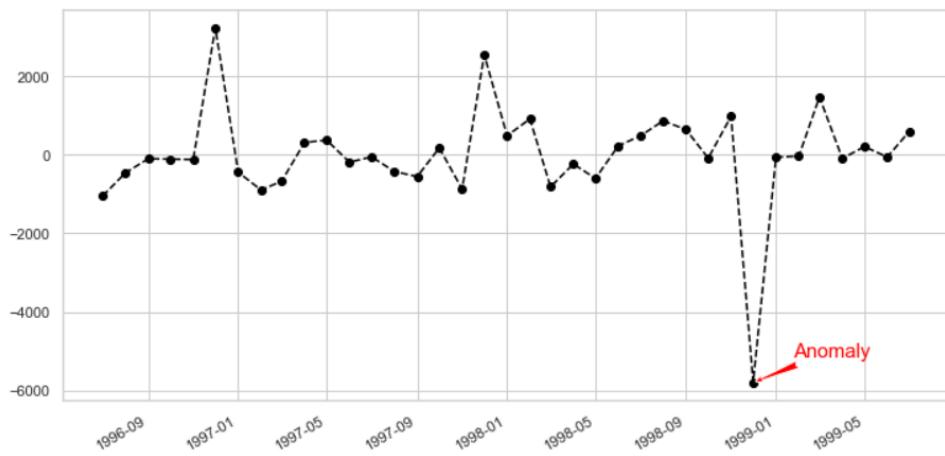


Figura 2.1: Ejemplo de anomalía puntual en una serie temporal. Fuente: [2]

- Anomalías colectivas: un bloque de valores consecutivos que, en su conjunto, constituyen un patrón atípico aunque cada punto por separado no lo sea. Un ejemplo es una elevación sostenida de contaminantes durante varias horas.

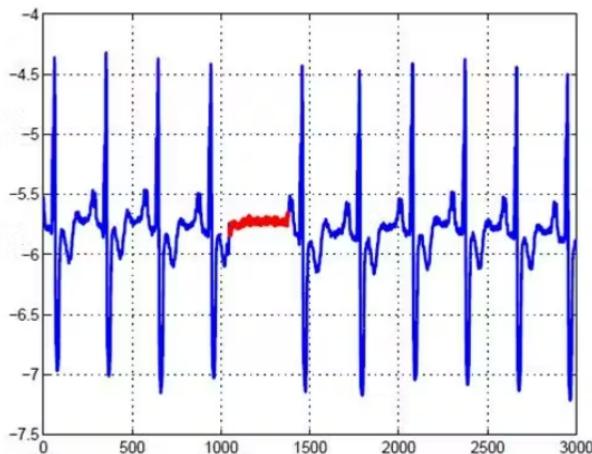


Figura 2.2: Ejemplo de anomalía colectiva en una serie temporal. Fuente: [3]

- Anomalías contextuales: valores que solo resultan anómalos en un contexto específico, como la estación del año o la hora del día. Un nivel elevado de ozono puede ser normal a mediodía en verano, pero atípico de madrugada o en invierno.

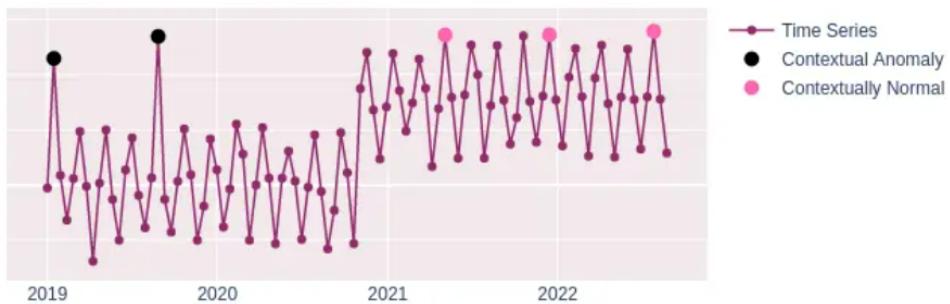


Figura 2.3: Ejemplo de anomalía colectiva en una serie temporal. Fuente: [4]

2.2. Métodos estadísticos clásicos

En este apartado se describen los métodos estadísticos de referencia utilizados para la detección de anomalías, todas ellas implementadas en el TFG.

- Regla de las 3 sigmas

Se considera anómalo todo valor x_t que satisfaga

$$|x_t - \mu| > 3\sigma,$$

donde μ y σ son la media y la desviación estándar de la serie histórica respectivamente. Es muy sencilla de computar y de bajo coste computacional, pero asume una distribución aproximadamente normal y es sensible a outliers previos que distorsionen μ y σ [5].

- Percentil 95

Define el umbral en el valor correspondiente al percentil 95 de la distribución histórica, de forma que cualquier observación superior se marca como anomalía.

Esta no requiere suposición de normalidad y captura directamente el comportamiento empírico extremo. Como limitación, no informa sobre la cola inferior y el valor de corte puede variar con el tamaño de la muestra y los cambios de tendencia [5].

- Regla de Hampel

Sustituye la media y la desviación estándar por la mediana \tilde{x} y la desviación absoluta mediana (MAD), definiendo anomalía cuando

$$\frac{|x_t - \tilde{x}|}{\text{MAD}} > k,$$

con $k \approx 3$. Es robusta frente a outliers ya presentes y adecuada para distribuciones no normales, aunque tiene un mayor coste computacional al calcular mediana y MAD, y la elección de k está menos estandarizada [6].

- Boxplot (IQR)

Basado en el rango intercuartílico $\text{IQR} = Q_3 - Q_1$. Se marcan como atípicos los valores fuera de

$$[Q_1 - 1,5 \text{ IQR}, Q_3 + 1,5 \text{ IQR}].$$

Este método está ampliamente utilizado en análisis exploratorio de datos y no depende de la forma de la distribución. Pero no distingue entre anomalías puntuales y colectivas y el factor 1.5 es arbitrario [7].

2.3. Algoritmos

2.3.1. ARIMA

El modelo ARIMA (AutoRegressive Integrated Moving Average) es uno de los enfoques más consolidados para el análisis y predicción de series temporales. Su estructura combina tres componentes:

- AR(p): parte autorregresiva, donde el valor actual se modela como una combinación lineal de los p valores pasados.
- I(d): parte de integración, que diferencia la serie d veces para alcanzar estacionariedad.
- MA(q): parte de media móvil, en la que el valor actual depende de los q errores de predicción anteriores.

Se representa mediante la notación $ARIMA(p, d, q)$, donde los parámetros se ajustan en función de las características de la serie. El modelo es ampliamente reconocido en la literatura como combinación de autoregresión, diferenciación e integración para series con tendencia [8].

Para la detección de anomalías, se sigue el siguiente procedimiento:

1. Se entrena el modelo ARIMA con la serie temporal de referencia.
2. Se genera la predicción \hat{x}_t para cada instante t .
3. Se calcula el residuo $\varepsilon_t = x_t - \hat{x}_t$.
4. Se considera que una observación es anómala si su residuo ε_t excede un umbral, normalmente definido como un múltiplo de la desviación estándar de los residuos ($k\sigma$) [9].

Selección de parámetros y `auto_arima`

La elección óptima de los hiperparámetros (p, d, q) puede realizarse mediante el análisis visual de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF). No obstante, para automatizar este proceso se emplea la función `auto_arima`, disponible en la librería `pmdarima`, que explora múltiples combinaciones de parámetros utilizando criterios como el AIC (Akaike Information Criterion) o BIC (Bayesian Information Criterion) [10].

Esta función evalúa distintos modelos candidatos y selecciona el que minimiza el criterio de información, ajustándose automáticamente al orden de diferenciación necesario y al mejor conjunto (p, d, q) .

Ventajas y limitaciones

Entre las principales ventajas del modelo ARIMA destaca su capacidad para modelar y prever tendencias y ciclos recurrentes en series temporales, lo que lo convierte en una herramienta eficaz para análisis con patrones estacionales definidos. Se trata de un modelo ampliamente estudiado y validado, lo que facilita su interpretación desde una perspectiva estadística. Además, ofrece la posibilidad de detectar anomalías directamente a partir del análisis de sus residuos, simplificando así el proceso de identificación de comportamientos inusuales.

No obstante, también presenta ciertas limitaciones. El modelo requiere que la serie temporal sea estacionaria, lo que implica aplicar procesos de diferenciación previos en muchos casos. Asimismo, puede mostrar poca robustez ante cambios abruptos en la dinámica de la serie si estos no fueron representativos durante la fase de entrenamiento. Finalmente, su rendimiento disminuye cuando se enfrenta a series con patrones no lineales o cuando intervienen múltiples fuentes de variabilidad difíciles de modelar de forma lineal.

Este modelo se utiliza como punto de partida en muchos sistemas de detección de anomalías, especialmente en contextos donde la interpretabilidad es un requisito importante o donde se dispone de series históricas suficientemente largas para un ajuste fiable.

2.3.2. Isolation Forest

Isolation Forest es un algoritmo basado en árboles de aislamiento, diseñado específicamente para detectar valores atípicos sin requerir un modelo probabilístico del comportamiento normal [11]. Su funcionamiento se basa en la hipótesis de que las anomalías, por su rareza, requieren menos divisiones aleatorias para ser aisladas que las observaciones habituales.

El modelo genera múltiples árboles binarios en los que cada nodo divide aleatoriamente una variable y un valor de corte. La longitud media del camino necesario para aislar un punto sirve como métrica: cuanto más corto es el camino, mayor es la probabilidad de que el dato sea anómalo. Esta lógica se aplica eficientemente incluso en espacios de alta dimensionalidad, lo que hace que Isolation Forest sea especialmente útil en contextos multivariantes [12, 13].

Preparación de variables

Para aplicar este método a series temporales ambientales, se diseñó una matriz de características que incorpora tanto información histórica como estacional. En concreto, se incluyeron:

- Variables de retardo: se añadieron rezagos (lags) de 1 y 7 días para capturar dependencia temporal reciente y semanal.
- Estadísticas móviles: se calcularon la media y desviación estándar móviles en ventanas de 7 días, permitiendo identificar desviaciones respecto al comportamiento reciente.
- Z-score local: se añadió una puntuación estandarizada basada en la ventana móvil anterior, como medida relativa de alejamiento.
- Variables temporales: se incorporaron variables categóricas como el día de la semana, el mes y una variable indicadora de fin de semana, posteriormente codificadas como variables *dummy*.

Estas variables permitieron al modelo trabajar con una representación enriquecida del estado del sistema atmosférico diario, capturando tanto patrones persistentes como fluctuaciones estacionales.

Una vez preparada la matriz de características, se entrenó el modelo Isolation Forest para cada contaminante de forma independiente. Las observaciones con puntuaciones de anomalía superiores a un umbral específico fueron etiquetadas como valores atípicos.

El enfoque permitió detectar no solo picos individuales, sino también configuraciones inusuales de varios atributos, como días en los que el valor absoluto de un contaminante no era extremadamente alto pero sí lo era su desviación respecto al patrón reciente.

Ventajas y limitaciones

Entre las ventajas más destacadas del algoritmo Isolation Forest se encuentra el hecho de que no requiere realizar suposiciones previas sobre la distribución de los datos, lo que lo hace especialmente útil en contextos reales con estructuras complejas. Su capacidad para detectar patrones multivariantes y relaciones no lineales le permite identificar tanto anomalías puntuales como contextuales o colectivas. Además, es un modelo altamente escalable, lo que lo hace adecuado para su aplicación en grandes volúmenes de datos sin comprometer significativamente el rendimiento computacional.

Sin embargo, también presenta ciertas limitaciones. Isolation Forest es sensible a la escala de las variables, por lo que resulta imprescindible aplicar técnicas de normalización adecuadas antes de su entrenamiento. Por otro lado, la interpretación de los resultados puede no ser tan intuitiva como en los modelos estadísticos clásicos, debido a la naturaleza de su funcionamiento basado en árboles de aislamiento. Finalmente, su rendimiento puede verse afectado por la selección de hiperparámetros y por la calidad y estructura de las variables utilizadas como entrada al modelo.

En resumen, Isolation Forest ofrece una solución robusta y flexible para detectar episodios anómalos en series temporales ambientales cuando se dispone de un conjunto rico de atributos derivados. Su capacidad para modelar relaciones complejas lo convierte en un complemento útil frente a métodos estadísticos más tradicionales [14].

2.3.3. Autoencoder LSTM

Los autoencoders LSTM son una arquitectura de redes neuronales recurrentes orientada a la reconstrucción de secuencias temporales, ampliamente utilizada en tareas de detección de anomalías no supervisada. Su principio se basa en que un modelo entrenado sobre datos normales será capaz de reconstruir correctamente esas secuencias, mientras que presentará errores notables cuando se enfrente a patrones anómalos.

Un autoencoder LSTM consta de dos componentes principales: un codificador que transforma la secuencia de entrada en una representación latente de menor dimensión, y un decodificador que intenta reconstruir la secuencia original a partir de dicha representación. Dado que las LSTM (*Long Short-Term Memory*) son especialmente adecuadas para capturar dependencias a largo plazo en datos secuenciales, este modelo aprende patrones temporales complejos que no podrían capturarse con métodos clásicos de series temporales [15, 16].

Preparación de variables

Para su aplicación en series temporales ambientales, los datos se estructuraron en ventanas de longitud fija (por ejemplo, 14 días) que contienen las concentraciones sucesivas de un contaminante. Previamente, se aplicó una normalización *min–max* para asegurar la estabilidad numérica durante el entrenamiento.

Cada ventana temporal representa una instancia de entrada al autoencoder. El modelo se entrena únicamente con datos considerados normales, optimizando su capacidad para reconstruir secuencias típicas. Posteriormente, cualquier secuencia que arroje un error de reconstrucción superior a un umbral, generalmente definido como la media más cierta cantidad de desviaciones estándar de los errores sobre el conjunto de entrenamiento, se considera anómala [15, 17].

Ventajas y limitaciones

Los autoencoders LSTM presentan una serie de ventajas relevantes para el análisis de series temporales ambientales. En primer lugar, son capaces de modelar relaciones temporales complejas y no lineales, superando así las limitaciones de los modelos lineales o basados en estadística clásica. Su carácter no supervisado les permite trabajar sin necesidad de disponer de etiquetas, lo que facilita su aplicación en entornos donde no se cuenta con un conjunto de datos anotado. Además, destacan por su capacidad para detectar tanto anomalías puntuales como secuencias anómalas prolongadas, gracias a su estructura basada en memoria de largo plazo. Esta versatilidad los convierte en modelos altamente adaptables a distintos contaminantes y escalas temporales.

No obstante, también presentan ciertas limitaciones. El entrenamiento efectivo de un autoencoder LSTM requiere una cantidad considerable de datos, lo que puede suponer una barrera en situaciones con registros incompletos o fragmentados. Asimismo, el rendimiento del modelo depende en gran medida de la adecuada elección de hiperparámetros como el tamaño de ventana, la dimensión del espacio latente o el número de épocas de entrenamiento. Por último, este tipo de arquitectura implica un mayor coste computacional respecto a los métodos estadísticos más tradicionales, tanto en tiempo de entrenamiento como en la detección posterior de anomalías.

En estudios recientes, los autoencoders LSTM han demostrado un alto rendimiento en la detección de anomalías en series temporales de calidad del aire interior, superando a enfoques tradicionales en detección de cambios sutiles y dependencias a largo plazo en las secuencias de datos [15]. Asimismo, han sido aplicados con éxito en entornos multivariantes, mostrando robustez frente a datos ruidosos y alta capacidad de generalización [17].

En conclusión, los autoencoders LSTM constituyen una herramienta poderosa para la detección de anomalías en datos ambientales con fuerte estructura temporal. Su capacidad para reconstruir secuencias completas los convierte en una solución eficaz frente a anomalías sutiles que no se detectan fácilmente con métodos basados en umbrales fijos o estadísticas simples.

Capítulo 3

Metodología

3.1. Datos del estudio

3.1.1. Fuentes y variables medidas

El conjunto de datos utilizado en este trabajo procede de la *Red Valenciana de Vigilancia y Control de la Contaminación Atmosférica* (RVVCCA), gestionada por la Generalitat Valenciana. Esta red fue establecida mediante el Decreto 161/2003 del Consell[18], y cuenta con múltiples estaciones de medida distribuidas por la ciudad y provincia de València.

En el año 2017, la red disponía de aproximadamente 65 estaciones de medición automáticas, cada una registrando de manera continua parámetros relacionados con la calidad del aire, generando decenas de miles de lecturas diarias[19]. La recogida y validación de los datos es realizada por el *Centro de Estudios Ambientales del Mediterráneo* (CEAM), bajo encargo de la Conselleria de Medio Ambiente[20].

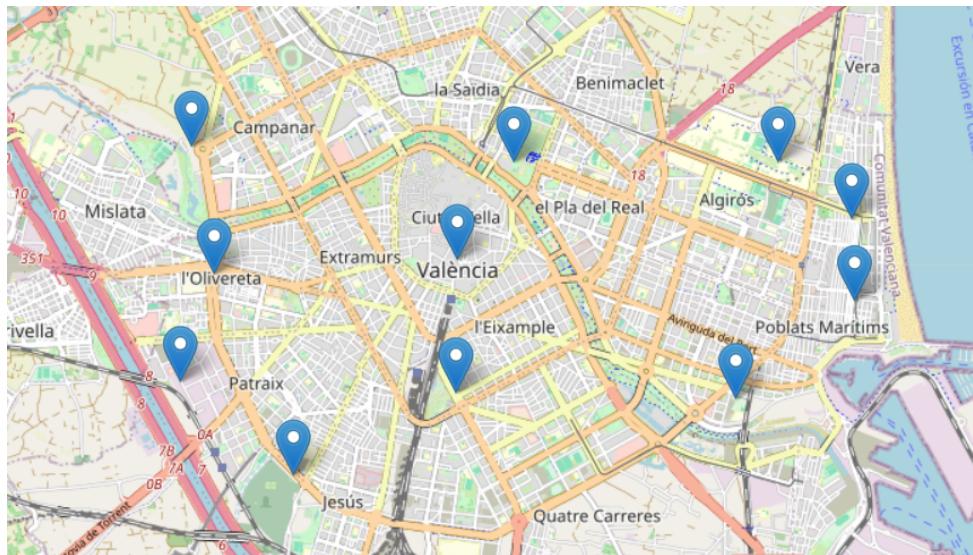


Figura 3.1: Localización de las estaciones de medición.

El mapa anterior muestra la ubicación de las estaciones de vigilancia de la calidad del aire en Valencia (marcadores azules) utilizadas en este trabajo, que forman parte de la RVVCCA. El portal oficial “Datos On-line” de la Generalitat Valenciana[21] permite

descargar registros históricos y en tiempo real procedentes de las estaciones de la red. En el presente trabajo, se ha realizado una recopilación y análisis de los datos horarios correspondientes al periodo 2022–2024. Los archivos, disponibles en formato **JSON**, se obtuvieron directamente desde dicho portal web, y posteriormente fueron procesados y tratados siguiendo una metodología detallada en el apartado correspondiente.

Las mediciones incluyen los principales contaminantes regulados en el aire ambiente. Según las fuentes oficiales, las estaciones miden continuamente compuestos como dióxido de azufre (SO_2), óxidos de nitrógeno (dióxido NO_2 , óxido NO , NO_x en total), ozono (O_3) y partículas en suspensión de distintos diámetros (PM_{10} , $\text{PM}_{2.5}$ y PM_1) [19].

3.1.2. Estructura de los datos utilizados

Los datos descargados están en formato **JSON** y, tras cargarlos con **pandas**, resultan en un único **DataFrame** con 63 columnas comunes para el periodo 2022–2024 (alrededor de 1.3 millones de registros en total). Cada registro corresponde a una medición puntual en una estación (campo `entityId`) y hora determinada (campos `dateObserved` o `dateObservedGMT0`).

Entre las columnas destacan: identificadores (`_id`, `entityId`, `entityType`), fecha y hora de la observación (`dateObserved`, `dateObservedGMT0`, `recvTime`), estado de la estación (`operationalStatus`), nombres y descripciones de parámetros (`N02Name`, `N02Description`, etc.), y los valores medidos de cada contaminante (`N02Value`, `O3Value`, `S02Value`, `PM10Value`, `PM25Value`, `PM1Value`, `NOXValue`, `NOValue`).

Para cada valor también se almacenan indicadores de calidad (`ValueFlag`), origen (`ValueOrigin`), tipo (`Type`) y factores de corrección (p. ej. `PM10CorrectionFactor`). Además, hay campos auxiliares como correos de mantenimiento (`maintenanceOwnerEmail`) o referencias geográficas (`refPointOfInterest`).

3.2. Preprocesamiento de datos

El preprocesado de los datos ha constituido una fase fundamental del trabajo, cuyo objetivo principal fue garantizar la calidad, coherencia y estructura adecuada del conjunto de observaciones antes de aplicar cualquier técnica de análisis o modelado.

En primer lugar, se procedió a la carga sistemática de los registros obtenidos de la Generalitat Valenciana, correspondientes a los años 2022, 2023 y parte de 2024. Estos archivos, en formato **JSON**, fueron unificados y transformados en un único **DataFrame** mediante la biblioteca **pandas**, conservando todas las columnas relevantes. La columna `dateObserved` se convirtió al tipo de dato `datetime64`, permitiendo su manipulación como serie temporal.

A continuación, se abordó la conversión de los datos de medición, ya que muchas variables aparecían inicialmente como texto (`string`). Se aplicó una conversión explícita a tipo `float64` para facilitar el cálculo de estadísticas, resampleos y modelos posteriores.

Una parte importante del preprocesado fue la selección de variables y columnas útiles. Se eliminaron aquellas que no aportaban valor al análisis temporal, como `location`, que contenía estructuras tipo `dict` difíciles de tratar y cuya información geoespacial no

era prioritaria en esta fase. Del mismo modo, se descartaron columnas redundantes como `*_Name`, `*_Description`, y se seleccionó trabajar únicamente con los campos `Value`, desechando `ValueOrigin` tras comprobar que ambas medidas eran prácticamente idénticas.

En cuanto a la limpieza estructural, se eliminaron los registros duplicados. Para ello, se comprobó la unicidad de la combinación `entityId` (estación) y `dateObserved`, conservando únicamente la primera aparición de cada par único. Esto garantizó que no existieran observaciones repetidas para una misma estación en una misma fecha y hora.

Posteriormente, se aplicaron procesos de agregación temporal. Se calcularon promedios diarios y horarios para cada contaminante, tanto a nivel global (promedio de todas las estaciones) como de forma individual para cada estación. Esto permitió obtener una visión general del comportamiento medio de la ciudad y también conservar la granularidad necesaria para análisis espaciales más finos.

A continuación, se presentan algunos gráficos representativos que ilustran la evolución temporal de los contaminantes más relevantes, tanto a nivel global como desglosado por estación, con el fin de visualizar las diferencias espaciales y temporales antes del modelado:

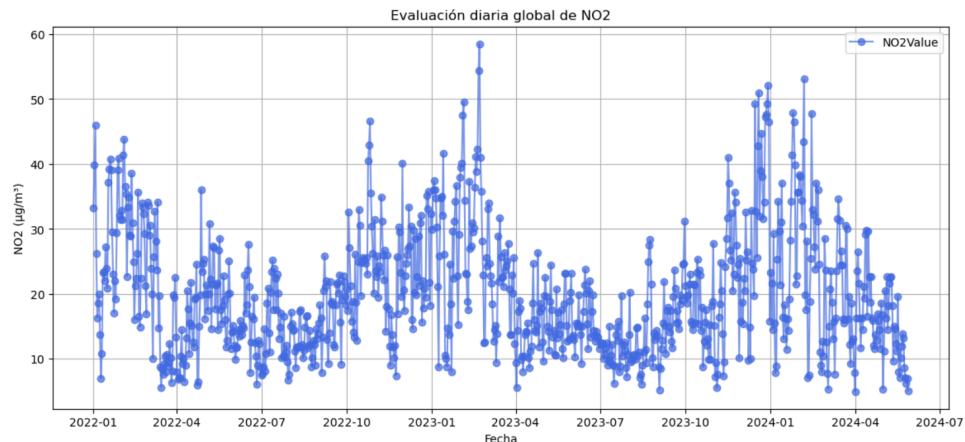


Figura 3.2: Evaluación diaria global de NO₂

En la Figura 3.2 se observa la evolución temporal del dióxido de nitrógeno (NO₂) para el conjunto total de estaciones entre 2022 y 2024. Se aprecian picos significativos en los meses invernales, consistentes con un mayor uso de calefacciones y condiciones atmosféricas que dificultan la dispersión de contaminantes.

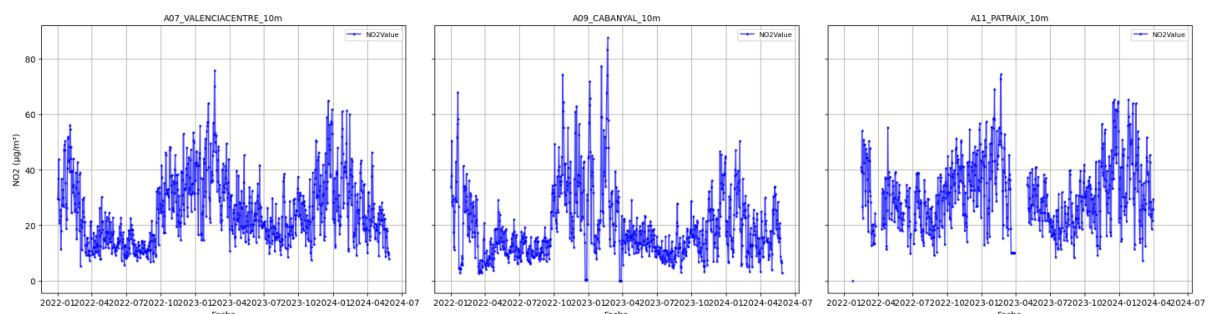


Figura 3.3: Comparativa diaria de NO₂ por estación

En la Figura 3.3 se comparan las concentraciones diarias de NO₂ registradas en tres estaciones representativas: VALENCIACENTRE, CABANYAL y PATRAIX. Se evidencian

diferencias de comportamiento entre zonas, siendo la estación central (VALENCIACENTRE) la que presenta niveles más elevados de forma más frecuente, reflejando una mayor carga vehicular o menor ventilación atmosférica.

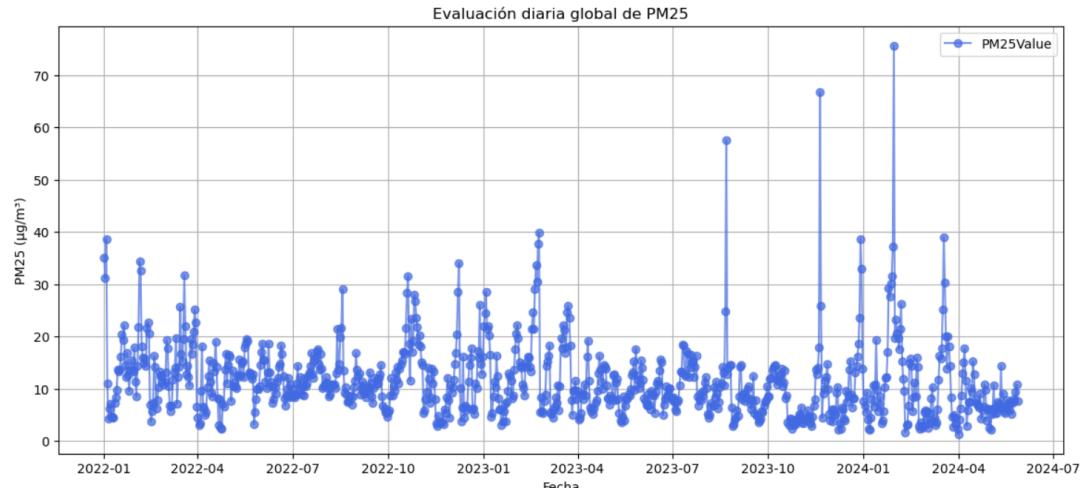


Figura 3.4: Evaluación diaria global de PM_{2.5}

La Figura 3.4 representa la evolución temporal de las concentraciones medias diarias de partículas finas (PM_{2.5}). Se observan múltiples picos aislados, posiblemente asociados a episodios de alta contaminación, obras o condiciones meteorológicas adversas. Aunque la mayoría de los valores se mantienen dentro del umbral recomendado, destacan algunos valores extremos que justifican su análisis posterior.

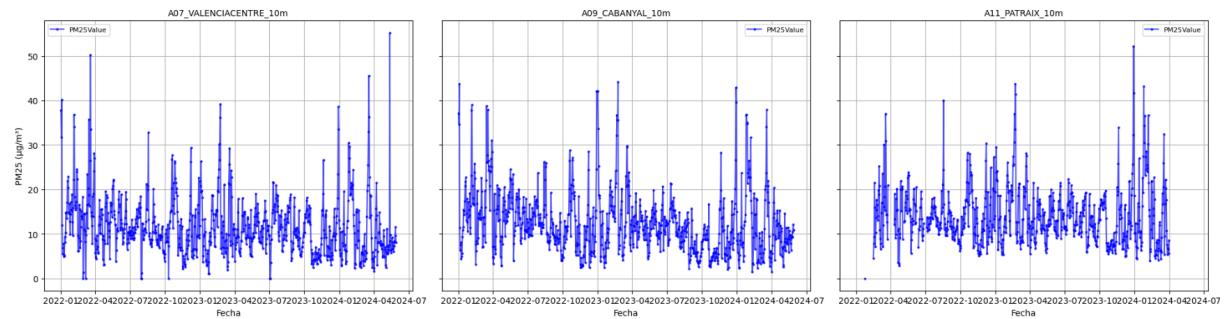


Figura 3.5: Comparativa diaria de PM_{2.5} por estación

La Figura 3.5 muestra la distribución diaria de las concentraciones de PM_{2.5} en las mismas estaciones. Aunque se mantiene una tendencia general homogénea entre ubicaciones, se identifican diferencias puntuales que podrían estar asociadas a condiciones locales específicas o eventos contaminantes transitorios.

Este proceso permitió construir un conjunto de datos limpio, consistente y estructurado sobre el que se pudieran aplicar con garantías los métodos de análisis exploratorio y modelado de anomalías.

3.3. Análisis exploratorio

El análisis exploratorio tiene como objetivo comprender la estructura de los datos, evaluar su calidad y extraer patrones iniciales que puedan guiar el modelado posterior.

3.3.1. Análisis inicial

Una de las primeras comprobaciones realizadas fue la comparación entre las columnas Value y ValueOrigin, disponibles para cada contaminante tanto en los datos diarios como horarios. El objetivo era verificar si existían diferencias significativas entre ambas medidas.

Las Tablas 3.1 y 3.2 recogen la diferencia relativa media entre ambas variables para cada contaminante. En todos los casos, las diferencias fueron inferiores al 3%, siendo la mayoría de ellas considerablemente menores, incluso iguales a cero en algunos casos como el ozono (O_3).

Variable	Media Value	Media ValueOrigin	Dif. relativa media (%)
SO_2	2.85	2.83	1.330
PM_1	7.57	7.53	1.207
$PM_{2.5}$	11.50	11.46	0.514
PM_{10}	23.10	23.05	0.233
NO	10.81	10.81	0.124
NO_2	19.93	19.93	0.021
NO_x	34.55	34.55	0.015
O_3	53.03	53.03	0.000

Tabla 3.1: Diferencia relativa entre *Value* y *ValueOrigin* para datos diarios

Variable	Media Value	Media ValueOrigin	Dif. relativa media (%)
PM_1	7.59	7.54	3.018
SO_2	2.86	2.84	2.067
NO	10.81	10.81	0.890
$PM_{2.5}$	11.51	11.46	0.707
PM_{10}	23.11	23.06	0.306
NO_2	19.96	19.96	0.101
NO_x	34.52	34.52	0.078
O_3	52.98	52.98	0.000

Tabla 3.2: Diferencia relativa entre *Value* y *ValueOrigin* para datos horarios

Dado que las discrepancias eran mínimas en ambos casos, se concluyó que *ValueOrigin* no aportaba información adicional significativa. Por tanto, para simplificar el conjunto de datos sin comprometer la calidad del análisis, se optó por utilizar únicamente la variable *Value* en el resto del trabajo.

A continuación, se compararon las estadísticas descriptivas entre los datos horarios y diarios. Tal como se muestra en la Tabla 3.3, las medias fueron prácticamente iguales en

todas las variables, con diferencias inferiores al 0.2 %. En cambio, la desviación estándar fue considerablemente mayor en los datos horarios, lo cual es esperable debido a la mayor variabilidad intra-diaria que se suaviza al promediar por día.

Variable	Media diaria	Media horaria	Dif. media (%)	Dif. std (%)
PM ₁	7.57	7.59	0.16	26.91
NO ₂	19.93	19.96	0.12	45.91
SO ₂	2.85	2.86	0.07	33.13
PM _{2.5}	11.50	11.51	0.06	30.44
PM ₁₀	23.10	23.11	0.04	41.63
NO	10.81	10.81	-0.06	70.28
O ₃	53.03	52.98	-0.09	47.05
NO _x	34.55	34.52	-0.11	49.60

Tabla 3.3: Comparación de estadísticas descriptivas entre datos diarios y horarios

Dado que el objetivo principal del estudio es analizar tendencias y patrones en una escala diaria, y considerando que los datos diarios son más estables y computacionalmente manejables, se decidió trabajar exclusivamente con estos. Además, como los datos por estación provienen del mismo conjunto original, se asumió que presentarían el mismo comportamiento, por lo que no se repitió la comparación a nivel de cada estación individual.

En cuanto a la calidad de los datos, se realizó un análisis de valores nulos, tanto a nivel global como por estación. Se identificó una cobertura muy desigual entre contaminantes: mientras que NO₂, PM₁₀ y PM_{2.5} presentaban una cobertura excelente o aceptable, otras variables como PM₁, NO_x o NO tenían más del 75 % de valores faltantes.

Variable	Nulos (global)	% (global)
PM ₁	321	36.52
NO	318	36.18
NO _x	318	36.18

Variable	Nulos (por estación)	% (por estación)
PM ₁	8433	88.31
NO	7442	77.93
NO _x	7442	77.93
O ₃	4320	45.24
SO ₂	4320	45.24
PM ₁₀	1935	20.26
PM _{2.5}	1935	20.26
NO ₂	210	2.20

Tabla 3.4: Porcentaje de valores nulos por variable en datos diarios y por estación

Como se observa en la Tabla 3.4, las variables PM₁, NO y NO_x presentan un elevado porcentaje de valores nulos, especialmente en el análisis por estación, donde superan ampliamente el 75 % de datos faltantes. En cambio, contaminantes como NO₂, PM₁₀ y PM_{2.5} muestran una cobertura muy superior, especialmente destacable en el caso de NO₂, cuya ausencia de datos es prácticamente despreciable (2.20 %). Estas diferencias en la disponibilidad justifican la exclusión de algunas variables con baja cobertura en los análisis

posteriores, priorizando aquellas con mayor calidad de datos y relevancia ambiental.

Para ilustrar la distribución de las concentraciones de contaminantes, se elaboraron diagramas de caja tanto a nivel global como por estación. El gráfico global (Figura 3.6) permite visualizar el comportamiento general de cada variable en términos de dispersión, valores extremos y simetría.

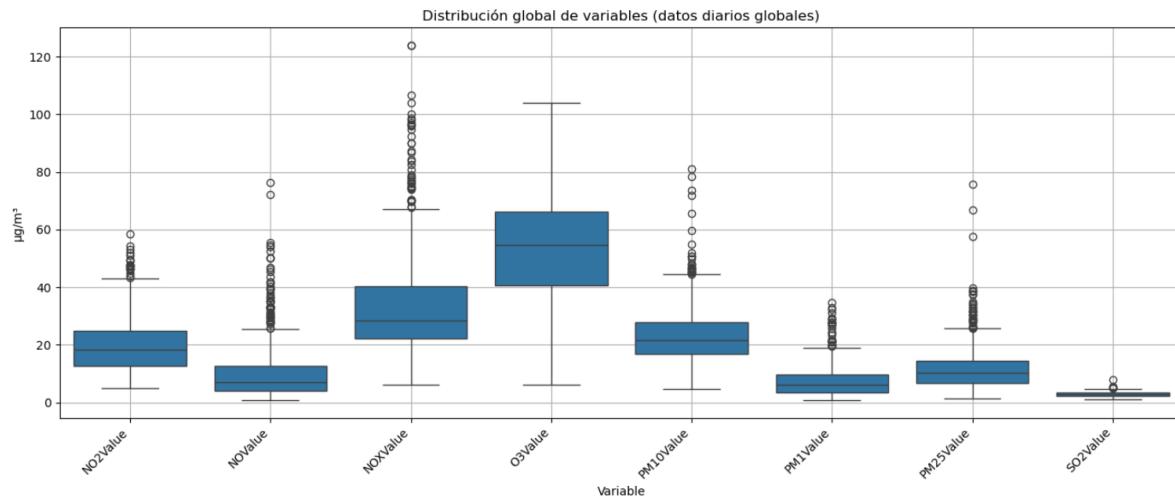


Figura 3.6: Distribución global de contaminantes (datos diarios)

Se observa que O_3 presenta las concentraciones más elevadas y regulares, con una mediana cercana a los $55 \mu\text{g}/\text{m}^3$ y una dispersión moderada. En cambio, contaminantes como NO_X , PM_{10} y $\text{PM}_{2.5}$ muestran una mayor variabilidad, con numerosos valores atípicos y colas largas hacia concentraciones más altas, indicativas de posibles episodios puntuales de alta contaminación. SO_2 es, por el contrario, la variable más baja y estable.

Además, se analizaron las distribuciones por estación para evaluar la variabilidad espacial. En el caso de NO_2 , se detectan diferencias notables entre estaciones, con valores más altos y dispersos en ubicaciones como OLIVERETA y PATRAIX, posiblemente debido a una mayor influencia del tráfico o actividad urbana. Para O_3 , en cambio, la distribución es bastante homogénea entre estaciones, aunque con algunos valores mínimos aislados en ubicaciones como MOLISOL o PISTASILLA. Por su parte, PM_{10} muestra una gran dispersión y presencia de valores extremos en varias estaciones, especialmente en AVFRANCIA y CABANYAL, lo que podría reflejar aportes locales o eventos específicos como resuspensión de polvo.

Este análisis gráfico proporciona una primera aproximación útil para detectar patrones anómalos y diferencias espaciales relevantes que serán consideradas en los análisis posteriores.

3.3.2. Correlaciones

Con el objetivo de entender la relación entre los distintos contaminantes registrados, se llevó a cabo un análisis exhaustivo de correlaciones, tanto a nivel global como desagregado por estación. Este análisis permite identificar redundancias, complementariedades y comportamientos particulares que podrían influir en la posterior detección de anomalías.

En primer lugar, se elaboró una matriz de correlación global (Figura 3.7) a partir de los datos diarios agregados para todas las estaciones. En dicha matriz se observa una fuerte correlación positiva entre los óxidos de nitrógeno (NO , NO_2 , NO_x), con coeficientes superiores a 0.85. Esto es coherente desde un punto de vista físico-químico, ya que NO_x incluye a NO y NO_2 , y todos ellos suelen tener una fuente común: el tráfico rodado. Esta alta redundancia motivó la decisión de conservar solo NO_2 como variable representativa del grupo, al tratarse del más regulado y con mayor impacto en salud pública.

Asimismo, las partículas en suspensión también presentaron fuertes correlaciones: PM_{10} , $\text{PM}_{2.5}$ y PM_1 están altamente relacionadas entre sí ($r > 0.75$), siendo $\text{PM}_{2.5}$ la variable seleccionada para representar este conjunto, debido a su mayor relevancia sanitaria (penetración en vías respiratorias profundas).

Por otro lado, el ozono (O_3) mostró correlaciones negativas significativas con los óxidos de nitrógeno, alrededor de -0.78 , lo que es indicativo de su comportamiento como contaminante secundario. Esta relación inversa se debe a que el ozono se forma a partir de reacciones fotoquímicas en presencia de luz solar y precursores como NO_2 , pero también se destruye en presencia de NO . Esta dinámica justifica su tratamiento como variable independiente y de interés en análisis posteriores.

El dióxido de azufre (SO_2), por su parte, mostró correlaciones muy bajas con el resto de contaminantes (valores inferiores a 0.25), lo que sugiere una fuente emisora distinta (posiblemente industrial o marítima). Su relativa independencia estadística apoya su inclusión como variable complementaria en los modelos de detección de anomalías.

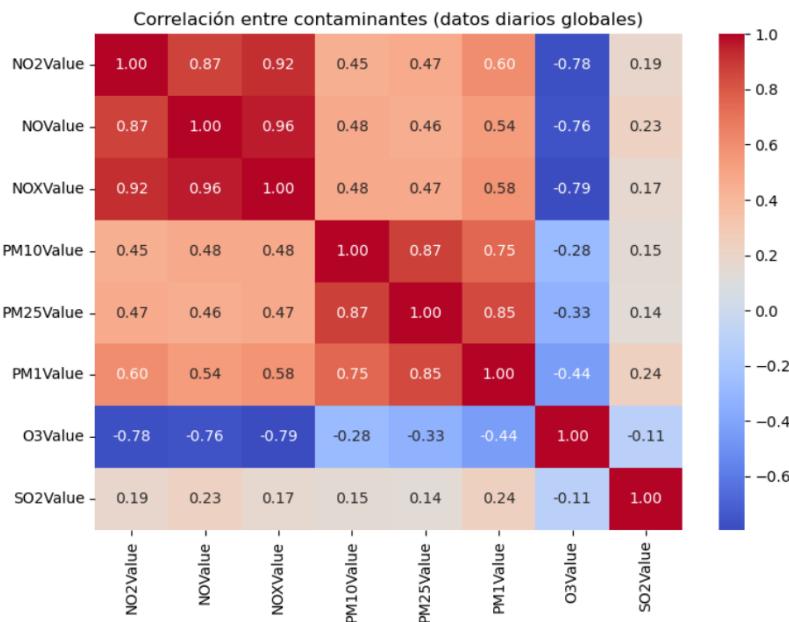


Figura 3.7: Matriz de correlación entre contaminantes (datos diarios globales)

Además del análisis global, se estudió la correlación entre contaminantes de forma desagregada por estación, con el fin de identificar posibles variaciones espaciales en las relaciones observadas. Previamente, se evaluó la cobertura de datos por estación y contaminante (Figura 3.8), lo cual evidenció que variables como NO_2 y $\text{PM}_{2.5}$ presentaban una excelente cobertura ($>90\%$) en la mayoría de estaciones, mientras que otras como O_3 y SO_2 estaban presentes únicamente en una parte del conjunto.

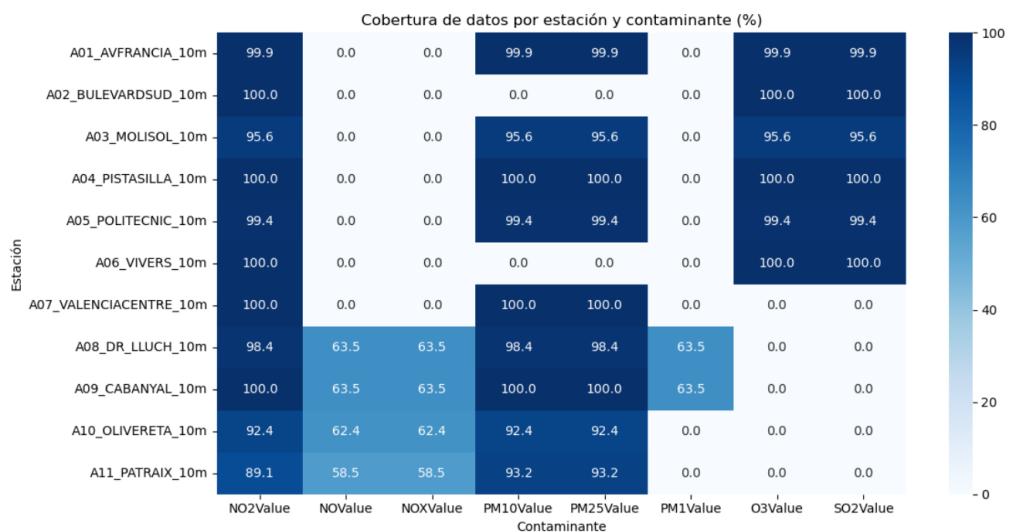


Figura 3.8: Cobertura de datos por estación y contaminante (%)

En estaciones con datos completos como A01_AVFRANCIA o A05_POLITECNIC se mantuvo el patrón observado a nivel global: correlación positiva entre NO₂ y PM_{2.5} (entre 0.3 y 0.6), correlación negativa entre NO₂ y O₃, y ausencia de relación destacable con SO₂. En otras estaciones con datos más limitados (como A07_VALENCIACENTRE o A08_DR_LLUCH), solo fue posible analizar las correlaciones entre NO₂ y PM_{2.5}, que también resultaron positivas y consistentes.

Estas diferencias locales en la fuerza y sentido de las correlaciones podrían explicarse por factores como la densidad del tráfico, las condiciones atmosféricas específicas o la presencia de fuentes contaminantes adicionales, y refuerzan la necesidad de análisis espaciales diferenciados cuando se aborden modelos más complejos.

A partir del análisis conjunto de la cobertura de datos y las relaciones entre contaminantes, se seleccionaron las variables NO₂, PM_{2.5}, O₃ y SO₂ como principales objetos de estudio. Esta elección responde a varios criterios complementarios.

En primer lugar, todas ellas presentan una cobertura de datos suficiente en la mayoría de las estaciones (ver Figura 3.8), lo que garantiza una base sólida para el análisis temporal y espacial.

En segundo lugar, desde el punto de vista estadístico, estas variables muestran comportamientos diferenciados entre sí y respecto a otros contaminantes, tal como revela la matriz de correlación global (Figura 3.7). Por ejemplo, O₃ presenta correlaciones negativas con NO₂ y NO, lo que indica dinámicas atmosféricas distintas que pueden enriquecer el análisis.

Finalmente, se priorizó la diversidad de perfiles: NO₂ como contaminante primario relacionado con el tráfico [22]; O₃ como contaminante secundario formado por reacciones fotoquímicas [23]; SO₂ como marcador de emisiones industriales o energéticas [24]; y PM_{2.5} como indicador del material particulado fino, asociado de forma sólida con aumentos de mortalidad y morbilidad respiratoria y cardiovascular [23, 25]. Esta selección equilibrada permite abordar múltiples fuentes y procesos de contaminación, facilitando así un estudio más representativo y útil para el diagnóstico ambiental.

3.3.3. Análisis temporal

El análisis temporal tiene como objetivo caracterizar el comportamiento dinámico de los contaminantes atmosféricos a lo largo del tiempo. Para ello, se han considerado tanto patrones estacionales generales como diferencias entre días laborables y fines de semana, utilizando los datos horarios agregados de las estaciones oficiales de medición en Valencia durante el periodo 2022–2024.

Evolución mensual global

En primer lugar, se ha representado la evolución mensual media de las concentraciones de los principales contaminantes: NO₂, O₃, PM_{2.5} y SO₂, agregando todas las estaciones de forma conjunta. El propósito es detectar tendencias estacionales generales.

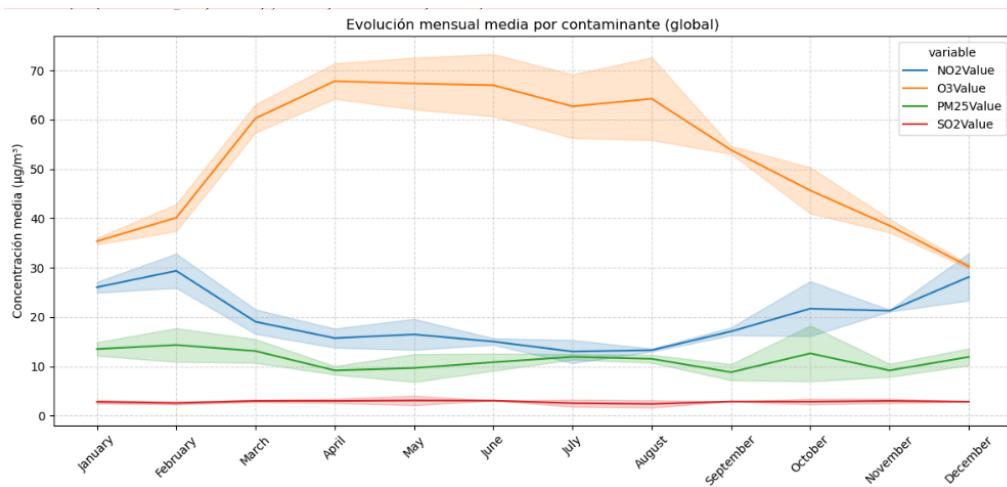


Figura 3.9: Evolución mensual media por contaminante (global)

En la Figura 3.9 se aprecian patrones diferenciados: el NO₂ muestra máximos durante los meses fríos, consistente con el uso de calefacción y menor dispersión atmosférica, mientras que el O₃ presenta un comportamiento inverso, con picos en los meses cálidos debido a su formación fotoquímica favorecida por la radiación solar. El PM_{2.5} refleja una distribución más homogénea, aunque con ciertos descensos estacionales, y el SO₂ mantiene concentraciones bajas sin estacionalidad clara.

Para una visión más detallada, se estudió la evolución mensual por estación para los cuatro contaminantes clave. Estos gráficos permiten identificar diferencias geográficas y comportamientos atípicos.

En el caso del NO₂, estaciones como A01_AVFRANCIA, A07_VALENCIACENTRE y A10_OLIVERETA mostraron valores elevados sostenidos, mientras que otras como A08_DR_LLUCH o A09_CABANYAL presentaron niveles más bajos, compatibles con ubicaciones menos urbanizadas. Para el O₃, se observó una marcada estacionalidad con máximos durante la primavera y el verano; sin embargo, la cobertura de datos fue incompleta en varias estaciones, lo que limitó su análisis. El PM_{2.5} presentó curvas irregulares con picos puntuales, especialmente en estaciones como A04_PISTASILLA, lo que podría haberse relacionado con intrusiones de polvo o condiciones meteorológicas locales.

En general, su estacionalidad fue menos evidente. Por último, el SO₂ mostró una baja cobertura y una señal débil, con valores dispersos e inestables en estaciones como A02_BULEVARDSUD o A03_MOLISOL, lo que redujo significativamente su utilidad analítica.

Por razones de espacio y claridad, se ha optado por describir los patrones observados sin incluir en la memoria los gráficos mensuales por estación. No obstante, estos fueron analizados durante el desarrollo del trabajo y se encuentran disponibles como material complementario.

Comparativa entre días laborables y fines de semana

Se analizó la diferencia de concentración media de contaminantes entre días laborables y fines de semana como proxy de la influencia de la actividad humana.

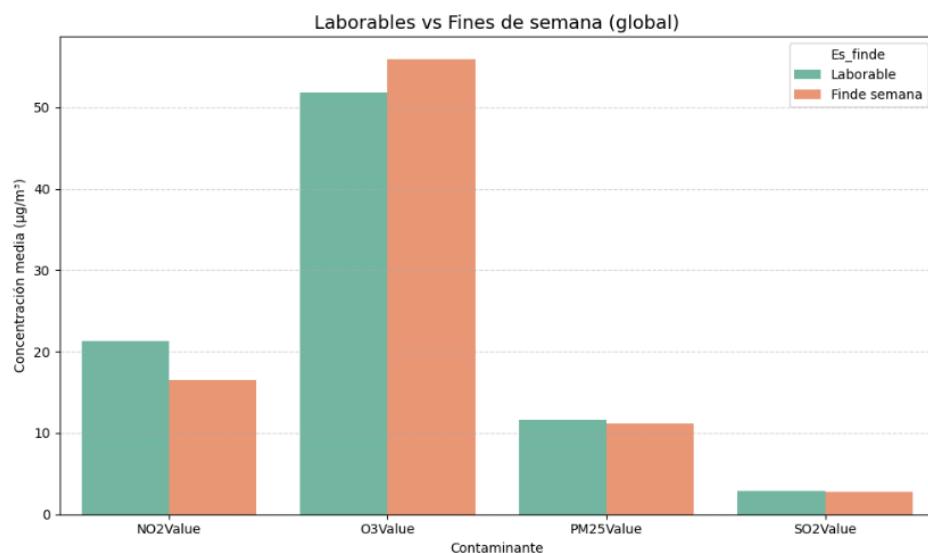


Figura 3.10: Comparativa entre días laborables y fines de semana (global)

Como se observa en la Figura 3.10, el NO₂ disminuye de forma significativa en fines de semana, lo que refleja su fuerte vínculo con el tráfico urbano. Por el contrario, el O₃ tiende a aumentar, debido al llamado “efecto fin de semana” causado por la reducción de NO, el cual reacciona con ozono y lo elimina. Los contaminantes PM_{2.5} y SO₂ no muestran diferencias sustanciales, lo que apunta a fuentes más constantes o no ligadas directamente a la movilidad.

También se exploró la variación de concentración por estación según el tipo de día (laborable o fin de semana). De nuevo, se identifican diferencias claras en NO₂ y O₃, aunque la magnitud y consistencia de estas variaciones difieren según la ubicación. Dado que los resultados son consistentes con el análisis global, se ha decidido no incluir estas figuras en la memoria para no sobrecargar visualmente el documento, pero pueden consultarse en los anexos.

3.3.4. Análisis STL

Se aplicó la descomposición STL a las series temporales de concentración de NO₂, O₃, PM_{2.5} y SO₂ en cada estación de monitoreo, separando en cada caso la tendencia de

largo plazo, el componente estacional recurrente anual y el residuo irregular. Este análisis permitió identificar patrones temporales diferenciados por contaminante y evidenció contrastes espaciales significativos entre estaciones.

Análisis STL del contaminante NO₂. El STL global de NO₂ (Figura 3.11) evidenció un descenso urbano suave, atribuible a la renovación del parque móvil y la implantación de Zonas de Bajas Emisiones [22?]. Se mantuvo la clásica estacionalidad invierno–verano y el residuo reflejó el fuerte desplome durante el confinamiento por la COVID-19, seguido de picos breves ligados a estabilidad atmosférica u obra viaria [?].

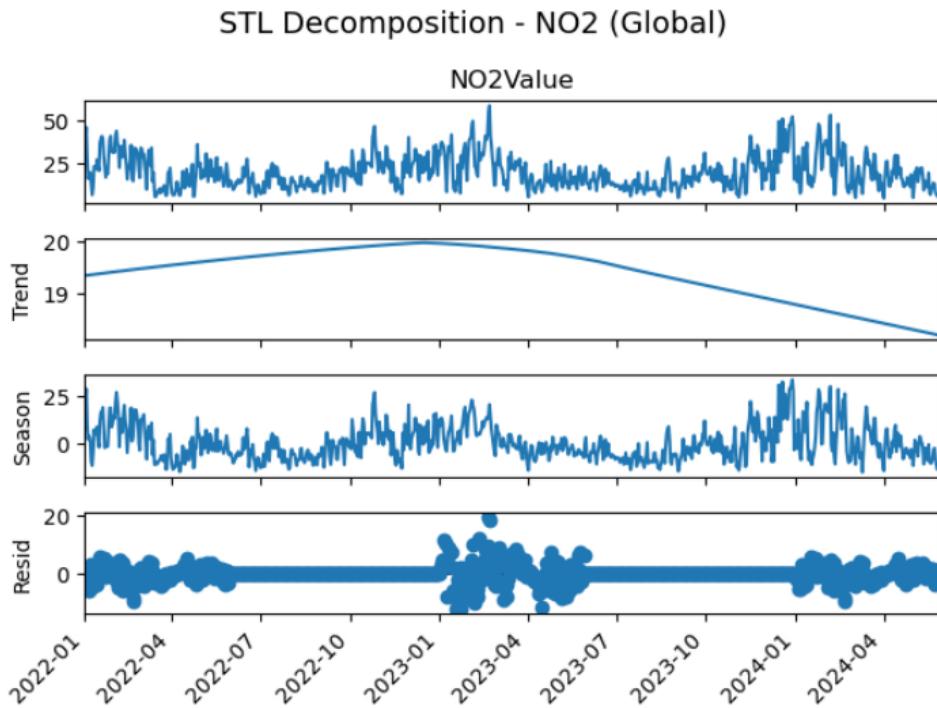


Figura 3.11: Descomposición STL global de NO₂

En el análisis por estación se distinguieron tres comportamientos: (i) las estaciones de tráfico intenso, A02_BULEVARD SUD y el eje A10_OLIVERETA - A11_PATRAIX, registraron un descenso continuo, reflejo directo de las restricciones viarias y la renovación del parque móvil; (ii) puntos de fondo urbano como A01_AVFRANCIA y A03_MOLISOL mostraron incrementos suaves, indicativos de menor “titulación” de NO₂ cuando el tráfico se desplazó a ejes periféricos; y (iii) estaciones con dinámica mixta presentaron patrones específicos: A04_PISTASILLA describió una trayectoria en “U” coronada por un pico asociado a la intrusión sahariana de marzo 2022 [?], A05_POLITECNIC descendió tras un máximo inicial ligado al ciclo académico, A06_VIVERS repuntó con la reactivación turística, A07_VALENCIACENTRE se estabilizó tras un leve ascenso y A09_CABANYAL moderó su incremento inicial al afianzarse los accesos portuarios. En todos los casos se mantuvo la marcada estacionalidad invierno–verano y se reprodujo el desplome extraordinario del confinamiento por la COVID-19 [?].

Análisis STL del contaminante O₃. El STL global de O₃ (Figura 3.12) reflejó un comportamiento estable, con una leve cresta a mediados de 2023 seguida de un descenso suave. La componente estacional mantuvo su patrón habitual: máximos claros en

primavera-verano y mínimos invernales, algo ya descrito por la OMS y los servicios meteorológicos nacionales para la cuenca mediterránea [23?]. El residuo fue relativamente pequeño y apenas registró picos puntuales durante olas de calor, lo que indica que la estacionalidad explica la mayor parte de la variabilidad diaria.

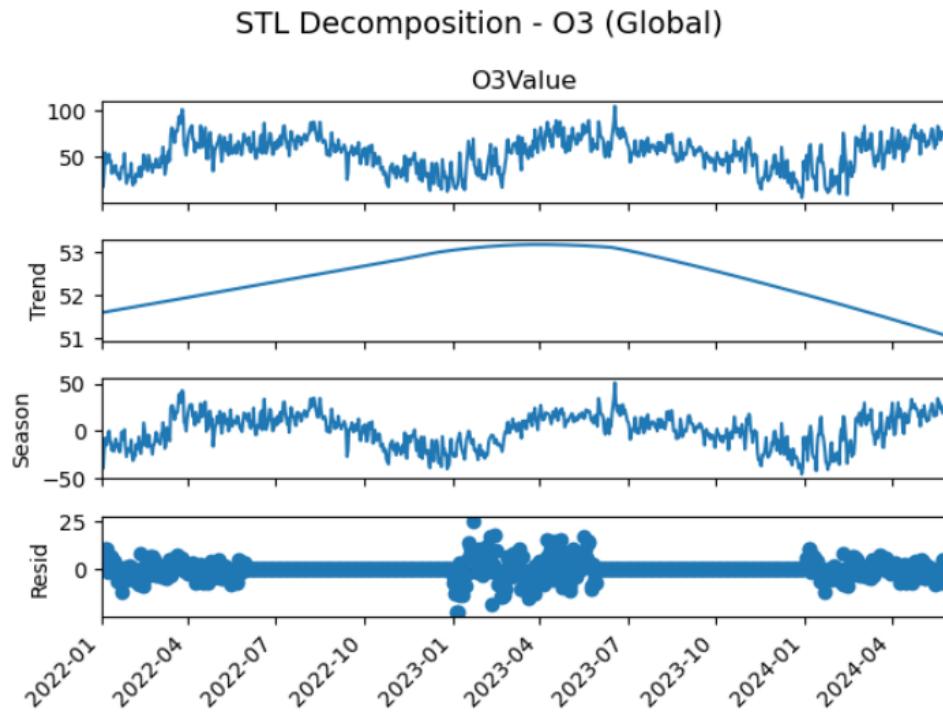


Figura 3.12: Descomposición STL global de O_3

El análisis por estación, limitado a los puntos con datos de O_3 , mostró ligeros contrastes. En A01_AVFRANCIA la tendencia presentó una suave joroba y terminó en valores similares a los iniciales, mientras que A02_BULEVARDSUD reprodujo el mismo perfil con amplitud ligeramente inferior. A03_MOLISOL mostró un descenso gradual durante todo el periodo, indicativo de una titulación algo mayor por la reducción local de precursores; por el contrario, A04_PISTASILLA evidenció un aumento entre 2022–2023 y un descenso posterior, probablemente moldeado por la advección de masas de aire más limpias al final del intervalo. A05_POLITECNIC permaneció estable hasta mediados de 2023 y luego decreció con suavidad, mientras que A06_VIVERS fue la única estación con tendencia claramente ascendente, quizás influida por la menor titulación de NO_x derivada del desvío progresivo del tráfico hacia otros ejes. El resto de estaciones, incluida A07_VALENCIACENTRE, carecieron de registros de O_3 , por lo que no se profundizó en su análisis. Pese a estas diferencias, todas las series locales conservaron la estacionalidad primavera-verano característica del ozono mediterráneo.

Análisis STL del contaminante PM_{2.5}. El análisis de la serie temporal de PM_{2.5} a escala global (Figura 3.13) mostró una tendencia decreciente, aunque con un comportamiento irregular, reflejo de episodios puntuales de contaminación. Este patrón es consistente con lo esperado en zonas urbanas, donde las partículas finas presentan incrementos esporádicos debido a diversas fuentes, como la circulación del tráfico y las actividades industriales, entre otras. La estacionalidad fue notoria, con picos en invierno y menores concentraciones en verano. Además, el residuo presentó variaciones significativas, lo que

indica la presencia de episodios extremos, como intrusiones de polvo sahariano o eventos meteorológicos adversos.

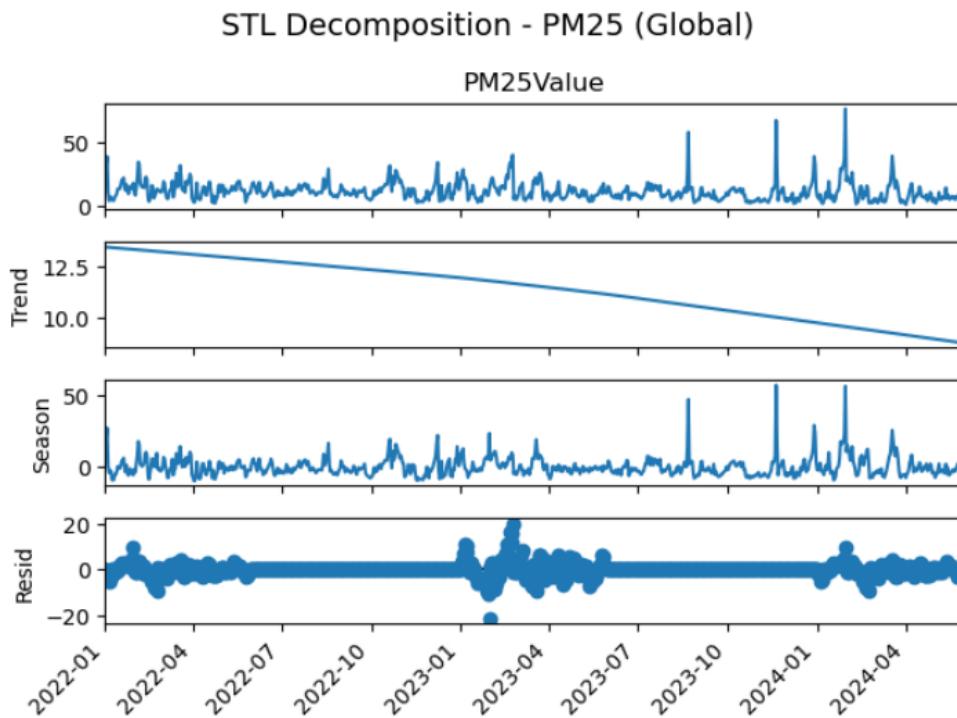


Figura 3.13: Descomposición STL global de PM_{2.5}

Por estación, el análisis de PM_{2.5} mostró diferencias significativas. En A01_AVFRANCIA, la tendencia fue descendente de forma continua, mientras que la estacionalidad mostró un ciclo claro, con picos durante los meses más fríos. En A02_BULEVARD_SUD, la tendencia siguió el mismo patrón, pero con menor intensidad, lo que sugiere una reducción más eficiente en las emisiones locales. A03_MOLISOL mostró una tendencia decreciente más pronunciada, pero con picos residuales que sugieren eventos más extremos de contaminación. A04_PISTASILLA presentó picos inusuales que podrían estar relacionados con episodios de contaminación transitoria, como el polvo sahariano documentado en 2022 [?].

En A05_POLITECNIC, la tendencia fue más estable, con fluctuaciones moderadas asociadas a la actividad universitaria, mientras que A06_VIVERS mostró un repunte hacia el final del período, posiblemente vinculado al retorno del turismo y los cambios en la dinámica del tráfico. A07_VALENCIACENTRE y A08_DR_LLUCH presentaron una tendencia descendente moderada, lo que sugiere una disminución en las emisiones, pero con una estacionalidad muy marcada. En A09_CABANYAL y A10 OLIVERETA, la tendencia fue moderadamente descendente, pero con picos residuales que podrían estar relacionados con la estacionalidad del tráfico y eventos locales.

Análisis STL del contaminante SO₂. El comportamiento global de SO₂ en Valencia mostró una tendencia ascendente leve, lo que podría estar relacionado con un aumento gradual de las emisiones o cambios en las condiciones de dispersión del aire. La estacionalidad fue especialmente pronunciada en los meses de primavera y verano, con picos en los valores de SO₂ debido a la acumulación de contaminantes por fenómenos de inversión

térmica. Los residuos reflejan picos aislados ligados a eventos meteorológicos, intrusiones de aire limpio y la reducción de emisiones locales [? ?].

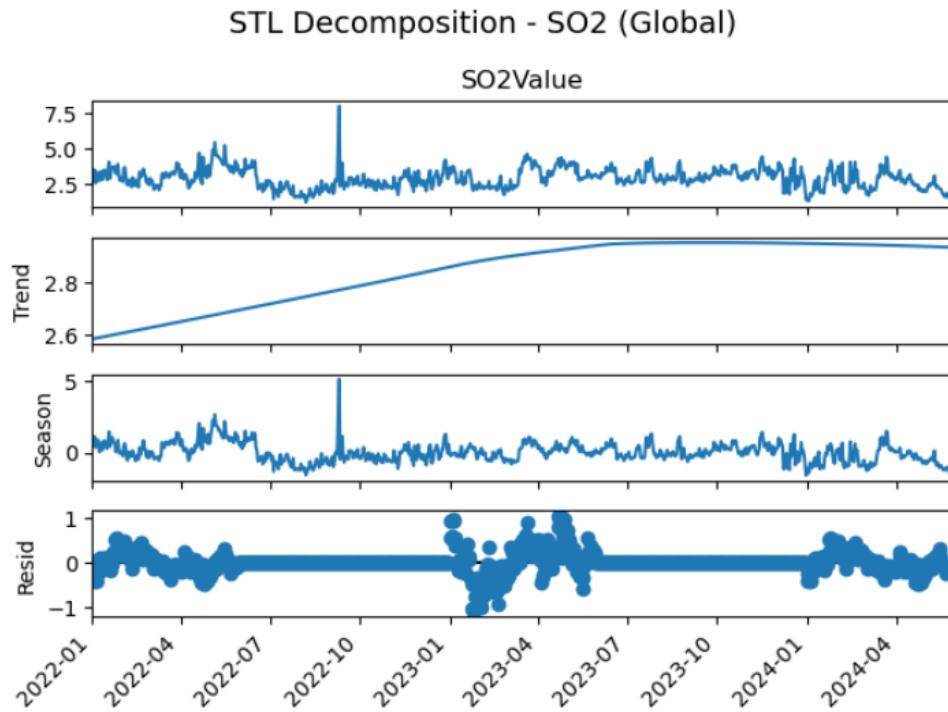


Figura 3.14: Descomposición STL global de SO₂

En las estaciones analizadas, se observó una variabilidad en las tendencias de SO₂. En general, las estaciones de zonas más cercanas al tráfico vehicular, como A01_AVFRANCIA y A02_BULEVARDSUD, mostraron una ligera tendencia ascendente, lo que podría reflejar una mayor acumulación de contaminantes debido a las emisiones del tráfico y las condiciones de ventilación del aire. Las estaciones más alejadas, como A04_PISTASILLA, presentaron un comportamiento algo errático, con picos asociados a fenómenos meteorológicos específicos, como intrusiones de aire limpio o cambios en la actividad local. A03_MOLISOL mostró una disminución en la tendencia, lo que sugiere una menor influencia de las fuentes locales de SO₂, mientras que A05_POLITECNIC se mantuvo estable, con ligeras fluctuaciones relacionadas con las variaciones del tráfico universitario.

A06_VIVERS mostró una ligera tendencia descendente, posiblemente debido a una mejora en las condiciones de dispersión y una reducción en las emisiones locales. En términos de estacionalidad, todas las estaciones mantuvieron el patrón típico de picos en los meses de primavera y verano, con fluctuaciones asociadas a la actividad local y a las condiciones meteorológicas [22?].

3.4. Umbral de anomalías

3.4.1. Umbrales estadísticos

Tras revisar los métodos de umbral estadístico descritos en el estado del arte, se aplicaron las cuatro técnicas (3 Sigmas, Percentil 95, Hampel y Boxplot) a las series diarias de NO₂, O₃, PM_{2.5} y SO₂.

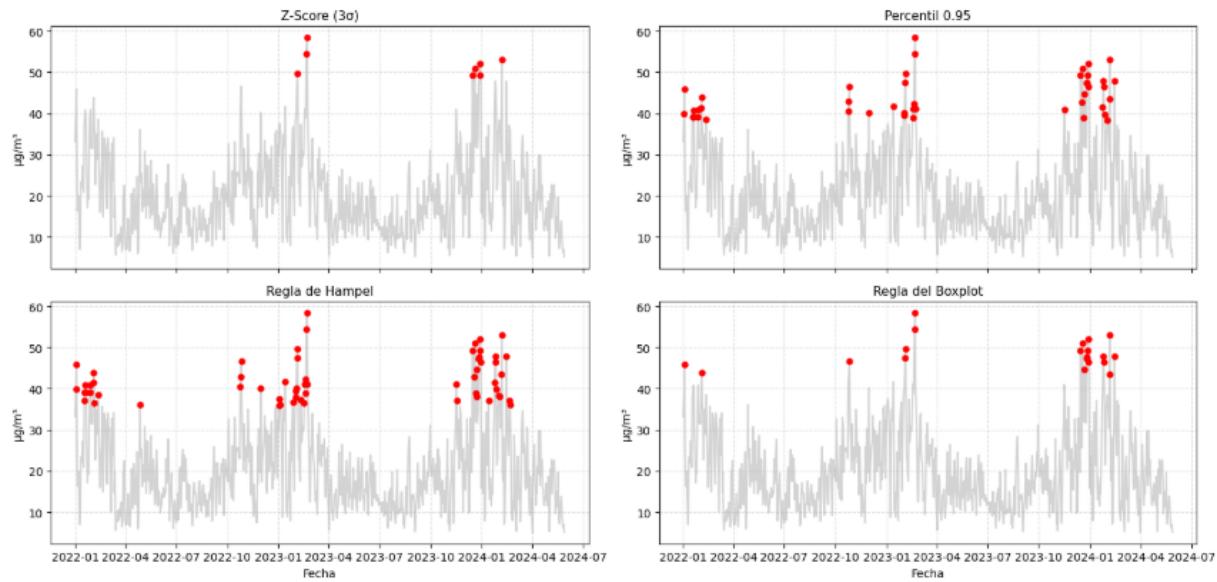


Figura 3.15: Detección de anomalías en NO_2 mediante 4 umbrales estadísticos.

En la Figura 3.15 se aprecia que la regla de las *3 Sigmas* únicamente identifica los picos invernales más pronunciados, coincidentes con episodios de inversión térmica y tráfico intenso. El enfoque basado en el *percentil 95* amplía la detección al 5% superior de la serie, de modo que clasifica como anómalos numerosos días fríos con concentración elevada de NO_2 . Por su parte, la regla de *Hampel* reproduce prácticamente todos los eventos extremos detectados por el Z-score y, además, incorpora valores moderados que sobresalen del rango intercuartílico sin verse afectada por la asimetría de la distribución. Finalmente, la regla del *Boxplot* (IQR) ofrece un resultado intermedio: captura tanto picos ligados a inversiones térmicas como repuntes asociados a incidencias de tráfico, sin llegar a la sobre-detección del percentil 95.

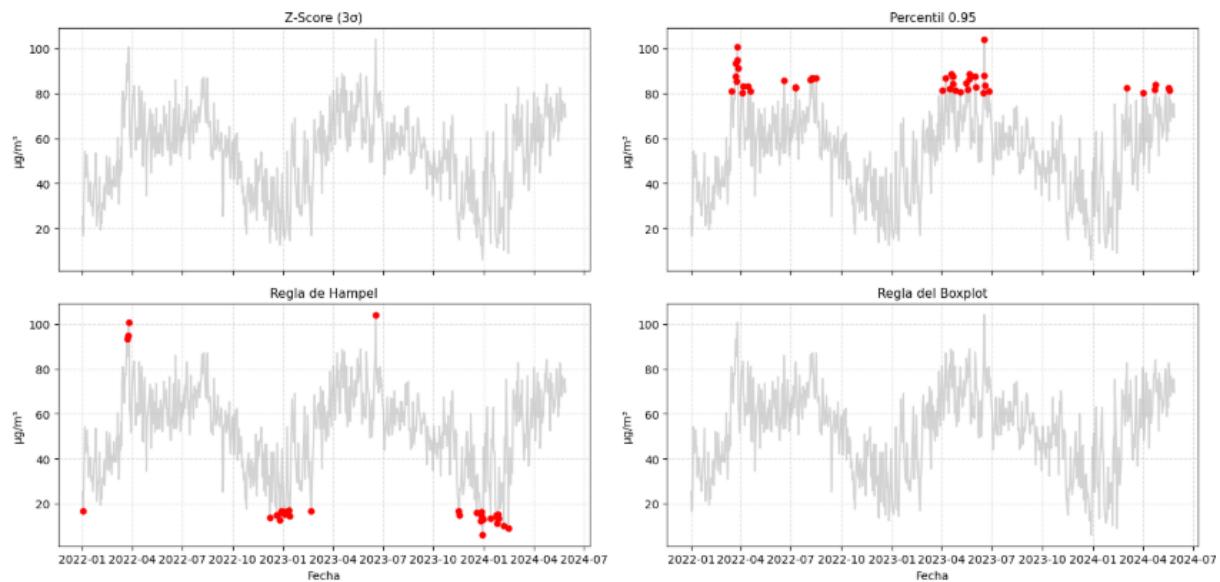


Figura 3.16: Detección de anomalías en O_3 mediante 4 umbrales estadísticos.

En la serie de O_3 , la regla de *3 Sigmas* sólo identifica los máximos estivales más extremos, típicos de episodios de ola de calor. El *percentil 95* amplía la detección a prá-

ticamente todos los días de verano con alta insolación, por lo que tiende a sobre-etiquetar valores propios de la estacionalidad normal. La regla de *Hampel* actúa como término medio: resalta los episodios fotoquímicos significativos sin un exceso de falsos positivos, mientras que el *IQR* añade algunos repuntes primaverales altos pero descarta los veraniegos moderados..

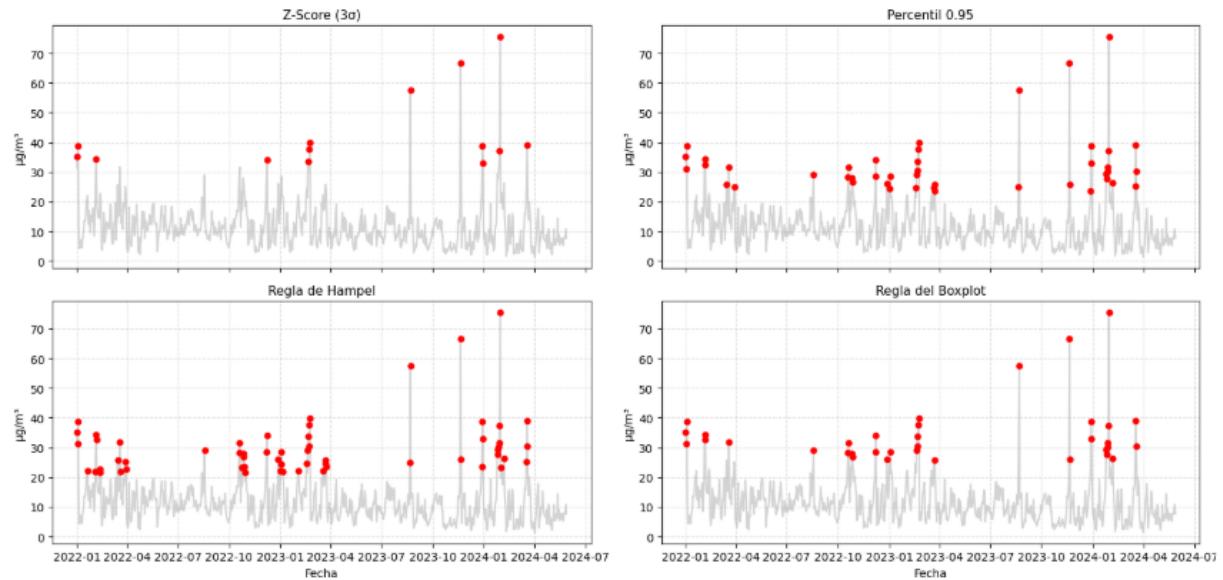


Figura 3.17: Detección de anomalías en PM_{2.5} mediante 4 umbrales estadísticos.

Para PM_{2.5}, el método de las *3 Sigmas* señala sobre todo los picos asociados a intrusiones de polvo sahariano u otros episodios muy intensos. El *percentil 95* incorpora además numerosos días invernales dominados por calefacción y estabilidad atmosférica. La regla de *Hampel* detecta los mismos eventos extremos y añade repuntes intermedios sin distorsión por la cola larga de la distribución, mientras que el *IQR* coincide en gran medida con Hampel gracias a su robustez frente a asimetrías.

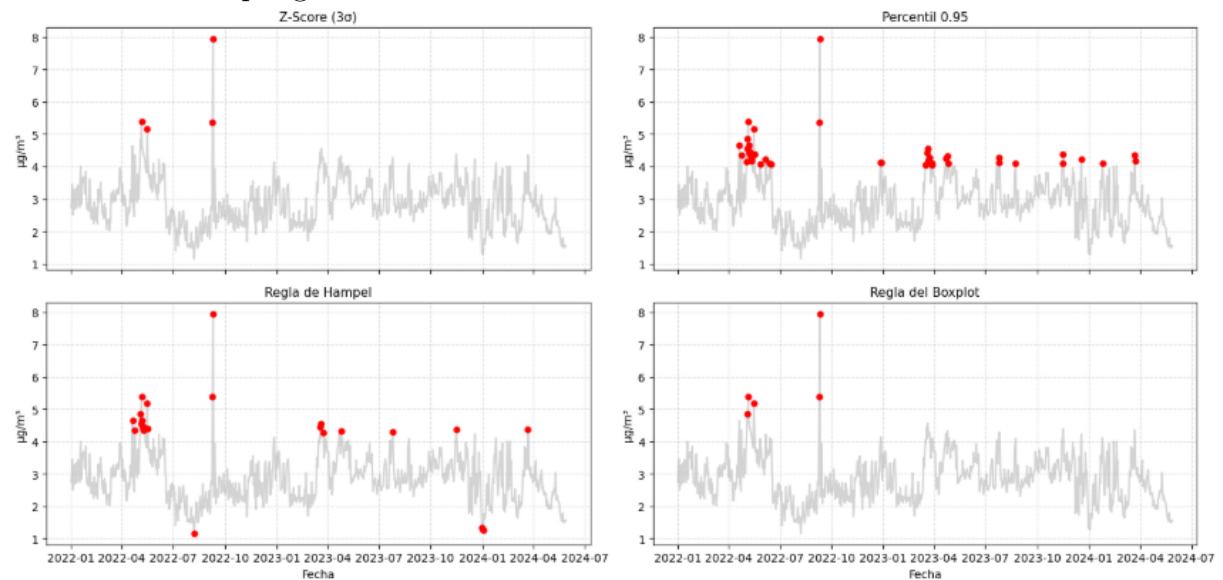


Figura 3.18: Detección de anomalías en SO₂ mediante 4 umbrales estadísticos.

En SO₂, donde los valores medios son bajos y los picos responden a fuentes puntuales, la regla de *3 Sigmas* sólo marca las descargas industriales más acusadas. El *percentil 95*

fija un umbral relativamente bajo y clasifica de forma constante el 5 % superior de días como anómalos. La regla de *Hampel* captura los mismos picos industriales y algunos incrementos moderados sin verse afectada por la cola asimétrica, mientras que el *IQR* ofrece un compromiso similar aunque descarta los repuntes de menor magnitud.

A nivel local se observa un patrón recurrente: las estaciones de tráfico (A01_AVFRANCIA, A02_BULEVARDSUD) concentran la mayor parte de outliers invernales de NO₂, mientras que los picos de O₃ se distribuyen de forma más homogénea en verano. PM_{2.5} y SO₂ muestran outliers dispersos y dominados por episodios puntuales.

Para no sobrecargar la memoria con todas las combinaciones estación-contaminante, se han elegido cuatro ejemplos claros: (i) A01_AVFRANCIA-SO₂, situada junto a vías muy transitadas y con picos ocasionales de emisiones; (ii) A03_MOLISOL-PM_{2.5}, un entorno suburbano donde las intrusiones de polvo elevan puntualmente las partículas; (iii) A08_DR_LLUCH-NO₂, zona costera que muestra fuertes subidas de NO₂ en los meses fríos; y (iv) A04_PISTASILLA-O₃, un área verde con poco tráfico donde el ozono alcanza sus máximos en verano.

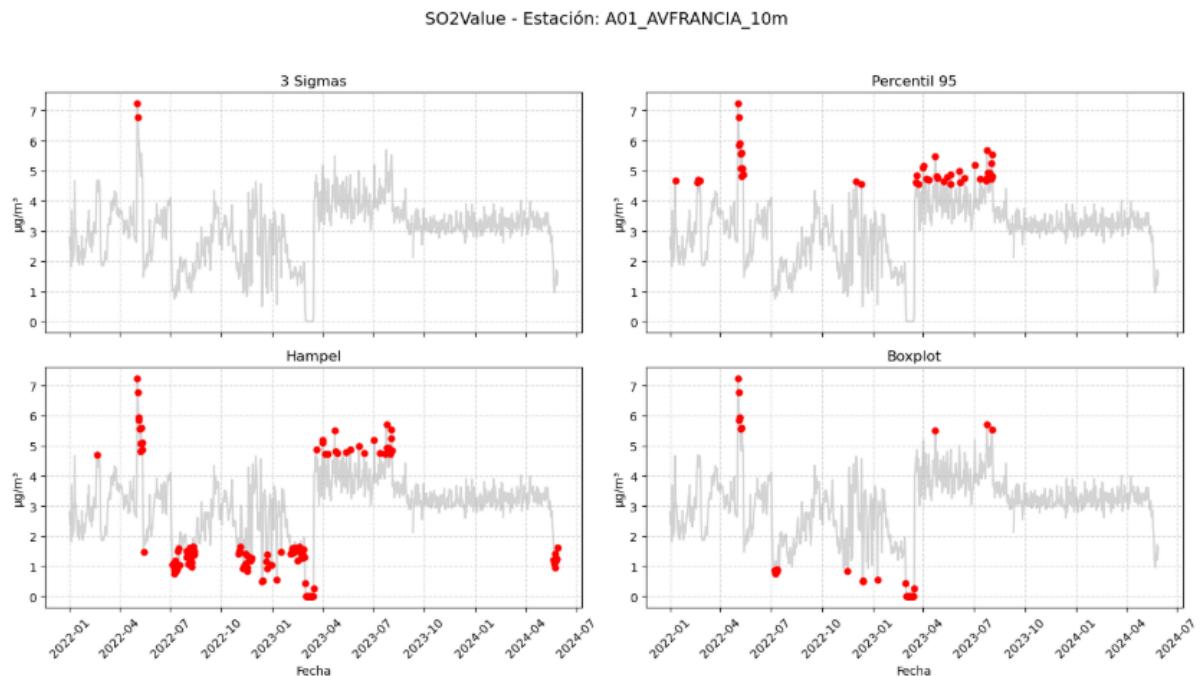


Figura 3.19: Detección de anomalías en SO₂, estación A01_AVFRANCIA.

Los cuatro métodos coinciden en señalar dos picos claros de SO₂ al inicio de la serie, atribuibles a emisiones puntuales; 3 Sigmas apenas marca esos valores extremos, mientras que Hampel e IQR añaden repuntes menores a mitad de año, mostrando mayor sensibilidad sin sobredetección.

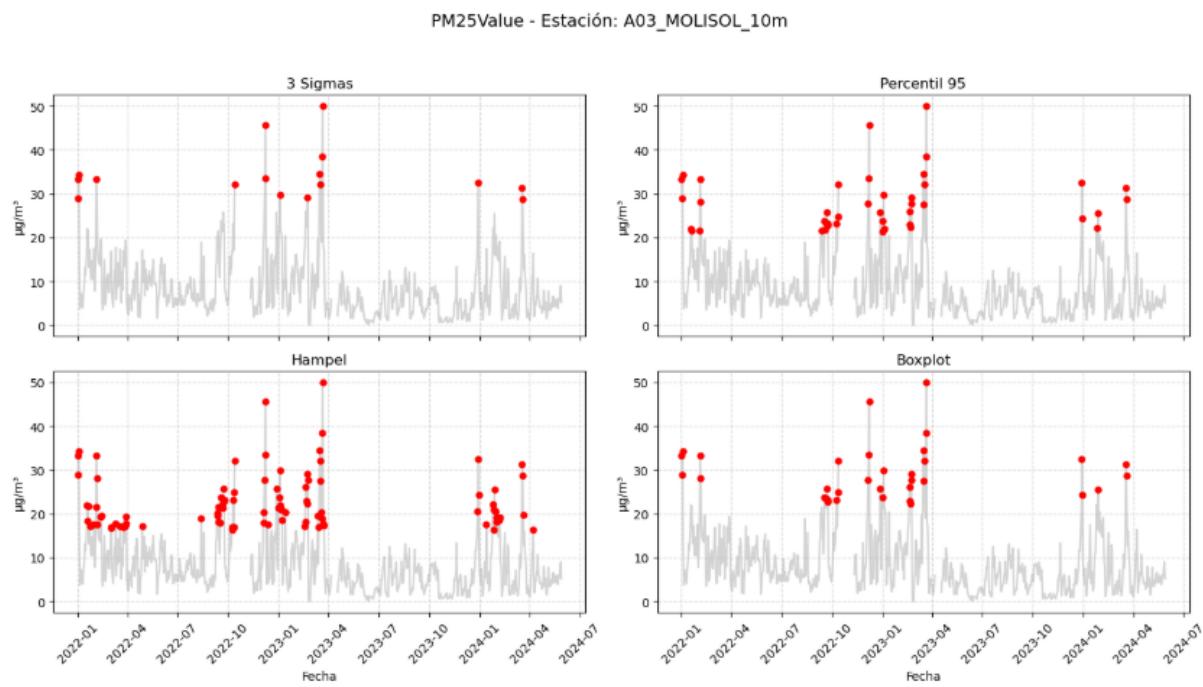


Figura 3.20: Detección de anomalías en PM_{2.5}, estación A03_MOLISOL.

En A03_MOLISOL, la regla de las *3 Sigmas* se limita a los picos más altos, mientras que el *percentil 95* extiende la etiqueta a un mayor número de días invernales con partículas elevadas. Los métodos robustos, *Hampel* y *IQR*, capturan tanto los máximos de polvo como varios repuntes otoñales sin llegar a sobre-detectar, ofreciendo un equilibrio entre sensibilidad y especificidad.

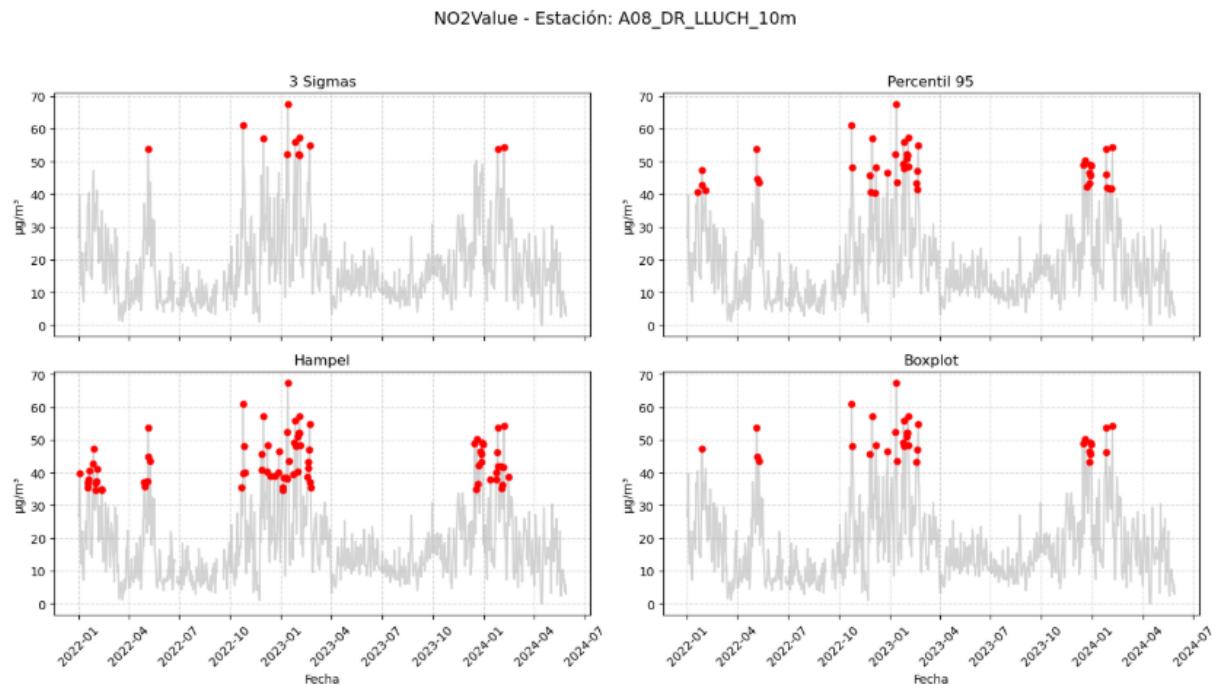


Figura 3.21: Detección de anomalías en NO₂, estación A08_DR_LLUCH.

Los picos invernales de NO₂ aparecen en los cuatro algoritmos, pero el percentil 95,

Hampel e IQR marcan muchos valores típicos de temporada, mientras que 3 Sigmas sólo destaca los días de concentración más alta.

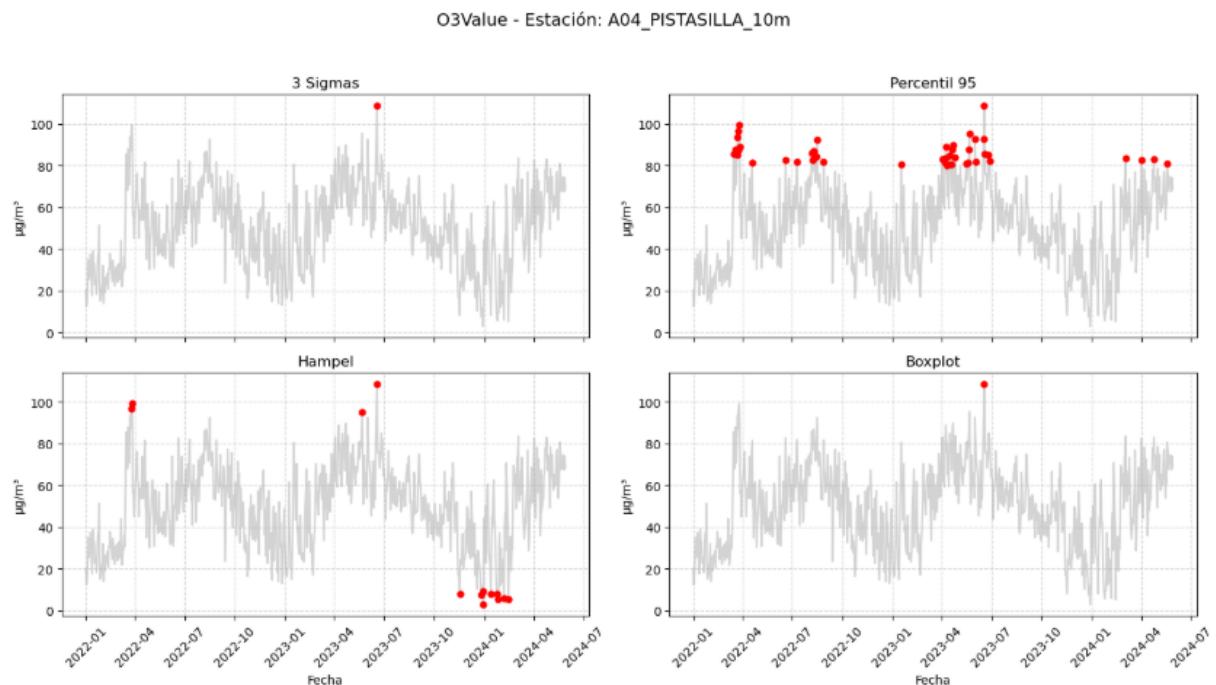


Figura 3.22: Detección de anomalías en O₃, estación A04_PISTASILLA.

El *percentil 95* marca numerosos días de verano con concentraciones altas de ozono, mientras que la regla de *3 Sigmas* se fija únicamente en el valor máximo absoluto. El método *Hampel* señala ese pico y, además, varios descensos invernales inusualmente bajos, mientras que el umbral *IQR* se comporta de forma muy conservadora y sólo detecta el día más alto del verano.

El método de las *3 Sigmas* resultó ser el más conservador, ya que únicamente detectó unos pocos eventos muy pronunciados, principalmente picos vinculados a intrusiones de polvo o incidencias industriales. El enfoque basado en el *percentil 95*, al señalar siempre el 5 % superior de los datos, identificó numerosos días invernales de NO₂ y veraniegos de O₃, pero tendió a sobredimensionar las anomalías en series con estacionalidad clara. La regla de *Hampel* ofreció un buen equilibrio entre robustez y sensibilidad, capturó los mismos episodios extremos que el Z-score, pero añadió valores moderados fuera del rango intercuartílico sin verse afectada por la forma de la distribución. Por último, la regla del *Boxplot* (IQR) funcionó especialmente bien para PM_{2.5} y SO₂, caracterizadas por distribuciones asimétricas y colas largas.

3.4.2. Comité de anomalías

Para reducir la tasa de falsos positivos generada por la aplicación aislada de cada umbral estadístico, se construyó un *comité de anomalías*.

El procedimiento combina los cuatro detectores empleados (3 Sigmas, percentil 95, Hampel e IQR) y marca un valor como anómalo únicamente si, al menos, tres de ellos (75 %) coinciden. Este umbral estricto se eligió para disminuir la influencia de métodos

demasiado permisivos y preservar la diversidad de criterio que aportan los detectores robustos frente a distribuciones sesgadas.

	NO ₂	PM _{2.5}	O ₃	SO ₂
3 Sigmas	8	14	0	4
Percentil 95	44	44	44	44
Hampel	60	58	31	24
IQR	20	35	0	5
Comité (≥ 3 votos)	20	35	0	5

Tabla 3.5: N° de valores anómalos detectados por cada método (global) y por el comité.

El resultado global (Tabla 3.5) revela el efecto filtrador del comité: de los 44–60 eventos marcados por los métodos más permisivos (Percentil 95 y Hampel), sólo entre 5 y 35 superan el consenso mínimo de tres detectores.

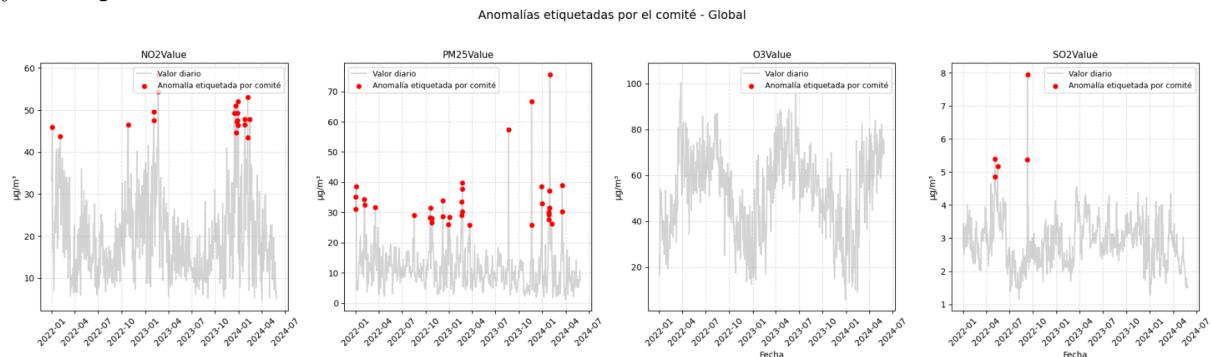


Figura 3.23: Anomalías detectadas por el comité para cada contaminante (global).

El comité valida 20 episodios de NO₂, la mayoría concentrados en los meses fríos; reconoce 35 repuntes de PM_{2.5}, distribuidos sobre todo en primavera y en algunos días de invierno; no confirma anomalías de ozono al no superar éste el umbral de tres votos, y señala únicamente cinco picos de SO₂ en todo el periodo analizado.

Vemos ahora un análisis del comportamiento del comité para los datos separados por estación. La Tabla 3.6 cuantifica, para cada estación, cuántas observaciones marca cada detector y cuántas superan el consenso del comité.

Estación	3 Sigmas	Percentil 95	Hampel	IQR	Comité
A01_AVFRANCIA	36	175	300	106	82
A02_BULEVARSDSUD	13	132	136	40	40
A03_MOLISOL	30	168	255	86	84
A04_PISTASILLA	25	176	177	77	77
A05_POLITECNIC	23	176	263	111	98
A06_VIVERS	17	131	204	50	37
A07_VALENCIACENTRE	20	88	112	41	41
A08_DR_LLUCH	26	88	135	61	61
A09_CABANYAL	40	88	165	73	73
A10 OLIVERETA	13	78	87	29	27
A11_PATRAIX	15	74	78	29	28

Tabla 3.6: Total de valores señalados como anómalos por cada método (suma de NO₂, PM_{2.5}, O₃ y SO₂) frente a los validados por el comité (≥ 3 votos) en cada estación.

La Tabla 3.6 pone de manifiesto un patrón claro. Hampel es siempre el detector más generoso, alcanza 300 avisos en A01_AVFRANCIA, mientras que la regla de las 3 Sigma marca apenas una fracción de esos valores; Percentil 95 e IQR se mantienen en cifras intermedias, con variaciones más moderadas entre estaciones. Cuando se impone el consenso de tres votos, el comité elimina, de media, casi siete de cada diez alertas iniciales: la poda abarca desde un 55 % en A04_PISTASILLA y A08_DR_LLUCH hasta un 82 % en A06_VIVERS, con un recorte medio del 68 % en toda la red. Aun así, persisten diferencias claras entre emplazamientos: A05_POLITECNIC (98 episodios validados), A03_MOLISOL (84) y A01_AVFRANCIA (82) se mantienen como los puntos con mayor carga residual de anomalías, seguidos por A04, A08 y A09 (61–77 casos). En el extremo opuesto, A10 OLIVERETA y A11_PATRAIX apenas superan la veintena de incidencias confirmadas. Estas cifras ofrecen un mapa objetivo que ayuda a priorizar las estaciones y contaminantes que precisan una revisión más detallada o medidas de control específicas.

Tras comparar cuántas alertas genera cada detector y cuántas sobrevive el filtrado del comité, conviene desglosar ahora dónde y en qué contaminante se concentran los episodios validados. La Tabla 3.7 muestra, para cada estación, el número final de anomalías aprobado por consenso en NO₂, PM_{2.5}, O₃ y SO₂.

Estación	NO ₂	PM _{2.5}	O ₃	SO ₂
A01_AVFRANCIA	32	41	0	9
A02_BULEVARDSUD	31	0	1	8
A03_MOLISOL	17	34	2	31
A04_PISTASILLA	36	27	1	13
A05_POLITECNIC	44	40	1	13
A06_VIVERS	19	0	0	18
A07_VALENCIACENTRE	9	32	0	0
A08_DR_LLUCH	34	27	0	0
A09_CABANYAL	38	35	0	0
A10 OLIVERETA	5	22	0	0
A11_PATRAIX	6	22	0	0

Tabla 3.7: Número de valores anómalos validados por el comité en cada estación.

La distribución deja varias lecturas rápidas. NO₂ presenta sus recuentos más altos en los puntos de tráfico denso (hasta 44 episodios en A05_POLITECNIC y 38 en A09_CABANYAL), mientras que varias estaciones periféricas apenas superan la decena. PM_{2.5} domina numéricamente en casi la mitad de los emplazamientos, con picos de 41 eventos en A01 y 40 en A05, lo que confirma que las partículas finas siguen siendo el contaminante con más episodios extremos validados. Sin embargo, el ozono apenas rebasa el umbral de anomalía: sólo se registran cuatro casos en toda la red. El SO₂, por último muestra cifras apreciables solo en A03, A04 y A06 (9–31 eventos), mientras que en el resto las anomalías se reducen prácticamente a cero. En conjunto, estos resultados delimitan con claridad dónde conviene focalizar los análisis de causa y las acciones de control específicas por contaminante.

El comité de decisión demuestra ser una estrategia eficaz para integrar la información de varios detectores y generar una etiqueta de anomalía de mayor fiabilidad. A escala urbana reduce en más de un 60 % los outliers señalados por los métodos individuales y,

a escala local, clarifica qué estaciones y contaminantes requieren atención prioritaria en posteriores modelos supervisados o en la programación de campañas de control.

3.4.3. Umbral legal

Muchos valores límite de la Directiva 2008/50/CE sobre calidad del aire [?] y de su transposición al ordenamiento español mediante el Real Decreto 102/2011 [26] están formulados para promedios cada hora o cada ocho horas. Como los datos tratados en este trabajo son medias diarias, ha sido necesario reajustar dichos umbrales con el fin de conservar su carácter preventivo sin penalizar en exceso la suavización que introduce el promedio de 24 h. La Figura 3.24 se toma como referencia gráfica de los rangos originales.

SO ₂		PM _{2,5}		PM10		O ₃		NO ₂		CATEGORÍA DEL ÍNDICE
0	100	0	10	0	20	0	50	0	40	BUENA
101	200	11	20	21	40	51	100	41	90	RAZONABLEMENTE BUENA
201	350	21	25	41	50	101	130	91	120	REGULAR
351	500	26	50	51	100	131	240	121	230	DESFAVORABLE
501	750	51	75	101	150	241	380	231	340	MUY DESFAVORABLE
751-1250		76-800		151-1200		381-800		341-1000		EXTREMADAMENTE DESFAVORABLE

Figura 3.24: Rangos oficiales del Índice de Calidad del Aire para los cinco contaminantes principales en la Comunitat Valenciana.

Para NO₂ la legislación fija un valor límite horario de 200 µg m⁻³ que no debe superarse más de 18 veces al año; al promediar datos a 24 h el pico se atenúa, de modo que se adopta un umbral diario de 100 µg m⁻³, suficiente para señalar aquellas jornadas con varias horas críticas. En O₃ el objetivo sanitario se sitúa en 120 µg m⁻³ sobre la máxima móvil de ocho horas (límite: 25 días al año, media trianual); trasladado a promedios diarios se emplea 90 µg m⁻³, valor que reproduce razonablemente esa octohoraria típica de los días estivales. Para PM_{2,5} el tope legal anual es 25 µg m⁻³, mientras que las directrices de la OMS 2021 lo rebajan a 15 µg m⁻³ [23]; se adopta un umbral diario intermedio de 30 µg m⁻³, suficientemente estricto para aislar episodios anómalos sin resultar excesivamente conservador. Por último, en SO₂ el límite ya está definido en media diaria (125 µg m⁻³ con un máximo de tres superaciones anuales), por lo que se mantiene sin modificaciones.

Contaminante	Umbral legal original	Umbral diario usado
NO ₂	200 $\mu\text{g m}^{-3}$ (1 h)	100 $\mu\text{g m}^{-3}$
O ₃	120 $\mu\text{g m}^{-3}$ (8 h máx.)	90 $\mu\text{g m}^{-3}$
PM _{2.5}	25 $\mu\text{g m}^{-3}$ (24 h)	30 $\mu\text{g m}^{-3}$
SO ₂	125 $\mu\text{g m}^{-3}$ (24 h)	125 $\mu\text{g m}^{-3}$

Tabla 3.8: Umbrales legales originales y valores adaptados para series diarias.

Estos límites adaptados se emplean en paralelo a los umbrales estadísticos examinados en la sección anterior; juntos ofrecen un marco coherente para distinguir entre *anomalías estadísticamente extremas* e *incumplimientos normativos*, facilitando la priorización de la detección de distintas anomalías..

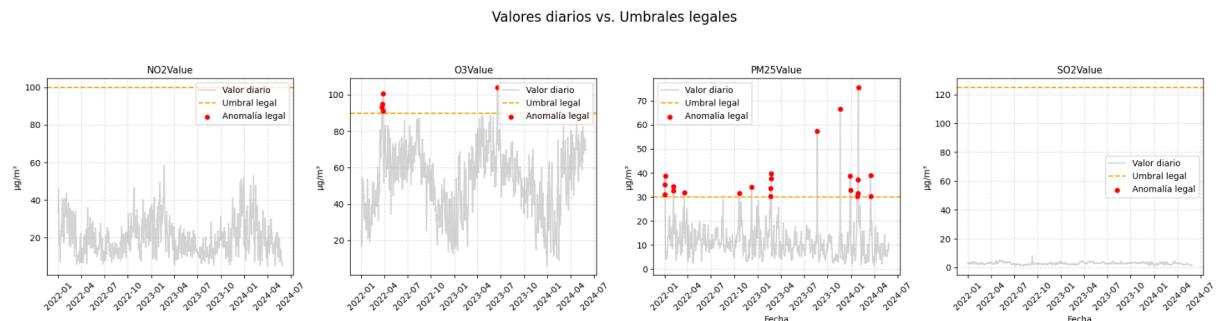


Figura 3.25: Valores diarios de NO₂, O₃, PM_{2.5} y SO₂ frente a los umbrales diarios adoptados (red completa, 2022–2024). El trazo discontinuo marca el límite legal adaptado; los puntos rojos señalan superaciones.

El gráfico global (Figura 3.25) permite comprobar de un vistazo qué contaminantes superan con más frecuencia los límites sanitarios. Durante el periodo 2022-2024 ningún día rebasó el umbral diario de NO₂ (100 $\mu\text{g m}^{-3}$), mientras que el ozono mostró algunos excesos puntuales en episodios fotoquímicos de primavera-verano.

Las partículas PM_{2.5} son, con diferencia, el contaminante que más a menudo sobrepasa el umbral de 30 $\mu\text{g m}^{-3}$, asociado sobre todo a intrusiones de polvo y episodios de combustión.

En cuanto al SO₂, las concentraciones diarias se mantuvieron muy por debajo del límite de 125 $\mu\text{g m}^{-3}$, sin una sola superación.

Estos resultados concuerdan con la tendencia general descrita en la Directiva europea de calidad del aire y su transposición española, donde las partículas finas y el ozono siguen siendo los principales retos sanitarios en el ámbito urbano [? 26, 23].

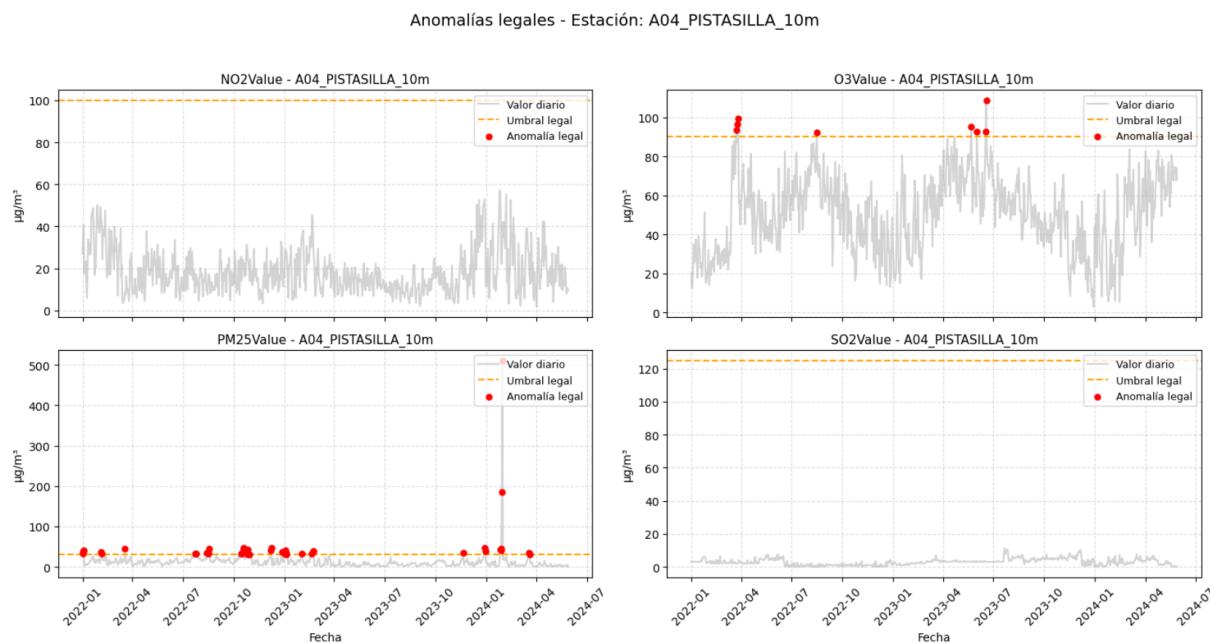


Figura 3.26: Ejemplo de aplicación de los umbrales diarios en la estación A04_PISTASILLA (2022-2024).

La Figura 3.26 ilustra cómo se traducen los umbrales en una estación concreta (A04_PISTASILLA, de carácter periurbano). NO₂ permanece siempre por debajo del límite, lo que confirma la escasa influencia del tráfico local. En O₃ se detectan varios picos estivales, todos ellos señalados como anomalía legal, que coinciden con episodios de alta radiación y estabilidad atmosférica.

PM_{2.5} muestra numerosas superaciones, incluido un repunte excepcional (enero 2024), evidenciando la utilidad operativa del umbral diario para identificar rápidamente días de mala calidad del aire.

Por último, SO₂ se mantiene muy por debajo del límite legal, corroborando la ausencia de fuentes significativas de azufre en la zona.

En conjunto, la combinación del análisis global y la verificación estación-a-estación confirma que los umbrales diarios adaptados son lo bastante sensibles para captar los episodios relevantes sin generar un número excesivo de falsas alarmas, facilitando así la labor de seguimiento operativo y la comunicación de riesgos a la población.

3.5. Análisis no supervisado

La exploración de patrones sin supervisión permite resumir la información multidimensional y detectar semejanzas entre estaciones sin necesidad de variables predictoras. Entre las técnicas más utilizadas destacan el Análisis de Componentes Principales (PCA), que proyecta los datos en un subespacio ortogonal de varianza máxima [27], y el agrupamiento *k*-means, que partitiona las observaciones en conjuntos compactos y disjuntos [?].

En los contaminantes diarios de Valencia se aplicaron ambos métodos con dos estrategias de tratamiento de ausencias: conjunto completo, donde los valores nulos se imputaron por la media de la columna, y conjunto filtrado, descartando las estaciones con más del 50 % de datos ausentes antes de la imputación. En ambos casos los contaminantes se

estandarizaron (media 0, desviación 1) para eliminar diferencias de escala.

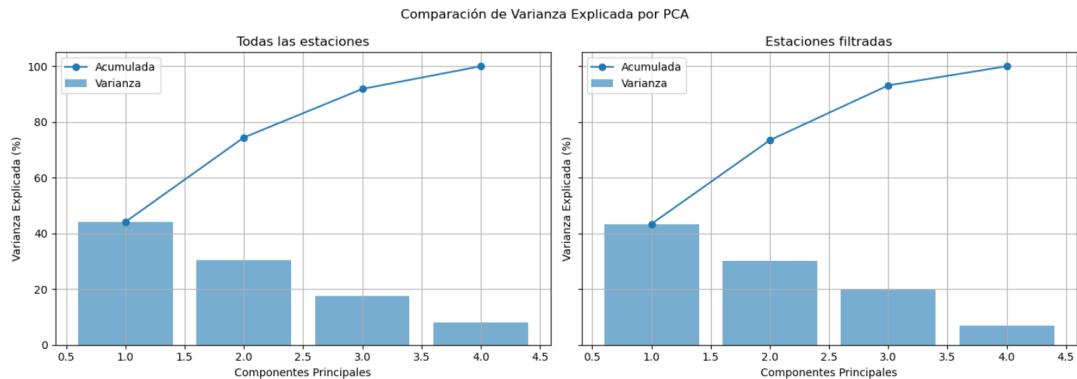


Figura 3.27: Porcentaje de varianza explicada por las componentes principales con todos los puntos de medida (izquierda) y con el subconjunto filtrado (derecha).

La Figura 3.27 muestra que las dos primeras componentes principales capturan cerca del 75 % de la variabilidad tanto en el conjunto completo como en el filtrado, lo que justifica su empleo para representar los datos en un plano bidimensional.

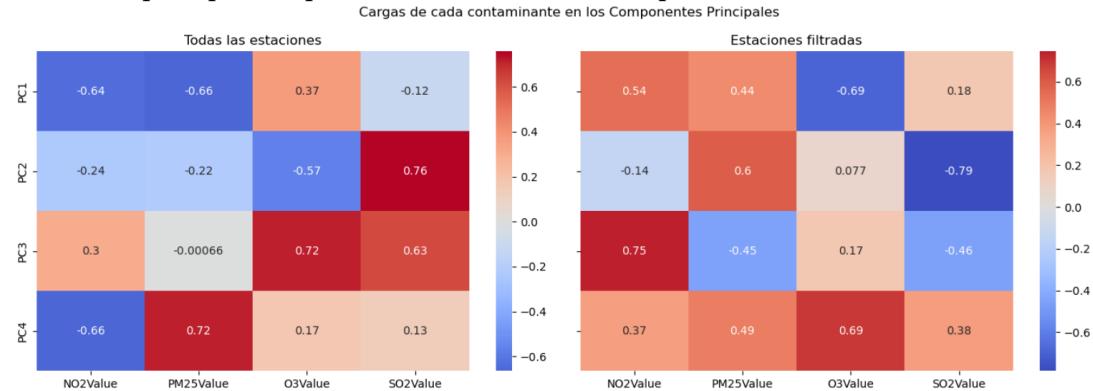


Figura 3.28: Matriz de cargas (contribución de cada contaminante) en las cuatro primeras componentes principales.

Las cargas de la Figura 3.28 permiten interpretar las nuevas dimensiones. En el análisis con *todas* las estaciones, la PC1 contrapone NO₂ / PM_{2.5} (*cargas negativas*) frente a O₃ (*carga positiva*), revelando el contraste habitual entre contaminación de tráfico y ozono. Tras eliminar los puntos con grandes vacíos de datos, la PC1 queda dominada por NO₂ y PM_{2.5} (*cargas positivas*) y O₃ (*carga negativa*), mientras la PC2 separa sobre todo la señal de SO₂, lo que sugiere una mayor diferenciación de las fuentes industriales.

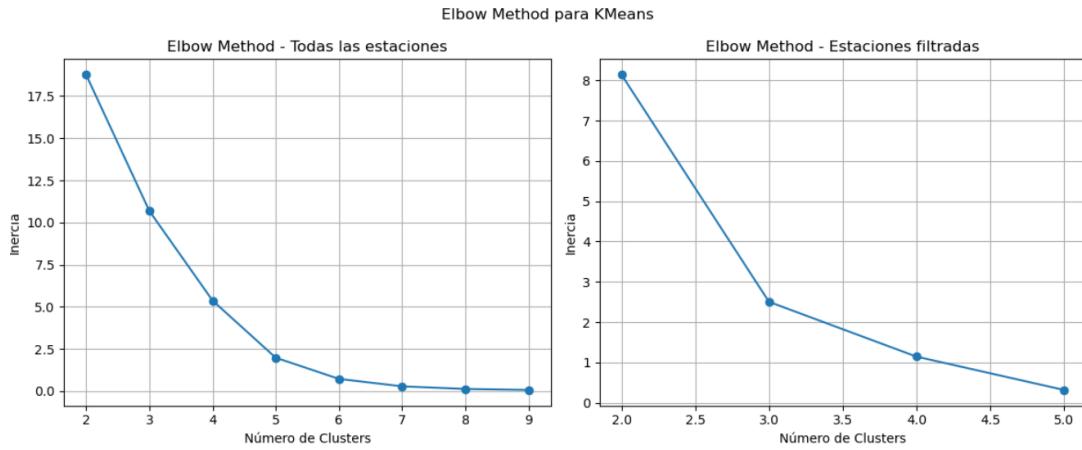


Figura 3.29: Método Elbow aplicado a las dos primeras componentes: inercia frente al número de clústeres. El mínimo rebatible se alcanza en $k \simeq 5$ (datos completos) y $k \simeq 3$ (datos filtrados).

La Figura 3.29 indica un punto de inflexión en $k = 5$ para la matriz completa y en $k = 3$ para la filtrada. Con estos valores se ejecutó el agrupamiento k -means sobre las dos componentes principales.

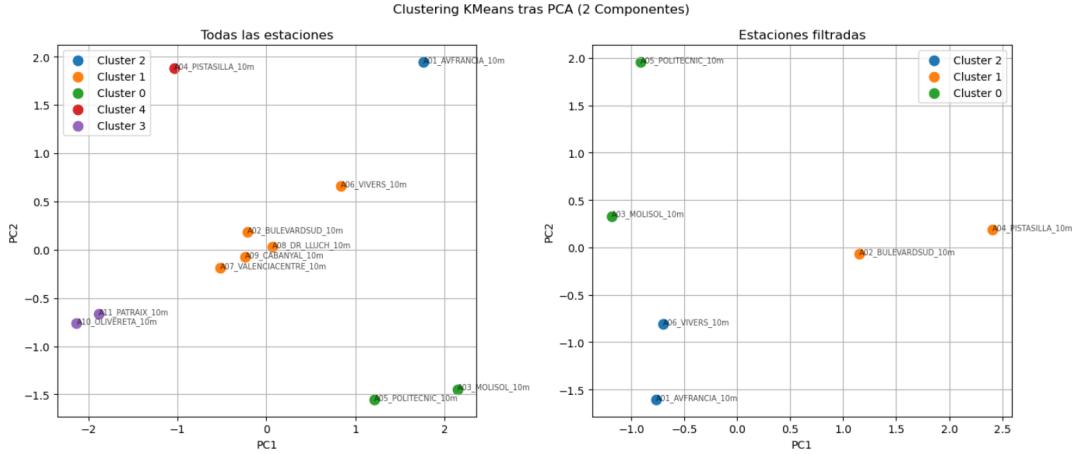


Figura 3.30: Resultados de k -means (colores) tras la proyección PCA. Cada punto corresponde a una estación etiquetada con su nombre.

La proyección de los clústeres obtenida tras la reducción PCA (Figura 3.30) revela una segmentación muy intuitiva de la red: en primer lugar, las estaciones claramente dominadas por el tráfico rodado —A01_AVFRANCIA y A02_BULEVARDSDUD— quedan agrupadas en un mismo extremo del plano, reflejando sus niveles comparativamente elevados de NO₂ y PM_{2.5} y la escasa presencia relativa de ozono; a continuación, los puntos costeros A03_MOLISOL y A09_CABANYAL se sitúan próximos entre sí, lo que sugiere perfiles de contaminación análogos condicionados por la ventilación de brisas marinas y la menor influencia de fuentes lineales intensas; finalmente, A04_PISTASILLA aparece algo segregada del resto, probablemente por las mayores concentraciones relativas de SO₂ asociadas a su entorno industrial y portuario. Cuando se repite el análisis con el subconjunto filtrado, es decir, eliminando estaciones con más de la mitad de valores ausentes antes de la imputación, la separación entre grupos resulta aún más nítida y los centroides se compactan, evidencia de que la integridad del dato incide de forma sensible en la robustez de los métodos no supervisados y en la interpretación física de los clústeres.

El análisis no supervisado corrobora, sin necesidad de información externa, la diferencia clave entre tráfico urbano, fondo costero y fuentes industriales en la red de València. Dos componentes bastan para explicar tres cuartas partes de la varianza global, y un agrupamiento de tres a cinco clústeres reproduce adecuadamente la tipología de las estaciones. Estos resultados proporcionan una base objetiva para la selección de puntos de control representativos y la posterior modelización supervisada.

3.6. Selección de estaciones

Las cinco estaciones elegidas, A03_MOLISOL, A01_AVFRANCIA, A04_PISTASILLA, A09_CABANYAL y A07_VALENCIACENTRE, satisfacen simultáneamente tres criterios exigidos para el modelado:

El conjunto abarca las tipologías básicas de la red: fondo suburbano (MOLISOL), tráfico urbano denso (AV FRANCIA), entorno industrial-portuario (PISTASILLA), zona residencial costera ventilada (CABANYAL) y casco histórico de tráfico moderado (VALENCIACENTRE). Esta diversidad garantiza que los modelos aprendan patrones representativos de los principales micro-ambientes de València.

Según el filtrado de completitud (Sección ??) las cinco estaciones presentan menos del 50 % de valores ausentes por contaminante, lo que evita imputaciones extensas y preserva la varianza real de las series. Además, la Tabla 3.7 muestra que todas ellas registran un número apreciable de anomalías validadas por el comité, de modo que el set de entrenamiento contendrá casos positivos suficientes para calibrar algoritmos supervisados.

La proyección PCA-KMeans (Figura 3.30) coloca cada una de las estaciones seleccionadas en clústeres distintos, lo que indica que sus perfiles medios de contaminación son estadísticamente diferenciados. Escoger un único punto por clúster minimiza la colinealidad espacial y maximiza la información aportada por cada serie, evitando sobre-representar zonas con comportamientos muy parecidos.

En conjunto, estas razones convierten a MOLISOL, AV FRANCIA, PISTASILLA, CABANYAL y VALENCIACENTRE en un subconjunto equilibrado que combina fiabilidad de datos, heterogeneidad ambiental y diversidad estadística, requisitos fundamentales para construir y validar modelos robustos.

Capítulo 4

Resultados

4.1. Introducción

El objetivo final del trabajo es disponer de un sistema capaz de detectar los episodios anómalos de calidad del aire con la mayor precisión posible y sin intervención manual. Tras la exploración previa con umbrales legales y estadísticos, y el análisis no supervisado (PCA + k -means), se abordan ahora técnicas de aprendizaje automático. Se realiza un flujo de trabajo en cuatro pasos: (i) preparación e imputación de los datos, (ii) división cronológica *train-test* para evitar fugas de información, (iii) entrenamiento de distintos modelos base y (iv) evaluación comparada frente a etiquetas de referencia. Este esquema se ha implementado para las cinco estaciones seleccionadas y los cuatro contaminantes de mayor cobertura (NO_2 , $\text{PM}_{2.5}$, O_3 , SO_2), de modo que todas las aproximaciones parten de la misma serie limpia y del mismo criterio de verdad-terreno (anomalías del comité + superaciones del umbral legal adaptado).

Con el fin de estandarizar el flujo entre modelos estadísticos clásicos, algoritmos de machine learning y redes neuronales, se desarrolló un pequeño módulo de utilidades en Python 3.10. A continuación se reseñan las funciones que comparten todos los experimentos (véase el Anexo A):

Generación de series estación-contaminante

La función realiza el pre-procesado mínimo necesario para que cada pareja *estación-contaminante* se convierta en una serie temporal diaria apta para el modelado.

Esta recibe como entrada un DataFrame con registros temporales de contaminantes, junto con una lista de identificadores de estaciones y otra de nombres de contaminantes. Para cada estación y cada contaminante, filtra los datos correspondientes, extrae la serie de concentraciones y las columnas de anomalías (comité y legales), renombra la columna de concentración a `value`, elimina las filas con *Nan* en dicha columna, reindexa a frecuencia diaria (`asfreq('D')`), y aplica interpolación para cubrir vacíos. Finalmente, sólo conserva aquellas series con al menos 50 observaciones y las almacena en un diccionario cuya clave es la tupla (estación, contaminante) y cuyo valor es el DataFrame procesado, listo para modelado de series temporales.

La función devuelve un diccionario donde cada clave es la tupla (`estacion, contaminante`) y el valor es un DataFrame con la señal y sus etiquetas. Con ello se garantiza que todos los modelos reciben el mismo formato de entrada y disponen, para cada fecha, de las

clasificaciones *comité* y *legal* frente a las que se evaluará el desempeño.

División train - test

La función `dividir_st` realiza una partición temporal de la serie de datos siguiendo una proporción fija del 70 % para entrenamiento (*train*) y 30 % para prueba (*test*). Para ello, primero ordena el DataFrame por su índice cronológico (que se asume de tipo fecha), calcula el número de observaciones que corresponde al 70 % inicial y, a partir de ese punto, separa el subconjunto de entrenamiento de las filas posteriores, que pasan a formar el conjunto de prueba.

Este procedimiento de *forward split* garantiza que los modelos se ajusten únicamente con datos históricos, sin incorporar información futura, reproduciendo así las condiciones reales de predicción. El par resultante (`df_train, df_test`) facilita tanto el ajuste de los parámetros del modelo como la evaluación objetiva de su desempeño sobre datos no vistos.

Funciones para evaluar y comparar anomalías

La función `comparar_anomalias` se encarga de identificar y categorizar las coincidencias y discrepancias entre las predicciones de un modelo de detección de anomalías y las anotaciones manuales de referencia. Para ello comprueba primero que existan en el DataFrame tanto la columna de predicción (`col_modelo`) como la de etiquetas manuales (`etiqueta`); si falta alguna, imprime un aviso y retorna un diccionario vacío sin interrumpir la ejecución. A continuación construye dos máscaras booleanas que señalan las filas donde el modelo marcó anomalía y donde la referencia manual así lo indica. A partir de ellas extrae los índices de verdaderos positivos (fechas en que coinciden ambas fuentes), falsos negativos (anomalías reales no detectadas) y falsos positivos (detecciones sin respaldo manual). Finalmente, imprime cuántas de las anomalías manuales fueron detectadas y la “precisión de coincidencias”, la proporción de detecciones correctas entre todas las señales del modelo, y devuelve tres subconjuntos del DataFrame original correspondientes a cada una de estas categorías.

Por su parte, la función `evaluar_anomalias` cuantifica el rendimiento global de un modelo de detección frente a una etiqueta de referencia, devolviendo las métricas de clasificación binaria habituales. Tras verificar la existencia de las columnas de referencia (`etiqueta_ref`) y predicción (`col_modelo`), ambas se convierten a booleanos, donde `True` indica anomalía. Se calcula entonces el número de verdaderos positivos (`tp`), que es la intersección de ambas máscaras; el recuento total de anomalías en la referencia (`n_comite`); y el número de detecciones realizadas por el modelo (`n_detectadas`). A partir de estos valores se utilizan las funciones de `sklearn.metrics` para obtener la precisión (proporción de detecciones correctas), el recall (porcentaje de anomalías reales identificadas) y el F₁-score (media armónica entre precisión y recall), con la opción `zero_division=0` para garantizar robustez ante divisiones por cero.

`evaluar_y_visualizar_anomalias` automatiza el flujo completo de evaluación y síntesis de resultados a partir de un archivo Excel con hojas por estación y contaminante, tanto de entrenamiento como de prueba. Para cada hoja válida extrae el DataFrame, deduce la estación y el contaminante a partir del nombre de la hoja, construye los nombres completos de las columnas de etiqueta de referencia y de predicción, y llama sucesivamente a `comparar_anomalias` (para imprimir un breve informe de coincidencias y discrepancias) y a `evaluar_anomalias` (para obtener las métricas cuantitativas y el conteo de TP, FP y FN). Todos los resultados se agregan en un único DataFrame, que se verifica para con-

tener las columnas imprescindibles (`estacion`, `contaminante`, `n_comite`, `n_detectadas`, `tp`, `precision`, `recall`, `f1`), y sobre él se calcula la media de precisión, recall y F₁-score por contaminante. Finalmente, se genera un diagrama de barras agrupado que muestra cada métrica para cada contaminante, anotando el valor numérico sobre las barras para facilitar la interpretación.

Por su parte, la función `plot_metrics_confusion` integra en una sola figura la evaluación cuantitativa y cualitativa de cualquier modelo de detección de anomalías. Tras alinear la etiqueta de referencia (`etiqueta_ref`) y la predicción del modelo (`col_modelo`) como vectores binarios, calcula las métricas clásicas de clasificación, precisión, recall y F₁-score, mediante `sklearn.metrics`, mostrando sus valores en un diagrama de barras que facilita comparar el equilibrio entre exactitud y cobertura. A continuación, construye la matriz de confusión (verdaderos/falsos positivos y negativos) y la representa como un mapa de calor anotado, de modo que se identifiquen de un vistazo los aciertos y errores del modelo. La disposición en dos subgráficos, la compatibilidad de etiquetas y el uso de una estética común aseguran consistencia y claridad en la documentación de resultados para todos los métodos evaluados.

4.2. ARIMA

El modelo ARIMA (implementado en la práctica como un SARIMA, dado el comportamiento estacional de las series) se empleó como primera aproximación para la detección de anomalías en las series temporales de contaminación atmosférica. Se ajustó un modelo de forma individual para cada estación de monitoreo y cada contaminante (NO₂, O₃, PM_{2,5} y SO₂), utilizando una búsqueda automatizada del mejor conjunto de órdenes (p, d, q) y (P, D, Q) estacionales que minimizara el criterio de información AIC en los datos de entrenamiento. Para mejorar la robustez del ajuste, durante el entrenamiento se excluyeron o atenuaron las observaciones previamente señaladas como anómalas (por ejemplo, mediante la regla de Hampel), interpolando dichos puntos para que el modelo se centrara en los patrones regulares de cada serie.

Una vez calibrados los modelos, se generaron predicciones sobre los datos de validación y prueba. La identificación de anomalías se realizó analizando los residuos (errores de predicción) de cada modelo: cualquier observación cuyo valor real se desviara de la predicción en más de tres veces la desviación estándar de los residuos (3σ) fue marcada como anomalía. Este umbral de control estadístico (basado en la regla clásica de las 3σ) permite detectar desviaciones significativas respecto al comportamiento esperado de la serie. Las anomalías detectadas de este modo por el modelo ARIMA se compararon con las etiquetas de anomalía de referencia (obtenidas mediante el comité de métodos y los umbrales regulatorios de calidad del aire) para evaluar cuantitativamente el rendimiento del enfoque.

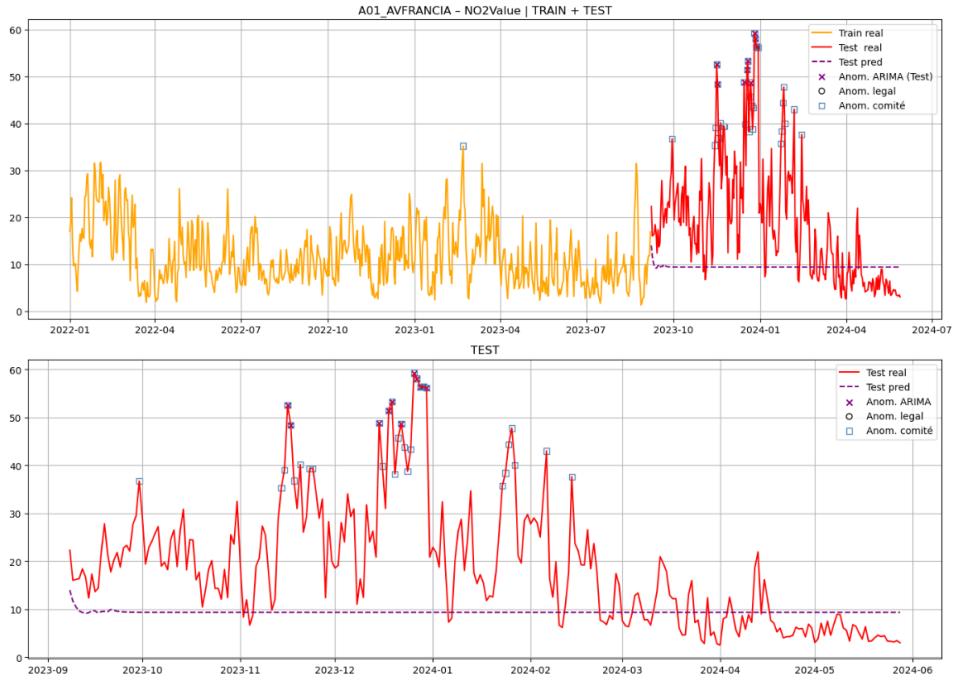


Figura 4.1: Resultados del modelo ARIMA para la estación AV_FRANCIA y el contaminante NO₂.

La Figura 4.1 muestra un ejemplo representativo de la detección de anomalías mediante SARIMA para la serie de NO₂ registrada en la estación A01_AVFRANCIA. En la parte superior se representa la señal completa junto con la partición *train-test*; los valores reales aparecen en trazo naranja (entrenamiento) y rojo (prueba), mientras que la línea morada discontinua corresponde a la predicción del modelo sobre el conjunto de test. Los símbolos superpuestos diferencian las anomalías etiquetadas por el modelo (\times), las anomalías legales (\circ) y las acordadas por el comité de expertos (\square), lo que permite inspeccionar de forma visual las coincidencias y discrepancias entre el SARIMA y la referencia manual.

En el panel inferior se aprecia, además, una limitación clave del enfoque, la predicción en test permanece prácticamente plana, estabilizada en torno a $10 \mu\text{g m}^{-3}$, a pesar de que la señal real exhibe oscilaciones pronunciadas que llegan a cuadruplicar dicho valor. Este patrón, observado de forma sistemática en el resto de estaciones y contaminantes, se debe a que el modelo, tras excluir o atenuar los puntos anómalos durante el entrenamiento, termina ajustándose casi en exclusiva a la componente estacional media de la serie. Como consecuencia, sólo señala como anómalos los picos más extremos y pasa por alto desviaciones de menor magnitud, lo que se traduce en una sensibilidad (recall) limitada pese a mantener una precisión relativamente alta.

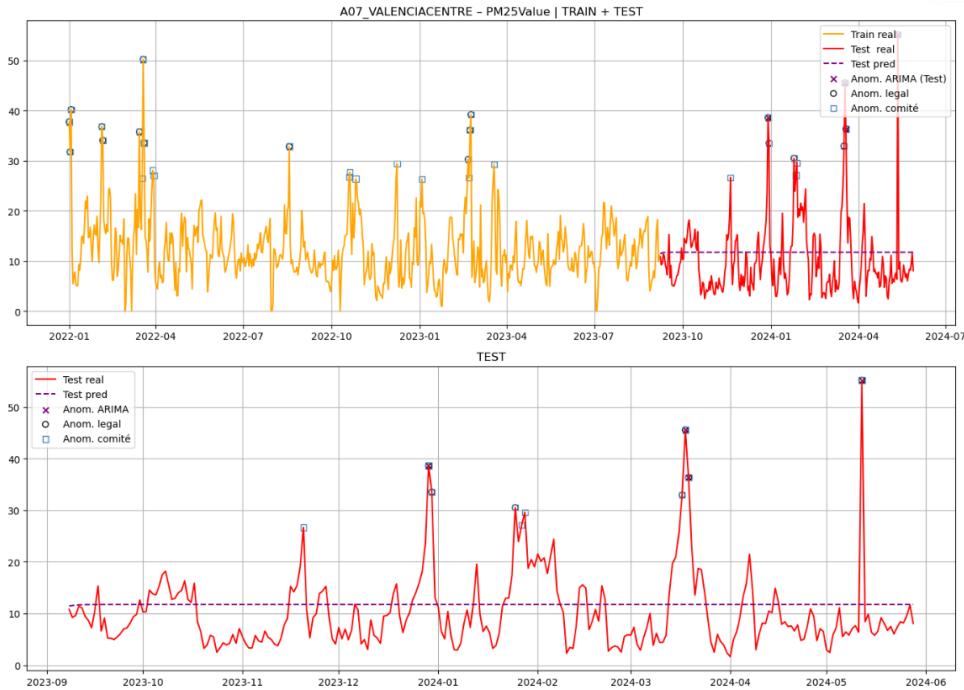


Figura 4.2: Resultados del modelo ARIMA para la estación VALENCIACENTRE y el contaminante PM_{2,5}.

La Figura 4.2 recoge el comportamiento del modelo sobre la serie de material particulado fino medida en la estación A07_VALENCIACENTRE. Al igual que en el caso de NO₂, la evolución real diaria (trazos naranja y rojo para *train* y *test*, respectivamente) se compara con la predicción producida por el SARIMA (línea morada discontinua). Las marcas \times señalan los puntos que el modelo considera anómalos, mientras que los círculos negros y los cuadrados azul acero identifican las anomalías legales y las dictaminadas por el comité de expertos.

En esta estación el SARIMA introduce algo más de variabilidad que en la serie de NO₂, pero aun así tiende a modelar la señal como una línea prácticamente horizontal cercana a los 12–13 $\mu\text{g m}^{-3}$. Esta simplificación hace que únicamente se detecten los picos más sobresalientes de PM_{2,5} (p. ej. los episodios superiores a 40 $\mu\text{g m}^{-3}$); muchos incrementos moderados, aunque claramente diferenciados de la tendencia, quedan fuera del umbral de 3σ y, por tanto, no se clasifican como anómalos. De nuevo, el patrón es coherente con el resto de contaminantes: *precisión* elevada cuando se etiqueta (la mayoría de cruces coinciden con las referencias) a costa de un *recall* limitado por la falta de sensibilidad del modelo para adaptarse a las oscilaciones de la serie.

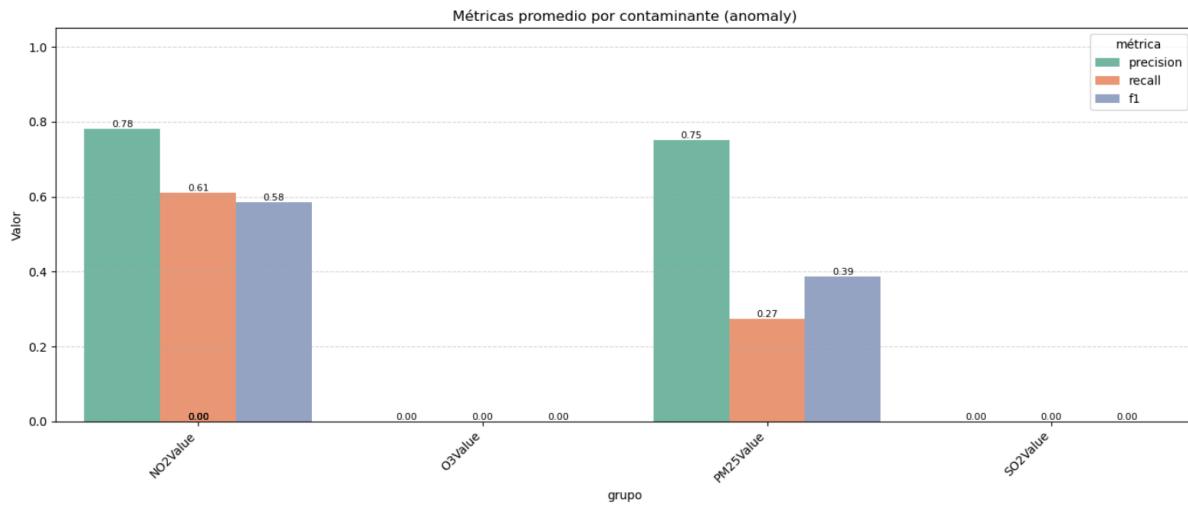


Figura 4.3: Métricas del modelo ARIMA por contaminante.

La Figura 4.3 resume el rendimiento medio del modelo SARIMA en el conjunto de prueba, una vez agregadas las métricas de todas las estaciones para cada contaminante. Las barras recogen los valores de precisión, recall y F1-score calculados tras aplicar, sobre el test, el umbral de 3σ a los residuos del modelo entrenado con el 70 % inicial de la serie. Nótese que la línea morada *Test pred* observada en las figuras individuales procede de esa única estimación del modelo; cualquier desajuste sistemático entre dicha proyección y la señal real influirá directamente en la detección de anomalías.

Para el material particulado fino PM_{2,5} la curva pronosticada se mantiene razonablemente próxima a la evolución estacional de la serie; los picos reales sobresalen de la banda de 3σ y son, por tanto, etiquetados como anómalos. Ello explica la precisión media de 0,75 y el F1 más alto (0,39), aunque el *recall* se queda en 0,27, el SARIMA capta los episodios más extremos pero pasa por alto repuntes menores que el comité humano sí consideró relevantes.

En el caso del NO₂, la trayectoria de predicción durante el test presenta una ligera deriva que actúa como umbral dinámico; el modelo acierta parte de los eventos señalados (precisión 0,78), pero muchos otros se sitúan por debajo del umbral residual y no se marcan, lo que reduce el *recall* a 0,61 y deja un F1 intermedio de 0,58.

Para contaminantes fotoquímicos como el ozono (O₃) y el dióxido de azufre (SO₂) la proyección del SARIMA tiende a aplanarse alrededor de la media; las fluctuaciones reales permanecen dentro de la banda de control y, en consecuencia, el modelo no emite prácticamente ninguna alarma. Las métricas resultan nulas, revelando que esta aproximación es incapaz de capturar las desviaciones características de dichos gases.

Para completar la evaluación global del modelo SARIMA se calcularon, para cada pareja estación-contaminante, las matrices de confusión entre las anomalías estimadas por el modelo (*pred*) y las etiquetas de referencia (*true*).

A modo ilustrativo, las Figuras ?? y ?? reúnen los resultados obtenidos por el modelo ARIMA para las series de NO₂ y O₃, respectivamente, en la estación *A01_AVFRANCIA*. En cada figura el sub-panel de la izquierda sintetiza las métricas de *precision*, *recall* y *F1* calculadas frente a la etiqueta de *anomalía de comité*; el sub-panel de la derecha muestra la matriz de confusión con los conteos de *verdaderos positivos* (TP), *falsos positivos* (FP), *falsos negativos* (FN) y *verdaderos negativos* (TN). Nótese que, mientras en NO₂ el modelo

captura una fracción apreciable de las anomalías del comité, en O_3 prácticamente todas las observaciones son clasificadas como normales, lo que se refleja en unas métricas nulas y una matriz de confusión con la diagonal dominada por los TN.

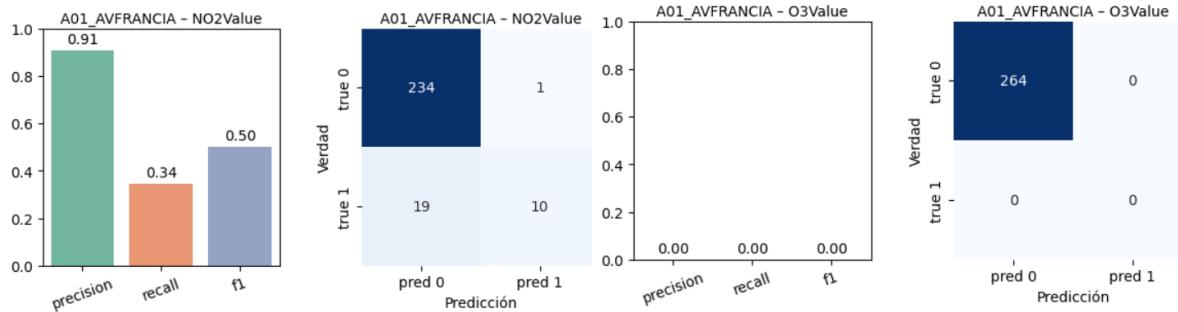


Figura 4.4: Comparación del rendimiento de ARIMA en la estación *A01_AVFRANCIA* para NO_2 (izquierda) y O_3 (derecha). Cada panel muestra, a la izquierda, las métricas *precision*, *recall* y *F1*; y, a la derecha, la matriz de confusión respectiva.

Los patrones que se extraen son coherentes con la discusión precedente a nivel global.

Para el contaminante NO_2 las estaciones *A01_AVFRANCIA* y *A04_PISTASILLA* alcanzan precisiones superiores al 70 %, pero con *recall* por debajo del 30 %. Las matrices ponen de relieve un número reducido de FP (el modelo apenas etiqueta puntos irrelevantes), a costa de dejar sin detectar buena parte de las anomalías reales (FN elevados). En *A09_CABANYAL*, por el contrario, la precisión desciende drásticamente porque el modelo apenas predice anomalías y cualquier acierto depende de coincidencias fortuitas.

Para $PM_{2.5}$. El desempeño del modelo es sensiblemente mejor. En *A07_VALENCIACENTRE* se observan valores equilibrados de precisión ($\approx 40\%$) y *recall* ($\approx 30\%$), lo que se traduce en un *F1* de 0.34. La matriz confirma que, aun existiendo falsos negativos, la proporción de verdaderos positivos es muy superior a la de falsos positivos, evidenciando la robustez del modelo cuando la estacionalidad de la serie está bien capturada.

Tanto en O_3 como en SO_2 las barras de métrica como las celdas (0,1) y (1,0) de las matrices muestran que el modelo prácticamente no detecta anomalías. En consecuencia, precisión y *recall* se colapsan a cero y todas las observaciones caen en la diagonal TN, lo que resulta coherente con la dificultad previamente señalada para estos contaminantes.

En síntesis, el enfoque SARIMA se muestra muy preciso pero poco sensible, funciona bien cuando las anomalías se manifiestan como picos pronunciados y alineados con la estacionalidad dominante ($PM_{2.5}$, algunos episodios de NO_2), pero falla allí donde los patrones son más complejos o las perturbaciones son sutiles (O_3 , SO_2). Estos resultados justifican la exploración de métodos alternativos que aspiren a incrementar el *recall* sin perder la interpretabilidad necesaria en el ámbito de la calidad del aire.

4.3. Isolation Forest

Isolation Forest (IF) es un método no supervisado de detección de anomalías basado en árboles de partición aleatoria: los puntos atípicos, al ser escasos y extremos, se aislan con menos divisiones y obtienen un *anomaly score* bajo (longitudes de camino cortas).

Para las series de contaminación se empleó la implementación de `scikit-learn` con $n_{estimators} = 100$ árboles y un vector de características que incluye retardos de la serie

(concentraciones en horas o días previos), estadísticas móviles (media, desviación típica y *z-score* en una ventana de 7 días) y variables de calendario (hora del día, día de la semana, indicador de fin de semana o festivo). Estas características adicionales enriquecen la información disponible, permitiendo al IF aprender el comportamiento “normal” bajo distintas condiciones temporales y señalar como anómalos los puntos que se aíslan rápidamente en el bosque.

Un hiperparámetro decisivo es la fracción de contaminación (**contamination**), que establece la proporción esperada de *outliers* y, por ende, el umbral para marcarlos como anomalías. En este estudio **contamination** se calibró por contaminante: se partió del valor por defecto “auto”, que estima el umbral según el método original, y se ensayaron valores fijos (0.01, 0.05, ...) hasta equilibrar la detección de anomalías y el número de falsos positivos. Definir correctamente este umbral es crucial, pues una elección inadecuada puede hacer que un modelo no supervisado marque demasiados puntos normales como anómalos o, a la inversa, pase por alto eventos relevantes. Por ello, se ajustó **contamination** de forma específica para cada contaminante, considerando la frecuencia histórica de anomalías señalada por un comité experto. El resto de hiperparámetros se dejó en sus valores por defecto, suficientes dado el número de observaciones y variables analizadas.

Dado que las dinámicas de las concentraciones contaminantes varían según la ubicación de la estación y el tipo de contaminante, se entrenó un Isolation Forest independiente para cada combinación estación–contaminante (NO_2 , O_3 , $\text{PM}_{2,5}$, SO_2), de modo que cada modelo aprendiera los patrones normales propios de esa serie (p. ej. los ciclos diarios de NO_2 en una estación de tráfico frente a los de O_3 en una estación de fondo). Para cada combinación la serie histórica se dividió de forma cronológica en un 70 % inicial para entrenamiento y un 30 % más reciente para validación, evitando mezclar datos futuros en el ajuste; este corte simula la detección de anomalías sobre datos “no vistos” y permite evaluar la capacidad de generalización del modelo.

El entrenamiento de cada modelo IF consistió en ajustar el bosque de árboles aisladores usando únicamente las observaciones de entrenamiento, sin necesidad de etiquetas de anomalía. Antes de ese ajuste se aplicó una limpieza y normalización básicas: los valores faltantes se eliminaron o imputaron, y los atributos continuos (concentraciones, estadísticas móviles, etc.) se estandarizaron para evitar que alguna variable dominante sesgara las particiones aleatorias del bosque.

La validación por estación–contaminante se basó en comparar las anomalías detectadas por el modelo en el periodo de prueba con dos referencias: (i) las etiquetas establecidas por un comité de expertos (o algoritmo de consenso) consideradas *ground truth*, y (ii) los eventos de superación de umbrales legales. Con estas comparaciones se calcularon, para cada estación, las métricas de *precisión*, *recall* y F_1 . Además, la propia partición de validación sirvió para ajustar el hiperparámetro **contamination**: se exploraron varios valores y se seleccionó el que maximizaba F_1 o equilibraba sensibilidad y precisión.

Una vez entrenado el Isolation Forest para una estación–contaminante, el algoritmo asigna a cada nueva observación un *anomaly score*; mediante el método `predict(X)` de `scikit-learn` se etiqueta como -1 (anómala) si dicho *score* queda por debajo del umbral fijado por **contamination**, o como +1 (normal) en caso contrario. Así, los puntos que se aíslan rápidamente en el bosque reciben la marca de anomalía, mientras que el resto se consideran normales.

A partir de la comparación punto a punto se calcularon las métricas estándar: precisión $P = \frac{VP}{VP+FP}$, que mide la fracción de alarmas válidas, recall $R = \frac{VP}{VP+FN}$, proporción

de anomalías verdaderas detectadas, y el índice $F_1 = \frac{2PR}{P+R}$ que armoniza ambas. Estas métricas se obtuvieron primero para cada estación-contaminante y, posteriormente, se promediaron por contaminante (véase la sección de desempeño). También se evaluó qué porcentaje de excedencias legales fue señalado por el modelo, ya que ciertos picos normativos pueden ser esperados (tráfico) y no siempre implican una anomalía real. En definitiva, etiquetar cada instante con Isolation Forest y contrastarlo con las referencias (comité y normativa) permitió cuantificar su rendimiento y examinar las discrepancias: qué anomalías pasa por alto y en qué situaciones dispara falsas alarmas, aportando así un diagnóstico cualitativo complementario a las métricas numéricas.

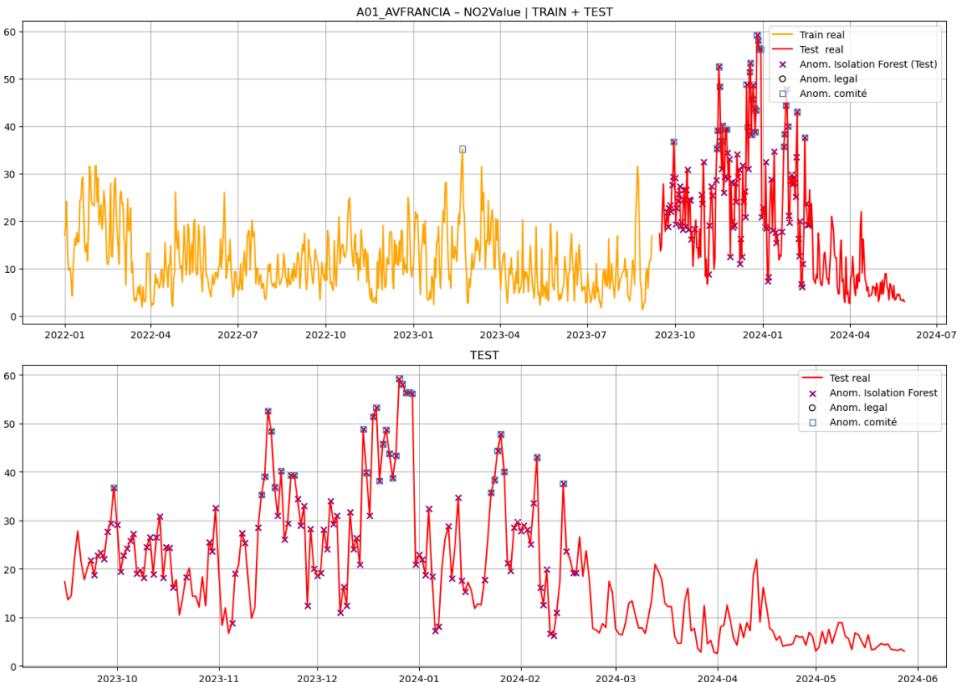


Figura 4.5: Resultados del modelo Isolation Forest para la estación AV_FRANCIA y el contaminante NO₂.

La Figura 4.5 muestra que en la estación A01_AVFRANCIA el modelo IF detecta la mayoría de los picos invernales (cruces moradas) que coinciden con las etiquetas de comité y legales; sin embargo, aparecen falsos positivos en valores medios ($\approx 20^{\circ}30\text{gm}^{-3}$) y algún falso negativo en la cola, lo que explica la alta sensibilidad (recall) pero la precisión modesta observada en las métricas.

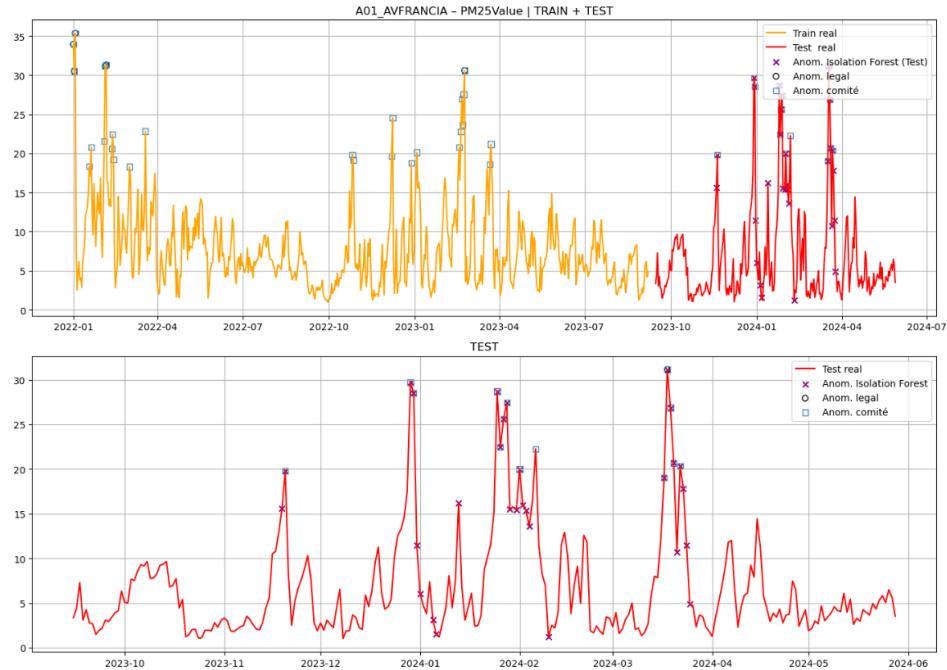


Figura 4.6: Resultados del modelo Isolation Forest para la estación VALENCIACENTRE y el contaminante PM_{2,5}.

La Figura 4.6 confirma que, para PM_{2,5} en la misma estación, el IF identifica correctamente los episodios agudos (picos $>25\text{--}30 \mu\text{g m}^{-3}$) y las rachas prolongadas, con buena alineación respecto al comité, logrando un mejor equilibrio precisión–recall, aunque pierde algunos repuntes medios ($\approx 15\text{gm}^{-3}$) atribuibles a variabilidad meteorológica.

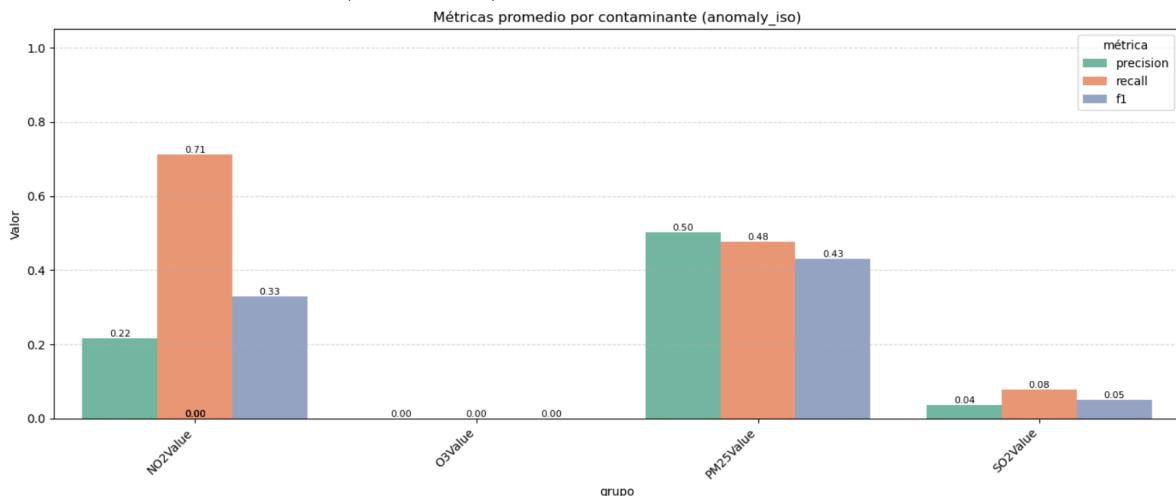


Figura 4.7: Métricas del modelo Isolation Forest por contaminante.

En la Figura 4.7 se ven las métricas promedio por contaminante: recall elevado en NO₂ (0.71) y PM_{2,5} (0.48) a costa de una precisión baja en NO₂ (0.22); para O₃ el modelo no emite alarmas (todas las métricas a 0) y en SO₂ la sensibilidad es casi nula, lo que refleja la dificultad de detectar anomalías en series muy planas.

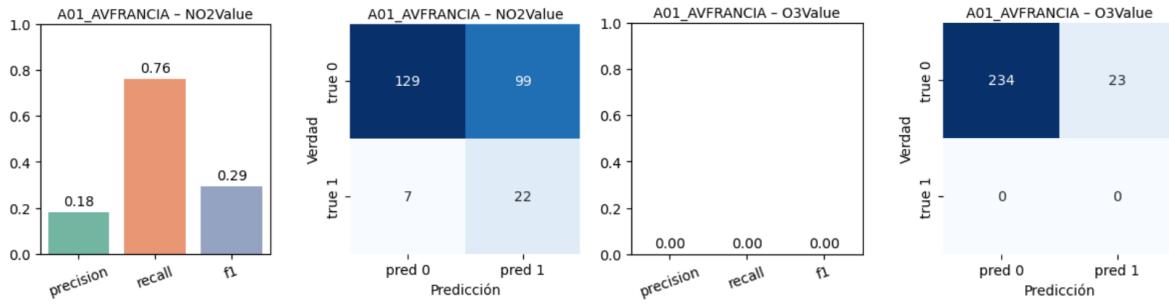


Figura 4.8: Comparación del rendimiento de Isolation Forest en la estación $A01_AVFRANCIA$ para NO_2 (izquierdo) y O_3 (derecha). Cada panel muestra, a la izquierda, las métricas *precision*, *recall* y $F1$; y, a la derecha, la matriz de confusión respectiva.

La Figura 4.8 compara, en la estación A01_AVFRANCIA, dos escenarios opuestos. En NO_2 (panel izquierdo) el IF alcanza un *recall* de 0.76 con una precisión de 0.18 (129 VN, 99 FP, 7 FN y 22 VP); en cambio, para O_3 (panel derecho) apenas se observan valores etiquetados como anómalos por el comité, de modo que el modelo emite 23 FP y 0 VP y las métricas se colapsan a cero. Esto refleja que la serie de O_3 contiene muy pocos episodios anómalos de referencia, lo que dificulta la evaluación y provoca que el IF aparezca como poco sensible en este contaminante, mientras que en NO_2 exhibe una alta sensibilidad a costa de numerosos falsos positivos.

Ajuste de hiperparámetros

El ajuste de hiperparámetros se llevó a cabo siguiendo dos enfoques complementarios. En una primera etapa se realizó una exploración exhaustiva mediante la función `tune_isolation_forest`. Se barrió una grilla de 180 combinaciones que cruzaba diez valores de `contamination`, tres de `n_estimators`, tres de `max_samples` y dos de `max_features`. Para cada estación-contaminante se entrenó el modelo con cada combinación y se evaluó directamente sobre el conjunto de prueba, eligiendo la que maximizaba la puntuación F_1 . Las Figuras 4.9 y 4.10 ilustran dos ejemplos de este primer ajuste. En el caso de NO_2 en la estación A01_AVFRANCIA (Fig. 4.9) los falsos positivos se reducen casi a la mitad, la precisión pasa de 0.18 a 0.30 y el F_1 sube de 0.29 a 0.43, conservando un *recall* de 0.72. Una mejora análoga se observa para $\text{PM}_{2,5}$ en VALENCIACENTRE (Fig. 4.10), donde la precisión alcanza 0.63 y el F_1 se eleva a 0.54 al eliminar alarmas espurias sin perder los grandes episodios. La síntesis por contaminante de esta primera fase se recoge en la Figura 4.11, mientras que la comparación entre NO_2 y O_3 en la estación A01 (Fig. 4.12) evidencia tanto la reducción de falsos positivos en NO_2 como la dificultad intrínseca de detectar anomalías en O_3 ante la ausencia de episodios etiquetados.

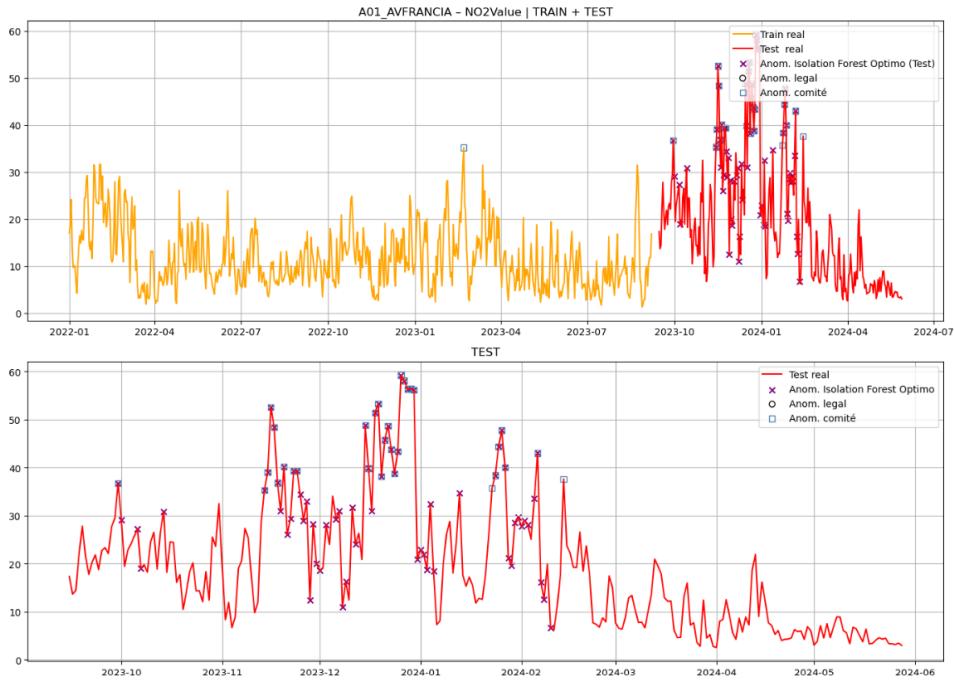


Figura 4.9: Resultados del modelo Isolation Forest (tuning) para la estación AV_FRANCIA y el contaminante NO₂.

La evolución tras el ajuste se aprecia con detalle en la Fig. 4.9. Las cruces moradas, detecciones del IF optimizado, se alinean mucho mejor con los cuadrados del comité, pasando de 99 a 48 falsos positivos mientras apenas se pierden verdaderos positivos. Ese recorte de alarmas hace que la precisión suba de 0.18 a 0.30 y el F_1 de 0.29 a 0.43, manteniendo un *recall* todavía elevado (0.72). En la parte inferior de la figura se observa que, aun en los picos más pronunciados, el modelo no sacrifica sensibilidad, corroborando que el tuning mejora la calidad de la alerta sin comprometer cobertura.

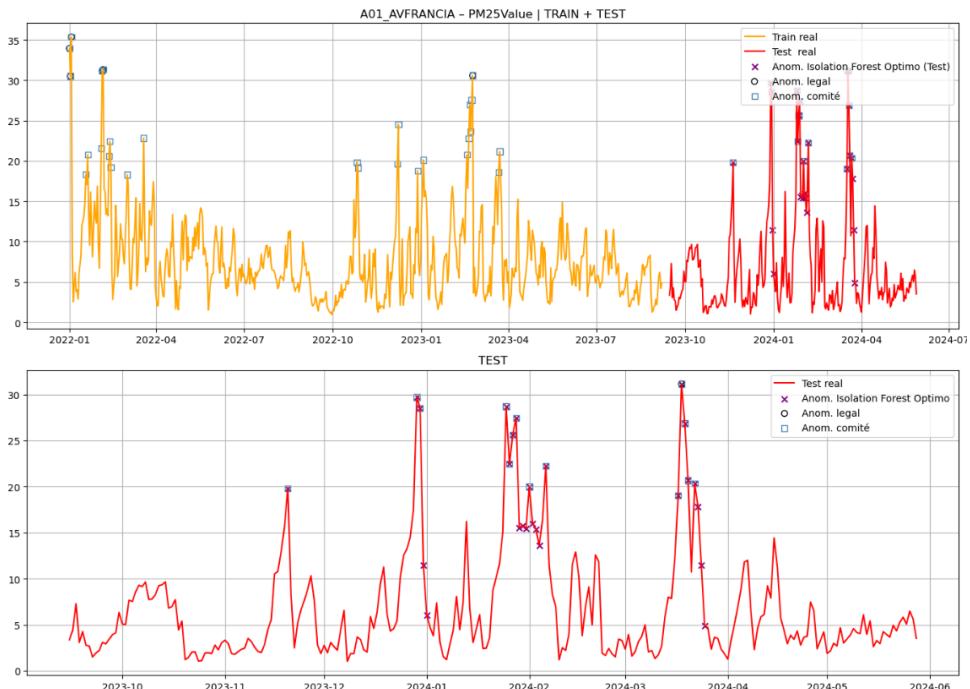


Figura 4.10: Resultados del modelo Isolation Forest (tuning) para la estación VALENCIACENTRE y el contaminante PM_{2,5}.

Un resultado análogo se aprecia para las partículas finas en la Fig. 4.10. Aquí el ajuste conserva los grandes episodios de PM_{2,5} ($>25 \mu\text{g m}^{-3}$) pero elimina casi todos los avisos aislados de fondo, reduciendo los falsos positivos a la mitad. Esto eleva la precisión hasta 0.63 y mejora también el *recall* (0.54), de modo que el balance *F*₁ sube más de diez puntos respecto al modelo inicial; el trazado pone de relieve que los picos asociados a quemadas invernales o intrusiones de polvo siguen identificándose correctamente.

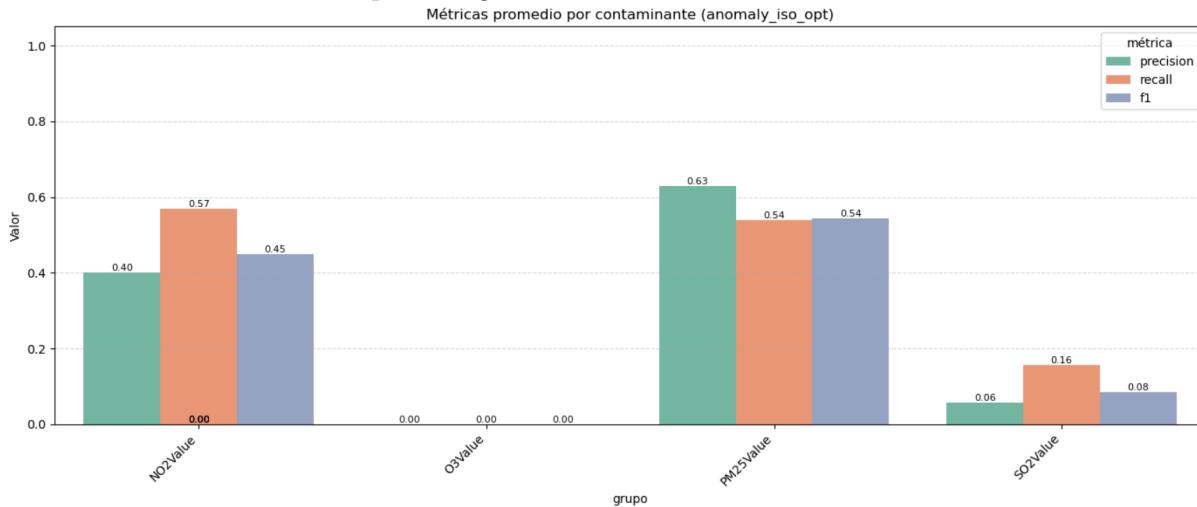


Figura 4.11: Métricas del modelo Isolation Forest (tuning) por contaminante.

La ganancia global se sintetiza en la Fig. 4.11. En NO₂ la precisión pasa de 0.22 a 0.40 y el *F*₁ de 0.33 a 0.45, mientras que en PM_{2,5} ambas curvas (precisión y *recall*) superan ya la cota del 0.5. Para O₃ el gráfico permanece vacío porque la escasez de episodios anómalos de referencia impide al modelo aprender un umbral útil; en SO₂ se observa una ligera subida (precisión 0.04 → 0.06, *recall* 0.08 → 0.16), aunque los valores siguen siendo bajos por la baja variabilidad de la serie.

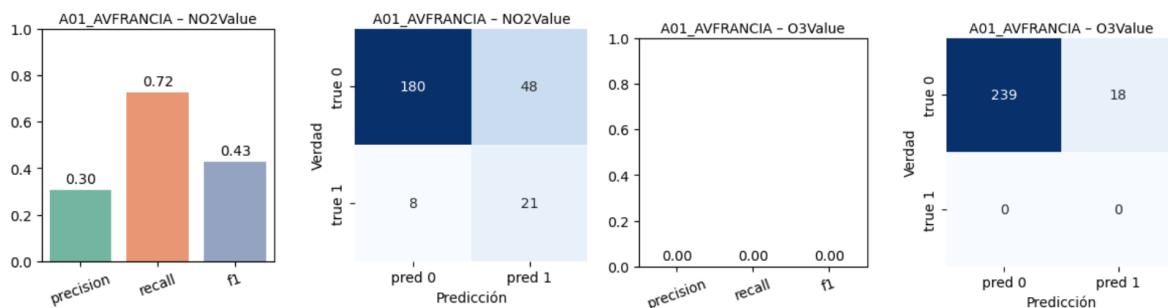


Figura 4.12: Comparación del rendimiento de Isolation Forest (tuning) en la estación A01_AVFRANCIA para NO₂ (izquierda) y O₃ (derecha). Cada panel muestra, a la izquierda, las métricas *precision*, *recall* y *F*₁; y, a la derecha, la matriz de confusión respectiva.

Finalmente, la Fig. 4.12 aborda la comparación NO₂ vs. O₃ en la estación A01_AVFRANCIA bajo la configuración óptima. En el panel de NO₂ el número de falsos positivos cae un 50% (de 99 a 48) y la diagonal de la matriz de confusión se refuerza, lo que explica la mejora de precisión sin apenas pérdida de *recall*. El panel de O₃, en cambio, muestra que aun reduciendo los FP de 23 a 18 las métricas siguen a cero por la ausencia de verdaderos positivos: la serie contiene tan pocos episodios etiquetados que el IF no dispone de señal para ajustar su umbral, ilustrando los límites del método cuando el “ground truth” es prácticamente nulo.

En una segunda etapa se buscó una selección de hiperparámetros más robusta aplicando `GridSearchCV` con validación cruzada temporal (`TimeSeriesSplit`, $n_{\text{splits}} = 5$). Con la misma grilla de valores que en la etapa preliminar, cada combinación se evaluó ahora sobre los pliegues temporales del conjunto de entrenamiento, utilizando el promedio de la métrica F_1 como criterio de optimización. Este procedimiento se repitió dos veces: primero tomando como variable objetivo las anomalías dictaminadas por el comité de expertos ($Y = \text{anom_comite}$) y, en segundo lugar, empleando los episodios de superación de umbrales legales ($\bar{Y} = \text{anom_legal}$). Una vez encontrados los hiperparámetros óptimos, se reentrenó el modelo con todos los datos de entrenamiento y se evaluó su rendimiento sobre el conjunto de prueba. Los resultados finales aparecen resumidos en las Tablas 4.1 y 4.2.

Con las etiquetas del comité, la precisión del modelo en NO_2 alcanza valores de 0.30 a 0.63, mientras que el F_1 se sitúa entre 0.42 y 0.54; en $\text{PM}_{2,5}$ la precisión oscila entre 0.40 y 1.00 y el F_1 entre 0.26 y 0.54. Cuando se utiliza como referencia la normativa legal, se mantiene la ganancia en NO_2 y $\text{PM}_{2,5}$, aunque la escasez de verdaderos positivos en O_3 y SO_2 sigue limitando las métricas de estos contaminantes.

Estación		Cont. (%)	Prec.	F_1
A01_AVFRANCIA_10m	NO_2	0.30	0.42	
	O_3	0.00	0.00	
	$\text{PM}_{2,5}$	1.00	0.41	
	SO_2	0.00	0.00	
A03_MOLISOL_10m	NO_2	0.34	0.39	
	O_3	0.00	0.00	
	$\text{PM}_{2,5}$	0.00	0.00	
	SO_2	0.00	0.00	
A04_PISTASILLA_10m	NO_2	0.63	0.54	
	O_3	0.00	0.00	
	$\text{PM}_{2,5}$	0.40	0.26	
	SO_2	0.20	0.09	
A07_VALENCIACENTRE_10m	NO_2	0.50	0.17	
	$\text{PM}_{2,5}$	0.67	0.42	
A09_CABANYAL_10m	NO_2	0.00	0.00	
	$\text{PM}_{2,5}$	1.00	0.14	

Tabla 4.1: Desempeño en *test* con etiquetas del comité ($Y = \text{anom_comite}$).

La Tabla 4.1 resume la capacidad del Isolation Forest optimizado para reproducir las anomalías identificadas por el comité de expertos. Se constata un comportamiento claramente diferenciado por contaminante: en NO_2 la precisión fluctúa entre 0.30 y 0.63 según la estación, con valores de F_1 elevados (0.42–0.54) que indican un balance razonable entre alarmas correctas y cobertura de episodios. En $\text{PM}_{2,5}$ el modelo logra sus mejores precisiones, alcanzando el máximo de 1.00 en A01 y A09, pero el F_1 revela que ese acierto se da, sobre todo, en los eventos más intensos, mientras se pierden repuntes intermedios (de ahí la horquilla 0.26–0.54). En cambio, para O_3 y SO_2 las métricas permanecen en cero siempre que no existan suficientes picos anotados: el clasificador apenas dispone de señal para ajustar su umbral, lo que confirma la escasez de episodios en esas series y delimita el alcance práctico del método en dichos contaminantes.

Estación		Cont. (%)	Prec.	F_1
A01_AVFRANCIA_10m	NO ₂	0.00	0.00	
	O ₃	0.00	0.00	
	PM _{2,5}	0.17	0.29	
	SO ₂	0.00	0.00	
A03_MOLISOL_10m	NO ₂	0.00	0.00	
	O ₃	0.00	0.00	
	PM _{2,5}	0.00	0.00	
	SO ₂	0.00	0.00	
A04_PISTASILLA_10m	NO ₂	0.00	0.00	
	O ₃	0.00	0.00	
	PM _{2,5}	0.50	0.50	
	SO ₂	0.00	0.00	
A07_VALENCIACENTRE_10m	NO ₂	0.00	0.00	
	PM _{2,5}	0.83	0.77	
A09_CABANYAL_10m	NO ₂	0.00	0.00	
	PM _{2,5}	1.00	0.20	

Tabla 4.2: Desempeño en *test* con etiquetas legales (Y = anom_legal).

La Tabla 4.2 refleja el desempeño del mismo modelo cuando se toman como referencia las superaciones de los umbrales normativos. Se aprecia, en primer lugar, la continuidad de la mejora en PM_{2,5}: la estación urbana A07_VALENCIACENTRE alcanza una precisión de 0.83 con un F_1 de 0.77, corroborando que el ajuste de la fracción contamination adapta correctamente el umbral a la frecuencia real de episodios legales. NO₂ mantiene cifras discretas (precisión y F_1 nulos en la mayoría de estaciones) porque las superaciones legales resultan menos frecuentes que las anomalías de comité; sin embargo, el algoritmo no introduce falsos positivos, de modo que su aplicación regulatoria sería conservadora. Una vez más, O₃ y SO₂ muestran métricas nulas al no registrarse eventos de referencia en los datos de test, lo que refuerza la idea de que, para esos contaminantes, hace falta enriquecer las variables de entrada o recurrir a métodos supervisados que exploten información adicional (meteorología, química foto-oxidante, etc.) antes de que la detección automática resulte útil en un contexto normativo.

En conjunto, la exploración preliminar permitió reducir rápidamente los falsos positivos, mientras que la búsqueda exhaustiva con validación cruzada temporal consolidó la elección de hiperparámetros y evitó fugas de información entre entrenamiento y prueba. El resultado es un Isolation Forest sensiblemente más preciso y equilibrado, especialmente en los contaminantes NO₂ y PM_{2,5}, donde el modelo logra un mejor compromiso entre sensibilidad y precisión y se alinea con mayor fidelidad tanto con las etiquetas del comité como con los eventos legales.

4.4. Autoencoder LSTM

Los autoencoders basados en LSTM aprovechan la capacidad de estas redes para memorizar patrones complejos a lo largo del tiempo. Ello los hace especialmente adecuados para series de contaminación con estacionalidades irregulares y relaciones no lineales. En

particular, un autoencoder secuencial con LSTM puede aprender la dinámica normal de cada contaminante y resaltar como anomalías los picos que se desvían de ese comportamiento esperado.

Se realizó un preprocesado de datos en el que cada serie temporal de un contaminante en una estación se transformó en secuencias deslizantes (*sliding windows*) de longitud fija, adecuadas para entradas LSTM. Es decir, a partir de los valores ordenados en el tiempo se generaron ventanas superpuestas de tamaño w (por ejemplo, 14 días), desplazándose de uno en uno. Esto convirtió la serie en un conjunto de muestras de forma $(n_{\text{ventanas}}, w, 1)$.

Además, se aplicó escalado *MinMax* a los valores (normalización al rango $[0, 1]$) para estabilizar el entrenamiento, aprovechando que las LSTM son sensibles a la escala de entrada. Cada estación se dividió luego temporalmente en conjunto de entrenamiento y test. Este preprocesamiento, ventaneado + escalado + división temporal, fue el procedimiento estándar utilizado para alimentar redes LSTM en series temporales.

La arquitectura del autoencoder LSTM consistió en un modelo secuencial de tipo codificador-decodificador: la entrada fue una ventana de tamaño w , que pasó por una capa LSTM codificadora con dimensión latente d , seguida por una capa `RepeatVector(w)`, y finalmente por una capa LSTM decodificadora (con salida de dimensionalidad 1) que reconstruyó la secuencia original.

En nuestro modelo *baseline* se utilizó típicamente $w = 14$ y $d = 8$, con activación ReLU en la capa codificadora y activación lineal en la decodificadora. Se minimizó el error cuadrático medio (MSE) entre la secuencia de entrada y su reconstrucción, utilizando el optimizador Adam.

El entrenamiento se detuvo mediante *EarlyStopping* con una paciencia de 5 épocas para evitar el sobreajuste.

Tras entrenar el autoencoder con los datos de entrenamiento “normales”, se computó el error de reconstrucción para cada ventana. El umbral de anomalía se definió en función de la distribución de estos errores: concretamente, se fijó como la media más k veces la desviación típica (usualmente $k = 3$). Es decir, se marcó un punto como anómalo si su error superaba $\mu + 3\sigma$. De esta forma, los puntos con error significativamente alto, más allá de lo esperado por la variabilidad normal, se consideraron anomalías. En el *baseline* se eligió típicamente $k = 3$ sin ajuste adicional, como umbral inicial para todas las series.

Una vez entrenado el modelo LSTM-AE con la configuración base y aplicado el umbral de detección descrito anteriormente, se evaluó su capacidad para identificar anomalías reales en el conjunto de prueba. Esta evaluación se llevó a cabo comparando las predicciones del modelo (columna `anomaly_ae`) con las etiquetas de referencia proporcionadas por el comité experto.

Para cada pareja estación-contaminante se calcularon las métricas estándar de clasificación binaria: precisión, *recall* y *F1-score*. A fin de obtener una visión global del rendimiento, dichas métricas se agregaron por contaminante, promediando los valores obtenidos en todas las estaciones disponibles.

A continuación, se presenta un análisis detallado de estos resultados y de los casos representativos seleccionados para ilustrar visualmente el comportamiento del modelo en esta primera fase, sin ajuste de hiperparámetros.

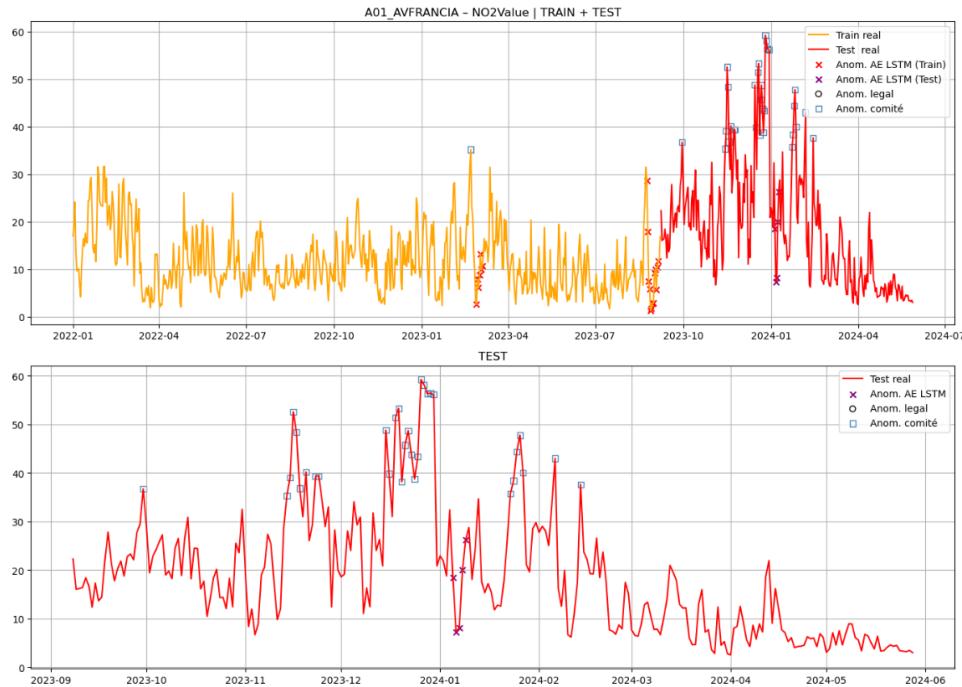


Figura 4.13: Resultados del modelo autoencoder LSTM para la estación AV_FRANCIA y el contaminante NO₂.

La Figura 4.13 muestra la serie temporal para NO₂ en la estación A01_AVFRANCIA en el que se observan en rojo los datos de test reales a lo largo del tiempo, con marcas (cruces rojas) donde el AE-LSTM del baseline indicó anomalías. También se señalan las anomalías «legales» (círculos) y las anotadas por el comité (cuadrados). Se observa que el modelo marcó sólo unos pocos picos como anómalos, e incluso en algunos casos las cruces se sitúan lejos de las etiquetas reales. Muchos eventos altos señalados por el comité no fueron detectados, mientras que algunos falsos picos menores sí fueron marcados. Esto evidencia que la sensibilidad del modelo es insuficiente: omite varias anomalías reales (bajo recall) y activa en momentos irrelevantes (baja precisión). La evolución general de la serie reconstruida por el autoencoder tiende a suavizar los picos pronunciados, por lo que muchos picos reales no producen error suficiente. En conjunto, este ejemplo pone de manifiesto la limitación del baseline en capturar la abrupta variabilidad del NO₂; refuerza la necesidad de reajustar parámetros para mejorar sensibilidad.

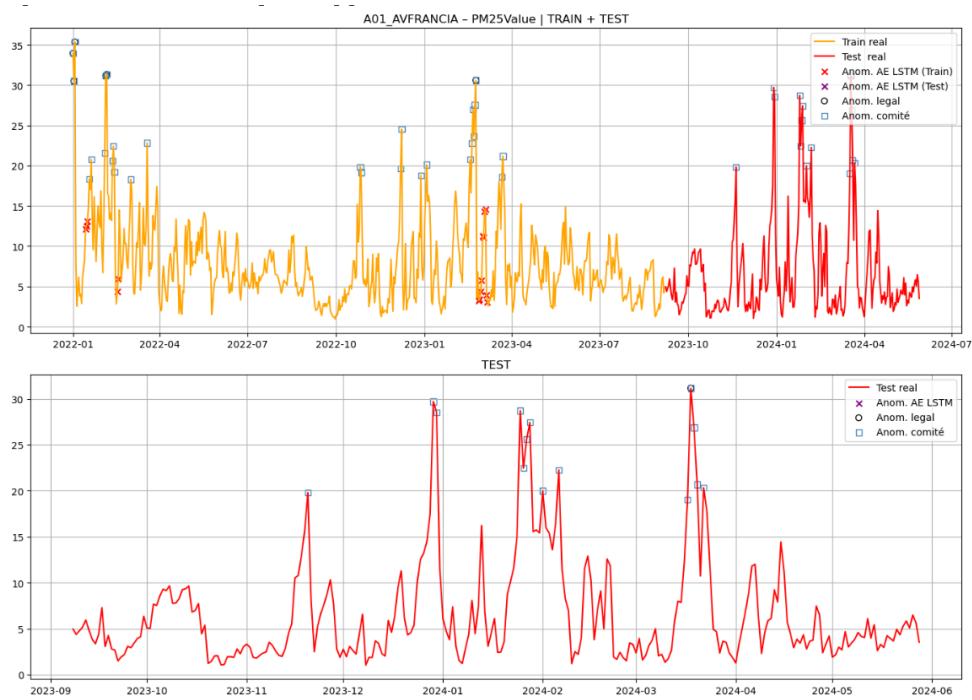


Figura 4.14: Resultados del modelo autoencoder LSTM para la estación VALENCIACENTRE y el contaminante PM_{2,5}.

La Figura 4.14 muestra, de forma análoga, la serie de PM_{2,5} también de la estación A01_AVFRANCIA, en la que se observan múltiples picos elevados (marcados por cuadrados) en los que el contaminante supera límites. El modelo *baseline*, sin embargo, sólo detectó algunas de estas crestas como anomalías (cruces rojas) y dejó otras sin marcar. Se aprecian lagunas donde hay etiquetas reales pero no hay detección, y pocas falsas alarmas cerca de picos más bajos.

Esto confirma nuevamente que el *recall* es bajo: sólo una fracción de las anomalías reales se capturan. Además, algunos picos detectados no siempre coinciden en el tiempo exacto. En suma, el *baseline* acertó parcialmente la detección de PM2.5, pero al igual que en NO₂, subestimó la mayoría de los eventos extremos y necesita ajuste de sensibilidad.

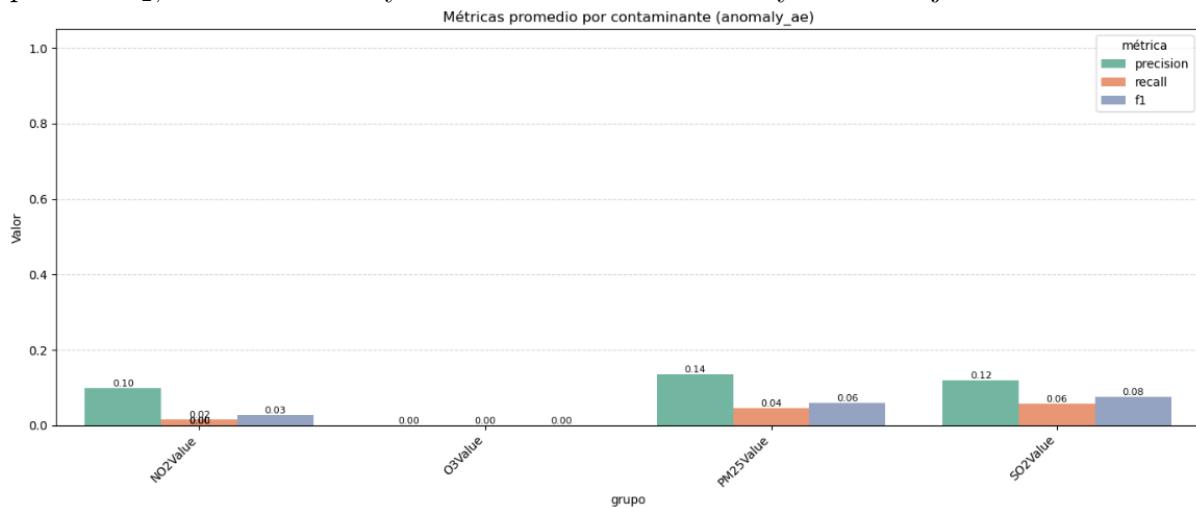


Figura 4.15: Métricas del modelo Isolation Forest por contaminante.

La Figura 4.15 muestra que todos los contaminantes presentaron valores de *recall* cercanos a cero y precisiones del orden de décimas (por ejemplo, $\sim 0,10$ en NO₂, $\sim 0,14$ en

PM2.5). Esto significa que el modelo identificó pocos eventos como anomalías (precisión baja) y capturó casi ninguno de los verdaderos (recall casi nulo). En esencia, el *baseline* falló en la detección sensible de anomalías en cualquier contaminante, especialmente en O₃, donde no se detectó ningún evento.

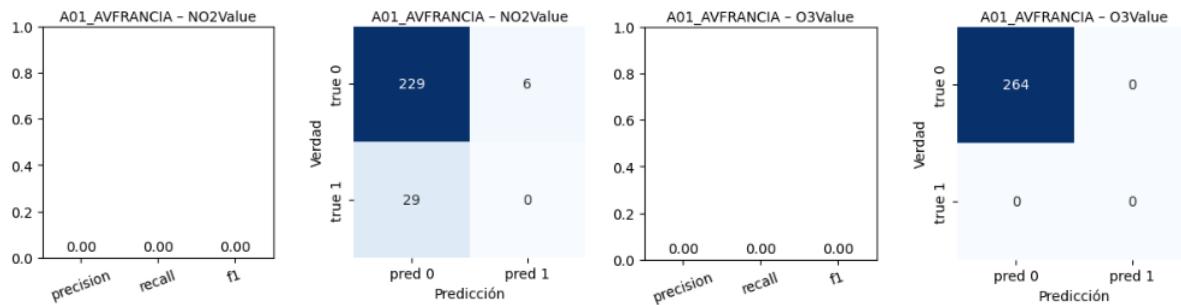


Figura 4.16: Comparación del rendimiento del autoencoder LSTM en la estación A01_AVFRANCIA para NO₂ (izquierda) y O₃ (derecha). Cada panel muestra, a la izquierda, las métricas *precision*, *recall* y *F1*; y, a la derecha, la matriz de confusión respectiva.

La Figura 4.16 muestra las matrices de confusión para los contaminantes NO₂ y O₃ en la estación A01_AVFRANCIA, confirmando las métricas pobres observadas previamente.

En el caso de NO₂, el modelo detectó 0 de las 29 anomalías reales (*recall* = 0) y marcó 6 falsos positivos, sin hallar ningún verdadero positivo. La diagonal principal muestra únicamente verdaderos negativos (229 aciertos) y los 6 falsos positivos; la fila correspondiente a anomalías reales está vacía en positivos. Este resultado evidencia que el umbral fue demasiado alto o el modelo excesivamente conservador, el modelo no logró capturar las variaciones agudas en NO₂, perdiendo todas las anomalías y limitando gravemente el *recall*.

Para O₃ en la misma estación, la matriz de confusión muestra que no hubo detecciones de ningún tipo (ningún verdadero positivo ni falso positivo). Esto confirma las métricas nulas: o bien no hubo anomalías presentes en el conjunto de prueba, o el modelo no marcó ninguna de las pocas existentes. En cualquier caso, el modelo fue completamente inefectivo con O₃; tanto su precisión como su *recall* son cero.

En conclusión, los resultados de la fase inicial indican un funcionamiento muy conservador del modelo base: se capturan pocas anomalías reales y se generan pocos falsos positivos. La precisión moderada, acompañada de un *recall* casi nulo, implica que el *F1-score* es muy bajo en todos los contaminantes.

Estos defectos observados en NO₂, O₃ y PM2.5 ponen en evidencia la necesidad de un ajuste de hiperparámetros para aumentar la capacidad de detección, especialmente mejorando el *recall* en aquellos casos donde el modelo no logra detectar anomalías existentes.

Ajuste de hiperparámetros

El conjunto de entrenamiento completo se dividió en dos subconjuntos temporales: un bloque “FIT” (aproximadamente 80 %) para ajustar los pesos del autoencoder y otro “VAL” (aproximadamente 20 %) para evaluar su desempeño. Esta separación respetando el orden cronológico evitó *futuros invasores*, de modo que el modelo se validó en datos posteriores no vistos durante el ajuste.

Así, se entrenó el LSTM-AE en FIT y se calculó el umbral (por ejemplo, $\mu + 3\sigma$) también en FIT. Luego, se evaluó la detección en VAL usando ese mismo umbral.

Se exploraron combinaciones de hiperparámetros fundamentales: longitud de ventana (w), dimensión latente (d) y coeficiente k del umbral (regla sigma). Por ejemplo, se probaron ventanas de 14 y 28 días, dimensiones latentes de 4, 8 y 16, y valores de k entre 2 y 3.5.

Para cada tupla (w, d, k) , se entrenó el autoencoder en FIT, se fijó el umbral en FIT según $\mu + k\sigma$, y se midió el desempeño en VAL calculando precisión, *recall* y *F1-score*. El uso de validación garantizó que la selección fuera robusta al sobreajuste. Esta estrategia de *grid search* sistemático suele generar mejoras sustanciales en modelos no supervisados, y es habitual en tareas de detección de anomalías ajustar los parámetros clave del autoencoder.

Como criterio de selección del mejor modelo se utilizó la puntuación F1 en el conjunto VAL (anomalías del comité). Se eligió la combinación (w, d, k) que maximizó F1 en VAL. Este enfoque pondera tanto los aciertos positivos como la cobertura de detección; de este modo se buscó un compromiso adecuado entre precisión y *recall*.

En general, se observó que valores más pequeños de k (umbral más bajo) aumentan el *recall* (más anomalías detectadas), pero reducen la precisión (más falsos positivos). La elección óptima equilibró estos efectos. Tras identificar el mejor modelo en VAL, este se reentrenó o se utilizó directamente para detectar anomalías en el conjunto de prueba final.

Se realizó un ajuste de hiperparámetros individual para cada combinación estación-contaminante, utilizando validación temporal (FIT/VAL) y seleccionando la configuración que maximizó el F1-score en el conjunto de validación. Sin embargo, los resultados obtenidos fueron en general muy bajos: la mayoría de las combinaciones presentan valores de F1 nulos o cercanos a cero, incluso tras la optimización. Solo en casos muy puntuales, como O₃ en A01_AVFRANCIA, se alcanzó una mejora moderada (F1 = 0,5).

Estos resultados confirman que el modelo base LSTM-AE, en su configuración inicial, tiene una capacidad limitada para detectar anomalías reales en las series analizadas.

Dado que los valores detallados por serie no aportan información adicional significativa, no se incluyen en el cuerpo del informe. No obstante, todos los experimentos, combinaciones evaluadas y métricas asociadas están disponibles en el repositorio de código y resultados adjunto para su consulta reproducible.

4.5. Comparación

A lo largo del estudio se implementaron tres enfoques no supervisados, ARIMA, Isolation Forest y Autoencoder LSTM, para detectar anomalías en series temporales de contaminación. Aunque cada uno parte de una lógica distinta (modelado estadístico, partición aleatoria y reconstrucción neuronal, respectivamente), su desempeño se evaluó de forma comparable mediante métricas estándar (precisión, recall, F1-score) frente a dos tipos de referencia: anomalías del comité experto y superaciones legales.

Modelo	Prec. (NO_2)	Recall	F_1
ARIMA	Alta (~0.78)	Moderada	Media
IF (base)	Baja	Alta (~0.71)	Media
IF (ajustado)	Media (~0.40)	Alta (~0.72)	Alta (~0.45)
LSTM-AE (base)	Baja (~0.10)	Muy baja (~0)	Muy baja
LSTM-AE (ajustado)	Baja–media	Leve mejora	Ligera mejora

Tabla 4.3: Comparación de modelos en NO_2 (frente a etiquetas del comité).

Modelo	Prec. (PM2.5)	Recall	F_1
ARIMA	Alta (~0.75)	Baja (~0.27)	Baja (~0.39)
IF (base)	Moderada (~0.48)	Media	Media
IF (ajustado)	Alta (~0.63)	Media (~0.54)	Alta (~0.54)
LSTM-AE (base)	Baja (~0.14)	Muy baja	Muy baja
LSTM-AE (ajustado)	Nulo	Nulo	Nulo

Tabla 4.4: Comparación de modelos en PM_{2.5} (frente a etiquetas del comité).

Modelo	Prec. (O_3)	Recall	F_1
ARIMA	Nula (~0)	Nula	0
IF (base)	Nula	Nula	0
IF (ajustado)	Muy baja	Muy baja	~0
LSTM-AE (base)	Nula	Nula	0
LSTM-AE (ajustado)	Nulo	Nulo	Nulo

Tabla 4.5: Comparación de modelos en O_3 (frente a etiquetas del comité).

Modelo	Prec. (SO_2)	Recall	F_1
ARIMA	Nula (~0)	Nula	0
IF (base)	Nula	Nula	0
IF (ajustado)	Muy baja	Muy baja	~0
LSTM-AE (base)	Nula	Nula	0
LSTM-AE (ajustado)	Nulo	Nulo	Nulo

Tabla 4.6: Comparación de modelos en SO_2 (frente a etiquetas del comité).

Las Tablas 4.3–4.6 resumen el rendimiento relativo de cada modelo por contaminante. En conjunto, el modelo Isolation Forest tras ajuste de hiperparámetros mostró el mejor equilibrio entre precisión y recall, particularmente en NO_2 y PM_{2.5}. El modelo ARIMA presentó una alta precisión, pero menor cobertura (recall), mientras que el autoencoder LSTM, incluso tras ajustes, ofreció el rendimiento más limitado en todos los contaminantes, destacando su baja sensibilidad y alta omisión de anomalías reales. Para contaminantes como O_3 y SO_2 , todos los modelos fracasaron en ofrecer detecciones útiles, lo que señala la necesidad de incorporar información adicional o adoptar enfoques supervisados.

Además de evaluar el rendimiento mediante métricas cuantitativas, es fundamental analizar cómo se alinean los modelos con los criterios de referencia utilizados en el estudio, tanto desde un enfoque técnico como desde una perspectiva normativa. En este caso, se consideraron dos tipos de referencia para validar las anomalías detectadas: (i) las etiquetas proporcionadas por el comité de expertos y (ii) los eventos que suponen una superación

de los umbrales legales de calidad del aire.

El modelo ARIMA mostró una buena adaptación a ambos criterios. Su funcionamiento conservador y su precisión elevada le permiten coincidir con un número considerable de eventos marcados tanto por el comité como por la normativa, especialmente aquellos de carácter extremo. Esto lo convierte en una opción adecuada cuando se prioriza la interpretabilidad y el cumplimiento normativo.

Isolation Forest también logró una buena coincidencia con ambas referencias, especialmente tras el ajuste del hiperparámetro **contamination**. Este ajuste permitió al modelo adaptarse mejor a la frecuencia esperada de anomalías reales, lo que incrementó su sensibilidad y capacidad para capturar eventos significativos, sin perder precisión de forma drástica. Esta flexibilidad lo hace útil para aplicaciones operativas, donde se requiere una detección más reactiva y adaptada a los cambios en los datos.

Por el contrario, el autoencoder LSTM no mostró una adaptación satisfactoria. Debido a su baja sensibilidad y a la ausencia de una integración explícita de criterios legales o conocimiento experto en su entrenamiento, el modelo fue incapaz de detectar correctamente los eventos de interés definidos por el comité o por la normativa. En consecuencia, su uso en contextos regulatorios o de apoyo a la toma de decisiones resulta, al menos en esta configuración, poco recomendable.

En resumen, tanto ARIMA como Isolation Forest demostraron ser compatibles con los criterios de evaluación considerados, aunque con diferencias en su comportamiento: ARIMA destaca por su precisión y rigor normativo, mientras que Isolation Forest ofrece mayor sensibilidad y versatilidad. El autoencoder LSTM, en cambio, requiere ajustes más profundos o información supervisada adicional para ser útil en este tipo de aplicaciones.

A partir del análisis integral de los modelos implementados, se identificaron diferencias significativas en su rendimiento según el criterio de evaluación y el contexto de aplicación. Esta comparación permite valorar cuál de los enfoques resulta más adecuado en función de los objetivos específicos del sistema de detección de anomalías.

El modelo ARIMA destacó por su comportamiento conservador, emitiendo alertas únicamente ante desviaciones muy pronunciadas respecto al patrón esperado. Esta prudencia lo convierte en una herramienta útil en contextos normativos o institucionales, donde se prioriza la precisión y se busca minimizar la emisión de falsas alarmas. Además, su base estadística lo hace fácilmente interpretable, facilitando su validación por parte de expertos y su integración en entornos donde la trazabilidad del modelo es un requisito.

Por otro lado, el modelo Isolation Forest ajustado fue el más eficaz en términos de sensibilidad, mostrando una alta capacidad para detectar un número amplio de anomalías reales, incluyendo aquellas de menor intensidad. Esta ventaja se deriva de su flexibilidad, especialmente tras el ajuste del parámetro **contamination**, lo que le permitió adaptarse al comportamiento específico de cada serie. Además, demostró ser especialmente robusto frente a series temporales complejas, donde otros modelos tienden a fallar. La incorporación de estadísticas móviles y variables temporales adicionales reforzó su capacidad de generalización, haciendo de este enfoque una opción potente para sistemas de vigilancia con elevada exigencia operativa.

En contraste, el autoencoder LSTM mostró el rendimiento más limitado entre los modelos evaluados. A pesar de su complejidad y del potencial teórico que presenta para modelar relaciones no lineales y patrones de largo plazo, su baja sensibilidad y escasa

coincidencia con las etiquetas de referencia lo hicieron poco efectivo en la práctica. Incluso tras ajustar sus hiperparámetros, el modelo mantuvo un *recall* cercano a cero en la mayoría de los casos, lo que compromete seriamente su utilidad como herramienta autónoma de detección.

En conjunto, los resultados obtenidos confirman que no existe un único modelo óptimo para todas las situaciones. La elección del enfoque más adecuado depende de múltiples factores, como el tipo de contaminante, la complejidad estructural de la serie temporal y los objetivos específicos del sistema (ya sea priorizar la precisión, la cobertura, la interpretabilidad o la robustez ante diferentes condiciones). Esta conclusión refuerza la necesidad de adoptar estrategias flexibles, posiblemente combinando distintos modelos o ajustando su configuración a cada contexto particular.

El modelo más efectivo globalmente fue Isolation Forest ajustado, que alcanzó el mejor equilibrio entre precisión y *recall* en contaminantes como NO₂ y PM2.5. Logró aprender adecuadamente los patrones de comportamiento usando características temporales, y su rendimiento mejoró significativamente con la calibración del umbral (**contamination**).

Por su parte, ARIMA ofreció un enfoque robusto y preciso, útil para detectar los eventos más extremos, especialmente cuando la serie muestra estacionalidades claras. Sin embargo, su limitada sensibilidad lo hace insuficiente en escenarios donde se requiere una cobertura amplia de anomalías.

El autoencoder LSTM, pese a su complejidad y potencial teórico, resultó ser el menos útil en la práctica con los datos disponibles. Incluso tras el ajuste de sus hiperparámetros, mostró un rendimiento pobre, con *recall* cercano a cero, lo que lo hace inadecuado como detector autónomo en este contexto.

Capítulo 5

Conclusiones

5.1. Conclusiones

Este Trabajo de Fin de Grado planteó como objetivo central la detección automática de anomalías en series temporales de contaminación atmosférica en Valencia, comparando varios modelos no supervisados. Para ello, se definieron metas específicas: (1) evaluar el rendimiento de diferentes algoritmos en distinguir episodios contaminantes significativos de la variabilidad habitual; (2) caracterizar las anomalías según su duración y distribución estacional; (3) comparar la incidencia de eventos anómalos entre distintas estaciones de monitoreo; y (4) proponer criterios operativos para la identificación automática de anomalías. En función de los resultados obtenidos, estos objetivos se han cumplido en buena medida. Se lograron implementar y probar tres enfoques evaluando cuantitativamente su precisión y cobertura en la detección de picos atípicos. Asimismo, se caracterizaron las anomalías en el dominio temporal y estacional, observando, por ejemplo, que los contaminantes como PM_{2.5} presentaron picos anómalos frecuentes en invierno, mientras que en O₃ apenas se registraron eventos fuera de lo común, salvo algunas superaciones puntuales en verano. También se identificaron diferencias por ubicación, con estaciones de tráfico mostrando mayores repuntes de NO₂, frente a estaciones de fondo donde predominan anomalías de O₃ o partículas. Finalmente, se establecieron criterios de detección operativos basados en umbrales estadísticos, sentando las bases para un sistema de alerta en tiempo real. En conjunto, el estudio provee evidencia de que es viable detectar anomalías de calidad del aire de forma no supervisada con buena precisión y razonable capacidad descriptiva, aunque con matices importantes en función del método.

La comparación entre ARIMA, Isolation Forest y LSTM-AE muestra que cada modelo presenta ventajas y limitaciones. En particular, el modelo ARIMA (SARIMA) adoptó un enfoque conservador: logró alta precisión en la detección de picos extremos al ajustarse bien a la estacionalidad de la serie, pero mostró baja sensibilidad, omitiendo numerosos repuntes moderados. Esto se tradujo en un número reducido de falsas alarmas, pero también en una considerable cantidad de falsos negativos. En contaminantes como NO₂ y PM_{2.5}, solo detectó los valores más altos, alcanzando precisiones superiores al 75 %, aunque con *recall* entre 0.27 y 0.61. En el caso de O₃ y SO₂, prácticamente no detectó anomalías, dado que la serie se mantenía cerca de su media y no superaba el umbral de 3 σ . En resumen, ARIMA ofrece un buen filtro de anomalías evidentes, pero resulta insuficiente para una detección completa, lo que motivó la incorporación de modelos más sensibles.

El modelo Isolation Forest (IF) mostró un enfoque complementario al de ARIMA: presentó una alta sensibilidad, detectando la mayoría de los picos relevantes (por ejemplo, con *recall* ~ 0.76 en NO₂), pero a costa de una precisión más baja (entre 0.18 y 0.22 en algunos casos). Al entrenarse con múltiples características temporales y sin necesidad de etiquetas, fue capaz de identificar patrones anómalos con eficacia, aunque también generó varias falsas alarmas, especialmente ante fluctuaciones intermedias. En estaciones como A01_AVFRANCIA, detectó bien los picos invernales reales, pero también marcó valores no críticos como anómalos. Este desequilibrio entre verdaderos y falsos positivos se reflejó en un F1-score modesto. Además, en contaminantes con poca variabilidad como O₃, el modelo fue excesivamente conservador, resultando en métricas prácticamente nulas.

Una de las principales ventajas de Isolation Forest es su flexibilidad para ajustar el umbral de detección mediante el parámetro **contamination**. En este trabajo se exploraron diferentes configuraciones (umbral, número de árboles, tamaño de muestra) buscando maximizar el F1-score por serie. Tras este ajuste, el modelo mejoró notablemente: en NO₂, la precisión aumentó del 22 % al 40 %, manteniendo un *recall* alto (0.72), y el F1 pasó de 0.33 a 0.45. En PM_{2.5}, ambos indicadores superaron el 50 %. Esto demuestra que, bien calibrado, el modelo puede detectar eficazmente la mayoría de los eventos relevantes, reduciendo sustancialmente las falsas alarmas. Sin embargo, en contaminantes con baja variabilidad o sin etiquetas representativas, como O₃ y SO₂, la mejora fue limitada. En conjunto, Isolation Forest resultó ser el modelo más sensible, aunque su uso efectivo requiere una calibración cuidadosa para evitar sobreseñalar fluctuaciones normales.

El modelo de autoencoder LSTM presentó un rendimiento muy limitado en la detección de anomalías. En su configuración base, con una ventana de 14 días, dimensión latente 8 y un umbral de detección basado en $\mu + 3\sigma$, resultó ser excesivamente conservador. Su reconstrucción tendía a suavizar las variaciones, lo que impedía que muchos picos reales generaran errores suficientemente altos como para ser considerados anómalos. Esto se tradujo en una sensibilidad extremadamente baja (*recall* ~ 0), como en el caso de NO₂ en la estación A01_AVFRANCIA, donde no detectó ninguna de las 29 anomalías reales. Las métricas globales iniciales fueron muy bajas en todos los contaminantes, con precisiones de apenas 0,10–0,15.

Se realizó un ajuste de hiperparámetros (ventana, dimensión latente, umbral) mediante validación temporal, pero los resultados siguieron siendo pobres, con valores de F1 cercanos a cero en casi todos los casos. Solo se observó alguna mejora aislada (por ejemplo, F1 ~ 0.5 en un caso de O₃), insuficiente para considerar el modelo funcional. En resumen, el LSTM-AE no supervisado no fue eficaz en este contexto, probablemente debido a una arquitectura poco adecuada, falta de etiquetas y una reconstrucción que prioriza la señal general sobre los detalles anómalos.

A pesar de las limitaciones individuales señaladas, en términos generales los enfoques no supervisados demostraron ser herramientas valiosas para la detección de anomalías en series de contaminación atmosférica, especialmente dada la carencia de etiquetas *a priori*. Su principal virtud es que permiten identificar patrones inusuales sin necesidad de un conjunto de entrenamiento etiquetado, algo crucial en problemas donde los eventos anómalos históricos no están catalogados exhaustivamente.

En este trabajo, ARIMA, Isolation Forest y LSTM pudieron entrenarse únicamente con datos no anotados, aprendiendo cada uno a su manera el “comportamiento normal” de la serie y señalando desviaciones significativas. Esto habilita la detección temprana de

episodios anómalos potencialmente peligrosos que podrían pasar inadvertidos si solo se aplicaran criterios estáticos, como los umbrales legales fijos. En este sentido, los métodos no supervisados complementan los enfoques tradicionales basados en reglas duras, al proporcionar una mirada adaptativa a lo que constituye un comportamiento inusual.

La eficacia real de estos enfoques depende en gran medida de una calibración adecuada y del contexto en que se apliquen. Al no contar con una guía externa (etiquetas verdaderas) durante el entrenamiento, los modelos requieren supuestos o umbrales bien fundamentados para distinguir anomalías de variaciones normales. Tal como se observó, pequeñas decisiones como la selección del umbral de 3σ en ARIMA, el valor de `contamination` en Isolation Forest o el factor k en el LSTM-AE tienen un impacto crítico en el equilibrio entre falsos positivos y falsos negativos.

Un umbral demasiado laxo deriva en numerosas falsas alarmas que podrían desensibilizar el sistema (como ilustró el caso inicial de Isolation Forest, con una precisión cercana a 0,2), mientras que un umbral demasiado estricto hace que solo se detecten las anomalías más extremas (como ocurrió con ARIMA y el autoencoder, que ignoraron muchos eventos relevantes). Por ello, en la práctica, la utilidad de los métodos no supervisados está condicionada a un riguroso ajuste y validación cruzada, idealmente incorporando conocimiento experto para fijar parámetros razonables.

Uno de los desafíos centrales enfrentados en este proyecto fue cómo evaluar y validar los modelos en ausencia de un conjunto completo de etiquetas de anomalía. A diferencia de un problema supervisado clásico, aquí no existía un *ground truth* completamente confiable sobre qué instantes fueron anómalos en la historia, más allá de excedencias regulatorias puntuales.

Para superar esta dificultad, se adoptó una estrategia de construcción de referencia combinando criterios automáticos y conocimiento experto. En concreto, se definió un etiquetado de referencia “semi-supervisado” mediante: (i) un comité de detectores estadísticos clásicos (p. ej., umbral por percentiles, regla de Hampel, etc.), donde un punto se consideró anómalo si al menos tres métodos coincidían; y (ii) las superaciones de umbrales legales diarios u horarios establecidos por la normativa de calidad del aire. Todo instante señalado por al menos tres de cuatro criterios y/o asociado a una violación normativa fue tomado como anomalía de referencia.

Este esquema proporcionó un *ground truth* aproximado contra el cual medir las detecciones de los modelos. Gracias a ello, se pudieron calcular métricas objetivas (precisión, *recall*, F1) para comparar métodos y afinar sus parámetros. No obstante, es importante resaltar que esta referencia no es perfecta: es posible que ciertos eventos anómalos sutiles no fueran capturados por los detectores simples del comité y, por tanto, falten en la etiqueta; o que algunas “anomalías” señaladas por consenso no revistan en realidad una importancia ambiental relevante.

En conjunto, el análisis comparativo llevado a cabo permite concluir que no existe un modelo único que domine en todos los aspectos, sino que cada enfoque no supervisado aportó una perspectiva complementaria en la detección de anomalías. El esquema de predicción y residuales (ARIMA) brindó confiabilidad al emitir solo alarmas ante desviaciones muy marcadas (útil para evitar sobresaltos infundados), mientras que el Isolation Forest ofreció una mirada más amplia, capaz de descubrir muchos más eventos atípicos (útil para no perder posibles episodios de interés). Por su parte, el LSTM-AE, aunque

deficiente en su estado actual, representa la línea de métodos de aprendizaje profundo que podrían capturar relaciones temporales complejas inadvertidas por los enfoques clásicos.

La utilidad de los métodos no supervisados en este contexto se considera probada, siempre que se apliquen con calibración cuidadosa y se interpreten a la luz de la normativa ambiental y del conocimiento experto de quienes gestionan la calidad del aire. Este TFG demuestra la viabilidad de implementar un sistema de alerta temprana basado en datos, capaz de rastrear patrones anómalos en tiempo real y servir de apoyo a la toma de decisiones ambientales. Al mismo tiempo, pone de manifiesto las áreas de mejora y consideraciones necesarias para que dicho sistema sea confiable: desde la obtención de mejores datos de referencia hasta la combinación de múltiples métodos para reforzar la detección.

5.2. Trabajo futuro

Considerando las limitaciones y hallazgos de este trabajo, se proponen varias direcciones para futuras investigaciones y desarrollos. Una de las más relevantes es la construcción de un *ground truth* de anomalías más completo y preciso. Para ello, se sugiere implementar un proceso semiautomatizado de etiquetado, combinando detección algorítmica y validación experta, con el fin de generar conjuntos de datos anotados de mayor calidad. Por ejemplo, se podrían aplicar múltiples detectores no supervisados junto con reglas de negocio sobre datos históricos para identificar posibles anomalías, y posteriormente someter estos resultados a revisión por parte de expertos. Este enfoque permitiría consolidar un corpus fiable tanto para la evaluación objetiva de nuevos modelos como para el entrenamiento de enfoques supervisados en trabajos futuros.

Otra línea de trabajo futuro consiste en la incorporación de variables externas al modelo, con el objetivo de mejorar sustancialmente la detección de anomalías. Variables exógenas como las condiciones meteorológicas y factores de actividad humana influyen directamente en las concentraciones de contaminantes y permiten contextualizar mejor sus fluctuaciones. Incluir esta información como entradas adicionales ayudaría a los algoritmos a distinguir entre incrementos esperables y verdaderas anomalías no explicadas por factores conocidos. Esta integración podría aplicarse tanto en modelos como Isolation Forest, mediante la ampliación del vector de entrada, como en redes neuronales multimodales que combinen series de contaminación con variables meteorológicas. Incorporar este tipo de contexto contribuiría a reducir falsos positivos debidos a variabilidad normal, aumentando la precisión sin perjudicar la sensibilidad del sistema.

Una tercera línea de trabajo futuro es la exploración ampliada del espacio de hiperparámetros y arquitecturas del modelo LSTM-AE, dado su limitado rendimiento en la configuración básica utilizada. Esta mejora implicaría una búsqueda más exhaustiva que incluya tamaños de ventana más largos o adaptativos, diferentes dimensiones latentes, un mayor número de capas (posiblemente con arquitecturas LSTM encoder-decoder apiladas), variantes en funciones de activación y en métricas de error, así como métodos alternativos para definir umbrales, por ejemplo usando percentiles en lugar de múltiples de la desviación estándar. Asimismo, se podrían evaluar técnicas de regularización más robustas y arquitecturas alternativas, como autoencoders convolucionales o modelos secuencia-a-secuencia con mecanismos de atención, que podrían adaptarse mejor a la complejidad temporal de las series ambientales.

Apéndice A

Anexos

Generación de series temporales por estación y contaminante

```
def series_estacion_contaminante(df, estaciones, contaminantes):
    df = df.copy()
    df.index = pd.to_datetime(df.index)
    series_dict = {}
    for estacion in estaciones:
        df_est = df[df['entityId'] == estacion].copy()
        for cont in contaminantes:
            if cont not in df_est.columns:
                continue

            # Identifica etiquetas de anomalía disponibles
            etiquetas = [f"{cont}_anom_comite", f"{cont}_anom_legal"]
            columnas = [cont] + [e for e in etiquetas if e in df_est.columns]

            # Renombra y filtra valores nulos
            df_sub = df_est[columnas].rename(columns={cont: 'value'}).dropna(subset=['value'])
            df_sub = df_sub.asfreq('D') # reindexa con frecuencia diaria
            df_sub['value'] = df_sub['value'].interpolate()

            if df_sub.shape[0] >= 50:
                series_dict[(estacion, cont)] = df_sub

    return series_dict
```

División train / test

```
def dividir_st(df, fecha_col='dateObserved'):
    df = df.sort_index()
    total_dias = len(df)
    n_train = int(0.7 * total_dias)

    df_train = df.iloc[:n_train]
    df_test = df.iloc[n_train:]

    return df_train, df_test
```

Funciones para evaluar y comparar anomalías

```
def evaluar_anomalias(df, etiqueta_ref='anom_comite', col_modelo='anomaly'):
```

```

if etiqueta_ref not in df.columns or col_modelo not in df.columns:
    return None

y_true = df[etiqueta_ref].fillna(False).astype(bool)
y_pred = df[col_modelo].fillna(False).astype(bool)
tp = ((y_true == True) & (y_pred == True)).sum()
return {
    'precision': precision_score(y_true, y_pred, zero_division=0),
    'recall': recall_score(y_true, y_pred, zero_division=0),
    'f1': f1_score(y_true, y_pred, zero_division=0),
    'n_comite': y_true.sum(),
    'n_detectadas': y_pred.sum(),
    'tp': tp}

```

```

def comparar_anomalias(df, etiqueta, col_modelo):
    # Validar columnas
    faltan = [c for c in (col_modelo, etiqueta) if c not in df.columns]
    if faltan:
        print(f" Saltando comparación: faltan columnas {faltan}")
        return {}
    # Máscaras booleanas
    mask_modelo = df[col_modelo].fillna(False).astype(bool)
    mask_manual = df[etiqueta].fillna(False).astype(bool)
    # Índices
    idx_modelo = df.index[mask_modelo]
    idx_manual = df.index[mask_manual]
    # Coincidencias y discrepancias
    coinc = idx_modelo.intersection(idx_manual)
    fn    = idx_manual.difference(idx_modelo)
    fp    = idx_modelo.difference(idx_manual)
    # Informe
    print(f"    Coinciden {len(coinc)} de {len(idx_manual)} anomalías manuales.")
    if len(idx_modelo) > 0:
        prec = len(coinc) / len(idx_modelo)
        print(f"    Precisión de coincidencias: {prec:.2f}")
    else:
        print("    El modelo no detectó ninguna anomalía.")
    # Devolver subconjuntos
    return {
        'coincidencias': df.loc[coinc].copy(),
        'falsos_negativos': df.loc[fn].copy(),
        'falsos_positivos': df.loc[fp].copy(),}

```

```

def evaluar_y_visualizar_anomalias(ruta_excel, col_modelo='anomaly', etiqueta_base='anom_comite'):
    xlsx = pd.ExcelFile(ruta_excel)
    hojas = xlsx.sheet_names
    resultados = []
    for hoja in hojas:
        df = xlsx.parse(hoja, index_col=0, parse_dates=True)
        partes = hoja.split('_')
        if len(partes) < 3:
            continue

        estacion = "_".join(partes[:2])
        contaminante_base = partes[2]
        contaminante = contaminante_base + 'Value'

```

```

etiqueta = f"{contaminante}_{etiqueta_base}"
print(f"Estación: {estacion} | Contaminante: {contaminante}")

# Comparar anomalías (opcional)
comparar_anomalias(df, etiqueta=etiqueta, col_modelo=col_modelo)

# Evaluar
resultado = evaluar_anomalias(df, etiqueta_ref=etiqueta, col_modelo=col_modelo)
if resultado:
    resultado.update({
        'estacion': estacion,
        'contaminante': contaminante})
    resultados.append(resultado)

# Si no hay resultados, salir
if not resultados:
    print("No se encontraron resultados válidos para evaluar.")
    return

# Convertir a DataFrame
df_resultados = pd.DataFrame(resultados)
columnas_esperadas = ['estacion', 'contaminante', 'n_comite', 'n_detectadas', 'tp',
                      'precision', 'recall', 'f1']
if not all(col in df_resultados.columns for col in columnas_esperadas):
    print("El DataFrame de resultados no contiene todas las columnas necesarias.")
    print("Columnas disponibles:", df_resultados.columns.tolist())
    return
df_resultados = df_resultados[columnas_esperadas]

# Promedios por contaminante
df_summary = df_resultados.groupby(['contaminante'])[['precision', 'recall', 'f1']].mean().reset_index()

# Preparar para gráfico
df_melt = df_summary.melt(id_vars=['contaminante'],
                           value_vars=['precision', 'recall', 'f1'],
                           var_name='métrica', value_name='valor')
df_melt['grupo'] = df_melt['contaminante']

# Gráfico
plt.figure(figsize=(14, 6))
ax = sns.barplot(data=df_melt, x='grupo', y='valor', hue='métrica', palette='Set2')
for p in ax.patches:
    altura = p.get_height()
    if not pd.isna(altura):
        ax.annotate(f'{altura:.2f}', (p.get_x() + p.get_width() / 2., altura),
                    ha='center', va='bottom', fontsize=8)
plt.title(f'Métricas promedio por contaminante ({col_modelo})')
plt.ylabel("Valor")
plt.ylim(0, 1.05)
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

```

Métricas por estación y contaminante + matriz de confusión

```

def plot_metrics_confusion(df, etiqueta_ref, col_modelo, title):
    # datos
    y_true = df[etiqueta_ref].fillna(False).astype(int)

```

```

y_pred = df[col_modelo].fillna(False).astype(int)

prec = precision_score(y_true, y_pred, zero_division=0)
rec = recall_score(y_true, y_pred, zero_division=0)
f1 = f1_score(y_true, y_pred, zero_division=0)
cm = confusion_matrix(y_true, y_pred, labels=[0, 1])

# figura + axes
fig, (ax_bar, ax_cm) = plt.subplots(
    1, 2, figsize=(7, 3), dpi=100, gridspec_kw={"wspace": 0.35})

# barra métricas
sns.barplot(x=["precision", "recall", "f1"],
             y=[prec, rec, f1],
             palette="Set2",
             ax=ax_bar)
ax_bar.set_ylim(0, 1)
ax_bar.set_title(title, fontsize=10, pad=4)
for i, v in enumerate([prec, rec, f1]):
    ax_bar.text(i, v + 0.03, f"{v:.2f}", ha="center")
ax_bar.tick_params(axis="x", rotation=20)

# matriz de confusión
sns.heatmap(cm, annot=True, fmt="d", cbar=False, ax=ax_cm, cmap="Blues",
            xticklabels=["pred 0", "pred 1"],
            yticklabels=["true 0", "true 1"])
ax_cm.set_title(title, fontsize=10, pad=4)
ax_cm.set_xlabel("Predicción")
ax_cm.set_ylabel("Verdad")
plt.tight_layout()
plt.show()

```

Visualizar anomalías detectadas

```

def visualizar_anomalias(ruta_excel, estacion, contaminante,
                         col_anom_modelo='anomaly', modelo_label='ARIMA'):
    # Preparar nombres de hoja
    base = contaminante.replace("Value", "")
    hoja_train = f"{estacion}_{base}_train"
    hoja_test = f"{estacion}_{base}_test"
    xlsx = pd.ExcelFile(ruta_excel)

    # Cargar datos
    df_train = xlsx.parse(hoja_train, index_col=0, parse_dates=True)
    df_test = xlsx.parse(hoja_test, index_col=0, parse_dates=True)
    est = estacion + '_10m'
    col_legal = f"{contaminante}_anom_legal"
    col_comite = f"{contaminante}_anom_comite"
    fig, axes = plt.subplots(2, 1, figsize=(14, 10), sharex=False)

    # TRAIN + TEST
    ax = axes[0]
    ax.set_title(f"{estacion} \ {contaminante} | TRAIN + TEST")
    ax.plot(df_train['value'], label='Train real', color='orange')
    ax.plot(df_test['value'], label='Test real', color='red')
    if 'pred' in df_test.columns:
        ax.plot(df_test['pred'], label='Test pred', linestyle='--', color='purple')
    # Anomalías del modelo
    for df, label, color in zip([df_train, df_test], ['Train', 'Test'], ['red', 'purple']):

```

```

if col_anom_modelo in df.columns:
    mask = df[col_anom_modelo].fillna(False).astype(bool)
    ix = df.index[mask]
    ax.scatter(ix, df.loc[ix, 'value'], marker='x', color=color,
               label=f'Anom. {modelo_label} ({label})')

# Anomalías legales
df_all = pd.concat([df_train, df_test]).sort_index()
if col_legal in df_all.columns:
    mask = df_all[col_legal].fillna(False).astype(bool)
    ix = df_all.index[mask]
    ax.scatter(ix, df_all.loc[ix, 'value'],
               facecolors='none', edgecolors='black',
               marker='o', label='Anom. legal')

# Anomalías del comité
if col_comite in df_all.columns:
    fechas = comite_indices_estacion[est][contaminante]
    fechas_format = [ts.tz_convert(None) if ts.tzinfo else ts for ts in fechas]
    valores = df_all.loc[fechas_format, 'value']
    ax.scatter(fechas, valores,
               facecolors='none', edgecolors='steelblue',
               marker='s', label='Anom. comité')

ax.legend(loc='upper right')
ax.grid(True)

# TEST
ax = axes[1]
ax.set_title("TEST")
ax.plot(df_test['value'], label='Test real', color='red')
if 'pred' in df_test.columns:
    ax.plot(df_test['pred'], label='Test pred', linestyle='--', color='purple')

# Modelo
if col_anom_modelo in df_test.columns:
    mask = df_test[col_anom_modelo].fillna(False).astype(bool)
    ix = df_test.index[mask]
    ax.scatter(ix, df_test.loc[ix, 'value'],
               marker='x', color='purple',
               label=f'Anom. {modelo_label}')

# Legal
if col_legal in df_test.columns:
    mask = df_test[col_legal].fillna(False).astype(bool)
    ix = df_test.index[mask]
    ax.scatter(ix, df_test.loc[ix, 'value'],
               facecolors='none', edgecolors='black',
               marker='o', label='Anom. legal')

# Comité
if col_comite in df_test.columns:
    fechas = comite_indices_estacion[est][contaminante]
    fechas_format = [ts.tz_convert(None) if ts.tzinfo else ts for ts in fechas]
    index_test = df_test.index.intersection(fechas_format)
    valores = df_test.loc[index_test, 'value']
    ax.scatter(index_test, valores,
               facecolors='none', edgecolors='steelblue',
               marker='s', label='Anom. comité')

ax.legend()
ax.grid(True)
plt.tight_layout()
plt.show()

```


Bibliografía

- [1] V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. *ACM Computing Surveys*, 2009.
- [2] Complex Systems AI. Detección de anomalías en series temporales: sistemas complejos e IA. <https://complex-systems-ai.com/es/pronostico-de-prediccion/deteccion-de-anomalias/#Point-Outlier>. Accedido el 17 de julio de 2025.
- [3] Stylianos (Stelios) Kampakis. 3 Types of Anomalies in Anomaly Detection. <https://hackernoon.com/3-types-of-anomalies-in-anomaly-detection>, 2022. Accedido el 17 de julio de 2025.
- [4] VictoriaMetrics. Anomaly Detection for Time Series Data: Anomaly Types (Chapter 2). <https://victoriametrics.com/blog/victoriametrics-anomaly-detection-handbook-chapter-2/>. Accedido el 17 de julio de 2025.
- [5] V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, Chichester, England, 3rd edition, 1994.
- [6] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, NY, 1986.
- [7] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.
- [8] Chang C.-K. Chang et al. Detection and quantification of anomalies in communication ... *PLOS ONE / PubMed Central*, 2022. ARIMA: combinación de AR, I y MA componentes.
- [9] Neptune.ai. Anomaly detection in time series. <https://neptune.ai/blog/anomaly-detection-in-time-series>. Accedido el 17 de julio de 2025.
- [10] Alkaline-ML developers. Manual de pmdarima auto_arima. https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html. Accedido el 17 de julio de 2025.
- [11] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. *Isolation-Based Anomaly Detection*. Association for Computing Machinery (ACM), 2008.
- [12] scikit-learn developers. Isolationforest — scikit-learn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>. Accedido el 17 de julio de 2025.
- [13] DigitalOcean. Anomaly detection in python with isolation forest. <https://www.digitalocean.com/community/tutorials/anomaly-detection-isolation-forest>. Accedido el 17 de julio de 2025.
- [14] OneNet Project. Literature review of anomaly detection on time series data. https://onenet-project.eu/wp-content/uploads/2023/05/OneNet-D1.1_Presify_v1.0.pdf. Accedido el 17 de julio de 2025.

- [15] Yuanyuan Wei, Julian Jang-Jaccard, Wen Xu, Fariza Sabrina, Seyit Camtepe, and Mikael Boulic. Lstm-autoencoder based anomaly detection for indoor air quality time series data. *IEEE Sensors Journal*, 2022.
- [16] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *Proceedings of ACM/IEEE Conference on Big Data Analytics*, 2016.
- [17] Unnamed Author. Unsupervised outlier detection for time-series data of indoor air. *Journal of Big Data*, 2023. uses LSTM-AE for multivariate time-series anomaly detection.
- [18] Generalitat Valenciana. Decreto 161/2003, del consell, por el que se establece la red valenciana de vigilancia y control de la contaminación atmosférica. Diario Oficial de la Generalitat Valenciana (DOGV), 2003.
- [19] Ajuntament de València. Contaminación atmosférica en valència. <https://www.valencia.es/cas/calidadaire/contaminacion-atmosferica>. Accedido el 17 de julio de 2025.
- [20] Generalitat Valenciana. Evaluación de la calidad del aire. zona es1007 – turia. Àrea costera. <https://rvvcca.pica.gva.es/sites/default/files/2024-06/ZONA%20ES1007.%20TURIA.%20%C3%80REA%20COSTERA.pdf>. Accedido el 17 de julio de 2025.
- [21] Generalitat Valenciana. Datos on-line — calidad del aire. <https://mediambiente.gva.es/es/web/calidad-ambiental/datos-on-line>. Accedido el 17 de julio de 2025.
- [22] European Environment Agency. Air quality in europe - 2023 report, 2023.
- [23] World Health Organization. *WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization, Geneva, 2021.
- [24] U.S. Environmental Protection Agency. Sulfur dioxide basics. <https://www.epa.gov/so2-pollution/sulfur-dioxide-basics>, n.d. Accedido el 17 de julio de 2025.
- [25] U.S. Environmental Protection Agency. Integrated science assessment (isa) for particulate matter, 2019.
- [26] Gobierno de España. Real decreto 102/2011, de 28 de enero, relativo a la mejora de la calidad del aire. <https://www.boe.es/buscar/act.php?id=BOE-A-2011-1645>, 2011. Accedido el 17 de julio de 2025.
- [27] I. T. Jolliffe and J. Cadima. *Principal Component Analysis: A Review and Recent Developments*. Philosophical Transactions of the Royal Society A, 2016. Accedido el 17 de julio de 2025.