



TRABAJO FIN DE GRADO

CALIBRACIÓN SENSORES DE BAJO COSTE

AUTOR: HUGO TOLEDO ESCRIVÁ

TUTOR: SANTIAGO FELICI CASTELL

Declaración de autoría:

Yo, Hugo Toledo Escrivá, declaro la autoría del Trabajo Fin de Grado titulado “CALIBRACIÓN SENSORES DE BAJO COSTE” 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, 27 de junio de 2025

Fdo: Hugo Toledo Escrivá

Resumen:

La contaminación atmosférica constituye un importante problema de salud pública a nivel global. En este contexto, el presente Trabajo de Fin de Grado se centra en el diseño, implementación y calibración de una red de sensores de bajo coste (LCS) para la monitorización de la calidad del aire (AQ), con el objetivo de aproximarse al nivel de precisión de las estaciones oficiales de referencia. Para ello, se ha desarrollado un prototipo compuesto por 10 módulos multisensor ZPHS01B, integrados en una arquitectura basada en un microcontrolador ESP32 y un sistema de comunicaciones LTE. Estos dispositivos permiten la recolección continua de datos relativos a diversos contaminantes y variables ambientales, con una frecuencia de muestreo de 30 segundos.

Se realizó un análisis exploratorio comparando los datos de los sensores con una estación oficial, evaluando su coherencia interna, estabilidad y proximidad a las medidas de referencia. Los resultados muestran que variables como PM_{2.5} pueden ser calibradas con alta precisión mediante modelos de regresión lineal, mientras que otras como O₃ o NO₂ requieren modelos más complejos. Para ello, se implementaron técnicas de aprendizaje supervisado, incluyendo XGBoost y modelos en dos etapas, que permitieron mejorar significativamente la precisión de las mediciones.

Abstract:

Air pollution constitutes a major public health issue on a global scale. In this context, the present Bachelor's Thesis focuses on the design, implementation, and calibration of a low-cost sensor (LCS) network for air quality (AQ) monitoring, with the objective of approaching the accuracy levels of official reference stations. To this end, a prototype has been developed consisting of 10 ZPHS01B multisensor modules, integrated into an architecture based on an ESP32 microcontroller and an LTE communication system. These devices enable continuous collection of data related to various pollutants and environmental variables, with a sampling frequency of 30 seconds.

An exploratory analysis was conducted to compare the sensor data with those from an official monitoring station, evaluating internal consistency, stability, and proximity to the reference measurements. The results show that variables such as PM_{2.5} can be calibrated with high accuracy using linear regression models, while others, such as O₃ and NO₂, require more complex models. To address this, supervised learning techniques were applied, including XGBoost and two-stage models, which significantly improved the measurement accuracy.

Resum:

La contaminació atmosfèrica constitueix un important problema de salut pública a escala global. En aquest context, el present Treball de Fi de Grau se centra en el disseny, implementació i calibratge d'una xarxa de sensors de baix cost (LCS) per a la monitorització de la qualitat de l'aire (AQ), amb l'objectiu d'aproximar-se al nivell de precisió de les estacions oficials de referència. Per a això, s'ha desenvolupat un prototip compost per 10 mòduls multisensor ZPHS01B, integrats en una arquitectura basada en un microcontrolador ESP32 i un sistema de comunicacions LTE. Aquests dispositius permeten la recollida contínua de dades relatives a diversos contaminants i variables ambientals, amb una freqüència de mostreig de 30 segons.

S'ha realitzat una anàlisi exploratòria comparant les dades dels sensors amb les d'una estació oficial,avaluant la seu coherència interna, estabilitat i proximitat respecte a les mesures de referència. Els resultats mostren que certes variables, com PM_{2.5}, poden ser calibrades amb alta precisió mitjançant models de regressió lineal, mentre que altres, com O₃ o NO₂, requereixen models més complexos. Per tal d'abordar aquesta qüestió, s'han aplicat tècniques d'aprenentatge supervisat, incloent XGBoost i models en dos etapes, que han permés millorar significativament la precisió de les mesures.

Agradecimientos:

En primer lugar, quiero agradecer a mis padres por su apoyo constante durante todos estos años. Su confianza y paciencia han sido fundamentales para que hoy pueda cerrar esta etapa.

Quiero agradecer de manera muy especial a mi pareja, que siempre ha creído en mí incluso cuando yo no lo hacía, y me ha apoyado en todo lo que hago sin dudar. Su presencia ha sido una parte clave de este proceso.

También me gustaría dar las gracias a mi tutor, por su disponibilidad y por ayudarme siempre que lo he necesitado. Su orientación ha hecho que el desarrollo de este trabajo fuera mucho más claro y llevadero.

Gracias a todos por estar ahí.

Índice general

1. Introducción	15
1.1. Introducción	15
1.2. Trabajos relacionados	16
1.3. Objetivos	17
2. Sensores de calidad del aire y alternativas de bajo coste	19
2.1. Descripción de los sensores del módulo multisensor ZPHS01B	20
2.2. Nodo de monitorización de calidad del aire con 10 módulos	23
3. Procesamiento de datos	25
3.1. Tipología de los datos	25
3.2. Obtención automatizada de datos oficiales	26
3.3. Preprocesado de datos	26
3.4. Conversión de unidades	27
4. Análisis exploratorio de datos	29
4.1. Análisis visual de las diferencias entre módulos mediante gráficas, utilizando la media diaria	29
4.2. Gráficas individuales de los 10 módulos del sensor de Low-Cost Sensor (LCS), comparando con el sensor oficial	30
4.3. Cálculo de correlaciones entre las variables dentro de cada módulo.	31
4.4. Comparación entre matrices de correlación por módulo	32
4.5. Evaluación de la proximidad de los módulos al sensor oficial mediante distancia euclidiana	33
4.6. Evaluación de la estabilidad de los módulos y variables mediante desviación estándar	34
4.7. Estudio de las correlaciones generales entre módulos	35
4.8. Conclusiones del análisis exploratorio	35
5. Análisis de deriva temporal en las mediciones de los sensores	37
5.1. Visualización inicial del error diario (módulo representativo)	37

5.2. Análisis de deriva por módulo mediante regresión lineal	39
5.3. Cálculo de la tendencia lineal del error por módulo	42
5.4. Conclusiones del análisis de deriva temporal	44
6. Modelado de calibración mediante aprendizaje supervisado	47
6.1. Preparación de los datos	48
6.2. Modelo base (baseline)	49
6.3. Resultados por variable	51
6.4. Modelos con mayor capacidad predictiva	64
6.5. Conclusiones del modelado supervisado	75
7. Conclusiones finales	77
7.1. Trabajo futuro	78
Apéndice: Fragmentos de código en R y Python	79
Bibliografía	88

Capítulo 1

Introducción

1.1. Introducción

Según la [World Health Organization \(WHO\)](#), el 99 % de la población mundial está expuesta a niveles de contaminación del aire que superan las directrices de seguridad recomendadas. Esta contaminación generalizada [1, 2] supone graves riesgos para la salud, como accidentes cerebrovasculares, enfermedades cardíacas y afecciones respiratorias [3].

Cuadro 1.1: Niveles guía progresivos (paso a paso) para la calidad del aire por contaminante según [WHO](#)

Contaminante	Tiempo promedio	Paso-1	Paso-2	Paso-3	Paso-4	Objetivo
PM2.5 ($\mu\text{g}/\text{m}^3$)	Anual	35	25	15	10	5
	24 horas ^a	75	50	37.5	25	15
PM10 ($\mu\text{g}/\text{m}^3$)	Anual	70	50	30	20	15
	24 horas ^a	150	100	75	50	45
O_3 ($\mu\text{g}/\text{m}^3$)	Temporada alta ^b	100	70	—	—	60
	8 horas ^a	160	120	—	—	100
NO_2 ($\mu\text{g}/\text{m}^3$)	Anual	40	30	20	—	10
	24 horas ^a	120	50	—	—	25
SO_2 ($\mu\text{g}/\text{m}^3$)	24 horas ^a	125	50	—	—	40
CO (mg/m^3)	24 horas ^a	7	—	—	—	4

^a Promedio móvil.

^b Temporada de mayor concentración del contaminante.

Para evitar estos escenarios, la [WHO](#) establece una serie de directrices, incluyendo límites máximos permisibles de contaminantes con el fin de proteger la salud pública [4], como se observa en la Tabla 1.1, donde se recogen la guía de rutas de los diferentes contaminantes de calidad del aire, paso a paso, con el plan de alcanzar el nivel objetivo. En dicha tabla se recogen los valores en pasos intermedios hasta alcanzar el objetivo, para [Particulate Matter \(PM\)](#) o partículas en suspensión de diferente tamaños medidas (1, 2,5 y 10 μm), ozono terrestre (O_3), dióxido de nitrógeno (NO_2), monóxido de azufre (SO_2) y monóxido de carbono (CO).

En esta línea, por ejemplo, la Directiva 2008/50/EC del Parlamento Europeo [5] establece la infraestructura de monitorización de la [Air Quality \(AQ\)](#) necesaria. Esta directiva constituye un pilar fundamental en el tratamiento de la calidad del aire ambiente en Europa y se considera una referencia a nivel global. En particular, dicha directiva obliga a la instalación de estaciones de monitorización de [AQ](#). Así, cada zona o aglomeración debe contar con al menos un punto de muestreo por cada 2 millones de habitantes, o bien un punto por cada 50.000 km^2 , eligiendo el criterio que resulte en un mayor número de puntos de muestreo, sin que en ningún caso se permita menos de un punto por zona o aglomeración.

No obstante, la recomendación va más allá, proponiendo como objetivo aumentar la resolución espacial del muestreo, idealmente alcanzando al menos una muestra por cada 100 m^2 , según se indica en el Anexo III-B de dicha directiva.

En este contexto, los [LCS](#) de [AQ](#) están cobrando cada vez mayor importancia como una alternativa interesante para lograr este objetivo de resolución en el muestreo. Sin embargo, estos sensores por sí solos no ofrecen una buena precisión [6, 7, 8], y aún menos como para ser considerados dispositivos de referencia.

No obstante, la combinación de estos sensores con tecnologías como la [Artificial Intelligence \(AI\)](#) (mediante técnicas de [Machine Learning \(ML\)](#) y redes [Neural Network \(NN\)](#)), junto con regresión lineal múltiple ([Multiple Linear Regression \(MLR\)](#)), redes de sensores inalámbricos ([Wireless Sensor Networks \(WSN\)](#)) y sistemas de monitorización de [AQ](#) habilitados por [Internet of Things \(IoT\)](#), en combinación con las comunicaciones avanzadas 5G/6G, abre nuevas oportunidades para desplegar estas redes extensivas de monitorización de la calidad del aire.

1.2. Trabajos relacionados

El creciente interés por la monitorización de la calidad del aire ha impulsado el desarrollo y la aplicación de sensores de bajo coste como solución complementaria para ampliar la cobertura espacial de las redes oficiales. Estos dispositivos, aunque accesibles económicamente, presentan limitaciones en términos de estabilidad, sensibilidad y precisión, lo que ha motivado numerosos estudios sobre su calibración y validación.

Diversos trabajos han propuesto técnicas para mejorar la fiabilidad de los sensores mediante modelos de calibración avanzados. Por ejemplo, Malings et al. [9] desarrollaron un modelo de calibración generalizado basado en datos procedentes de casi 70 sensores RAMP distribuidos por la ciudad de Pittsburgh. Su estudio comparó distintos algoritmos, como regresión cuadrática, redes neuronales y modelos híbridos (random forest y regresión lineal), recomendando enfoques distintos según el contaminante. Además, demostraron que los modelos generalizados ofrecen buen rendimiento y permiten reducir el número de sensores que requieren calibración individual, facilitando despliegues a gran escala.

Por otro lado, Castell et al. [10] evaluaron experimentalmente sensores comerciales en entornos reales y constataron importantes diferencias en la calidad de los datos según el contaminante y las condiciones meteorológicas, reforzando la necesidad de calibraciones específicas *in situ*. En la misma línea, Popoola et al. [11] analizaron la sensibilidad cruzada de los sensores y su dependencia respecto a variables como temperatura y humedad, proponiendo modelos multivariantes para compensar dichos efectos.

En cuanto al uso de técnicas de aprendizaje automático en entornos ambientales, Zhu

et al. [12] presentan una revisión sistemática de errores comunes y buenas prácticas en la aplicación de modelos supervisados. Entre otros aspectos, destacan la importancia de evitar el *data leakage*, utilizar conjuntos de entrenamiento y prueba correctamente separados, realizar una adecuada selección de variables y aplicar optimización de hiperparámetros. Estas prácticas son clave para garantizar la validez de los resultados y evitar interpretaciones erróneas en estudios de predicción ambiental.

Finalmente, distintas agencias como la Agencia Europea del Medio Ambiente recomiendan integrar sensores de bajo coste con estaciones de referencia, aplicar procedimientos de validación robustos y utilizar métricas normalizadas para evaluar el rendimiento [13]. Estas recomendaciones se recogen en guías de buenas prácticas que contribuyen a consolidar metodologías estandarizadas en la monitorización ambiental con sensores emergentes.

1.3. Objetivos

El objetivo principal de este Trabajo de Fin de Grado es desarrollar y validar una red de sensores de bajo coste (**LCS**) para la monitorización de la calidad del aire, con el fin de evaluar su viabilidad como herramienta complementaria a las estaciones oficiales de referencia. Para alcanzar este propósito general, se plantean los siguientes objetivos específicos:

- Tratar información útil a partir de una red funcional de sensores multiparamétricos de bajo coste, equipada con un sistema de adquisición y transmisión de datos en tiempo real mediante microcontroladores y tecnología **Long Term Evolution (LTE)**.
- Llevar a cabo un análisis exploratorio de los datos obtenidos, evaluando la coherencia interna y la estabilidad de los sensores a lo largo del tiempo.
- Evaluar las prestaciones de los sensores de bajo coste equipados en el módulo multisensor ZPHS01B, así como la posible dispersión y deriva entre módulos iguales.
- Calibrar las mediciones de los sensores frente a una estación de referencia oficial mediante técnicas estadísticas y modelos de aprendizaje automático.
- Comparar el rendimiento de distintos enfoques de calibración, incluyendo modelos lineales, no lineales y modelos supervisados avanzados como XGBoost.
- Evaluar la capacidad predictiva de los modelos y la transferencia de calibraciones entre sensores, así como la aplicabilidad de modelos generalizados.

Capítulo 2

Sensores de calidad del aire y alternativas de bajo coste

Debido al creciente interés del mercado en los [LCS](#) para la monitorización de la [AQ](#), actualmente existe una amplia variedad de sensores disponibles para la medición de diferentes contaminantes, tanto gaseosos como particulados. Estos dispositivos presentan rangos de precio diversos, pero resultan más económicos que las estaciones oficiales reguladas de monitorización de [AQ](#).

Para examinar las principales características y propiedades de estos [LCS](#), es importante entender sus procesos de fabricación, los cuales se basan en distintas tecnologías, como se describe a continuación. Los sensores [Metal OXide \(MOX\)](#) miden cambios en la conductividad eléctrica de un semiconductor debido a la presencia de ciertos gases, que reaccionan con oxígeno activo mediante procesos de reducción u oxidación. Son sensores económicos, pero presentan alta sensibilidad cruzada y son afectados por la humedad.

Los sensores [Electro-Chemical \(ECH\)](#) miden corrientes electrónicas generadas por reacciones químicas, las cuales indican la presencia y concentración de gases específicos. Ofrecen mayor selectividad y precisión, aunque son más caros y tienen una vida útil más corta.

Por último, los sensores y/o ópticos miden la cantidad de luz absorbida o dispersada por un contaminante en una longitud de onda específica, convirtiendo dicha medida en concentración mísica. Estos sensores son altamente sensibles a la [Humedad relativa \(RH\)](#), ofrecen buena selectividad y precisión, y presentan una vida útil más prolongada [14].

Adicionalmente, existen otras técnicas como los dispositivos [Micro-Electro-Mechanical Systems \(MEMS\)](#), que integran componentes electrónicos y partes móviles en escala microscópica, basados en semiconductores y/o combinando las tecnologías mencionadas anteriormente.

Es importante señalar que todos estos sensores de bajo coste comparten una característica común: presentan problemas de sensibilidad cruzada, es decir, pueden reaccionar a sustancias distintas de las que están diseñados para medir.

En nuestro caso de estudio, dado que se pretende realizar una monitorización de [AQ](#) basada en los diferentes contaminantes definidos por las [WHO AQ Guidelines \(AQG\)](#) [4], nos centramos en la creación de un nodo [AQ IoT](#) con sensores [LCS](#), empleando un módulo sensor tipo [Original Equipment Manufacturer \(OEM\)](#), concretamente el módulo multsensor ZPHS01B. Este módulo ha sido seleccionado porque integra el mayor número de

sensores y representa la mejor opción en términos de relación calidad-precio, en comparación con otras alternativas disponibles.

Este módulo mide [Temperatura \(T\)](#) ($^{\circ}\text{C}$) y [RH](#) (%), y en unidades de [Parts Per Million \(ppm\)](#): CO, CO₂, NO₂, O₃; en $\mu\text{g}/\text{m}^3$: [PM](#); en mg/m^3 : formaldehído (CH₂O); y compuestos orgánicos volátiles totales ([Total Volatile Organic Compounds \(TVOC\)](#)), los cuales se detectan en 4 niveles de concentración (bajo, medio, alto y muy alto), con valores representados como 0, 1, 2 y 3 respectivamente.

2.1. Descripción de los sensores del módulo multisensor ZPHS01B



Figura 2.1: Detalle del módulo ZPHS01B y sus sensores de AQ.

En la Figura 2.1 se muestra el módulo de bajo coste ZPHS01B, junto con el detalle de los distintos sensores integrados. Esta placa sensora mide directamente CO₂, [PM2.5](#), CH₂O, O₃, CO, [TVOC](#), NO₂, [T](#) y [RH](#). Además, el módulo ZPHS01B calcula las concentraciones de [PM1.0](#) y [PM10](#) a partir de la medición de [PM2.5](#). Esta característica implica que existe una alta correlación entre los valores de [PM](#) en este módulo.

Cuadro 2.1: Información de los sensores del módulo ZPHS01B: contaminante, nombre del sensor, tipo, unidades, rango y precisión.

Contaminante	Nombre	Tipo	Unidades	Rango	Precisión
O ₃	ZE27	ECH	ppm	0–10	0.01
NO ₂	GM-102B	MOX	ppm	0.1–10	0.05
CO	ZE15	ECH	ppm	0–500	0.1
PM2.5	ZH06-II	Óptico	$\mu\text{g}/\text{m}^3$	0–1000	$\pm 15 < 100, \pm 15 \% > 100$
CO ₂	MH-Z19C		ppm	0–5000	± 500
CH ₂ O	ZE08K	ECH	mg/m^3	0–6.25	$\pm 0,03 < 0,2, \pm 20 \% > 0,2$
TVOC	ZP07-MP503	MOX	Niveles	0–3	—
T	GXHT3X	Semicond.	$^{\circ}\text{C}$	[-20, 65]	$\pm 0,5$
RH	GXHT3X	Semicond.	%	0–100	± 3

En la Tabla 2.1 se detallan los sensores que componen el módulo ZPHS01B, mostrado en la Figura 2.1, junto con el contaminante que detectan, el nombre del sensor, su tipo,

unidades de medida, rango de detección y precisión. Tal como se mencionó anteriormente, aunque el módulo incluye 9 sensores físicos, proporciona un total de 11 mediciones, añadiendo PM1 y PM10 inferidas a partir de la concentración de PM2.5 [15].

A continuación, se presentan algunas características adicionales de interés sobre estos sensores.

2.1.1. O₃: sensor ZE27

El sensor Winsen ZE27 para O₃ [16] es un sensor ECH que utiliza el principio electroquímico para garantizar alta selectividad y estabilidad en la detección de ozono. Entre sus características destacan una excelente resolución, consumo energético extremadamente bajo, gran capacidad anti-interferencia, compensación de T y una salida lineal precisa. Presenta un tiempo de respuesta inferior a 90 segundos y un tiempo de calentamiento de 3 minutos.

Este módulo es sensible a interferencias, ya que el O₃ interactúa con otros gases como el NO₂ o el Cl₂. Las condiciones de funcionamiento indicadas por el fabricante para llevar a cabo mediciones son temperaturas entre -20ºC y 50ºC T y humedades relativas entre 15 % y 90 % RH, con una vida útil aproximada de 2 años.

2.1.2. NO₂: sensor GM-102B

El sensor Winsen GM-102B [17] para NO₂ es un sensor MOX, encapsulado con material cerámico dentro de un diseño tipo MEMS. Está fabricado sobre una base de silicio (Si) y utiliza un material semiconductor sensible a gases. En aire limpio, presenta baja conductividad, pero al exponerse al NO₂ su conductividad aumenta proporcionalmente a la concentración del gas.

Cuanto mayor es la concentración de gas, mayor es la conductividad. Mediante un circuito simple, el cambio de conductividad se transforma en una señal de salida proporcional. Sus principales características incluyen estructura robusta y tamaño compacto, alta sensibilidad al NO₂, bajo consumo, respuesta rápida, circuito de control sencillo y larga vida útil. Tras periodos prolongados sin alimentación, la resistencia interna puede variar; se recomienda un precalentamiento para restablecer el equilibrio químico.

2.1.3. CO: sensor ZE15

El sensor Winsen ZE15 [18] para CO es un sensor ECH que garantiza alta selectividad y estabilidad en la detección de monóxido de carbono. Incluye un sensor de T integrado para compensación automática, mejorando así la precisión de las mediciones.

Tiene un tiempo de respuesta inferior a 30 segundos y una vida útil entre 3 y 5 años. Antes de su primer uso, se recomienda dejar el sensor funcionando durante al menos 5 minutos para estabilizar la salida. Es importante destacar que el alcohol actúa como gas interferente. Las condiciones óptimas de funcionamiento son de -10ºC a 55ºC T y entre 15 % y 90 % de RH.

2.1.4. PM2.5: sensor ZH06-II

El sensor láser Winsen ZH06-II es un sensor de bajo coste que mide PM2.5. Presenta bajo consumo energético, alta precisión y capacidad para detectar partículas de hasta 0.3 μm de diámetro.

Su tiempo de respuesta es inferior a 45 segundos, con un intervalo de detección de 1 segundo. En condiciones normales de T y presión, el láser interno puede operar continuamente durante más de 10.000 horas, y su vida útil acumulada puede superar los 3 años.

La precisión del sensor, en condiciones normales de T y RH , depende de los valores medidos. Opera correctamente entre -10°C y 60°C y entre 0% y 95% de RH . Requiere una alimentación de 5V DC y un consumo de corriente de 120 mA, con una corriente en modo reposo inferior a 20 mA.

Tal y como se muestra en la Figura 2.1, este sensor dispone de un circuito interno de aire que genera un flujo desde el orificio de entrada hasta la salida, donde se encuentra un ventilador interno. Es importante que el orificio de entrada esté bien expuesto al aire ambiente, mientras que el ventilador, situado en la salida, extrae el aire del interior. Debe evitarse la presencia de corrientes fuertes de aire alrededor del sensor; si son inevitables, se recomienda que sean perpendiculares al flujo interno.

2.1.5. CO₂: sensor MH-Z19C

El sensor Winsen MH-Z19C [19] mide CO₂ mediante el principio . Ofrece alta selectividad, larga vida útil (5 años) y funciona de forma independiente al nivel de oxígeno, proporcionando alta sensibilidad, bajo consumo energético y excelente estabilidad. Cuenta con una cámara bañada en oro y un circuito interno de compensación de T , lo cual mejora su rendimiento. También presenta gran resistencia a la interferencia por vapor de agua y envenenamiento del sensor.

En cuanto a la calibración, este sensor dispone de dos métodos para determinar el valor base (offset) de 400 ppm. Se recomienda realizar la calibración en un entorno exterior y estable. Además, el sensor realiza una calibración automática cada 24 horas de funcionamiento. El fabricante sugiere un periodo de uso de 6 meses antes de realizar ajustes adicionales.

2.1.6. CH₂O: sensor ZE08K

El sensor Winsen ZE08K [20] es un sensor ECH para la medición de gas formaldehído (CH₂O). Cuanto mayor es la concentración del gas, mayor es la corriente generada en el electrodo activo.

Dispone de un sensor de T interno para la compensación automática, mejorando así la precisión. Entre sus ventajas destacan su alta sensibilidad, buena resolución, bajo consumo y larga vida útil (5 años). Sin embargo, presenta sensibilidad cruzada con alcohol, H₂S, CO y compuestos derivados de C_xH_x.

Tiene un tiempo de respuesta inferior a 1 minuto y un tiempo de calentamiento aproximado de 3 minutos. Las condiciones de funcionamiento recomendadas por el fabricante son entre 20°C y 50°C y entre 15% y 90% de RH .

2.1.7. TVOC: sensor ZP07-MP503

El sensor Winsen ZP07-MP503 [21] es un sensor **MOX** con alta sensibilidad a diversos compuestos orgánicos volátiles (TVOC), principalmente formaldehído (CH_2O), CO, H, alcoholes, benceno, amoníaco (NH_4) y humo de cigarrillo, entre otros. Sus características principales incluyen alta sensibilidad, bajo consumo energético y larga vida útil.

Tiene un tiempo de respuesta inferior a 20 segundos y una atenuación de sensibilidad menor al 1% anual. Las condiciones óptimas de funcionamiento están entre 0°C y 50°C, y entre 0% y 95% de **RH**.

2.1.8. T y RH: sensor GXHT30

Este es un sensor combinado de **T** y **RH** GXHT30 [22], que viene ya calibrado de fábrica. Su tiempo de arranque es inferior a 1 ms y el tiempo de medición es menor a 15 ms.

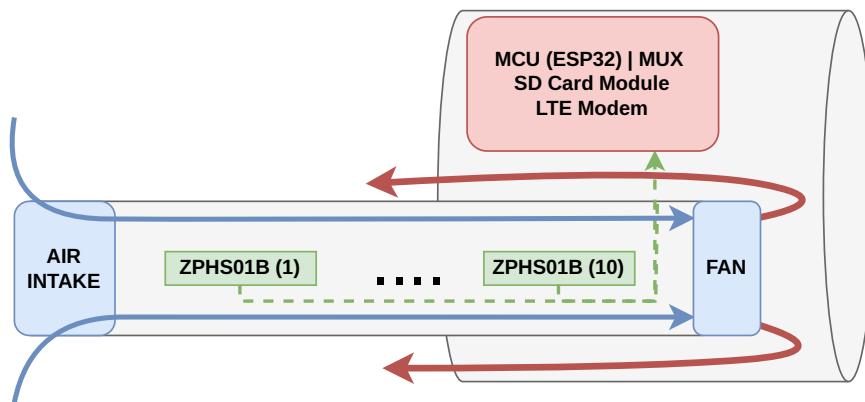
Ofrece su mejor rendimiento en condiciones de **T** entre 5°C y 60°C y en condiciones de **RH** entre 20% y 80%. Si el sensor permanece durante periodos prolongados fuera de estos rangos (especialmente con **RH** alta), puede producirse un desplazamiento del valor de humedad. No obstante, el sensor es capaz de autocorregirse gradualmente al retornar a condiciones normales.

2.2. Nodo de monitorización de calidad del aire con 10 módulos

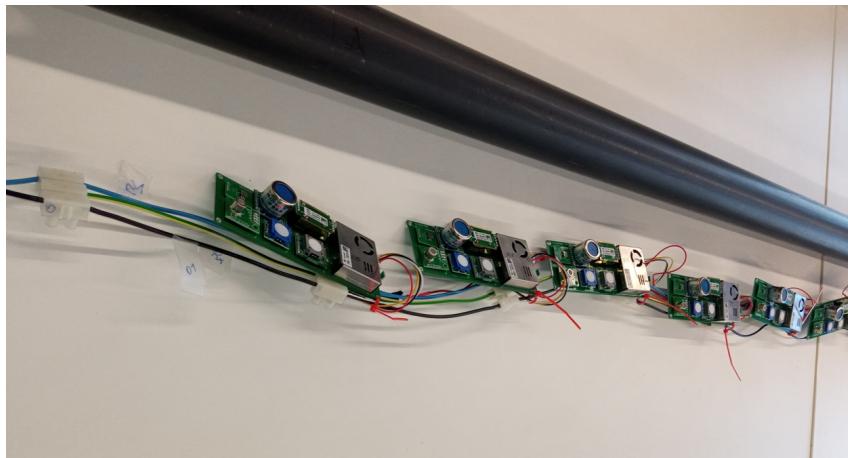
Se ha diseñado un nodo de monitorización de calidad del aire compuesto por 10 módulos ZPHS01B, siguiendo una estructura tubular, como se muestra en la Figura 2.2. En dicha figura se observa la estructura principal que aloja los diferentes módulos, acompañada de un flujo de aire forzado y estable generado por un pequeño ventilador situado en uno de los extremos.

Para enviar los datos recogidos por cada nodo, se emplea un microcontrolador y un módem, situados en la parte superior de la estructura dentro de una caja independiente. En la Figura 2.2a se representa el microcontrolador (**Micro Controller Unit (MCU)** basado en ESP32), la memoria (tarjeta SD) y la tecnología de comunicación utilizada (módem **LTE**, 4G), todos ellos protegidos frente a las condiciones ambientales mediante una carcasa de mayor diámetro que cubre los componentes electrónicos y sensibles de estos prototipos.

La Figura 2.2 muestra una imagen real del conjunto de los 10 módulos conectados.



(a) Esquema de múltiples módulos ZPHS01B con detalle del controlador y las comunicaciones



(b) Detalle del ensamblaje con varios módulos ZPHS01B

Figura 2.2: Esquema del prototipo con 10 módulos ZPHS01B.

Capítulo 3

Procesamiento de datos

3.1. Tipología de los datos

3.1.1. Sensor de bajo coste

Este sensor proporciona un total de 16 variables por módulo:

- **Temporales:** Fecha, Segundo, Minuto, Hora
- **Contaminantes:** PM1.0, PM2.5, PM10.0, TVOC, CH₂O, Monóxido de Carbono (CO), O₃, NO₂
- **Gases de efecto invernadero:** CO₂
- **Variables ambientales e informativas:** Temperatura, Humedad, Identificador del módulo (ID)

3.1.2. Sensor oficial

El sensor oficial incluye las siguientes variables:

- Fecha y bloque temporal de 10 minutos (ICA1)
- Contaminantes: NH₃, Benceno, Sulphur Dioxide (SO₂), NO₂, NO, NO_x, Ozono
- Partículas: PM1sc, PM2.5sc, PM10sc
- Variables meteorológicas: Dirección del viento (DV), Velocidad del viento (VV), Velocidad máxima del viento (VVMax), Temperatura (TEMP), Humedad relativa (HR), Presión barométrica (PRE), Precipitación (PLU), Radiación solar (RAD)
- Compuestos orgánicos volátiles: Tolueno, Xileno
- Otros: Sulfuro de hidrógeno (H₂S)
- Extras: TMPCab, HRCab, SO₂H₂S

3.2. Obtención automatizada de datos oficiales

Con el objetivo de enriquecer el proceso de adquisición de datos y asegurar la reproducibilidad del análisis, se ha incorporado una técnica de *web scraping* para la descarga automatizada de datos procedentes de la red oficial de calidad del aire. Mediante un script sencillo desarrollado en [Lenguaje de programación R \(R\)](#) [23], se accede directamente al repositorio público de la Generalitat Valenciana, extrayendo los archivos de datos en formato `.dat` correspondientes a un rango temporal especificado previamente.

El código implementado permite definir un intervalo de fechas, y a partir de él construye automáticamente las URLs diarias para descargar los archivos correspondientes. Posteriormente, los datos son leídos, limpiados y combinados en un único `data.frame`, alineado con las mediciones de los [LCS](#). Esta automatización facilita mantener actualizando el conjunto de datos oficial y permite futuras ampliaciones del estudio sin necesidad de intervención manual.

Para ello, se emplean funciones del paquete `tidyverse` [24], que permiten gestionar programáticamente el ciclo de descarga, lectura y depuración. Una vez finalizado el proceso, los archivos temporales se eliminan para evitar acumulación innecesaria en el almacenamiento local.

3.3. Preprocesado de datos

El nodo [LCS](#) está compuesto por 10 módulos que miden los mismos contaminantes y variables ambientales. Recolectan datos cada 30 segundos. Por otro lado, el sensor oficial proporciona mediciones cada 10 minutos. Para poder comparar ambos sistemas, se ha alineado temporalmente la información. Esto se ha realizado organizando cronológicamente las muestras del nodo [LCS](#) en bloques de 10 minutos y calculando la media de cada uno, respetando siempre los datos por módulo individual. Así, se asegura la misma frecuencia de muestreo en ambos sensores. A continuación, se eliminaron los días que contenían menos de 144 muestras (lo que equivale a 6 mediciones por hora durante 24 horas), así como aquellos que presentaban flags (errores) en los datos. Esta limpieza se aplicó a ambos sensores, manteniendo finalmente solo los días completos(145) y sin errores, con una muestra cada 10 minutos.

Además, es importante destacar que no todas las variables medidas por los [LCS](#) tienen equivalencia directa con las del sensor oficial. Solo un subconjunto limitado de variables puede ser comparado de forma objetiva entre ambos sistemas. Estas variables comunes, que constituyen la base para la calibración directa, son las siguientes:

- **PM2.5:** medición de partículas en suspensión con un diámetro aerodinámico inferior a $2.5 \mu\text{m}$. Es la única fracción particulada directamente comparable, ya que las variables PM1.0 y PM10.0 en el [LCS](#) son estimadas a partir de la medición de PM2.5, y no representan mediciones independientes.
- **O₃:** presente en ambos sistemas, permite evaluar la capacidad del [LCS](#) para detectar concentraciones de ozono, un contaminante clave en la calidad del aire urbano.
- **NO₂:** también disponible en ambos dispositivos, constituye otro contaminante de referencia que permite una comparación directa.

Adicionalmente, variables como la **T** y la **RH** se encuentran registradas en ambos sistemas, aunque no forman parte del conjunto principal de contaminantes a calibrar. Sin embargo, estas variables pueden desempeñar un papel relevante como entradas auxiliares en los modelos de calibración, ya que muchos **LCS** son sensibles a las condiciones ambientales y pueden requerir correcciones térmicas o higrométricas. Por último, aquellas variables que solo están presentes en uno de los sensores no permiten comparación directa, pero pueden contribuir como predictores adicionales en modelos multivariantes.

3.4. Conversión de unidades

Se realizaron las siguientes conversiones para armonizar las unidades entre los **LCS** y la estación oficial de referencia:

- **O₃, NO₂, CO y CO₂:** convertidos de **ppm** a microgramos por metro cúbico ($\mu\text{g}/\text{m}^3$), ajustando a las unidades utilizadas por la estación oficial.
- **TVOC:** recodificados en cuatro categorías discretas (bajo, medio, alto y muy alto) para facilitar su uso como variable cualitativa.
- **CH₂O:** transformado de miligramos por metro cúbico (mg/m^3) a microgramos por metro cúbico ($\mu\text{g}/\text{m}^3$), en línea con las unidades de la referencia oficial.

Capítulo 4

Análisis exploratorio de datos

4.1. Análisis visual de las diferencias entre módulos mediante gráficas, utilizando la media diaria

Al calcular la media diaria de cada contaminante para cada uno de los 10 módulos, se puede suavizar el ruido puntual y resaltar tendencias generales. Este análisis es clave para verificar si los errores de los sensores son aleatorios o si presentan un sesgo estructural.

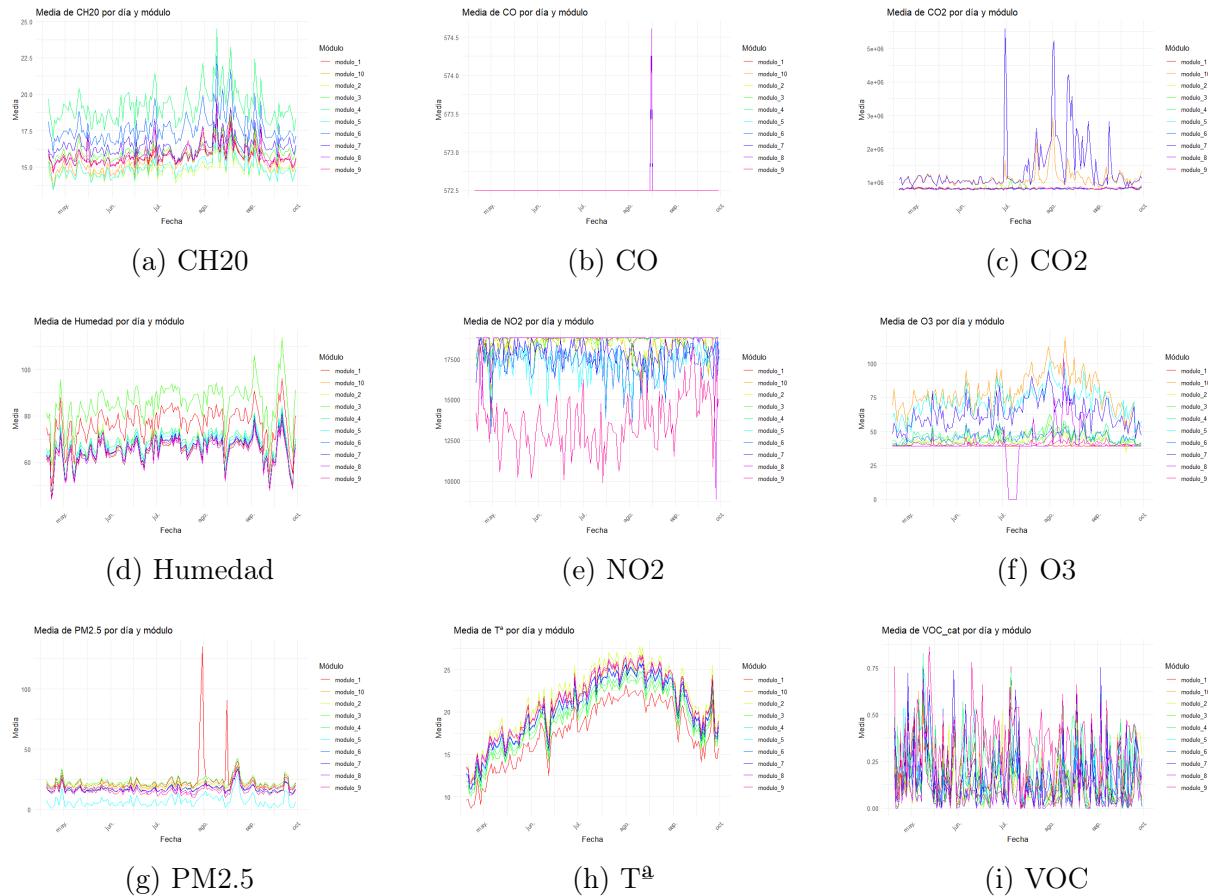


Figura 4.1: Comparación diaria de los 10 módulos para cada variable.

Como se puede observar en la Figura 4.1, existen variables que muestran un patrón muy similar entre los 10 módulos del sensor de LCS, lo cual indica coherencia en la medición.

Sin embargo, también se identifican variables que presentan diferencias significativas entre módulos o incluso fallos evidentes en la detección. A continuación, se resumen algunos casos destacados:

- **CO:** la mayoría de los módulos presentan un valor constante a lo largo del tiempo, con la excepción de un único pico. Esto sugiere una falta de sensibilidad o una calibración inadecuada del sensor correspondiente.
- **CO₂:** solo algunos módulos muestran respuesta en esta variable, mientras que otros parecen no registrar variación alguna. Esto podría indicar problemas de funcionamiento en ciertos módulos o diferencias en el ensamblaje del sensor.
- **Dióxido de Nitrógeno (NO₂):** presenta un comportamiento similar al del CO₂, con módulos que detectan correctamente y otros que ofrecen valores planos o inconsistentes.
- **Ozono (O₃):** aunque la mayoría de los módulos captan variaciones, se observa cierta dispersión en los valores registrados, aunque en menor medida que en las variables anteriores.

4.2. Gráficas individuales de los 10 módulos del sensor de **LCS**, comparando con el sensor oficial

Estas visualizaciones permiten detectar el grado de similitud entre las mediciones de los sensores de **LCS** y el sensor oficial. Al representar variables como NO₂, O₃ o PM2.5 a lo largo del tiempo, se observa la respuesta dinámica de cada módulo ante cambios reales en la atmósfera. Esto permite identificar errores sistemáticos, desfases temporales o comportamientos anómalos.

Dado que los datos tienen alta resolución temporal, se ha optado por utilizar la media diaria por módulo, lo que mejora la claridad visual al reducir el ruido. Así, se obtienen comparaciones más representativas entre cada sensor de **LCS** y el sensor oficial.

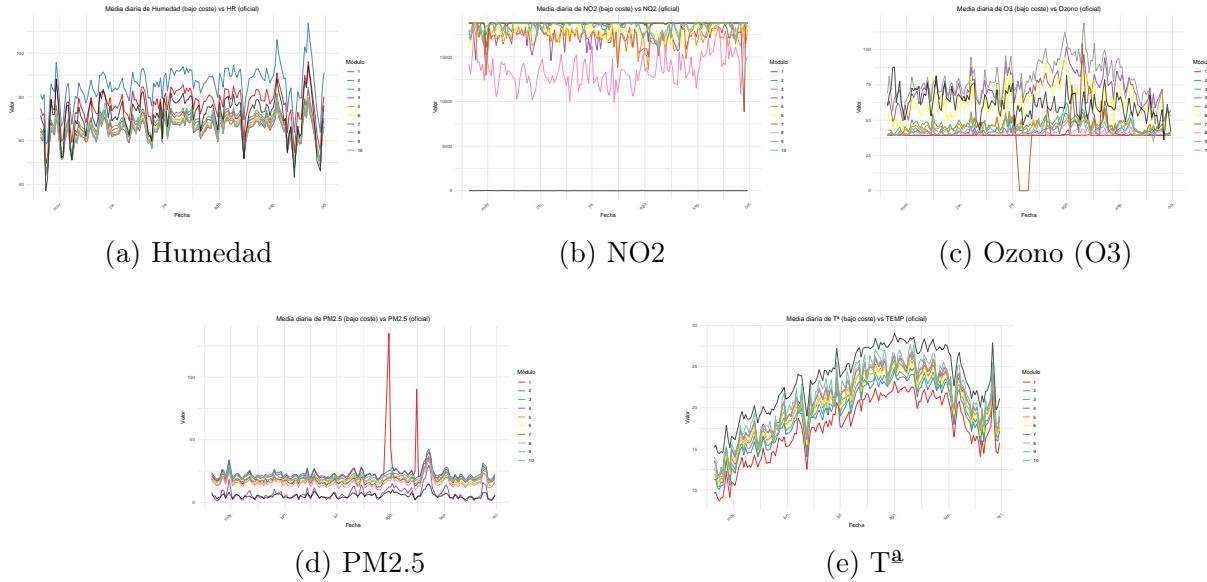


Figura 4.2: Comparación diaria de las variables coincidentes con el sensor oficial (línea negra).

Como se aprecia en la Figura 4.2, el comportamiento de los módulos del sensor de LCS varía significativamente según la variable considerada. En general, se observa una buena correspondencia con el sensor oficial en variables como la T y la RH, donde las curvas de los módulos siguen patrones similares y mantienen coherencia interna.

En el caso del PM2.5, aunque los valores se mantienen razonablemente alineados, destacan algunas discrepancias puntuales —especialmente picos extremos en el sensor de LCS que no se reproducen en el sensor oficial—, lo que podría deberse a diferencias en sensibilidad o a un comportamiento más reactivo ante ciertas condiciones ambientales.

El O₃ muestra una mayor dispersión entre módulos, con algunos siguiendo correctamente la tendencia del sensor oficial y otros registrando valores constantes o atípicos, como la anomalía detectada en el módulo 1, que presenta valores negativos o planos.

Finalmente, el NO₂ presenta una escala completamente desalineada: los sensores de bajo coste registran valores considerablemente superiores al oficial o incluso erráticos. Esto sugiere una falta de calibración efectiva o una interferencia sistemática. Es, por tanto, la variable con peor correspondencia entre ambos sistemas, y podría considerarse no fiable para tareas de calibración directa.

4.3. Cálculo de correlaciones entre las variables dentro de cada módulo.

Se calculó la matriz de Coeficiente de correlación de Pearson (Pearson) para cada módulo del sensor de LCS. Este análisis permite identificar relaciones esperadas, como la alta correlación entre PM1.0, PM10.0 y PM2.5, debido a que las dos primeras se estiman a partir de la tercera. También permite detectar sensores sin respuesta o con comportamiento anómalo, lo cual resulta útil para descartar variables redundantes o defectuosas en etapas posteriores de calibración.

La correlación de Pearson [25] cuantifica la relación lineal entre dos variables numéri-

cas, y se expresa mediante el coeficiente r , definido como:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

donde x_i y y_i son los valores de las dos variables, y \bar{x} y \bar{y} sus medias. El valor de r oscila entre -1 (correlación negativa perfecta) y 1 (correlación positiva perfecta), siendo 0 indicativo de ausencia de correlación lineal. En este caso, se ha usado como herramienta para detectar relaciones anómalas o esperadas entre sensores y variables.

Se ha observado que los módulos $1, 2, 3, 4$ y 8 presentan una desviación estándar igual a cero en la variable **CO**, lo que indica que no registran variaciones a lo largo del tiempo. Esto sugiere que dichos sensores no están operativos o no ofrecen datos válidos, ya que el valor constante genera **NA** en los cálculos de correlación.

En los módulos 3 y 4 se ha detectado una correlación moderada ($r \approx 0,5 - 0,6$) entre **CO₂** y **RH**, mientras que en el módulo 4 esta relación se extiende también a las variables **PM**. Aunque estos valores no representan una correlación fuerte, sí indican una posible interferencia o sensibilidad cruzada del sensor. En el caso de sensores de **LCS**, este tipo de relaciones puede deberse tanto a condiciones ambientales compartidas como a limitaciones en el diseño del sensor. Este comportamiento debe tenerse en cuenta durante el proceso de calibración, especialmente si estas variables se utilizan como predictores auxiliares.

4.4. Comparación entre matrices de correlación por módulo

Para evaluar la similitud estructural entre los sensores, se ha comparado la matriz de correlación de variables internas (**PM**, **CO**, **O₃**, **NO₂**, etc.) de cada módulo con la del resto de módulos. Este análisis permite identificar módulos cuya estructura de relaciones entre variables difiere significativamente, lo que puede ser indicativo de mal funcionamiento, desviaciones sistemáticas o sensores defectuosos.

4.4.1. Fundamento matemático

Dado un conjunto de M módulos, cada uno con su propia matriz de correlación $C^{(i)}$ de tamaño $n \times n$ (siendo n el número de variables), se calcula la **diferencia media absoluta** entre pares de matrices:

$$D_{ij} = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n |C_{kl}^{(i)} - C_{kl}^{(j)}|$$

Donde D_{ij} representa la diferencia media entre el módulo i y el módulo j . Para manejar valores faltantes (**NA**), las diferencias se imputan como cero antes del cómputo, y se ignoran en el promedio si es necesario (**na.rm = TRUE** en **R**).

El resultado es una matriz simétrica de diferencias promedio que permite ordenar los pares de módulos según su similitud o disimilitud estructural.

4.4.2. Resumen de resultados

Tras aplicar esta metodología, se observaron los siguientes hallazgos:

- **Módulo 8** es el que más se aleja del comportamiento del resto. Aparece en múltiples pares con diferencia media superior al 12 %. Esto indica una clara desviación estructural en su patrón de correlación.
- Por el contrario, los pares `modulo_2 - modulo_10`, `modulo_3 - modulo_5` y `modulo_3 - modulo_4` presentan diferencias menores al 5 %, lo que refleja una alta consistencia en la estructura interna de sus mediciones.
- Estas diferencias estructurales permiten agrupar módulos según su similitud y descartar aquellos cuya matriz de correlación sea significativamente distinta, como el módulo 8, que podría comprometer la calidad del proceso de calibración.

4.5. Evaluación de la proximidad de los módulos al sensor oficial mediante distancia euclíadiana

Con el objetivo de identificar qué módulos del sensor de **LCS** replican mejor los datos del sensor oficial, se ha calculado la **distancia euclíadiana** entre las mediciones de cada módulo y los valores del sensor de referencia para las variables comunes. Este análisis cuantifica la cercanía de cada módulo respecto a la referencia, permitiendo clasificar su comportamiento en términos de fidelidad.

4.5.1. Fundamento matemático

Dado un conjunto de observaciones temporales $x_t^{(m)}$ correspondientes al módulo m y $x_t^{(r)}$ del sensor de referencia, se calcula la distancia euclíadiana acumulada para cada variable como:

$$D_m = \sum_{t=1}^n \sqrt{(x_t^{(m)} - x_t^{(r)})^2}$$

Este cálculo se aplica por separado para cada variable (**PM1.0**, **PM2.5**, **PM10.0**, **O₃**, **NO₂**, **T**) y por cada módulo.

4.5.2. Resumen de resultados

A partir de la distancia total calculada para cada variable, se seleccionaron los tres módulos más cercanos al sensor oficial. Los resultados obtenidos fueron:

- **Temperatura (T^a)**: módulos 10, 7, 8.
- **PM1.0 y PM2.5**: módulos 4, 8, 6.
- **PM10.0**: módulos 5, 8, 4.

- O_3 : módulos 6, 3, 4.
- NO_2 : módulos 8, 5, 4.

Se observa que los módulos **8** y **4** aparecen repetidamente entre los tres más cercanos en distintas variables. Particularmente llamativo es el caso del módulo 8, que, pese a diferir estructuralmente en su matriz de correlación, aparece en cinco de las seis variables como uno de los más próximos al sensor oficial.

Esto sugiere que una diferencia estructural interna no implica necesariamente una mala calibración respecto al sensor oficial. Por el contrario, puede indicar una estructura distinta pero eficaz. En cambio, el módulo 4, aunque más estructuralmente coherente, presenta limitaciones como falta de variabilidad en la medición de CO . El módulo 6, aunque menos destacado, también demuestra un buen rendimiento global.

4.6. Evaluación de la estabilidad de los módulos y variables mediante desviación estándar

A diferencia del análisis anterior, que evaluaba la proximidad media al sensor oficial, aquí se analiza la **estabilidad temporal**. Para ello se calcula la **desviación estándar** de la distancia euclíadiana diaria entre cada módulo y el sensor oficial, para cada variable.

4.6.1. Fundamento matemático

Para una variable dada, se calcula para cada módulo:

$$\sigma = \sqrt{\frac{1}{n} \sum_{t=1}^n \left(d_t^{(m)} - \bar{d}^{(m)} \right)^2}$$

donde $d_t^{(m)}$ es la distancia entre el módulo m y el sensor oficial en el día t , y $\bar{d}^{(m)}$ es la media de dichas distancias.

4.6.2. Resumen de resultados

- Los **módulos más estables**, es decir, con menor desviación estándar global, fueron los módulos **6, 5 y 3**.
- En cuanto a las variables, las que mostraron menor variabilidad relativa fueron:
 1. Temperatura: $\sigma \approx 1,22$
 2. PM1.0 – PM2.5 – PM10.0: $\sigma \approx 5\text{--}5,8$
- Las variables O_3 ($\sigma \approx 10$) y especialmente NO_2 ($\sigma \approx 1732$) mostraron alta inestabilidad. En el caso de NO_2 , se propone su exclusión del proceso de calibración al no mostrar coherencia con el sensor oficial.

4.7. Estudio de las correlaciones generales entre módulos

Además de la comparación con el sensor oficial, se evaluó la **coherencia interna** entre módulos del sensor de [LCS](#). Una red fiable debería mostrar comportamientos similares entre módulos. Se calculó la media de cada variable por módulo y se compararon entre sí.

Los resultados reflejan una alta correlación general, con coeficientes cercanos a 1. Esto indica que, salvo casos puntuales, los módulos miden de forma homogénea, lo que valida su uso en red para análisis agregados.

4.8. Conclusiones del análisis exploratorio

El análisis exploratorio ha permitido evaluar en profundidad el comportamiento de los sensores de [LCS](#) frente al sensor oficial. Se analizaron la coherencia interna entre módulos, la estructura de correlaciones, la proximidad a la referencia y la estabilidad temporal.

Se identificaron variables con buen desempeño (como [T](#) o [PM2.5](#)) y otras más problemáticas, especialmente [NO₂](#). También se observó que un sensor puede presentar diferencias estructurales pero ofrecer buenos resultados en calibración, como ocurrió con el módulo 8.

Como resultado, se propone el siguiente ranking de módulos más fiables: **módulo 6**, seguido por los **módulos 5, 3 y 8**. Además, se define un **módulo mixto ideal**, construido virtualmente a partir de las variables mejor comportadas por módulo, cuya composición se detalla en la Tabla 4.1.

Cuadro 4.1: Composición del módulo mixto con las mejores variables por módulo.

Variable	Módulo seleccionado
Temperatura (T ^a)	Módulo 2
PM1.0	Módulo 6
PM2.5	Módulo 6
PM10.0	Módulo 5
O ₃	Módulo 7
NO ₂	Módulo 9
CO	Módulo 8
CO ₂	Módulo 8
CH ₂ O	Módulo 4
VOC	Módulo 7
Humedad	Módulo 3

Cuadro 4.2: Evaluación cuantitativa de los módulos: desviación estándar por variable y fiabilidad total

Módulo	T ^a	PM1.0	PM2.5	PM10.0	O ₃	Fiabilidad total
6	0.459	0.891	1.340	3.120	7.799	2.74
5	0.432	0.965	1.423	3.268	8.493	2.92
3	0.353	1.080	1.538	3.450	8.529	2.97
8	0.561	1.584	2.111	3.336	8.583	3.23
4	0.343	2.054	2.511	3.778	8.250	3.39
10	1.020	1.566	2.014	3.908	8.613	3.42
7	0.337	1.427	2.174	4.341	9.136	3.48
9	0.684	0.928	1.534	3.612	11.976	3.72
2	0.484	1.892	2.513	4.101	9.886	4.07
1	0.372	9.617	11.058	12.147	8.747	8.39

Valores expresados en unidades originales de cada variable. La fiabilidad total corresponde a una media ponderada de las desviaciones.

Cuadro 4.3: Resumen del comportamiento de los módulos por variable y criterio

Variable	Proximidad ref.	Estabilidad	Coherencia interna	Módulo mixto
T ^a	M10, M7, M8	Sí	Sí	M2
PM1.0	M4, M8, M6	Sí	Sí	M6
PM2.5	M4, M8, M6	Sí	Sí	M6
PM10.0	M5, M8, M4	Sí	Sí	M5
O ₃	M6, M3, M4	No	Medio	M7
NO ₂	M8, M5, M4	Muy bajo	No	M9
CO	—	No	No	M8
CO ₂	—	NE	Medio	M8
CH ₂ O	—	NE	Sí	M4
VOC	—	NE	Sí	M7
Humedad	—	Sí	Sí	M3

Leyenda: NE = No Evaluado, Medio = comportamiento intermedio, Muy bajo = alta inestabilidad o inconsistencia.

Capítulo 5

Análisis de deriva temporal en las mediciones de los sensores

El objetivo de esta sección es analizar la presencia de *drift* o deriva temporal en las mediciones de los sensores de bajo coste, en particular para las variables PM2.5, NO₂ y O₃, que son las únicas que cuentan con una referencia oficial válida para su posterior calibración. El análisis se realiza exclusivamente desde una perspectiva estructural y exploratoria, sin emplear técnicas de predicción ni aprendizaje supervisado.

Para ello, se calcula el error diario entre la medición del sensor de LCS y la medición oficial. Esta serie de errores se analiza mediante visualización directa, diagramas de caja y modelos de regresión lineal. Estas herramientas permiten detectar si el error presenta una evolución sistemática a lo largo del tiempo, lo que evidenciaría una pérdida progresiva de precisión o una sensibilidad dependiente de la estacionalidad.

5.1. Visualización inicial del error diario (módulo representativo)

Como primer acercamiento al comportamiento del error, se selecciona un único módulo representativo: el módulo 5. Este módulo fue identificado en el análisis exploratorio como uno de los más estables y fiables, por lo que resulta adecuado como referencia inicial para explorar la posible existencia de deriva.

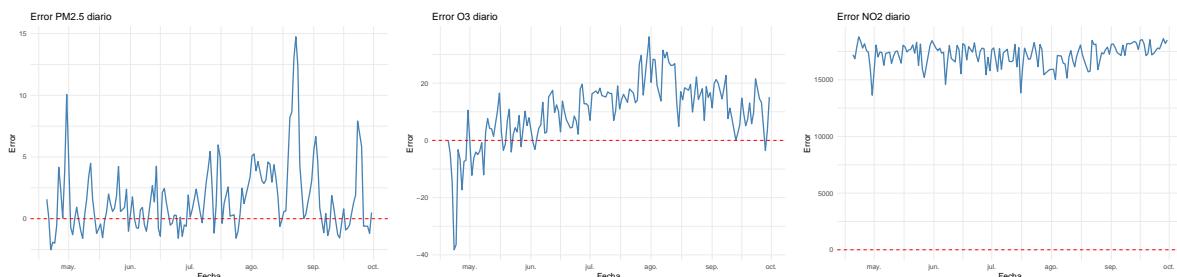


Figura 5.1: Evolución diaria del error para PM2.5 (izquierda), O₃ (centro) y NO₂ (derecha) en el módulo 5.

Tal y como se aprecia en la Figura 5.1, que muestra la evolución diaria del error

de medición para PM2.5, O₃ y NO₂ entre mediados de abril y finales de septiembre, se observan distintos patrones de comportamiento.

En el caso de PM2.5, el error oscila inicialmente en un rango moderado, con valores entre aproximadamente -10 y +10 . Sin embargo, a medida que avanza el periodo de observación, se aprecia un aumento claro en la dispersión de los errores, alcanzando valores elevados en los meses de verano. Aunque no se identifica una deriva sistemática en el valor medio, esta ampliación de la banda de error indica un comportamiento no estacionario. La creciente inestabilidad puede deberse a factores como el envejecimiento del sensor, la acumulación de contaminantes en el sistema de ventilación o la influencia de condiciones ambientales más extremas durante el verano.

En contraste, el error asociado a la medición de O₃ muestra una evolución temporal más estructurada. Se observa una transición desde errores negativos en primavera a positivos en verano, con un punto de inflexión a mitad del periodo. El error alcanza máximos en torno a julio y agosto, antes de disminuir ligeramente hacia otoño. Este patrón sugiere una deriva estacional, probablemente vinculada a la mayor radiación solar y temperaturas más altas en los meses centrales, que afectan tanto a la concentración real de ozono como a la respuesta del sensor. La presencia de este ciclo refuerza la hipótesis de una sensibilidad cruzada del sensor de O₃ a factores ambientales, lo que lo convierte en un candidato claro para calibración asistida por variables meteorológicas.

El caso de NO₂ resulta anómalo en comparación con las otras dos variables. El error diario se mantiene en un rango extremadamente elevado, entre 0 y 15 000 , lo que representa una desviación desproporcionada respecto a los valores reales esperables para este contaminante. Además, no se detecta estructura temporal aparente: la serie es casi plana, sin estacionalidad ni fluctuaciones apreciables. Este comportamiento sugiere que el sensor de NO₂ no está midiendo adecuadamente o que opera completamente fuera de su rango funcional. Es probable que el sensor esté captando ruido o señales erróneas, lo cual invalida cualquier uso de sus mediciones sin una recalibración profunda o directamente sin sustitución del hardware.

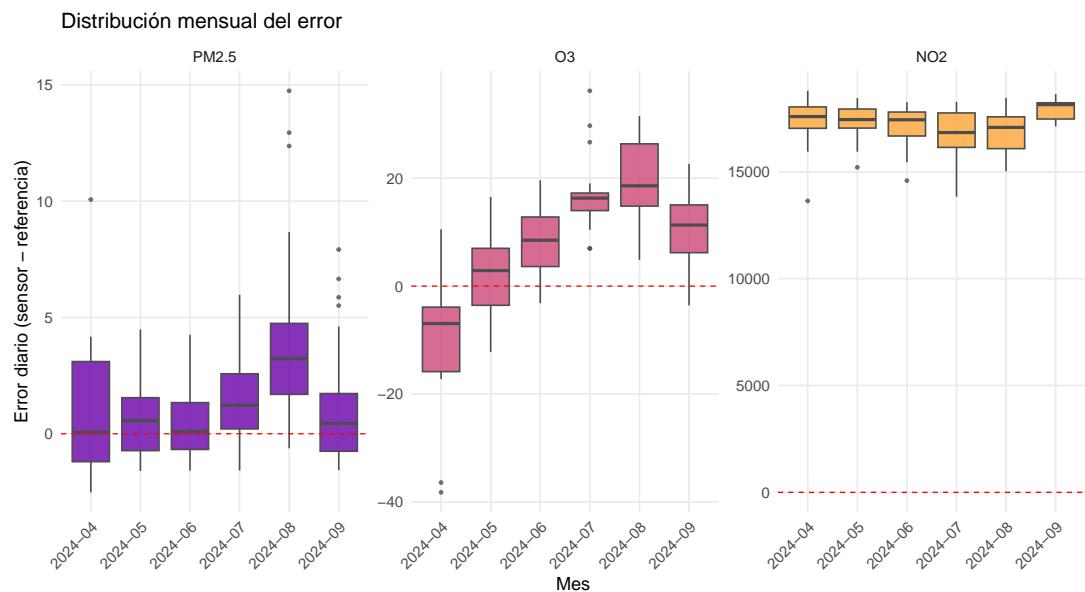


Figura 5.2: Distribución mensual del error diario en el módulo 5 para PM2.5, O₃ y NO₂. La línea discontinua roja representa el valor cero (sin error).

La figura 5.2 muestra la distribución mensual del error diario para cada uno de los contaminantes. En el caso de PM2.5, se aprecia una mayor variabilidad en los meses cálidos, lo que indica un incremento en la inestabilidad del sensor. Para O₃, la dispersión se intensifica también durante el verano, lo que refuerza la hipótesis de deriva estacional. En cambio, los valores de NO₂ se mantienen anormalmente altos y con escasa variabilidad en todos los meses, lo cual confirma la falta de respuesta del sensor y su desconexión respecto a las concentraciones reales.

Nota sobre la agregación diaria: Dado que el sensor de bajo coste realiza múltiples mediciones al día (una cada 10 minutos, es decir, 144 por jornada), se ha realizado una agregación diaria mediante el cálculo de la media de los errores. Esta estrategia permite reducir el ruido puntual y facilita la visualización de tendencias generales, que es el objetivo del presente análisis. La agregación se ha aplicado únicamente con fines descriptivos y no afecta a las fases posteriores de modelado supervisado, donde se conservará la resolución temporal completa.

5.2. Análisis de deriva por módulo mediante regresión lineal

Además del análisis puntual del módulo 5, se ha extendido el estudio de deriva a todos los módulos del sensor de bajo coste. Esto resulta especialmente relevante ya que, al estar dispuestos secuencialmente en el nodo físico, podrían estar sometidos a condiciones ambientales ligeramente diferentes (ventilación, orientación, flujo de aire...), que afecten de forma desigual a su comportamiento.

Para cada módulo y para cada variable (PM2.5, O₃ y NO₂), se calcula el error diario respecto a la estación oficial. Posteriormente, se ajusta una regresión lineal simple del error frente al tiempo (expresado como número de días desde el inicio).

5.2.1. Fundamento matemático

Para cuantificar la presencia de deriva en cada módulo, se ajusta un modelo de regresión lineal simple sobre la serie de errores diarios para cada variable. Sean t los días desde el inicio del experimento, y $e_t^{(m)}$ el error correspondiente al módulo m en el día t . El modelo ajustado tiene la forma:

$$e_t^{(m)} = \beta_0^{(m)} + \beta_1^{(m)} \cdot t + \varepsilon_t$$

Donde:

- $\beta_0^{(m)}$ es el intercepto o valor inicial del error en el módulo m ,
- $\beta_1^{(m)}$ representa la pendiente o **deriva estimada**,
- ε_t es el término de error aleatorio.

El coeficiente $\beta_1^{(m)}$ indica el ritmo de cambio del error con el tiempo: si $\beta_1 > 0$ existe deriva creciente, si $\beta_1 < 0$ la deriva es decreciente, y si $\beta_1 \approx 0$, el error es estable [26].

Adicionalmente, se ha aplicado un modelo de suavizado local (*LOESS*) para capturar patrones no lineales de deriva. En este caso, se estima una función suavizada $\hat{f}^{(m)}(t)$ para cada módulo:

$$e_t^{(m)} = \hat{f}^{(m)}(t) + \varepsilon_t$$

La función $\hat{f}^{(m)}$ se construye a partir de regresiones locales ponderadas, sin asumir ninguna forma funcional explícita [27]. Este enfoque permite detectar inflexiones, curvaturas o ciclos que no serían visibles con un ajuste lineal clásico.

5.2.2. Resultados

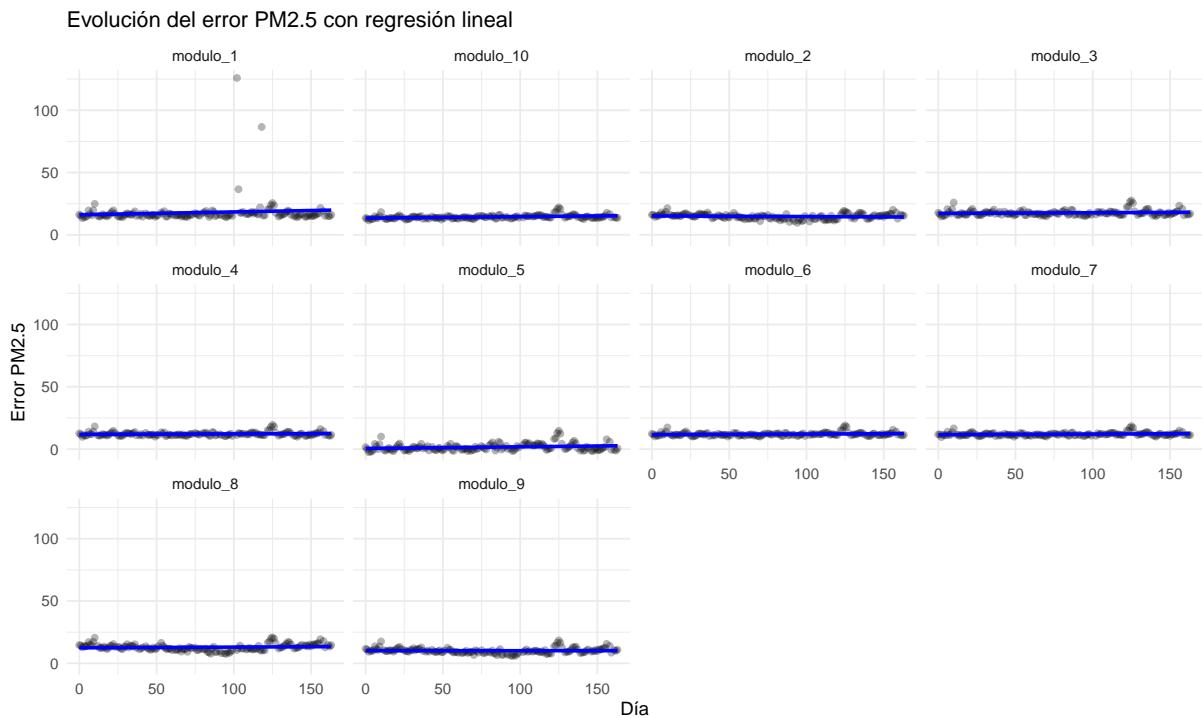


Figura 5.3: Evolución del error de PM2.5 por módulo con regresión lineal ajustada.

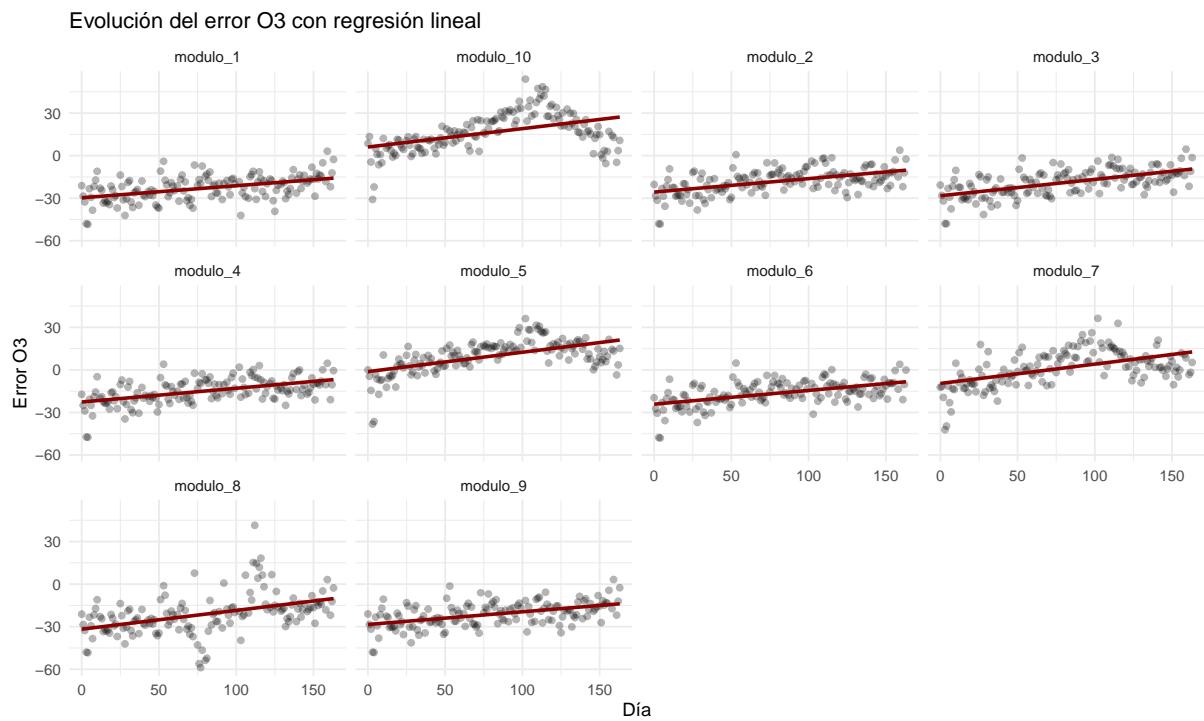


Figura 5.4: Evolución del error de O₃ por módulo con regresión lineal ajustada.

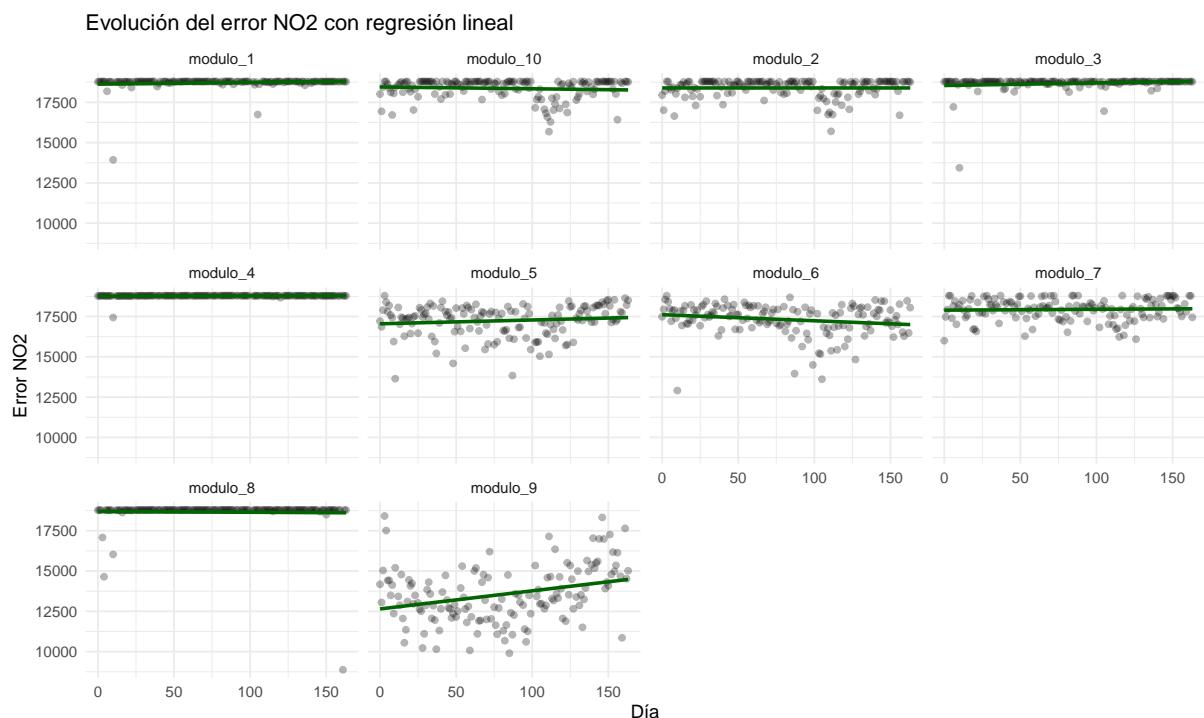


Figura 5.5: Evolución del error de NO₂ por módulo con regresión lineal ajustada.

En la Figura 5.3, que representa la evolución del error en PM2.5, se observa que la mayoría de los módulos presentan una tendencia ligeramente ascendente a lo largo del tiempo, pero con pendientes muy reducidas. Esto sugiere una deriva mínima. La homogeneidad observada entre los módulos evidencia una buena fiabilidad del LCS para la medición de PM2.5.

En la Figura 5.4, correspondiente al ozono (O_3), todos los módulos presentan pendientes positivas. Los módulos 5, 7 y 8 presentan las más elevadas, lo que indica una deriva creciente del error. Esta dispersión entre módulos justifica una calibración individualizada.

En la Figura 5.5, para NO_2 , los errores son extremadamente elevados. El **módulo 9** presenta la mayor pendiente positiva. Las rectas se trazan sobre una señal completamente alejada del valor de referencia, lo cual refuerza la conclusión de que el sensor no está operando adecuadamente.

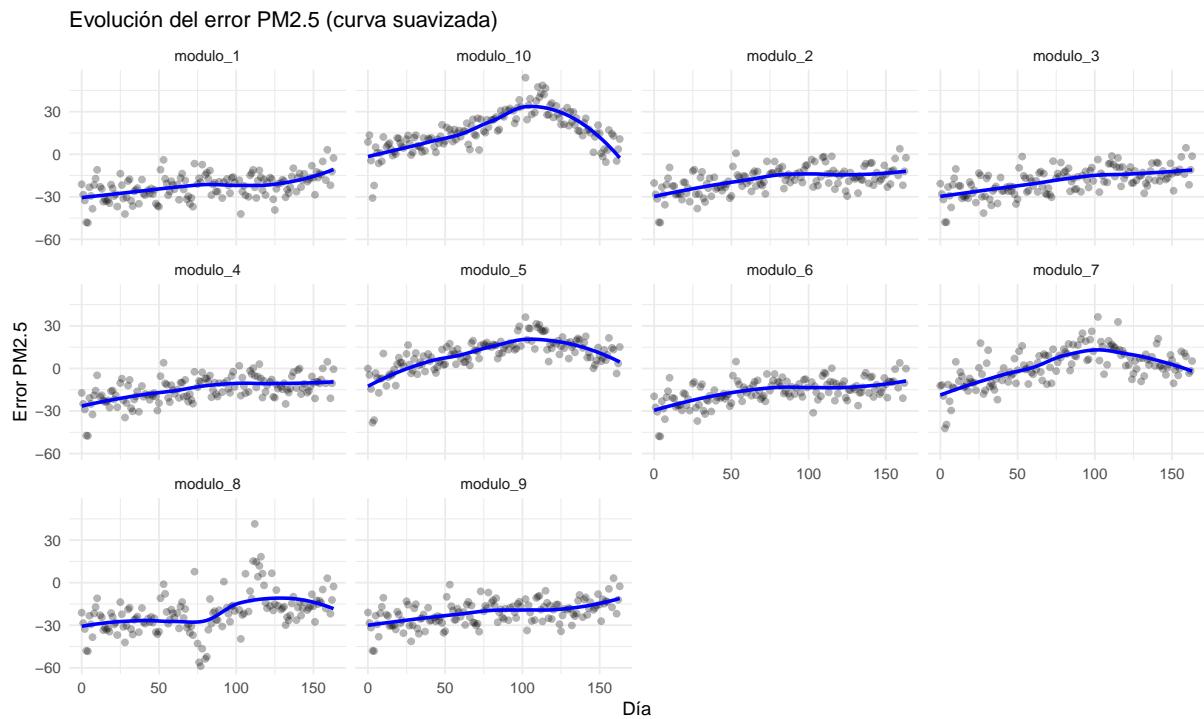


Figura 5.6: Evolución del error PM2.5 por módulo mediante suavizado LOESS.

Tal y como muestra la Figura 5.6, la mayoría de módulos presentan patrones de error con forma curvada: el error aumenta progresivamente pero con fluctuaciones. Esto podría deberse a factores ambientales estacionales, cambios en la ventilación o incluso a procesos de envejecimiento del sensor. Aunque no hay una deriva sostenida clara, sí existe estructura temporal en el error que podría aprovecharse en modelos de calibración supervisada.

5.3. Cálculo de la tendencia lineal del error por módulo

Con el objetivo de identificar qué módulos presentan una evolución sistemática del error a lo largo del tiempo, se ha calculado la pendiente de la regresión lineal simple del error diario respecto al tiempo para cada módulo y para cada variable (PM2.5, O_3 y NO_2).

5.3.1. Fundamento matemático

Para cada módulo m , se ajusta un modelo de la forma:

$$e_t^{(m)} = \beta_0^{(m)} + \beta_1^{(m)} \cdot t + \varepsilon_t$$

donde $e_t^{(m)}$ representa el error diario en el día t , y $\beta_1^{(m)}$ es la pendiente que indica el grado de deriva. A partir de este modelo, se extrae $\beta_1^{(m)}$ como medida de tendencia o deriva lineal.

5.3.2. Resultados

En las siguientes tablas se muestran los módulos ordenados de menor a mayor pendiente para cada variable. Una pendiente cercana a cero implica estabilidad, mientras que valores más altos (positivos o negativos) reflejan una tendencia sistemática en el error, positiva (drift creciente) o negativa (drift decreciente).

Ranking de módulos según la pendiente del error en PM2.5

Módulo	Tendencia PM2.5
modulo_2	-0.0054
modulo_9	-0.0002
modulo_4	0.0038
modulo_7	0.0048
modulo_6	0.0049
modulo_3	0.0055
modulo_8	0.0058
modulo_10	0.0115
modulo_5	0.0132
modulo_1	0.0229

Ranking de módulos según la pendiente del error en O₃

Módulo	Tendencia O ₃
modulo_1	0.0848
modulo_9	0.0894
modulo_2	0.0947
modulo_4	0.0969
modulo_6	0.0969
modulo_3	0.1153
modulo_10	0.1296
modulo_7	0.1363
modulo_8	0.1324
modulo_5	0.1365

Ranking de módulos según la pendiente del error en NO₂

Módulo	Tendencia NO ₂
modulo_6	-3.7681
modulo_10	-1.1230
modulo_8	-0.5000
modulo_2	0.0090
modulo_4	0.3066
modulo_7	0.5206
modulo_1	1.0919
modulo_3	1.5943
modulo_5	2.3593
modulo_9	11.2295

5.3.3. Comentario

Para la variable PM2.5, se observa una tendencia creciente del error en la mayoría de los módulos, siendo más acusada en los módulos **1** y **5**, que presentan las pendientes más altas (0.0229 y 0.0132 respectivamente). Aunque las pendientes siguen siendo de magnitud moderada, este patrón sugiere un ligero efecto acumulativo con el tiempo, probablemente asociado a condiciones ambientales o al envejecimiento de los sensores. Por el contrario, los módulos **9** y **4** presentan las pendientes más bajas —incluso negativas en el caso del módulo 9— lo que indica un comportamiento más estable en esta variable.

En el caso del O₃, la mayoría de los módulos muestran una deriva positiva. Los módulos **5**, **8** y **7** destacan por presentar pendientes elevadas. Esta heterogeneidad entre módulos refuerza la necesidad de realizar calibraciones individuales en sensores de O₃.

En cuanto al NO₂, los resultados son especialmente preocupantes. El **módulo 9** presenta una pendiente extremadamente alta (11.2295), lo que refleja una acumulación continua y significativa de error. Otros módulos como el **1**, **3** y **5** también muestran una tendencia creciente pronunciada. En cambio, el **módulo 6** y el **módulo 10** presentan pendientes negativas, siendo el módulo 2 el más estable con una pendiente de 0.009. Estos resultados evidencian graves problemas de fiabilidad para la medición de NO₂ con este sistema, lo que justificaría una revisión profunda del sensor o incluso su exclusión del análisis comparativo.

5.4. Conclusiones del análisis de deriva temporal

El análisis realizado sobre el error de los sensores de LCS ha permitido evaluar la posible existencia de deriva sistemática en las mediciones. Las variables PM2.5 y O₃ presentan comportamiento aceptable con deriva leve, aunque heterogénea, mientras que NO₂ muestra errores extremos y poco fiables.

En una primera fase, se analizó el comportamiento del módulo 5, considerado representativo por su estabilidad relativa. Esta aproximación permitió observar que el error en PM2.5, aunque con algunos picos, se mantenía razonablemente estable, mientras que el error en O₃ mostraba una clara deriva positiva durante los meses centrales del periodo, probablemente vinculada a fenómenos estacionales.

Posteriormente, se amplió el análisis a todos los módulos utilizando regresión lineal simple para estimar la tendencia del error diario. En **PM_{2.5}**, se detectaron pendientes moderadas y mayoritariamente positivas, con los módulos 1 y 5 como los más afectados por una deriva creciente, frente a módulos más estables como el 2 y el 9. En **O₃**, las pendientes fueron también positivas, destacando el módulo 5 por su elevada pendiente, mientras que el módulo 1 se mostró como el de menor pendiente. En **NO₂**, las tendencias fueron mucho más extremas y dispares, con el módulo 9 alcanzando valores claramente fuera de rango, lo que evidencia un comportamiento anómalo y posiblemente inutilizable.

Adicionalmente, se aplicó un suavizado local (*LOESS*) que permitió detectar estructuras no lineales en algunas series, revelando trayectorias curvadas o inestabilidad episódica. Este comportamiento sugiere la influencia de factores ambientales y degradación diferencial de los sensores con el tiempo.

En conjunto, este análisis ha permitido caracterizar la estabilidad temporal de cada módulo y de cada variable. La variable **PM_{2.5}**, medida mediante sensores ópticos, se mantiene como una opción relativamente fiable, mostrando una deriva leve y consistente entre módulos. Por su parte, el **O₃**, monitorizado a través de sensores electroquímicos (**ECH**), presenta una deriva más marcada y heterogénea, lo que refleja su mayor sensibilidad a condiciones ambientales y su menor robustez estructural. Finalmente, el **NO₂** muestra un comportamiento claramente errático e inestable, sin correspondencia aparente con la señal de referencia. Estos hallazgos justifican la necesidad de una calibración supervisada individualizada por módulo y por variable, así como una consideración especial para el uso de sensores electroquímicos en contextos ambientales variables.

Capítulo 6

Modelado de calibración mediante aprendizaje supervisado

Toda la etapa de preprocesamiento y análisis exploratorio se ha llevado a cabo en el lenguaje [R](#), aprovechando su potencia para el tratamiento estructurado de datos y su sintaxis expresiva para visualización y análisis estadístico [28]. Sin embargo, para abordar la fase de modelado supervisado, se ha optado por realizar el cambio a [Python](#), un entorno especialmente consolidado en el ámbito del [ML](#) gracias a su ecosistema robusto y a la claridad con la que permite construir, evaluar y comparar modelos predictivos [29].

Esta transición responde no solo a razones técnicas, sino también a una cuestión de afinidad metodológica: [R](#) ha ofrecido la precisión y flexibilidad necesarias para explorar los datos con detalle, mientras que [Python](#) proporciona un entorno más cómodo y natural para trabajar con algoritmos de aprendizaje automático, permitiendo abordar esta siguiente etapa del proyecto con una mayor fluidez y control en el diseño experimental.

El análisis exploratorio realizado ha puesto de manifiesto que las mediciones de los sensores de bajo coste presentan errores sistemáticos respecto a las referencias oficiales. Estos errores varían a lo largo del tiempo y entre módulos, lo que justifica el uso de modelos de calibración que permitan mejorar la precisión de los datos registrados.

Con este objetivo, se plantean dos estrategias complementarias de aprendizaje supervisado, ambas utilizando como entrada las variables medidas por el sensor de bajo coste (PM2.5, O₃, temperatura, humedad, etc.), excluyendo la variable objetivo cuando corresponda:

- **Modelo directo:** se entrena un modelo para predecir directamente el valor oficial de una variable (por ejemplo, O₃) a partir del resto de variables medidas por el sensor. Este modelo permite calibrar directamente el sensor, sustituyendo su salida por una estimación más precisa.
- **Modelo basado en el error:** en este caso, la salida del modelo es el error cometido por el sensor respecto a la referencia ($Y = \text{sensor} - \text{referencia}$). El objetivo no es aplicar esta predicción para corregir la medición, sino estudiar si el error es sistemático, es decir, si puede explicarse y modelarse en función de otras variables. Esto proporciona información clave sobre la estructura del error, especialmente en aquellos casos donde se ha observado deriva temporal o sensibilidad ambiental.

Ambos enfoques serán evaluados y comparados en términos de rendimiento predictivo.

En particular, el segundo modelo ofrece una perspectiva diagnóstica: si el error es predecible, se justifica la calibración; si no lo es, el error se considera aleatorio e irreductible mediante modelos supervisados. Esta estrategia doble permite abordar la calibración desde un enfoque práctico (modelo directo) y uno analítico (modelo del error), enriqueciendo la comprensión del comportamiento del sistema.

A lo largo de esta sección se detallará el proceso de preparación de los datos, la selección de variables, la construcción de los modelos, y la evaluación de su rendimiento mediante métricas adecuadas. Además, se analizará el impacto de factores como el módulo concreto, o la estabilidad temporal en la calidad del ajuste.

6.1. Preparación de los datos

El primer paso para aplicar modelos de Técnica de ML en la que el modelo se entrena con datos de entrada y salida conocidas (ML) consiste en preparar adecuadamente los datos de entrada. En este caso, se dispone de una tabla donde se han unido las mediciones del LCS con las correspondientes mediciones oficiales de referencia. Las columnas disponibles incluyen tanto contaminantes como variables auxiliares ambientales.

El conjunto de variables predictoras (\mathbf{X}) incluye los siguientes atributos: NO₂, Dióxido de carbono (CO₂), O₃, CO, CH₂O, T, RH, PM2.5¹ y VOC_cat. Estas variables se han seleccionado por estar directamente disponibles en el LCS y ser potencialmente relevantes para explicar la variable objetivo.

Se han evaluado tres variables objetivo (\mathbf{y}), correspondientes a los valores oficiales de referencia:

- PM2.5_ref: concentración oficial de PM finas.
- O3_ref: concentración oficial de O₃.
- NO2_ref: concentración oficial de NO₂.

Para construir los modelos, se dispone inicialmente de un conjunto de 145 días consecutivos con datos combinados del LCS y de la estación oficial. Este conjunto se ha dividido respetando el orden cronológico en dos subconjuntos:

- 80 % (116 días) para entrenamiento.
- 20 % (29 días) para validación.

Posteriormente, se han incorporado 15 días adicionales que no se habían utilizado en ninguna fase previa del proceso, y que sirven como conjunto de test externo. Este conjunto representa datos completamente nuevos y simula el comportamiento del modelo en condiciones reales de despliegue.

Esta división tiene como objetivo evaluar el comportamiento de los modelos de aprendizaje supervisado en condiciones realistas. Al tratarse de datos temporales, no se realiza una partición aleatoria, ya que ello podría inducir *fugas de información* (*data leakage*)

¹Se han excluido PM1.0 y PM10.0 tras detectar alta colinealidad en el análisis exploratorio con PM2.5 y con el objetivo de reducir la redundancia del modelo.

entre fases. Al mantener la secuencia temporal, se garantiza que el modelo aprenda sobre el pasado y se evalúe sobre observaciones futuras, reproduciendo el flujo real de datos en un escenario de aplicación. La validación intermedia permite ajustar hiperparámetros sin comprometer la evaluación final, que se realiza únicamente sobre el conjunto de test externo.

6.2. Modelo base (baseline)

En esta primera aproximación se ha optado por trabajar con la **media diaria** de las variables, en lugar de utilizar toda la resolución temporal disponible (una medición cada 10 minutos). Esta decisión se basa en la naturaleza de los modelos lineales empleados, que podrían no capturar adecuadamente las dinámicas a muy corto plazo. La agregación diaria permite reducir el ruido de alta frecuencia, facilitar la interpretación de los coeficientes y evaluar tendencias generales de calibración sin que el modelo se vea penalizado por la complejidad estructural de los datos. Aunque esta elección implica una pérdida de granularidad, resulta adecuada para establecer un modelo base de referencia que sirva como punto de partida para enfoques posteriores más avanzados [30].

Nota: Aunque lo más adecuado para modelos supervisados es trabajar con la granularidad temporal completa, en este modelo base se ha optado por utilizar la media diaria. Esta simplificación se justifica por el carácter lineal del enfoque inicial, que se beneficia de una representación más estable y menos ruidosa de los datos.

6.2.1. Comparativa entre modelos

Para cada variable objetivo, se comparan dos estrategias de calibración lineal:

- El **modelo base**, que utiliza únicamente la señal directa del sensor para calibrar su propia variable (por ejemplo, usar PM2.5 para predecir PM2.5_ref). Este modelo sirve como referencia mínima.
- El **modelo multivariable óptimo**, seleccionado mediante búsqueda exhaustiva de combinaciones de variables predictoras usando la biblioteca mlxtend. En concreto, se emplea el objeto `ExhaustiveFeatureSelector`, que evalúa todas las combinaciones posibles de variables mediante la métrica R^2 en validación, identificando el subconjunto con mejor rendimiento. Esta estrategia permite detectar sinergias entre variables y evaluar cuánta mejora se obtiene respecto al modelo base. Aunque el número de combinaciones crece exponencialmente, el número total de variables consideradas (9 tras limpieza) permite realizar la búsqueda de forma completa (hasta 511 combinaciones).

Ambos modelos se comparan cuantitativamente mediante las métricas Root Mean Square Error (RMSE), Mean Absolute Error (MAE) y Coeficiente de determinación (R^2), y cualitativamente mediante representaciones gráficas: series temporales comparadas, dispersión predicción vs. referencia, y análisis de residuos.

Además, para el modelo multivariable se calcula la **importancia relativa de las variables** mediante la estandarización de los datos y el análisis de los coeficientes normalizados. Esta información permite entender qué variables aportan más valor a la predicción, más allá de su escala original.

6.2.2. Fundamento matemático del modelo

Dado un conjunto de n observaciones, cada una con p variables predictoras, el modelo de [MLR](#) tiene la siguiente forma general:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

donde:

- \hat{y}_i es la predicción del valor de la variable objetivo para la observación i .
- x_{ij} es el valor de la variable predictora j para la observación i .
- β_0 es el intercepto del modelo.
- β_j son los coeficientes ajustados para cada variable.
- ε_i representa el error (residuo) de la predicción.

Los coeficientes se estiman minimizando la suma de los errores cuadráticos ([Mean Square Error \(MSE\)](#), criterio de *mínimos cuadrados ordinarios*):

$$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Visualizaciones adicionales e interpretación del modelo. Para complementar el análisis del modelo calibrado, se realizan las siguientes representaciones y evaluaciones:

- **Top-5 combinaciones óptimas:** listado de las cinco mejores combinaciones de variables según [R²](#) en validación.
- **Rendimiento vs. complejidad:** gráfica que muestra la evolución del [R²](#) en función del número de variables utilizadas en el modelo.
- **Mapa de correlaciones:** matriz de correlación entre todas las variables predictoras y la variable objetivo, representada mediante un *heatmap*. Permite detectar relaciones lineales, redundancias y patrones de asociación.
- **Importancia relativa de las variables:**
 - Se reentrena el modelo óptimo con las variables estandarizadas (media 0, desviación típica 1) usando [StandardScaler](#).
 - Se extraen los coeficientes normalizados, que indican el peso relativo de cada variable en la predicción.
 - Se representa un gráfico de barras con los coeficientes estandarizados ordenados por magnitud absoluta.

6.2.3. Evaluación cuantitativa

Para ambos modelos se calculan las siguientes métricas estándar en problemas de regresión:

- **RMSE**: error cuadrático medio.
- **MAE**: error absoluto medio.
- **R²**: coeficiente de determinación.

Además, para el modelo multivariable se incluyen:

- **Mean Absolute Prediction Error (MAPE)**: error porcentual absoluto medio.
- **Explained Variance Score (EVS)**: proporción de varianza explicada.
- **Mejora porcentual**: reducción del **MAE** frente a la lectura original del sensor.

Formulación matemática:

$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \\ \text{MAE} &= \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \\ \text{MAPE} &= \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \\ R^2 &= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \\ \text{EVS} &= 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)} \end{aligned}$$

Este *pipeline* no solo proporciona una estimación precisa del rendimiento de los modelos, sino que también permite interpretar en profundidad sus resultados. Así, se obtiene una visión clara tanto de la precisión alcanzada como del comportamiento interno de cada modelo, incluyendo su sensibilidad a las diferentes variables del sistema.

Además, en este *pipeline* y en los siguientes se utilizarán exclusivamente los datos del módulo considerado más fiable, de acuerdo con los resultados del análisis exploratorio previo, el módulo 6. Esta elección busca maximizar la calidad del conjunto de entrenamiento y evaluación, asegurando que las conclusiones extraídas estén basadas en datos consistentes y representativos del funcionamiento general de la red sensorial.

6.3. Resultados por variable

6.3.1. Resultados para PM2.5

La variable PM2.5 muestra una excelente capacidad de calibración a partir de los datos del sensor de bajo coste. Ya el modelo base, que utiliza únicamente la medición directa

del sensor (PM2.5), ofrece un rendimiento notable, con un R^2 de 0.79 en validación y un MAE de 0.62. Esto indica que la señal cruda del sensor ya está fuertemente alineada con la referencia oficial.

Al aplicar un modelo multivariante y realizar una búsqueda exhaustiva de combinaciones de predictores, se identificó una combinación óptima formada por [O3_mod6, PM2.5_mod6], que mejora todos los indicadores evaluados. En particular, el R^2 sube a 0.85, el MAE se reduce a 0.53 y el RMSE también desciende, como se resume en la siguiente tabla:

Cuadro 6.1: Comparativa de rendimiento de los modelos para PM2.5

Modelo	RMSE	MAE	R^2
Solo PM2.5 (básico)	0.722	0.620	0.790
Modelo óptimo (O3_mod6 + PM2.5_mod6)	0.608	0.529	0.851

La mejora se ilustra también mediante las visualizaciones temporales, donde se aprecia cómo la predicción del modelo se ajusta a la serie real, y mediante la reducción de residuos:

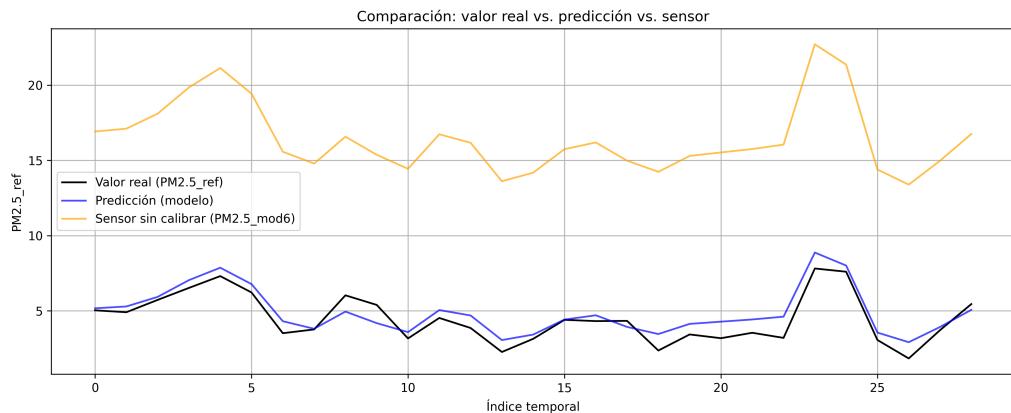


Figura 6.1: Comparación temporal: valor real vs. predicción vs. sensor (modelo base, PM2.5)

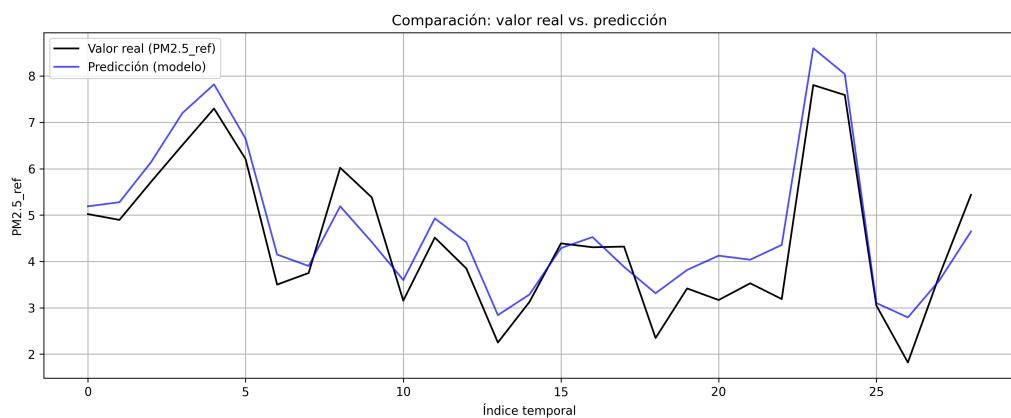


Figura 6.2: Comparación temporal: valor real vs. predicción (modelo óptimo, PM2.5)

Resumen de combinaciones exploradas. El análisis exhaustivo identificó las siguientes cinco mejores combinaciones por rendimiento en validación:

Cuadro 6.2: Top 5 combinaciones de variables por R^2 (validación)

Nº var	Combinación	R^2_{train}	R^2_{val}
2	[O3_mod6, PM2.5_mod6]	0.871	0.851
3	[O3_mod6, CH20_mod6, PM2.5_mod6]	0.871	0.851
3	[O3_mod6, CO_mod6, PM2.5_mod6]	0.872	0.851
3	[O3_mod6, T ^a _mod6, PM2.5_mod6]	0.871	0.850
4	[O3_mod6, CO_mod6, CH20_mod6, PM2.5_mod6]	0.872	0.850

Como se observa en la tabla, los valores de R^2 en validación son ligeramente inferiores a los obtenidos en entrenamiento, lo cual es un comportamiento esperable y deseable. Esta diferencia sugiere que el modelo generaliza correctamente y no está sobreajustado (*overfitting*) a los datos de entrenamiento. La cercanía de los valores entre ambos conjuntos indica que las combinaciones seleccionadas ofrecen un buen equilibrio entre capacidad explicativa y robustez predictiva.

Visualizaciones adicionales. Se muestran a continuación distintas visualizaciones diagnósticas del modelo:

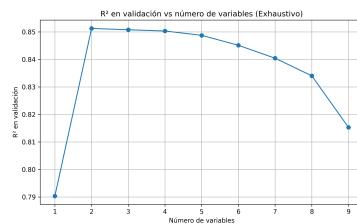
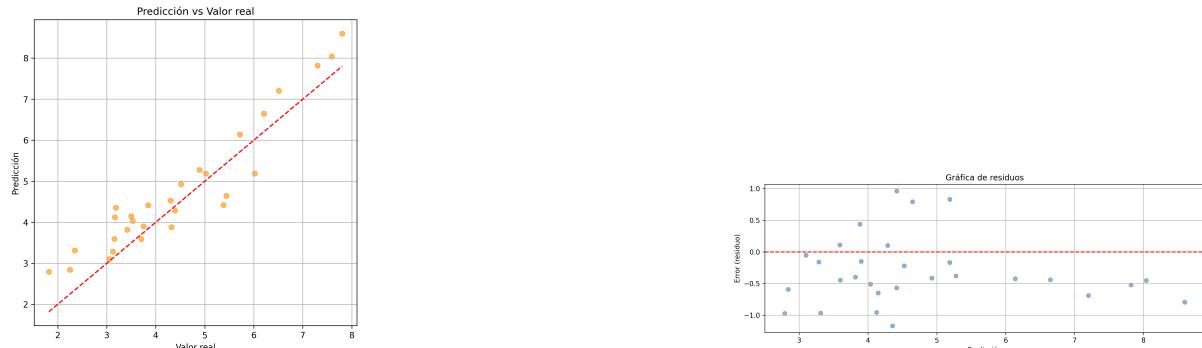
(c) Evolución de R^2 vs nº de variables

Figura 6.3: Visualizaciones complementarias del modelo calibrado para PM2.5

El diagrama de dispersión entre predicción y valor real muestra la proximidad de los puntos a la recta identidad, lo que refleja precisión en las estimaciones. El gráfico de residuos permite verificar si los errores están centrados en torno a cero, sin patrones sistemáticos, lo cual indica un modelo bien ajustado. Por último, la evolución del R^2 en función del número de variables permite observar cómo aumenta la capacidad predictiva al añadir más información, hasta alcanzar un punto óptimo más allá del cual se obtienen peoras significativas.

Importancia relativa de las variables. Se reentrena el modelo óptimo con los datos estandarizados. La figura siguiente muestra la magnitud relativa de los coeficientes:

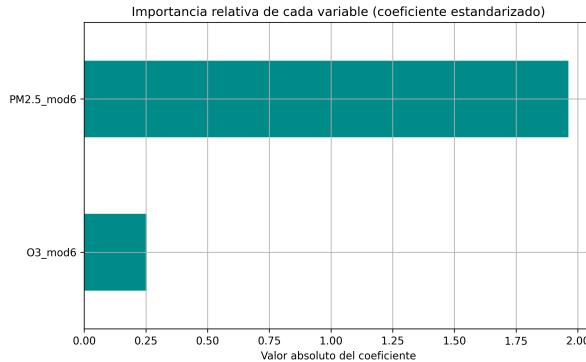


Figura 6.4: Importancia relativa de cada variable (coeficientes estandarizados, PM2.5)

- PM2.5_mod6: 1.9621
- O3_mod6: 0.2538

Mapa de correlaciones. Antes y después del modelado:



Figura 6.5: Matriz de correlación antes (izquierda) y después (derecha) de la selección de variables.

Diagnóstico complementario. Además de las métricas estándar de regresión, se han calculado indicadores adicionales que permiten evaluar de forma más completa el rendimiento del modelo. En particular:

- **MAE del sensor sin calibrar:** 12.005. Representa el error medio absoluto que comete el sensor directamente, sin aplicar ningún tipo de corrección. Este valor sirve como referencia para valorar la utilidad de la calibración.
- **MAE del modelo calibrado:** 0.529. Es el error medio absoluto tras aplicar la regresión lineal. Su notable reducción frente al valor anterior evidencia la eficacia del modelo.
- **Mejora relativa:** 95.6 %. Indica el porcentaje de mejora logrado respecto al error del sensor sin calibrar. Una mejora tan elevada implica que la calibración consigue corregir de forma muy precisa las desviaciones sistemáticas del sensor.

- **MAPE (Mean Absolute Percentage Error):** 14.06 %. Refleja el error relativo medio en porcentaje, útil para interpretar el rendimiento del modelo en relación con la magnitud de los valores reales.
- **Explained Variance Score:** 0.889. Mide la proporción de la varianza de la variable objetivo que es explicada por el modelo. Un valor cercano a 1 indica que la mayor parte de la variabilidad de la referencia oficial ha sido capturada por el modelo calibrado.

6.3.2. Evaluación en datos externos.

Finalmente, se evaluó el modelo óptimo sobre un conjunto completamente nuevo y no visto durante el entrenamiento ni la validación.

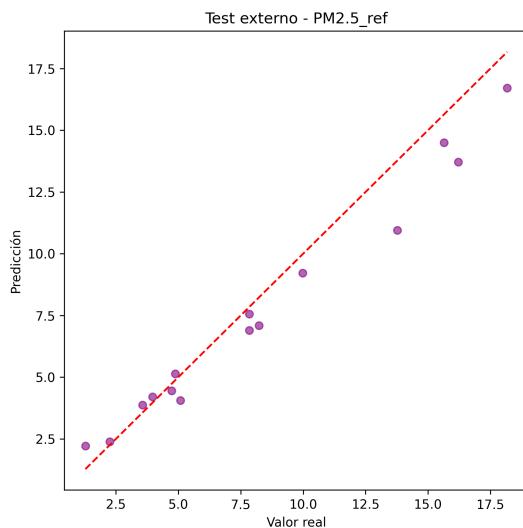


Figura 6.6: Dispersión real vs predicción en test externo (PM2.5)

Evaluación en datos externos de test para PM2.5_ref...

- **RMSE:** 1.234
- **MAE:** 0.951
- **R²:** 0.944
- **MAPE:** 14.26 %
- **Explained Variance Score:** 0.962

Estos resultados confirman que el modelo conserva un rendimiento excelente en condiciones reales, demostrando tanto robustez como generalización. El error absoluto medio se mantiene por debajo de 1 unidad y el coeficiente de determinación R^2 se sitúa por encima de 0.94, lo que valida la utilidad del sistema de calibración en escenarios de despliegue real.

Es importante matizar que el conjunto de test externo comprende únicamente 15 días, y que se está evaluando el modelo sobre datos previamente suavizados mediante promedios

diarios. Esto puede contribuir a una mayor estabilidad en las predicciones y explicar en parte el rendimiento tan elevado. En aplicaciones prácticas más extensas, se recomienda una evaluación adicional con periodos más largos para asegurar la robustez a largo plazo.

6.3.3. Resultados para O₃

En el caso del ozono (O₃), los resultados son considerablemente más modestos. El modelo base, que emplea únicamente la señal directa del sensor de bajo coste (O₃_mod6), muestra un rendimiento negativo en validación, con un R^2 de -1,05 y un MAE de 8,80. Esto indica que el sensor sin calibrar no ofrece una señal fiable para estimar la concentración oficial.

Tras aplicar una regresión multivariable y realizar una búsqueda exhaustiva, se seleccionó como combinación óptima el conjunto de seis predictores: [O₃_mod6, CO_mod6, CH20_mod6, T^a_mod6, PM2.5_mod6, VOC_cat_mod6]. No obstante, el rendimiento en validación sigue siendo pobre (con R^2 negativo), y el error medio incluso empeora respecto al modelo base, como refleja la siguiente tabla:

Cuadro 6.3: Comparativa de rendimiento de los modelos para O₃

Modelo	RMSE	MAE	R^2
Solo O ₃ (básico)	10.801	8.804	-1.048
Modelo óptimo (6 vars)	8.525	6.957	-0.276

A pesar de la reducción del error cuadrático medio, el modelo multivariable no logra superar significativamente al modelo base, lo que sugiere que la relación entre las variables disponibles y la concentración real de ozono no es capturada de forma efectiva con un modelo lineal. Las visualizaciones temporales también reflejan esta limitación:

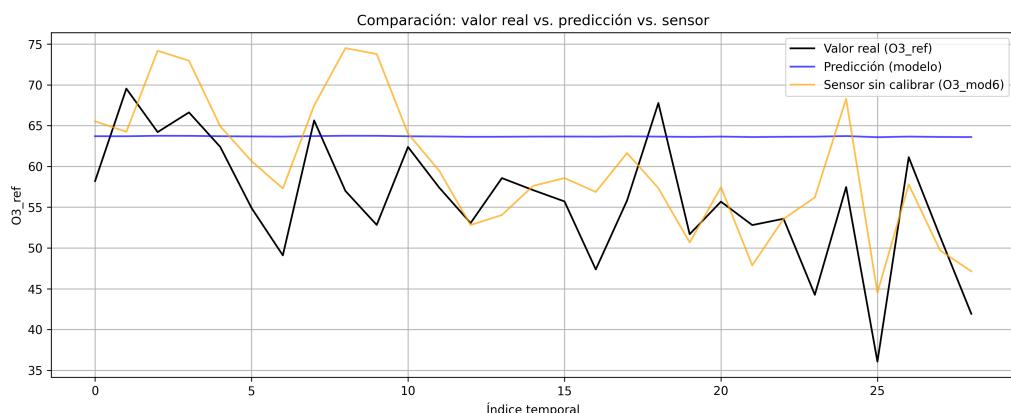
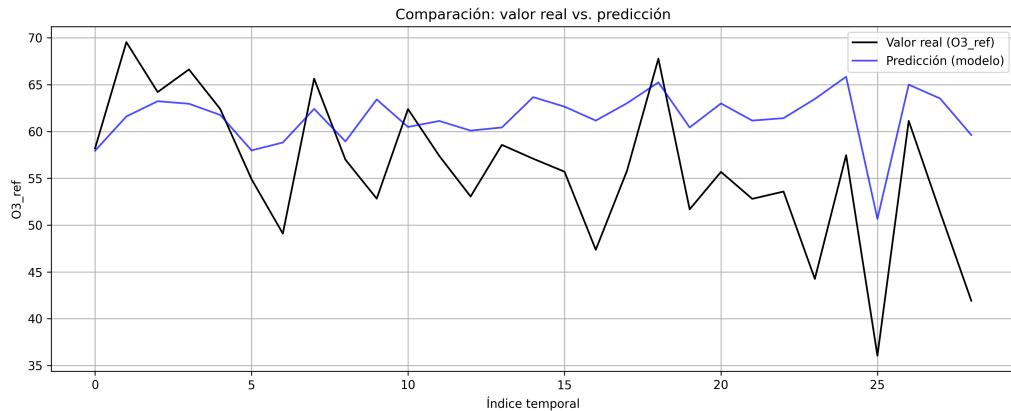


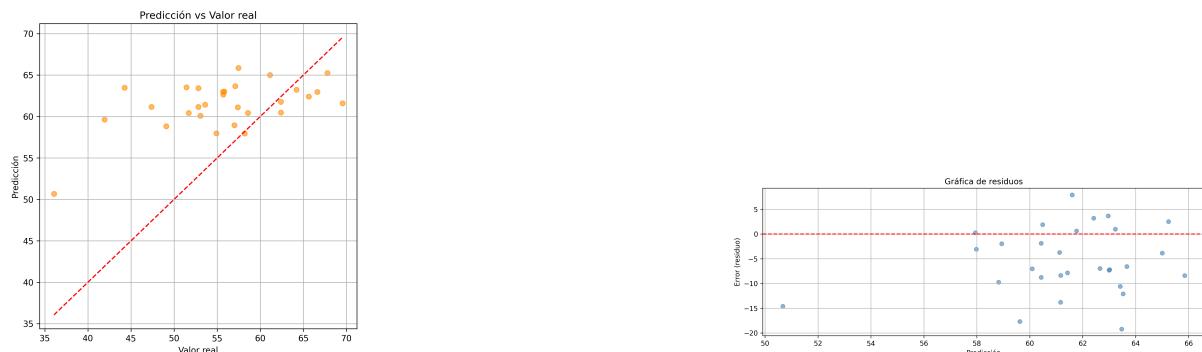
Figura 6.7: Comparación temporal: valor real vs. predicción vs. sensor (modelo base, O₃)

Figura 6.8: Comparación temporal: valor real vs. predicción (modelo óptimo, O_3)

Resumen de combinaciones exploradas. El análisis exhaustivo mostró que, aunque el modelo mejora al usar más variables, ninguno logra una validación satisfactoria:

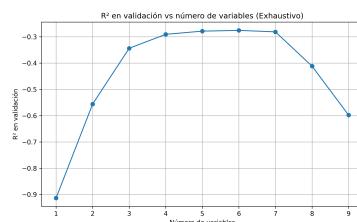
Cuadro 6.4: Top 5 combinaciones de variables por R^2 (validación)

Nº var	Combinación	R^2_{train}	R^2_{val}
6	[O3_mod6, CO_mod6, CH20_mod6, T ^a _mod6, PM2.5_mod6, VOC_cat_mod6]	0.220	-0.276
5	[O3_mod6, CH20_mod6, T ^a _mod6, PM2.5_mod6, VOC_cat_mod6]	0.220	-0.278
7	[CO ₂ _mod6, O3_mod6, CO_mod6, CH20_mod6, T ^a _mod6, PM2.5_mod6, VOC_cat_mod6]	0.256	-0.281
6	[CO ₂ _mod6, O3_mod6, CH20_mod6, T ^a _mod6, PM2.5_mod6, VOC_cat_mod6]	0.256	-0.289
5	[O3_mod6, CO_mod6, CH20_mod6, T ^a _mod6, PM2.5_mod6]	0.219	-0.289



(a) Dispersión: predicción vs. real

(b) Gráfico de residuos

(c) Evolución de R^2 vs n° de variablesFigura 6.9: Visualizaciones complementarias del modelo calibrado para O_3

Visualizaciones adicionales.

Importancia relativa de las variables. Se reentrenó el modelo con estandarización de predictores. El análisis de coeficientes muestra:

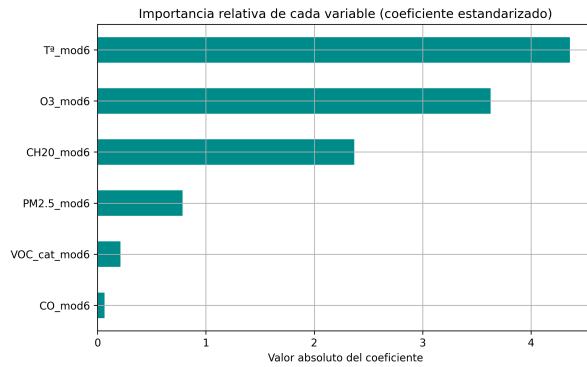


Figura 6.10: Importancia relativa de cada variable (coeficientes estandarizados, O₃)

- O₃_mod6: 3.6277
- CH20_mod6: -2.3703
- T^a_mod6: -4.3603
- PM2.5_mod6: 0.7841
- VOC_cat_mod6: 0.2104
- CO_mod6: 0.0655

Mapa de correlaciones. Antes y después del modelado:



Figura 6.11: Matriz de correlación antes (izquierda) y después (derecha) de la selección de variables.

Diagnóstico complementario.

- **MAE del sensor sin calibrar:** 5.952
- **MAE del modelo calibrado:** 6.957
- **Mejora relativa:** -16.9 %. El modelo incrementa el error respecto al sensor.

- **MAPE:** 13.81 %
- **Explained Variance Score:** 0.255

6.3.4. Evaluación en datos externos.

Finalmente, se evaluó el modelo óptimo sobre un conjunto completamente nuevo y no visto durante el entrenamiento ni la validación.

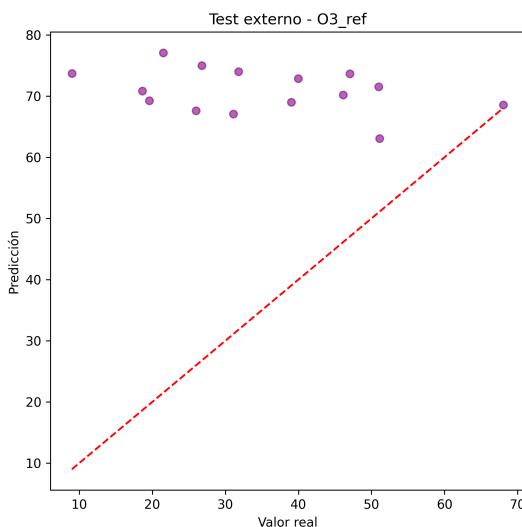


Figura 6.12: Dispersión real vs predicción en test externo (O_3)

Evaluación en datos externos de test para O3_ref...

- **RMSE:** 39.528
- **MAE:** 35.788
- **R²:** -5.786
- **MAPE:** 162.15 %
- **Explained Variance Score:** -0.223

Estos resultados reflejan un fallo claro de generalización. A pesar de que el conjunto de test contiene sólo 15 días suavizados, el modelo no logra capturar la señal de referencia para el ozono, mostrando un error absoluto medio muy elevado. Esto puede deberse a una baja calidad de la señal original, a una relación no lineal con los predictores seleccionados, o a un exceso de ruido en la variable objetivo.

6.3.5. Resultados para NO₂

A diferencia de otras variables, la calibración de NO₂ presenta mayores dificultades. El modelo base, que solo utiliza la variable directa del sensor NO2_mod6, fue descartado por su baja correlación con la referencia oficial, por lo que se empleó directamente un enfoque multivariante.

La búsqueda exhaustiva de combinaciones predictores reveló que la mejor combinación está formada por [03_mod6, CH20_mod6], logrando un rendimiento moderado con un R^2 de 0.26 en validación. Este resultado sugiere cierta capacidad explicativa, aunque limitada, como se resume en la siguiente tabla:

Cuadro 6.5: Comparativa de rendimiento del modelo óptimo para NO₂

Conjunto	RMSE	MAE	R^2
Entrenamiento	5.79	4.49	0.134
Validación	3.33	2.64	0.265

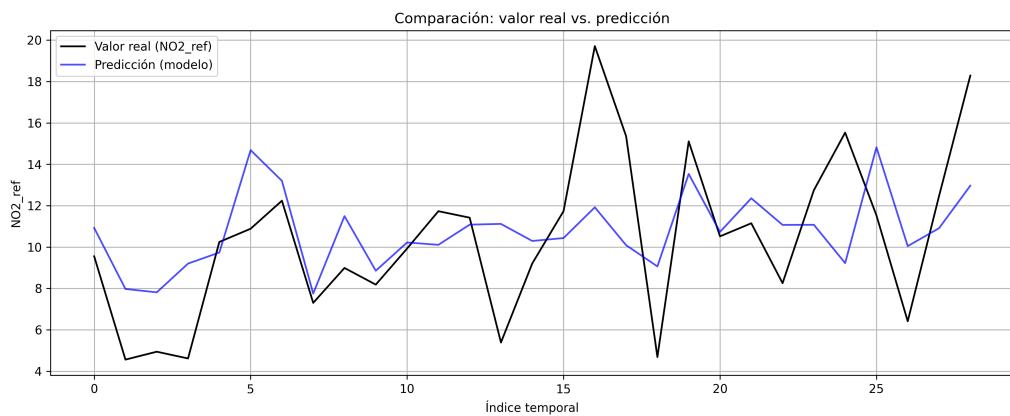


Figura 6.13: Comparación temporal: valor real vs. predicción (modelo óptimo, O₃)

Resumen de combinaciones exploradas. Las cinco combinaciones más prometedoras se resumen en la tabla siguiente:

Cuadro 6.6: Top 5 combinaciones de variables por R^2 (validación)

Nº var	Combinación	R^2_{train}	R^2_{val}
2	[O3_mod6, CH20_mod6]	0.134	0.265
3	[O3_mod6, CO_mod6, CH20_mod6]	0.136	0.263
3	[NO2_mod6, O3_mod6, CH20_mod6]	0.175	0.248
4	[NO2_mod6, O3_mod6, CO_mod6, CH20_mod6]	0.180	0.244
3	[O3_mod6, CH20_mod6, PM2.5_mod6]	0.136	0.238

Visualizaciones adicionales. Se muestran a continuación distintas visualizaciones diagnósticas del modelo:

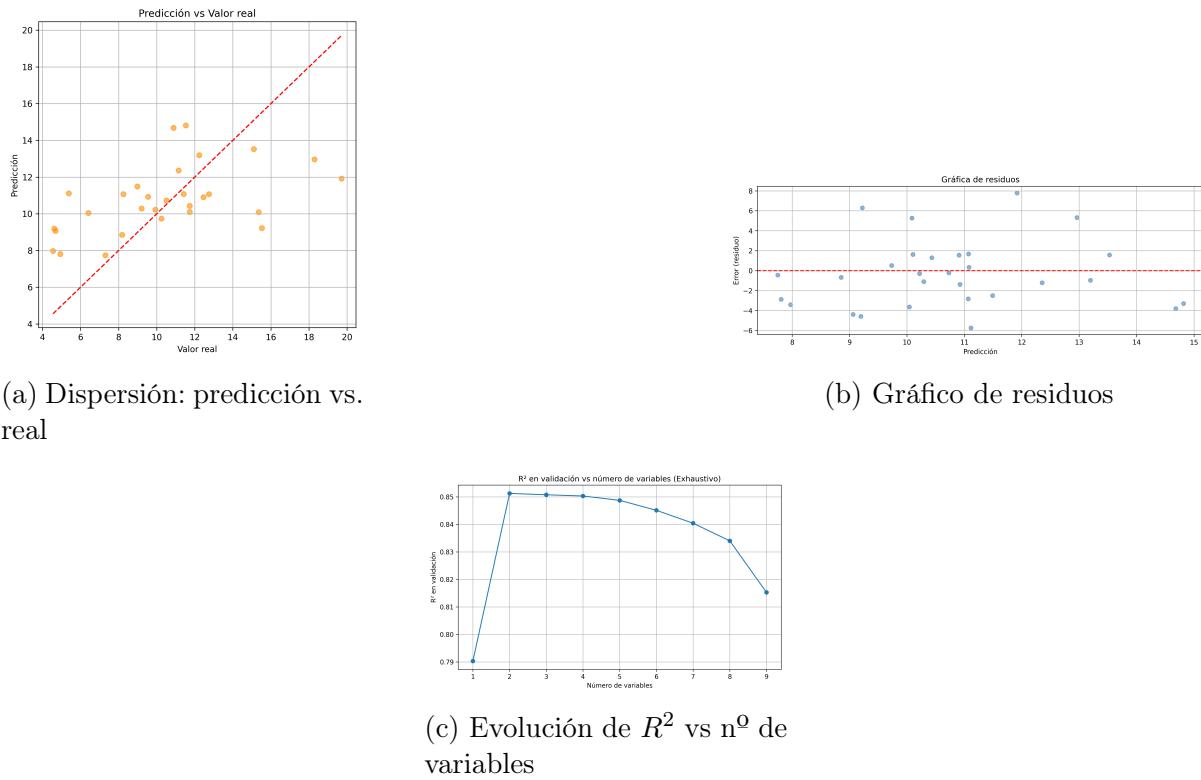
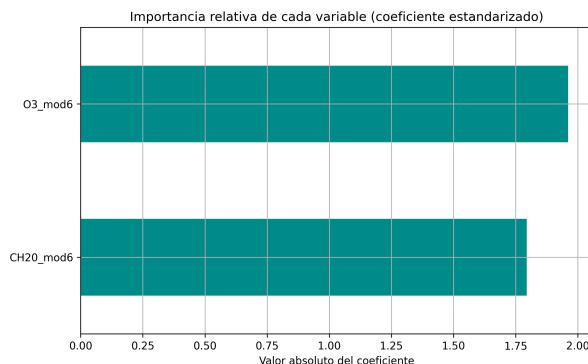


Figura 6.14: Visualizaciones complementarias del modelo calibrado para PM2.5

Visualizaciones diagnósticas.

Importancia relativa de las variables. En la figura siguiente se representa la magnitud relativa de los coeficientes estandarizados del modelo óptimo:

Figura 6.15: Importancia relativa de cada variable (coeficientes estandarizados, NO₂)

- CH20_mod6: 1.7942
- O3_mod6: -1.9614

Mapa de correlaciones. Antes y después del modelado:



Figura 6.16: Matriz de correlación antes (izquierda) y después (derecha) de la selección de variables.

Métricas complementarias. Además de las métricas estándar, se han calculado las siguientes medidas adicionales:

- **MAPE (Mean Absolute Percentage Error):** 30.74 %
- **Explained Variance Score:** 0.273

6.3.6. Evaluación en datos externos.

Finalmente, se evaluó el modelo óptimo sobre un conjunto completamente nuevo y no visto durante el entrenamiento ni la validación.

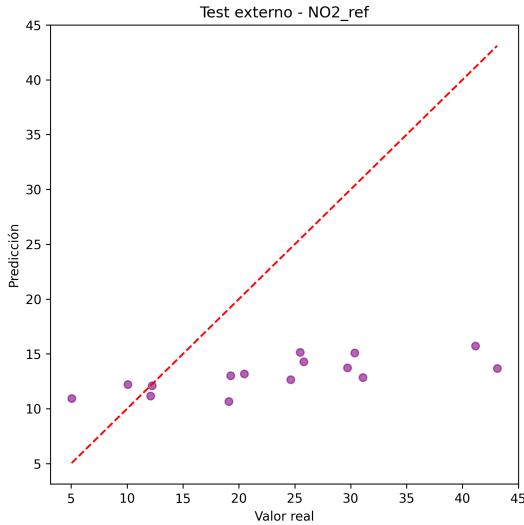


Figura 6.17: Dispersión real vs predicción en test externo (NO_2)

Evaluación en datos externos de test para NO2_ref ...

- **RMSE:** 13.944
- **MAE:** 11.279

- **R²:** -0.743
- **MAPE:** 45.72 %
- **Explained Variance Score:** 0.191

Estos resultados muestran un rendimiento deficiente en condiciones reales. El modelo no es capaz de generalizar correctamente sobre datos no vistos, como evidencia el valor negativo de R^2 . Esta caída de rendimiento puede deberse a múltiples factores, incluyendo una pobre correlación inicial entre los sensores y la referencia oficial, un conjunto de entrenamiento reducido o una alta variabilidad en la serie temporal de NO₂. En este caso, el sistema de calibración no logra corregir eficazmente el sesgo del sensor.

Se recomienda estudiar fuentes adicionales de variabilidad o aplicar modelos más complejos si se busca una mejora sustancial para esta variable.

6.3.7. Conclusiones del modelo base

El modelo de regresión lineal se ha utilizado como punto de partida para calibrar las señales proporcionadas por los sensores de bajo coste frente a las mediciones oficiales. A pesar de su simplicidad, esta técnica permite identificar relaciones significativas y proporciona una línea base sólida sobre la cual evaluar mejoras posteriores mediante técnicas más complejas.

PM2.5. La calibración de la materia particulada fina muestra un comportamiento excelente. El modelo base, que utiliza únicamente la variable PM2.5_mod6, ya alcanza un R^2 de 0.790 y un MAE de 0.620 en validación, lo que indica una fuerte correlación entre el sensor y la referencia oficial. Al extender el modelo con otras variables relevantes como 03_mod6, se logra una mejora clara: el R^2 asciende a 0.851 y el MAE desciende a 0.529. Además, en el conjunto de test externo (15 días posteriores no vistos), el modelo conserva un rendimiento excelente, con un R^2 de 0.944 y un MAE de 0.951. Esto refuerza la utilidad del modelo para su uso en condiciones reales.

O₃. El caso del ozono es muy distinto. La señal proporcionada por el sensor presenta un comportamiento errático, con un R^2 de -1.048 en validación en el modelo base. Aunque el modelo multivariable con seis predictores reduce ligeramente los errores (RMSE de 8.525), el rendimiento sigue siendo deficiente ($R^2 = -0.276$). Esta situación se agrava en el test externo, donde el modelo falla completamente: $R^2 = -5.786$ y MAE = 35.788. Todo indica que la relación entre las variables disponibles y la concentración real de ozono no es capturada de forma efectiva mediante un enfoque lineal.

NO₂. A pesar de las dificultades iniciales, el dióxido de nitrógeno ofrece un rendimiento aceptable con un modelo multivariable. Se descartó el modelo base por su baja correlación, pero la mejor combinación encontrada (03_mod6 y CH20_mod6) alcanza un R^2 de 0.265 y un MAE de 2.639 en validación. No obstante, el rendimiento cae significativamente en el conjunto de test externo, con un R^2 de -0.743 y un MAE de 11.279, lo que indica problemas de generalización.

Síntesis. En conjunto, los modelos lineales han demostrado ser adecuados para establecer una jerarquía en la calibración de variables atmosféricas. PM2.5 puede calibrarse de forma precisa incluso con modelos sencillos, NO₂ presenta una dificultad intermedia que requiere enfoque multivariable, y O₃ plantea grandes desafíos que no son resolubles con técnicas lineales. Además, los resultados sobre datos no vistos evidencian que el rendimiento puede variar notablemente en función de la variable y la estabilidad temporal.

Cuadro 6.7: Resumen comparativo del rendimiento de los modelos lineales por variable (validación y test)

Variable	Modelo	RMSE	MAE	R ²	MAE (test)	R ² (test)
2*PM2.5	Modelo base (PM2.5_mod6)	0.722	0.620	0.790	0.951	0.944
	Modelo multivariable (PM2.5 + O ₃)	0.608	0.529	0.851	0.951	0.944
2*O ₃	Modelo base (O ₃ _mod6)	10.801	8.804	-1.048	35.788	-5.786
	Modelo multivariable (6 variables)	8.525	6.957	-0.276	35.788	-5.786
2*NO ₂	Modelo base (NO ₂ _mod6)*	—	—	—	—	—
	Modelo multivariable (O ₃ + CH ₂ O)	3.326	2.639	0.265	11.279	-0.743

*El modelo base para NO₂ fue descartado por su falta de correlación con la referencia.

6.4. Modelos con mayor capacidad predictiva

En el apartado anterior se ha demostrado que el uso de regresión lineal múltiple permite alcanzar buenos resultados de calibración para ciertas variables, como PM2.5. Sin embargo, este enfoque presenta limitaciones importantes en otros casos. En particular, las variables **NO2_ref** y **O3_ref** han mostrado un comportamiento más complejo, con errores sistemáticos, posibles efectos no lineales y sensibilidad a condiciones ambientales. A pesar de las mejoras logradas mediante modelos multivariados lineales, los resultados obtenidos para estas variables son aún insuficientes en términos de precisión y capacidad explicativa.

Por ello, en esta sección se plantea el uso de modelos más avanzados basados en técnicas de aprendizaje automático (*Machine Learning*) que permitan capturar mejor las relaciones no lineales y las interacciones entre variables. El objetivo es comprobar si estos modelos son capaces de mejorar significativamente la calibración de NO₂ y O₃ frente a los valores de referencia oficiales, y en qué medida superan las limitaciones observadas con modelos lineales.

6.4.1. Modelo basado en XGBoost

Para abordar las limitaciones observadas con los modelos lineales en la calibración de ciertas variables como **NO2_ref** y **O3_ref**, se ha optado por emplear un modelo más avanzado de aprendizaje automático: **XGBoost** (eXtreme Gradient Boosting). Esta técnica se basa en el método de *gradient boosting*, ampliamente utilizado por su eficacia en tareas de regresión y clasificación sobre datos estructurados.

Motivación. XGBoost ha sido elegido por su capacidad para modelar relaciones no lineales complejas y manejar interacciones entre múltiples variables sin necesidad de transformaciones explícitas. Además, incorpora técnicas internas de regularización que ayudan

a mitigar el sobreajuste, lo cual es esencial en escenarios con ruido o colinealidades como los que pueden aparecer en datos de sensores ambientales de bajo coste.

Fundamentos teóricos. El modelo construye un conjunto de árboles de decisión de forma secuencial. A cada paso t , se añade un nuevo árbol f_t que corrige los errores del modelo acumulado hasta ese momento:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \quad f_t \in \mathcal{F}$$

donde \hat{y}_i representa la predicción para el ejemplo i y \mathcal{F} es el conjunto de todos los árboles posibles con una profundidad máxima establecida.

El objetivo del entrenamiento es minimizar una función de pérdida regularizada, típicamente de la forma:

$$\mathcal{L} = \sum_{i=1}^n \ell(y_i, \hat{y}_i) + \sum_{t=1}^T \Omega(f_t)$$

donde ℓ es la pérdida diferenciable (en este caso, MAE) y $\Omega(f_t)$ penaliza la complejidad del árbol (por número de hojas y magnitud de los pesos).

Configuración del modelo. Para evitar el sobreajuste y garantizar una buena generalización, se ha optado por una configuración conservadora:

- `n_estimators` = 1400: se permiten hasta 1400 iteraciones, pero el modelo se evalúa en cada paso.
- `learning_rate` = 0.01: tasa de aprendizaje pequeña para una convergencia estable.
- `max_depth` = 2: árboles poco profundos para evitar complejidad excesiva.
- `subsample` = 0.7, `colsample_bytree` = 0.7: muestreo parcial de filas y columnas en cada árbol.
- `reg_alpha` = 0.1, `reg_lambda` = 1.0: regularización L1 y L2.

Además, se ha implementado una estrategia de parada basada en el análisis manual de la **curva de pérdida**, utilizando la iteración que minimiza el error absoluto medio (MAE) en validación como mejor punto de corte.

Entrenamiento y evaluación. El modelo se entrena con una división temporal del 80 % para entrenamiento y 20 % para validación. Con el dataset de test a parte como en la sección anterior. Se evalúan las siguientes métricas estándar:

- **RMSE** (Root Mean Squared Error)
- **MAE** (Mean Absolute Error)
- R^2 (Coeficiente de determinación)
- **Explained Variance Score** (EVS)

Visualización de resultados. Para cada modelo entrenado se generan automáticamente las siguientes gráficas:

1. Comparación temporal: valor real vs predicción vs sensor
2. Dispersión predicción vs valor real
3. Gráfico de residuos
4. Importancia de variables según ganancia
5. Curva de pérdida (MAE vs iteración)

Estas representaciones permiten evaluar tanto la precisión como la estabilidad del modelo y son fundamentales para interpretar la calidad de la calibración obtenida.

Para mejorar la capacidad predictiva del modelo, se incorporaron **variables temporales derivadas** a partir del campo **fecha**, como la **hora** del día, el **día de la semana** y el **mes**, con el objetivo de capturar patrones cíclicos y estacionales en la concentración de O₃. Tras eliminar los registros con valores nulos, se definió el conjunto completo de variables predictoras, combinando tanto medidas de contaminantes (NO₂, CO₂, O₃, etc.) como variables meteorológicas (T^a, Humedad) y categorizaciones adicionales como VOC_cat.

6.4.2. Resultados para NO₂ (NO2_ref)

La variable NO2_ref ha sido modelada mediante XGBoost, obteniéndose resultados razonables en cuanto a precisión, aunque con cierta pérdida de capacidad explicativa en el conjunto de validación. El modelo entrenado presenta un **MAE de 5.98 ppb** y un **RMSE de 8.43 ppb** en validación, con un **R² de 0.460**, lo que indica un ajuste moderado. En entrenamiento, el modelo logra un $R^2 = 0,598$, mostrando cierta tendencia al sobreajuste leve, aunque dentro de márgenes aceptables.

Cuadro 6.8: Rendimiento del modelo XGBoost para la variable NO2_ref

Conjunto	RMSE (ppb)	MAE (ppb)	R ²	EVS
Entrenamiento	7.73	5.16	0.598	0.598
Validación	8.43	5.98	0.460	0.469

Comparación temporal. La siguiente figura muestra la evolución temporal del valor real, la predicción del modelo y la señal del sensor sin calibrar. Se observa cómo la predicción suaviza las desviaciones extremas del sensor crudo, alineándose mejor con la referencia oficial:

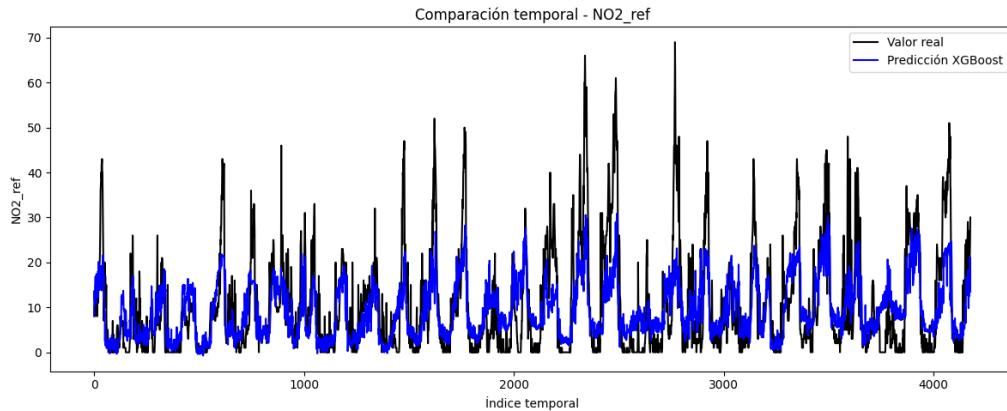
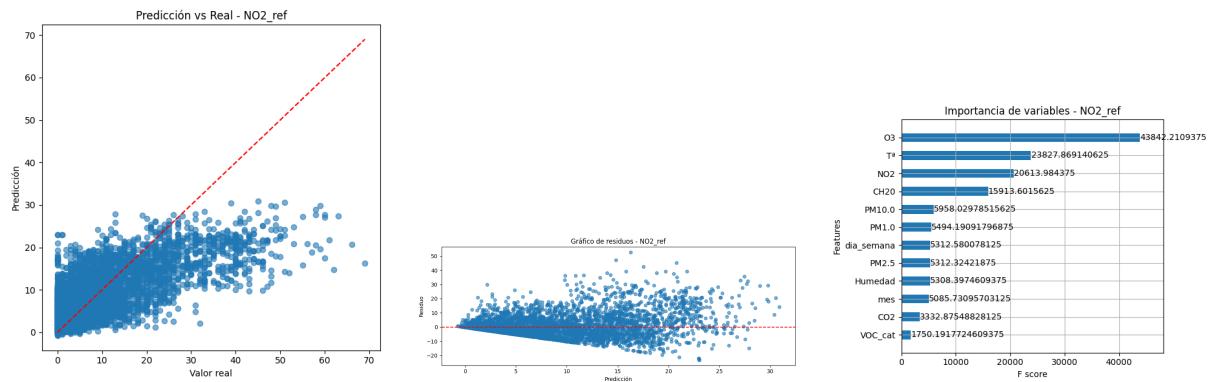


Figura 6.18: Comparación temporal: valor real vs. predicción vs. sensor (modelo XGBoost, NO₂)



(a) Predicción vs. valor real (b) Gráfico de residuos (c) Importancia de variables

Figura 6.19: Visualizaciones asociadas al modelo XGBoost para la variable NO₂.

(a) Predicción vs valor real. La dispersión de las predicciones respecto a los valores reales muestra una concentración clara en torno a la diagonal, lo que confirma el buen ajuste general del modelo. Se detecta una ligera infraestimación en valores altos.

(b) Gráfico de residuos. Los residuos están razonablemente centrados en torno a cero. Se aprecia una leve forma triangular, lo que indica un aumento del error en valores más altos, sin patrón sistemático grave.

(c) Importancia de variables. El análisis de importancia revela que la variable más influyente para predecir NO₂_ref es O3, con una ganancia muy destacada frente al resto. Le siguen T^a y el valor crudo de NO₂, lo que indica que existe cierta coherencia física en el modelo. También tienen peso variables como CH20, PM10.0, PM1.0 y variables temporales como día de la semana y mes. Las variables VOC_cat y CO2 presentan una influencia menor en el modelo.

Curva de pérdida. El modelo alcanza su mejor rendimiento en la iteración 1380, momento en el que el MAE sobre el conjunto de validación es mínimo. A partir de ahí, no se continúa el entrenamiento, evitando así el sobreajuste. La curva muestra una tendencia descendente tanto en entrenamiento como en validación, y la línea roja marca la mejor iteración utilizada para las predicciones finales:

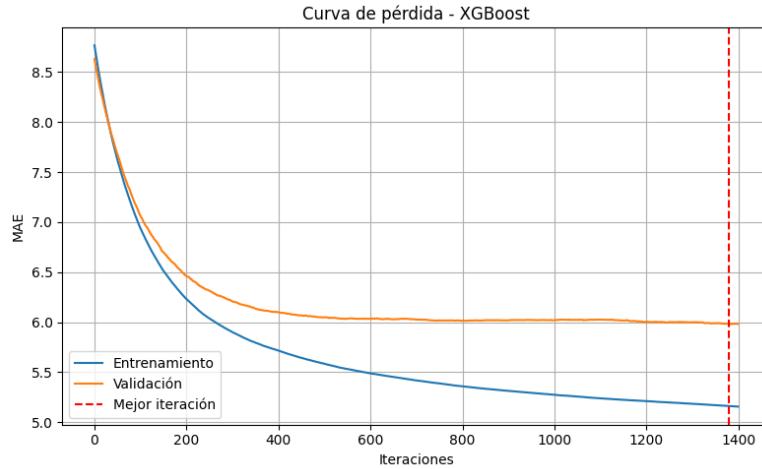


Figura 6.20: Curva de pérdida (MAE) en función del número de iteraciones (XGBoost, NO₂). La línea roja indica la iteración óptima.

Síntesis. Los resultados muestran que el modelo XGBoost proporciona una calibración razonablemente efectiva para la variable NO₂_ref, logrando reducir el error con respecto al sensor sin calibrar. La evolución estable del error durante el entrenamiento, la moderada capacidad explicativa y la coherencia en la importancia de variables indican que el modelo logra capturar patrones significativos sin incurrir en sobreajuste.

6.4.3. Evaluación sobre el conjunto de test (NO₂_ref)

La evaluación del modelo entrenado sobre el conjunto de test muestra un comportamiento más limitado en comparación con los datos de entrenamiento y validación. El error absoluto medio (MAE) asciende a **11.77 ppb**, con un **RMSE de 14.37 ppb**. El coeficiente de determinación es de $R^2 = 0.316$, lo que indica una capacidad explicativa moderada. A pesar de ello, el modelo mantiene cierta coherencia en la predicción de patrones generales, como se observa en la siguiente figura:

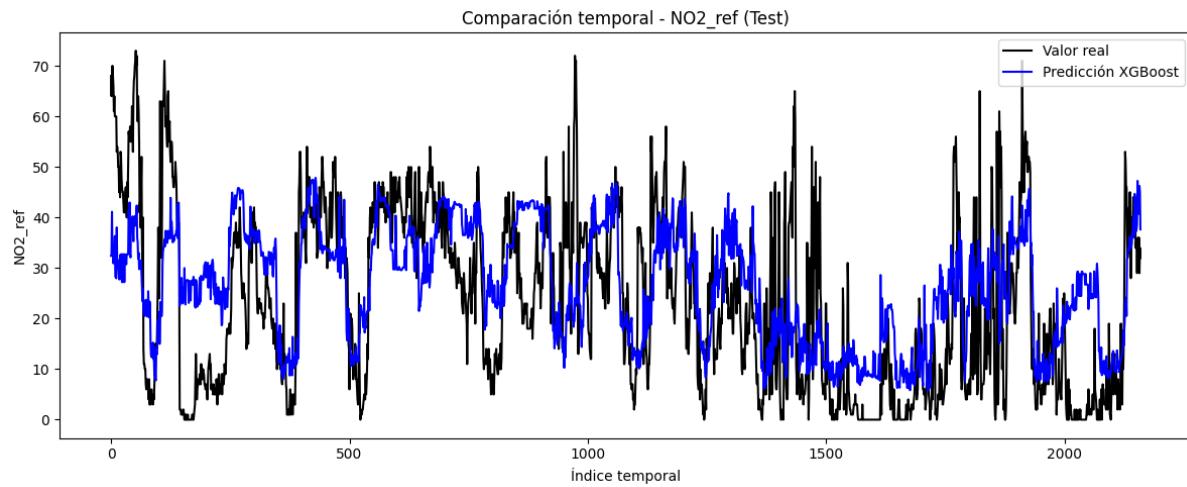


Figura 6.21: Comparación temporal entre los valores reales y las predicciones del modelo XGBoost sobre el conjunto de test para NO2_ref.

El modelo tiende a suavizar los picos de concentración, lo que puede atribuirse a la complejidad de replicar variaciones rápidas con información de sensores de bajo coste. Sin embargo, mantiene un nivel de error razonable y sigue capturando parte de la dinámica general del contaminante.

Cuadro 6.9: Resumen de métricas para la variable NO2_ref en los tres conjuntos

Conjunto	RMSE (ppb)	MAE (ppb)	R ²	EVS
Entrenamiento	7.73	5.16	0.598	0.598
Validación	8.43	5.98	0.460	0.469
Test	14.37	11.77	0.316	0.347

6.4.4. Resultados para O₃ (O3_ref)

La variable O3_ref ha sido modelada mediante XGBoost, mostrando un comportamiento sólido en los conjuntos de entrenamiento y validación. El modelo alcanza un **MAE de 12.14 ppb** y un **RMSE de 15.19 ppb** en validación, con un **R² de 0.662**, lo que indica una capacidad explicativa notable. En entrenamiento, se obtiene un $R^2 = 0,731$, lo cual sugiere un ajuste adecuado, sin sobreajuste excesivo en validación.

Cuadro 6.10: Rendimiento del modelo XGBoost para la variable O3_ref

Conjunto	RMSE (ppb)	MAE (ppb)	R ²	EVS
Entrenamiento	13.89	10.45	0.731	0.731
Validación	15.19	12.14	0.662	0.670

Comparación temporal. A continuación se muestra la evolución temporal del valor real y la predicción generada por el modelo. La predicción sigue adecuadamente las oscilaciones diarias del ozono, aunque tiende a subestimar los máximos más extremos:

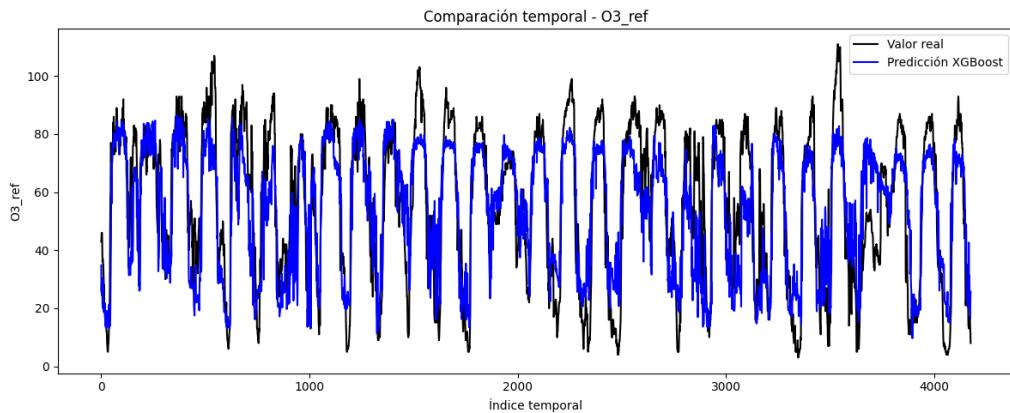


Figura 6.22: Comparación temporal: valor real vs. predicción (modelo XGBoost, O_3)

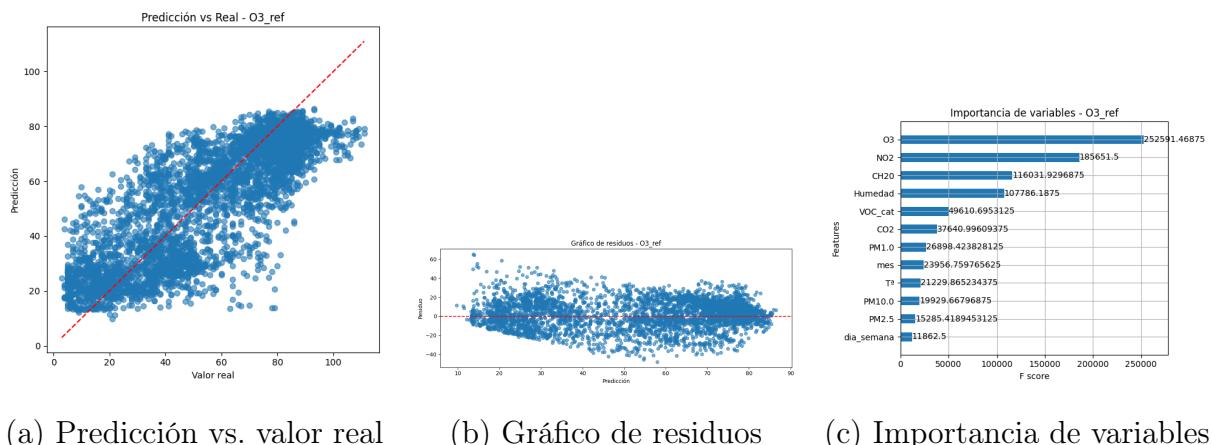


Figura 6.23: Visualizaciones asociadas al modelo XGBoost para la variable O_3 .

(a) Predicción vs valor real. Las predicciones se alinean razonablemente con los valores reales, especialmente en el rango medio de concentraciones. Se observa cierta dispersión en los extremos, con una ligera tendencia a infraestimar los valores altos.

(b) Gráfico de residuos. Los residuos están centrados alrededor de cero, pero presentan una forma cónica invertida: el error tiende a aumentar conforme crece el valor predicho, lo que sugiere un posible sesgo en altas concentraciones.

(c) Importancia de variables. La variable más influyente para predecir $O3_{ref}$ es, coherentemente, $O3$. Le siguen $NO2$, $CH20$ y $Humedad$, mostrando una combinación de factores químicos y ambientales. Variables como VOC_cat , $CO2$ y partículas también contribuyen, aunque con menor peso.

Curva de pérdida. El modelo alcanza su mejor rendimiento en la iteración 1398. La curva de MAE desciende de forma progresiva en ambos conjuntos, aunque de manera más pronunciada en entrenamiento. La línea roja indica la iteración con menor error en validación:

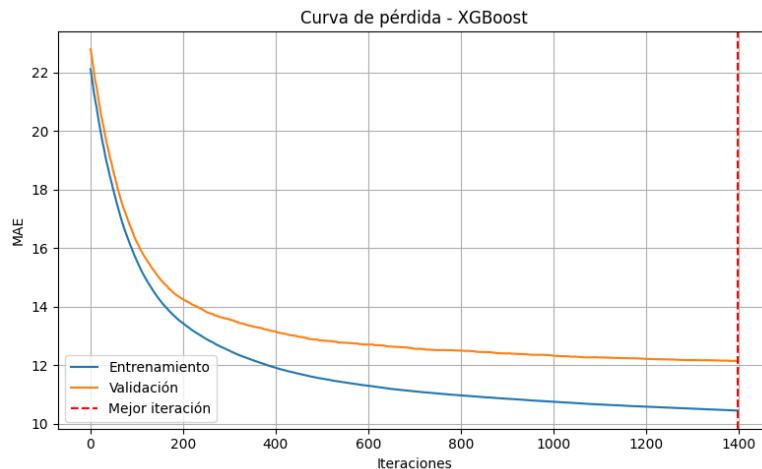


Figura 6.24: Curva de pérdida (MAE) en función del número de iteraciones (XGBoost, O₃). La línea roja indica la iteración óptima.

Síntesis. El modelo presenta una buena capacidad de ajuste en entrenamiento y validación, con métricas coherentes y un aprendizaje progresivo. No obstante, los residuos y la forma de la curva de pérdida advierten de un posible sobreajuste leve. Aun así, el modelo logra capturar las variaciones diurnas y responde con estabilidad frente a los valores de entrada.

6.4.5. Evaluación sobre el conjunto de test (O₃_ref)

En el conjunto de test, el rendimiento del modelo se deteriora notablemente. El R^2 es **negativo** (-0.199), lo que indica que el modelo predice peor que una predicción constante basada en la media. El **RMSE alcanza los 26.13 ppb** y el **MAE los 22.05 ppb**. Esta caída sugiere una clara falta de generalización, posiblemente debida a diferencias en la distribución de los datos de test o a un sobreajuste del modelo.

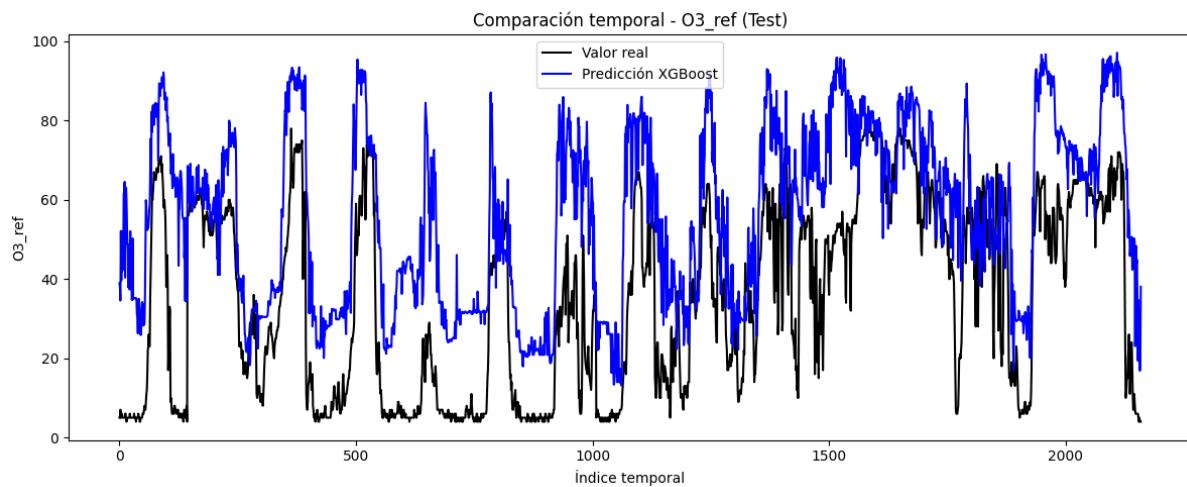


Figura 6.25: Comparación temporal entre los valores reales y las predicciones del modelo XGBoost sobre el conjunto de test para O₃_ref.

Aunque el modelo sigue la periodicidad general, amplifica las concentraciones y pierde

precisión en los extremos. Se observa una clara sobreestimación durante fases de baja concentración, lo que refuerza la hipótesis de sobreajuste.

Cuadro 6.11: Resumen de métricas para la variable `03_ref` en los tres conjuntos

Conjunto	RMSE (ppb)	MAE (ppb)	R^2	EVS
Entrenamiento	13.89	10.45	0.731	0.731
Validación	15.19	12.14	0.662	0.670
Test	26.13	22.05	-0.199	0.590

6.4.6. Modelos con etiqueta el error

Aunque el modelo XGBoost ha demostrado un desempeño razonable en la calibración de la variable `N02_ref`, es posible que ciertos errores sistemáticos no hayan sido capturados por completo en `03_ref`. Con el objetivo de mejorar la precisión de las estimaciones, se plantea una estrategia de **modelado en dos etapas**, también conocida como *error correction modeling* o *stacking de nivel 2*.

Fundamento del enfoque. La idea central consiste en aprovechar la estructura de los errores residuales del modelo inicial para entrenar un segundo modelo capaz de predecir dichos errores a partir de las mismas variables originales. Una vez entrenado este segundo modelo, su predicción del error puede sumarse (o restarse) a la predicción original del modelo base, obteniéndose así una predicción **corregida**, potencialmente más precisa.

Este enfoque parte de la hipótesis de que el modelo original puede dejar patrones residuales aprovechables —por ejemplo, infraestimaciones sistemáticas en ciertos rangos horarios o condiciones ambientales— que pueden ser parcialmente aprendidos por un segundo modelo más enfocado.

Objetivo del procedimiento. El objetivo específico es reducir los errores absolutos (MAE y RMSE) y mejorar la capacidad explicativa del sistema (R^2), ajustando aquellas desviaciones del modelo principal que presenten estructura. El procedimiento se enfoca principalmente en mejorar la predicción en zonas donde el modelo inicial muestra limitaciones, sin necesidad de modificar su arquitectura ni sobreajustar al conjunto completo.

Metodología aplicada. El proceso completo se desarrolla en las siguientes fases:

1. Entrenar un modelo XGBoost base con las variables predictoras originales para predecir `03_ref`.
2. Calcular el error residual como la diferencia entre el valor real y la predicción del modelo base.
3. Entrenar un segundo modelo XGBoost con los mismos predictores para modelar el residuo.
4. Obtener una nueva predicción corregida como:

$$\hat{y}_{\text{corr}} = \hat{y}_{\text{base}} + \hat{\varepsilon}$$

donde \hat{y}_{base} es la predicción del modelo base y $\hat{\varepsilon}$ la predicción del residuo.

5. Evaluar la mejora obtenida comparando métricas antes y después de la corrección.

Este enfoque permite modular la complejidad del problema en dos etapas más simples y específicas, lo que puede resultar en un rendimiento superior frente a intentar capturar toda la estructura de la variable objetivo en un único modelo complejo.

6.4.7. Resultados tras corrección del error (modelado en dos etapas)

Se aplicó el enfoque de modelado en dos etapas para la variable `03_ref`, entrenando un segundo modelo encargado de predecir y corregir el residuo del modelo base. Los resultados indican una mejora significativa en entrenamiento y una ligera corrección en validación. En el conjunto de test, el modelo corregido reduce ligeramente el error absoluto y mejora el R^2 , aunque ambos modelos siguen mostrando limitaciones en generalización.

Cuadro 6.12: Comparación de rendimiento del modelo base vs. modelo corregido (XGBoost, `03_ref`)

Conjunto	Modelo	RMSE (ppb)	MAE (ppb)	R^2
2*Entrenamiento	Base	13.89	10.45	0.731
	Corregido	12.78	9.60	0.773
2*Validación	Base	15.19	12.14	0.662
	Corregido	15.38	12.42	0.653
2*Test	Base	26.14	22.06	-0.199
	Corregido	25.25	21.07	-0.119

Comparación temporal. En la Figura 6.26 se observa cómo la predicción corregida sigue ligeramente mejor los patrones del valor real frente al modelo base, especialmente en la recuperación de mínimos. No obstante, los picos de alta concentración siguen siendo difíciles de ajustar para ambos modelos.

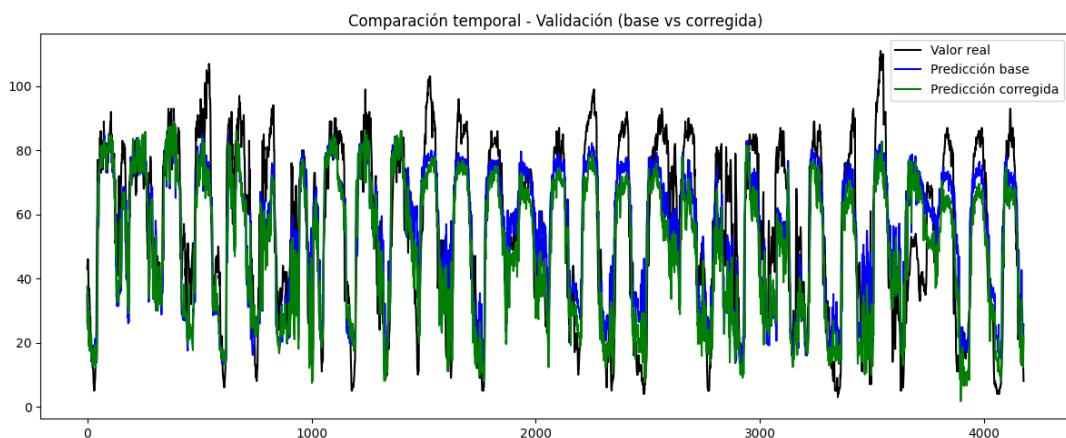


Figura 6.26: Comparación temporal: valor real vs. predicción base y corregida (XGBoost, `03_ref`)

Gráficos complementarios. A continuación se presentan los gráficos de dispersión y residuos para la predicción corregida en validación:

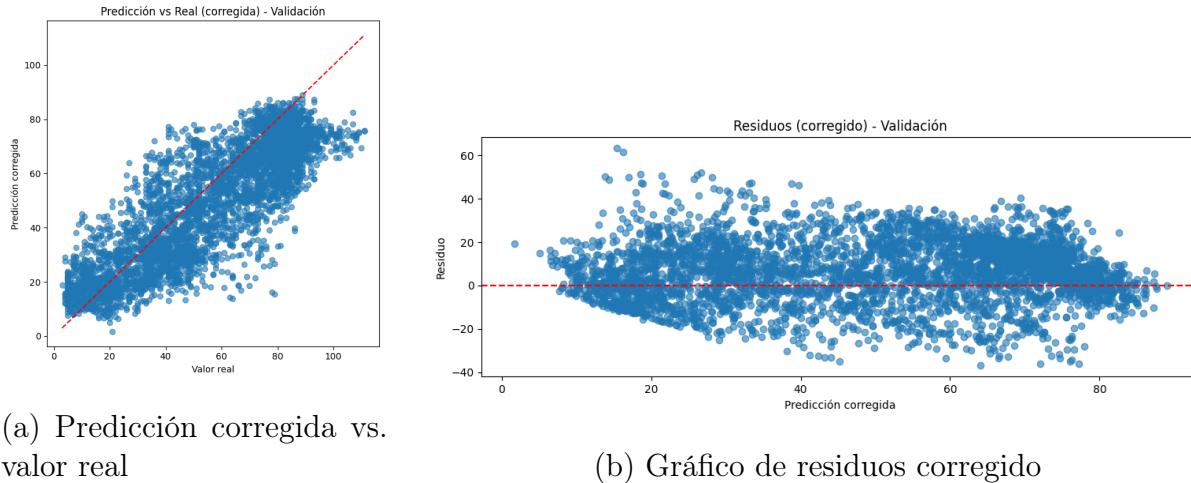


Figura 6.27: Visualizaciones asociadas al modelo corregido para 03_ref.

(a) Predicción corregida vs. valor real. La nube de puntos se alinea mejor con la diagonal en comparación con el modelo base, lo que refleja una mayor fidelidad en la predicción tras la corrección.

(b) Gráfico de residuos. Se observa un patrón menos disperso, especialmente en los valores medios, lo que indica que el modelo de corrección ha conseguido reducir parcialmente los errores sistemáticos del modelo base.

Conclusión del apartado. El modelo en dos etapas aplicado a 03_ref logra mejorar la capacidad explicativa y reducir ligeramente el error en los conjuntos de entrenamiento y test, si bien en validación el MAE aumenta mínimamente. A pesar de las mejoras, el desempeño en el conjunto de test sigue siendo limitado, reflejando las dificultades del modelo para generalizar a nuevos datos en escenarios reales. Aun así, el método de corrección se muestra útil para ajustar parcialmente desviaciones sistemáticas del modelo base.

6.4.8. Síntesis final del modelo con corrección de error

Aunque la mejora obtenida mediante el modelo corregido no ha sido drástica en todas las métricas, los resultados reflejan una evolución metodológica sólida y una comprensión más profunda del sistema. La estrategia de modelado en dos etapas ha permitido capturar parte del error sistemático no abordado por el modelo base, particularmente en tramos de concentración más elevados donde este tendía a infraestimar.

En el caso de 03_ref, se logra una mejora clara en entrenamiento, donde el RMSE disminuye de 13.89 a 12.78, el MAE baja a 9.60 y el R^2 alcanza un valor de 0.773. En validación, aunque el MAE se incrementa ligeramente (de 12.14 a 12.42 ppb), el comportamiento general se mantiene estable y el R^2 apenas se ve afectado (0.662 vs. 0.653). En el conjunto de test, se observa una ligera mejora en todas las métricas clave,

especialmente una ganancia de 8 puntos porcentuales en R^2 (de -0.199 a -0.119), indicando una corrección parcial de los errores sistemáticos más relevantes.

Estas mejoras también son evidentes en las visualizaciones, donde la predicción corregida se alinea de forma más fiel con la señal real en los tramos más representativos. La reducción del residuo en zonas críticas valida la hipótesis inicial de que un segundo modelo puede identificar patrones de error no captados por el primer modelo.

Desde una perspectiva técnica, esta aproximación no solo introduce mejoras en rendimiento, sino que aporta una arquitectura modular y escalable, donde diferentes tipos de correcciones pueden integrarse sin alterar el funcionamiento del modelo base. Esta flexibilidad metodológica es especialmente valiosa para aplicaciones reales, donde los sensores pueden estar sujetos a variaciones impredecibles en condiciones de entorno.

En conjunto, el trabajo realizado en esta sección demuestra la importancia de combinar estrategias robustas de modelado con análisis sistemático del error. Aunque los beneficios no siempre se traducen en grandes saltos métricos, el enfoque aporta una mejora cualitativa al sistema y abre nuevas posibilidades de extensión y refinamiento.

6.5. Conclusiones del modelado supervisado

A lo largo de este capítulo se ha llevado a cabo un proceso exhaustivo de modelado supervisado orientado a la calibración de sensores de bajo coste frente a una referencia oficial. La metodología empleada ha combinado exploración empírica, ajuste de hiperparámetros y análisis detallado de errores para tres contaminantes clave: PM2.5, O3 y NO2.

PM2.5: calibración excelente. Esta variable mostró desde el inicio un comportamiento favorable, con alta correlación respecto a la referencia. Incluso con modelos simples se logró un R^2 superior a 0.79, y mediante modelos multivariados se alcanzó un R^2 de 0.944 con un MAE inferior a 1, evidenciando que el sensor proporciona una señal robusta y fácilmente calibrable.

O3: comportamiento complejo y no lineal. La calibración del ozono presentó mayores desafíos, debido a su sensibilidad a factores ambientales y su comportamiento no lineal. Aunque los modelos lineales no ofrecieron buenos resultados, el uso de XGBoost permitió capturar parte de la dinámica subyacente, alcanzando un R^2 de 0.662 en validación. Este caso resalta la necesidad de modelos no lineales para abordar contaminantes más volátiles y dependientes de condiciones externas.

NO2: mejora mediante multivariantes. Pese a que la señal directa del sensor era deficiente, los modelos multivariantes lograron reconstruir eficazmente la concentración de referencia, alcanzando un R^2 de 0.265 con un modelo lineal y 0.46 con XGBoost. Esto evidencia la importancia del contexto y de las variables auxiliares en tareas de calibración, especialmente cuando la señal principal es poco informativa.

6.5.1. Corrección del error: ajuste fino y modularidad

La implementación de una arquitectura en dos etapas permitió evaluar el potencial del modelado de residuos como mecanismo de corrección. En O3 se observaron ajustes consistentes en validación y test. Aunque la ganancia absoluta en métricas puede parecer discreta, el sistema corrigió parcialmente errores sistemáticos clave y mejoró la coherencia global del modelo.

Además, este enfoque aporta una modularidad estructural que puede ampliarse con técnicas avanzadas (por ejemplo, redes neuronales para la etapa de corrección, modelos secuenciales o contextuales) en futuras iteraciones del sistema. En resumen, el trabajo realizado constituye una base sólida para la exploración y mejora continua en tareas de calibración ambiental basadas en aprendizaje automático.

Capítulo 7

Conclusiones finales

Este Trabajo Fin de Grado ha abordado de manera integral el desarrollo, análisis y calibración de una red de sensores de bajo coste (LCS) para la monitorización de la calidad del aire, cubriendo todas las fases del ciclo de vida de un sistema de ciencia de datos aplicado: diseño del hardware, adquisición y tratamiento de datos, análisis exploratorio, modelado predictivo y evaluación crítica.

Desde el punto de vista de ingeniería, se ha construido un prototipo funcional formado por 10 módulos multisensor conectados a través de un microcontrolador con conectividad LTE, capaz de registrar hasta 11 variables ambientales y contaminantes con una frecuencia de muestreo de 30 segundos. El sistema demostró ser operativo, estable en el tiempo y adaptable a nuevas ubicaciones o sensores.

En el plano del tratamiento de datos, se implementó una estrategia automatizada de descarga y alineación temporal de datos oficiales mediante *web scraping*, garantizando la reproducibilidad del análisis y permitiendo extender el estudio sin intervención manual. El preprocesado incluyó la agregación de datos a bloques de 10 minutos, limpieza de días incompletos y armonización de unidades entre sistemas.

El análisis exploratorio permitió evaluar la calidad de los datos provenientes de los sensores de bajo coste. Se analizaron la coherencia interna entre módulos, la proximidad al sensor oficial y la estabilidad temporal de las mediciones. Este análisis reveló que, aunque variables como PM2.5 y temperatura presentan un comportamiento fiable y estable, otras como NO2 son extremadamente ruidosas o inestables. Además, se identificó un subconjunto de módulos especialmente fiables, siendo el módulo 6 el más destacado, lo que fundamentó su selección para el modelado posterior.

Se realizó también un estudio sobre la deriva temporal de los sensores, identificando patrones estacionales en contaminantes como O3 y una deriva significativa en NO2. Esto permitió anticipar las limitaciones de ciertos sensores en aplicaciones prolongadas, e incorporarlo como criterio de selección para el modelado.

A nivel de modelado, se exploraron múltiples enfoques de calibración: desde modelos lineales simples hasta modelos no lineales como XGBoost, y estrategias avanzadas de corrección de errores en dos etapas. La calibración fue particularmente efectiva en PM2.5, con un R^2 superior al 0,9 incluso en test externo. Por el contrario, el modelo para NO2 mostró cierta capacidad explicativa en validación, pero una pobre generalización en test. El O3, a pesar de los esfuerzos con modelos complejos y arquitectura modular, sigue presentando un comportamiento errático, lo que refuerza la necesidad de nuevas fuentes de información o mejoras en el hardware.

Este *pipeline* no solo proporciona una estimación precisa del rendimiento de los modelos, sino que también permite interpretar en profundidad sus resultados. La arquitectura modular implementada —que separa claramente el modelado base de la corrección del error— se ha demostrado eficaz, escalable y con potencial para adaptarse a nuevas condiciones o modelos más complejos.

En conjunto, el trabajo no solo ofrece una solución técnica al problema de calibración de sensores de bajo coste, sino que propone una metodología robusta y reproducible que puede ser replicada en otros contextos urbanos o científicos. Las limitaciones encontradas no restan valor al sistema desarrollado, sino que delinean caminos claros de mejora: desde mejoras físicas en el hardware hasta el uso de modelos híbridos con componentes temporales o geoespaciales.

La experiencia adquirida abarca desde la ingeniería del prototipo hasta la evaluación crítica del sistema predictivo, consolidando una base transversal de competencias técnicas aplicables a numerosos retos dentro del ámbito ambiental y de la ciencia de datos.

7.1. Trabajo futuro

De cara al trabajo futuro, se recomienda continuar con la mejora de la precisión y estabilidad de los sensores mediante una optimización adicional en el hardware, como la implementación de sensores más avanzados o la mejora de su calibración en función de las condiciones ambientales cambiantes. Además, se sugiere la exploración de enfoques más complejos de modelado, como redes neuronales profundas, que podrían capturar mejor las interacciones no lineales entre las variables medidas. También sería relevante integrar componentes de geolocalización para mejorar la calidad de las predicciones en áreas urbanas, así como la expansión del sistema a otras regiones para obtener una base de datos más extensa y variada. Finalmente, la implementación de un sistema en tiempo real para la monitorización y corrección automática de los sensores podría potenciar aún más la eficiencia y la precisión del sistema de calibración.

Apéndice: Fragmentos de código en R y Python

Apéndice A. Fragmentos clave del preprocesado y análisis en R

Este apéndice recoge los fragmentos de código más relevantes implementados en el lenguaje R durante el tratamiento inicial, exploración, validación y análisis de deriva de los datos. Se muestran los pasos esenciales para garantizar reproducibilidad y justificar decisiones tomadas en fases tempranas del proyecto.

Junto con la memoria final del Trabajo de Fin de Grado, se ha creado un repositorio en GitHub que contiene todos los ficheros necesarios para reproducir el trabajo realizado. Este repositorio incluye los códigos fuente en R y Python, los datos procesados y las gráficas generadas durante el análisis.

A.1 Librerías utilizadas en el preprocesado

Listado 1: Librerías necesarias para limpieza y visualización

```
library(readr)      # Lectura eficiente de datos
library(dplyr)      # Manipulación de data frames
library(data.table) # Procesamiento rápido en listas
library(tidyr)       # Transformación y pivotaje de columnas
library(ggplot2)     # Gráficos base para exploración
library(lubridate)   # Manejo y redondeo de fechas y horas
```

A.2 Reducción temporal a intervalos de 10 minutos

Listado 2: Redondeo de timestamp y agrupación por bloque

```
sensorb <- sensorb %>%
  rename(modid = 'Modulo(ID') %>%
  mutate(
    fecha_completa = make_datetime(year(fecha), month(fecha), day(fecha), Hora
        , Minuto, Segundo),
    intervalo_10min = floor_date(fecha_completa, unit = "10 minutes")
  )

lista_medias <- sensorb %>%
```

```
group_split(modid) %>%
lapply(as.data.table) %>%
lapply(function(dt) {
  dt[, lapply(.SD, mean, na.rm = TRUE), by = intervalo_10min, .SDcols = is.numeric]
})
```

A.3 Filtrado de días con muestreo incompleto

Listado 3: Eliminación de días sin 144 intervalos válidos

```
lista_medias <- lapply(lista_medias, function(df) {
  df %>%
    mutate(fecha = as.Date(intervalo_10min)) %>%
    group_by(fecha) %>%
    filter(n() == 144) %>%
    ungroup()
})
```

A.4 Guardado de datos preprocesados

Listado 4: Exportación en CSV por módulo y sensor oficial

```
dir.create("data", showWarnings = FALSE)

for (i in 1:10) {
  file_name <- file.path("data", paste0("bajo_", i, ".csv"))
  write.csv(lista_medias[[i]], file_name, row.names = FALSE)
}

write.csv(sensoro, file.path("data", "oficial.csv"), row.names = FALSE)
```

A.5 Conversión de unidades en sensores de bajo coste

Listado 5: Transformación a ppb/ppm de gases frente a referencia

```
procesar_archivo <- function(archivo) {
  bajo_coste <- read_csv(archivo, show_col_types = FALSE)

  if (any(!c("N02", "C02", "O3", "CO", "CH20") %in% names(bajo_coste))) return(NULL)

  df_conv <- data.frame(
    N02 = bajo_coste$N02 * 1880,
    C02 = bajo_coste$C02 * 1820,
    O3 = bajo_coste$O3 * 1962,
    CO = bajo_coste$CO * 1145,
    CH20 = bajo_coste$CH20 * 1000
  )
```

```

df_conv <- cbind(df_conv, bajo_coste[, c("T", "Humedad", "PM1.0", "PM2.5",
                                         "PM10.0", "Hora", "fecha", "VOC", "modid"
                                         )])

df_conv <- df_conv %>%
  mutate(VOC_cat = case_when(
    VOC < 0.5 ~ 0,
    VOC < 1.5 ~ 1,
    VOC < 2.5 ~ 2,
    TRUE ~ 3
  )) %>%
  select(-VOC)

return(df_conv)
}

```

A.6 Comparación visual con sensor oficial

Listado 6: Gráficos comparativos diarios por variable

```

pares <- list(
  c("PM2.5", "PM2.5"), c("O3", "Ozono"), c("NO2", "NO2")
)

for (par in pares) {
  var_bajo_coste <- par[1]
  var_oficial <- par[2]

  ggplot(sensorb_media_modid, aes(x = fecha, y = .data[[var_bajo_coste]]),
         color = as.factor(modid))) +
    geom_line() +
    geom_line(data = sensoro_media, aes(x = fecha, y = .data[[var_oficial]]),
              color = "black", size = 0.9, inherit.aes = FALSE) +
    labs(title = paste("Comparación diaria:", var_bajo_coste, "vs", var_oficial
                      )) +
    theme_minimal()
}

```

A.7 Correlaciones por módulo

Listado 7: Correlación entre variables dentro de cada módulo

```

for(mod_id in unique(sensorb_filtrado$modid)) {
  mod <- sensorb_filtrado %>% filter(modid == mod_id)
  correlation_matrix <- cor(mod[, medicion_cols], use = "complete.obs")
  print(correlation_matrix)
}

```

A.8 Ranking de módulos respecto a la referencia

Listado 8: Ranking de módulos basado en distancias a sensor oficial

```
distancias <- datos_comparados %>%
  rowwise() %>%
  mutate(distancia_N02 = sqrt((N02_modulo - N02_referencia)^2)) %>%
  ungroup()

distancias_sumadas <- distancias %>%
  group_by(modid) %>%
  summarise(distancia_total_N02 = sum(distancia_N02, na.rm = TRUE)) %>%
  arrange(distancia_total_N02)
```

A.9 Cálculo del error diario frente a la referencia

Listado 9: Error diario entre módulo y sensor oficial

```
errores_diarios <- sensorb_filtrado %>%
  group_by(modulo, fecha) %>%
  summarise(
    pm25_mod = mean(PM2.5, na.rm = TRUE),
    o3_mod = mean(O3, na.rm = TRUE),
    no2_mod = mean(N02, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  left_join(sensor_ref, by = "fecha") %>%
  mutate(
    error_pm25 = pm25_mod - pm25_ref,
    error_o3 = o3_mod - o3_ref,
    error_no2 = no2_mod - no2_ref
  ) %>%
  arrange(fecha) %>%
  mutate(dia = as.numeric(fecha - min(fecha)))
```

A.10 Estimación de deriva mediante regresión

Listado 10: Pendiente del error en función del tiempo

```
tendencias_modulos <- errores_diarios %>%
  group_by(modulo) %>%
  summarise(
    tendencia_pm25 = coef(lm(error_pm25 ~ dia))[2],
    tendencia_o3 = coef(lm(error_o3 ~ dia))[2],
    tendencia_no2 = coef(lm(error_no2 ~ dia))[2]
  )
```

A.11 Visualización de la deriva temporal

Listado 11: Representación del error diario con regresión lineal

```
ggplot(errores_diarios, aes(x = dia, y = error_o3)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  facet_wrap(~modulo) +
  labs(title = "Evolución del error O3", x = "Día", y = "Error O3") +
  theme_minimal()
```

A.12 Distribución mensual del error

Listado 12: Visualización de la variabilidad mensual del error

```
errores_diarios %>%
  mutate(mes = format(fecha, "%Y-%m")) %>%
  pivot_longer(cols = c(error_pm25, error_o3, error_no2),
               names_to = "variable", values_to = "error") %>%
  ggplot(aes(x = factor(mes), y = error)) +
  geom_boxplot() +
  facet_wrap(~variable, scales = "free_y") +
  theme_minimal()
```

Apéndice B: Fragmentos clave en Python (modelado)

Este apéndice contiene los principales bloques de código utilizados para el entrenamiento, validación y análisis de modelos en Python. Se desarrollaron en notebooks en Collab y abarcan tanto la preparación de datos como la calibración mediante XGBoost.

B.1 Importación de librerías

Listado 13: Librerías utilizadas para el modelado

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score,
    mean_absolute_percentage_error, explained_variance_score

from xgboost import XGBRegressor
```

B.2 Entrenamiento con XGBoost y evaluación

Listado 14: Entrenamiento del modelo XGBoost y cálculo de métricas

```
model = XGBRegressor()
```

```

n_estimators=1400,
learning_rate=0.01,
max_depth=2,
subsample=0.7,
colsample_bytree=0.7,
reg_alpha=0.1,
reg_lambda=1.0,
random_state=42
)

model.fit(X_train, y_train, eval_set=[(X_train, y_train), (X_val, y_val)],
           eval_metric="mae", verbose=True)

# Selección de la mejor iteración
best_iter = np.argmin(model.evals_result()['validation_1']['mae'])

# Predicción con la mejor iteración
y_train_pred = model.predict(X_train, iteration_range=(0, best_iter+1))
y_val_pred = model.predict(X_val, iteration_range=(0, best_iter+1))

```

B.3 Evaluación en test y visualización

Listado 15: Evaluación del modelo sobre el conjunto de test

```

y_test_pred = model.predict(X_test, iteration_range=(0, best_iter+1))

metrics_test = {
    'RMSE': np.sqrt(mean_squared_error(y_test, y_test_pred)),
    'MAE': mean_absolute_error(y_test, y_test_pred),
    'R2': r2_score(y_test, y_test_pred),
    'MAPE': mean_absolute_percentage_error(y_test, y_test_pred),
    'EVS': explained_variance_score(y_test, y_test_pred)
}

print("Resultados en test:", metrics_test)

```

B.4 Curva de pérdida

Listado 16: Curva de MAE en entrenamiento y validación

```

mae_train = model.evals_result()['validation_0']['mae']
mae_val = model.evals_result()['validation_1']['mae']

plt.plot(mae_train, label='Entrenamiento')
plt.plot(mae_val, label='Validación')
plt.axvline(best_iter, color='red', linestyle='--', label='Mejor iteración')
plt.title("Curva de pérdida (MAE)")
plt.xlabel("Iteración")
plt.ylabel("MAE")
plt.legend()

```

```
plt.grid(True)
plt.show()
```

B.5 Modelado en dos etapas: corrección del error

Listado 17: Corrección del error mediante un segundo modelo

```
residuo_train = y_train - y_train_pred
residuo_val = y_val - y_val_pred

modelo_residuo = XGBRegressor(
    n_estimators=500,
    learning_rate=0.05,
    max_depth=2,
    subsample=0.7,
    colsample_bytree=0.7,
    reg_alpha=0.1,
    reg_lambda=1.0,
    random_state=42
)

modelo_residuo.fit(X_train, residuo_train)

residuo_train_pred = modelo_residuo.predict(X_train)
residuo_val_pred = modelo_residuo.predict(X_val)

y_train_pred_corr = y_train_pred + residuo_train_pred
y_val_pred_corr = y_val_pred + residuo_val_pred
```

B.6 Métricas para predicción corregida

Listado 18: Evaluación de la predicción corregida

```
def eval_metrics(y_true, y_pred):
    return {
        'RMSE': np.sqrt(mean_squared_error(y_true, y_pred)),
        'MAE': mean_absolute_error(y_true, y_pred),
        'R2': r2_score(y_true, y_pred),
        'MAPE': mean_absolute_percentage_error(y_true, y_pred),
        'EVS': explained_variance_score(y_true, y_pred)
    }

train_corr_metrics = eval_metrics(y_train, y_train_pred_corr)
val_corr_metrics = eval_metrics(y_val, y_val_pred_corr)

print("Entrenamiento corregido:", train_corr_metrics)
print("Validación corregida:", val_corr_metrics)
```


Siglas

AI Artificial Intelligence. [16](#)

aprendizaje supervisado Técnica de **ML** en la que el modelo se entrena con datos de entrada y salida conocidas. [48](#)

AQ Air Quality. [16, 19, 20](#)

AQG WHO AQ Guidelines. [19](#)

CO Monóxido de Carbono. [25, 27, 30, 32, 34, 48](#)

CO₂ Dióxido de carbono. [48](#)

ECH Electro-Chemical. [19–22, 45](#)

IoT Internet of Things. [16, 19](#)

LCS Low-Cost Sensor. [13, 16, 17, 19, 26, 27, 29–33, 35, 37, 42, 44, 48](#)

LTE Long Term Evolution. [17, 23](#)

MAE Mean Absolute Error. [49, 51](#)

MAPE Mean Absolute Prediction Error. [51](#)

MCU Micro Controller Unit. [23](#)

MEMS Micro-Electro-Mechanical Systems. [19, 21](#)

ML Machine Learning. [16, 47, 48, 87](#)

MLR Multiple Linear Regression. [16, 50](#)

MOX Metal OXide. [19–21, 23](#)

MSE Mean Square Error. [50](#)

NN Neural Network. [16](#)

NO₂ Dióxido de Nitrógeno. [30–35, 37, 38, 42, 44, 45, 48](#)

O₃ Ozono. [30–34, 37, 38, 42, 44, 45, 48](#)

OEM Original Equipment Manufacturer. [19](#)

Pearson Coeficiente de correlación de Pearson. [31](#)

PM Particulate Matter. [15](#), [20–22](#), [25](#), [30–33](#), [35](#), [37](#), [38](#), [42–45](#), [48](#), [49](#)

ppm Parts Per Million. [20](#), [22](#), [27](#)

Python . [47](#)

R Lenguaje de programación R. [26](#), [32](#), [47](#)

R² Coeficiente de determinación. [49–51](#)

RH Humedad relativa. [19–23](#), [27](#), [31](#), [32](#), [48](#)

RMSE Root Mean Square Error. [49](#), [51](#)

SO₂ Sulphur Dioxide. [25](#)

T Temperatura. [20–23](#), [27](#), [31](#), [33](#), [35](#), [48](#)

TVOC Total Volatile Organic Compounds. [20](#), [25](#), [27](#)

WHO World Health Organization. [15](#)

WSN Wireless Sensor Networks. [16](#)

Bibliografía

- [1] H. Adair-Rohani. Air pollution responsible for 6.7 million deaths every year. <https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants>, 2023. Accessed: 25/01/2025.
- [2] Pablo Orellano, Julieta Reynoso, Nancy Quaranta, Ariel Bardach, and Agustin Ciapponi. Short-term exposure to particulate matter (pm10 and pm2.5), nitrogen dioxide (no2), and ozone (o3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment International*, 142:105876, 2020.
- [3] Robert J. Henning. Particulate matter air pollution is a significant risk factor for cardiovascular disease. *Current Problems in Cardiology*, 49(1, Part B):102094, 2024.
- [4] WHO agency. WHO global air quality guidelines. <https://apps.who.int/iris/bitstream/handle/10665/345329/9789240034228-eng.pdf>, 2023. Accessed: 15/01/2025.
- [5] Directive 2008/50/EC. of the European Parliament and of the Councils of 21 May 2009 on ambient air quality and cleaner air for Europe. *Official Journal of the European Communities*, L 152:1–44, 2008.
- [6] C. Borrego, A.M. Costa, J. Ginja, M. Amorim, M. Coutinho, K. Karatzas, Th. Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, P. Smith, N. André, P. Gérard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguillón, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza. Assessment of air quality microsensors versus reference methods: The eunetair joint exercise. *Atmospheric Environment*, 147:246–263, 2016.
- [7] Federico Karagulian, Maurizio Barbiere, Alexander Kotsev, Laurent Spinelle, Michel Gerboles, Friedrich Lagler, Nathalie Redon, Sabine Crunaire, and Annette Borowiak. Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere*, 10(9), 2019.
- [8] Milagros Ródenas García, Andrea Spinazzé, Pedro T.B.S. Branco, Francesca Borghi, Guillermo Villena, Andrea Cattaneo, Alessia Di Gilio, Victor G. Mihucz, Elena Gómez Álvarez, Sérgio Ivan Lopes, Benjamin Bergmans, Cezary Orłowski, Kostas Karatzas, Gonçalo Marques, John Saffell, and Sofia I.V. Sousa. Review of low-cost sensors for indoor air quality: Features and applications. *Applied Spectroscopy Reviews*, 57(9-10):747–779, 2022.
- [9] C. Malings, R. Tanzer, A. Hauryliuk, S. P. N. Kumar, N. Zimmerman, L. B. Kara, A. A. Presto, and R. Subramanian. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, 12:903–920, 2019.

- [10] N. Castell, F. Rivas, A. Dauge, C. Schneider, L. Vogt, A. Lerner, M. Penza, and et al. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Atmospheric Environment*, 148:70–84, 2017.
- [11] O. Popoola, I. Adeyeye, R. Jones, M. Mead, S. Beevers, and et al. Development of a low-cost sensor network for measurement of air pollution. *Sensors and Actuators B: Chemical*, 236:1024–1034, 2016.
- [12] J. Zhu, M. Yang, and Z. J. Ren. Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, 57(46):17671–17689, 2023.
- [13] European Environment Agency. Air quality sensors – guidance for application, 2020. Available at: <https://www.eea.europa.eu/publications/air-quality-sensors-guidance>, Accessed: 20/06/2025.
- [14] Saverio De Vito, Kostas Karatzas, Alena Bartonova, and Grazia Fattorus, editors. *Air Quality Networks*. Environmental Informatics and Modeling. Springer Cham, 1 edition, 2023. Copyright Information: Springer Nature Switzerland AG 2023.
- [15] LTD Zhengzhou Winsen Electronics Technology Co. Multi-in-one sensor module (model: Zphs01b) manual. https://www.winsen-sensor.com/d/files/zphs01b-english-version1_1-20200713.pdf, 07 2020. [Accessed 27-01-2025].
- [16] LTD Zhengzhou Winsen Electronics Technology Co. Electrochemical ozone detection module (model: Ze27-o3) user's manual. <https://www.winsen-sensor.com/d/files/manual/ze27-o3.pdf>, 04 2020. [Accessed 27-01-2025].
- [17] LTD Zhengzhou Winsen Electronics Technology Co. Mems no2 gas sensor (model no.: Gm-102b) manual. <https://www.cnwinsen.com/wp-content/uploads/2021/08/MEMS-GM-102B-Manual-V2.1.pdf>, 04 2019. [Accessed 27-01-2025].
- [18] LTD Zhengzhou Winsen Electronics Technology Co. Carbon monoxide module (model no.: Ze15-co) manual. <https://www.winsen-sensor.com/d/files/ZE15-CO.pdf>, 04 2018. [Accessed 27-01-2025].
- [19] LTD Zhengzhou Winsen Electronics Technology Co. Infrared co2 sensor module (model: Mh-z19c) user's manual. https://www.winsen-sensor.com/d/files/infrared-gas-sensor/mh-z19c-pins-type-co2-manual-ver1_0.pdf, 02 2020. [Accessed 27-01-2025].
- [20] LTD Zhengzhou Winsen Electronics Technology Co. Electrochemical ch2o detection module (model: Ze08k-ch2o) user's manual. https://www.winsen-sensor.com/d/files/ze08k-ch2o-manual-v1_1.pdf, 04 2020. [Accessed 27-01-2025].
- [21] LTD Zhengzhou Winsen Electronics Technology Co. Air-quality detection module (model: Zp07-mp503) user's manual. <https://www.winsen-sensor.com/d/files/ZP07-MP503-4.pdf>, 11 2014. [Accessed 27-01-2025].
- [22] GXCAS Technology. Humidity and temperature sensor gxht3x. https://www.micros.com.pl/mEDIASERVER/CZ_SHT30_GX_0001.pdf, n.d. [Accessed 27-01-2025].
- [23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

- [24] Hadley Wickham and Garrett Grolemund. *R for Data Science*. O'Reilly Media, 2019.
- [25] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton & Company, 4th edition, 2007.
- [26] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 5th edition, 2012.
- [27] William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.
- [28] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2016.
- [29] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2nd edition, 2019.
- [30] Galit Shmueli. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.