CONAMA 2026

Integrar para comprender: datos de EDAR, fuentes abiertas y barreras de acceso a la información ambiental.

Edita: Fundación Conama

Año: 2026

Este documento está bajo una licencia de Creative Commons Reconocimiento-NoComercial 4.0 Internacional.

Personas autoras de la presente comunicación técnica

María Castrillo Melguizo. Personal Científico Titular. Instituto de Física de Cantabria (IFCA (CSIC-UC))

Otros Autores: Maria Fernanda Román Sánchez. Investigadora contratada. (IFCA (CSIC-UC))

Integrar para comprender: datos de EDAR, fuentes abiertas y barreras de acceso a la información ambiental.

CONAMA 2026

Índice

1. Resumen.....4

2. Introducción.....6

3. Métodos.....7

3.1. Caso de estudio.....7

3.2. Recolección de datos.....8

3.3. Comunicación e interfaz de usuario.....9

3.4. Preprocesado de datos.....10

3.5. Estrategia de validación.....11

3.6. Identificación de regímenes de afluente.....11

3.7. Clasificación en regímenes de afluente.....13

4. Resultados.....14

4.1. Identificación de regímenes.....14

4.2. Clasificación en regímenes de afluente.....17

5. Conclusión.....19

6. Agradecimientos.....20

7. Referencias.....20

Integrar para comprender: datos de EDAR, fuentes abiertas y barreras de acceso a la información ambiental.

CONAMA 2026

1. Resumen

En las Estaciones Depuradoras de Aguas Residuales (EDAR) se genera una gran cantidad de datos con enorme potencial para generar información de valor. Sin embargo, generalmente permanecen infrautilizados tanto por las dificultades para transformarlos en información útil como por las limitaciones de acceso, interoperabilidad e integración de fuentes de información complementarias. Se plantea así que el reto no es únicamente disponer de más datos, sino garantizar su accesibilidad y favorecer su interoperabilidad para transformar información dispersa en conocimiento útil para comprender la contaminación, mejorar la gestión del saneamiento y avanzar hacia una gobernanza ambiental más transparente y eficaz.

Los parámetros fisicoquímicos medidos directamente en el agua proporcionan una visión cuantitativa, localizada y con alta resolución temporal sobre su composición y su evolución, mientras que los datos procedentes de fuentes de información en abierto pueden aportar contexto sobre las posibles causas de la contaminación, como las actividades urbanas e industriales, los usos del suelo, las condiciones meteorológicas o las variaciones demográficas. La combinación de ambas perspectivas permite avanzar desde la observación de la contaminación hacia una mejor comprensión de los procesos que la generan y condicionan. Aunque parte de esta información está disponible en fuentes abiertas, su localización, acceso, calidad, resolución y compatibilidad no siempre permiten integrarla de forma sencilla con los datos de explotación de una EDAR.

Esta comunicación explora el valor complementario de ambas fuentes de datos, considerando sus capacidades, limitaciones y escalas de observación a través del proyecto PROWESS (Optimización de la eliminación de fósforo en agua residual mediante sensores basados en datos). En él, se abordan estos retos usando como caso de uso el desarrollo de sensores virtuales (*soft-sensors*), basados en datos e inteligencia artificial, capaces de inferir parámetros de calidad del agua en tiempo real a partir de variables medidas mediante sensores convencionales y de información procedente de fuentes abiertas. El enfoque centrado en el dato, junto con la aplicación de técnicas de explicabilidad y con el conocimiento previo del funcionamiento de los procesos que se dan en una EDAR, permite que estos modelos tengan sentido desde el punto de vista experto, reduciendo así la opacidad que habitualmente los caracteriza.

El proyecto se desarrolla a escala real en la una EDAR en Cantabria, en la cual durante un año se han recopilado cerca de 500 determinaciones de fósforo en laboratorio, junto con registros de alta resolución temporal de variables fisicoquímicas del agua, así como datos

abiertos meteorológicos y de movilidad poblacional. En esta comunicación se analiza en qué medida los diferentes tipos de datos permiten explicar distintos patrones de variación de la concentración de fósforo afluente, apoyado en el conocimiento experto, poniendo así en valor la complementariedad entre las distintas fuentes de datos y sus limitaciones. Este análisis permitirá avanzar hacia modelos más interpretables y fundamentados en el conocimiento del proceso, y contribuirá a identificar qué información y herramientas resultan relevantes para mejorar la monitorización y la gestión de las EDAR.

Palabras clave: digitalización, interoperabilidad, datos abiertos, sensores virtuales, agua residual

2. Introducción

En las Estaciones Depuradoras de Aguas Residuales (EDAR) se producen una gran cantidad de datos que pueden utilizarse para generar información de valor. En el ámbito académico, el desarrollo de aplicaciones basadas en datos para el apoyo en la operación y gestión de estas infraestructuras ha sido objeto de numerosos estudios (Corominas et al., 2017; Alvi et al., 2023). Sin embargo, en entorno operativo real, si bien la infraestructura de adquisición de datos y la automatización básica se hallan consolidadas, la adopción de tecnologías avanzadas de optimización basada en datos se encuentra aún en una fase incipiente, concentrada principalmente en grandes plantas urbanas y proyectos de demostración (CCI, 2021; MITECO, 2022). Por otro lado, la ya existente adquisición de datos en EDAR abre el debate sobre su soberanía y la utilidad de la digitalización si estos datos no se ponen al alcance de la ciudadanía. De acuerdo con López Martín, A. (2026) un verdadero proceso de digitalización ha de ampliar nuestro conocimiento, y para ello se hace necesaria la democratización de la información sobre este recurso esencial que es el agua. De hecho el acceso a la información ambiental no es solo una cuestión técnica, sino también un derecho reconocido en España por la Ley 27/2006, vinculada al Convenio de Aarhus. Sin embargo, el acceso a la información ambiental en el sector del agua residual en España se enfrenta con frecuencia a barreras estructurales: los datos sobre la operativa de las EDAR suelen encontrarse fragmentados, sujetos a opacidad o limitados por restricciones de las licencias de software propietario en las que habitualmente se basa su gestión. De este modo, en contra de otros campos como la meteorología, donde la puesta a disposición de la ciudadanía de la información de forma abierta y estandarizada ha impulsado un ecosistema dinámico de innovación y transparencia, el sector del agua residual arrastra un déficit histórico en la accesibilidad de sus datos. Por ello, esta falta de apertura no solo limita la fiscalización ciudadana y el cumplimiento efectivo del derecho de acceso a la información ambiental, sino que frena también la investigación colaborativa y el desarrollo de nuevos servicios de valor añadido por parte de terceros. Superar esta barrera pasa necesariamente por transicionar hacia estándares abiertos e interoperables que garanticen que la transformación digital de las depuradoras sirva tanto para optimizar el rendimiento industrial en el entorno operativo real como para devolver el conocimiento sobre un recurso tan estratégico al conjunto de la ciudadanía.

En el contexto científico, a esta limitación en la disponibilidad y calidad del dato se suma la complejidad intrínseca de los afluentes, caracterizados por una elevada variabilidad estacional y dinámicas no lineales que dificultan el análisis e interpretación de los datos. Ante este reto, el presente trabajo aborda la resolución de un problema práctico en una EDAR real, en concreto el conocimiento del fósforo (P) en el afluente, basándose en tres pilares fundamentales orientados a impulsar una verdadera digitalización del sector:

- Datos abiertos (Open Data): Se emplean fuentes de datos de acceso público relativas a procesos externos vinculados directamente a las variaciones del afluente, tales como la

meteorología local y la dinámica de la población flotante en los municipios a los que da servicio la planta.

- Integración de la información: Se combinan conjuntos de datos heterogéneos que operan a diferentes escalas temporales y espaciales. Mientras que las mediciones fisicoquímicas de alta resolución recopiladas en la EDAR describen qué ocurre en el agua, los datos abiertos (de menor frecuencia) aportan contexto sobre por qué ocurre. Aunque estos últimos no determinan causalidad por sí solos —dependiendo fuertemente de su resolución y acoplamiento con las mediciones operativas—, la convergencia de ambos niveles permite avanzar desde la simple observación hacia una comprensión integral de los procesos.

- Software abierto (Open Source): En coherencia con la apertura de datos y para contrarrestar las restricciones de opacidad y las barreras impuestas por las licencias de software propietario que dominan habitualmente la gestión de infraestructuras, toda la metodología se desarrolla mediante herramientas de código abierto. Esto garantiza la reproducibilidad científica, la transparencia técnica y facilita la transferencia real de las soluciones desarrolladas al conjunto de la industria y la comunidad investigadora.

- Aprendizaje automático informado (Informed ML): Para mitigar la brecha de conocimiento entre la ciencia de datos y la ingeniería de procesos, se adopta el paradigma de Informed Machine Learning (von Rueden et al., 2023). Este enfoque híbrido integra de forma explícita el conocimiento del dominio dentro del flujo de trabajo (ya sea en la preparación de datos, en la arquitectura del modelo o en la evaluación), garantizando la coherencia física de las predicciones. Como resultado, no solo se incrementa la precisión en las tareas de modelado, sino que se potencia la interpretabilidad de las soluciones y, consecuentemente, se favorece su aceptación y adopción por parte de los operadores finales.

El presente trabajo se limita al análisis de datos de afluente, su agrupamiento semi-supervisado en regímenes diferentes y el entrenamiento de un modelo de clasificación acompañado de selección de variables para asignar cada nueva observación a un determinado régimen.

3. Métodos

3.1. Caso de estudio

La instalación experimental de PROWESS se ha implementado a escala real en una EDAR en el interior de Cantabria (España), que da servicio a aproximadamente 7.000 habitantes equivalentes. La EDAR consta de una etapa de pretratamiento, una cámara de homogeneización, un reactor biológico anóxico-aerobio para la eliminación de carbono y nitrógeno y un proceso de ultrafiltración. Además, el P se elimina por precipitación mediante la adición de $FeCl_3$ en la cámara anóxica del

reactor biológico. El punto de muestreo y medición se situó en el canal de entrada, donde se colocaron los sensores, por lo que se utilizó el afluente bruto como matriz para el análisis.

La EDAR sufre fuertes cambios en su carga de P debido a diversos factores (vertidos, fluctuación de la población, etc.). Por este motivo, una estimación en tiempo real de la carga de P podría ayudar en la gestión de la dosificación de $FeCl_3$ u otras estrategias de explotación, garantizando el cumplimiento de los requisitos de concentración en el efluente.

3.2. Recolección de datos

A lo largo de todo un año, se han recopilado datos procedentes de tres fuentes diferentes para componer el conjunto de datos que se utilizará para entrenar los modelos del soft-sensor.

En primer lugar, las concentraciones de fósforo como ortofosfato (PO_4-P), se obtuvieron mediante muestreo puntual y análisis de laboratorio. Se usó el método fotométrico utilizando un espectrofotómetro DR2800® (HACH) previo filtrado al vacío utilizando un sistema de filtración de policarbonato FS04725OP® (Fisher Scientific) con filtros de membrana de fibra de vidrio (47 mm, 1,6 µm; VWR n.º de cat. 516-0861).

En cuanto a las variables fisicoquímicas que actúan como predictores durante el entrenamiento y posteriormente en la clasificación, se instalaron cuatro sondas digitales (Endress+Hauser) en la cámara de entrada para registrar sus valores con una frecuencia minutal. En concreto, se seleccionaron seis variables para su medición en línea. La conductividad (cond), considerada una de las variables más importantes debido a su esperada correlación con las concentraciones de nutrientes, se monitorizó con una sonda inductiva (CLS50D). El pH, el potencial de oxidación-reducción (redox) y la temperatura (temp) se midieron mediante un electrodo de vidrio (CPS16E). El oxígeno disuelto (OD) y la turbidez (turb) se midieron con sensores ópticos de acero inoxidable (COS61D y CUS52D, respectivamente). Los sensores se colocaron en pértigas de PVC para garantizar que permanecieran sumergidos sin tocar el fondo de la cámara.



Figura 1. Cuadro de control en el muro del canal de entrada a la EDAR.

Como datos en abierto, se han recopilado datos meteorológicos y datos de población. En cuanto a los datos meteorológicos, estos se han obtenido de la plataforma de datos abiertos de la Agencia Estatal de Meteorología (AEMET). En concreto, se han recopilado los

datos de la estación meteorológica más cercana. Se dispone de datos horarios de las siguientes variables: temperatura media horaria (ta) (ºC), precipitación acumulada horaria (prec)(mm) y humedad relativa (hr) (%). Los datos de población se han obtenido del *Estudio de movilidad con Big Data* del Ministerio de Transportes y Movilidad Sostenible (MITMA, 2022). En particular se han utilizado los datos de pernoctaciones diarias (personas) en los municipios a los que sirve la EDAR de estudio. Como los datos de movilidad han dejado de estar disponibles a partir de julio de 2025, se ha llevado a cabo un proceso de extrapolación para estimar los datos de 2026.

3.3. Comunicación e interfaz de usuario

Las cuatro sondas se conectaron al controlador de Endress+Hauser (CM444) y al gestor de datos (Ecograph T RSG35), lo que permitió leer y almacenar de forma segura las mediciones de las sondas, además de proporcionar mediciones in-situ a través de sus interfaces visuales. Se conectó un ordenador industrial de bajo consumo al gestor de datos Ecograph, del que lee datos de forma continua. Este ordenador está equipado con un módulo 4G LTE, conectado a la red de Movistar, y realiza tareas adicionales que son clave para todo el proceso de monitorización. Se encarga de recopilar y preprocesar los datos de los sensores a través de una conexión Ethernet cableada directa. A continuación, estos datos se transfieren y sincronizan a intervalos mediante un método de transferencia seguro al sistema en la nube.

Un componente clave del sistema es la plataforma basada en la nube que se utiliza para gestionar de forma centralizada los datos, el entrenamiento de los modelos y la visualización. Para ello, PROWESS aprovecha la infraestructura de la Nube Científica del IFCA-CSIC, donde se agregan y almacenan los datos de los sensores y los datos meteorológicos de la AEMET tanto en una base de datos relacional como en disco, con el fin de utilizarlos para seguir entrenando el sensor virtual. Para crear visualizaciones, PROWESS utiliza la herramienta de código abierto Grafana, como se muestra en la Figura 2.



Figura 2. Vista del panel en Grafana.

3.4. Preprocesado de datos

Los datos obtenidos de los sensores se procesaron para detectar y eliminar o corregir los datos anómalos. En primer lugar, en el caso de la conductividad, la temperatura y la turbidez, se eliminaron los valores negativos, ya que son físicamente imposibles. Además, se aplicó el método del rango intercuartílico (IQR) a las series diferenciadas para eliminar los valores anómalos identificados como extremadamente diferentes al valor anterior. Se eliminaron los valores atípicos y se sustituyeron mediante interpolación lineal entre los valores anterior y siguiente. En el caso del OD y el ORP, se eliminaron uno y tres valores claramente anómalos, respectivamente, mediante inspección visual. Por último, la serie temporal de pH requirió corrección, ya que se observó una deriva. Para ello, se asumió que la deriva se había producido a una tasa constante desde 0 hasta la deriva detectada, basándose en la comparación con el patrón (pH = 7) observado en marzo. La serie se corrigió restando la deriva esperada para el tiempo transcurrido desde que la serie comenzó a derivar. Esta deriva no se observó en las otras dos variables medidas por la misma sonda: ORP y temperatura.

En cuanto a los datos meteorológicos, se crearon variables derivadas. En particular la temperatura media de los últimos dos días (ta_avg_2880) y la precipitación acumulada en los últimos dos días (prec_acc_2880). Por último, en el caso de la población se diferenció a la población local (personas_local) seleccionando sólo aquellos registros en los que el código postal de procedencia coincidía con el de pernoctación.

3.5. Estrategia de validación

Como se ha indicado previamente, se dispone de datos recogidos a lo largo de un año. Aunque los datos tengan una marca temporal (día y mes), el problema no se aborda como un problema de series temporales clásico. Las variables predictoras con las que contamos son variables ambientales discretas y no se han introducido variables autorregresivas o tipo lag. El modelo aprende a mapear causas o indicadores físicos con concentraciones, no a extrapolar el futuro basándose en el pasado inmediato. En este caso, llevar a cabo una validación cronológica, es decir, separando la última fracción de los datos disponibles como fracción de validación, no resulta óptimo para capturar la variabilidad estacional completa de un ciclo anual. Por este motivo, se ha implementado una estrategia de validación cruzada dejando un mes fuera (Leave-One-Month-Out Cross-Validation (LOMO-CV)), evaluando secuencialmente cada uno de los 12 meses del año como conjunto de test independiente frente al resto de meses utilizados para entrenamiento. Al llevar a cabo la validación LOMO-CV se busca comprobar si el modelo es capaz de entender el comportamiento de un mes entero (con sus dinámicas específicas condicionadas por factores externos representados por las variables meteorológicas y poblacionales) que nunca ha visto durante el entrenamiento, en lugar de aprender a predecir un futuro cercano en función de lo que ha ocurrido en el pasado.

3.6. Identificación de regímenes de afluente

En el caso particular del estudio del $P\text{-}PO_4$ en el afluente de una EDAR, se aprovecha el conocimiento previo en la correlación entre la variable de interés y las variables predictoras para preparar el espacio de datos y condicionar los modelos de inferencia. En la bibliografía científica se ha reportado habitualmente la existencia de una fuerte correlación lineal positiva entre la conductividad y la concentración de $P\text{-}PO_4$ (Yu et al., 2019, Castrillo et al., 2020). A la vista del gráfico de dispersión entre estas dos variables (Figura 3), se intuye cierta correlación positiva pero fuera del grupo de datos principal se observa una dispersión en dos direcciones. Por un lado, las observaciones con mayores valores de conductividad tienen asociados valores de concentración de P-PO4 que no son tan altos, en general menores de 6 mg L^{-1}. Por otro lado, las observaciones correspondientes a las mayores concentraciones de $P\text{-}PO_4$ tienen asociados valores moderados de conductividad, menores de 1000 µS cm^{-1}.

De acuerdo con la taxonomía propuesta por von Rueden (2023), la introducción de conocimiento previo procedente de leyes físicas y físico-químicas que rigen el comportamiento del afluente de una EDAR y la relación estequiométrica/hidráulica entre la conductividad y el fósforo, se trata de conocimiento científico. Además, este es formalmente representable a través de la ecuación de la recta. En cuanto al modo de introducir el conocimiento en el flujo de trabajo, tiene un peso muy fuerte en la preparación de los datos de entrenamiento, pero con una conexión directa con el planteamiento de la hipótesis y cómo se estructura el modelo de identificación de regímenes de afluente en una arquitectura de regímenes múltiples o mezcla de expertos para guiar al clasificador.

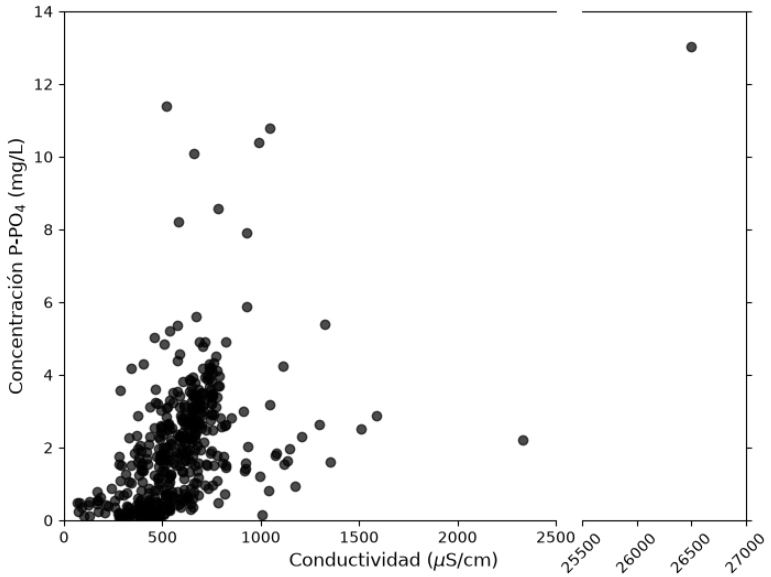


Figura 3. Gráfico de dispersión entre la concentración de P-PO_4 y conductividad de las muestras analizadas.

La observación del gráfico de la Figura 3 sugiere la existencia de dos regímenes de afluente en los que se dan diferentes relaciones entre la conductividad y la concentración de P-PO_4, una con mayor pendiente y otra con menor pendiente. Para tratar de identificar estos regímenes se han aplicado distintas estrategias de agrupamiento, entre ellas:

-Identificación de un régimen principal y observaciones excepcionales mediante agrupamiento basado en densidad: Para ello se utilizó el algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise), con el cual se separó el grupo principal de las observaciones excepcionales, pero no se pudo conservar la esperada relación lineal entre conductividad y concentración de P-PO_4, por lo que se descartó este método.

-Identificación de un régimen principal y otros secundarios mediante regresión robusta mediante el algoritmo RANSAC (*Random Sample Consensus Regressor*). Esta técnica permite respetar la correlación lineal identificando una linea principal y outliers a ambos lados de la misma. Esta estrategia considera outliers los puntos que quedan en las zonas más dispersas, lo que no es coherente con la intuición de los dos regímenes.

-Agrupamiento no supervisado en dos clústers mediante el algoritmo K-Means. En lugar de aplicar el clustering directamente sobre los valores de la variable, en este caso de la conductividad, este se ha ejecutado sobre el cociente individual (Y/X) de cada punto, permitiendo separar los datos en función de sus regímenes de pendiente antes de ajustar las rectas de regresión.

3.7. Clasificación en regímenes de afluente

Una vez que se conoce el régimen al que pertenece cada punto, es necesario desarrollar un modelo de clasificación que permita asignar un régimen a los puntos de la fracción de validación. Entre los diferentes modelos de clasificación basados en datos se ha seleccionado un clasificador basado en árboles (Gradient Boosting), que es capaz de trazar fronteras de decisión complejas y adaptativas para distinguir cuándo el sistema salta de un régimen a otro. Uno de los principales efectos que se buscan con la división en regímenes, es dotar al modelo de mayor explicabilidad de cara a su uso por gestores del agua. Desde este punto de vista, llevar a cabo el entrenamiento del modelo de clasificación acompañado de una selección de variables, permite conocer qué variables son importantes a la hora de identificar estos regímenes, lo que ayudaría a entender qué mecanismos externos dan lugar a los mismos. Para alcanzar este objetivo, se aplicó una estrategia de selección de variables inversa (*backward selection)* con una complejidad de modelo fija y conservadora (regularización estática) durante todo el proceso. Al restringir la profundidad de los árboles de decisión (max_depth = 3) y la tasa de aprendizaje (learning_rate = 0.05) a valores constantes, se evitó el sobreajuste local que típicamente ocurre al evaluar subconjuntos reducidos de datos, garantizando así que la eliminación de variables respondiera exclusivamente a su aportación física y predictiva, y no a ajustes artificiales del hiperparámetro. Partiendo del conjunto completo de predictores disponibles, se fueron eliminando variables una a una hasta que solo quedó una variable en el subconjunto. En cada iteración, mediante validación cruzada, se evaluaron todos los posibles subconjuntos candidatos obtenidos al eliminar una sola variable. Dentro de cada *fold* de entrenamiento, los datos se fraccionan de forma aleatoria para evitar que la distribución de los datos dentro de cada uno esté sesgada por la estacionalidad de los procesos que condicionan las características del agua residual. La métrica utilizada para evaluar cada modelo correspondiente a cada subconjunto de variables es su exactitud o *accuracy*, aunque también se han calculado otra variables que ayudan a evaluar los resultados, como son el *F1 score macro* y *ponderado*.

4. Resultados

4.1. Identificación de regímenes

La aplicación de k-means para el agrupamiento de las observaciones en función del valor de la relación X/Y dio lugar a dos clústers que permiten diferenciar claramente los dos grupos de datos que se intuían en un principio: un grupo (Clúster 0) que consta de 217 observaciones, en el que se alcanzan mayores concentraciones de $P\text{-}PO_4$ manteniendo una conductividad de hasta 1000 µS cm^{-1}, y un grupo (Clúster 1) que incluye 235 observaciones en las que se da un rango más amplio de conductividad y valores bajos de concentración de $P\text{-}PO_4$ y . Para cada uno de estos grupos se ajustó una recta de regresión lineal, obteniéndose pendientes bien diferenciadas: 0,0053 y 0,0005 respectivamente para los grupos citados. En la Figura 4 se muestra de nuevo el gráfico de dispersión entre conductividad y concentración de $P\text{-}PO_4$, esta vez diferenciando los dos clústers obtenidos y mostrando las rectas de regresión. Es importante destacar cómo el punto con los valores más altos para las dos variables es recogido por la recta de regresión del clúster 1, por lo que a pesar de sus valores excepcionales no ha de ser considerado outlier.

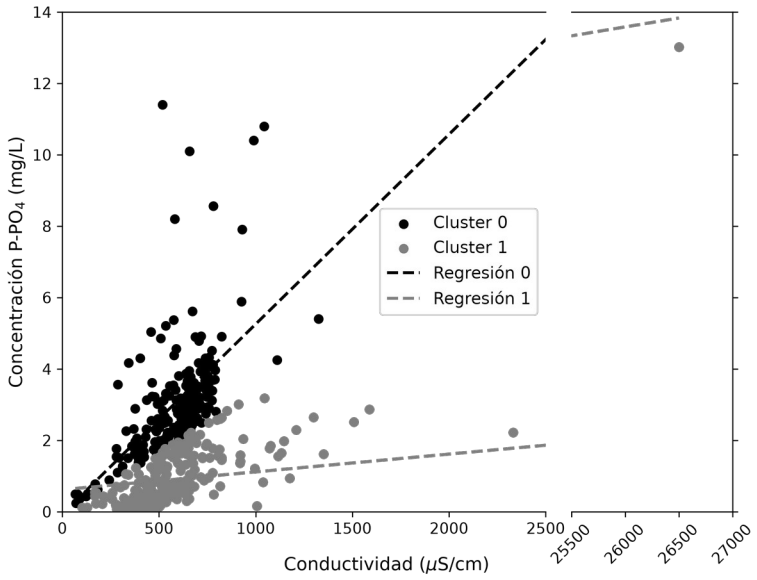


Figura 4. Gráfico de dispersión entre la concentración de $P\text{-}PO_4$ y conductividad de las muestras analizadas diferenciado en regímenes de afluente.

Para poder interpretar mejor a qué fenómenos responde la división de las observaciones en los dos clústers obtenidos, en la Figura 5 se han representado los valores de concentración a lo largo de la campaña experimental diferenciados en función del clúster de pertenencia. Se puede observar que las observaciones pertenecientes al clúster 0, caracterizado por una mayor pendiente en la correlación conductividad–P-PO_4, se concentran principalmente de mediados de mayo a octubre, mientras que las del clúster 1, caracterizado por una menor pendiente en la correlación conductividad–P-PO_4, se concentran entre noviembre y mediados de mayo, si bien durante el resto del año aparecen también observaciones pertenecientes a este clúster. Este primer análisis visual sugiere por tanto, que las condiciones estivales inducen a una mayor pendiente en la relación conductividad–P-PO_4.

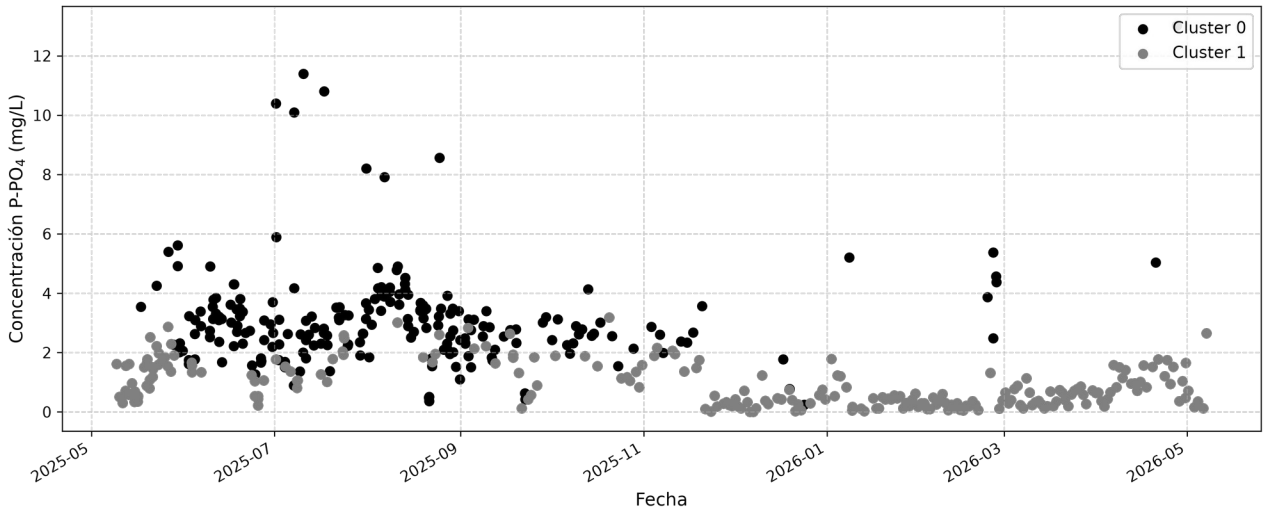


Figura 5. Representación de la concentración de P-PO_4 en las muestras analizadas en escala temporal.

La representación de la distribución de las variables predictoras en cada uno de los dos clústers, como se muestra en la Figura 6 mediante diagramas de cajas, proporciona información sobre cuáles son las variables que diferencian a cada clúster. Se puede observar que las variables relacionadas con la temperatura (temp, ta y ta_avg_2880) son las que determinan en mayor medida esa diferenciación, ya que las cajas correspondientes a cada clúster no se solapan. Esta observación es coherente con lo mostrado por el gráfico de la Figura 5, donde se observa un cambio estacional entre ambos regímenes. El potencial redox y la concentración de oxígeno disuelto tienen un comportamiento similar entre ellas, lo que es coherente con el carácter oxidante del oxígeno. En los dos casos, valores altos de estas variables aparecen principalmente en el clúster 1. Asimismo, los valores altos de precipitación acumulada, más habituales en periodo invernal, aparecen también principalmente en el clúster 1. Por último, las variables relacionadas con la población también parecen estar relacionadas con la diferenciación de los grupos, especialmente la variable personas. Se observa como los valores más altos de pernoctaciones de personas, incluyendo las personas de

fuera de los municipios a los que da servicio la EDAR, están asociados al clúster 0. Teniendo en cuenta que es esperable que la mayor afluencia de personas se produzca en el periodo estival, en el cual también se da un incremento de temperatura, no es posible diferenciar entre correlación y causalidad sólo con esta información. Es necesario profundizar en el análisis para comprobar si la variable de la población realmente aporta información adicional o se trata de un fenómeno de colinealidad con la temperatura.

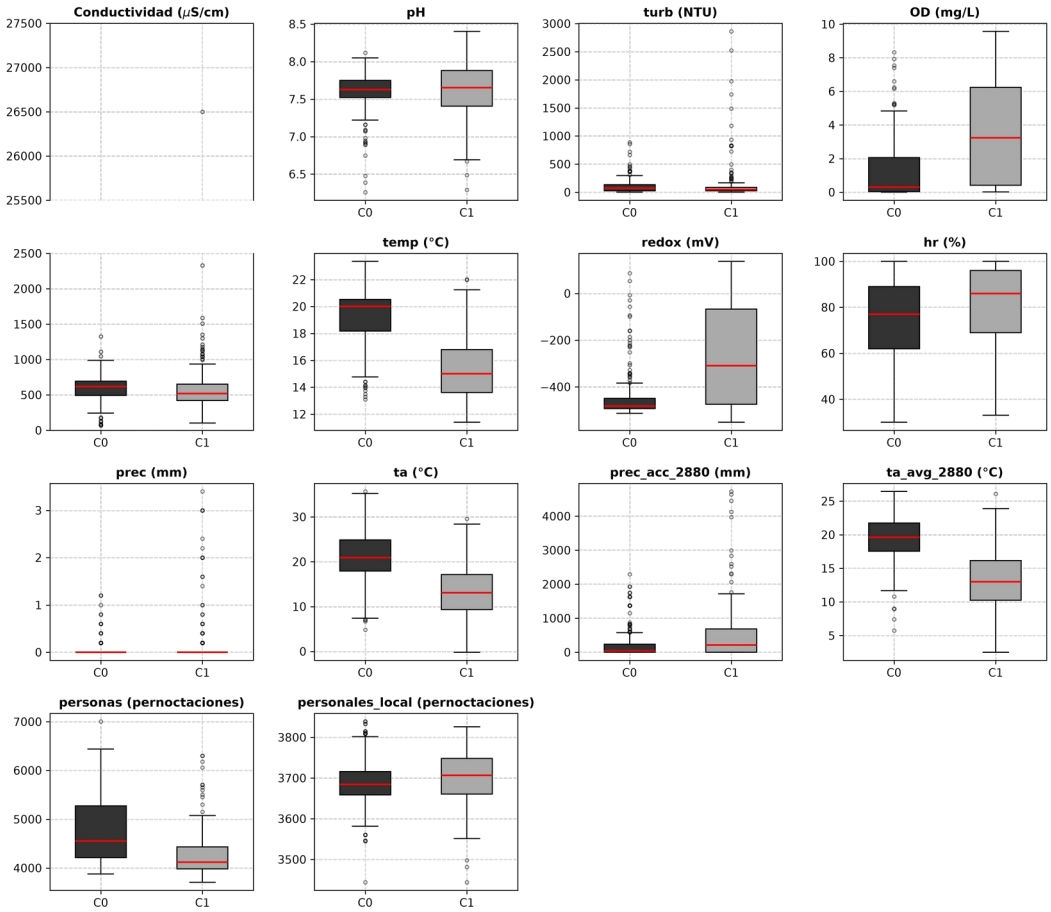


Figura 6. Diagramas de cajas de las variables utilizadas para entrenar el modelo de clasificación de regímenes de afluente.

4.2. Clasificación en regímenes de afluente

La observación de que las temperaturas apenas se solapan entre clústers podría llevar a pensar que el agrupamiento no supervisado debería haberse hecho en función de la temperatura. Esta decisión probablemente hubiera dado como resultado dos clústers para los cuales las nuevas observaciones serían más fáciles de clasificar, ya que se basaría en un punto de corte de la temperatura. Sin embargo, esto perjudicaría la linealidad de la relación conductividad–P-PO_4 dentro de cada clúster. El agrupamiento por pendientes trata de representar dos situaciones funcionales o estados biogeoquímicos, más que dividir en función de una variable determinada. Es por ello que como modelo de clasificación se ha seleccionado Gradient Boosting, caracterizado por ser capaz de encontrar fronteras de decisión no lineales y complejas.

Para conocer qué variables son realmente útiles a la hora de clasificar una nueva observación se ha realizado una selección de variables en el entrenamiento del modelo de clasificación que permita realizar esta tarea. En la Tabla 1 se muestran, para cada mes usado como validación, las variables que proporcionaron el menor error de CV durante el entrenamiento. La variable que representa la precipitación acumulada en los últimos dos días aparece como una de las seleccionadas para todos los modelos entrenados. La siguiente variable que más modelos utilizan es la temperatura del agua, que aparece en todos menos en el modelo que usa el mes de junio como test. Otras variables que aparecen en un número elevado de modelos, en particular en más de la mitad, son la temperatura media en los últimos dos días, la precipitación instantánea y el número de personas que pernoctan en los municipios a los que da servicio la EDAR.

Mes de test	Exactitud	cond	pH	temp	OD	redox	turb	ta	ta_avg_2880	prec	prec_acc_2880	hr	personas	personas local
Enero	0,9630	✓		✓		✓	✓	✓		✓	✓	✓	✓	✓
Febrero	0,8214	✓	✓	✓				✓			✓	✓	✓	✓
Marzo	1	✓		✓	✓	✓			✓	✓	✓	✓		
Abril	0,9643			✓					✓		✓	✓	✓	
Mayo	0,8776			✓	✓				✓	✓	✓		✓	
Junio	0,3898		✓		✓		✓	✓		✓	✓	✓	✓	
Julio	0,8621		✓	✓		✓			✓	✓	✓		✓	
Agosto	0,8281			✓					✓		✓			✓
Septiembre	0,6111			✓	✓						✓		✓	
Octubre	0,6000	✓		✓	✓	✓	✓			✓	✓		✓	✓

Noviembre	0,7273	✓	✓	✓		✓	✓	✓		✓	✓
Diciembre	0,9231	✓	✓	✓		✓	✓	✓			

Tabla 1. Variables seleccionadas mediante selección backward para la clasificación en regímenes según el mes de validación.

Finalmente, se selecciona un conjunto de variables único que proporcione las mejores métricas de clasificación globales para todos los meses del año, resultando este conjunto en el compuesto por las cinco variables citadas.

Mes de test	Exactitud	F1-Macro	F1-ponderado	Matrices de confusión	
Enero	0,9630	0,4906	0,9448	0	1
				0	26
Febrero	0,8214	0,4510	0,7409	23	0
				5	0
Marzo	1,0000	1,0000	1,0000	30	
Abril	0,9643	0,4909	0,9468	0	1
				0	27
Mayo	0,8776	0,7759	0,8714	5	4
				2	38
Junio	0,8305	0,5362	0,7679	48	0
				10	1
Julio	0,8448	0,4579	0,7738	49	0
				9	0
Agosto	0,9219	0,4797	0,8844	59	0
				5	0
Septiembre	0,7222	0,5982	0,6602	23	0
				10	3
Octubre	0,6400	0,6180	0,6363	11	4
				5	5
Noviembre	0,7727	0,7270	0,7677	13	2
				3	4
Diciembre	0,8846	0,4694	0,8666	0	2
				1	23

Tabla 2. Métricas de los modelos de clasificación para cada uno de los meses de validación.

La exactitud es alta en la mayoría de los meses (alrededor del 80%-90%). Sin embargo, al analizar el F1-Macro, se observa que en algunos meses el clasificador predice siempre la clase mayoritaria.

Ocurre principalmente durante los meses estivales (Junio, Julio y Agosto), en las que la clase 0 domina casi por completo pero el modelo predice siempre ceros para la clase 1, y en algunos meses invernales (Diciembre, Enero y Abril) en los que domina la clase 1 y el modelo tiende a predecir 0. Esto se explica porque la representación de la clase minoritaria en cada uno de los dos regímenes que se diferencian en el año es muy baja, insuficiente para que el modelo pueda aprender a diferenciarla.

A pesar de incorporar variables de inercia temporal y meteorológica (medias móviles térmicas y acumulados de precipitación a dos días), el clasificador experimenta una degradación severa de su capacidad predictiva durante el periodo estival (meses 6 a 8). Este comportamiento evidencia que los regímenes de la EDAR en verano están dominados por dinámicas que no quedan plenamente reflejadas en las variables de contexto proxies disponibles. Este resultado pone de manifiesto una de las principales barreras de acceso a la información ambiental en la gestión del agua urbana: la escasez de variables socioeconómicas y demográficas de alta resolución temporal, limitando la capacidad de los modelos basados exclusivamente en datos físico-químicos e hidrometeorológicos para anticipar transiciones de régimen extremas.

5. Conclusión

En esta comunicación se ha mostrado, a través de un ejemplo, la utilidad de utilizar el conocimiento físico-químico del proceso para el etiquetado semi-supervisado de regímenes operacionales en una EDAR. Esta integración de conocimiento experto en la fase previa de *clustering* permite superar los modelos de caja negra puros, asegurando que los clústeres representen estados biogeoquímicos reales del sistema de depuración.

El empleo de un clasificador basado en árboles de decisión reforzados, optimizado mediante selección inversa de variables y una regularización estática conservadora, logra un rendimiento global muy elevado en los meses templados y fríos (con precisiones superiores al 80-90%). Esto confirma que el clasificador es capaz de aprender eficazmente las fronteras no lineales entre regímenes cuando las condiciones operativas y ambientales se mantienen dentro de los rangos habituales de entrenamiento.

No obstante, los resultados de la validación cruzada temporal (LOMO-CV) evidencian una degradación del rendimiento durante el periodo estival, principalmente en los meses centrales de verano. Dado que este fallo persiste incluso tras incorporar variables de inercia hidrometeorológica (medias móviles térmicas y acumulados de precipitación) y variables que tratan de representar las dinámicas poblacionales, se intuye la influencia de variables que no han podido ser recopiladas en este estudio. Este fenómeno pone de manifiesto una de las principales barreras de acceso a la información ambiental en la gestión del agua residual: la escasez de variables demográficas y socioeconómicas de alta resolución temporal, lo cual limita la capacidad de los modelos basados

exclusivamente en registros físico-químicos e hidrometeorológicos para anticipar transiciones de régimen extremas.

6. Agradecimientos

Optimización de la eliminación de fósforo en agua residual mediante sensores basados en datos (PROWESS). Plan Estatal de Investigación Científica y Técnica y de Innovación 2021-2023 (PID2023-150331OA-I00).

7. Referencias

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. (2017). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, 88, 185-199.

Alvi, M., Batstone, D., Mbamba, C. K., Keymer, P., French, T., Ward, A., Dwyer, J., & Cardell-Oliver, R. (2023). Deep learning in wastewater treatment: A critical review. *Water Research*, 245, 120518.

Centro de Ciberseguridad Industrial. (2021). Estado de la ciberseguridad y digitalización en las infraestructuras del sector del agua en España. CCI. https://www.cci-es.org/

Ministerio para la Transición Ecológica y el Reto Demográfico. (2022). PERTE de Digitalización del Ciclo del Agua: Memoria de la medida. Gobierno de España. https://www.miteco.gob.es/es/agua/temas/perte-ciclo-agua.html

López Martín, A. (2026, 6 de noviembre). Sin dato público de calidad no hay digitalización que valga. Blog iAgua. https://www.iagua.es/blogs/agustina-lopez/sin-dato-publico-de-calidad-no-hay-digitalizacion-que-valga

Ministerio de Transportes, Movilidad y Agenda Urbana. (2022). Estudio de movilidad con Big Data [Conjunto de datos]. Gobierno de España. https://www.mitma.gob.es/el-ministerio/campanas/open-data-movilidad

Yu, Q., Liu, R., Chen, J., Chen, L. (2019) Electrical conductivity in rural domestic sewage: An indication for comprehensive concentrations of influent pollutants and the effectiveness of treatment facilities, International Biodeterioration & Biodegradation, 143, 104719,

Castrillo, M., & García, Á. L. (2020, April). *Estimation of high frequency nutrient concentrations from water quality surrogates* using machine learning methods. Water Research. 172, 115490

L. von Rueden et al. (2023) "Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 614-633, doi: 10.1109/TKDE.2021.3079836.

CONAMA 2026

CONAMA 2026

Integrar para comprender: datos de EDAR, fuentes abiertas y barreras de acceso a la información ambiental.