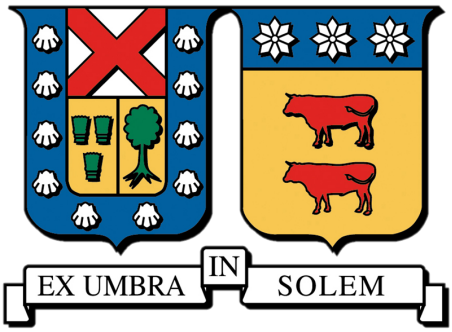
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INGENIERÍA QUÍMICA Y AMBIENTAL



PREDICCIÓN Y MITIGACIÓN DE EMISIONES DE N_2O EN PLANTAS DE TRATAMIENTO DE AGUAS RESIDUALES

JOAQUÍN FRANCISCO GALLARDO AVENDAÑO

TESIS PARA OPTAR AL GRADO DE

MAGÍSTER EN CIENCIAS DE LA INGENIERÍA QUÍMICA

Guía: Dr. Luis Bergh
Co-referentes: Dr. Santiago García
MSc. Martin Concha R.

DICIEMBRE-2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título ☒ Tesis de Postgrado

Título del trabajo: Predicción y mitigación de emisiones de N_2O en plantas de tratamiento de aguas residuales.

Nombre del candidato(a): Joaquín Francisco Gallardo Avendaño

Carrera / Grado: Magíster en Ciencias de la Ingeniería Química

Campus: Casa Central **Departamento:** Ingeniería Química y Ambiental

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Luis Bergh Olivares,en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente

DEJO CONSTANCIA que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☐ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☒ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

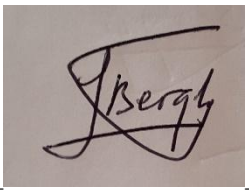
☒ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

Presencia de datos operacionales y de emisiones de gases de efecto invernadero de una planta de tratamiento de aguas residuales que fue cliente de la empresa con la cual se realiza la investigación.

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 20-12-2015 **Firma:** 

Estudiante o Candidato(a):

Fecha: 19/12/25 **Firma:** 

Resumen

Las plantas de tratamiento de aguas residuales son una fuente significativa de N_2O, un potente gas de efecto invernadero. La predicción precisa de estas emisiones es un desafío debido a la complejidad y naturaleza no lineal de los procesos biológicos involucrados. Esta investigación se enfoca en el desarrollo de un modelo predictivo para las emisiones de N_2O en una PTAR, utilizando datos operacionales en tiempo real e incorporando la dinámica temporal del afluente para considerar la variabilidad de la alimentación y el tiempo de residencia hidráulico.

Inicialmente, se constató que la planta opera en un estado predominantemente no estacionario , lo que impidió el uso de Análisis de Componentes Principales para la reducción de dimensionalidad. Como alternativa, se evaluó un modelo de Mínimos Cuadrados Parciales (PLS) dinámico, sin embargo, fue descartado debido a que el análisis de sus residuos reveló un comportamiento no normal, indicando que no lograba capturar adecuadamente la complejidad del sistema.

Posteriormente, se implementó un modelo de machine learning basado en XGBoost. Este enfoque demostró ser altamente eficaz, logrando una predicción precisa de las emisiones de N_2O. El análisis de importancia de características identificó al pH, el flujo de aire, el oxígeno disuelto y los sólidos suspendidos como las variables más determinantes en la predicción de emisiones. La validación del modelo confirmó que los residuos siguen una distribución normal y están centrados en cero. Además, se obtuvo un coeficiente de variación entre la desviación estándar del error y la media de los datos de tan solo 0.6%, lo que subraya la alta precisión y capacidad predictiva del modelo. Adicionalmente, se validó un horizonte predictivo robusto de aproximadamente 20 horas, el cual coincide con el tiempo de residencia hidráulico del reactor, delimitando así la ventana temporal confiable para la aplicación de estrategias de control.

Se concluye que el modelo XGBoost es una herramienta robusta y precisa para la modelación de emisiones en sistemas dinámicos y no lineales como las PTAR. Finalmente, se sugiere la integración de este modelo en estrategias de control operativo, permitiendo la modificación de las variables críticas identificadas para reducir activamente la generación de gases de efecto invernadero. Además, se sugiere evaluar la posibilidad de expandir el modelo a los demás gases principales de efecto invernadero (CO_2 y CH_4).

Tabla de Contenidos

Lista de Figuras

Nomenclatura

Siglas/Símbolo	**Significado**
AOB	Bacterias Oxidadoras de Amonio (*Ammonia-Oxidizing Bacteria*)
CO_2	Dióxido de carbono
DQO	Demanda química de oxígeno
DQO/N	Relación adimensional entre la demanda química de oxígeno y el nitrógeno
EF	Factor de Emisión (*Emission Factor*)
GEI	Gases de Efecto Invernadero
N_2	Nitrógeno gaseoso
N_2O	Óxido nitroso
NDIR	Infrarrojo No Dispersivo (*Non-Dispersive Infrared*)
NH_2OH	Hidroxilamina
NO_2^-	Nitrito
NO_3^-	Nitrato
OD	Oxígeno disuelto
PC1	Primer componente principal
PC2	Segundo componente principal
PCA	Análisis de Componentes Principales (*Principal Component Analysis*)
PLC	Controlador Lógico Programable (*Programmable Logic Controller*)
PLS	Mínimos Cuadrados Parciales (*Partial Least Squares*)
PTAR	Planta de Tratamiento de Aguas Residuales
RAS	Lodo Activado de Retorno (*Return Activated Sludge*)
SST	Sólidos Suspendidos Totales
WAS	Lodo Activado de Purga (*Waste Activated Sludge*)
XGBoost	*Extreme Gradient Boosting*

Capítulo 1

Introducción

1.0.1 ¿Qué es una planta de tratamiento de aguas residuales?

Las aguas residuales se definen como la combinación de los desechos líquidos o aguas portadoras de residuos, provenientes de residencias, establecimientos comerciales, instituciones e instalaciones industriales, a las que se pueden sumar, eventualmente, aguas subterráneas, aguas superficiales o aguas lluvias. La recolección y tratamiento de estas aguas es uno de los pilares fundamentales de la salud pública y la protección ambiental.

El propósito principal de una Planta de Tratamiento de Aguas Residuales (PTAR) es acelerar los procesos naturales de purificación del agua en un entorno controlado y contenido. Si se descargaran sin tratar, los contaminantes presentes en ellas agotarían el oxígeno disuelto de los cuerpos de agua receptores (ríos, lagos, océanos), provocando la muerte de la vida acuática, generando olores y propagando enfermedades hídricas (Nguyen et al., 2020).

En un contexto de creciente escasez hídrica y urbanización acelerada, la necesidad de un tratamiento efectivo de las aguas residuales se ha intensificado. En Chile, por ejemplo, la escasez hídrica es un problema estructural, lo que impulsa el desarrollo y la optimización de las PTAR no solo como un sistema de saneamiento, sino como una potencial fuente de agua recuperada para usos no potables, como el riego agrícola o industrial. Este impulso hacia la reutilización exige estándares de tratamiento cada vez más estrictos.

1.0.2 Contaminantes principales en aguas residuales

Las aguas residuales son una mezcla heterogénea compleja. Para diseñar un tratamiento efectivo y entender la generación de subproductos, como los Gases de Efecto Invernadero (GEI), es crucial caracterizar los contaminantes a remover. Los principales constituyentes de interés en las aguas residuales municipales son:

- Materia Orgánica: compuesta por proteínas, carbohidratos, grasas y otros compuestos de origen biológico o industrial. Su presencia es problemática ya que su descomposición microbiana en los cuerpos de agua consume oxígeno disuelto. Se caracteriza con dos parámetros principales, DBO y DQO:

  - Demanda Bioquímica de Oxígeno (DBO): mide la cantidad de oxígeno disuelto requerido por los microorganismos para oxidar biológicamente la materia orgánica en un tiempo determinado (usualmente 5 días, DBO_5). Es el indicador clave de la carga contaminante biodegradable.
  - Demanda Química de Oxígeno (DQO): mide la cantidad total de materia oxidable (biodegradable y no biodegradable) usando un oxidante químico fuerte.

  La relación DBO/DQO indica la tratabilidad biológica del agua.

- Sólidos: se clasifican en Sólidos Totales (ST), que a su vez se dividen en Sólidos Disueltos Totales (SDT) y Sólidos Suspendidos Totales (SST). Los SST son la fracción que puede removerse por medios físicos como la sedimentación. Una fracción de los SST es materia orgánica, cuantificada como Sólidos Suspendidos Volátiles (SSV), que representa la biomasa microbiana y la materia orgánica particulada.

- Nitrógeno: presente en formas orgánicas e inorgánicas. El Nitrógeno Total Kjeldahl (TKN) representa la suma del nitrógeno orgánico y el amonio. El amonio (NH_4^+) es la forma inorgánica predominante en el agua residual cruda. Los procesos de tratamiento lo convierten en nitrito (NO_2^-) y nitrato (NO_3^-) (Campins-Falco et al., 2008).

  En la Figura 1.1 se observa el ciclo del nitrógeno general en una PTAR.

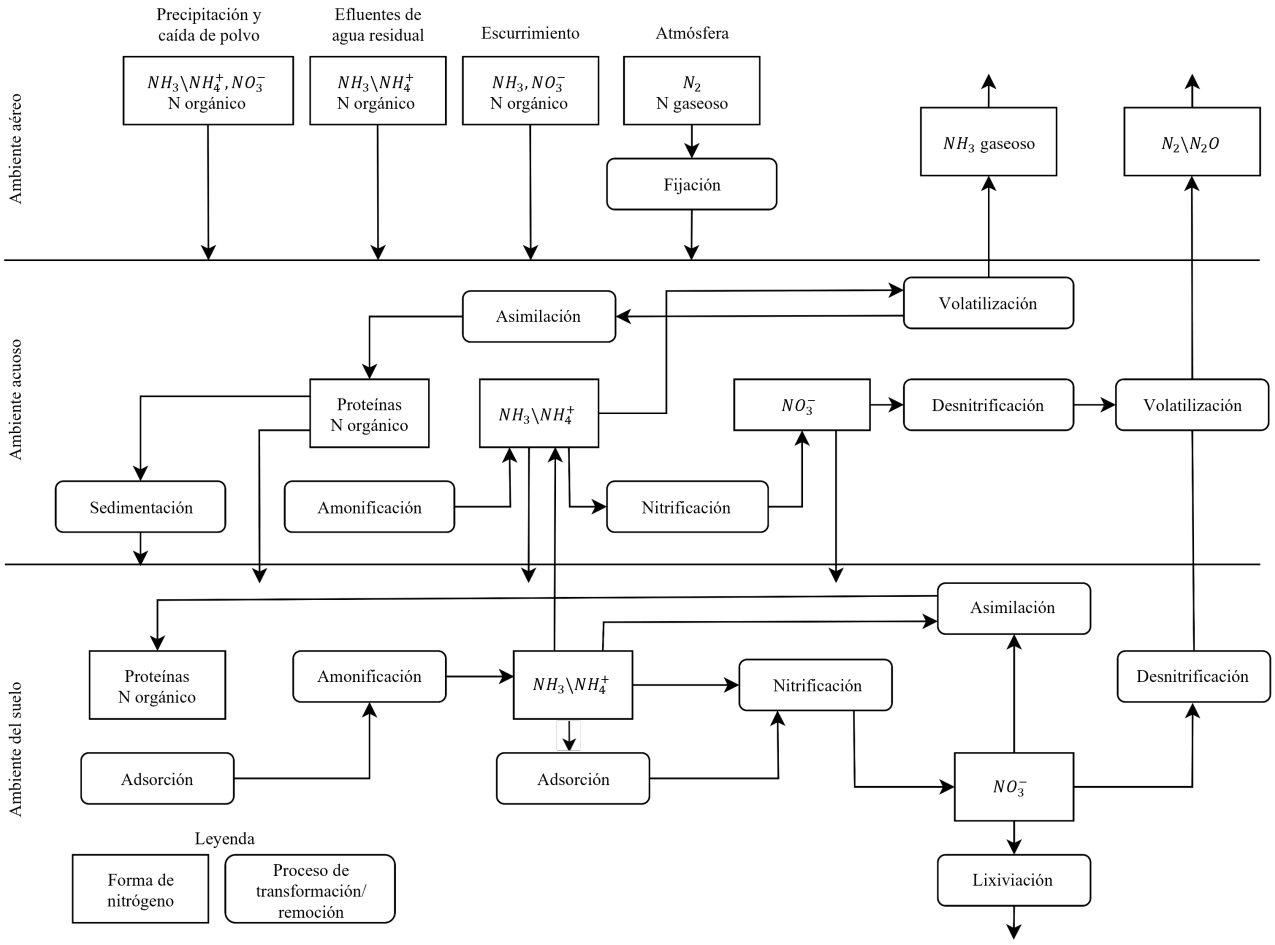


Figura 1.1: Procesos generales del nitrógeno en una PTAR (Metcalf & Eddy Inc. et al., 2014)

- Fósforo: presente como ortofosfato (PO_4^{3-}), polifosfatos y fósforo orgánico.
- Patógenos: microorganismos causantes de enfermedades, incluyendo bacterias (como E. coli), virus y protozoos.
- Contaminantes Emergentes y Tóxicos: incluyen metales pesados, pesticidas, productos farmacéuticos y otros compuestos químicos que pueden ser tóxicos para la vida acuática o humana, aunque su remoción no siempre es el objetivo primario de las PTAR convencionales.

1.0.3 Las PTAR como foco de emisiones de GEI

Históricamente, el éxito de una PTAR se medía casi exclusivamente por la calidad de su efluente. Sin embargo, en las últimas décadas, ha surgido una preocupación creciente por el impacto ambiental secundario de estas instalaciones: su contribución a las emisiones de gases de efecto invernadero

(GEI). Las PTAR son ahora reconocidas como fuentes antropogénicas significativas de dióxido de carbono (CO_2), metano (CH_4) y, de manera crítica, óxido nitroso (N_2O).

Mientras que el CO_2 es el GEI más abundante, el CH_4 y el N_2O son mucho más potentes. Las emisiones de N_2O son particularmente problemáticas, ya que se estima que pueden ser responsables de más del 50% de la huella de carbono total de una PTAR que implementa eliminación biológica de nitrógeno (BNR). A nivel global, se estima que el sector del tratamiento de aguas residuales es responsable de aproximadamente el 2% de las emisiones de carbono globales (Huang et al., 2024).

Por lo tanto, las PTAR modernas enfrentan un desafío complejo: deben cumplir con estándares de efluente cada vez más estrictos (especialmente para nutrientes como el nitrógeno), pero los mismos procesos biológicos implementados para remover el nitrógeno (nitrificación y desnitrificación) son las principales fuentes de emisión de N_2O. Esta dualidad hace que la cuantificación precisa y el monitoreo fiable de estas emisiones sean un primer paso esencial para desarrollar estrategias de mitigación efectivas.

1.1 Proceso de tratamiento de aguas residuales

Para entender dónde y cómo se generan los GEI, es esencial primero comprender el flujo de procesos en una PTAR convencional. El proceso de tratamiento de aguas residuales se encuentra explicado de manera detallada en el libro de Metcalf & Eddy Inc. et al. (2014).

En resumen, el tratamiento se divide en dos corrientes principales: la línea de agua, que procesa el agua residual hasta convertirla en un efluente apto para su descarga o reutilización, y la línea de lodos, que gestiona los sólidos removidos durante el proceso.

1.1.1 Línea de agua

Pre-tratamiento

El pre-tratamiento o tratamiento preliminar tiene como objetivo remover los materiales gruesos, pesados o flotantes que podrían dañar equipos mecánicos (bombas, aireadores) o interferir con los procesos posteriores. Incluye:

- Desbaste: consiste en pasar el agua a través de pantallas o rejas (finas o gruesas) para retener sólidos de gran tamaño como plásticos, trapos, maderas y otros desechos voluminosos que podrían obstruir tuberías o dañar bombas.
- Desarenado: el agua fluye a través de "desarenadores" (canales o tanques aireados) donde se reduce la velocidad del flujo para permitir que los sólidos inorgánicos pesados (arena, grava, cáscaras de huevo) sedimenten por gravedad, mientras la materia orgánica más ligera

permanece en suspensión. Esto previene la abrasión en equipos mecánicos y la acumulación de inertes en los reactores biológicos.

- Desengrasado: remoción de aceites y grasas, que son menos densos que el agua y flotan en la superficie.

Tratamiento primario

Tras el pre-tratamiento, el agua fluye hacia los sedimentadores primarios (también llamados clarificadores primarios). Estos son grandes tanques, usualmente circulares o rectangulares, donde el agua permanece en condiciones de quietud (baja velocidad de flujo) durante varias horas. Durante este tiempo, la gravedad actúa sobre los sólidos suspendidos (SST) que son más densos que el agua, haciéndolos decantar hacia el fondo. Esta masa de sólidos acumulada se denomina lodo primario y es bombeada fuera para su tratamiento en la línea de lodos. La sedimentación primaria puede remover entre el 50% y 70% de los SST y entre el 30% y 40% de la DBO de la corriente de agua de proceso.

Tratamiento secundario (biológico)

Esta es la etapa principal y más compleja de la PTAR, diseñada para eliminar la materia orgánica disuelta y coloidal (que no sedimentó en el primario) y los nutrientes (N y P). El método más común es el de Lodos Activados (CAS, en inglés *Conventional Activated Sludge*).

El principio del CAS es cultivar un consorcio microbiano denso (el "lodo activado" o "licor de mezcla", medido como Sólidos Suspendidos Volátiles del Licor de Mezcla, SSVLM) en un reactor biológico o tanque de aeración. Este consorcio microbiano utiliza los contaminantes orgánicos disueltos como "alimento" para su crecimiento y respiración. Se suministra oxígeno (generalmente mediante difusores de aire) para mantener la actividad aeróbica. Posteriormente, el licor de mezcla pasa a un sedimentador secundario, donde la biomasa se separa del agua tratada. Una gran parte de esta biomasa (lodo) se recircula al inicio del reactor biológico (Recirculación de Lodos Activados, RAS) para mantener una alta concentración de microorganismos. El exceso de biomasa producido (Lodo Activado de Purga, WAS) se retira del sistema para su tratamiento en la línea de lodos.

1. Remoción de Materia Orgánica (DBO/DQO): En presencia de oxígeno (condiciones aeróbicas), los microorganismos heterótrofos oxidan la materia orgánica para obtener energía, produciendo CO_2, agua y nueva biomasa (más lodo):

$$\text{Materia Orgánica}(C_xH_yO_z) + O_2 \xrightarrow{\text{Microorganismos}} CO_2 + H_2O + \text{Biomasa} \tag{1.1}$$

   Este es el proceso fundamental de tratamiento del agua y la principal fuente de CO_2 biogénico.

2. Eliminación Biológica de Nitrógeno (BNR): la remoción de nitrógeno es un proceso biológico más complejo que requiere dos etapas distintas, realizadas por dos tipos diferentes de bacterias en ambientes opuestos:

   (a) Nitrificación (aeróbica): ocurre en la zona oxigenada del reactor. Es un proceso autótrofo (los microorganismos obtienen carbono del CO_2 o el bicarbonato) en dos pasos:

   - Amonio-oxidación: bacterias Amonio-Oxidantes (AOB), como Nitrosomonas, oxidan el amonio a nitrito: $2NH_4^+ + 3O_2 \rightarrow 2NO_2^- + 4H^+ + 2H_2O$
   - Nitrito-oxidación: bacterias Nitrito-Oxidantes (NOB), como Nitrospira o Nitrobacter, oxidan el nitrito a nitrato: $2NO_2^- + 1O_2 \rightarrow 2NO_3^-$ El resultado neto es la conversión del amonio, el cual es tóxico para los peces en nitrato, siendo este menos tóxico, pero aún un nutriente.

   (b) b) Desnitrificación (anóxica): ocurre en una zona anóxica (sin oxígeno disuelto, pero con presencia de nitrato). Microorganismos heterótrofos facultativos (que pueden respirar O_2 o NO_3^-) utilizan la materia orgánica del agua residual (DBO) como fuente de carbono y el nitrato (producido en la nitrificación) como aceptor de electrones, liberando N_2, gas inerte y no contaminante, a la atmósfera: NO_3^- + Materia Orgánica $\rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow \mathbf{N_2} + CO_2 + H_2O$

Para lograr ambas etapas, las PTAR que incluyen la eliminación biológica de nitrógeno deben tener configuraciones que incluyan zonas óxicas y anóxicas. Ejemplos de estas configuraciones, relevantes para la generación de GEI, incluyen:

- A2O (anaeróbica/anóxica/oxica): el agua fluye secuencialmente a través de un tanque anaerobio, un tanque anóxico (para desnitrificación, recibiendo nitrato recirculado desde la zona óxica) y un tanque óxico (para remoción de DBO y nitrificación). Se puede observar un diagrama de un proceso asi en la Figura 1.2.

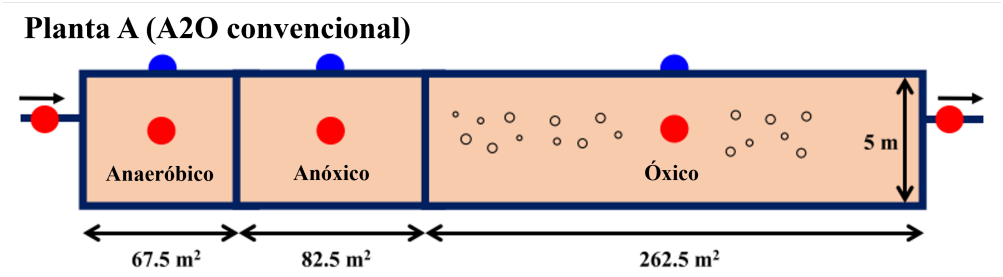


Figura 1.2: Ejemplo de configuración A2O (Song et al., 2020)

- SBR (*Sequencing Batch Reactor*): todo el tratamiento ocurre en un solo tanque, pero en ciclos de tiempo. Se alternan fases de reacción aeróbica (con aeración, para nitrificar y remover DBO) y reacción anóxica (solo mezcla, sin aeración, para desnitrificar). En la Figura 1.3 se observa un ejemplo de operación con ciclos de aireación.

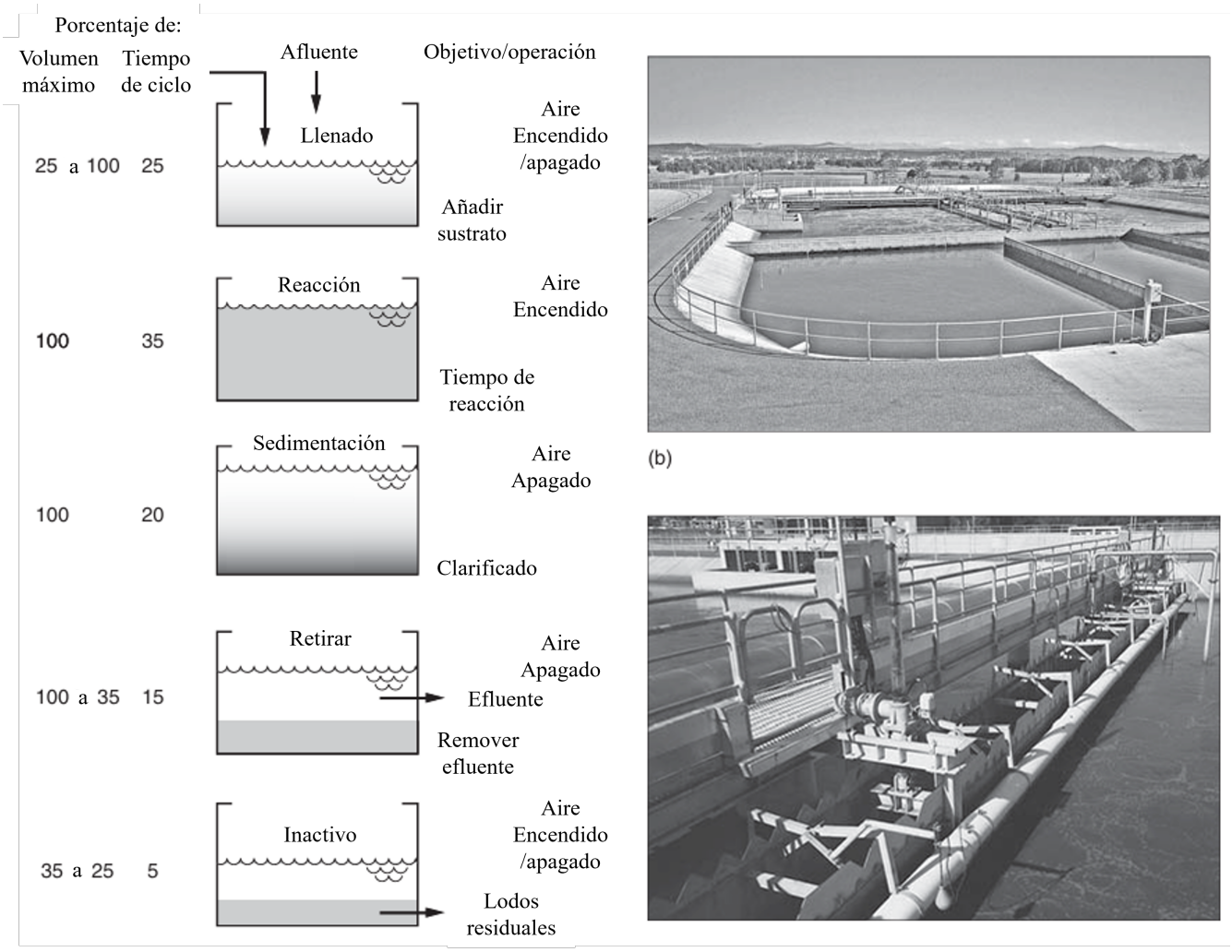


Figura 1.3: Ejemplo de configuración SBR (Metcalf & Eddy Inc. et al., 2014)

Tratamiento terciario

Tras una etapa simple de sedimentación secundaria, donde la biomasa se separa del agua tratada, el efluente puede pasar a un tratamiento terciario para maximizar su calidad, especialmente si se destina a reutilización. Esto incluye:

- Filtración: remoción de los últimos sólidos suspendidos finos que no sedimentaron.
- Desinfección: Eliminación de patógenos residuales, comúnmente mediante cloración (adición de hipoclorito) o radiación ultravioleta (UV).

1.1.2 Línea de lodos

El tratamiento del agua genera grandes volúmenes de lodo (primario y secundario), que es una suspensión acuosa de sólidos orgánicos e inorgánicos. La línea de lodos tiene como objetivo reducir su volumen, estabilizar la materia orgánica (para reducir olores y patógenos) y prepararlo para su disposición final.

En la Figura 1.4 se puede ver el proceso general del tratamiento en la línea de lodos para diferentes tipos de PTAR.

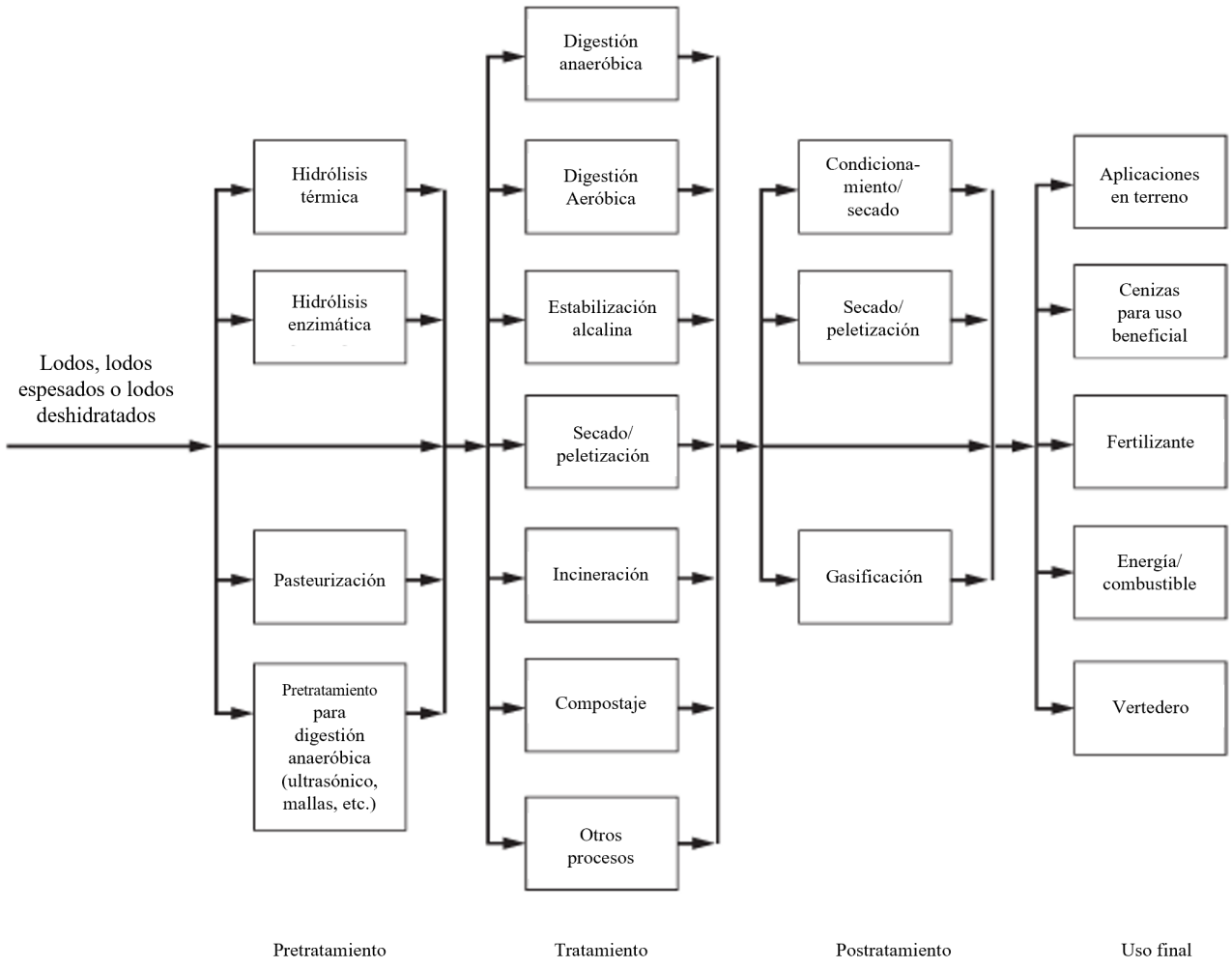


Figura 1.4: Procesamiento de lodos en PTAR (Metcalf & Eddy Inc. et al., 2014)

Espesamiento

El lodo combinado tiene un alto contenido de agua (95-99%). El espesamiento (por gravedad o medios mecánicos como flotación o centrifugación) es un primer paso para reducir el volumen, eliminando parte del agua libre.

Estabilización

Este es el proceso central de la línea de lodos y su foco principal de emisiones de GEI.

- Digestión Anaerobia (DA): es el proceso más común para estabilizar lodos en PTAR medianas y grandes. El lodo se introduce en un gran reactor sellado (digestor), en ausencia de oxígeno y a temperatura controlada (usualmente 35-40°C) durante 15-30 días. Aquí ocurre una compleja descomposición biológica en cuatro etapas:

    1. Hidrólisis: enzimas extracelulares rompen las macromoléculas orgánicas complejas (proteínas, lípidos, polisacáridos) en compuestos solubles más simples (aminoácidos, ácidos grasos, azúcares).
    2. Acidogénesis (o fermentación): bacterias fermentativas convierten estos compuestos solubles en una mezcla de Ácidos Grasos Volátiles (AGV) (como acético, propiónico, butírico), alcoholes, CO_2 e hidrógeno (H_2).
    3. Acetogénesis: bacterias acetogénicas convierten los AGV de cadena larga y alcoholes en acetato, H_2 y CO_2.
    4. Metanogénesis: Arqueas completan el proceso. Existen dos grupos principales:
        (a) Metanógenas acetoclásticas: convierten acetato en CH_4 y CO_2.
        (b) Metanógenas hidrogenotróficas: combinan H_2 (como donador de electrones) y CO_2 (como aceptor) para producir CH_4. El producto final es un biogás (compuesto principalmente por 60-70% de CH_4 y 30-40% de CO_2) y un lodo estabilizado llamado digestato.

- Digestión Aerobia: el lodo se airea en un tanque abierto durante un período prolongado. Los microorganismos consumen la materia orgánica y, una vez agotada, entran en fase de respiración endógena, consumiendo su propia biomasa. Este proceso consume mucha energía (aeración) y produce CH_4 como GEI principal.

Deshidratación y disposición final

El lodo estabilizado (digestato) aún contiene mucha agua. Se deshidrata mecánicamente (usando centrífugas, filtros prensa o tornillos) para convertirlo en un biosólido. Este biosólido final se transporta para su disposición final, que puede ser la aplicación agrícola (como mejorador de suelos), el compostaje, la incineración o el depósito en rellenos sanitarios.

1.2 Métodos de cuantificación de GEI

Dada la alta variabilidad de las emisiones, especialmente de N_2O, su cuantificación precisa es un desafío. Los métodos para estimar o medir las emisiones de GEI de las PTAR se dividen en tres categorías principales:

1.2.1 Método de Factores de Emisión (IPCC)

Este enfoque, propuesto por el Panel Intergubernamental sobre el Cambio Climático (IPCC), es el método más simple. Consiste en multiplicar datos de actividad (e.g., cantidad de nitrógeno que entra a la planta, cantidad de DBO removida) por factores de emisión (FE) estandarizados. (Huang et al., 2024)

- Ventajas: es un método sencillo, de bajo costo, y útil para realizar inventarios nacionales o regionales y comparaciones a gran escala.
- Desventajas: su principal inconveniente es la alta incertidumbre y la falta de especificidad. Los FE son promedios generales que no capturan la especificidad del proceso, las condiciones operativas, ni la variabilidad temporal/estacional . Numerosos estudios de monitoreo directo han reportado FE que difieren en órdenes de magnitud de los valores por defecto del IPCC.

1.2.2 Método de Modelación (MCM y ML)

Este enfoque utiliza modelos matemáticos para simular los procesos biológicos y químicos dentro de la PTAR y predecir las emisiones resultantes.

Modelos Mecanísticos (MCM): se basan en el conocimiento fundamental de los procesos. Los Modelos de Lodos Activados (ASM) se han extendido para incluir las rutas de producción de GEI. Por ejemplo, el modelo ASM2d-N_2O (Massara et al., 2018a) integra las tres rutas biológicas de producción de N_2O (AOB-hidroxilamina, AOB-desnitrificación y HD-incompleta). Szelag et al. (2023) utiliza un MCM (ASM2d-N_2O) como un modelo de referencia para verificar y validar la lógica de los modelos de *machine learning*.

Modelos de Machine Learning (ML): son modelos basados en datos que no requieren un conocimiento profundo de los mecanismos subyacentes. Utilizan algoritmos como *Random Forest* (RF), Redes Neuronales de Aprendizaje Profundo (DNN), *Long Short-Term Memory* (LSTM), XGboost o Máquinas de Vectores de Soporte (SVM) para encontrar correlaciones complejas entre los parámetros operativos de la planta (variables de entrada, como OD, T, NH_4^+, caudal) y las emisiones de GEI (variable de salida) (Song et al., 2020).

- Ventajas: permiten simular escenarios, optimizar operaciones y capturar relaciones complejas no lineales.

- Desventajas: ambos enfoques son altamente dependientes de la disponibilidad y calidad de los datos para su calibración y validación . Los MCM requieren una calibración extensa y los modelos de ML necesitan grandes conjuntos de datos de entrenamiento para ser fiables.

1.2.3 Método de Monitoreo Directo (*Field Monitoring*)

Este método implica la medición física de las concentraciones de gases en los puntos de emisión y la medición del caudal de gas (*off-gas*) para calcular el flujo másico (tasa de emisión). Se puede realizar de forma discontinua (toma de muestras en bolsas y análisis posterior en laboratorio, comúnmente por Cromatografía de Gases) o de forma continua/semicontinua (*online*) utilizando sensores instalados in situ, como los sensores NDIR.

- Ventajas: es el método estándar que proporciona los datos más precisos y específicos del sitio. Es indispensable para la calibración de los modelos MCM y ML.

- Desventajas: es costoso, complejo, y la variabilidad es un desafío. Estudios de monitoreo a corto plazo (por ejemplo, días o semanas) pueden ser engañosos, ya que no capturan eventos esporádicos de alta emisión o variaciones estacionales. Vasilaki et al. (2019) y Huang et al. (2024) han demostrado que las campañas de monitoreo a corto plazo (1 mes) tienden a subestimar significativamente las emisiones anuales en comparación con las campañas a largo plazo (1 año), que son esenciales para capturar la variabilidad estacional.

La necesidad de un monitoreo a largo plazo, fiable y rentable para capturar la verdadera dinámica de las emisiones de GEI ha impulsado la adopción de sensores en línea como los NDIR. Sin embargo, para que estos datos a largo plazo sean válidos, la precisión del sensor es fundamental. Un error sistemático en el sensor, como el causado por la sensibilidad cruzada, invalidaría los datos recopilados. Por lo tanto, comprender, cuantificar y corregir este fenómeno, es un paso crítico y necesario para permitir un monitoreo de GEI preciso y fiable en las PTAR.

1.3 Motivación principal

La gestión de las emisiones de gases de efecto invernadero (GEI) es un desafío crítico en el contexto del cambio climático. Como se mencionó anteriormente, las plantas de tratamiento de aguas residuales (PTAR) representan una fuente significativa de emisiones de óxido nitroso (N_2O), un gas con un potencial de calentamiento global 298 veces mayor que el del dióxido de carbono (CO_2). El

N_2O afecta directamente a la capa de ozono, reaccionando con radicales de oxígeno gracias a la radiación UV, generando moléculas de NO. Estas terminan reaccionando con el ozono y generando oxígeno, destruyendo asi la capa protectora de la atmósfera (Crutzen, 1978).

Estas emisiones son el resultado de procesos biológicos complejos relacionados con la eliminación de nutrientes, donde factores operacionales y ambientales influyen de manera significativa en su generación.

Actualmente, la cuantificación precisa y la predicción de estas emisiones presentan desafíos metodológicos, ya que los datos disponibles suelen ser limitados o inexactos debido a la complejidad inherente de los sistemas de tratamiento de aguas residuales. En este contexto, los sensores específicos para N_2O y las cámaras de flujo flotantes emergen como herramientas clave para registrar emisiones en tiempo real y con alta resolución espacial y temporal.

El objetivo principal de esta investigación es desarrollar un modelo basado en aprendizaje automático que permita predecir las emisiones de N_2O en una PTAR, integrando datos obtenidos de sensores de N_2O y otros parámetros operacionales y no operacionales relevantes. Este enfoque no solo busca mejorar la precisión de la estimación de emisiones, sino también proporcionar recomendaciones prácticas para mejorar el desempeño ambiental de las PTAR y mitigar el impacto de sus emisiones en el cambio climático.

Esta tesis contribuirá al entendimiento y la gestión sostenible de las emisiones de N_2O en PTAR, promoviendo soluciones innovadoras y aplicables en el ámbito de la ingeniería ambiental y la mitigación del cambio climático.

1.4 Pregunta de investigación e hipósitesis

1.4.1 Pregunta de investigación

¿Es viable desarrollar un modelo de emisión de N_2O basado en el análisis de datos, que combine técnicas de *Machine Learning*, PCA y fenomenología, desarrollando un modelo híbrido, para predecir y controlar en línea las emisiones de N_2O, superando las limitaciones de los modelos mecanicistas y de Deep Learning con la información disponible?

1.4.2 Hipósitesis

Si se aplican herramientas de *Machine Learning* a registros de sensores NDIR de concentración y variables de operación en una PTAR, entonces es posible desarrollar un modelo híbrido capaz de predecir las emisiones de N_2O de manera más efectiva que supera las limitaciones de los modelos mecanicistas y de Deep Learning.

Capítulo 2

Objetivos y Alcances

2.1 Objetivo General

Desarrollar un modelo híbrido para predecir las emisiones de N_2O en una planta de tratamiento de aguas residuales, integrando el comportamiento mecanicista de la generación de N_2O y herramientas de aprendizaje automático, basado en datos experimentales.

2.2 Objetivos Específicos

1. Aplicar análisis de componentes principales (PCA) como herramienta de reducción de dimensionalidad para simplificar las variables de entrada del modelo en caso de ser posible.

2. Comparar el rendimiento de diferentes modelos de aprendizaje automático, evaluando métricas como precisión, capacidad de generalización y eficiencia computacional.

3. Analizar y cuantificar el impacto de la sensibilidad cruzada en los sensores de tipo infrarrojo no dispersivo (NDIR) en la calidad de los datos utilizados para el modelo.

Capítulo 3

Marco Teórico

3.1 Tratamiento de aguas residuales

La evolución en el monitoreo de N_2O en las PTAR ha sido notable en las últimas décadas, pasando de métodos teóricos a técnicas avanzadas de medición directa en sistemas a escala real. Inicialmente, las emisiones se estimaban mediante factores de emisión (*Emission Factors*, EF) calculados de forma teórica, lo que frecuentemente resultaba en una subestimación significativa de su impacto ambiental. Este enfoque fue paulatinamente reemplazado por campañas de monitoreo directo, lo que permitió recopilar datos más precisos sobre la dinámica de las emisiones y los factores que las condicionan.

Los estudios recientes han mostrado que la duración de los monitoreos es crucial para capturar la variabilidad temporal, particularmente la asociada a los cambios estacionales. Los monitoreos de largo plazo, que abarcan al menos un año, son esenciales para identificar patrones estacionales que influyen en la formación de N_2O y para obtener un panorama más representativo de las emisiones a escala real. En contraste, los monitoreos de corto plazo suelen ofrecer datos limitados y, en ocasiones, subestiman las emisiones reales debido a su incapacidad de reflejar fluctuaciones a largo plazo.

El desarrollo de herramientas y técnicas avanzadas ha enriquecido el entendimiento de los procesos involucrados en la producción de N_2O. La integración de datos operativos y mediciones de N_2O se ha logrado mediante modelos mecanísticos, análisis estadístico multivariado y técnicas de minería de datos (Vasilaki et al., 2019). Estos enfoques permiten identificar relaciones complejas entre variables operativas y las emisiones, proporcionando una base para optimizar los procesos y mitigar las emisiones.

La creciente preocupación por el impacto ambiental de las PTAR ha llevado a priorizar el desarrollo de estrategias de mitigación de emisiones de N_2O. Estas estrategias buscan optimizar las condiciones operativas y reducir las emisiones asociadas. A pesar de los avances logrados, persisten

desafíos importantes. Estos desarrollos no solo mejoran la comprensión de las emisiones de N_2O, sino que también fortalecen la capacidad de las PTAR para mitigar su impacto ambiental (Gruber et al., 2020).

3.2 Producción de N_2O

Según los estudios de Massara et al. (2018b), la producción de N_2O está vinculada a los procesos bioquímicos de nitrificación y desnitrificación. Se han identificado tres rutas microbianas principales para su formación:

1. Oxidación de hidroxilamina (NH_2OH): Esta es una de las dos rutas metabólicas llevadas a cabo por las Bacterias Oxidadoras de Amonio (AOB).

2. Desnitrificación nitrificante: Es la segunda ruta dependiente de las AOB. En este proceso, el nitrito (NO_2^-) actúa como aceptor de electrones en condiciones de oxígeno limitado, lo que conduce a la producción de N_2O.

   Estas dos vías metabólicas de las AOB se consideran las principales contribuyentes a la producción total de N_2O en las PTAR.

3. Desnitrificación heterotrófica: En esta vía, el N_2O es un producto intermedio en la reducción de nitrato (NO_3^-) a nitrógeno gaseoso (N_2) por parte de organismos heterótrofos.

Diversos factores operativos influyen significativamente en la generación de N_2O. Los más relevantes son:

- Bajos niveles de oxígeno disuelto (OD): Concentraciones insuficientes de OD inhiben la nitrificación completa, favoreciendo la acumulación de intermediarios y la activación de rutas productoras de N_2O.

- Altas concentraciones de nitrito (NO_2^-): La acumulación de nitrito, tanto en la nitrificación como en la desnitrificación, es un conocido detonante para la producción de N_2O.

- Baja relación DQO/N: Una relación baja entre la demanda química de oxígeno (DQO) y el nitrógeno durante la desnitrificación también contribuye a mayores emisiones.

Un fenómeno clave es que, aunque el N_2O es un intermedio de la desnitrificación (un proceso anóxico), son los compartimentos aeróbicos de las PTAR los que a menudo se identifican como los principales focos de emisión. Esto se debe al efecto de arrastre, donde la aireación necesaria para la nitrificación transfiere el N_2O disuelto en el licor mixto hacia la fase gaseosa, liberándolo a la atmósfera.

3.3 Análisis de datos

En el ámbito de la investigación sobre tratamiento de aguas residuales, se han desarrollado diversos enfoques analíticos para modelar y predecir las emisiones de N_2O.

Se han empleado técnicas de análisis multivariante para comprender la relación entre las emisiones de N_2O y las condiciones operativas durante la eliminación biológica de nitrógeno en una planta de tratamiento de aguas residuales. Se aplica PCA para reducir las dimensiones de los datos y revelar patrones que explican la variabilidad en las emisiones. En la investigación de Vasilaki et al. (2018), se utilizaron métodos de *clustering* jerárquico k-means, lo que permitió categorizar los datos en grupos asociados a diferentes perfiles de emisiones. Asimismo, se aplicó un algoritmo de segmentación binaria para identificar cambios en el comportamiento de las emisiones y dividir el conjunto de datos en subperíodos específicos. Además, se realizó un análisis de correlación de Spearman que mostró cómo las relaciones entre las variables operativas y las emisiones. Este enfoque integró datos de largo plazo obtenidos mediante sensores en línea.

Tambinén se han probado métodos de recopilacoón de datos para desarrollar estrategias de mitigación de emisiones de N_2O en tiempo real, utilizando datos obtenidos de sensores en una planta. En el trabajo de Bellandi et al. (2020) el análisis comenzó con un preprocesamiento basado en percentiles, que permitió construir patrones diarios representativos de las variables monitoreadas, como el amonio, los óxidos de nitrógeno, el oxígeno disuelto y el flujo de aire. Posteriormente, se aplicó PCA para identificar los componentes principales que explicaban la mayor parte de la varianza en los datos. Las concentraciones de N_2O se utilizaron como referencia para interpretar los resultados del PCA, lo que reveló la existencia conjunta de diferentes rutas de producción de N_2O, como la denitrificación de bacterias oxidadoras de amonio (AOB) y la oxidación incompleta de hidroxilamina. Finalmente, se emplearon técnicas de *clustering* jerárquico basado en densidad (HDBSCAN) para agrupar los datos según niveles altos y bajos de emisiones, lo que permitió identificar patrones específicos asociados a diferentes vías de producción de N_2O. Este enfoque demostró ser eficaz para analizar datos en tiempo real y desarrollar herramientas prácticas para la mitigación de emisiones en plantas de tratamiento de aguas residuales.

3.4 Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA), es una técnica estadística de aprendizaje no supervisado fundamental para la reducción de la dimensionalidad. Su objetivo principal es transformar un conjunto de datos con un gran número de variables, correlacionadas entre sí, en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales. Esta transformación se realiza de tal manera que se conserva la máxima varianza posible de los datos

originales en un número reducido de componentes.

El mecanismo del PCA se basa en un cambio de base del espacio de características. El primer componente principal (PC1) se define como la dirección en el espacio de datos a lo largo de la cual la varianza es máxima. El segundo componente principal (PC2) se calcula como la dirección ortogonal (perpendicular) al PC1 que captura la mayor parte de la varianza restante. Este proceso se repite sucesivamente, donde cada nuevo componente es ortogonal a los anteriores y captura la máxima varianza residual posible. El resultado es un nuevo sistema de coordenadas donde las variables (los componentes principales) son, por definición, linealmente independientes (Jolliffe, 2011).

La implementación matemática del PCA se fundamenta en el cálculo de los vectores propios y valores propios de la matriz de covarianza de los datos. Los vectores propios determinan la dirección de los nuevos ejes (los componentes principales), mientras que los valores propios indican la magnitud de la varianza capturada por cada componente.

El valor práctico del PCA reside en su capacidad para simplificar sistemas complejos con múltiples variables monitoreadas. Al reducir la dimensionalidad, no solo se facilita la visualización de patrones, tendencias y grupos ocultos en los datos, sino que también se mitigan problemas como la dependecia entre variables, se reduce el ruido y se puede mejorar la eficiencia y el rendimiento de los modelos predictivos posteriores.

Una restricción fundamental de la reducción de dimensiones por PCA es que el sistema debe estar en estado estacionario para aplicarlo. En caso de que la planta no se encuentre en estado estacionario, la alternativa para la reducción dimensional es la modelación por PLS dinámico.

3.5 *Partial Least Squares* (PLS)

La regresión de tipo *Partial Least Squares* es un método estadístico que combina aspectos del análisis de componentes principales y la regresión múltiple. Su utilidad radica especialmente cuando se requiere predecir un conjunto de variables dependientes a partir de un gran conjunto de variables independientes.

En lugar de simplemente buscar correlaciones entre las variables, como en la regresión ordinaria, PLS busca construir nuevos conjuntos de variables (llamados variables latentes) que capturen la mayor cantidad posible de información de las variables independientes. Esto es útil cuando lidiamos con conjuntos de datos grandes y relaciones complejas. El mecanismo detallado de este método se puede encontrar en el articulo de Abdi (2010).

3.6 Modelación por *Machine Learning*

El *Gradient Boosting* es un método de *Machine Learning* que construye modelos de forma secuencial, donde cada nuevo modelo, usualmente un árbol de decisión, se entrena para corregir los errores o residuos de los modelos anteriores. Este proceso sigue el gradiente de una función de pérdida para mejorar progresivamente la precisión del conjunto. XGBoost (Extreme Gradient Boosting) es una implementación optimizada de esta técnica, reconocida por su eficiencia y rendimiento.

Para comprender el mecanismo de modelado utilizando XGBoost y la naturaleza de los hiperparámetros de manera mas detallada se recomienda visitar el articulo por Bartz et al. (2023).

El ajuste de XGBoost se realiza mediante la configuración de sus hiperparámetros, los cuales son:

- `nrounds` y `eta`: nrounds o n_estimators establece el número de árboles (pasos de *boosting*) en el modelo, mientras que eta (tasa de aprendizaje) modera la contribución de cada árbol para evitar el sobreajuste. Existe un compromiso entre ambos.
- `lambda` y `alpha`: son parámetros de regularización L2 y L1, respectivamente, que controlan la complejidad del modelo para prevenir el sobreajuste.
- `subsample` y `colsample_bytree`: introducen aleatoriedad muestreando un subconjunto de los datos (*subsample*) y de las características (colsample_bytree) para entrenar cada árbol, lo que mejora la generalización.
- `max_depth`, `gamma` y `min_child_weight`: controlan la complejidad de cada árbol de decisión individual. Max_depth limita la profundidad máxima, gamma exige una mejora mínima para realizar una partición, y min_child_weight restringe el número de divisiones.

El resumen de la estructura de un modelo XGBoost es el siguiente:

1. El modelo comienza construyendo un primer árbol de decisión que realiza una predicción inicial del objetivo (en este caso, el N_2O). Naturalmente, esta primera predicción contendrá errores, conocidos como residuos.
2. Se construye un segundo árbol. La tarea de este segundo árbol no es predecir la variable objetivo directamente, sino **predecir los residuos** cometidos por el primer árbol. Su objetivo es corregir las equivocaciones de su predecesor.
3. La predicción del primer árbol se actualiza sumándole una fracción (controlada por el hiperparámetro `learning_rate`) de la predicción del segundo árbol. Esto genera un nuevo conjunto de predicciones, ahora más precisas.

4. Se calculan los nuevos residuos del modelo combinado. Posteriormente, se entrena un tercer árbol para predecir estos nuevos errores. El proceso se repite de forma iterativa (un número de veces definido por el hiperparámetro `n_estimators`), donde cada nuevo árbol se especializa en corregir los errores residuales del ensamblaje hasta ese momento.

La estructura de un modelo XGBoost es de una secuencia de árboles de decisión. La estructura general de un árbol de decisión se puede ver en la Figura 3.1.

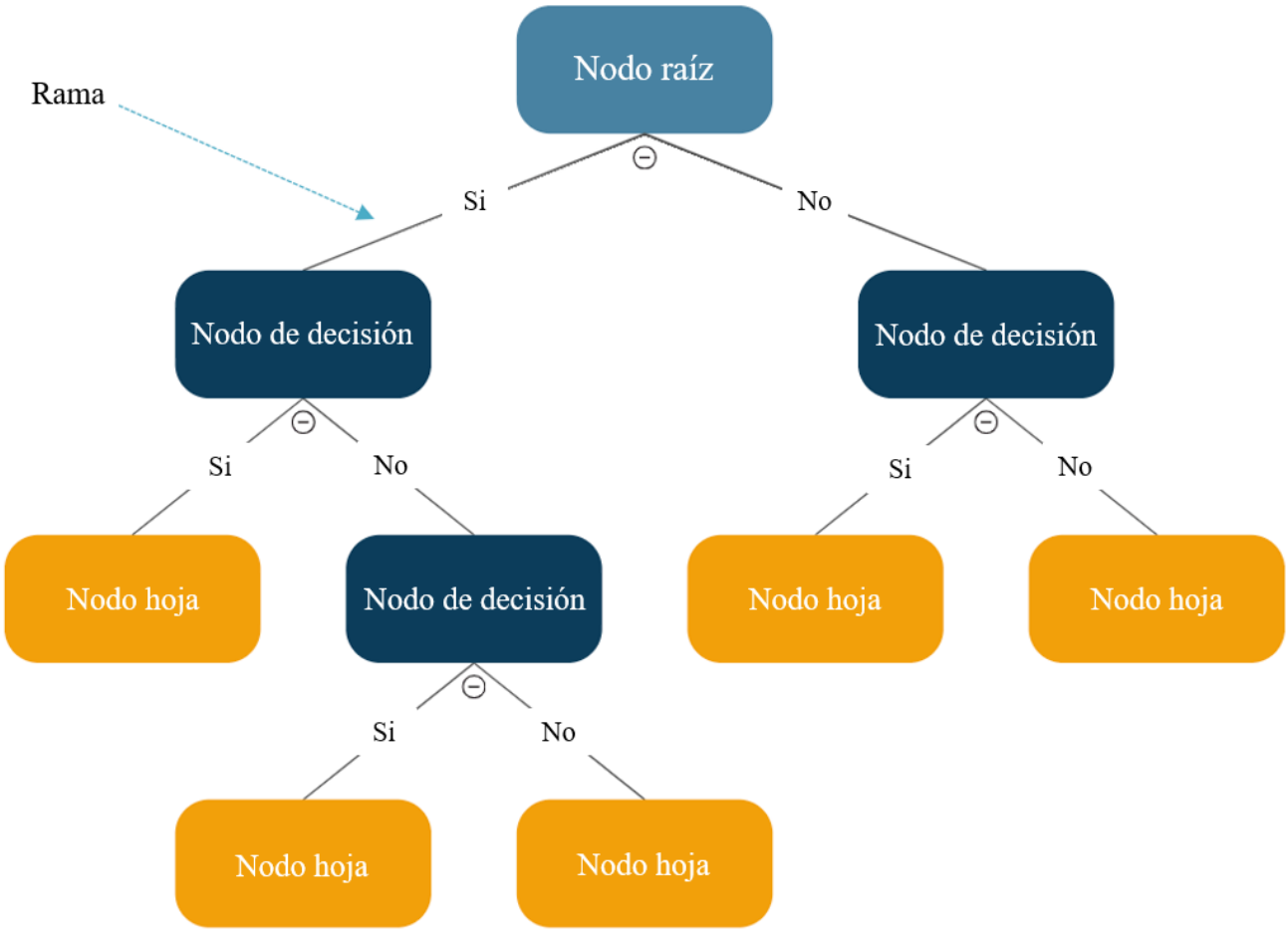


Figura 3.1: Diagrama general de un árbol de decisión

Los nodos tienen condiciones que dictan que camino del árbol se sigue. Por ejemplo un nodo en el caso del modelo a desarrollar podría ser:

- Nodo raiz: flujo de aire $>$ 1000 lpm
  - Si: OD $>$ 2 mg/L
    - Si: nodo hoja (200 ppm)

    - No: nodo hoja (100 ppm)
- No: pH > 8
    - Si: nodo hoja (300 ppm)
    - No: nodo hoja (400 ppm)

Este podría ser un árbol que prediga la concentración dependiendo de el valor de algunas variables. En el modelo real este modelo sería mucho más complejo, ya que toma en cuenta todas las variables disponibles, y además, cambia la profundidad de los arboles. Los umbrales son definidos automáticamente al realizarse la modelación por *Machine Learning*

Lo más importante es que el modelo XGBoost se forma con una secuencia de árboles de decisión (hiperparámetro *nrounds*) donde el primer árbol es el encargado de predecir directamente la concentración de gas, mientras que los demás se encargan de predecir el error o desviación que el primer árbol dentra, para que finalmente se obtenga una medición precisa luego de pasar por todos los árboles que se definan al momento de ajustar los hiperparámetros del modelo de *Machine Learning*.

La estructura general del modelo de *Machine Learning* se observa en la Figura 3.2, donde f(u,t) es la función que se utiliza para predecir las emisiones, y esta misma se puede incluir en un lazo de control para mitigarlas.

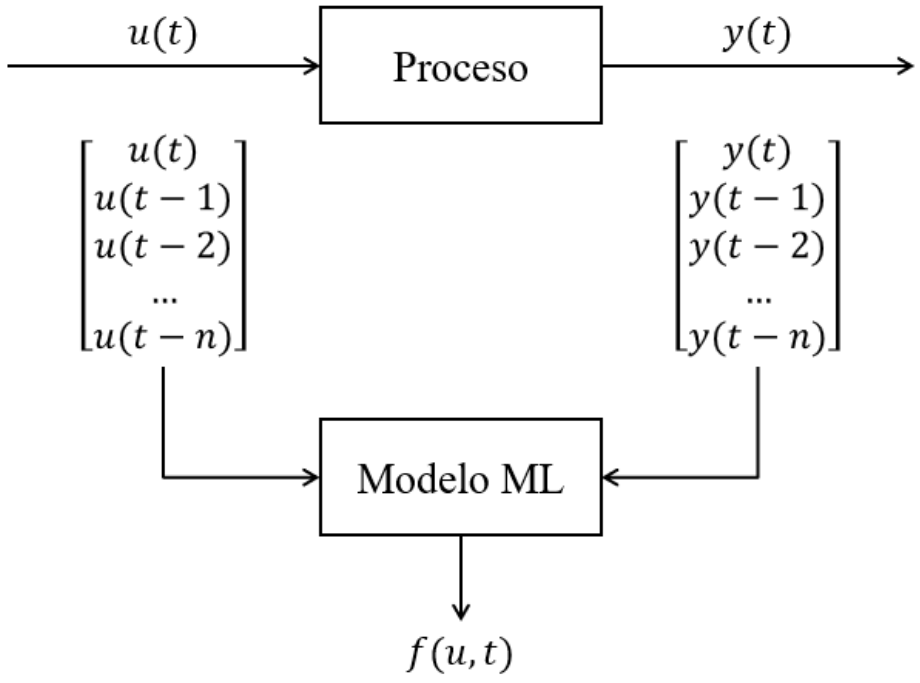


Figura 3.2: Diagrama general de un modelo ML

Capítulo 4

Materiales y Métodos

4.1 Descripción del sitio de estudio

El sitio donde se realiza el estudio es una planta de tratamiento de aguas residuales ubicada en Temuco, región de La Araucanía. La planta pemitió a Ainwater instalar las cámaras de flujo flotante en los reactores aireados del proceso. De igual manera, se permitió realizar modificaciones al código del PLC de la planta para poder tener acceso a los datos históricos.

4.2 Variables y datos recopilados

El diagrama de flujo de la Figura 4.1 detalla el proceso de la planta de tratamiento objeto de estudio. Las variables que se utilizan para el desarrollo del modelo son todas las que afectan directamente las emisiones de N_2O. Se puede observar como la alimentación se separa en flujos de lodos y de agua para el tratamiento. Las mediciones de concentración de gas son provenientes de los reactores aeróbicos, los cuales tratan el flujo de agua clarificada. Como se puede observar, algunos recirculadores devuelven lodos secundarios a los reactores, lo que hace necesario recopilar datos de ambos flujos para obtener una visión integral del proceso. Al aplicar el análisis PCA, es posible evaluar si esta decisión es adecuada o si ciertos datos pueden ser omitidos sin afectar significativamente el análisis.

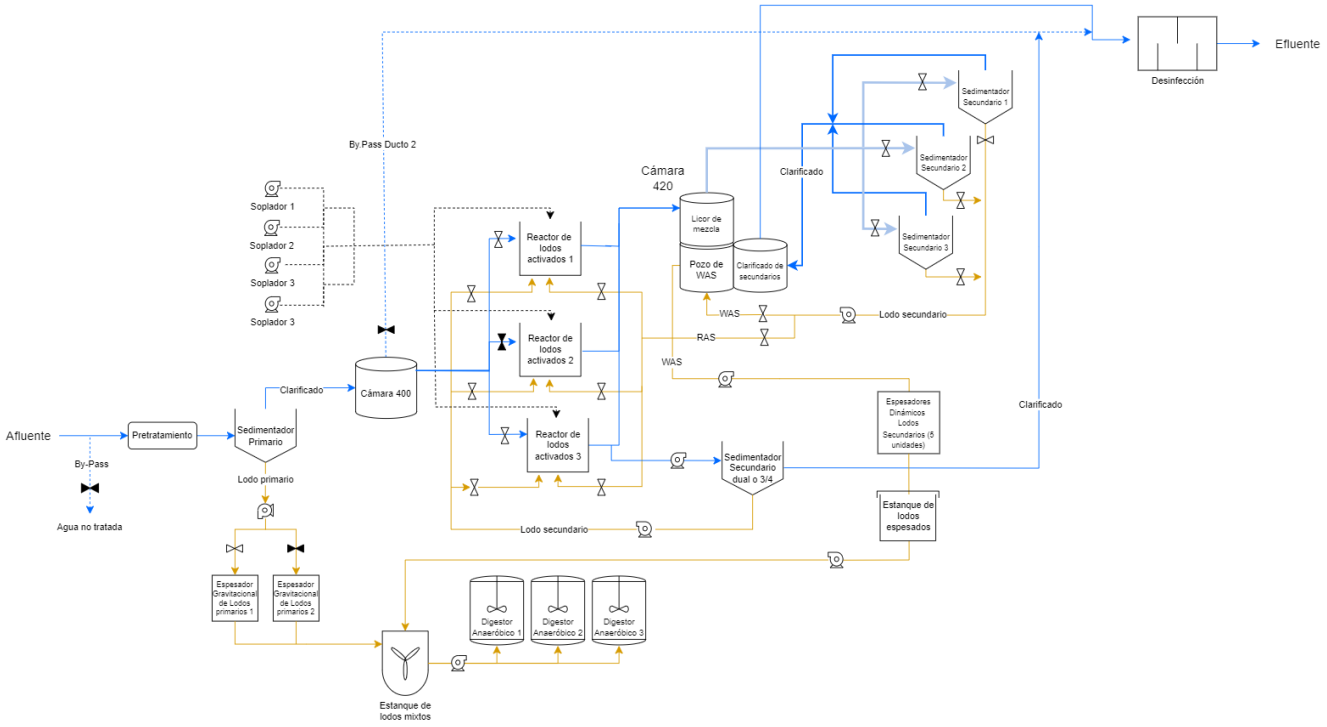


Figura 4.1: Diagrama de flujo PTAR

La recopilación de datos se lleva a cabo en dos plantas. La primera es una planta de tratamiento de aguas residuales, Aguas Araucanía, Temuco. Y la otra es la planta de tratamiento de aguas residuales de Bimbo, Quilicura. En este caso se tienen datos suficientes para realizar la modelación en la planta de Aguas Araucanía. Las emisiones de N_2O están siento monitoreadas en el Reactor 3 de la planta. En esta planta se trabaja en paralelo con 2 reactores mientras que el tercero queda como respaldo cuando se le debe hacer mantenimiento a alguno de ellos.

La lista completa de variables monitoreadas son:

- Flujo Purga Primaria
- Flujo Afluente PTAR
- Flujo ByPass
- Flujo Camara de contacto 1
- Flujo Camara de contacto 2
- Flujo Alimentacion Digestor 1
- Flujo Alimentacion Digestor 3
- Nivel Pozo WAS
- Flujo Entrada a Sedimentador Secundario 1
- Flujo Entrada a Sedimentador Secundario 2
- Flujo Entrada a Sedimentador Secundario 3
- pH Camara 400
- SST Camara 400

- Temperatura (sensor SST) Camara 400
- Flujo Salida de Lodo Sedimentador Secundario 1
- Flujo Salida de Lodo Sedimentador Secundario 2
- Flujo Salida de Lodo Sedimentador Secundario 3
- SST Clarificado de Secundarios (Camara 420)
- Temperatura Clarificado de Secundarios (Camara 420)
- Conductividad electrica Reactor 3
- Flujo Aire Reactor 3
- Flujo RAS a Reactor 3
- OD Reactor 3
- SST Reactor 3
- Temperatura Reactor 3
- Flujo RAS Total
- Flujo Entrada WAS a Espesadores Secundarios
- SST RAS/WAS
- Temperatura RAS/WAS

La frecuencia disponible de los datos varía dependiendo del sensor, hay algunas entre 1 y 5 minutos, mientras que la mayoría de mediciones ocurren principalmente cada 30 minutos y 1 hora.

4.3 Preprocesamiento de datos

La plataforma da la opción de descargar los datos crudos o de descargar promedios en ciertos periodos de tiempo. En este caso se elige descargar los datos ajustados a un promedio de 1 hora. Debido a que en este periodo de tiempo se tienen mediciones para prácticamente todas las variables de interés en el estudio.

La fecha en la que los datos estan disponibles parte desde el 21 de julio de 2025, por lo que se descargan los datos desde esa fecha hasta el 31 de julio, siendo así un poco mas de una semana de datos de concentración de N_2O, lo que abre las puertas a realizar un análisis preliminar del comportamiento de estos.

4.4 Pretratamiento de Datos

El pretratamiento de los datos crudos se realiza en dos etapas. Primero, se lleva a cabo un proceso de limpieza para manejar los valores ausentes. Se aplica una estrategia de eliminación por lista, que consiste en descartar todas las filas que contienen valores nulos o no definidos (NaN) en cualquiera de las variables monitoreadas.

Segundo, se aborda la heterogeneidad en las frecuencias de muestreo. Para homogeneizar el conjunto de datos de salida, es decir, las mediciones de concentración se realiza un promedio para ajustar los datos de manera de tener una medición cada hora y así tener una base de datos de periodo homogéneo. Se sabe que esto significa una pérdida de información, pero debido a la gran diferencia entre periodos de medición de las variables de entrada, esta es una opción muy sólida. Aunque también se puede realizar nuevamente el proceso de homogenización, dependiendo de las variables principales y su frecuencia disponible.

4.5 Detección de estado estacionario para PCA

Para realizar el PCA, se debe confirmar si efectivamente la planta cumple con estar en estado estacionario de manera consistente para así poder definir las relaciones entre las variables utilizando el análisis de componetes principales.

Se utiliza una técninca mencionada en (Rhinehart, 1995), que consiste en realizar un análisis de la media y la varianza de la variable que se quiere observar en un periodo de tiempo determinado.

Los resultados obtenidos en esta detección de estado estacionario se pueden observar en la Figura 4.2

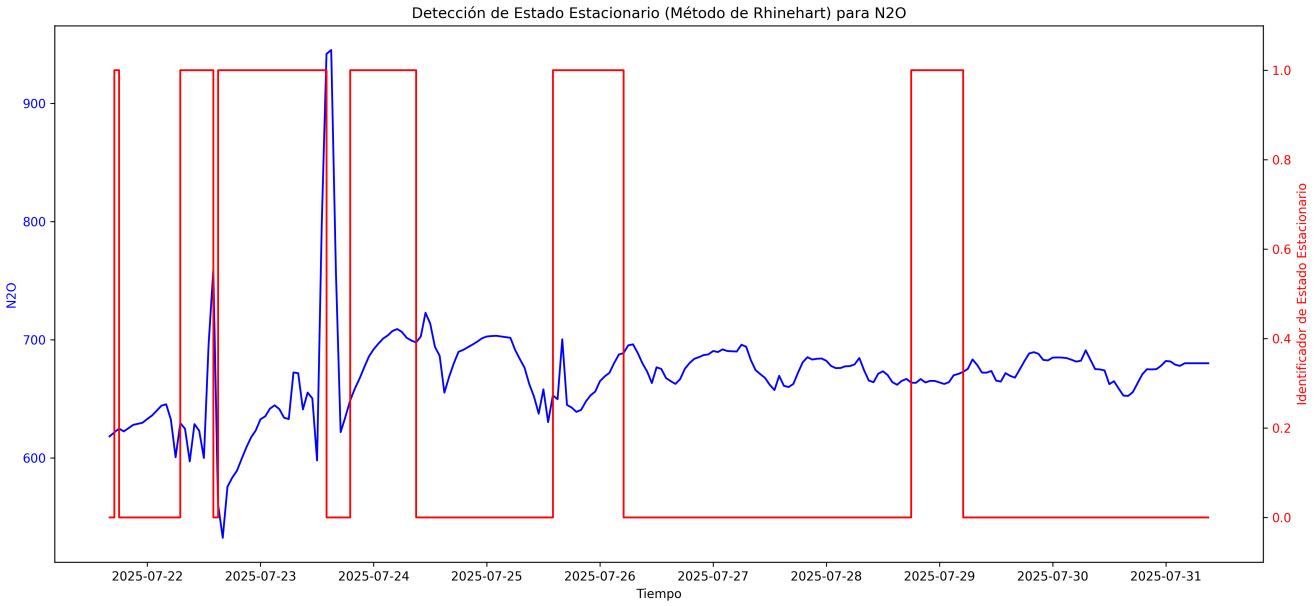
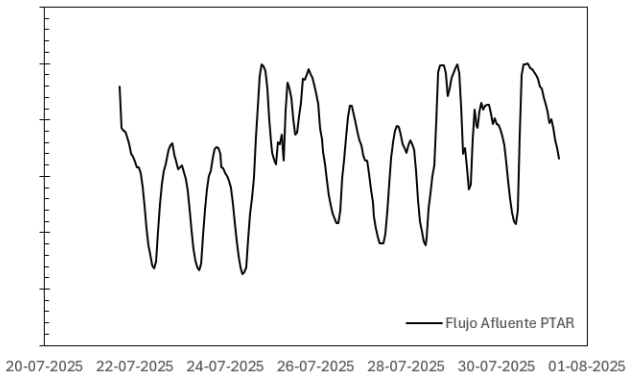
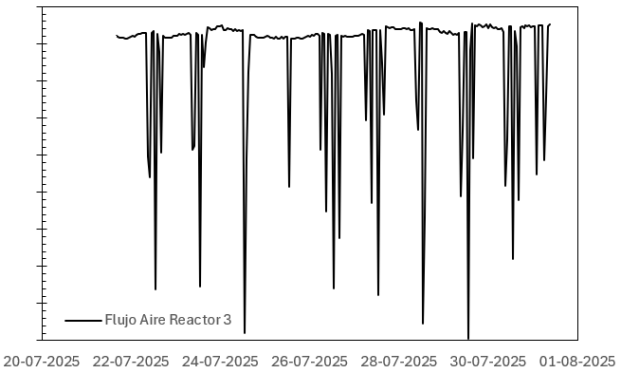


Figura 4.2: Detección de estado estacionario para PTAR (concentración de N_2O)

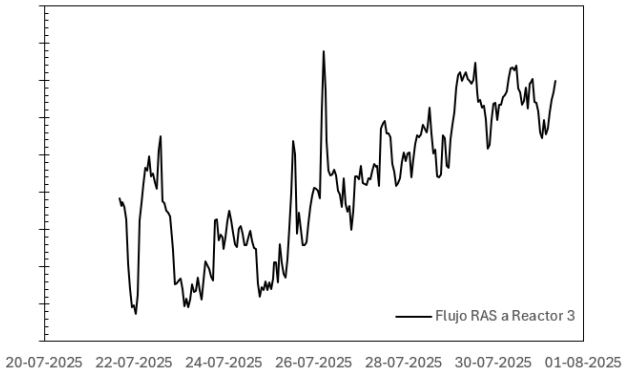
Adicionalmente se grafican las variables principales para ver su comportamiento.



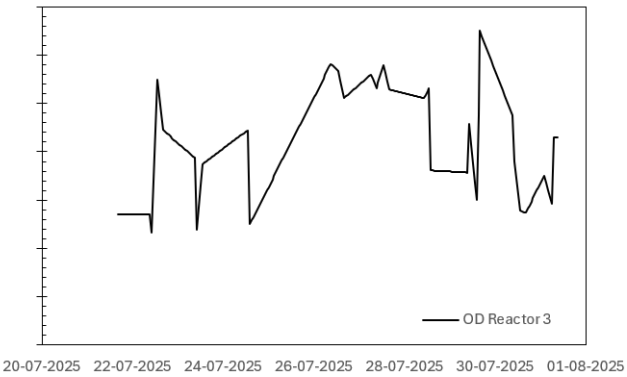
(a) Flujo Afluente PTAR



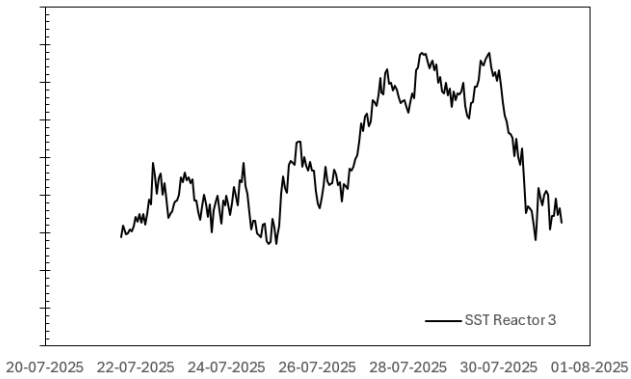
(b) Flujo Aire Reactor 3



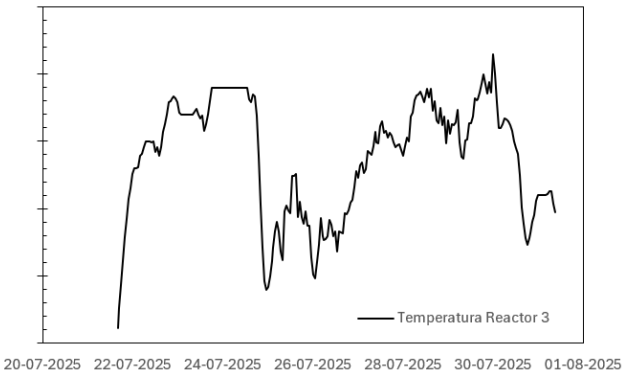
(c) Flujo RAS a Reactor 3



(d) OD Reactor 3



(e) SST Reactor 3



(f) Temperatura Reactor 3

Figura 4.3: Evolución de las principales variables y control del sistema durante el periodo de estudio. Los gráficos (a-f) ilustran la variabilidad operativa a la que está sometido el proceso biológico en el Reactor 3. Por motivos de confidencialidad y acuerdos con la planta industrial, los valores numéricos absolutos del eje vertical no se muestran.

Como se puede observar, el estado estacionario no es un estado común en el funcionamiento de la planta, por lo que no es posible aplicar PCA para reducir las dimensiones de las variables de entrada.

Esto no significa un problema, ya que se pueden utilizar herramientas como PLS dinámicos y modelos de *Machine Learning* dinámicos basados en XGBoost utilizando todas las variables disponibles.

4.6 Técnicas de modelación

Para la selección del modelo predictivo utilizado en esta investigación, se tomaron como base los resultados obtenidos en un estudio previo realizado en conjunto con Ainwater. En dicho trabajo, se analizaron bases de datos de acceso público de emisiones de N_2O provenientes de reactores biológicos en tres plantas de tratamiento de aguas residuales (Gruber et al., 2020). El objetivo fue evaluar distintas arquitecturas de modelación para comprender los comportamientos relacionales de las variables en fase líquida y gas.

Inicialmente, se evaluó la viabilidad de modelos de Regresión Lineal. Sin embargo, los resultados demostraron un rendimiento deficiente, con coeficientes de determinación bajos y una incapacidad para capturar la fenomenología no lineal de la generación de emisiones de N_2O.

Posteriormente, se exploraron modelos basados en Árboles de Decisión. Si bien estos mejoraron el ajuste respecto a la regresión lineal, el análisis de residuos evidenció problemas significativos de sobreajuste, donde los residuos tendían a cero de manera artificial, indicando una baja capacidad de generalización ante datos nuevos.

Con el fin de mitigar el sobreajuste y mejorar la capacidad predictiva, se implementaron modelos de ensamble tipo *Gradient Boosting*, específicamente LightGBM y XGBoost. El modelo LightGBM presentó un rendimiento regular; aunque superó a la regresión lineal, el análisis de residuos reveló que estos no seguían una distribución normal, sugiriendo dificultades para manejar el ruido blanco presente en ciertas variables. Por el contrario, el modelo XGBoost demostró ser la técnica más robusta y eficiente para este tipo de sistema.

Las pruebas realizadas indicaron que:

- Ajuste y Generalización: XGBoost presentó los mejores indicadores de rendimiento (R^2 y MSE), con una distribución de residuos cercana a la normalidad, lo que sugiere estabilidad en las predicciones.

- Dinámica Temporal (Lags): la incorporación de variables con retardos temporales (lags) de hasta 5 periodos (equivalente a 2.5 horas de historia en este caso) mejoró sustancialmente la calidad del modelo. Esto permitió capturar la evolución temporal del proceso, elevando la precisión del modelo a rangos de R^2 entre 0.71 y 0.82 en los reactores estudiados. Esto es un indicador de que los procesos estudiados para la modelación se encuentran en estado no estacionario.

- Validación Fenomenológica: el análisis de importancia de variables realizado sobre el modelo XGBoost fue coherente con la fenomenología física y química del proceso de tratamiento de aguas.

Considerando que XGBoost fue capaz de adaptarse a las diferentes condiciones operacionales de los reactores y demostró un rendimiento superior al integrar dinámicas temporales, se decide utilizar esta técnica para la modelación actual.

4.7 Validación de los modelos

La validación del modelo se lleva a cabo por un análisis de residuos, donde se comprueba la normalidad de estos y si están centrados en cero. Por otro lado se realiza un cálculo y comparación del error porcentual que tienen los errores para así realizar una comparación.

Por otro lado, cuando se tenga disponibilidad de más datos se planea aislar una cantidad considerable de datos para probar la predicción del modelo con datos que no han sido vistos antes.

4.8 Software y herramientas

Las herramientas principales son los sensores a instalar por la empresa y la plataforma diseñada para acceder a los datos históricos de la PTAR.

Los softwares principales a utilizar son Microsoft Excel, MATLAB, SOLO (Eigenvector Research Inc., 2025) y Visual Studio Code, este último siendo utilizado con el idioma de programación Python y sus respectivas librerias de *Machine Learning* y análisis de datos.

Capítulo 5

Resultados

5.1 Modelación de emisiones de N_2O

El hecho de que la planta no alcance un estado estacionario estable justifica la necesidad de utilizar modelos dinámicos. Estos son esenciales para capturar el cambio y las interacciones de las variables en el tiempo.

El número de periodos de medición que se utilizan en la modelación dinámica es de 6. Se sabe que el tiempo de residencia promedio del proceso es de 18 a 24 horas.

Para mantener la confidencialidad de la empresa con respecto a los datos, todos los gráficos y datos presentados a continuación están escalados.

5.2 Modelación de emisiones de N_2O por PLS dinámico

En el caso del PLS dinámico, es necesario escoger el número de variables latentes, por lo que se realiza esta selección según el criterio del punto de inflexión en la gráfica de varianza acumulada v/s el número de variables latentes.

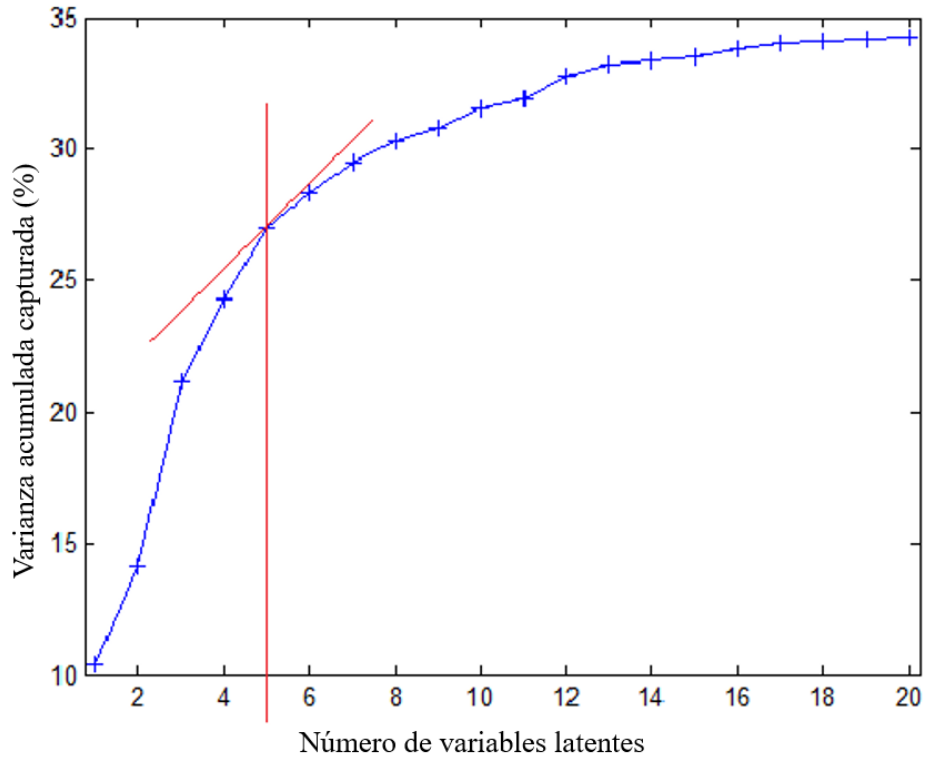


Figura 5.1: Punto de inflexión en gráfico de varianza acumulada

El gráfico en cuestión se puede ver en la Figura 5.1. Como se indica en esta figura, el punto de inflexión ocurre con 5 variables latentes, esto concuerda con el criterio de los valores propios mayores a 1, por lo que ese es el número que se utiliza en la regresión PLS dinámica.

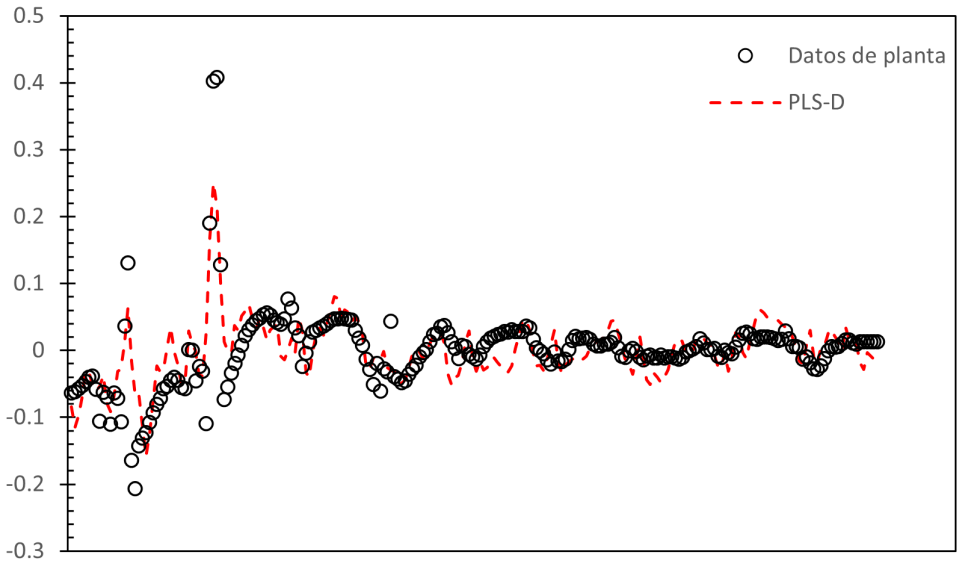


Figura 5.2: Datos de planta v/s predicción PLS-D

En la Figura 5.2 se observa el modelo obtenido en este caso, a primera vista se puede pensar que es un rendimiento aceptable, por lo que se procede a realizar un análisis de residuos para comprobar esto.

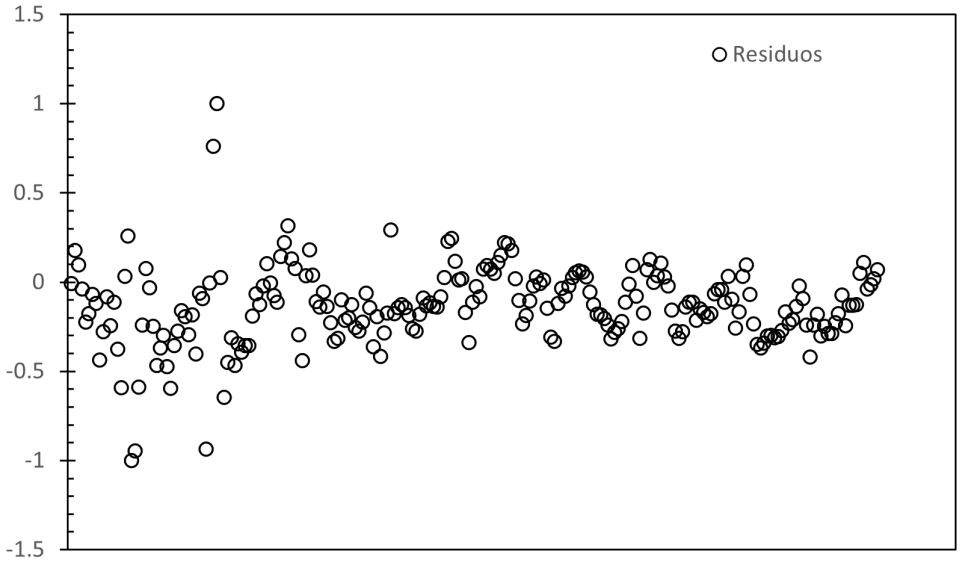


Figura 5.3: Residuos predicción PLS-D

En la Figura 5.3 se observa un comportamiento no normal en los residuos, por lo que se puede sospechar que este modelo no captura la totalidad de las relaciones existentes en la PTAR, por lo que se realiza un gráfico de la función de auto correlación, para así comprobar si en realidad queda información en estos.

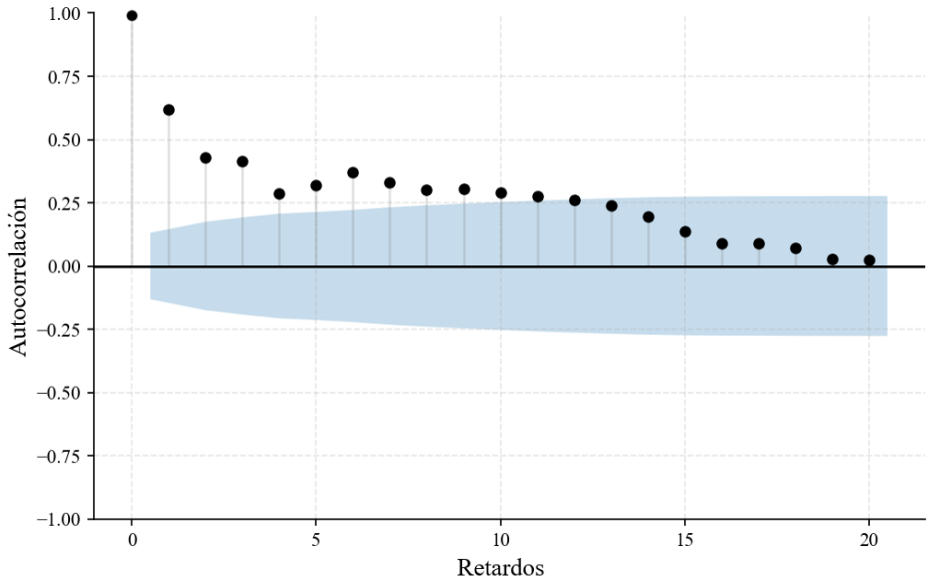


Figura 5.4: Auto correlación residuos predicción PLS-D

Como se observa en la Figura 5.4, se tienen valores mayores a 0 y mayores al umbral límite en múltiples retardos, por lo que se concluye que el modelo PLS dinámico es incapaz de abarcar la complejidad de las relaciones entre las variables.

5.3 Modelación de emisiones de N_2O por XGBoost

Para el caso de modelación por *Machine Learning* basada en XGBoost, se debe realizar el ajuste de los hiperparámetros. Esto se optimiza mediante iteración en código, lo que ocurre es una regresión preliminar variando los hiperparámetros del modelo entre los valores estándar mencionados en literatura y se quedan los hiperparámetros que producen un mejor ajuste preliminar.

Los hiperparámetros encontrados son: 'learning rate': 0.05, 'max depth': 4, 'n estimators': 100

Con estos hiperparámetros se realiza en entrenamiento del modelo, el cual se puede observar en la Figura 5.5

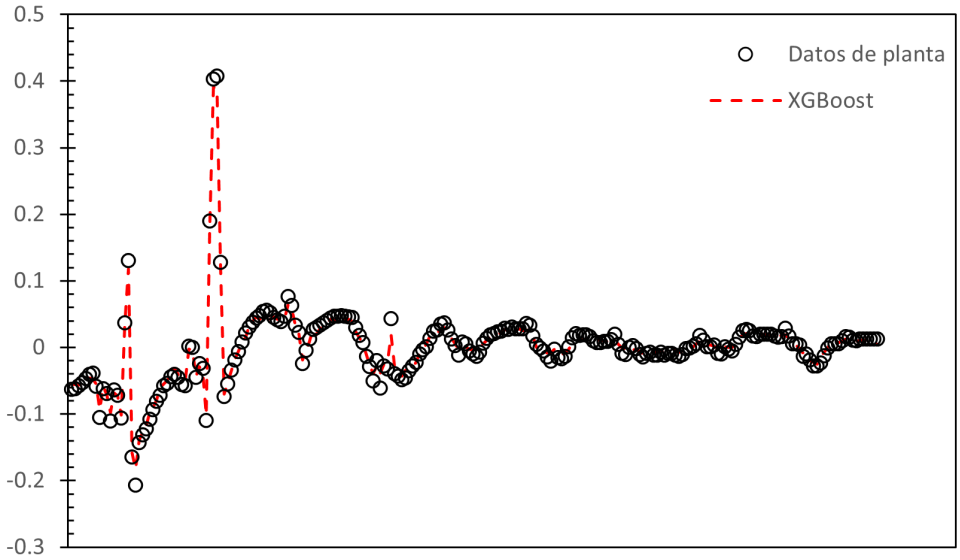


Figura 5.5: Datos de planta v/s predicción XGBoost

Se puede observar que este modelo es mucho más preciso que el estimado anteriormente, pero de igual manera se debe validar analizando los residuos.

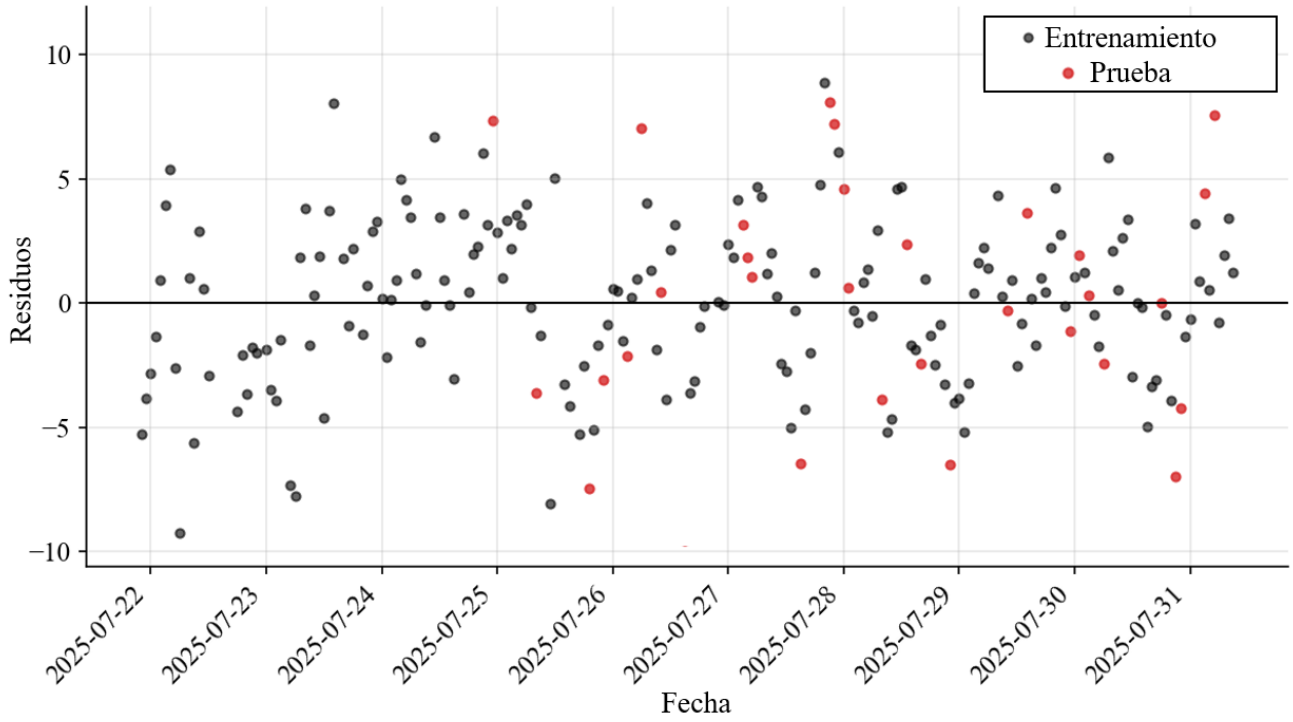


Figura 5.6: Residuos predicción XGBoost

En la Figura 5.6 se observa un gráfico de los residuos escalado, el cual de manera superficial se ve como si tuvieran una distribución normal centrada en 0. También se realiza la distinción entre los residuos del entrenamiento y la validación.

Esto se comprueba utilizando MATLAB para definir si estos residuos tienen un comportamiento normal. En la Figura 5.7 se observa que los residuos efectivamente cumplen con una distribución normal, por lo que se proponen los análisis posteriores.

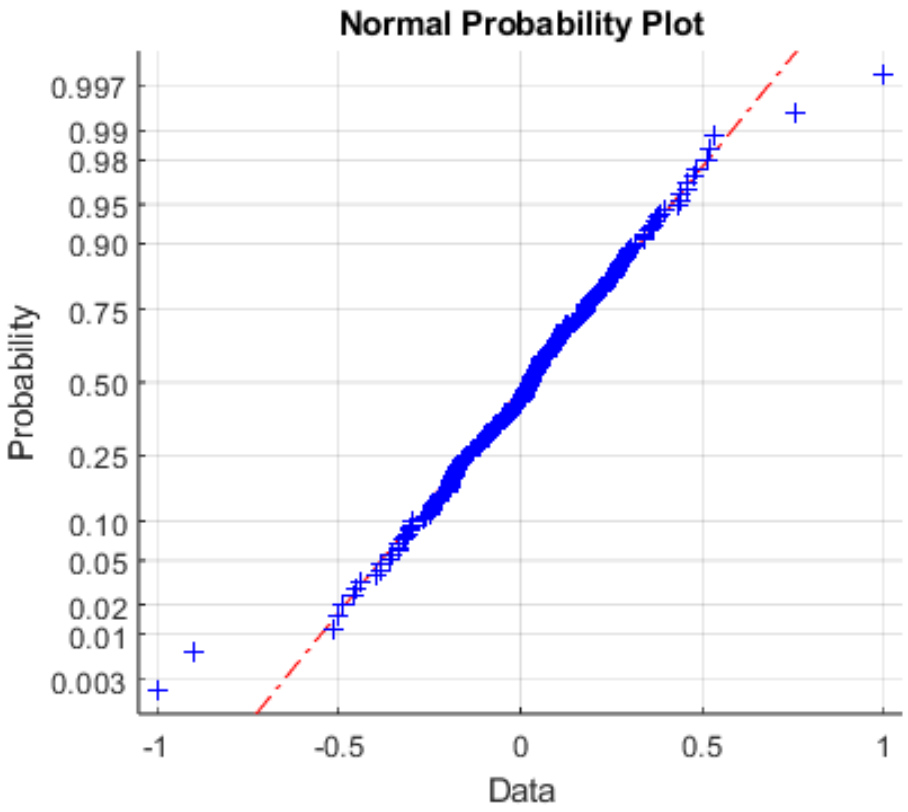


Figura 5.7: Distribución residuos predicción XGBoost

Lo último necesario es ver el error relativo que tiene el modelo con respecto a los datos reales, para eso se decide utilizar el valor del coeficiente de variación entre la desviación estándar de los residuos del modelo y la media de los datos reales. De esta manera se tiene un indicador porcentual de la desviación que puede existir en la predicción.

Al calcular este valor se obtiene que el valor del coeficiente de variación es de 0.6% lo que indica que la varianza de los residuos reales del modelo son una proporción muy baja a comparación del valor real de la concentración de N_2O en ese momento.

Finalmente con el modelo definitivo se realiza el análisis de características principales del modelo, obteniéndose el resultado de la Figura 5.8.

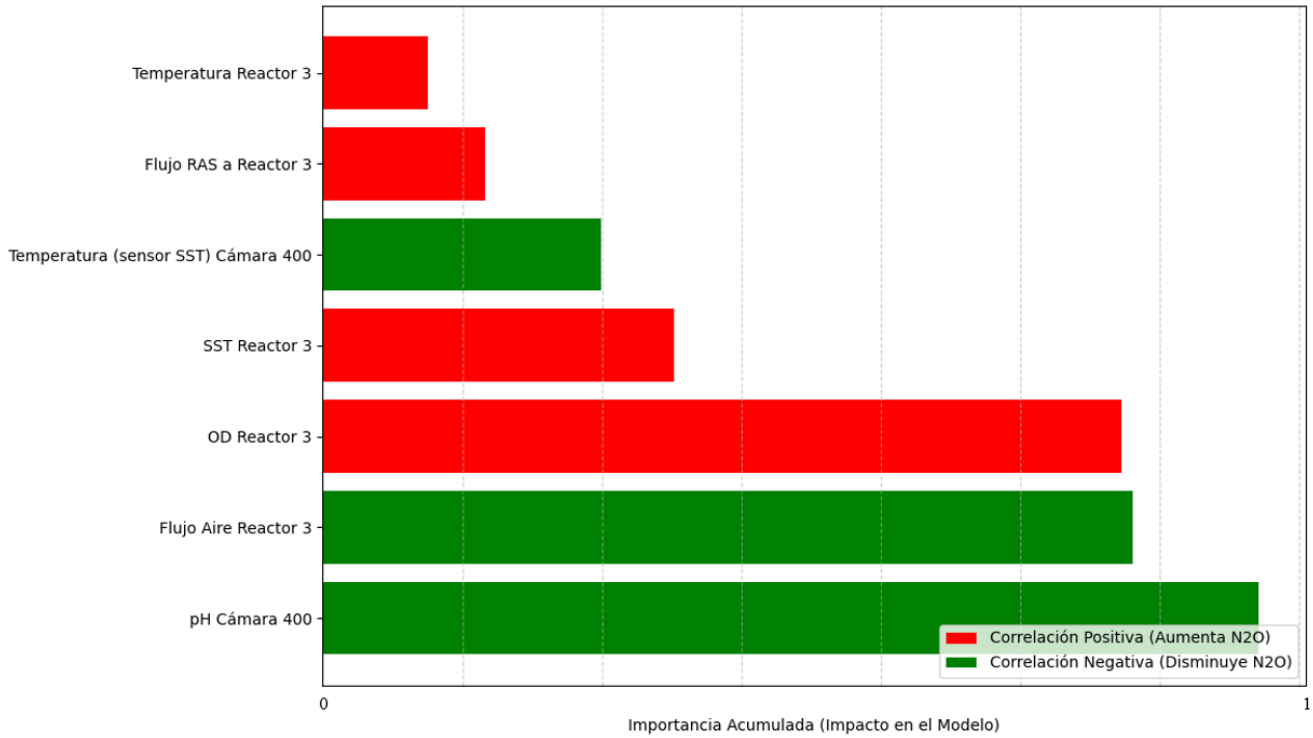


Figura 5.8: Definición de las variables más importantes en el modelo XGBoost

Las variables principales del modelo XGBoost son consistentes con la fenomenología del proceso, pero hay dos variables mencionadas en el marco teórico que no se tienen, el DQO y el TKN. Estas variables dependen de la alimentación, y esta es una variable estacional del proceso, por lo que el modelo es válido hasta que la alimentación de la planta cambie, y en ese momento habrá que entrenar un nuevo modelo y realizar una nueva validación.

Capítulo 6

Discusión

6.1 Dinámica del proceso y selección de retardos

La elección de una ventana de tiempo de 6 periodos temporales de las mediciones para la modelación dinámica, demostró ser una aproximación que funcionó de manera correcta. En reactores biológicos, la respuesta de las variables de salida (como la emisión de N_2O) ante perturbaciones en la entrada no es instantánea debido a los fenómenos de transporte y a la cinética de crecimiento bacteriano. En este caso el tiempo de residencia es cercano a las 24 horas.

El hecho de que se requieran múltiples retardos temporales para capturar la variabilidad del sistema confirma que la planta opera bajo un régimen transitorio constante. Esto valida la hipótesis inicial de que los modelos estáticos serían insuficientes para describir el comportamiento de las emisiones, ya que ignorarían la "memoria" del sistema acumulada durante los tiempos de residencia anteriores.

6.2 Limitaciones de la linealidad: Análisis del PLS

El desempeño deficiente del modelo PLS dinámico ofrece una visión importante sobre la naturaleza de los datos. Si bien el PLS es una herramienta robusta para procesos químicos convencionales, el análisis de residuos (Figura 5.3) reveló un comportamiento no normal y una autocorrelación significativa. Desde el punto de vista de la ingeniería de procesos, esto confirma que la relación global entre las variables operacionales y la generación de N_2O no pudo ser capturada por una única estructura lineal. La producción de N_2O en etapas de nitrificación y desnitrificación responde a rutas metabólicas complejas gobernadas por umbrales de activación (por ejemplo, niveles críticos de oxígeno disuelto o ratios C/N) más que por proporcionalidades directas.

Sin embargo, la insuficiencia de un modelo PLS global no descarta necesariamente la utilidad de las aproximaciones lineales si se cambia la estrategia de modelado. Dado que los residuos sugieren

que la no-linealidad es estructural, una alternativa viable sería la identificación de regímenes de operación diferenciados. En lugar de ajustar un único modelo para todo el horizonte temporal, se podrían establecer rangos de condiciones operacionales (baja carga y alta carga, o condiciones aeróbicas y anóxicas) para construir múltiples modelos PLS de validez local. Esta estrategia de modelado por tramos permitiría linealizar el comportamiento del proceso dentro de zonas de operación acotadas, mejorando la capacidad predictiva sin abandonar la interpretabilidad de las variables latentes.

6.3 Capacidad predictiva del modelo XGBoost

A diferencia del PLS, el modelo basado en XGBoost demostró una capacidad superior para manejar la complejidad del sistema. La validación de los residuos, que en este caso sí siguieron una distribución normal (Figura 5.7), implica que el modelo ha logrado extraer exitosamente la información sistemática de los datos, dejando en el error únicamente el "ruido" aleatorio inherente a la medición y al proceso.

El coeficiente de variación obtenido (0.6%) es notablemente bajo. Esto indica que el modelo es capaz de predecir las emisiones con un error relativo mínimo respecto a la media de los datos reales. Este nivel de precisión sugiere que, a pesar de la complejidad biológica, las variables monitorizadas en la planta contienen la información suficiente para describir el fenómeno, siempre y cuando se utilice un algoritmo capaz de establecer interacciones no lineales de alto orden, como es el caso de los árboles de decisión potenciados por gradiente.

Para verificar la utilidad del modelo, se distingue entre estimación y predicción. En la estimación, el modelo calcula y_t utilizando el vector de entradas reales medidas $[u_{t-1}, ..., u_{t-6}]$. En cambio, para una predicción a un paso ($k = 1$), el modelo utiliza la información disponible hasta el instante anterior, proyectando el estado y_{t+1} antes de que este ocurra. Esta distinción es crítica para evaluar la aplicabilidad del modelo en estrategias de control preventivo.

Se utiliza una proyección de 50 periodos de tiempo, ya que el tiempo de residencia del reactor biológico es de 24 horas, por lo que se quiere ver la calidad de predicción en aproximadamente 2 tiempos de residencia. El resultado obtenido se observa en la Figura 6.1.

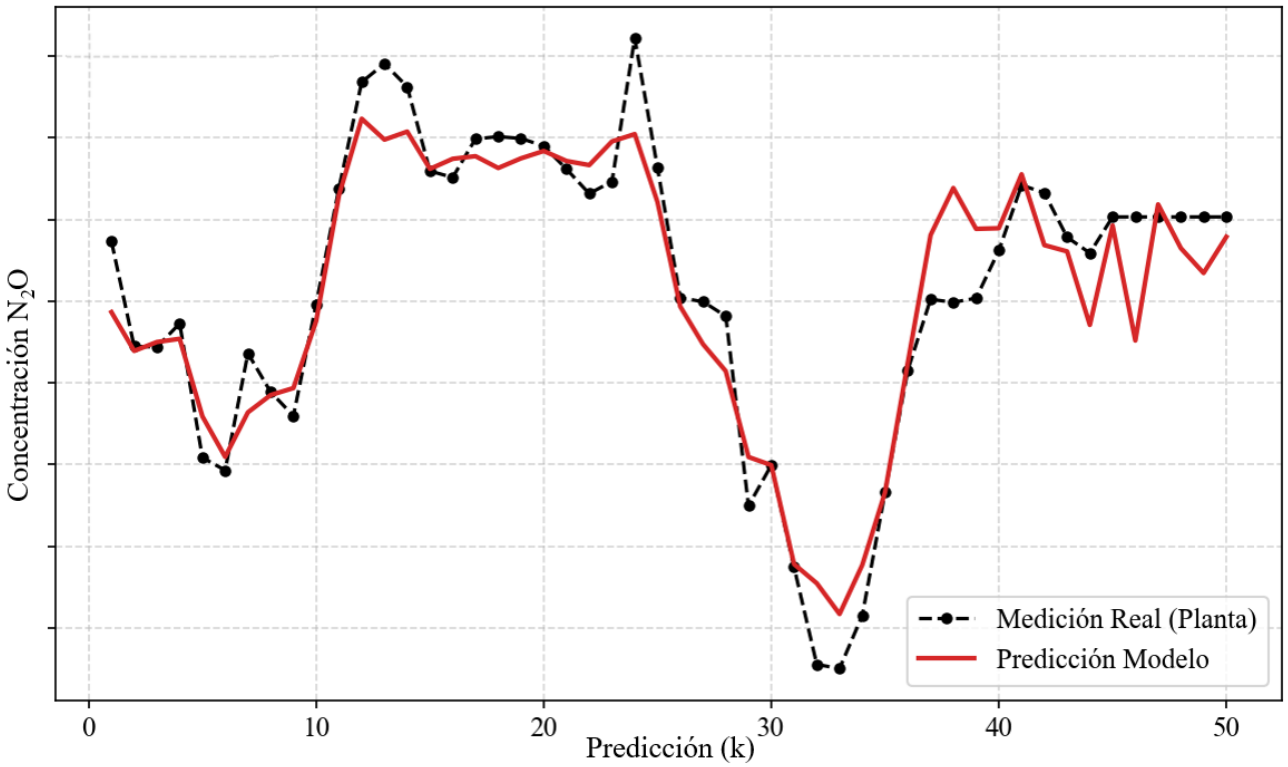


Figura 6.1: Predicción de 50 periodos con modelo XGBoost

Al analizar los residuos de las predicciones, se puede observar que hay un punto donde los residuos comienzan a estar relacionados entre sí. En la Figura 6.2 se observa cómo esto comienza a ocurrir al momento de estar prediciendo más de 20 periodos de tiempo con el modelo. Esto tiene sentido, ya que se está alcanzando el tiempo de residencia del reactor biológico, por lo que es esperable que la calidad de las predicciones disminuya de manera significativa.

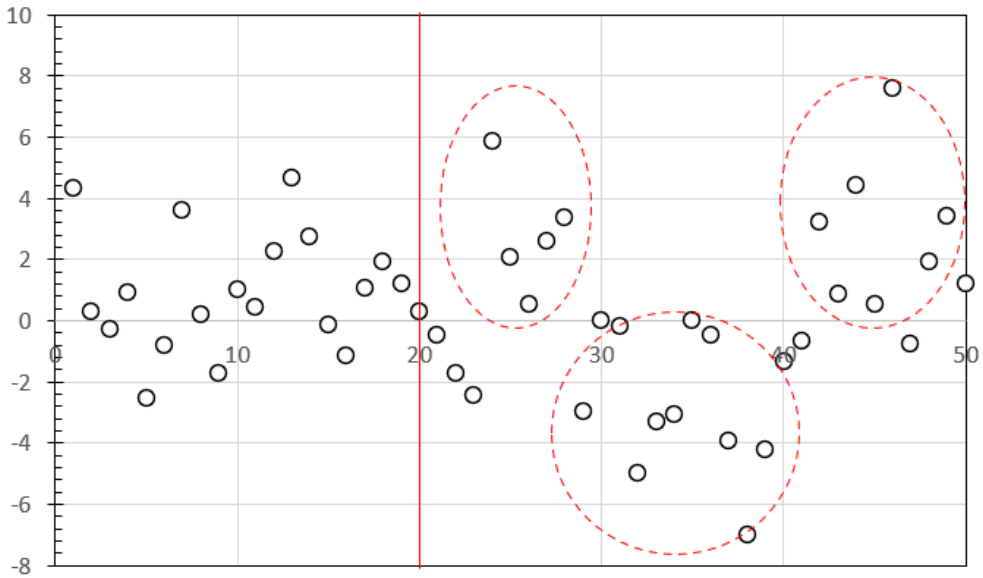
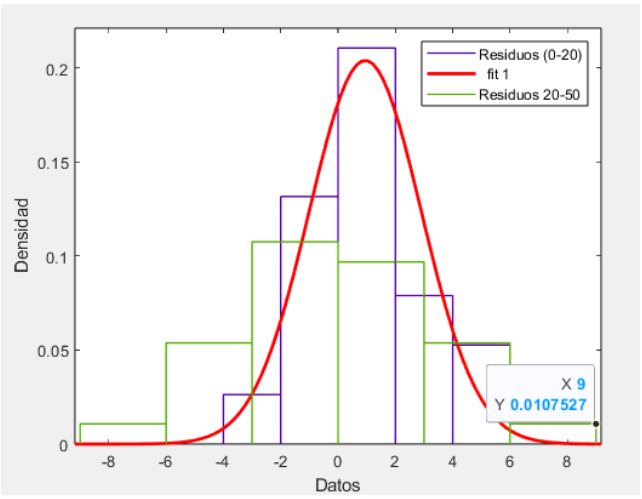


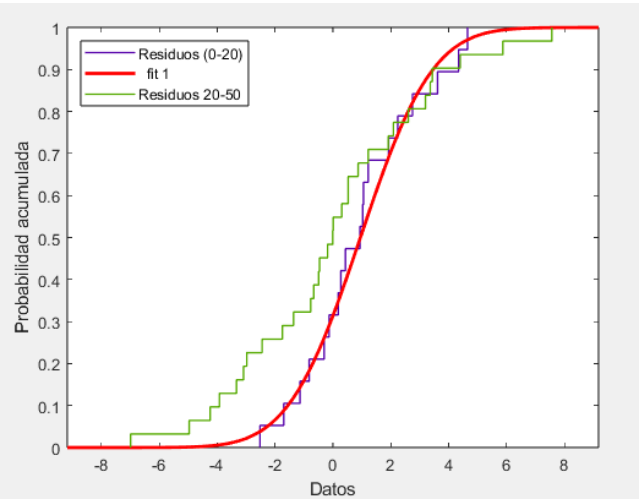
Figura 6.2: Residuos predicción de 50 periodos con modelo XGBoost

Para comprobar esto, se realiza un análisis de normalidad entre los residuos, separando los residuos en 2 grupos, hasta la predicción 20 y los 30 restantes.

En la Figura 6.3 se observa cómo los primeros 20 residuos presentan efectivamente un comportamiento normal, mientras que los demás no, demostrando así la limitación de la capacidad predictiva del modelo desarrollado.



(a) Gráfico de densidad



(b) Gráfico de probabilidad acumulada

Figura 6.3: Prueba de normalidad para ambos grupos de residuos definidos, la primera figura (a) ilustra el gráfico de densidad y la segunda (b) el gráfico de probabilidad acumulada, mostrando ambos grupos de datos en las figuras

6.3.1 Estructura matemática del modelo

Matemáticamente, el modelo predictivo final no es una única ecuación polinómica, sino un ensamble aditivo de K árboles de regresión. La predicción de la emisión de N_2O ($\hat{y}_i$) para una observación dada x_i se define como la suma de las contribuciones de cada árbol:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{6.1}$$

Donde $\mathcal{F}$ es el espacio de árboles de regresión. Cada función f_k representa un árbol independiente que divide el espacio de las variables operacionales en regiones distintas, asignando un peso constante w a cada hoja. A diferencia del PLS que busca hiperplanos, cada $f_k(x_i)$ captura interacciones locales y umbrales biológicos no lineales.

El modelo se entrenó minimizando una función objetivo regularizada $\mathcal{L}(\phi)$ que garantiza el equilibrio entre precisión y simplicidad (para evitar sobreajuste):

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{6.2}$$

Donde l es la función de pérdida (en este caso, error cuadrático medio) y Ω penaliza la complejidad de los árboles (número de hojas y magnitud de los pesos).

6.3.2 Explicitación matemática del primer estimador

Dado que el modelo XGBoost completo consta de una suma de $K = 100$ funciones, su representación explícita total es inabarcable. Sin embargo, es posible formular matemáticamente el primer estimador ($f_1(\mathbf{x})$) para demostrar cómo el algoritmo segmenta el espacio operativo basándose en condiciones físicas críticas.

A partir del análisis estructural del primer árbol de decisión, se identificó que la variable dominante para la segregación inicial de los datos es la concentración de Sólidos Suspendidos Totales (SST). La función de corrección $f_1(\mathbf{x})$ adopta la siguiente estructura lógica simplificada para sus ramas principales:

$$f_1(\mathbf{x}) \approx \begin{cases} +8.39 & \text{si } x_{SST} \geq 1958.2 \text{ y } x_{salida} < 112.3 \\ +0.82 & \text{si } x_{SST} \geq 1958.2 \text{ y } x_{salida} \geq 112.3 \\ \approx 0.005 & \text{si } x_{SST} < 1958.2 \text{ y } x_{entrada} \geq 39.8 \\ -1.68 & \text{si } x_{SST} < 1958.2 \text{ y } x_{entrada} < 39.8 \text{ y } x_{OD} < 17.8 \end{cases} \tag{6.3}$$

Donde las variables de estado y control (con sus respectivos retardos temporales) se definen

como:

- x_{SST}: Sólidos Suspendidos Totales $[t-6]$.
- x_{salida}: Flujo Salida Reactor 3 $[t-2]$.
- $x_{entrada}$: Flujo Entrada Reactor 3 $[t-1]$.
- x_{OD}: Oxígeno Disuelto Reactor 3 $[t-1]$.

Interpretación del mecanismo de control: la primera rama de la ecuación revela un hallazgo fenomenológico coherente: el modelo asigna el mayor valor positivo de emisión (+8.39) cuando coexisten una alta concentración de sólidos ($x_{SST} \geq 1958$) y una baja tasa de flujo de salida ($x_{salida} < 112$). Esta combinación sugiere operativamente una acumulación de biomasa o retención excesiva de sólidos, condiciones que favorecen la generación de zonas anóxicas no controladas y, consecuentemente, picos en la producción de N_2O. Por el contrario, cuando los niveles de SST están controlados (tercera rama), la corrección del modelo es cercana a cero (+0.005), indicando una operación estable.

6.4 Interpretación de las variables influyentes

El análisis de importancia de variables presentado en la Figura 5.8 permite abrir la "caja negra" del modelo de *Machine Learning* y contrastarlo con la teoría de bioprocesos.

Se observa que entre las variables con mayor peso en la predicción se encuentra el oxígeno disuelto, el cual se controla con el flujo de aire, que también fue reconocido dentro de las principales variables. Esto es consistente con la literatura, ya que el oxígeno disuelto es el factor principal que controla si la ruta metabólica favorece la producción de N_2O o N_2, y además se puede observar que la relación es inversa, ya que, según fenomenología, si hay oxígeno disuelto insuficiente, entonces ocurren los procesos incompletos que generan N_2O como emisión.. El hecho de que el modelo XGBoost haya identificado estas variables como las más relevantes sin intervención externa valida aún más su coherencia fenomenológica, demostrando que no solo ajusta datos matemáticamente, sino que captura las relaciones causales subyacentes del proceso de tratamiento de aguas.

Capítulo 7

Conclusiones y Recomendaciones

Una de las conclusiones fundamentales de esta investigación es que las plantas de tratamiento de aguas residuales estudiadas operan consistentemente en un estado no estacionario. Este comportamiento es inherente a la naturaleza del proceso, el cual, al estar gobernado por la dinámica de poblaciones bacterianas responsables de la nitrificación y desnitrificación, presenta una inercia temporal significativa. Se determinó que considerar el tiempo de residencia (y sus múltiplos) es esencial para una modelación correcta, validando el uso de enfoques dinámicos sobre los estáticos.

Respecto a las técnicas de modelado, se desestimó la efectividad de los modelos lineales para este propósito. Aunque se aplicó una regresión mediante Mínimos Cuadrados Parciales (PLS) dinámico, el análisis de residuos demostró una distribución no normal, evidenciando que las relaciones lineales son insuficientes para describir la complejidad de la generación de N_2O. Esto confirma lo descrito en la literatura sobre la alta no linealidad en las interacciones de las variables de una PTAR.

Por el contrario, la implementación del modelo de *Machine Learning* basado en el algoritmo XGBoost arrojó resultados altamente satisfactorios. El modelo no solo superó las pruebas de normalidad en los residuos, sino que alcanzó un coeficiente de variación de apenas 0.6% entre la desviación estándar del error y la media de los datos. Esto permite concluir que el algoritmo es capaz de capturar con alta precisión la variabilidad del proceso.

En cuanto a la capacidad predictiva y su validez temporal, las pruebas de predicción revelaron un hallazgo crítico. El modelo demostró sostener predicciones estadísticamente robustas durante un horizonte aproximado de 20 horas, punto a partir del cual la calidad de la estimación se degrada debido a la propagación del error. Este umbral coincide físicamente con el tiempo de residencia de un reactor biológico, lo que permite concluir que el modelo puede predecir las emisiones de manera válida hasta que ocurre un ciclo de renovación de masa. Por tanto, su utilidad para estrategias de control predictivo es confiable siempre que el horizonte de acción se mantenga dentro de esta ventana temporal.

Adicionalmente, el análisis de importancia de variables derivado del modelo XGBoost permitió

identificar los parámetros críticos del proceso. Se concluye que estas variables tienen la mayor influencia estadística sobre las emisiones de N_2O, lo cual es coherente con los principios teóricos del metabolismo biológico en el tratamiento de aguas. Esto dota al modelo de interpretabilidad física, más allá de su precisión matemática.

Como recomendación, se sugiere evaluar la robustez del modelo XGBoost desarrollado mediante la incorporación de nuevos conjuntos de datos temporales (validación cruzada en el tiempo) a medida que estén disponibles. Asimismo, se recomienda estudiar la posibilidad de integrar este modelo predictivo en un lazo de control, utilizando las variables identificadas como más influyentes, como el flujo de aire, para minimizar activamente las emisiones de gases de efecto invernadero en tiempo real, para así poder concluir respecto a la capacidad predictiva del modelo. Asimismo, se sugiere incorporar la cuantificación de otros gases de efecto invernadero generados en la planta, lo que permitiría una estimación más integral de la huella de carbono.

Finalmente, se debe tener en cuenta las demás posibilidades existentes al utilizar *Machine Learning*, ya que en este trabajo se utilizaron modelos de aprendizaje automático, pero también se podrían haber utilizado redes neuronales, como LSTM por ejemplo, por lo que se recomienda explorar estas opciones en trabajos futuros.

Referencias

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *WIREs Computational Statistics*, 2(1):97–106.

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., and Mersmann, O. (2023). *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*. Springer Singapore.

Bellandi, G., Weijers, S., Gori, R., and Nopens, I. (2020). Towards an online mitigation strategy for n2o emissions through principal components analysis and clustering techniques. *Journal of Environmental Management*, 261.

Campins-Falco, P., Meseguer-Lloret, S., Climent-Santamaria, T., and Molins-Legua, C. (2008). A microscale kjeldahl nitrogen determination for environmental waters. *Talanta*, 75(4):1123–1126.

Crutzen, P. J. (1978). The role of NO and NO_2 in the chemistry of the troposphere and stratosphere. *Annual Review of Earth and Planetary Sciences*, 7.

Eigenvector Research Inc. (2025). Eigenvector research, inc. — chemometrics and machine learning software. Accedido: 15-12-2025.

Gruber, W., Villez, K., Kipf, M., Wunderlin, P., Siegrist, H., Vogt, L., and Joss, A. (2020). N_2O emission in full-scale wastewater treatment: Proposing a refined monitoring strategy. *Science of the Total Environment*, 699.

Huang, L., Li, H., and Li, Y. (2024). Greenhouse gas accounting methodologies for wastewater treatment plants: A review. *Journal of Cleaner Production*, 448.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

Massara, T. M., Solís, B., Guisasola, A., Katsou, E., and Baeza, J. A. (2018a). Development of an asm2d-n2o model to describe nitrous oxide emissions in municipal wwtps under dynamic conditions. *Chemical Engineering Journal*, 335:185–196.

Massara, T. M., Solís, B., Guisasola, A., Katsou, E., and Baeza, J. A. (2018b). Development of an asm2d-n2o model to describe nitrous oxide emissions in municipal wwtps under dynamic conditions. *Chemical Engineering Journal*, 335:185–196.

Metcalf & Eddy Inc., Tchobanoglous, G., Stensel, H. D., Tsuchihashi, R., and Burton, F. (2014). *Wastewater Engineering: Treatment and Resource Recovery*. McGraw-Hill Education, New York, NY, fifth edition.

Nguyen, T. K., Ngo, H. H., Guo, W. S., Chang, S. W., Nguyen, D. D., Nghiem, L. D., and Nguyen, T. V. (2020). A critical review on life cycle assessment and plant-wide models towards emission control strategies for greenhouse gas from wastewater treatment plants. *Journal of Environmental Management*, 264.

Rhinehart, R. (1995). A novel method for automated identification of steady-state. In *Proceedings of 1995 American Control Conference - ACC'95*, volume 6, pages 4065–4066 vol.6.

Song, M. J., Choi, S., Bae, W. B., Lee, J., Han, H., Kim, D. D., Kwon, M., Myung, J., Kim, Y. M., and Yoon, S. (2020). Identification of primary effecters of n2o emissions from full-scale biological nitrogen removal systems using random forest approach. *Water Research*, 184.

Szelag, B., Zaborowska, E., and Makinia, J. (2023). An algorithm for selecting a machine learning method for predicting nitrous oxide emissions in municipal wastewater treatment plants. *Journal of Water Process Engineering*, 54.

Vasilaki, V., Massara, T. M., Stanchev, P., Fatone, F., and Katsou, E. (2019). A decade of nitrous oxide monitoring in full-scale wastewater treatment processes: A critical review. *Water Research*, 161:392–412.

Vasilaki, V., Volcke, E. I., Nandi, A. K., van Loosdrecht, M. C., and Katsou, E. (2018). Relating n2o emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. *Water Research*, 140:387–402.

Apéndices

Apéndice A

Material Suplementario: Conjunto de Datos de Operación de la PTAR

El conjunto de datos completo utilizado para el entrenamiento y validación de los modelos presentados en esta tesis se encuentra disponible como material suplementario en un archivo externo.

Nombre del archivo: `Base_de_datos.xlsx`

Descripción: El archivo contiene los registros variables operacionales de la planta de tratamiento de Aguas Araucanía y la concentración de N_2O de las emisiones monitoreadas en el Reactor 3 de esta planta, abarcando el periodo del 21 de julio al 31 de julio de 2025.

Disponibilidad: El archivo es confidencial, para acceso a este favor comunicarse con luis.bergh@usm.cl