Universidad de Oviedo

Centro Internacional de Postgrado

Programa de Doctorado en Ingeniería de Producción, Minero-Ambiental y de Proyectos

Tesis doctoral

Optimización de la operación y tratamiento de sólidos en depuradoras de aguas residuales

Vanesa Mateo Pérez

Octubre 2021

AGRADECIMIENTOS

Después de esta dura travesía por fin llegó el momento de los agradecimientos. Pensé en hacer unos agradecimientos más convencionales dándole las gracias a mi Director de Tesis (gracias Fran por animarme a vivir esta aventura y por guiarme y apoyarme en ella y por aspirar siempre, no a lo bueno, sino a lo excelente), a API, mi área (sobre todo gracias Joaquín por enseñarme tanto de esos locos algoritmos, gracias Mesa por ayudarme en el arduo camino de los artículos y mil gracias Cris por sacarme siempre una sonrisa), al personal de la EDAR de Villapérez (Iván santa paciencia que has tenido conmigo), a mis compañeros del Principado (especialmente gracias a Xavi y a Bárbara por aguantar mis monólogos infinitos), a mis amigos (menos mal que esto ya acaba porque podría quedarme sin ninguno), a mis padres (rendirse nunca es una opción), a mis pequeñas (por todo ese tiempo robado) y a mi costilla (que bonito es ver que alguien cree más en ti que tú mismo). Pero luego pensé que unos agradecimientos convencionales eran más aburridos. Hablando con mi querida Bárbara (en uno de mis muchos monólogos) pensamos en buscar una frase chula, de alguien poco conocido, pero llegamos a la conclusión de que para citar a alguien tendría que leerme como mínimo su bibliografía completa, para citar con rigor, y creo que el esfuerzo no vale la pena. Así que he decido citar a mi abuela y a uno de sus refranes que forman parte de la sabiduría popular, como ella decía "es de bien nacidos ser agradecidos" así que muchas gracias a todos, a los que están, a los que estuvieron, a los que me ayudaron con la tesis, a los que me aguantaron contando mis avances y a los que simplemente creyeron en mí, gracias a todos.

Índice

INDICE DE FIGURAS

INDICE DE TABLAS

LISTA DE ABREVIATURAS Y ACRÓNIMOS

AEMET	Agencia Estatal de Meteorología
ANFIS	*Adaptive Neuro-Fuzzy Inference System* Sistemas Neuronales de Inferencia Difusa Adaptativos
ANN	*Artificial Neural Networks* Redes Neuronales Artificiales
COD	Carbono Orgánico Disuelto
DBO5	Demanda Biológica de Oxígeno a los 5 días
DQO	Demanda Química de Oxígeno
EDAR	Estación Depuradora de Aguas Residuales
FFA	*Free Fatty Acids*. Ácidos grasos Libres
FIS	Sistema de Inferencia Difusa
FOG	*Fat, Oil and Grease*, variable grasa.
GTS	*Grease Tap Sludge* (sistema de captura de grasas)
GVC	*Generalized Cross Validation,(* Validación Cruzada Generalizada*)*
MARS	*Multi-Adaptive Regressiones Splines*
MAE	*Mean Absolute Error,* Error medio absoluto
MSE	*Mean Square Error,* Error Cuadrático Medio
PCA	*Principal Component Analysis,* Análisis de componentes principales
RBF	*Radial Basis Function* Función de Base Radial Gaussiana
RF	*Random Forest*, Árbol de decisión
RSS	*Residual Sum of Squares,* sumas de cuadrados residuales
SCADA	*Supervisory Control And Data Acquisition*, Supervisión, Control y Adquisición de Datos
SS	Sólidos en Suspensión
SVM	*Support Vector Machine*, Máquinas de Vector Soporte
VIF	Factor de inflación de la varianza
WCSS	*Within Clusters Summed Squares* Suma de los cuadrados de las distancias
WWTP	*Waste Water Treatment Plants* (Plantas depuradoras)

CRISP-DM *Cross-Industry Standard Process for Data Mining*

CART *Classification and Regression Trees*

Resumen

Los sistemas de tratamiento de aguas residuales incluyen procesos físicos, químicos y biológicos. Los procesos físicos separan los sólidos de mayor tamaño y son característicos de la primera etapa de tratamiento denominada pretratamiento. En ella se eliminan las grasas, las arenas y los sólidos de desbaste (toallitas, plásticos, productos de higiene íntima y materia orgánica no disuelta).

El pretratamiento es una de las etapas menos estudiadas de una estación depuradora y, sin embargo, si algo no funciona correctamente, ocasiona vertidos directos de agua sin tratar al medio receptor, con el impacto ambiental que ello conlleva. Internamente si no se eliminan esos residuos en la primera etapa genera problemas de operación en el resto de la estación de tratamiento.

El objeto del presente trabajo es predecir las cantidades de grasas, arenas y sólidos de desbaste recibidas a la entrada de las depuradoras que deberán ser eliminadas en el pretratamiento, con el objetivo de garantizar el funcionamiento óptimo de esta etapa y evitar vertidos al cauce de agua sin tratar.

Para poder llevar a cabo esta predicción de parámetros se utilizan distintos algoritmos. En el caso de los sólidos de desbaste se ha desarrollado un modelo *SVM* para predecir su llegada. La precisión alcanzada en la fase de validación es de R^2= 0,6869, ligeramente inferior a la alcanzada en el entrenamiento (R^2= 0,7093), se considera suficiente para detectar cambios de tendencia en la llegada de sólidos de desbaste a la depuradora. El modelo final presenta un *MSE* de 0,435 en la prueba de validación. Los mayores errores del modelo se producen en los extremos, es decir, por debajo de 2 toneladas y por encima de 4 toneladas de sólidos brutos, que son valores poco habituales, por lo que no suponen un gran inconveniente y se deben a la escasa presencia de este tipo de patrones en el conjunto de datos de entrenamiento.

En el caso de las arenas, para su predicción se utiliza un algoritmo tipo *MARS*. En este caso, el modelo alcanza un R^2=0,70 en los datos de validación, suficiente para predecir los cambios de tendencia en la recuperación de la arena durante las fases de pretratamiento. El modelo final presenta un *MSE* de 0,46 en la prueba de validación

Por último, en el caso de las grasas, la evaluación de la idoneidad del modelo de predicción se realizó mediante el coeficiente de determinación (R^2 ajustado) entre los valores predichos frente al conjunto de datos reales. En este caso la precisión del modelo *Random Forest* es muy alta, R^2=0,98 para los datos de entrenamiento y R^2=0,93 en los datos de validación, óptima para predecir los cambios de tendencia en la separación de grasas durante las fases de pretratamiento. El modelo final presenta un *MSE* de 0,037 en el entrenamiento mientras que alcanza 0,089 en la prueba de validación.

Globalmente, los modelos desarrollados demuestran que es posible determinar con antelación los contenidos sólidos que se recibirán en la depuradora, en función de la cuenca y de las condiciones meteorológicas, permitiendo así optimizar su funcionamiento, con los beneficios técnicos, económicos y operativos que de ellos se deriva.

Abstract

Wastewater treatment systems include physical, chemical and biological processes. Physical processes separate larger solids and are characteristic of the first stage of treatment called pretreatment. In this stage, grease, sand and solids (wipes, plastics, intimate hygiene products and undissolved organic matter) are removed.

Pretreatment is one of the least studied stages of a wastewater treatment plant and yet, if something goes wrong, it causes direct discharges of untreated water into the receiving environment, with the environmental impact that this entails. Internally, if these wastes are not eliminated in the first stage, it generates operating problems in the rest of the treatment plant.

The purpose of the present work is to predict the quantities of grease, sand and solids from roughing received at the inlet of the treatment plants that should be eliminated in the pretreatment, with the objective of guaranteeing the optimal operation of this stage and avoiding discharges of untreated water into the watercourse.

In order to carry out this parameter prediction, different algorithms are used, and in the case of the roughing solids, an SVM model has been developed to predict their arrival. The accuracy achieved in the validation phase is $R^2= 0.6869$, slightly lower than that achieved in the training ($R^2= 0.7093$), which is considered sufficient to detect trend changes in the arrival of solids in the treatment plant. The final model presents an *MSE* of *0.435* in the validation test. The largest model errors occur at the extremes, i.e., below 2 tons and above 4 tons of gross solids, which are unusual values, so they are not a major drawback and are due to the scarce presence of this type of patterns in the training data set.

In the case of sands, a *MARS*-type algorithm is used for their prediction. In this case, the model achieves an $R^2=0.70$ in the validation data, sufficient to predict the trend changes in sand recovery during the pretreatment phases. The final model presents an *MSE* of *0.46* in the validation test.

Finally, in the case of fats, the evaluation of the adequacy of the prediction model was performed by the coefficient of determination (adjusted R^2) between the predicted values versus the actual data set. In this case the accuracy of the Random Forest model is very high, $R^2=0.98$ for the training data and $R^2=0.93$ in the validation data, optimal for predicting trend changes in fat separation during the pretreatment phases. The final model presents an *MSE* of *0.037* in the training while it reaches *0.089* in the validation test.

Overall, the models developed demonstrate that it is possible to determine in advance the solid contents to be received at the WWTP, depending on the basin

and meteorological conditions, thus allowing to optimize its operation, with the resulting technical, economic and operational benefits.

CAPITULO 1. Introducción

1.1 Conceptos básicos del tratamiento de aguas residuales

La depuración de aguas tiene como objetivo tratar el agua residual para hacerla adecuada para su vertido a los cauces receptores.

Los procesos de depuración se dividen en varias fases, cada una de las cuales tiene como objetivo la eliminación de un tipo de residuo diferente. Además, las fases son secuenciales y el funcionamiento incorrecto de una de ellas condiciona el resultado de las fases sucesivas.

En una primera fase de tratamiento se elimina la contaminación no disuelta mediante medios mecánicos. En la segunda fase se elimina la contaminación disuelta mediante tratamientos biológicos que implican el uso de bacterias. Por último, los tratamientos de afino eliminan todo aquello que no ha podido ser eliminado en el tratamiento biológico, llevando a cabo posteriormente la desinfección del agua en aquellos casos en los que es necesario.

El agua residual, tras recibir todos estos tratamientos, pasa a ser apta para ser vertida en el medio receptor sin generar problemas ambientales y de salubridad.

1.2 Motivación de la tesis y objetivos

El objetivo final de este trabajo es optimizar el funcionamiento del pretratamiento en estaciones depuradoras considerando los sólidos recibidos. El pretratamiento se corresponde con la primera fase de depuración, en la que se eliminan los contaminantes no disueltos mediante procedimientos físicos.

El pretratamiento en estaciones depuradoras de un cierto tamaño está formado por varias líneas de tratamiento. El número de líneas que funcionan en cada momento depende de la experiencia del jefe de planta que decide en función de su experiencia si se debe incrementar el número de líneas activadas, adelantándose a la llegada de los residuos de mayor tamaño en el agua residual. Es, por tanto, un sistema de toma de decisión muy manual y dependiente del conocimiento, o simplemente la presencia, del personal adecuado en cada momento. Si el jefe de planta no tiene ese conocimiento o simplemente no se encuentra trabajando esos días, no se implementa el número de líneas hasta que ya se produce el problema, porque el incremento de residuos ya se ha producido.

Este trabajo de investigación tiene como objetivo disminuir la incertidumbre asociada a la intervención humana detectando los cambios de tendencia en la producción de residuos. Es decir, se trata de que el conocimiento deje de residir en

las personas para pasar a ser de la organización, homogeneizando la toma de decisiones, adelantándose a las posibles situaciones de malfuncionamiento y optimizando las actuaciones.

Para poder optimizar realmente el funcionamiento del pretratamiento es necesario predecir todos los residuos que van a tener que ser tratados en este, ya que su comportamiento y origen es diferente como se ha visto en apartados anteriores.

Los residuos que se generan en el pretratamiento, tras separarlos del agua residual son los sólidos de desbaste, las grasas y las arenas. En este trabajo se establece un objetivo por cada uno de dichos residuos, elaborando un algoritmo predictivo para cada uno de ellos con el objetivo último de optimizar cada una de las fases del pretratamiento.

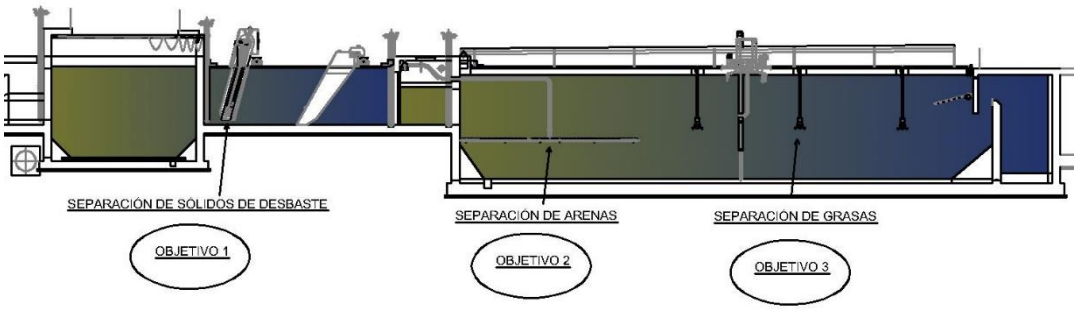


Figura 1: Relación entre las etapas de pretratamiento y los objetivos.

Cada una de esas aproximaciones se presenta en uno de los artículos que se incluyen en esta tesis por compendio de publicaciones. Los objetivos son, por tanto, los siguientes:

Objetivo 1: Predicción de la llegada de sólidos de desbaste a la depuradora. Se trata de los elementos más variables, generalmente procedentes de la falta de descomposición de productos de higiene. Suelen aparecer en la zona superior puesto que se transportan por flotación. Este objetivo se desarrolla en el artículo titulado "*Gross Solids Content Prediction in Urban WWTPs Using SVM*"[1].

Objetivo 2: Predicción de la llegada de arenas a la depuradora. Son elementos relacionados con las aguas de origen no doméstico. Su transporte se produce mediante arrastre puesto que su densidad es muy superior a la del agua. Se detalla en "*Sand Content Prediction in Urban WWTPs Using MARS*"[2]. Otros detalles se pueden encontrar en la comunicación "*Estimación y caracterización de sólidos en depuradoras de aguas residuales*", presentada en el 22nd International Congress on Project Management and Engineering [3].

Objetivo 3: Predicción de la llegada de grasas a la depuradora. Son materiales complejos procedentes tanto de las aguas residuales domésticas como industriales. Por su baja densidad se transportan por flotación si bien en muchas ocasiones se

adhieren a otros elementos (sólidos de desbaste, arenas), viajando con ellos. Su desarrollo se detalla en el artículo: *"A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs"*[4].

Es decir, de forma secuencial se van obteniendo los distintos modelos, cada uno de ellos con sus variables y características específicas. La obtención de los tres modelos predictivos permitirá optimizar el funcionamiento de cada uno de los elementos del pretratamiento dando lugar a un sistema global que puede ser utilizado para la correcta gestión de la plata depuradora. La visión global, aún no publicada en revista, se presenta de forma integrada en la comunicación a congreso titulada *"Mejora del pretratamiento de una EDAR mediante la predicción de parámetros del agua de entrada"* presentada en el 25nd International Congress on Project Management and Engineering.

Con la predicción de llegada de sólidos de desbaste se podrá determinar el número de líneas de rejas y tamices a poner en servicio, con la predicción de arenas se determinará el número de desarenadores en servicio y con la predicción de grasas se podrá ajustar la cantidad de aireación de los desarenadores y la velocidad necesaria de las rasquetas de recogida de grasas. Conocer el cambio de tendencia en la producción de residuos también permitirá preparar los sistemas de recogida de residuos para situaciones de llegadas masivas de dichos residuos.

Los objetivos no tienen por qué ser secuenciales ya que unos no condicionan a los otros, el esquema de trabajo de la investigación se puede representar en la siguiente figura:

Figura 2: Elementos de la investigación

1.3 Ámbito de aplicación de la tesis

La eliminación previa de los contaminantes no disueltos es una etapa común a todas las estaciones depuradoras de aguas residuales, independientemente de su tamaño o del tipo de medio receptor. Puede darse que no existan las etapas posteriores (cada vez menos) pero el pretratamiento existe siempre.

El resultado de la tesis genera unos modelos de predicción aplicables a cualquier depuradora de aguas residuales urbanas. Además, se han elegido para los casos de estudio depuradoras que incluyen tanto vertidos de agua residual urbana como industrial, obteniendo unos resultados excelentes. Por ello, se puede afirmar que los modelos aquí desarrollados son generalizables y pueden ser aplicados a prácticamente cualquier depuradora de uso urbano o mixto urbano-industrial, independientemente de sus características.

1.4 Estructura de la memoria

En este primer capítulo se ha realizado una breve introducción al trabajo realizado, incluyendo la motivación de la tesis y los principales objetivos a conseguir con ella, además de definir su ámbito de aplicación.

En el Capítulo 2 se analizan los elementos clave del estado del arte y se explica el caso de estudio y las variables analizadas. En el Capítulo 3 se describe la metodología utilizada. El Capítulo 4 incluye la parte central de la investigación mediante la incorporación de los 3 artículos que forman el núcleo de este trabajo. El Capítulo 5 resume los resultados más relevantes de la investigación, así como su interpretación. El Capítulo 6 muestra las conclusiones y líneas de desarrollo futuras. Los capítulos 7 y 8 contienen un informe sobre el factor de impacto y las referencias bibliográficas respectivamente. Por último, en el anexo I se incluyen diversas comunicaciones adicionales sobre el mismo tema y las publicaciones que se han llevado a cabo con las líneas futuras de investigación pero que excedían el alcance de la presente tesis.

CAPITULO 2. Estado del Arte y caso de estudio

2.1 Estado del arte

Aunque la captación y drenaje de agua pluviales data de tiempos antiguos, la recogida de aguas residuales no aparece como tal hasta principios del siglo XIX y su tratamiento hasta finales del siglo XIX y principios del XX [5].

Los primeros tratamientos biológicos del agua residual se llevaron a cabo mediante filtros percoladores [6], aunque el impulso definitivo del tratamiento de las aguas residuales se produce gracias al desarrollo de la teoría del germen a cargo de Koch y Pasteur [7] en la segunda mitad del siglo XIX. Este desarrollo marcó el inicio de una nueva era en el campo del saneamiento. Hasta ese momento se había profundizado poco en la relación entre contaminación y enfermedades y no se había aplicado la bacteriología al tratamiento de aguas, disciplina entonces en sus inicios [8]. Las aguas se vertían directamente a los medios receptores sin tratar previamente o se aplicaban directamente al terreno. La aparición de enfermedades, como el brote de cólera producido en Inglaterra en la segunda mitad del SXIX, hace que se empiece a avanzar en el tratamiento de las aguas residuales y que se analice la relación entre el contacto con aguas residuales y la transmisión de enfermedades [9].

En los inicios se acudía a tratamientos locales debido a la dispersión de la población, como el uso de los tanques Imhoff [10] pero el aumento de la concentración de la población en las ciudades hizo necesario buscar avances hacia tratamientos más completos [11].

Las aguas residuales están compuestas por una fracción orgánica y una fracción inorgánica. Los compuestos orgánicos están formados principalmente por combinaciones de carbono, hidrógeno y oxígeno, destacando como principales grupos de sustancias orgánicas presentes en el agua residual las proteínas (40-60 %), hidratos de carbono (25-50%) y grasas y aceites (10%) [12]. El agua residual en su fracción orgánica también contiene pequeñas cantidades de gran número de moléculas orgánicas sintéticas cuya estructura puede ser desde muy simple a extremadamente compleja [13].

Para eliminar estos contaminantes se lleva a cabo la depuración de las aguas residuales. El tratamiento convencional consiste en una combinación de procesos y operaciones físicas, químicas y biológicas para eliminar los sólidos, la materia orgánica y los nutrientes de las aguas residuales. Los términos generales utilizados

para describir los diferentes grados de tratamiento, en orden de aumento del nivel de tratamiento, son:

Fase 1: tratamiento preliminar o pretratamiento.

Fase 2: tratamiento primario.

Fase 3: tratamiento secundario y

Fase 4: tratamiento terciario y/o avanzado de las aguas residuales [14].

Cuando se realiza el diseño de las estaciones depuradoras de aguas residuales, se acude a modelos no lineales centrados en la optimización económica, pero, generalmente, se modeliza el tratamiento biológico, usando valores medios de los parámetros de entrada. Esto supone una aproximación simplificada, derivada de la falta de información existente en el momento inicial, donde aún no se pueden reproducir las condiciones que se van a encontrar en la operación real, pero provoca que las modelizaciones se ajusten poco a la realidad de las estaciones depuradoras de aguas residuales en fase de uso [15].

La evolución de las técnicas computacionales desde los años 90 provoca la proliferación de los modelados del funcionamiento de las estaciones depuradoras de aguas residuales, limitados a la zona de los tratamientos biológicos o línea de fango [16], [17], [18], [19] pero, de nuevo, estos modelos únicamente analizan el comportamiento de la depuradora a partir de los parámetros medidos en su interior al considerar el tratamiento biológico clave para la calidad de salida del agua tratada. Como consecuencia, no pretenden determinar de forma previa el influente que va a llegar a la depuradora lo que impide adaptar el proceso de depuración a las distintas circunstancias operacionales.

Esta falta de fiabilidad de las modelizaciones provoca la búsqueda de otros sistemas que sirvan para el control de las depuradoras.

Olsson [20] hace referencia por primera vez al potencial de los sistemas basados en el conocimiento para el diagnóstico de procesos y aparecen las primeras modelizaciones con procedimientos estocásticos con una búsqueda aleatoria, integrada y controlada que utiliza como datos de entrada los valores medidos por los sensores en continuo de caudal, oxígeno disuelto y sólidos en suspensión. Los datos de entrada no incluyen la materia orgánica ya que no existen sensores en el mercado para proporcionar esta medida *in-situ* y de continuo. La modelización parte de los datos de entrada al reactor biológico, no desde la entrada a la depuradora.

Con la rápida ascensión de los métodos basados en datos, dentro del campo de las depuradoras, las primeras simulaciones con redes neuronales artificiales las lleva a cabo Krovvidy [21] utilizando un algoritmo basado en un árbol de decisión para determinar el funcionamiento de un tratamiento físico-químico a partir de las medidas en continuo de varios compuestos.

A partir de esta primera aproximación, comenzaron los desarrollos de modelos más complejos del tratamiento biológico basados en sus ecuaciones básicas de funcionamiento mediante el modelo matemático tradicionalmente utilizado en el tratamiento de las aguas residuales [22]. Los modelos fueron haciéndose progresivamente más complejos, como en el caso de Peguero que, en su tesis doctoral [23], desarrolla un modelo de predicción de funcionamiento de una depuradora basado en redes neuronales artificiales, mediante la simulación del comportamiento del reactor biológico en función de la variabilidad de los parámetros de entrada medidos y actuando sobre la recirculación de fangos. Fang et al [24] desarrollan una red neuronal que busca obtener los parámetros de salida a partir de unos parámetros de entrada conocidos y medidos in situ.

De la misma forma, Joong-Wen Lee [25] desarrolla 4 modelos de comportamiento de DBO_5, DQO, sólidos en suspensión y nitrógeno mediante redes neuronales combinadas con algoritmos genéticos. Utiliza una serie de 2 años de datos diarios con el objetivo es predecir los parámetros de salida a partir de datos conocidos de entrada. Entre las variables de entrada utilizadas están el caudal, DBO, COD, sólidos en suspensión y nitrógeno.

Hammed [26] desarrolla una red neuronal para simular el funcionamiento de una estación depuradora de aguas residuales, utilizando como parámetros los sólidos en suspensión y la DBO a partir de datos diarios durante 10 meses. Precisamente esta toma limitada de datos, que no llega a cubrir un año completo, es una de las debilidades del modelo, ya que no considera toda la variación estacional.

Además de todas aquellas cuestiones ligadas al proceso, la mejora de las operaciones en las plantas de tratamiento y su impacto en el rendimiento de las mismas, la reducción del consumo energético y la disminución de los costes de mantenimiento está recibiendo cada vez más atención por parte de los investigadores [26]–[29]. Las exigencias legales y medioambientales, cada vez más estrictas, obligan a buscar una mejora en el funcionamiento de estas instalaciones [30], [31]. Una forma importante de optimizar esta operación es el desarrollo de modelos matemáticos del proceso. Muchos autores han desarrollado modelos matemáticos de las diferentes etapas de tratamiento de las plantas de tratamiento de aguas residuales [32], [33].

Como se puede observar en todas las referencias anteriores, los estudios de modelización de estaciones depuradoras se centran en el funcionamiento de los tratamientos biológicos. Si bien es cierto que esta parte del tratamiento del agua residual es un elemento fundamental, existen otras fases que son críticas a la hora de plantear el diseño de las estaciones depuradoras y que, hasta el momento, no han sido estudiadas de forma exhaustiva, seguramente por estar más ligadas a la parte operativa que al diseño.

La fase inicial de pretratamiento es clave para el buen funcionamiento de estas instalaciones [34], sin embargo, tal y como indican varios autores [35], [36], se han estudiado más los siguientes pasos de las plantas de tratamiento dado su gran

impacto en la recuperación del agua. En esta etapa de pretratamiento se realizan varias operaciones como el desbaste, la eliminación de arenas y el desengrasado.

Aunque la etapa de tratamiento preliminar ha sido menos estudiada, en parte debido a su gran dependencia del diseño inicial de la planta, su impacto en el rendimiento de las etapas posteriores es incuestionable. Un correcto funcionamiento del pretratamiento repercute de forma directa en los costes de explotación y en la eficiencia energética de la depuradora. No se puede optimizar el funcionamiento de una depuradora considerando únicamente la optimización del reactor biológico o de la línea de fangos.

Con el objetivo de poder estudiar y optimizar el funcionamiento del pretratamiento, el presente trabajo pretende predecir la llegada de arenas, sólidos de desbaste y grasa a las estaciones depuradoras, es decir, pretende predecir todos aquellos residuos que son eliminados en la fase de pretratamiento.

2.2 Composición de los residuos eliminados en el pretratamiento.

Las aguas residuales urbanas contienen diversos componentes como sólidos en suspensión, materia orgánica, aceites y grasas, pero no suelen contener sustancias peligrosas. Esta composición es consecuencia de que la mayoría de los colectores suelen ser unitarios, es decir, no sólo recogen las aguas domésticas sino también todo tipo de residuos procedentes de las calles arrastrados por las aguas pluviales, el riego o la limpieza municipal.

Cuando las aguas residuales entran en una instalación de tratamiento, pasan por una etapa denominada tratamiento preliminar o pretratamiento. En esta etapa se eliminan los sólidos brutos, la materia gruesa en suspensión y la flotante. Esta etapa de la depuración no ha recibido mucha atención por parte de la investigación y depende en gran medida de las características de diseño iniciales de la planta [35], [37], [38]. Sin embargo, su impacto en la gestión, operación y mantenimiento de las estaciones depuradoras, así como la influencia en el rendimiento de las etapas de tratamiento posteriores, es muy importante. En esta etapa de pretratamiento se realizan diversas operaciones, como el desbaste, la eliminación de arena y el desengrasado. Si estas sustancias no se eliminan correctamente hacen que las siguientes etapas del tratamiento funcionen de forma incorrecta.

2.2.1 Sólidos de desbaste

Los sólidos de desbaste en los colectores puede contribuir a crear varios problemas como atascos de los colectores [39], [40]. En los tramos de la red de alcantarillado en los que el agua circula por gravedad, los sólidos se combinan con las grasas y generan tapones que bloquean los colectores. Cuando el agua circula por bombeo, la presencia de sólidos gruesos puede provocar atascos en las bombas y desbordamientos de los pozos de bombeo con los consiguientes problemas de contaminación.

Por lo general, un tamiz elimina los objetos flotantes de gran tamaño, como trapos, latas, botellas y palos que pueden obstruir las bombas, las tuberías pequeñas y los procesos posteriores. Si los sólidos gruesos no se eliminan, quedan atrapados en las tuberías y otras partes móviles de la planta de tratamiento y pueden causar daños importantes e ineficiencia en el proceso [12], [41]. Los tamices se colocan generalmente en una cámara o canal y se inclinan hacia el flujo de las aguas residuales. La inclinación del tamiz permite que los residuos queden atrapados en la superficie aguas arriba del tamiz, pero también permite el acceso para la limpieza manual o mecánica. El propio depósito de sólidos de desbaste en los tamices hace que el efecto filtro sea mayor y más eficiente. Cuando el grado de atascamiento es tan alto que empieza a subir el nivel del agua en la zona previa al tamiz, los sistemas de limpieza arrancan de forma automática.

La gestión operativa de esta etapa inicial suele enfrentarse a varios problemas, entre ellos los siguientes:

- Los sólidos brutos en los días sin lluvia se depositan en el fondo de los colectores y, cuando se produce un episodio de lluvia intensa, son arrastrados bruscamente a la estación depuradora [42]. Numerosos investigadores han estudiado las consecuencias de estos sólidos en los sistemas de alcantarillado [43]–[48]. La llegada de todos estos sólidos brutos a la EDAR puede provocar atascos en los equipos y, en consecuencia, provocar el vertido de aguas residuales no tratadas al cauce receptor. Conocer la llegada de sólidos lo antes posible permitiría anticiparse y poner en servicio más líneas de pretratamiento, evitando esos atascos.
- Desde el punto de vista operativo, es necesario preparar suficiente espacio y sistema de recepción de residuos para los sólidos brutos y evitar tener que apilarlos en el suelo de forma precaria. Previendo la llegada de los sólidos brutos con mayor antelación es posible asegurar la disponibilidad de sistemas de recepción de residuos suficientes y planificar los transportes a las zonas de vertido.

Los procesos de tratamiento de las plantas depuradoras se monitorizan continuamente pero, a menudo, los datos recogidos no se aprovechan lo suficiente [49]. Por lo tanto, la utilización de los datos disponibles para mejorar la gestión desde los primeros procesos de tratamiento en la EDAR, redundará en una mejora del rendimiento de las etapas posteriores, en una disminución del consumo energético, de los problemas de mantenimiento de las instalaciones y, finalmente, en una mejor calidad del agua de salida.

2.2.2 Arenas

Uno de los componentes fundamental del agua residual es la arena, a la que están adheridos diversos componentes inorgánicos y orgánicos [50].

Las arenas provienen fundamentalmente de los arrastres de materiales existentes en el exterior y se separan durante el pretratamiento en los desarenadores. Los desarenadores se diseñan de forma que el agua residual al recorrerlos esté el tiempo suficiente como para permitir la decantación de la arena.

La práctica general [12] indica que los sistemas de desarenado recuperan el 95% de la arena de más de 200 micras y, por tanto, eliminan la mayor parte de los posibles problemas de funcionamiento de las instalaciones [36] . Sin embargo, en días de lluvia, con un mayor caudal de entrada y un aumento de la cantidad de arena, los sistemas no son capaces de absorber en su totalidad el material y se incrementa su paso a las siguientes etapas de tratamiento, generando problemas como la abrasión de los equipos de la línea de fangos o la disminución del rendimiento de los procesos biológicos [35] . Una predicción precisa del contenido de arena en el flujo de entrada de las plantas de tratamiento de aguas residuales permitiría llevar a cabo diferentes acciones preventivas, como la puesta en marcha de más líneas de pretratamiento, el vaciado previo de los contenedores y, en definitiva, la preparación de la planta de tratamiento para la eliminación de esta arena durante el pretratamiento. Además, reduciría el consumo energético, al no ser necesario transportar esta arena a otras etapas de la depuradora, y aumentaría la vida útil de los equipos electromecánicos, especialmente de las bombas.

2.2.3 Grasas

Las grasas y aceites son uno de los componentes de las aguas residuales urbanas y el resultado de la preparación de alimentos tanto en los hogares como en diversas actividades comerciales e industriales. Las grasas son una preocupación creciente para los ayuntamientos y los operadores de las plantas de tratamiento de aguas residuales, debido a su tendencia a causar graves obstrucciones en las tuberías y alcantarillas [51]–[53].

Las características de las grasas pueden variar enormemente en función de los tipos y fuentes de recogida [54]. Las grasas pueden aparecer como líquidos o sólidos y se caracterizan por tener una textura grasienta y una densidad inferior a la del agua, por lo que flotan en la superficie. Además, pueden formar emulsiones en medios acuosos en presencia de jabón u otros agentes emulsionantes. Las grasas están compuestas por ácidos grasos, triacilglicerol e hidrocarburos liposolubles, siendo los FFA (ácidos grasos libres) procedentes de la hidrólisis del resto de componentes por efecto de la temperatura, tiempo y humedad, los componentes más importantes debido a su reactividad química. La presencia de una gran cantidad de FFA da lugar a un pH característicamente bajo [51], [55].

Aguas arriba de las plantas de tratamiento, las grasas se unen a otro tipo de residuos generando los llamados "*fatbergs*" [52] que causan diferentes problemas en las tuberías de las plantas de tratamiento [56]. Debido a su relevante papel en el proceso, se han desarrollado diferentes sistemas de prevención con distintos enfoques, desde campañas educativas para promover las buenas prácticas de gestión, la instalación de sistemas de captura de grasas (GTS), o la realización de

inspecciones periódicas para evitar la eliminación inadecuada [57], [58]. En varios países se han puesto en marcha numerosas iniciativas y programas de este tipo, aunque en general son de ámbito local o a escala piloto y no se han extendido a nivel nacional o internacional [48], [51]. Un ejemplo de ello es la gestión municipal en Suecia y Noruega, donde al instalar GTS en la mayoría de los restaurantes, el número de problemas y atascos debidos a las acumulaciones de grasas se redujo significativamente [56].

En las estaciones de tratamiento, para poder eliminar las grasas, estas se emulsionan mediante la adición de aire, lo que permite que se transformen en espumas que se eliminan mediante el rascado de las mismas.

Las grasas que no se eliminan en el proceso de desengrasado pueden provocar atascos y otros problemas en sus infraestructuras (tuberías, bombas, depósitos, digestores, sensores, ...). Esto aumenta el tiempo y el dinero necesarios para la limpieza y el mantenimiento. El proyecto europeo *RecOil* estimó que el 25% de los costes de tratamiento de las aguas residuales pueden atribuirse al componente graso [59]. Por otra parte, si no se eliminan, las grasas consumen oxígeno del agua y empeoran los resultados de los tratamientos biológicos posteriores, reduciendo la calidad del agua tratada. Todos estos problemas requieren capacidad y energía adicionales en las plantas de tratamiento de aguas residuales, aumentando los costes de operación y mantenimiento de las instalaciones [52]. En consecuencia, se utilizan diferentes métodos para eliminar y reciclar estas grasas y aceites al principio de los procesos de depuración [60]–[63].

En comparación con otros trabajos de investigación realizados en relación con las grasas, normalmente centrados en el estudio de sus características físicas y químicas, los procesos de uso o reciclaje posterior, o su efecto en los tratamientos biológicos de las aguas residuales, este trabajo se centra en la mejora de la operatividad de las EDAR.

La separación mecánica de las grasas en el pretratamiento ha recibido menos atención por parte de los investigadores en comparación con su uso energético [51] [60], la reducción del impacto ambiental en los vertederos [54], [65], [66] o su influencia en los tratamientos posteriores en las EDAR [61].

Las plantas de tratamiento tienen que gestionar cambios significativos en el caudal y las características (composición, temperatura, etc.) de las aguas residuales entrantes [67], [68]. Más concretamente, son muchos los factores que influyen en la cantidad, proporción y características del contenido de las grasas de las aguas residuales de entrada de dichas instalaciones:

- Los cambios meteorológicos, es decir, las lluvias, más o menos intensas, la temperatura ambiente y el número de días previos sin lluvia con la consiguiente reducción del caudal de entrada, entre otros, modifican la cantidad y las características de las grasas que llegan a la EDAR. La predicción de estos fenómenos meteorológicos y su influencia en las diferentes

infraestructuras de gestión del agua ha sido estudiada en numerosos trabajos [67], [69]–[72];
- La aportación de grasas procedente de las actividades domésticas, alterada por los días festivos, los periodos vacacionales, las diferentes estaciones del año o el propio clima [53];
- Las características de las fuentes comerciales de las grasas (tamaño, densidad y distribución geográfica), como los restaurantes, y el uso de sistemas de captura de grasa, por ejemplo [51], [58];
- Las actividades industriales, como las fábricas de procesamiento de alimentos o los mataderos [63], [73], [74];
- La presencia de otros tipos de residuos mezclados con las grasas presentes en las aguas residuales, como sólidos brutos (especialmente toallitas húmedas), arenilla y otros [75].

2.3 Caso de estudio

Para la realización de esta investigación se ha contado con datos de diversas EDAR. No obstante, los modelos finales se han desarrollado y probado sobre la depuradora de Villapérez por ser la más completa y compleja: altos caudales de tratamiento, tecnología moderna, buena sensorización y una gran cuenca que incluye tanto actividades industriales como aguas residuales domésticas de ciudades importantes, lo que aumenta la complejidad del agua residual recibida y la hace más representativa de cualquier otra depuradora.

La Estación Depuradora de Aguas Residuales de Villapérez, está situada al noreste de la ciudad de Oviedo (Asturias, España) y ocupa una extensión cercana a las 21 hectáreas (Figura 3). Proporciona servicio a una población aproximada de 723.000 habitantes equivalentes. Las aguas residuales llegan a Villapérez a través de una red unitaria de colectores que tiene una longitud aproximada de 75 km e incluye 44 aliviaderos. Los diámetros de los colectores oscilan entre 600 mm y 2.000 mm con tramos en gravedad y en impulsión.

La EDAR de Villapérez recoge tanto aguas residuales urbanas como industriales, incluyendo una industria láctea con una capacidad de producción de 500.000.000 millones de litros de leche al año y que vierte un caudal medio de 200m^3/h a la red de saneamiento.

Figura 3: Vista general de la EDAR de Villapérez

La línea de proceso de esta EDAR incluye pretratamiento, tratamiento de tormentas, decantación primaria, tratamiento biológico, decantación secundaria y tratamiento terciario de afino.

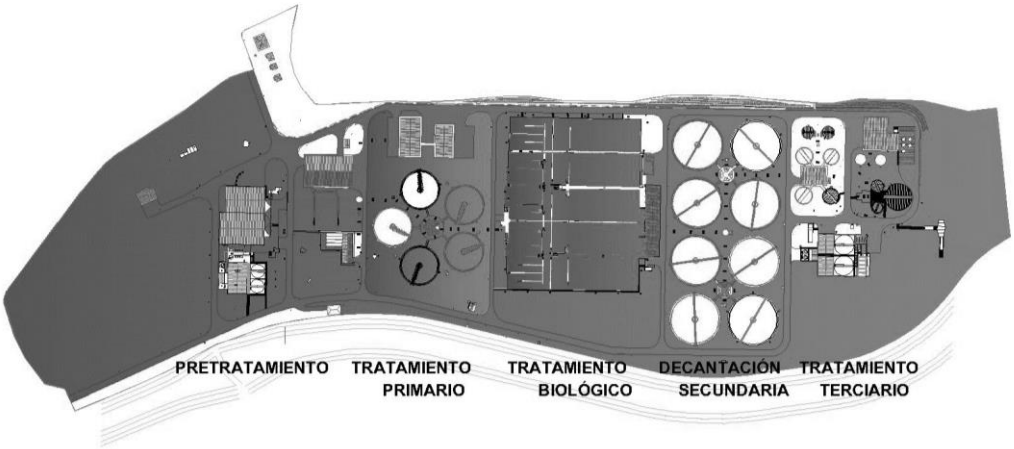


Figura 4: Línea de proceso EDAR de Villapérez

El pretratamiento tiene capacidad para tratar un caudal de 8,5 m^3/s y comienza con dos pozos de gruesos, equipados con una cuchara bivalva de 500 litros de capacidad. A continuación, la instalación dispone de cuatro canales de desbaste, que incluyen en cada uno de ellos una reja de limpieza automática de 60 mm de luz de paso y un tamiz de finos auto-limpiable de 3 mm de luz de paso e inclinado 50°.

Tras la etapa de desbaste, el agua llega a las instalaciones de separación de grasas y arenas del agua bruta que constan de 5 desarenadores rectangulares con un volumen útil unitario de 449,8 m³. Dichas unidades de desarenado garantizan la eliminación de partículas de tamaño igual o superior a 0,2 milímetros. En el canal de desarenado se realiza una inyección de aire mediante difusores de burbuja gruesa en el primer tercio de la longitud y aireadores sumergidos en los dos tercios restantes, asegurando un flujo giratorio y la flotación de las grasas. Este sistema persigue emulsionar las grasas en el agua, transformándolas en espumas y la separación de los flotantes además de ayudar a la decantación y limpieza de las arenas.

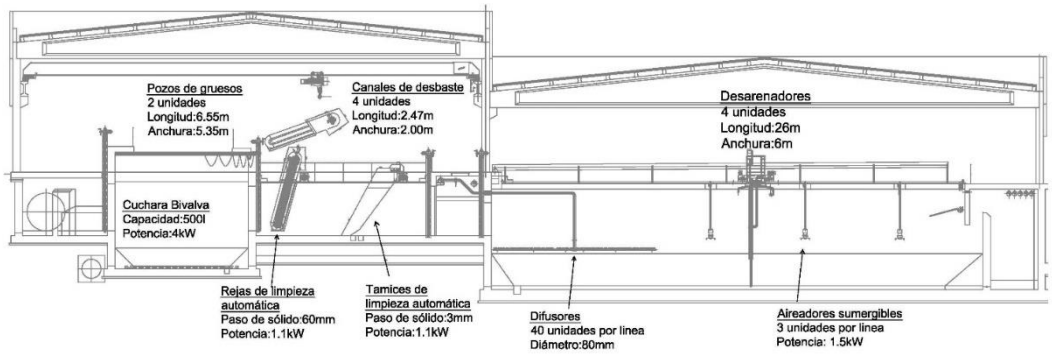


Figura 5: Pretratamiento de la EDAR de Villapérez

Los principales parámetros de diseño de la depuradora se muestran en la Tabla 1:

Tabla 1: Parámetros de diseño de la depuradora de Villapérez

Máximo caudal (tiempo de lluvia)	8,50 m³/s
Máximo caudal (tiempo seco)	2,89 m³/s
DBO5	418,00 mg/l
DQO	652,00 mg/l
SST	329,00 mg/l
N-NTK	47,40 mg/l
N-NO3	0,76 mg/l
Nt	48,20 mg/l
N-NH4+	25,00 mg/l
Pt	6,50 mg/l
P-PO4	3,24 mg/l

2.4 Caracterización de los residuos

Desafortunadamente, disponer de una medida general de los sólidos no es viable puesto que su comportamiento fluidodinámico es distinto y, en consecuencia, no serán predecibles de forma global. Los sólidos en suspensión son transportados

mediante flotación, mientras que las arenas son arrastradas por el fondo del lecho y las grasas lo hacen por distintos medios. Por ello, con el fin de realizar la modelización, es necesario en primer lugar un análisis de la composición de dichos sólidos. Para ello en la depuradora objeto del estudio se llevaron a cabo varias tomas de muestras de material realizando su posterior clasificación.

Figura 6: Clasificación de los sólidos de desbaste

Figura 7: Sólidos de desbaste ya clasificados

Tabla 2: Composición sólidos de desbaste

Fecha	Origen muestra	Hora inicio toma muestra	Hora fin toma muestra	Peso total húmedo	Toallitas %	Plásticos %	Higiene íntima %	Materia orgánica %
30/01/2018	Reja+tamiz	10:00:00	11:00:00	45.21 kg	30.57%	1.19%	2.92%	65.32%
01/02/2018	Reja	10:30:00	10:45:00	9.80 kg	90.31%	2.04%	7.65%	0.00%
1/02/2018	Tamiz	10:30:00	10:45:00	33.81 kg	9.52%	0.86%	2.57%	87.05%
01/02/2018	Reja+tamiz	10:30:00	10:45:00	43.61 kg	27.68%	1.12%	3.71%	67.48%
08/02/2018	Reja	10:45:00	11:00:00	1.13 kg	81.42%	0.88%	17.70%	0.00%
08/02/2018	Tamiz	10:45:00	11:00:00	8.12 kg	16.75%	0.00%	4.93%	78.33%
08/02/2018	Reja+tamiz	10:45:00	11:00:00	9.25 kg	24.65%	0.11%	6.49%	68.76%

Como se puede ver en la Tabla 2 aunque los porcentajes del residuo varían en función del momento en el que se toma la muestra, la composición se mantiene básicamente constante: toallitas, plásticos, productos de higiene íntima y materia orgánica no disuelta (restos de comida, etc.).

2.5 Recogida de datos

Las EDAR actuales recogen una gran cantidad de datos, muchas veces inutilizados para la gestión de las instalaciones, por lo que es necesario realizar un esfuerzo inicial de exploración, visualización y selección de la información relevante [49], [68].

Los datos empleados en este trabajo proceden de distintas fuentes:

- Los relativos al agua residual se obtuvieron mediante el software SCADA (*Supervisory Control and Data Acquisition*) de la EDAR. Este sistema registra cada 9 minutos 226 parámetros procedentes de equipos de medida y sensores distribuidos por la estación depuradora. De ese conjunto de datos se utilizaron los datos asociados a la medición de parámetros de entrada en el agua bruta en el pretratamiento. Los parámetros medidos en el agua bruta son el caudal de entrada, el pH, la temperatura del agua bruta, conductividad y amonio. Los datos asociados a esas variables se identifican mediante la hora y la fecha de medición del dato.
- Los datos de los distintos residuos (arenas, sólidos de desbaste y grasas) se llevan a cabo mediante medidas discontinuas. Estos datos se obtienen mediante la medición de las cantidades de residuos almacenados en depósitos que se vacían de forma regular. Para ello se recogieron los

albaranes de retirada de los respectivos residuos donde figura el dato real de peso total del residuo retirado en cada contenedor.

- Los datos de clima proceden de la página de Agencia Estatal de Meteorología (AEMET) de España y los datos pluviométricos (precipitación instantánea y precipitación acumulada) se obtienen de los registrados por la estación meteorológica de la propia planta. A partir de esos datos se crea además una variable calculada a partir de la precipitación instantánea, correspondiente al número de días previos sin lluvia.

CAPITULO 3. Metodología

3.1 Introducción

El elemento principal de la tesis es probar la viabilidad de desarrollar modelos basados en datos que sean capaces de predecir la cantidad de material sólido de entrada. Al tratarse de un problema consistente en desarrollar modelos a partir de la información procedente de datos, se ha seguido la metodología más habitual en estos procesos de *Machine Learning*, *CRISP-DM* [76].

Para el desarrollo de los modelos se ha decidido probar distintas técnicas puesto que no es posible conocer con antelación cuál de los algoritmos va a ser más eficiente en cada caso. Dado que en todos los casos se trata de problemas predictivos, se han evaluado diversas técnicas, primando en la selección aquellas que aportan más capacidad predictiva. Específicamente, para modelar los parámetros de las EDAR se han utilizado *ANN* (Redes Neuronales Artificiales), *FIS* (Sistema de Inferencia Difusa), *ANFIS* (Sistemas Neuronales de Inferencia Difusa Adaptativos) y *RF* (*Random Forest*) [26].

Finalmente se seleccionaron tres algoritmos: *MARS*, *SVM* y *Random Forest*, que serán explicados a continuación.

A continuación, se describen la metodología y técnicas mencionadas anteriormente, incluyendo sus criterios de aplicación y optimización de parámetros.

3.2 Metodología CRISP-DM

El modelado de un problema a partir de la información contenida en un conjunto de datos es un proceso iterativo e interactivo que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles. Aunque existen diversas metodologías, la más aceptada y extendida es *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) desarrollada en el año 2000 por un importante consorcio de empresas europeas liderado por Mercedes y SAS [15].

La metodología *CRISP-DM* se estructura en seis fases, de acuerdo a la Figura 8.

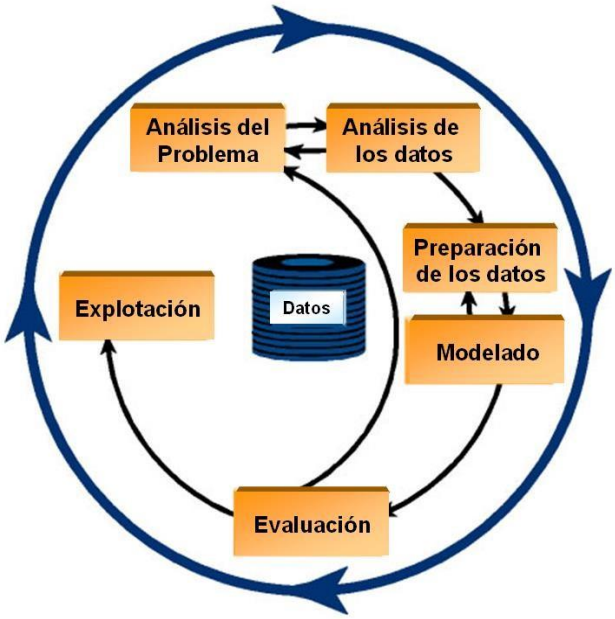


Figura 8: Fases del proceso de modelado según la metodología CRISP-DM

Las flechas indican relaciones más habituales entre las fases, aunque se admite que puedan establecerse otras relaciones. El círculo exterior representa la naturaleza iterativa del proceso de modelado.

Una descripción más completa de las principales tareas y subtareas a desarrollar en cada una de las fases se muestra en la siguiente tabla.

Tabla 3: Principales fases, tareas y subtareas de la metodología CRISP-DM

Análisis del problema	Análisis de los datos	Preparación de los datos	Modelado	Evaluación	Explotación
Determinación de los objetivos empresariales	**Adquisición de los datos**	**Procesado de datos**	**Selección de la técnica de modelado**	**Evaluación de los resultados**	**Planificación de la explotación**
Conocimiento previo *Objetivos* *Criterios de éxito*	*Análisis fuentes datos* *Estudio datos disponibles* *Instalación base de datos*	*Conversión a valores numéricos* *Rellenado de datos* *Identificación de valores no usuales*	*Técnicas de modelado* *Supuestos de la técnica de modelado*	*Valoración de los resultados* *Modelos válidos*	*Plan de utilización*
Evaluación de la situación	**Descripción de los datos**	**Reducción de la dimensionalidad**	**Diseño del método de evaluación**	**Revisión del proceso**	**Planificación de la monitorización y mantenimiento**
Recursos disponibles *Requerimientos, supuestos y restricciones* *Terminología*	*Tipos* *Unidades* *Significado* *Procedencia*	*Variables* *Muestras*	*Medidas de error*	*Detección de errores en el proceso de modelización*	*Plan de monitorización y mantenimiento*
Determinación de los objetivos técnicos	**Explotación de los datos**	**Transformación de datos**	**Generación del modelo**	**Determinación de las siguientes acciones**	**Revisión del proyecto**
Objetivos del modelado *Criterios de éxito*		*Normalización* *Transformaciones matemáticas* *Discretización*	*Parámetros del modelo* *Modelos* *Descripción del modelo*	*Lista de las posibilidades* *Decisión*	*Extracción de conclusiones*
Elaboración de la estrategia	**Verificación de la calidad de los datos**		**Evaluación del modelo**		
Planificación *Valoración inicial* *Técnicas y herramientas*			*Verificación de los resultados* *Obtención de más información*		

3.3 Técnicas utilizadas

En base a estos algoritmos se ha modelizado la predicción de residuos optando en cada caso por el tipo de técnica con la que se obtenía el menor error. La técnica que menor error ha dado ha sido diferente en cada tipo de residuo. Las técnicas aplicadas son SVM, MARS y RF.

3.3.1 SVM (*Support Vector Machine*)

El término SVM (*Support Vector Machine*) se refiere a un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T [77]. Aunque inicialmente se desarrolló como un método para la clasificación binaria, su aplicación se ha extendido a problemas de clasificación múltiple y regresión. Los algoritmos SVM se han utilizado con éxito en muchos campos diferentes, como la visión por ordenador, el reconocimiento de caracteres, la categorización de textos e hipertextos, la clasificación, el procesamiento del

lenguaje natural o el análisis de series temporales [78]–[80]. Esto se debe a que este método ha mostrado una buena capacidad de generalización, evitando problemas de sobreajuste del entrenamiento como tienen otros métodos similares [81]. Recientemente también se ha utilizado en el campo del tratamiento de aguas residuales para predecir diferentes parámetros del proceso de tratamiento [67], [68], [82]–[87] .

La forma más sencilla de resolver problemas de clasificación es utilizar técnicas lineales, es decir, separar las zonas que identifican los distintos grupos mediante líneas rectas, planos o hiperplanos en función de su dimensionalidad.

Un modelo SVM parte de un conjunto de muestras de entrenamiento donde cada una de ellas será marcada previamente como perteneciente a una de las categorías deseadas, con el objetivo de ser representadas como puntos del espacio que permanecerán divididos por una separación tan amplia como sea posible. Una vez lograda dicha separación, conocida como hiperplano, los nuevos datos de entrada serán clasificados en el mismo espacio, lo que permitirá predecir a que categoría pertenecen cada uno de ellos.

Así, dado un conjunto de observaciones o datos

$$S = \{(x_1, y_1), \dots (x_n, y_n)\} \tag{1}$$

donde,

$$x_i \epsilon R^d \tag{2}$$

$$y_i \epsilon \{+1, -1\} \tag{3}$$

En base a lo anterior se puede definir un hiperplano de separación:

$$D(x) = (\omega_1 x_1 + \cdots + \omega_d x_d) + b =< \omega, x > +b \tag{4}$$

Como se puede deducir, el hiperplano no es único, lo que lleva a buscar el hiperplano de separación óptimo (HSO), que será aquel cuyo margen o distancia mínima entre el hiperplano y el punto más cercano (τ) sea máximo.

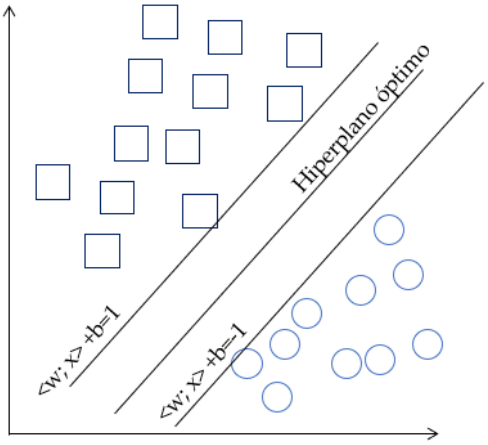


Figura 9: Hiperplanos en SVM

$$Distancia\ (D(x), x') = \frac{|D(x')|}{\|\omega\|} \tag{5}$$

$$\frac{y_i D(x_i)}{\|\omega\|} \geq \tau \qquad i = 1, .., n \tag{6}$$

Todos aquellos elementos que limitan con la frontera del hiperplano reciben el nombre de vectores soporte (V_s) y cumplen:

$$\frac{y_i D(x_i)}{\|\omega\|} = \tau \qquad \forall i \in V_s \tag{7}$$

Estos vectores son los más cercanos a los límites que marcan la zona de su categoría y, por tanto, más difíciles de clasificar por lo que se convierten en los elementos clave de trabajo.

El margen máximo que separa ambas categorías es igual a $\frac{2}{||w||}$, y viene dado por la menor distancia de los puntos al hiperplano. Por tanto, el objetivo será encontrar la función que permite maximizar $\frac{2}{||w||}$ con restricciones lineales, la cual permitirá mediante el signo resultante que el clasificador indique la categoría a la que pertenece la observación analizada.

Todo lo anterior se cumple siempre y cuando exista un hiperplano que permita clasificar la información. El problema es que los casos a estudiar suelen ser más complejos, no lineales, y no pueden ser separados de forma lineal más que en el caso de poder utilizar un número de casos infinito que sea capaz de generar cualquier curva a partir de elementos rectos. En estos casos es necesario acudir al uso de funciones que transformen los datos conocidas como *kernels*.

Las funciones *kernel* buscan proyectar la información en un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal, es decir, establecemos una nueva o nuevas

dimensiones en las que podremos encontrar el hiperplano de separación para maximizar los márgenes entre cada elemento y el hiperplano obtenido.

Para ello se utiliza una función ϕ de transformación que hace corresponder a cada punto x un punto en el espacio de características F, donde

$$\phi(x)=[\phi_1(x),..,\phi_m(x)] \tag{8}$$

donde $\exists\phi_i(x)$, $i=1,..,m$ tal que $\phi_i(x)$ es una función no lineal.

En base a lo anterior la función del hiperplano que es necesario determinar sería la siguiente:

$$D(x) = \left(\omega_1\phi_1(x) + \cdots + \omega_m\phi_m(x)\right) =< \omega, \phi(x) > \tag{9}$$

La elección del *kernel*, así como la selección particular de los parámetros ajustables del *kernel*, tienen una influencia fundamental sobre el rendimiento del modelo [88]. Para realizar esta transformación, se pueden utilizar diversas técnicas de análisis multivariado que se puedan formular en términos de productos Entre las más usadas se encuentran:

- *Kernel* lineal: $K(x_i, x_k) = x_i * x_k$
- *Kernel* polinómico: $K(x_i, x_k) = (x_i * x_k + c)^d$
- *Kernel* gaussiano radial: $K(x_i, x_k) = exp(-\gamma\, ||x_i - x_k||^2)$

En este trabajo se probaron distintos tipos de funciones *kernel* comúnmente utilizadas, como las funciones lineales, polinómicas, sigmoides y de base radial [89].

Se aplicó la metodología de búsqueda en cuadrícula con validación cruzada 10 veces sobre el conjunto de entrenamiento para establecer el mejor tipo de función *kernel* y recuperar los valores óptimos para los parámetros del modelo. Este procedimiento de validación cruzada *k-fold* se considera uno de los enfoques más utilizados para evaluar los valores de los parámetros de la arquitectura del modelo [90], [91]. Tras este proceso, el mejor *kernel* para la clasificación resultó ser la Función de Base Radial Gaussiana (*RBF*) porque produce la mayor precisión global [92] (ecuación 9).

$$k(x_i, x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} \tag{10}$$

donde, σ es un parámetro libre y $||x_1 - x_2||$ es la distancia euclídea entre dos puntos x_1 y x_2.

Para la realización de las simulaciones se utilizó la implementación de SVM sobre el entorno R [93].

3.3.2 MARS (*Multivariate adaptive regression splines*)

Desde que Jerome H. Friedman presentó el método de regresión no paramétrica conocido como MARS (*Multivariate adaptive regression splines*) se ha utilizado en

muchas aplicaciones en diversos campos [94]. Entre sus ventajas destacan su capacidad para identificar relaciones no lineales en los datos, generar modelos simples y más fácilmente interpretables a partir de un gran número de variables de entrada, mostrar su importancia relativa y ser computacionalmente eficiente en comparación con otras técnicas [95]–[97]. En el ámbito de las EDAR, se ha utilizado recientemente en diferentes estudios para predecir la demanda bioquímica y química de oxígeno, la concentración de nitrógeno, fósforo y sólidos totales en suspensión [98], [99], la capacidad de sedimentación de fangos activados [98] o la reducción de costes [100] .

El modelo MARS de una variable dependiente, con n funciones o términos base puede expresarse como [94]:

$$\hat{y} = c_0 + \sum_{i=1}^{n} c_i \cdot B_i\,(\vec{x}) \tag{11}$$

Donde $\hat{y}$ es la variable dependiente predicha por el modelo MARS, c_0 es una constante, Bi son las n funciones base y c_i son los coeficientes de cada una de ellas. Las funciones base Bi tienen el aspecto de $max(0,x-t)$ ó $max(0,t-x)$ donde t es una constante llamada nodo o punto de corte de las diferentes variables. Estos puntos de partición del espacio, así como los parámetros del modelo, se obtienen a partir de los datos analizados mediante un algoritmo *hacia delante/hacia atrás* de dos pasos. Primero, mediante el algoritmo *hacia adelante stepwise*, se genera un modelo sobreajustado con un gran número de funciones base y después, mediante el algoritmo *hacia atrás*, se eliminan los nodos que menos contribuyen al ajuste global. El espacio se divide en regiones, ajustando una función de base radial en cada una de ellas. El modelo final es una combinación de todas las funciones de base generadas, cuyo número indica la complejidad del modelo.

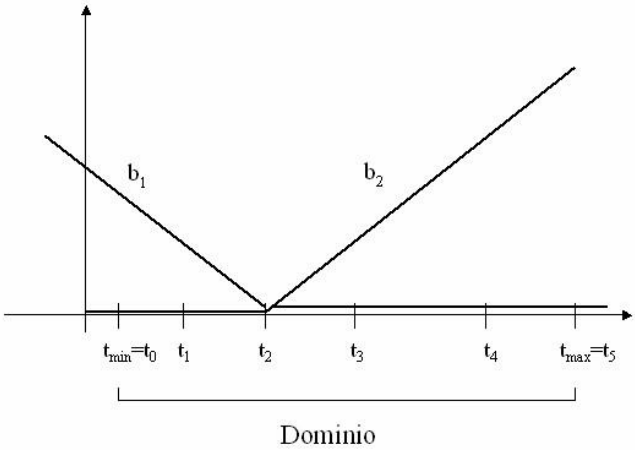


Figura 10: Funciones base asociadas a un nodo

Mediante este proceso, el método MARS realiza una selección automática de las variables de entrada, es decir, incluye las variables importantes en el modelo y excluye las no relevantes. Sin embargo, es necesario considerar previamente la

posible colinealidad de las variables predictoras. En este caso, se estudió la correlación entre las variables y se determinó el factor de inflación de la varianza (VIF), eliminando en el grupo de variables predictoras aquellas con un VIF superior a 10.

3.3.3 RF (*Random Forest*).

Los bosques aleatorios (RF en adelante) son un algoritmo popular y eficiente, basado en la agregación de modelos (*ensembles*), utilizado tanto para problemas de clasificación como de regresión, introducido por Breiman [101] .Pertenece a la familia de los métodos de conjunto, que aparecieron en el aprendizaje automático a finales de los noventa (véase, por ejemplo, Dietterich (1998) [102] y (2002) [103]). Recordemos brevemente el marco estadístico considerando un conjunto de aprendizaje L = {(X1, Y1), ...(Xn, Yn)} formado por n observaciones i.i.d. de un vector aleatorio (X, Y). El vector X = (X1..., Xp) contiene predictores o variables explicativas, digamos que X ∈ R p e Y ∈ Y, donde Y es una etiqueta de clase o una respuesta numérica. Para los problemas de clasificación, un clasificador t es un mapeo

t : Rp → Y mientras que para los problemas de regresión, suponemos que

$$Y = s(X) + \varepsilon \text{ y} \qquad (12)$$

s es la llamada función de regresión.

El principio de los *Random Forest* es combinar muchos *árboles de decisión* binarios construidos utilizando varias muestras *bootstrap* procedentes de la muestra de aprendizaje L y eligiendo aleatoriamente en cada nodo un subconjunto de variables explicativas X.

Con respecto a la conocida estrategia de construcción de modelos CART (véase Breiman et al. (1984) [104]) realizando un paso de crecimiento seguido de uno de poda, se pueden observar dos diferencias. En primer lugar, en cada nodo, un número determinado (denotado por *mtry*) de variables de entrada y se calcula la mejor división sólo dentro de este subconjunto. En segundo lugar, no se realiza ningún paso de poda, por lo que todos los árboles son árboles máximos.

La novedad introducida por la metodología *Random Forest* respecto a los CART es la aleatoriedad introducida para reducir la correlación entre los árboles.

Para formarse el árbol de decisión se aplican los siguientes pasos:

1. De forma aleatoria se generan los árboles de decisión mediante la técnica *bootstraping*, introduciendo la aleatoriedad.

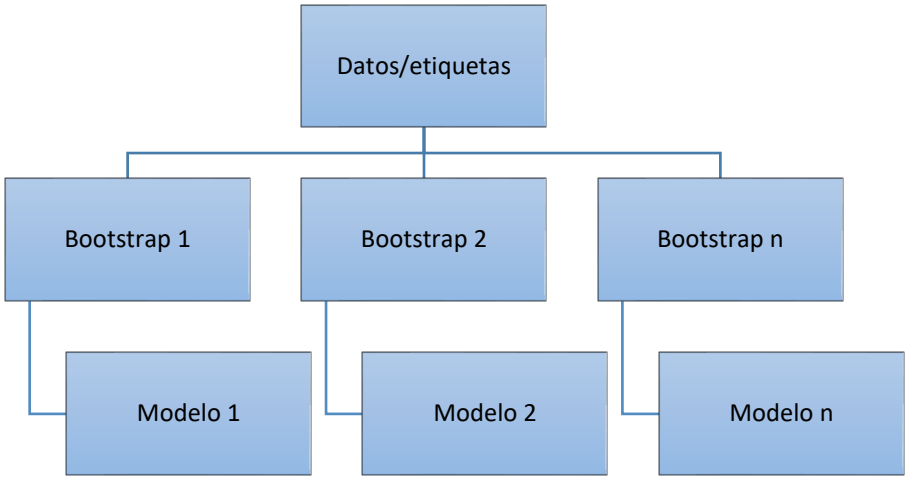


Figura 11: Técnica bootstraping

2. Dadas n variables de forma aleatoria se seleccionarán p variables tales que $p \lll M$, siendo M el número de muestras.
3. Por último, se dejará crecer el árbol.

En base a lo anterior se puede establecer que esta técnica depende de dos números fundamentales, el número de árboles de decisión y el número de variables, p, que se seleccionan en cada nodo.

El error generado está relacionado con estos parámetros, al reducir el número de variables se reduce la correlación entre los árboles, pero también se reduce la precisión del árbol y por lo tanto habrá que buscar un equilibrio con estos parámetros.

En este estudio, se utilizaron bosques aleatorios (RF), un enfoque de aprendizaje automático para la selección de características a partir de conjuntos de datos altamente multivariados, para desarrollar un modelo de predicción del contenido de la variable grasa (FOG) en las aguas residuales de entrada. El algoritmo RF llega a la predicción final a partir de la votación mayoritaria de las decisiones tomadas con múltiples árboles de decisión construidos con características y observaciones permutadas aleatoriamente a través de la partición recursiva [101].

El método RF se ha aplicado en una amplia gama de áreas de investigación debido a sus numerosas ventajas [105] y en los últimos años ha ganado gran importancia en la investigación relacionada con los recursos hídricos. La técnica de *Random Forest* se ha utilizado para abordar numerosos problemas de investigación en las EDAR como:

- estimar diferentes parámetros de la calidad del agua como la Demanda Química de Oxígeno (DQO) [106], los sólidos en suspensión totales (SST) [107], las concentraciones de nitrógeno (N) y fósforo (P) de las corrientes [108] o el caudal afluente de las EDAR [109].

- monitorizar diferentes procesos de tratamiento como, por ejemplo, para hacer predicciones de la capacidad de sedimentación de los lodos activados [110] o de los sistemas de eliminación de nitrógeno [111].
- generar modelos de coste energético [112] o de sistemas de bombeo [113] en las EDAR.
- obtener otras mejoras en el control de la planta [114] o en la fiabilidad de pequeñas plantas de tratamiento de aguas residuales [115].

Para su implementación y uso en este trabajo, se desarrolló el código utilizando R [93] y los paquetes *Cart* [116] y *RandomForest* [117].

3.4 Datos de partida

El histórico de datos empleados no es el mismo en cada uno de los modelos puesto que, según va aumentando la complejidad de las variables a predecir, es necesario ir ampliando el conjunto de datos. Por otra parte, la cantidad de residuo generado es diferente para los tres componentes estudiados: sólidos, arena y grasas.

Los datos empleados en este trabajo proceden de distintas fuentes:

- Los relativos al agua residual se obtuvieron mediante el software SCADA (*Supervisory Control and Data Acquisition*) de la EDAR. Este sistema registra cada 9 minutos 226 parámetros procedentes de equipos de medida y sensores distribuidos por la estación depuradora. De ese conjunto de datos se utilizaron los datos asociados a la medición de parámetros de entrada en el agua bruta en el pretratamiento. Los parámetros medidos en el agua bruta son el caudal de entrada, el pH, la temperatura del agua bruta, conductividad y amonio. Los datos asociados a esas variables se identifican mediante la hora y la fecha de medición del dato. Estos datos son datos continuos.
- Los datos de los distintos residuos (arenas, sólidos de desbaste y grasas) se llevan a cabo mediante medidas discontinuas. Estos datos se obtienen mediante la medición de las cantidades de residuos almacenados en depósitos que se vacían de forma regular. Para ello se recogieron los albaranes de retirada de los respectivos residuos donde figura el dato real de peso total del residuo retirado en cada contenedor.
- Los datos de clima proceden de la página de Agencia Estatal de Meteorología (AEMET) de España [118] y los datos pluviométricos (precipitación instantánea y precipitación acumulada) se obtienen de los registrados por la estación meteorológica de la propia planta. Todos ellos se agrupan también teniendo en cuenta los intervalos de llenado de los contenedores. A partir de esos datos se crea además una variable calculada a partir de la precipitación instantánea, correspondiente al número de días previos sin lluvia.

Hay que tener en cuenta que la calidad de los datos recogidos en las instalaciones industriales suele presentar diversos problemas de fiabilidad debido a las difíciles condiciones ambientales de trabajo de los sensores, lo que implica una alta

variación e incluso errores en las medidas obtenidas [38]. Por ello, se ha realizado un esfuerzo considerable en la recogida de datos durante más de dos años a partir de diversas fuentes. De este modo se consiguen dos mejoras:

1. La mayor representatividad de las muestras puesto que el mayor tamaño de la muestra permite acudir a datos medios y minimizar los errores puntuales.
2. Los datos incluyen información de carácter estacional, cambios en la actividad doméstica o industrial, largos periodos de lluvias intensas o tiempo seco, etc. Así, son representativos de todas las condiciones de funcionamiento de la instalación.

A partir de esas fuentes se recogió información para cada estudio. El primer elemento analizado es la arena; para este análisis se emplean los datos comprendidos entre el 1/03/2017 al 6/03/2018. En ese tiempo se generan 187 datos discretos de medición de cantidad generada de arena. Para la predicción de la cantidad de sólidos de desbaste se emplean los datos comprendidos entre el 1/03/2017 al 24/06/2019. En ese tiempo se generan 165 datos discretos de cantidad generada de sólidos de desbaste. Este mismo conjunto de datos se utiliza para la predicción de la cantidad de grasa. En ese tiempo se generan 89 datos discretos de cantidad generada de grasa.

3.5 Pretratamiento de datos

Otro reto importante de este trabajo ha sido la selección y posterior tratamiento de las variables de entrada para disponer de un número adecuado de patrones de entrenamiento y prueba. Las EDAR actuales recogen una gran cantidad de datos, muchas veces inutilizados para la gestión de las instalaciones, por lo que es necesario realizar un esfuerzo inicial de exploración, visualización y selección de la información relevante [49], [68].

El resultado de la simulación solo podrá ser representativo del proceso si se consigue que sólo se introduzcan en el sistema datos fiables. Por tanto, antes de su introducción en la base de datos se le aplicaron a los diferentes grupos de datos los siguientes tratamientos:

1. Tratamiento de datos ausentes o incompletos. En ocasiones algunos de los datos, especialmente los procedentes del SCADA, carecen de datos porque estos no han sido recogidos por problemas con los sensores o en la transmisión. En los casos de series temporales estos datos han sido completados mediante interpolación. En los datos no temporales se ha eliminado el patrón incompleto.
2. Detección de datos erróneos. Se establecieron unos valores máximos y mínimos de cada una de las variables, realizando una primera limpieza, eliminando aquellos datos que superaban los márgenes de funcionamiento. En los datos procedentes del SCADA, se debía a funcionamientos erróneos, en los casos de variables manuales se debían a errores humanos. Los datos

estadísticos de las variables consideradas inicialmente en el estudio se presentan en la Tabla 4. Como se ha indicado anteriormente, la referencia es el intervalo de tiempo desde que se coloca un contenedor vacío hasta que se retira. Cuando se retira cada contenedor, se pesa, y los datos se registran en el albarán correspondiente. Para la elaboración de los patrones de entrenamiento, se han calculado algunas variables. Los datos correspondientes a cada uno de estos periodos se han resumido calculando para cada variable su valor mínimo, medio y máximo, como se muestra en la Tabla 4.

Tabla 4: Variables empleadas

Variable	Descripción	Ud.	Media	Mínimo	Maximo
FOG	Aceites y grasas	ton	3,01	2,32	3,52
Interval	Intervalo de tiempo	h	228,91	1,28	6289,76
PDwR	Días previos sin lluvia	día	2,06	0,00	19,55
MxDwR	Máximo de días sin lluvia	día	4,42	0,07	20,68
Vol	Volumen de agua	m^3	731.056,32	3.946,58	4.886.022,43
PrecipTotal	Precipitación total	m^3	13,88	0,00	203,40
PrecipMax	Precipitación máxima	m^3	1,09	0,00	12,00
pH	pH	-	7,21	6,22	7,99
pHMax	Máximo pH en el periodo	-	8,20	7,01	11,65
MedTemperature	Temperatura media en el agua	°C	17,98	10,95	22,55
MaxTemperature	Temperatura máxima en el agua	°C	19,59	12,68	25,62
MedConductivity	Conductividad media	µS/cm	996,72	380,80	1439,72
MaxConductivity	Conductividad máxima	µS/cm	1995,47	757,62	3768,84
MedAmmonium	Amonio medio	mg/L	27,61	9,06	68,31
MaxAmmonium	Amonio máximo	mg/L	38,33	15,82	88,22
MedFlow	Caudal medio	m^3/h	4.193,95	2.446,96	12.608,21
MaxFlow	Caudal máximo	m^3/h	9.216,91	3.446,37	17.885,11
MinFlow	Caudal mínimo	m^3/h	1.779,97	975,59	6.803,43
TempExtMed	Temperatura ambiente media	°C	13,14	3,30	22,20
TempExtMax	Temperatura ambiente máxima	°C	17,50	4,60	28,20
TempExtMin	Temperatura ambiente mínima	°C	9,75	−0,20	17,60
MedPDwR	Media de días sin lluvia	día	2,12	0,01	19,52

3. Unificación de la frecuencia de muestreo. Los datos son muestreados en periodos muy diferentes. Las diferencias oscilan entre los dos minutos de algunas de las variables de proceso y los días de las variables de muestreo. Los muestreos se han realizado acumulando los valores en los puntos de muestreo físico, obteniendo así 165 patrones. Igualmente se generó un *dataset* completo

con frecuencias cada 9 minutos dividiendo las variables de muestreo en el intervalo de tiempo de forma proporcional al caudal.

CAPITULO 4. Publicaciones

water

MDPI

Article

Sand Content Prediction in Urban WWTPs Using MARS

Vanesa Mateo Pérez, José Manuel Mesa Fernández *, Francisco Ortega Fernández and Henar Morán Palacios

Project Engineering Area, University of Oviedo, Independencia, nº 13, 33012 Oviedo, Asturias, Spain; mateovanesa@uniovi.es (V.M.P.); fdeasis@uniovi.es (F.O.F.); moranhenar@uniovi.es (H.M.P.)

* Correspondence: jmmesa@uniovi.es

Received: 17 April 2020; Accepted: 8 May 2020; Published: 11 May 2020

Abstract: The pre-treatment stage of wastewater treatment plants (WWTP), where most of the larger waste, including sand and fat, is removed, is of great importance for the performance and durability of these plants. This work develops a model that predicts the sand content that reaches the plant. For this purpose, data were collected from one operation year of the Villapérez Wastewater Treatment Plant located in the northeast of the city of Oviedo (Asturias, Spain) and the MARS (Multivariate Adaptive Regression Splines) method was used for modelling. The accuracy of the MARS model developed using the determination coefficient is $R^2 = 0.74$ for training data and $R^2 = 0.70$ in validation data. These results indicate that it is possible to predict trend changes in sand production as a function of input variables changes such as flow rate, pH, ammonia, etc. This will prevent the plant from possible operational problems, as actions could be taken, such as starting up more pre-treatment lines or emptying the containers, so that the arrival of the sand can be assumed without any problem. In this way, the possibility of letting sand contents over the established limits pass that could affect the following processes of the treatment plant is avoided.

Keywords: wastewater; pre-treatment; sand; MARS method

1. Introduction

General urban wastewater contains various components such as suspended solids, organic matter, oils and fat, but usually no hazardous substances. This composition is a consequence of the fact that most collectors are usually unitary, that is, they not only collect domestic water but also all kinds of waste from the streets dragged by rainwater, irrigation or municipal cleaning.

The initial pre-treatment phase is of great importance for the proper functioning of WWTP (wastewater treatment plants) [1]; however, as indicated by several authors [2,3], the next steps of these facilities have been studied further given their great impact on water recovery. Parameters such as pH, chemical oxygen demand (COD), biochemical oxygen demand (BOD) and suspended solids (SS) in later stages have been the subject of numerous studies [4–7]. In contrast, the pretreatment stage has been studied much less, and efficient operation is considered to primarily depend on a good initial plant design and proper operations management.

In this pre-treatment stage, several operations are carried out such as roughing, sand removal and degreasing. The purpose of the pretreatment is to remove suspended solids in the water by mechanical processes [8]. The usual pretreatment stage in WWTP begins with automatic bar screens where larger wastes, such as rags and plastics, are removed. Then, depending on the WWTP design, gross solids are also collected by mean of different types of screens as, for example, rotor screens or statics screens. Finally, aerated grit chambers are generally used for the sand removal process. Included in this "sand" are various inorganic and organic components [9].

General practice [8] indicates that sand trap systems recover 95% of the sand above 200 microns and therefore eliminate most of the possible operational problems of the facilities [10]. However, on rainy days, with a higher input flowrate and an increase in the amount of sand, its passage to the following treatment stages increases, generating problems such as abrasion of the equipment of the sludge line or affecting the performance of the biological processes [2]. An accurate prediction of the sand content in the input flow of wastewater treatment plants would allow different preventive actions to be taken, such as commissioning more pre-treatment lines, emptying the containers beforehand and ultimately prepare the treatment plant for the elimination of this sand during the pre-treatment. Besides, this will reduce energy consumption, since it is not necessary to transport this sand to other stages of the treatment plant, and increase the service life of the electromechanical equipment, especially the pumps.

Therefore, the aim of this work is to predict the sand content in the WWTP inlet flow to anticipate the most convenient operational decisions and improve the efficiency of the pretreatment facilities. In order to achieve this prediction will be used a method based on data called MARS [11]. This technique is a novelty in this area of study but has been used successfully in many other applications as described in the next section.

2. Materials and Methods

2.1. Case Study

The Villapérez Wastewater Treatment Plant is located in the northeast of the city of Oviedo (Asturias, Spain) covers an area of nearly 21 hectares (Figure 1). It provides service to a population of approximately 723,000 inhabitants equivalent. The network of collectors feeding the treatment plant is unitary. The process line of this WWTP includes pre-treatment, storm treatment, primary decantation, biological treatment, secondary decantation and tertiary treatment of wastewater.

Figure 1. View of the wastewater treatment plant (WWTP) of Villapérez (Asturias, Spain).

The pre-treatment has the capacity to treat an inflow of 8.5 m^3/s and starts with two thick wells, equipped with a 500-L clamshell bucket. The plant then has four roughing channels, each of which

includes an automatic cleaning screen with a 60 mm clearance and a self-cleaning fines screen with a 3 mm clearance and an inclination of 50°. After the roughing stage, water arrives at the raw water fat and sand separation facilities, which consist of 5 rectangular sand traps with a unit useful volume of 449.8 m^3. These sand removal units guarantee the elimination of particles of size equal to or greater than 0.2 millimeters. Air is injected into the grit removal channel by means of coarse bubble diffusers in the first third and submerged aerators in the remaining two thirds, which ensure a rotating flow and the floating of the fat. This system aims to break up the emulsion of the fat in the water and the separation of the floating particles, as well as helping to decant and clean the sand.

The main design parameters of the treatment plant are included in Table 1.

Table 1. Design parameters of wastewater treatment plant of Villapérez (Asturias, Spain).

Maximum Inflow (rainy weather)	8.50 m^3/s
Maximum Inflow (dry weather)	2.89 m^3/s
DBO5	418.00 mg/L
DQO	652.00 mg/L
SST	329.00 mg/L
N-NTK	47.40 mg/L
N-NO_3	0.76 mg/L
Nt	48.20 mg/L
N-NH^{4+}	25.00 mg/L
Pt	6.50 mg/L
P-PO_4	3.24 mg/L

2.2. Data

All data used in this work were collected in the period from 1 March 2017 to 6 March 2018 and come from different sources:

- Data related to wastewater were obtained through the SCADA software (Supervisory Control and Data Acquisition) of the WWTP. This system registers every 9 min 226 parameters from measuring equipment and sensors distributed all over the treatment plant. From this set of data, the data associated to the measurement of input parameters in the raw water during the pre-treatment stage were used. The parameters measured in the raw water are the input flow rate, pH, raw water temperature, conductivity and ammonia. The data associated with these variables are identified by the time and date of the data measurement.
- Sand data were collected from the container removal delivery notes, which contain the actual data of the waste total weight inside each container. The number of containers in the study period was 187. Their filling time was used as time intervals to group the data of the SCADA system.
- Climate data comes from the Spanish State Agency for Meteorology website (Agencia Estatal de Meteorología, Aemet) and the pluviometry data (instantaneous and accumulated rainfall) is obtained from those recorded by the plant's own weather station. All of them are also grouped considering the intervals in which the containers are filled. From these data, a new calculated variable from the instantaneous precipitation is also created, corresponding to the number of previous days without rain.

The obtained data set (187 cases) was divided into two groups. Eighty percent of the data were used for training the MARS (Multivariate Adaptive Regression Splines) model and the remaining 20% were kept for validating the model. This method was selected over other data-based techniques for its ability to identify non-linear relationships in the data, for being easier to understand the importance of input variables, and for the relatively lower computational cost.

Statistical data for the variables initially considered in the study are presented in Table 2. As indicated above, the reference is the time interval (Time) from when an empty container is placed to when it is removed. When each container is removed, it is weighed, and the data is recorded on the

corresponding delivery note. The data corresponding to each of these periods was summarized by calculating for each variable its minimum, mean and maximum value, as shown in Table 2.

Table 2. Statistical description of the variables.

Variable	Description	Unit	Mean	Standard Deviation	Minimum	Maximum
Sand (Output)	Removed sand	ton	4.35	0.93	2.38	7.28
PDwR	Previous days without rain	day	1.06	2.39	0.00	19.79
Time	Time interval	h	44.70	40.70	0.32	264.64
Vol	Water volume	m^3	202,380.29	172,009.62	3404.22	990,263.20
TotalPrecip	Total precipitation	m^3	4.96	8.11	0.00	59.40
MaxPrecip	Maximum precipitation	mm	0.72	1.39	0.00	12.00
MedpH	Medium pH	-	7.10	0.47	5.59	8.05
MaxpH	Maximum pH	-	7.72	0.73	5.60	11.65
MedTemperature	Medium temperature	°C	17.24	3.39	9.11	22.85
MaxTemperature	Maximum temperature	°C	18.22	3.40	9.20	23.65
MedConductivity	Medium conductivity	µS/cm	880.99	282.38	147.19	1550.38
MaxConductivity	Maximum conductivity	µS/cm	1428.30	621.12	176.28	3350.60
MedAmmonium	Ammonium	mg/L	19.21	8.58	2.38	45.71
MaxAmmonium	Maximum ammonium	mg/L	24.56	10.99	2.97	57.99
MedFlow	Medium flow	m^3/h	5550.69	3107.41	2528.54	16,297.02
MaxFlow	Maximum flow	m^3/h	8602.23	3686.51	3224.42	17,885.11
MinFlow	Minimum flow	m^3/h	3634.63	3018.53	975.59	15,977.44

2.3. Methods

Since Jerome H. Friedman presented the non-parametric regression method known as MARS (Multivariate adaptive regression splines) it has been used in many applications in various fields [11]. Among its advantages are its ability to identify non-linear relationships in the data, to generate simple and more easily interpreted models from a large number of input variables, to show their relative importance, and to be computationally efficient compared to other techniques [12–14]. In the field of WWTPs, it has recently been used in different studies to predict the biochemical and chemical oxygen demand, the nitrogen, phosphorus and total suspended solids concentration [7,15], the activated sludge sedimentation capacity [15] or cost reduction [16].

The MARS model of a dependent variable, with n base functions or terms can be expressed as [11]:

$$\hat{y} = c_0 + \sum_{i=1}^{n} c_i \times B_i(\vec{x}) \tag{1}$$

where $\hat{y}$ is the dependent variable predicted by the MARS model, c_0 is a constant, B_i are the n basis functions and c_i are the coefficients of each one of them. The basis functions B_i look like $\max(0,\ x - t)$ or $\max(0,\ t - x)$ where t is a constant called node or cut-off point for the different variables. These space partition points, as well as the model parameters, are obtained from the analyzed data using a two-step forward/backward algorithm. First, using the forward stepwise algorithm, an over-fitted model with a large number of basis functions is generated and then, using the backward stepwise algorithm, the nodes that contribute the least to the overall fit are removed. The space is divided into regions; in each region a linear basis function is adjusted. The final model is a combination of all basis functions generated; whose number indicates the complexity of the model. Greater detail of the MARS method compared to other prediction techniques such as artificial neural networks (ANN) has been well described by Nalcaci et al. [13].

Through this process, the MARS method makes an automatic selection of input variables, that is, it includes the important variables in the model and excludes the non-relevant ones. However, it is necessary to previously consider the possible collinearity of the predictor variables. In this case, the correlation between the variables was studied and the variance inflation factor (VIF) was

determined. Consequently, those variables with a VIF greater than 10 were eliminated as possible predictor variables. Figure 2 includes the correlation values of all variables initially studied.

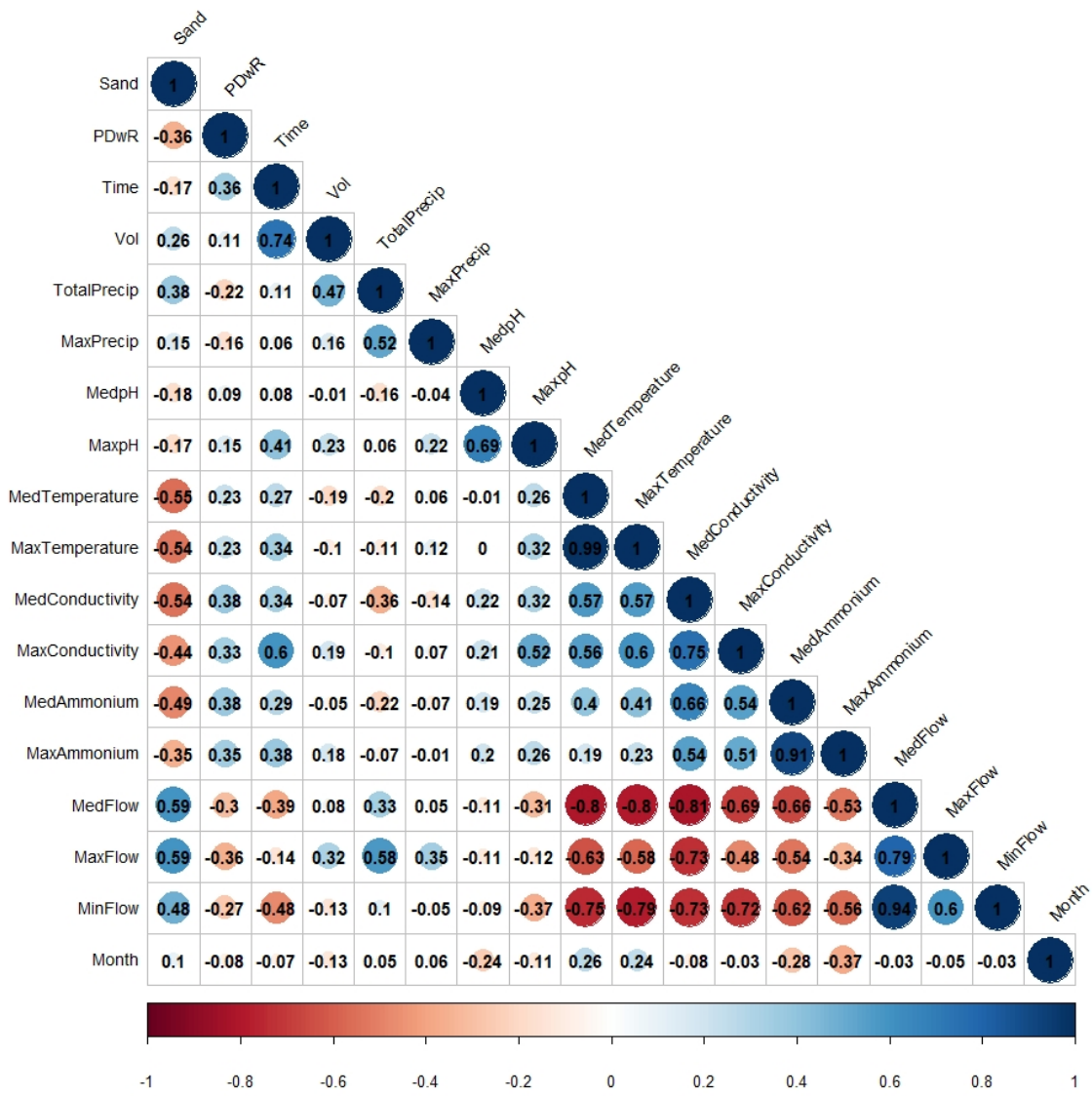


Figure 2. Correlation between variables.

3. Results

The assessment of the suitability of the prediction model, developed with the R earth package [17], was performed by using the determination coefficient (adjusted R^2) between the predicted values versus the actual dataset. In this case, although the accuracy of the MARS model obtained is not very high, $R^2 = 0.74$ (Figure 3) for training data and $R^2 = 0.70$ in validation data, it is enough for predicting trend changes in sand recovery during pre-treatment phases.

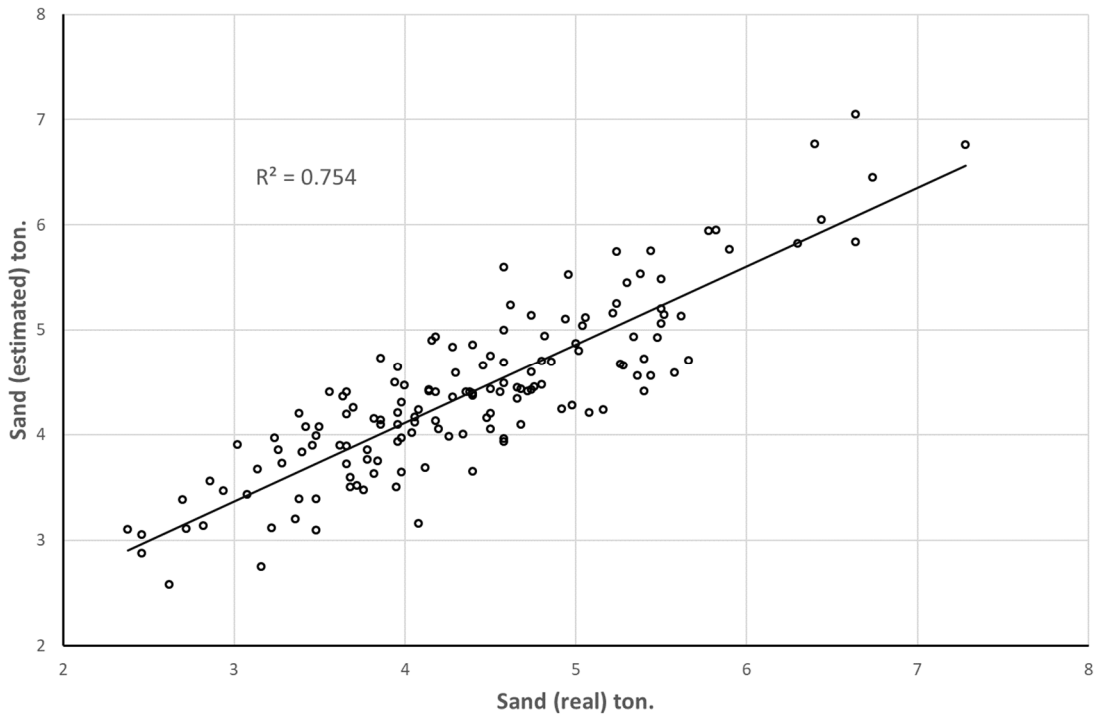


Figure 3. Prediction performance of the Multivariate Adaptive Regression Splines (MARS) model (training data).

Figure 4 shows the estimated and actual sand values over time, corresponding to the validation data set. It is possible to observe that the model is capable of detecting when changes in the amount of sand arrival to the treatment plant occur.

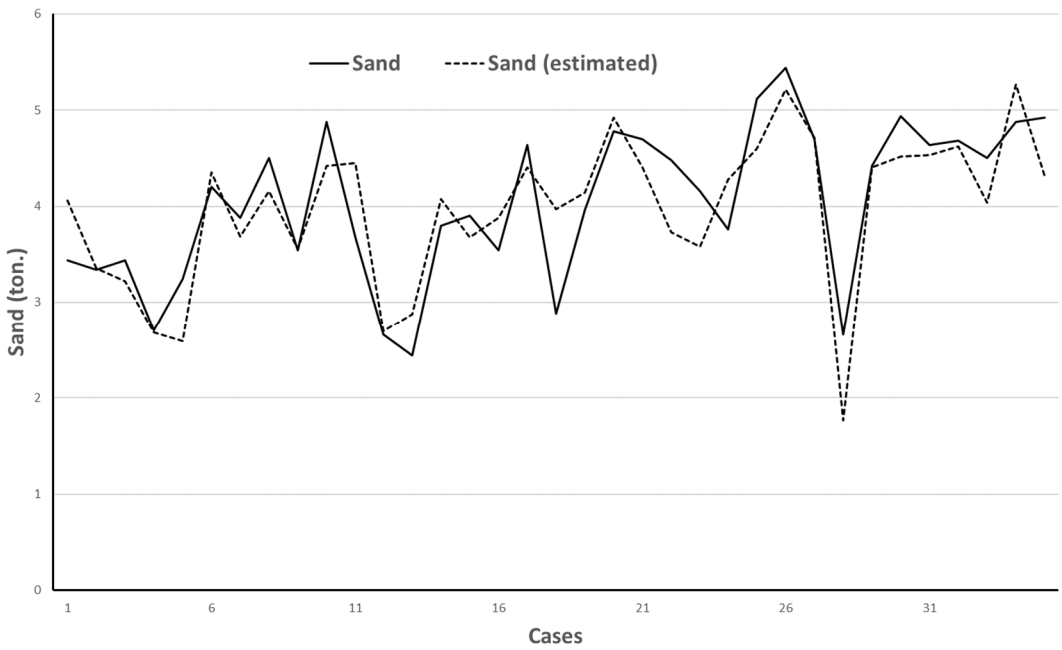


Figure 4. Prediction performance of MARS model (validation data).

The basis functions of the MARS model obtained, and their coefficients are shown in Table 3.

Table 3. List of basic functions of MARS model.

Basis Function	Definition	c_i
B_1	1	4.402074
B_2	max(0, MedFlow − 4784.547)	0.000155372
B_3	max(0, 1.4 − TotalPrecip) × max(0, 4784.547 − MedFlow)	−0.0004165734
B_4	max(0, MedpH − 6.841621) × max(0, 4784.547 − MedFlow)	−0.0008090167
B_5	max(0, MedpH − 7.193376) × max(0, Month − 4)	+0.9243512
B_6	max(0, Month − 4) × max(0, 18.58 − MedAmmonium)	+0.008425636
B_7	max(0, 24.32 − Time) × max(0, MedFlow − 4784.547) × Month	-1.45412×10^{-6}
B_8	max(0, 1.4 − TotalPrecip) × max(0, 4784.547 − MedFlow) × max(0, Month − 4)	$+5.940599 \times 10^{-5}$
B_9	max(0, MedpH − 7.193376) × max(0, Month − 4) × max(0, MedAmmonium − 12.56)	−0.2440937
B_{10}	max(0, MedpH − 7.193376) × max(0, Month − 4) × max(0,12.56 − MedAmmonium)	−0.2182406
B_{11}	max(0, MedpH − 7.193376) × max(0, Month − 4) × max(0, MedAmmonium − 16.22)	+0.2774549

Table 4 presents an assessment of the importance of each variable in the model according to the following criteria: the number of subsets of the model in which each variable is included (Nsubsets), GVC (Generalized cross-validation) and RSS (residual sums of squares). These last two parameters (GVC and RSS) are presented on a scale from 0 to 100. The GVC value shown in Table 4 can be understood as the model's ability to generalize and it is analyzed with the test data. Similarly, the RSS value represents the error that reduces a certain variable accumulated in each of the defined subsets. Therefore, the variables that cause a greater reduction of that error are considered more significant in the model.

Table 4. Importance assessment of variables in the model.

Variable	Nsubsets	GCV	RSS
MedFlow	10	100	100
MedpH	8	85.2	85.5
Time	8	55.8	61.3
Month	7	51.4	56.1
MedAmmonium	6	45.1	49.6
TotalPrecipitation	2	24.9	26.6

As shown in Table 4, the main variable that determines the amount of sand reaching the treatment plant is the flow (MedFlow). As expected, the greater the flow, the more sand is collected, although this relationship is not directly proportional to it. Sand content in wastewater is relatively low. Nevertheless, it is higher in rainwater, as may be indicated by the TotalPrecipitation variable, which also appears as significant. Experience indicates that, in periods of continuous rain, the arrival of sand is constant and there is no washing effect of the collector.

The pH, included in this case as an average value (MedpH), is another relevant variable since it indicates that there is a variation in the composition of the wastewater due to the presence of rainwater. While in dry weather the composition of the wastewater is fairly constant, when it rains, the pH value recorded is modified by the dilution effect.

The length of time intervals (Time) indicates the time it takes to fill the containers, while the variable corresponding to the month (Month) reflects the seasonality of sand production. The behaviour is different depending on the time of year, i.e., in rainy seasons or periods with a greater probability of rain, a greater quantity of sand is produced.

Like the pH, ammonium (MedAmmonium) is a variable conditioned by water dilution, but also by industrial activity. In other words, during working days, when there is greater industrial activity, the average value of ammonium increases. However, with the pH, this effect is not detected because certain industrial activities reduce the pH value while others increase it.

4. Discussion

An appropriate operation and control of wastewater treatment plants is receiving more and more attention due to the growing social concern for environmental issues [18]. As a result, different advanced plant control strategies are being implemented. Many of them are mainly based on monitoring different plant parameters [19]. Thus, variations in the composition and quantity of incoming wastewater, as well as the complexity of the treatment processes themselves, make it necessary to model different parameters to improve the operational control of the facility [4]. With this aim, and taking advantage of the data availability from the facilities' control systems (SCADA), different modelling techniques have been used, such as ANN (Artificial Neural Networks) [18], FIS (Fuzzy Inference System), ANFIS (Adaptive Neural Fuzzy Inference Systems), RF (Random Forest).

In this paper, the technique introduced is MARS methodology, which has been frequently used in other sectors, but which had not been used in the field of WWTPs parameter modelling. Achieving good results in all these modelling techniques, like MARS, depends fundamentally on the quantity and quality of the data used during training. In this work, great effort has been made in data collection and its initial preparation. Collecting a database corresponding to a full year of the variables indicated in Section 2 was the starting point. In addition to usual pre-processing of missing, wrong or incomplete data, in this case, due to the lack of continuous measurement of the sand received, it has been necessary to group the data of the remaining variables, according to data obtained from weighing sand containers when they are removed.

One of the positive aspects of using MARS is the greater ease in interpreting the resulting model compared to other techniques. For example, models based on ANN or RF cannot be easily interpreted [19]. This allows contrasting the results obtained with previous experience. In the presented model, as indicated in the Results section, the most significant variables correspond to the expected ones, considering the collected wastewater specific characteristics at the Villapérez station, with a unitary network that also receives a certain amount of industrial water.

The characteristics of wastewater entering the plant depend on the characteristics of the area it serves, population or industrial activities presence and they are affected by various factors, such as, seasonality or weather phenomena. Different works have studied cases in which plants operating conditions differ greatly from the designed working conditions, storms is a common example. On such occasions, as the facility performance gets worse, different treatment processes may be affected [20]. In the scope of this study, domestic wastewater does not contain sand, but it is associated with rainwater.

During these episodes of heavy rainfall and especially if it has not rained for some time, the arrival of sand increases notably, so having a model that detects these changes as soon as possible, makes it easier for the plant operators to take the most appropriate decisions. This is another of the outstanding aspects of this work, which focuses on predicting an input parameter, sand, in the initial treatment stage, which, as indicated above (Section 2), has an important influence on the performance and conservation of any treatment plant. In general, most studies focus on predicting other parameters such as COD, BOD, temperature, pH, in more advanced stages of the treatment process [18].

5. Conclusions

Although wastewater pre-treatment in WWTPs has been addressed by relatively little research, it is one of the stages most affected by untreated water discharges and on which the performance and durability of the rest of the plant also depends to a large extent. Not completing the pre-treatment process correctly i.e., not removing the larger waste, sand and fat, necessarily generates problems in the remaining treatments.

For the development of this model, data has been collected for approximately one year using the plant's SCADA system. This system registers 226 parameters every 9 min. In addition, climatic and sand data recovered in pretreatment have been collected in the same period. All this amount of data has been filtered and processed until obtaining 187 valid cases with the values corresponding to 17 variables.

In this work, the prediction of one of these components, sand, is addressed using a data-based model applying the MARS method. The accuracy achieved in the validation tests (R^2 = 0.70), which is similar to that obtained during training (R^2 = 0.74), provides a new tool for better management of WWTPs. Having an estimation of sand production will make it easier to open pretreatment lines based on the prediction of significant increases in sand production or, on the other hand, to close lines if a significant decrease is predicted. Similarly, regarding the filling of containers, warnings could be programmed into the SCADA plant control system, so that it is possible to predict when their removal will be necessary.

The MARS model obtained reflects the importance of certain variables and makes it possible to interpret, based on plant experience, that the variation in input values such as flow, pH, ammonia, etc., indicate changes due to rain or industrial activity, for example. Among all the variables introduced in the model, the mean flow (MedFlow) and the mean pH (MedpH), are the most outstanding variables according to the number of subsets, GVC and RSS.

Finally, it should be pointed out that it would also be possible to extend the prediction, with a modelling process similar to that developed in this work, to the other components of the pre-treatment stage as fats and gross solids, which would lead to operational improvements at the plant.

Author Contributions: Conceptualization, V.M.P. and F.O.F.; methodology, J.M.M.F.; data curation, V.M.P.; writing—original draft preparation, J.M.M.F. and V.M.P.; writing—review and editing, F.O.F. and H.M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science, Technology and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is part-funded by the European Regional Development Fund (ERDF).

Acknowledgments: The authors would like to thank Aguas de las Cuencas de España (ACUAES) and the joint venture formed by Dragados S.A. and Drace Infraestructuras S.A. for their collaboration in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Borges, N.B.; Campos, J.R.; Pablos, J.M. Characterization of residual sand removed from the grit chambers of a wastewater treatment plant and its use as fine aggregate in the preparation of non-structural concrete. *Water Pr. Technol.* **2015**, *10*, 164–171. [CrossRef]
2. He, L.; Tan, T.; Gao, Z.; Fan, L. The Shock Effect of Inorganic Suspended Solids in Surface Runoff on Wastewater Treatment Plant Performance. *Int. J. Environ. Res. Public Heal.* **2019**, *16*, 453. [CrossRef] [PubMed]
3. Do Prado, G.S.; Campos, J.R. O emprego da análise de imagem na determinação da distribuição de tamanho de partículas da areia presente no esgoto sanitário. *Eng. Sanit. e Ambient.* **2009**, *14*, 401–409. [CrossRef]
4. Yel, E.; Yalpir, S. Prediction of primary treatment effluent parameters by Fuzzy Inference System (FIS) approach. *Procedia Comput. Sci.* **2011**, *3*, 659–665. [CrossRef]
5. Zhang, J.; Du, C.; Feng, X. Research on a soft measurement model of sewage treatment based on a case-based reasoning approach. *Water Sci. Technol.* **2017**, *76*, 3181–3189. [CrossRef] [PubMed]
6. Civelekoglu, G.; Yigit, N.O.; Diamadopoulos, E.; Kitis, M. Modelling of COD removal in a biological wastewater treatment plant using adaptive neuro-fuzzy inference system and artificial neural network. *Water Sci. Technol.* **2009**, *60*, 1475–1487. [CrossRef] [PubMed]
7. Bakia, O.T.; Arasb, E.; Akdemirc, U.O.; Yilmaza, B. Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models. *Desalin. Water Treat.* **2019**, *157*, 79–89. [CrossRef]
8. Metcalf & Eddy, Inc.; Tchobanoglous, G.; Burton, F.; Stensel, H.D. *Wastewater Engineering: Treatment and Reuse*; McGraw-Hill Education: New York, NY, USA, 2002; ISBN 978-0-07-041878-3.
9. Office of Water. *Wastewater Technology Fact Sheet: Screening and Grit Removal*; Environmental Protection Agency (EPA): Washington, DC, USA, 2003; p. 11.
10. Do Prado, G.S.; Campos, J.R. Determinação da quantidade de areia no esgoto sanitário: Metodologia e estudo de caso. *Eng. Sanit. e Ambient.* **2008**, *13*, 306–312. [CrossRef]

11. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Statist.* **1991**, *19*, 1–67. [CrossRef]
12. Li, D.H.W.; Chen, W.; Li, S.; Lou, S. Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS)—A case study of Hong Kong. *Energy* **2019**, *186*, 115857. [CrossRef]
13. Nalcaci, G.; Özmen, A.; Weber, G.W. Long-term load forecasting: Models based on MARS, ANN and LR methods. *Central Eur. J. Oper. Res.* **2019**, *27*, 1033–1049. [CrossRef]
14. Zhang, X.; Fang, F.; Liu, J. Weather-Classification-MARS-Based Photovoltaic Power Forecasting for Energy Imbalance Market. *IEEE Trans. Ind. Electron.* **2019**, *66*, 8692–8702. [CrossRef]
15. Szeląg, B.; Bartkiewicz, L.; Studziński, J.; Barbusiński, K. Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear. *Arch. Environ. Prot.* **2017**, *43*, 74–81. [CrossRef]
16. Zadorojniy, A.; Wasserkrug, S.; Zeltyn, S.; Lipets, V. Unleashing Analytics to Reduce Costs and Improve Quality in Wastewater Treatment. *Inf. J. Appl. Anal.* **2019**, *49*, 262–268. [CrossRef]
17. R Core Team. *R: A language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
18. Hamed, M.M.; Khalafallah, M.G.; Hassanien, E.A. Prediction of wastewater treatment plant performance using artificial neural networks. *Environ. Model. Softw.* **2004**, *19*, 919–928. [CrossRef]
19. Dürrenmatt, D.J.; Gujer, W. Data-driven modeling approaches to support wastewater treatment plant operation. *Environ. Model. Softw.* **2012**, *30*, 47–56. [CrossRef]
20. El-Din, A.G.; Smith, D.W. A neural network model to predict the wastewater inflow incorporating rainfall events. *Water Res.* **2002**, *36*, 1115–1126. [CrossRef]

water

MDPI

Article

Gross Solids Content Prediction in Urban WWTPs Using SVM

Vanesa Mateo Pérez, José Manuel Mesa Fernández *, Francisco Ortega Fernández and Joaquín Villanueva Balsera

Project Engineering Area, University of Oviedo, Independencia, n° 13, 33012 Oviedo, Spain; mateovanesa@uniovi.es (V.M.P.); fdeasis@uniovi.es (F.O.F.); jmvillanueva@uniovi.es (J.V.B.)
* Correspondence: jmmesa@uniovi.es

Abstract: The preliminary treatment of wastewater at wastewater treatment plants (WWTPs) is of great importance for the performance and durability of these plants. One fraction that is removed at this initial stage is commonly called gross solids and can cause various operational, downstream performance, or maintenance problems. To avoid this, data from more than two operation years of the Villapérez Wastewater Treatment Plant, located in the northeast of the city of Oviedo (Asturias, Spain), were collected and used to develop a model that predicts the gross solids content that reaches the plant. The support vector machine (SVM) method was used for modelling. The achieved model precision (R^2_{adj} = 0.7 and MSE = 0.43) allows early detection of trend changes in the arrival of gross solids and will improve plant operations by avoiding blockages and overflows. The results obtained indicate that it is possible to predict trend changes in gross solids content as a function of the selected input variables. This will prevent the plant from suffering possible operational problems or discharges of untreated wastewater as actions could be taken, such as starting up more pretreatment lines or emptying the containers.

Keywords: wastewater; pretreatment; gross solids; SVM method

check for updates

Citation: Mateo Pérez, V.; Mesa Fernández, J.M.; Ortega Fernández, F.; Villanueva Balsera, J. Gross Solids Content Prediction in Urban WWTPs Using SVM. *Water* **2021**, *13*, 442. https://doi.org/10.3390/w13040442

Academic Editor: Andreas N. Angelakis
Received: 20 December 2020
Accepted: 5 February 2021
Published: 8 February 2021

1. Introduction

Municipal wastewater is derived from domestic, commercial, and industrial waste streams, along with storm water runoff. In addition to fecal matter, sewage contains a variety of suspended and floating debris, including sand and other entrained inert solids, paper, plastics, rags, and other debris. The presence of gross solids in the collectors can help to create several problems [1,2]. In the sections of the sewer network in which the water circulates by gravity, the solids combine with the fats and generate blockages. When water is circulated by pumping, the presence of gross solids can cause pump jams and pump well overflows with resulting contamination problems.

As wastewater enters a treatment facility, it typically flows through a step called preliminary treatment. This stage, which removes gross solids and coarse suspended and floating matter, has not received much research attention, and it is highly dependent on the initial design characteristics of the plant [3–5]. However, its impact on the management, operation, and maintenance of one of these wastewater treatment plants (WWTPs), as well as its influence on the performance of the subsequent treatment stages, is very important. In this pretreatment stage, various operations are carried out, such as roughing, sand removal, and degreasing. Generally, a screen removes large floating objects, such as rags, cans, bottles, and sticks, that may clog pumps, small pipes, and downstream processes. If gross solids are not removed, they become entrained in pipes and other moving parts of the treatment plant and can cause substantial damage and inefficiency in the process [6,7]. Screens are generally placed in a chamber or channel and inclined towards the flow of the wastewater. The inclined screen allows debris to be caught on the upstream surface of the screen, but it also allows access for manual or mechanical cleaning.

The operational management of this initial stage usually faces various problems, such as the following:

- Gross solids on days without rain are deposited in the bottom of the collectors, and when there is heavy rain, they are suddenly drawn into the treatment plant [8]. Numerous researchers have studied the consequences of these solids in sewage systems [9–14]. The arrival of all these gross solids at the WWTP can cause blockages in the equipment and, consequently, lead to discharge of untreated wastewater into rivers. Knowing of the arrival of solids as soon as possible would allow for anticipating and putting more pretreatment lines into service, avoiding those blockages.
- Another operational problem to be faced is the need to have enough containers for the gross solids and to avoid having to pile them on the ground in a precarious way. By predicting the arrival of gross solids earlier, it is possible to ensure the availability of empty containers.

The improvement of operations in treatment plants and its impact on their performance, the reduction of energy consumption, and the reduction of maintenance costs is receiving more and more attention from researchers [15–18]. The increasingly strict legal and environmental requirements force us to seek an improvement in the operation of these facilities [19,20]. An important way of optimizing this operation is the development of mathematical process models. Many authors have developed mathematical models of the different treatment stages of wastewater treatment plants [21,22]. Although the preliminary treatment stage has been less studied, in part due to its great dependence on the initial plant design, its impact on the performance of later stages is unquestionable.

Moreover, the treatment processes of sewage treatment plants are monitored continuously, but often the data collected are not sufficiently exploited [23]. Therefore, the use of the available data to improve management from the first treatment processes in the WWTP will result in an improvement in the performance of the later stages, a decrease in energy consumption, fewer installation maintenance problems, and, finally, in a better quality of the outlet water.

Therefore, the main objective of this work is to predict the gross solids content in wastewater to improve the operation of treatment plants. Having this new model will help the operators of the WWTPs make the most appropriate decisions, reducing the possibility of the problems described above. No reference to similar works (developing a prediction model for this operational parameter) was found in the literature review carried out by the authors, which indicates the novelty of this study.

This paper is divided into three main sections. Section 2 describes the characteristics of the WWTP under study, the acquisition and processing of data, and the mathematical techniques used in the development of the model. Next, in Section 3, the results obtained are presented and discussed, both in the model training process and in its validation. Finally, the main contributions of the study are highlighted in Section 4.

2. Materials and Methods

2.1. Case Study

The Villapérez Wastewater Treatment Plant is located in the northeast of the city of Oviedo (Asturias, Spain) and occupies an area of nearly 21 hectares (Figure 1). It provides service to an approximate population of 723,000 equivalent inhabitants. The wastewater to Villapérez arrives through a unitary network of collectors that has an approximate length of 75 km. This network includes 44 spillways. Collector diameters range from 600 to 2000 mm with sections in gravity and in impulsion.

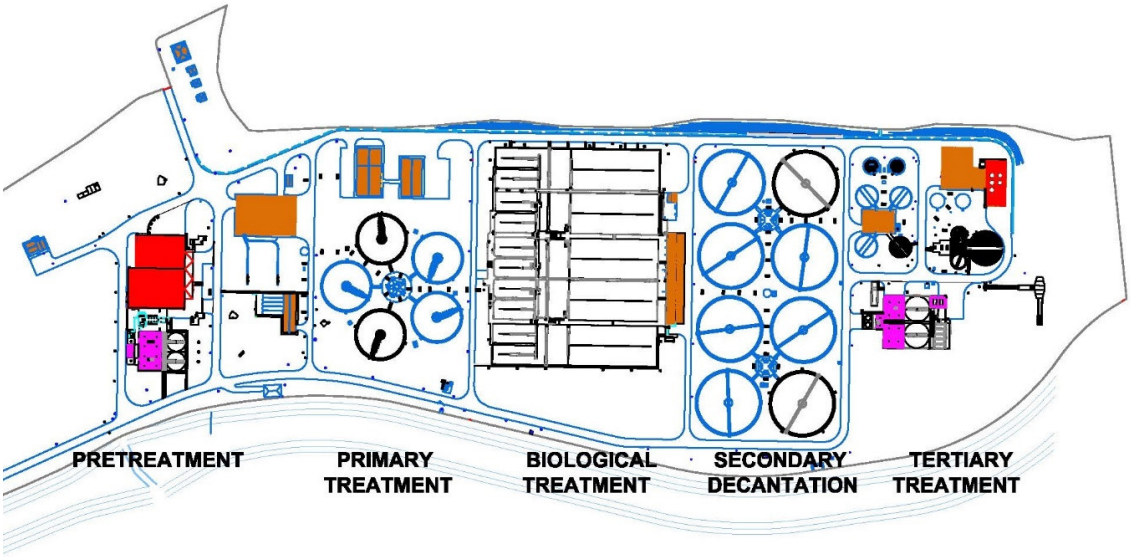


Figure 1. Plan view of the Villapérez wastewater treatment plant (WWTP) (Asturias, Spain).

As can be seen in Figure 2, the wastewater treatment in Villapérez WWTP begins with a pretreatment stage in which the larger solids, sands, and fats are removed. Subsequently, the water is taken to primary settling by gravity. The water then goes to biological treatment where organic matter, nitrogen, and phosphorus are removed. This treatment involves the passage of water through several anoxic chambers, anaerobic and aerobic. The next stage is secondary settling, which is carried out via gravity. Finally, the tertiary treatment stage consists of a physical–chemical treatment, lamellar settling, and filtration.

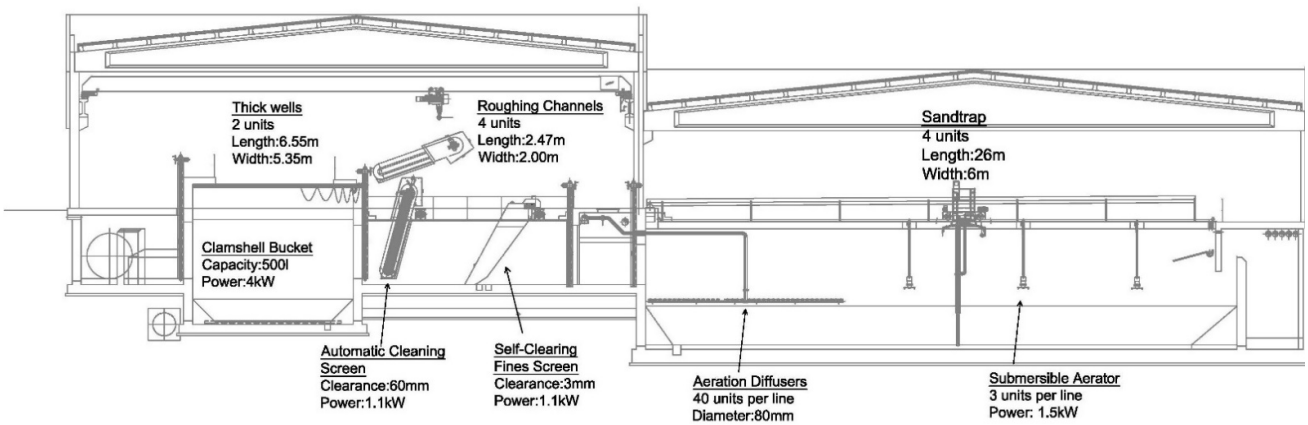


Figure 2. Sectional view of the pretreatment equipment at the Villapérez WWTP (Asturias, Spain).

The pretreatment section has the capacity to treat an inflow of 8.5 m^3/s (734,400 m^3/day) and starts with two thick wells, equipped with a 500-L clamshell bucket (Figure 2). The plant then has four roughing channels, each of which includes an automatic cleaning screen with a 60 mm clearance and a self-cleaning fines screen with a 3 mm clearance and an inclination of 50°.

In order to size the installation, Table 1 shows the main design parameters of the installation, including the legally established [24] values for the discharge of treated water.

Table 1. Design parameters of the Villapérez wastewater treatment plant (Asturias, Spain).

Parameter	Input	Output
Maximum inflow (rainy weather)	8.50 m^3/s	
Maximum inflow (dry weather)	2.89 m^3/s	
Five-day biological oxygen demand (BOD5)	418.00 mg/L	5 mg/L
Chemical oxygen demand (COD)	652.00 mg/L	30 mg/L
Total suspended solids (TSS)	329.00 mg/L	10 mg/L
Total Kjeldahl nitrogen (N-NTK)	47.40 mg/L	4 mg/L
Total phosphorus (Pt)	6.50 mg/L	0.5 mg/L

The Villapérez treatment plant receives around 19 tons of roughing solids monthly. As already indicated, although these roughing solids are produced continuously, they are stored at the bottom of the collectors and suddenly arrive at the treatment plant when heavy rains occur. In episodes of intense rains, the arrival of up to 4 tons of solids in one hour has been recorded.

Table 2 shows the composition of a few samples of solids collected in the bar and fine screens of the Villapérez plant. These samples represent the main materials included in the gross solids, such as wipes and hygiene products, plastics, and organic matter (Figure 3) from various sources.

Table 2. Composition of gross solids samples from pretreatment at Villapérez WWTP.

Sample	Total Wet Weight	Wipes		Plastics		Hygiene Products		Organic Material	
	kg	kg	%	kg	%	kg	%	kg	%
1	42.21	13.82	30.57	0.54	1.19	1.32	2.92	29.53	65.32
2	43.61	12.07	27.68	0.49	1.12	1.62	3.71	29.43	67.48
3	9.25	2.28	24.56	0.01	0.11	0.60	6.49	6.36	68.76

Figure 3. Gross solids sample removed in pretreatment at Villapérez WWTP.

2.2. Data

All data used in this work were collected in the period from 1 March 2017 to 24 June 2019 and come from different sources, as follows:

- Data related to wastewater were obtained through the SCADA software (Supervisory Control and Data Acquisition) of the WWTP. This system registers 226 parameters every 9 minutes from measuring equipment and sensors distributed all over the treatment plant. From this set of data, the data set associated to the measurement of input parameters in the raw water during the pretreatment stage was used. The parameters measured in the raw water are the input flow rate, pH, raw water temperature, conductivity, and ammonia. Data associated with these variables were identified by date and time of the data measurement.

- Gross solids data were collected from the container removal delivery notes (provided by the waste management entity), which contain the actual information of the waste total weight inside each container. The number of containers in the study period was 165. Their filling times were used as time intervals to group the data from the SCADA system.
- Climate data were obtained from the Spanish State Agency for Meteorology website (Agencia Estatal de Meteorología, Aemet) and pluviometry data (instantaneous and accumulated rainfall) were obtained from the plant's own weather station. All of them were also grouped according to the intervals in which the containers were filled. From these data, a new variable calculated from the instantaneous precipitation was also created, corresponding to the number of previous days without rain.

The obtained data set (165 cases) was divided into two groups. Eighty percent of the data were used for training the support vector machine (SVM) model, and the remaining 20% were kept for validating the model.

Statistical data for the variables initially considered in the study are presented in Table 3. As indicated above, the reference is the time interval (Time) from when an empty container was placed to when it was removed. When each container was removed, it was weighed, and the data were recorded on the corresponding delivery note. The data corresponding to each one of these periods were summarized by calculating for each variable its minimum, mean, and maximum values, as shown in Table 3.

Table 3. Statistical description of the variables.

Variable	Description	Unit	Mean	Standard Deviation	Min	Max
GrossSolids	Gross solids	ton	2.96	0.79	1.42	5.44
Interval	Time interval	h	123.40	468.56	1.28	6023.36
PDwR	Previous days without rain	day	1.15	2.69	0.00	20.41
MxDwR	Maximum previous days without rain in the time interval	day	2.86	3.48	0.01	20.68
Vol	Water volume	m^3	398,312.01	407,147.11	4254.39	2,600,377.05
PrecipTotal	Total precipitation	m^3	7.57	12.37	0.00	86.50
MaxpH	Maximum pH		7.99	0.66	6.44	11.65
MedConductivity	Medium conductivity	µS/cm	926.74	240.99	256.57	1578.82
MedFlow	Medium flow	m^3/h	4853.96	2404.04	2382.21	14,195.72
Month	Month		5.51	3.19	1.00	12.00
Week	Week		22.21	14.04	1.00	52.00
TempExtMed	Medium ambient temperature	°C	11.80	4.53	3.10	24.60
TempExtMax	Maximum ambient temperature	°C	16.05	5.42	4.20	31.50
TempExtMin	Minimum ambient temperature	°C	8.49	4.39	0.70	19.20
DayYear	Day of the year		151.98	98.53	2.00	363.00
DayWeek	Day of the week		3.18	1.71	1.00	6.00
MedRH	Medium relative humidity	%	78.91	9.07	46.17	96.81
MaxSolarRadiation	Maximum solar radiation	W/m^2	44.89	79.48	0.77	532.98
AtmosphericPressureMax	Maximum atmospheric pressure	millibars	1004.52	7.85	972.41	1021.96
MaxMedRH	Maximum relative humidity	%	94.86	9.03	49.99	99.92
MinMedRH	Minimum relative humidity	%	46.84	18.41	0.00	92.14

Different statistical analyses were performed to explore the initial data set in order to identify the existence of outliers, as well as to confirm the quality of the data. Among them, we can highlight the principal component analysis (PCA) projection shown in Figure 4. The data were projected in the two main dimensions, which are those that best represent the initial data set in terms of minimum squares. In this figure, on the left, each case of the study is represented with a different color depending on the month of the year in which the sample was taken. In addition, the graph on the right shows the same PCA projection but with the cases separated by month and the average flow (MedFlow) represented with a color scale. These monthly projections clearly reflect that the months with usually higher rainfall present higher inflow into the WWTP, which is a sign of the quality of the training patterns. On the other hand, it is possible to observe in Figure 4 that the cases that are isolated in the complete PCA projection (on the left in the figure), which could initially be considered outliers, correspond to a continuous trend in the cases of Month 12 (December).

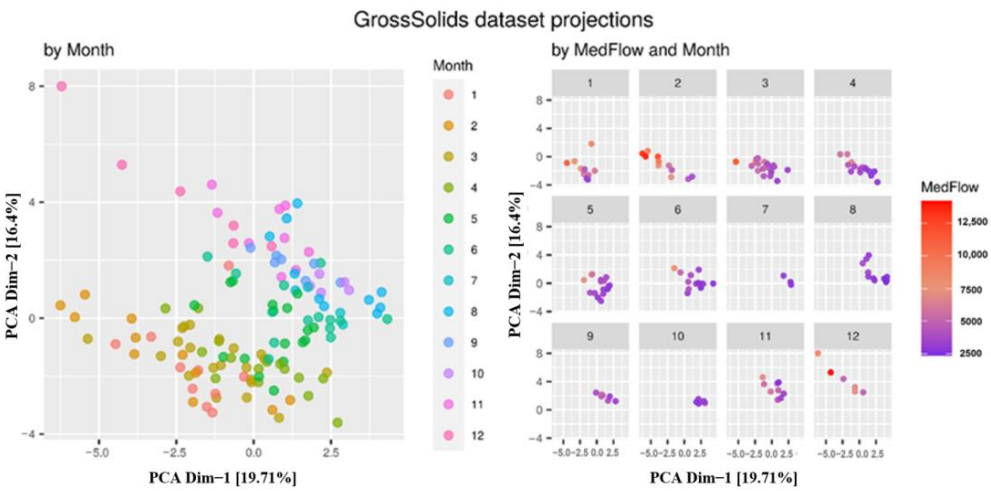


Figure 4. Principal component analysis (PCA) projection of the initial data set.

2.3. Methods

Different data-based techniques have been used to model different WWTP parameters, such as artificial neural networks (ANNs), fuzzy inference systems (FISs), adaptive neural fuzzy inference systems (ANFISs), and random forest (RF) [15]. In this paper, the method used was support vector machine (SVM), which has been successfully used in many different fields.

SVM refers to a set of supervised learning algorithms developed by Vladimir Vapnik and his team at AT&T laboratories [25]. Although initially developed as a method for binary classification, its application has been extended to multiple classification and regression problems. SVM has been successfully used in many different fields, such as computer vision, character recognition, text and hypertext categorization, classification, natural language processing, and time series analysis [26–28]. This is because this method has shown good generalization ability, avoiding the problems of training overfitting that occur in other similar methods [29]. Recently, it has also been used in the field of wastewater treatment to predict different parameters of the treatment process [30–37].

The core of this method is a kernel-based algorithm. Its predictions for new inputs depend on the kernel function evaluation for a subcategory of occurrences during a training stage. The objective of this method is to find a function to minimize the final error in Equation (1):

$$y(x) = w^T \cdot \phi(x) + b \quad (1)$$

where $y(x)$ is the predicted value, w is the vector of parameters that define the model, b is the value of the bias, and $\phi(x)$ fixes the feature space transformation. In this method, the error function that appears in the simple linear regression (Equation (2)) is replaced by

an ϵ-insensitive error function (Equation (3)). The latter assigns a zero to values when ϵ exceeds the difference between the target (t_n) and the predicted value (y_n). If the difference is not less than ϵ, the error function maintains its value.

$$\frac{1}{2}\sum_{n=1}^{N}[y_n - t_n]^2 + \frac{\lambda}{2}\|w\|^2 \tag{2}$$

$$E_\epsilon(y(x) - t) = \begin{cases} 0, & if\ |y(x) - t| < \epsilon \\ |y(x) - t| - \epsilon, & otherwise \end{cases} \tag{3}$$

$$C\sum_{n=1}^{N} E_\epsilon(y(x_n) - t_n) + \frac{1}{2}\|w\|^2 \tag{4}$$

To minimize Equation (4), a cost (C) is also assigned to the difference between the target and predicted values, where $y(x)$ is the value that Equation (2) predicts, t is the searched target function, ϵ is the margin where the function does not penalize, and C is the penalty. The process is optimized, but the initial function (Equation (2)) increases in complexity (Equation (5)), where α is one solution for the optimization problem that Lagrangian Theory makes possible.

$$y(x) = \sum_{n=1}^{N}(\alpha_i - \alpha_i{}^*)\langle x_i \cdot x\rangle + b \tag{5}$$

The data are transformed by the function to a higher-dimensional feature space. This increases the accuracy of the nonlinear problem. Thus, the final function resembles Equation (6).

$$y(x) = \sum_{n=1}^{N}(\alpha_i - \alpha_i{}^*)k(x_i, x) + b \tag{6}$$

Likewise, as in many other data-based modeling techniques, the quantity and quality of data greatly affect the results obtained. In this case, it is necessary to take into account that the quality of the data collected in these facilities usually presents various reliability problems due to the difficult environmental working conditions of the sensors, which implies a high variation and even errors in the measurements obtained [38]. Therefore, considerable effort was put into collecting data over more than two years from various sources. In this way, the data include information of a seasonal nature, changes in domestic or industrial activity, long periods of intense rains or dry weather, etc. Thus, they are representative of the normal operating conditions at the installation. Subsequently, these data were carefully processed to avoid missing, wrong, or incomplete data to obtain 165 verified patterns to train the model (80%) and to validate the results (20%).

The kernel choice and the particular selection of adjustable kernel parameters have an important influence on the performance of the model [39]. This work was developed by trying various commonly used types of kernel functions, such as linear, polynomial, sigmoid, and radial basis functions [40]. The best kernel for classification in general is the Gaussian radial basis function (RBF) because it produces the highest overall accuracy and highest overall kappa [41].

A grid search methodology with 10-fold cross-validation on the training set was applied to establish the best type of kernel function and to retrieve the optimal values for the model parameters. This k-fold cross-validation procedure is an extensively used approach for assessing the values of model architecture parameters [42,43]. After this process, the RBF was the kernel with the best results (Equation (7)):

$$k(x_i, x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} \tag{7}$$

where σ is a free parameter and $||x_1 - x_2||$ is the Euclidean distance between points x_1 and x_2.

R statistical software was selected to program the proposed methodology [44].

3. Results and Discussion

As a result of the training process, an SVM model was obtained that predicts the gross solids in tons based on the variables listed in Table 3.

Figure 5 presents different analyses carried out to validate the results of the training process of the SVM model obtained. At the top of the figure, the temporal evolution of the actual values is compared with that predicted from the training data set. It is possible to observe that the model can detect when changes occur in the content of gross solids arriving at the treatment plant.

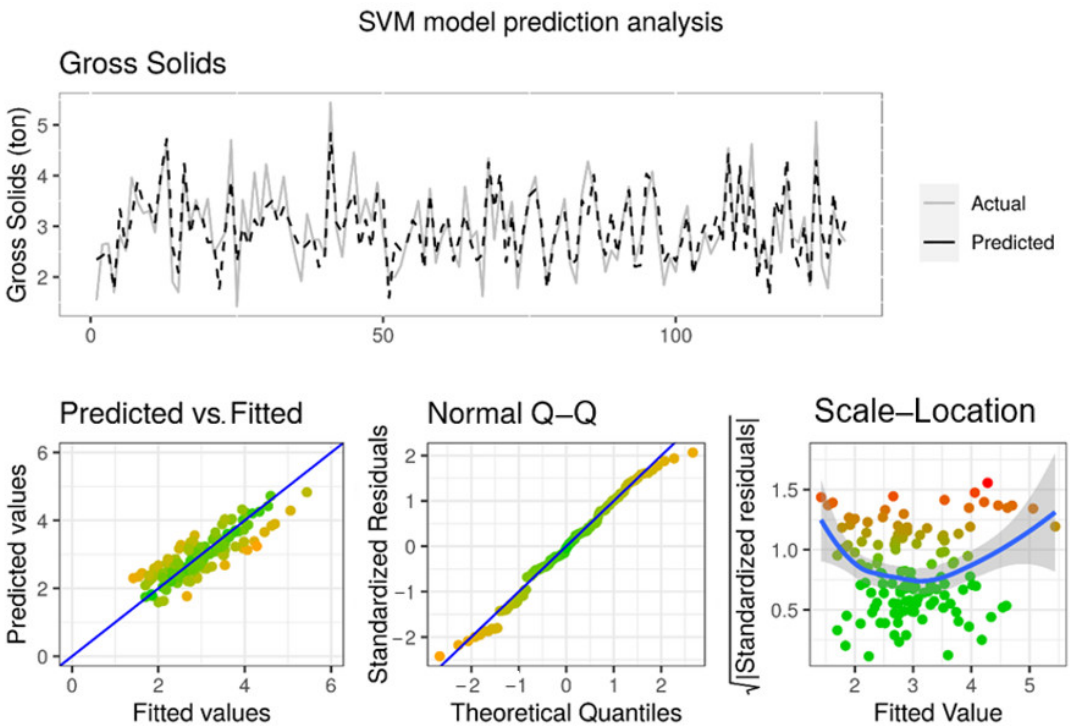


Figure 5. Analysis of results in the training phase.

At the bottom of Figure 5, several graphs are included to represent the error made by the SVM model. The "Prediction vs. Fitted" graph contrasts the actual measured values against the values predicted by the SVM model. It can be seen that all the estimated cases are around the blue line that represents the theoretical behavior of perfect prediction. In the "Normal Q–Q" graph it can be seen that the standardized errors generated by the SVM model in its estimation have a behavior almost identical to the expected theoretical behavior. A greater deviation can be seen at the ends of the line, which is confirmed in the "Scale–Location" graph that shows the estimation error made in each case. In this last graph, it can be seen that those gross solids values lower than 2 tons or higher than 4 tons show an increase in the standardized residuals.

In Figure 6, the curve of the cumulative percentage of successes by the SVM model is represented in blue with increasing tolerance of the estimation error (residuals). The control curve (in red in Figure 6) represents the cumulative success rate achieved by the sewage plant operators, estimated from the mean value of the historical data. A significant improvement can be observed in the results of the SVM model compared to the estimation of the plant control.

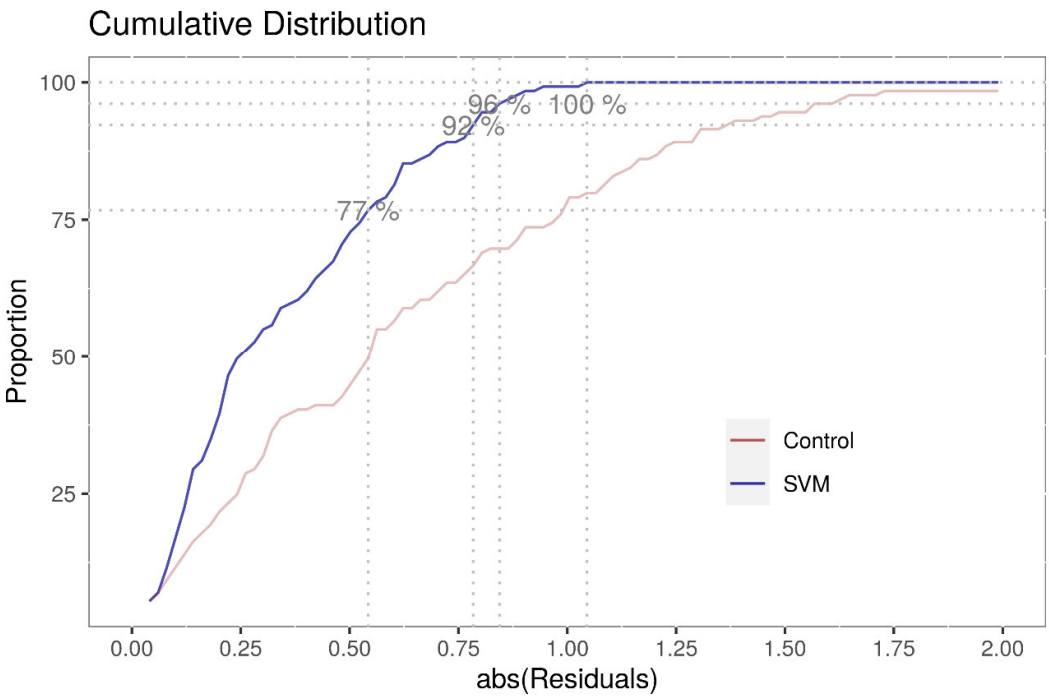


Figure 6. Cumulative distribution of the percentage of success of the model versus allowed error.

Figure 7 shows the estimated and actual gross solids values over time, corresponding to the validation data set. It is possible to observe that the model can detect when changes occur in the content of gross solids arriving at the treatment plant.

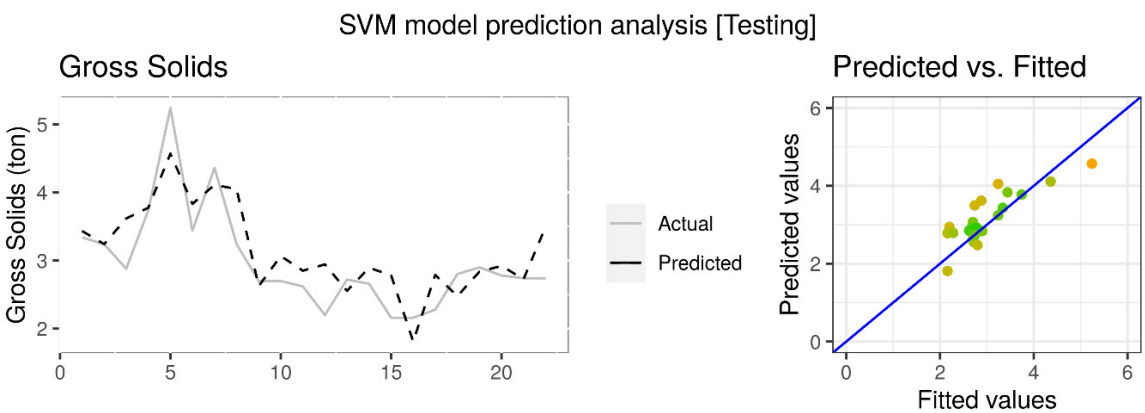


Figure 7. Analysis of results in the testing phase.

The coefficient of determination is a statistical indicator that compares the accuracy of the model to the accuracy of a trivial benchmark model wherein the prediction is just the mean of all the samples [45]. The performance of the SVM model was measured using the adjusted coefficient of determination (R^2_{adj}) an adjustment for the coefficient of determination that takes into account the number of variables in a data set [46]. It also penalizes you for points that do not fit the model.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right] \tag{8}$$

Here, n is the number of points in the data sample, k is the number of variables in the model, and R^2 is the coefficient of determination.

In this case, although the accuracy of the SVM model obtained was not very high, R^2_{adj} = 0.7093 for training data and R^2_{adj} = 0.6869 for validation data, it is enough for predicting trend changes in gross solids recovery during the pretreatment phases. The final model presented mean squared error (MSE) values of 0.426 in training and 0.435 in validation testing. With these results, the resulting final model will provide relevant information to the operators of the WWTP, anticipating problems such as blockages in the equipment or untreated wastewater discharges into the river.

Table 4 includes the most relevant variables for the SVM model when predicting the arrival of gross solids at the WWTP. The two first ones, the week and day of the year, are related to the seasonal component of this variable. An increase in the amount of rain supposes a greater drag on the solids deposited in the collectors, while the pH is an indicator of the amount of flow that reaches the treatment plant from industrial activities. The pH of water from domestic activities is relatively constant, while that from industrial activities alters it, sometimes raising it and sometimes lowering it. One of the consequences is the so-called "weekend effect". Since the Villapérez WWTP receives a significant portion of wastewater from industrial facilities, the activity of which decreases on weekends and holidays, the resulting reduction in flow modifies the pH; therefore, it is relevant to the SVM model.

Table 4. Importance of variables in the model.

Overall	%
Week	100
DayYear	98.87
PrecipTotal	93.84
MaxpH	79.2
MinMedRH	76.58
MedRH	63.56
TempExtMed	60.08
PDwR	59.79
MedFlow	54.75

The three parameters MinMedRH, MedRH, and TempExtMed characterize the weather, i.e., if a certain day is clear or rainy. Another significant parameter is the number of previous days without rain. Gross solids should accumulate at the bottom of the collectors on days without rain; therefore, this should be a very relevant variable. However, its influence on the estimation of the model is less than expected, perhaps because the time periods are relatively long (PDwR mean = 123.4 h), and a downpour may occur within that period that is not detected.

4. Conclusions

Gross solids (wipes, sanitary waste, swabs, etc.) dragged by rain into sanitation systems generate numerous problems both in the collectors and in the treatment plants, causing severe blockages as described in multiple references. Reducing those blockages in pretreatment equipment and avoiding the discharge of untreated water due to possible overflows was the main objective of this work. It should be noted that in studies prior to this work, no other scientific reference predicting a similar parameter was found to compare the results to, which reflects the novelty of this work.

An SVM model was developed for predicting the content of gross solids present in roughing wastewater. The SVM method has demonstrated good features in numerous previous works, and in this case, the precision achieved in the validation phase was R^2_{adj} = 0.6869, slightly lower than that achieved in training (R^2_{adj} = 0.7093); this is considered enough to detect change trends in the arrival of roughing solids at the treatment plant. Having this information in advance will make it possible to open pretreatment lines when necessary to receive the arrival of a greater quantity of gross solids and to

have enough containers for their storage. This good performance of the model was also endorsed in the comparison of the precision of the model with that of the current estimates based on historical average values. The model was observed to represent a considerable operational improvement.

The final model presented MSE values of 0.426 in training and 0.435 in validation testing. The largest errors in the model occurred at the extremes, that is, for below 2 tons and above 4 tons of gross solids; these are unusual values, since containers of less than 2 tons mean that they have left the installation without being completely full, and for those above 4 tons, the container runs the risk of overflowing. Therefore, they do not represent a major drawback, and the biggest errors of the model are due to the low presence of such patterns in the training data set.

Finally, it should be noted that, following a similar line of work, it would be convenient to estimate other operating parameters of the pretreatment stage; this would facilitate its operation, which would have an impact on the performance of the entire WWTP and, therefore, on the quality of the outgoing treated water.

Author Contributions: Conceptualization, V.M.P. and F.O.F.; methodology, J.M.M.F.; data curation, V.M.P.; writing—original draft preparation, J.M.M.F. and V.M.P.; writing—review and editing, F.O.F. and J.V.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science, Technology and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is partly funded by the European Regional Development Fund (ERDF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Aguas de las Cuencas de España (ACUAES) and the joint venture formed by Dragados S.A. and Drace Infraestructuras S.A. for their collaboration in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Collin, T.D.; Cunningham, R.; Asghar, M.Q.; Villa, R.; MacAdam, J.; Jefferson, B. Assessing the Potential of Enhanced Primary Clarification to Manage Fats, Oils and Grease (FOG) at Wastewater Treatment Works. *Sci. Total Environ.* **2020**, *728*, 138415. [CrossRef]
2. Roychand, R.; Li, J.; De Silva, S.; Saberian, M.; Law, D.; Pramanik, B.K. Development of Zero Cement Composite for the Protection of Concrete Sewage Pipes from Corrosion and Fatbergs. *Resour. Conserv. Recycl.* **2021**, *164*, 105166. [CrossRef]
3. Prado, G.S.D.; Campos, J.R. O emprego da análise de imagem na determinação da distribuição de tamanho de partículas da areia presente no esgoto sanitário. *Eng. Sanit. Ambient.* **2009**, *14*, 401–409. [CrossRef]
4. He, L.; Tan, T.; Gao, Z.; Fan, L. The Shock Effect of Inorganic Suspended Solids in Surface Runoff on Wastewater Treatment Plant Performance. *Int. J. Environ. Res. Public Health* **2019**, *16*, 453. [CrossRef]
5. Sidwick, J.M. The Preliminary Treatment of Wastewater. *J. Chem. Technol. Biotechnol.* **1991**, *52*, 291–300. [CrossRef]
6. Metcalf & Eddy, Inc.; Tchobanoglous, G.; Burton, F.; Stensel, H.D. *Wastewater Engineering: Treatment and Reuse*; McGraw-Hill Education: New York, NY, USA, 2002; ISBN 978-0-07-041878-3.
7. Office of Wastewater Management; United States Environmental Protection Agency (EPA). *Primer for Municipal Wastewater Treatment Systems*; Office of Wastewater Management: Washington, DC, USA, 2004.
8. Ashley, R.M.; Bertrand-Krajewski, J.-L.; Hvitved-Jacobsen, T.; Verbanck, M. *Solids in Sewers*; IWA Publishing: London, UK, 2004; ISBN 978-1-900222-91-4.
9. Brown, D.M.; Butler, D.; Orman, N.R.; Davies, J.W. Gross Solids Transport in Small Diameter Sewers. *Water Sci. Technol.* **1996**, *33*, 25–30. [CrossRef]
10. Eren, B.; Karadagli, F. Physical Disintegration of Toilet Papers in Wastewater Systems: Experimental Analysis and Mathematical Modeling. *Environ. Sci. Technol.* **2012**, *46*, 2870–2876. [CrossRef] [PubMed]
11. Butler, D.; Littlewood, K.; Orman, N. A Model for the Movement of Large Solids in Small Sewers. *Water Sci. Technol.* **2005**, *52*, 69–76. [CrossRef]

12. Digman, C.J.; Littlewood, K.; Butler, D.; Spence, K.; Balmforth, D.J.; Davies, J.; Schütze, M. A Model to Predict the Temporal Distribution of Gross Solids Loading in Combined Sewerage Systems. *Glob. Solut. Urban Drain.* **2012**, 1–13. [CrossRef]
13. Walski, T.; Edwards, B.; Heifer, E.; Whitman, B.E. Transport of Large Solids in Sewer Pipes. *Water Environ. Res.* **2009**, *81*, 709–714. [CrossRef] [PubMed]
14. Walski, T.; Falco, J.; McAloon, M.; Whitman, B. Transport of Large Solids in Unsteady Flow in Sewers. *Urban Water J.* **2011**, *8*, 179–187. [CrossRef]
15. Hamed, M.M.; Khalafallah, M.G.; Hassanien, E.A. Prediction of Wastewater Treatment Plant Performance Using Artificial Neural Networks. *Environ. Model. Softw.* **2004**, *19*, 919–928. [CrossRef]
16. Hernández-Chover, V.; Castellet-Viciano, L.; Hernández-Sancho, F. Preventive Maintenance versus Cost of Repairs in Asset Management: An Efficiency Analysis in Wastewater Treatment Plants. *Process Saf. Environ. Prot.* **2020**, *141*, 215–221. [CrossRef]
17. Hernández-Chover, V.; Bellver-Domingo, Á.; Hernández-Sancho, F. The Influence of Oversizing on Maintenance Cost in Wastewater Treatment Plants. *Process Saf. Environ. Prot.* **2021**, *147*, 734–741. [CrossRef]
18. Heo, S.; Nam, K.; Tariq, S.; Lim, J.Y.; Park, J.; Yoo, C. A Hybrid Machine Learning–Based Multi-Objective Supervisory Control Strategy of a Full-Scale Wastewater Treatment for Cost-Effective and Sustainable Operation under Varying Influent Conditions. *J. Clean. Prod.* **2021**, *291*, 125853. [CrossRef]
19. Ortiz-Martínez, V.M.; Martínez-Frutos, J.; Hontoria, E.; Hernández-Fernández, F.J.; Egea, J.A. Multiplicity of Solutions in Model-Based Multiobjective Optimization of Wastewater Treatment Plants. *Optim. Eng.* **2020**, 1–16. [CrossRef]
20. Pang, J.; Yang, S.; He, L.; Chen, Y.; Ren, N. Intelligent Control/Operational Strategies in WWTPs through an Integrated Q-Learning Algorithm with ASM2d-Guided Reward. *Water* **2019**, *11*, 927. [CrossRef]
21. Benedetti, L.; Langeveld, J.; Comeau, A.; Corominas, L.; Daigger, G.; Martin, C.; Mikkelsen, P.S.; Vezzaro, L.; Weijers, S.; Vanrolleghem, P.A. Modelling and Monitoring of Integrated Urban Wastewater Systems: Review on Status and Perspectives. *Water Sci. Technol.* **2013**, *68*, 1203–1215. [CrossRef]
22. Hreiz, R.; Latifi, M.A.; Roche, N. Optimal Design and Operation of Activated Sludge Processes: State-of-the-Art. *Chem. Eng. J.* **2015**, *281*, 900–920. [CrossRef]
23. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. [CrossRef]
24. The Council of The European Communities. *Council Directive 91/271/EEC of 21 May 1991 Concerning Urban Waste-Water Treatment*; The Council of the European Communities: Brussels, Belgium, 2014.
25. Vapnik, V. The Support Vector Method of Function Estimation. In *Nonlinear Modeling: Advanced Black-Box Techniques*; Suykens, J.A.K., Vandewalle, J., Eds.; Springer: Boston, MA, USA, 1998; pp. 55–85, ISBN 978-1-4615-5703-6.
26. Bishop, C. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2006; ISBN 978-0-387-31073-2.
27. Clarke, S.M.; Griebsch, J.H.; Simpson, T.W. Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses. *J. Mech. Des.* **2004**, *127*, 1077–1087. [CrossRef]
28. Chauhan, V.K.; Dahiya, K.; Sharma, A. Problem Formulations and Solvers in Linear SVM: A Review. *Artif. Intell. Rev.* **2019**, *52*, 803–855. [CrossRef]
29. Liu, Z.; Xu, H. Kernel Parameter Selection for Support Vector Machine Classification. *J. Algorithms Comput. Technol.* **2014**, *8*, 163–177. [CrossRef]
30. Cheng, T.; Dairi, A.; Harrou, F.; Sun, Y.; Leiknes, T. Monitoring Influent Conditions of Wastewater Treatment Plants by Nonlinear Data-Based Techniques. *IEEE Access* **2019**, *7*, 108827–108837. [CrossRef]
31. Yang, Y.H.; Guergachi, A.; Khan, G. Support Vector Machines for Environmental Informatics: Application to Modelling the Nitrogen Removal Processes in Wastewater Treatment Systems. *J. Environ. Inform.* **2015**, *7*, 14–23. [CrossRef]
32. Mahmoodi, N.M.; Abdi, J.; Taghizadeh, M.; Taghizadeh, A.; Hayati, B.; Shekarchi, A.A.; Vossoughi, M. Activated Carbon/Metal-Organic Framework Nanocomposite: Preparation and Photocatalytic Dye Degradation Mathematical Modeling from Wastewater by Least Squares Support Vector Machine. *J. Environ. Manag.* **2019**, *233*, 660–672. [CrossRef]
33. Abobakr Yahya, A.S.; Ahmed, A.N.; Binti Othman, F.; Ibrahim, R.K.; Afan, H.A.; El-Shafie, A.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios. *Water* **2019**, *11*, 1231. [CrossRef]
34. Najafzadeh, M.; Zeinolabedini, M. Prognostication of Waste Water Treatment Plant Performance Using Efficient Soft Computing Models: An Environmental Evaluation. *Measurement* **2019**, *138*, 690–701. [CrossRef]
35. Negara, M.P.; Cornelissen, E.; Geurkink, A.K.; Euverink, G.J.W.; Jayawardhana, B. Next Generation Sequencing Analysis of Wastewater Treatment Plant Process via Support Vector Regression. *IFAC-PapersOnLine* **2019**, *52*, 37–42. [CrossRef]
36. Cheng, H.; Liu, Y.; Huang, D.; Liu, B. Optimized Forecast Components-SVM-Based Fault Diagnosis With Applications for Wastewater Treatment. *IEEE Access* **2019**, *7*, 128534–128543. [CrossRef]
37. Harrou, F.; Dairi, A.; Sun, Y.; Senouci, M. Statistical Monitoring of a Wastewater Treatment Plant: A Case Study. *J. Environ. Manag.* **2018**, *223*, 807–814. [CrossRef] [PubMed]
38. Jover-Smet, M.; Martín-Pascual, J.; Trapote, A. Model of Suspended Solids Removal in the Primary Sedimentation Tanks for the Treatment of Urban Wastewater. *Water* **2017**, *9*, 448. [CrossRef]

39. Hsu, C.; Chang, C.; Lin, C. A Practical Guide to Support Vector Classification. 2010. Available online: www.csie.ntu.edu.tw/$\sim$cjlin/papers/guide/guide.pdf (accessed on 5 February 2021).
40. Campbell, C.; Ying, Y. Learning with Support Vector Machines. *Synth. Lect. Artif. Intell. Mach. Learn.* **2011**, *5*, 1–95. [CrossRef]
41. Kranjčić, N.; Medak, D.; Župan, R.; Rezo, M. Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns. *Remote Sens.* **2019**, *11*, 655. [CrossRef]
42. Duan, K.; Keerthi, S.S.; Poo, A.N. Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. *Neurocomputing* **2003**, *51*, 41–59. [CrossRef]
43. Budiman, F. SVM-RBF Parameters Testing Optimization Using Cross Validation and Grid Search to Improve Multiclass Classification. *Sci. Vis.* **2019**, *11*, 11. [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
45. El-Din, A.G.; Smith, D.W. A Neural Network Model to Predict the Wastewater Inflow Incorporating Rainfall Events. *Water Res.* **2002**, *36*, 1115–1126. [CrossRef]
46. Saunders, L.J.; Russell, R.A.; Crabb, D.P. The Coefficient of Determination: What Determines a Useful R2 Statistic? *Investig. Ophthalmol. Vis. Sci.* **2012**, *53*, 6830–6832. [CrossRef]

water

MDPI

Article

A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs

Vanesa Mateo Pérez, José Manuel Mesa Fernández *, Joaquín Villanueva Balsera and Cristina Alonso Álvarez

Project Engineering Area, University of Oviedo, 33012 Oviedo, Spain; mateovanesa@uniovi.es (V.M.P.); jmvillanueva@uniovi.es (J.V.B.); alonsocristina@uniovi.es (C.A.Á.)
* Correspondence: jmmesa@uniovi.es

Abstract: The content of fats, oils, and greases (FOG) in wastewater, as a result of food preparation, both in homes and in different commercial and industrial activities, is a growing problem. In addition to the blockages generated in the sanitary networks, it also represents a difficulty for the performance of wastewater treatment plants (WWTP), increasing energy and maintenance costs and worsening the performance of downstream treatment processes. The pretreatment stage of these facilities is responsible for removing most of the FOG to avoid these problems. However, so far, optimization has been limited to the correct design and initial installation dimensioning. Proper management of this initial stage is left to the experience of the operators to adjust the process when changes occur in the characteristics of the wastewater inlet. The main difficulty is the large number of factors influencing these changes. In this work, a prediction model of the FOG content in the inlet water is presented. The model is capable of correctly predicting 98.45% of the cases in training and 72.73% in testing, with a relative error of 10%. It was developed using random forest (RF) and the good results obtained (R^2 = 0.9348 and RMSE = 0.089 in test) will make it possible to improve operations in this initial stage. The good features of this machine learning algorithm had not been used, so far, in the modeling of pretreatment parameters. This novel approach will result in a global improvement in the performance of this type of facility allowing early adoption of adjustments to the pretreatment process to remove the maximum amount of FOG.

Keywords: wastewater; pre-treatment; FOG; random forest

check for updates

Citation: Mateo Pérez, V.; Mesa Fernández, J.M.; Villanueva Balsera, J.; Alonso Álvarez, C. A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs. *Water* **2021**, *13*, 1237. https://doi.org/10.3390/w13091237

Academic Editor: Fi-John Chang

Received: 23 March 2021
Accepted: 28 April 2021
Published: 29 April 2021

1. Introduction

Fats, oils, and greases (FOG) are some of the components of urban wastewater and the result of food preparation both in homes and in various commercial and industrial settings. FOG is a growing concern for municipalities and sewage plant operators, due to its tendency to cause severe blockages in pipes and sewers [1–3].

FOG characteristics can vary greatly depending on the types of fat, oil, and grease and their sources of collection [4]. FOGs can appear as liquids or solids and are characterized by a greasy texture and lower density than water, which is why they float on the surface. Furthermore, FOG can form emulsions in aqueous media in the presence of soap or other emulsifying agents. FOG is composed of fatty acids, triacylglycerol, and lipid-soluble hydrocarbons, with FFA (free fatty acids) being the most important components due to their chemical reactivity. The presence of a large amount of FFA results in a characteristically low pH [1,5].

Upstream of the treatment plants, the FOG with other types of waste generate the so-called "fatbergs" [2] that cause different problems in the pipes to the treatment plants [6]. Due to all the problems generated by the FOG, different prevention systems have been developed with different approaches, from educational campaigns to promote good management practices, the installation of grease trapping systems (GTSs), or the performance of periodic inspections to avoid improper disposal [7,8]. Numerous initiatives and programs

of this type have been implemented in various countries, although in general they are at the local level or pilot-scale and have not been extended nationally or internationally [2,5]. An example is in municipal management in Sweden and Norway, whereby installing GTSs in most restaurants, the number of problems and blockages due to accumulations of FOG significantly reduced [6].

Once in the treatment plants, the FOG that is not eliminated in the degreasing process can cause blockages and other problems in their infrastructures (pipes, pumps, tanks, digesters, sensors). This increases both the time and money required for cleaning and maintenance. The EU-RecOil project estimated that 25% of wastewater treatment costs can be attributed to the FOG component [9]. On the other hand, if they are not removed, FOGs consume oxygen from water and worsen the results of subsequent biological treatments, reducing the quality of the treated water. All these problems require additional capacity and energy in wastewater treatment plants, increasing operating and maintenance costs of the facilities [2]. As a consequence, different methods are used to remove and recycle these fats, oils, and greases at the beginning of the purification processes [10–13].

Compared to other research work carried out in relation to FOG, usually focused on studying their physical and chemical characteristics, the processes of subsequent use or recycling, or their effect on the biological treatments of wastewater, among other examples, the focus of this study is to improve the operability of WWTPs. The mechanical separation of FOGs in pretreatment has received less attention from researchers compared to their energy use [5,14], reducing environmental impact in landfills [4,15,16] or their influence on downstream treatments in WWTP [11]. The objective of this work is to improve the removal of FOG in the pretreatment stage, which will have an impact on the improvement of the performance of the subsequent stages and the general operation of the wastewater treatment plant.

Treatment plants have to manage significant changes in the flow rate and characteristics (composition, temperature, etc.) of the incoming wastewater [17,18]. More specifically, many factors influence the amount, proportion, and characteristics of the FOG content of the inlet wastewater from such facilities:

- Weather changes, i.e., rain, more or less intense, ambient temperature, number of previous days without rain with consequent reduction of the inflow, among others, modify the quantity and characteristics of FOG reaching the WWTP. Predicting these weather events and their influence on different management infrastructures water has been studied in numerous works [19–22];
- The part of FOG from domestic activities is altered by holidays, vacation periods, the different seasons of the year, or the weather itself [3];
- The features of commercial sources of FOG (size, density, and geographical distribution) such as restaurants, and the use of grease trapping systems, for example [1,8];
- Another important source of FOG is industrial activities, such as food processing or slaughterhouse factories [13,23,24];
- The presence of other types of residues mixed with FOG present in the wastewater, such as gross solids (especially wet wipes), grit, and others [25].

Another important challenge of this work involved the selection and subsequent processing of the input variables to have an adequate number of training and testing patterns. Current WWTPs collect a large amount of data, often unused for facility management, so it is necessary to make an initial effort of exploration, visualization, and selection of relevant information [17,26].

This paper is divided into three main sections. Section 2 describes the characteristics of the WWTP being studied, the acquisition and processing of data, and the mathematical techniques used in the development of the model. Collecting data from different sources and different frequencies, to have enough training and test patterns and subsequent processing to ensure quality and representativeness have been one of the initial challenges of this work. Next, in Section 3, the results obtained are presented and discussed, both in the model training process and in its validation. These results indicate that the FOG

prediction model developed has enough accuracy to provide valuable information that will improve the operation of the WWTP. Finally, the main contributions of the study are highlighted in Section 4.

2. Materials and Methods

2.1. Case Study

The Villapérez Wastewater Treatment Plant is located in the northeast of the city of Oviedo (Spain) and occupies an area of nearly 21 hectares (Figure 1). It provides service to an approximate population of 723,000 equivalent inhabitants. Wastewater arrives at Villapérez through a unitary network of collectors that has an approximate length of 75 km. This network includes 44 spillways. Collector diameters range from 600 mm to 2000 mm with sections in gravity and impulsion.

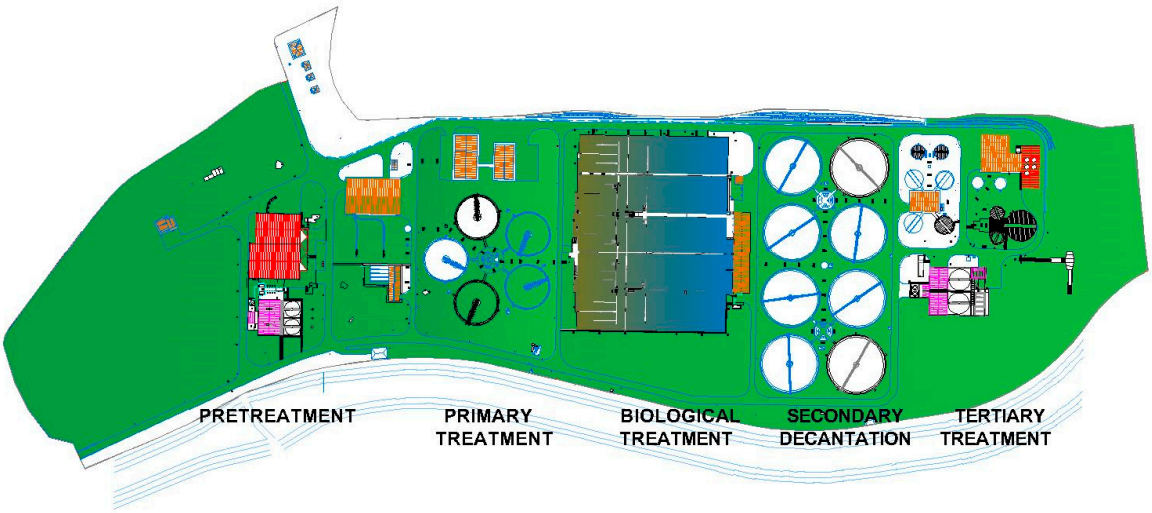


Figure 1. Plan view of the Villapérez wastewater treatment plant (WWTP) (Asturias, Spain).

The Villapérez WWTP collects both urban and industrial wastewater. One of the main industries that discharge to Villapérez is a dairy facility with a production capacity of 500,000,000 million liters of milk per year and that discharges an average flow of 200 m^3/h into the sanitation network. Therefore, the representativeness of this WWTP is given by being a medium-sized facility, which receives urban wastewater from a relatively large area and which must also treat industrial discharges with high FOG content such as dairy industries.

As can be seen in Figure 1, the wastewater treatment in Villapérez WWTP begins with a pretreatment stage in which the larger solids, sands, and fats are removed. Subsequently, water is taken to primary settling by gravity. Then, water goes to biological treatment where organic matter, nitrogen, and phosphorus are removed. This treatment involves passing the water through several anoxic chambers, anaerobic and aerobic. The next stage is secondary settling, which is carried out via gravity. Finally, the tertiary treatment stage consists of a physical-chemical treatment, lamellar settling, and filtration.

The pre-treatment has the capacity to treat an inflow of 8.5 m^3/s and starts with two, thick wells, equipped with a 500 L clamshell bucket. The plant then has four roughing channels, each of which includes an automatic cleaning screen with a 60 mm clearance and a self-cleaning fines screen with a 3 mm clearance and an inclination of 50°. After the roughing stage, the water reaches the facilities for separating FOG and sands from raw water, which consist of 5 rectangular grit traps with a unit useful volume of 449.8 m^3. To properly separate the FOG, they are first emulsified, and for this, the grit traps are aerated: 2/3 of the length of the grit remover using coarse bubble aerators, and 1/3 of the grit remover using fine bubble diffusers. Once the fat has been emulsified, it is collected by a scraper that cyclically runs the entire length of the sand trap.

After this separation, the emulsified FOG is sent to a fat concentrator by means of chains and scrapers that separate water from fat (Figure 2). These concentrators have a flow rate of 30 m^3/h and a power of 0.18 kW. The Villapérez WWTP removes an average of 5.25 tons of FOG per month, which is approximately 63 tons per year, or in other words, a container is filled every 9 days.

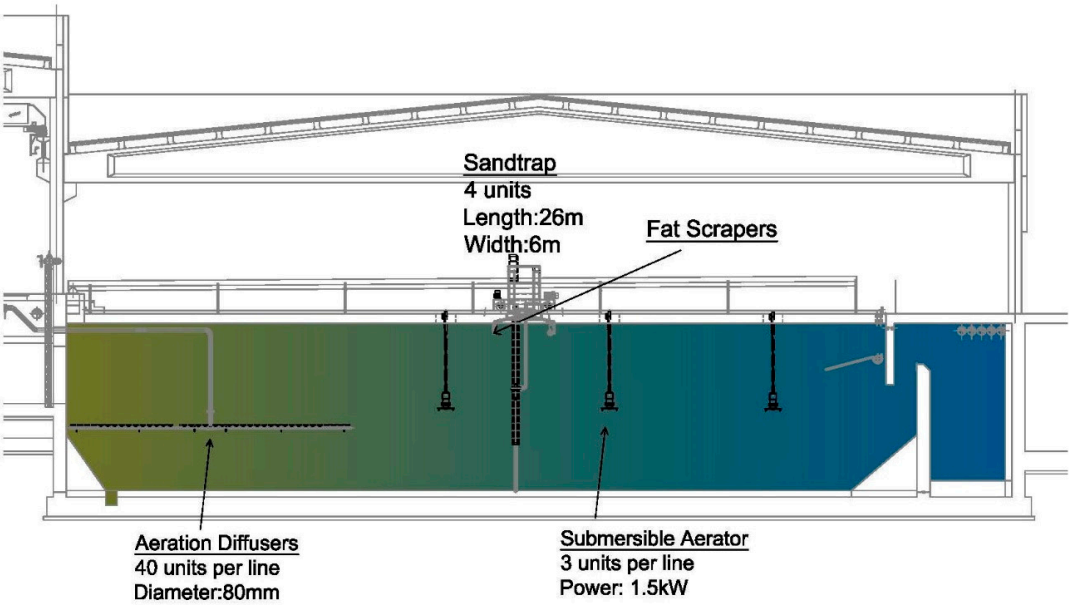


Figure 2. A sectional view of a FOG concentrator.

The main design parameters of the treatment plant are included in Table 1.

Table 1. Design parameters of wastewater treatment plant of Villapérez (Asturias, Spain).

Parameter	Design Value
Maximum inflow (rainy weather)	8.50 m^3/s
Maximum inflow (dry weather)	2.89 m^3/s
Five-day biological oxygen demand (BOD5)	418.00 mg/L
Chemical oxygen demand (QOD)	652.00 mg/L
Total suspended solids (TSS)	329.00 mg/L
Total Kjeldahl nitrogen (N-NTK)	47.40 mg/L
Total phosphorus (Pt)	6.50 mg/L

2.2. Data

All data used in this work were collected in the period from 1 March 2017 to 24 June 2019 and come from different sources:

- Data related to wastewater were obtained through the Supervisory Control and Data Acquisition software (SCADA) of the WWTP. This system registers 226 parameters every 9 min from measuring equipment and sensors distributed all over the treatment plant. From this set of data, the data associated with the measurement of input parameters in the raw water during the pre-treatment stage were used. The parameters measured in the raw water are the input flow rate, pH, raw water temperature, conductivity, and ammonia. The data associated with these variables are identified by the time and date of the data measurement.
- FOG data were collected from the container removal delivery notes, which contained the actual data of the waste total weight inside each container. The number of containers in the study period was 89. Their filling time was used as time intervals to group the data of the SCADA system.

- Climate data comes from the Spanish State Agency for Meteorology website (Agencia Estatal de Meteorología, Aemet) and the pluviometry data (instantaneous and accumulated rainfall) is obtained from those recorded by the plant's weather station. All of them are also grouped considering the intervals in which the containers are filled. From these data, a new calculated variable from the instantaneous precipitation is also created, corresponding to the number of previous days without rain.

Statistical data for the variables initially considered in the study are presented in Table 2. As indicated above, the reference is the time interval from when an empty container is placed to when it is removed. When each container is removed, it is weighed, and the data is recorded on the corresponding delivery note. For the elaboration of the training patterns, some variables have been calculated. The data corresponding to each of these periods was summarized by calculating for each variable its minimum, mean and maximum value, as shown in Table 2.

Table 2. Statistical description of the variables.

Variable	Description	Unit	Mean	Standard Deviation	Min	Max
FOG	Fats, oils, and greases	ton	3.01	0.26	2.32	3.52
Interval	Time interval	h	228.91	660.09	1.28	6289.76
PDwR	Previous days without rain	day	2.06	3.61	0.00	19.55
MxDwR	Maximum previous days without rain in the time interval	day	4.42	4.09	0.07	20.68
Vol	Water volume	m^3	731,056.32	817,413.43	3946.58	4,886,022.43
PrecipTotal	Total precipitation	m^3	13.88	26.27	0.00	203.40
PrecipMax	Maximum precipitation	m^3	1.09	1.95	0.00	12.00
pH	pH		7.21	0.32	6.22	7.99
pHMax	Maximum pH		8.20	0.64	7.01	11.65
MedTemperature	Wastewater medium temperature	°C	17.98	3.05	10.95	22.55
MaxTemperature	Wastewater maximum temperature	°C	19.59	2.76	12.68	25.62
MedConductivity	Medium conductivity	µS/cm	996.72	212.72	380.80	1439.72
MaxConductivity	Maximum conductivity	µS/cm	1995.47	541.84	757.62	3768.84
MedAmmonium	Medium ammonium	mg/L	27.61	12.36	9.06	68.31
MaxAmmonium	Maximum ammonium	mg/L	38.33	17.41	15.82	88.22
MedFlow	Medium flow	m^3/h	4193.95	1896.00	2446.96	12,608.21
MaxFlow	Maximunm flow	m^3/h	9216.91	3958.30	3446.37	17,885.11
MinFlow	Minimun flow	m^3/h	1779.97	1004.17	975.59	6803.43
TempExtMed	Medium Ambient Temperature	°C	13.14	4.41	3.30	22.20
TempExtMax	Maximunm Ambient Temperature	°C	17.50	5.12	4.60	28.20
TempExtMin	Minimun Ambient Temperature	°C	9.75	4.51	−0.20	17.60
MedPDwR	Medium previous days without rain in the time interval	day	2.12	3.10	0.01	19.52

A preliminary analysis by principal component analysis (PCA) [27] was carried out in order to study the initial data set. The graph in Figure 3 shows the contribution of the different variables to the dimensions of the PCA projection.

Some aspects that can be highlighted from this graph are:

- As might be expected, the temperature variables (ambient temperature, wastewater temperature) appear grouped.
- Conductivity is related to the number of days without rain. This is because wastewater, both urban and industrial, is not diluted by rainwater.
- Obviously, the flow variables are related to the level of precipitation, that is, the more rain, the higher the inlet flow.
- Finally, it can be seen how the amount of FOG (fat variable) is related to ammonium, and therefore this is an important parameter to consider in the modeling. This relationship may be due to industrial discharges since they provide both fat and nitrogen.

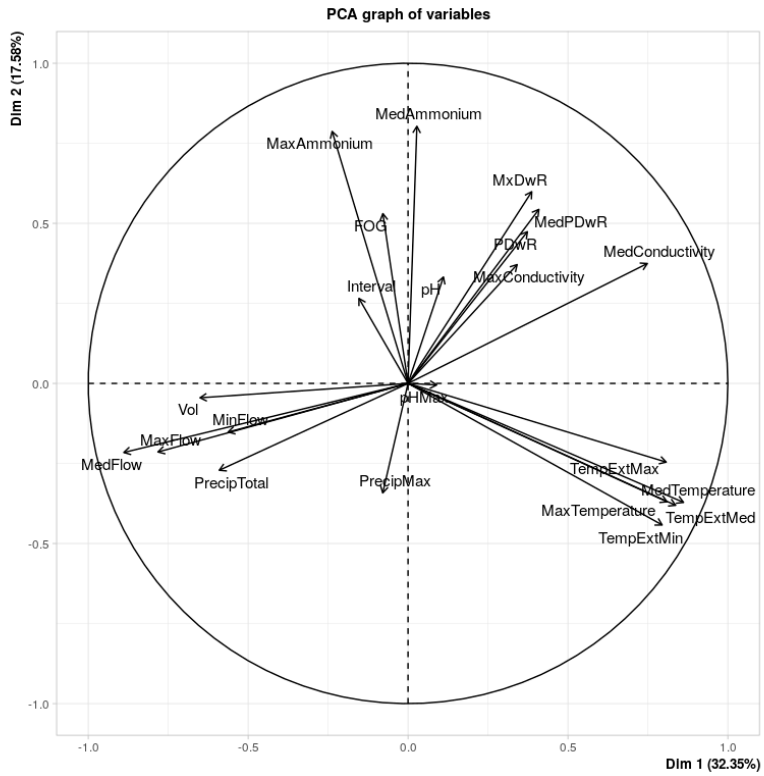


Figure 3. PCA graph of variables.

Figure 4 shows the contribution of each of the variables in the complete dataset. It can be seen that the FOG variable is one of the variables that least contributes to variability and this is because it has a fairly steady behavior. The dotted reference line in red corresponds to the expected value if the contributions were uniform.

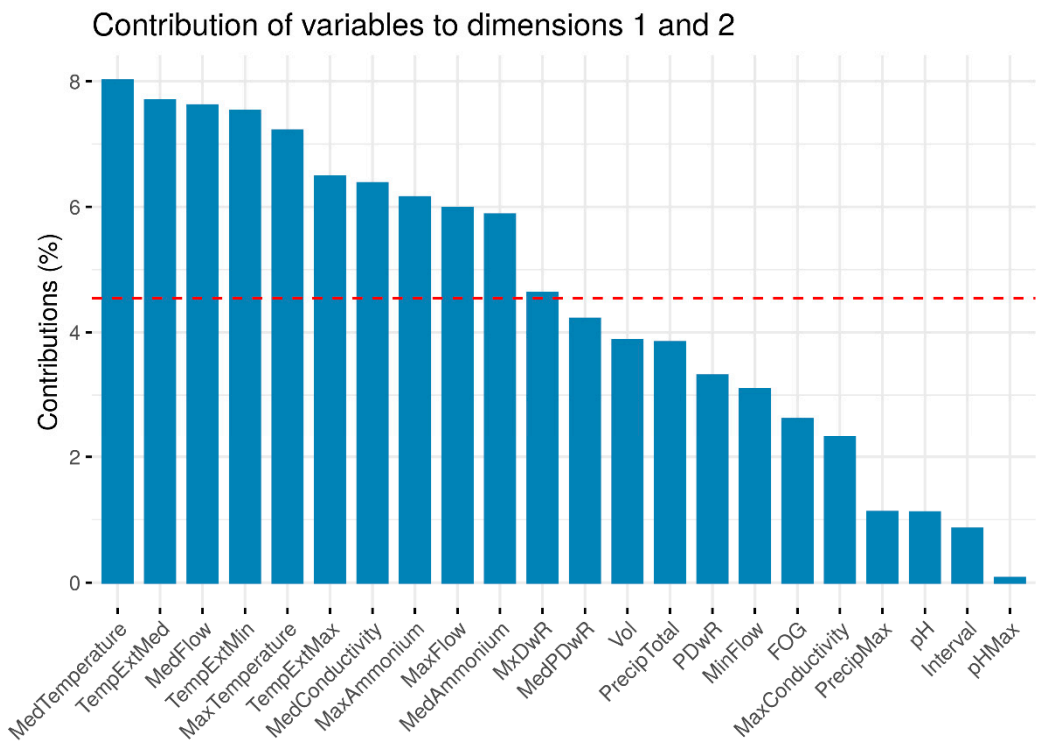


Figure 4. Contribution of variables to Dimensions 1 and 2.

Finally, a PCA plot (Figure 5) was performed in order to detect outliers and groups of cases with similar characteristics. After analyzing the within clusters summed squares (WCSS) and using the elbow method (a heuristic used in determining the number of clusters in a data set [28]), 4 was the optimal number of groups we decided to take. For group identification, hierarchical clustering [29] has been chosen, using complete linkage clustering [30] as the agglomeration method.

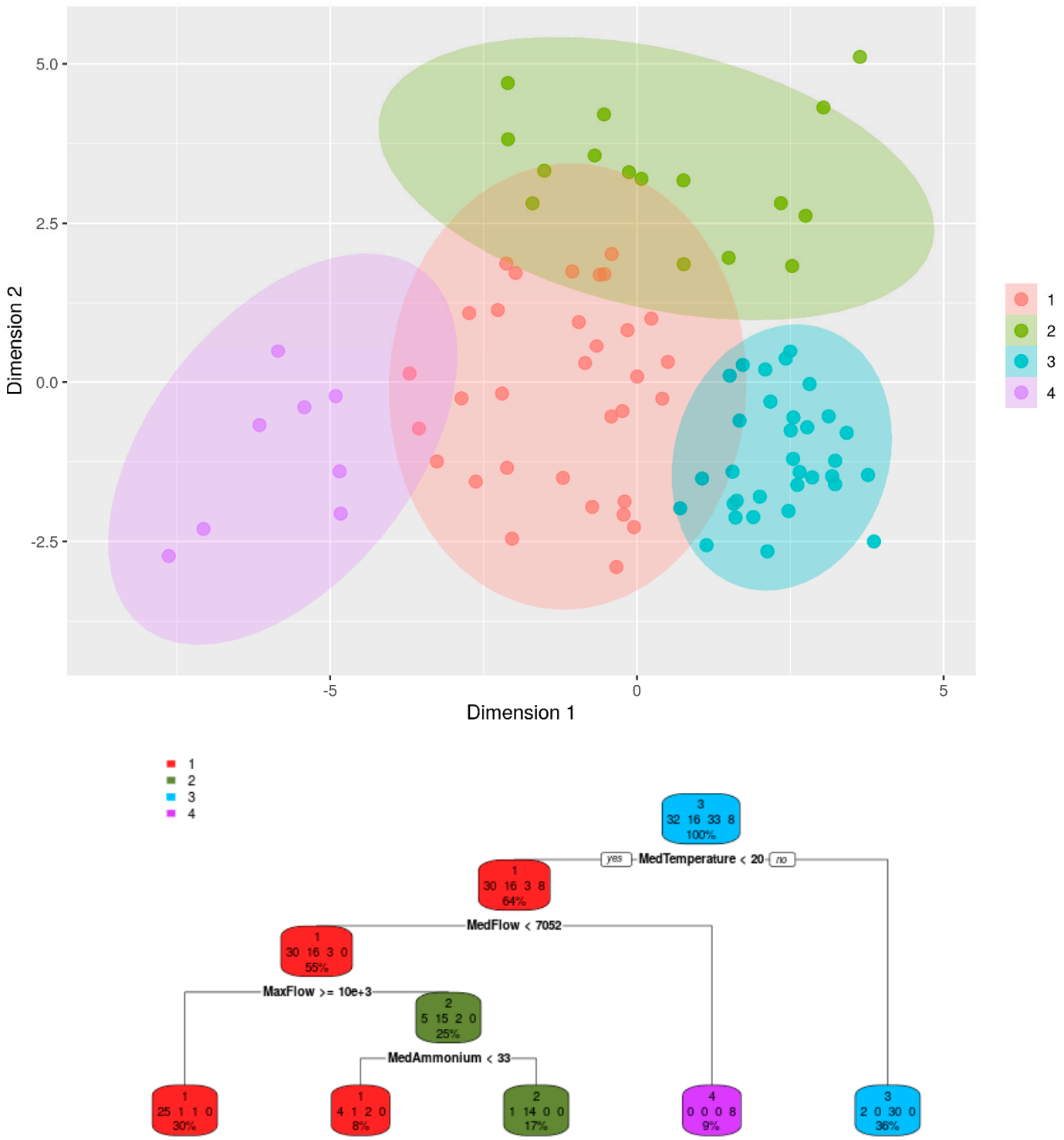


Figure 5. FOG Clusters of PCA projection.

Four groups can be observed (Figure 5) with the following characteristics:

- Cluster 1 includes those cases with a maximum flow value greater than 10,000 m^3/h;
- Cluster 2 part of the cases with a maximum flow greater than 10,000 m^3/h and also with average ammonium above 33 mg/L are included in this group;
- Cluster 3 consists of data with an average temperature greater than 20 °C;
- Cluster 4 is defined by an average flow greater than 7052 m^3/h and includes 100% of the cases in this cluster.

Figure 6 shows the same projection of the data of the previous figure (Figure 5), but representing the variables average temperature (MedTemperature), average flow (MedFlow), and average ammonia (MedAmmonium) in the same way. Comparing both graphs, it is possible to observe that the cases with the highest average temperature are in the area of cluster 3. In the graph at the bottom left, it can be seen how the points with the lowest average flow (MedFlow) values correspond to the cases of cluster 2 and 3. Finally, the points with the highest average ammonium values (MedAmmonium) correspond to cluster 2.

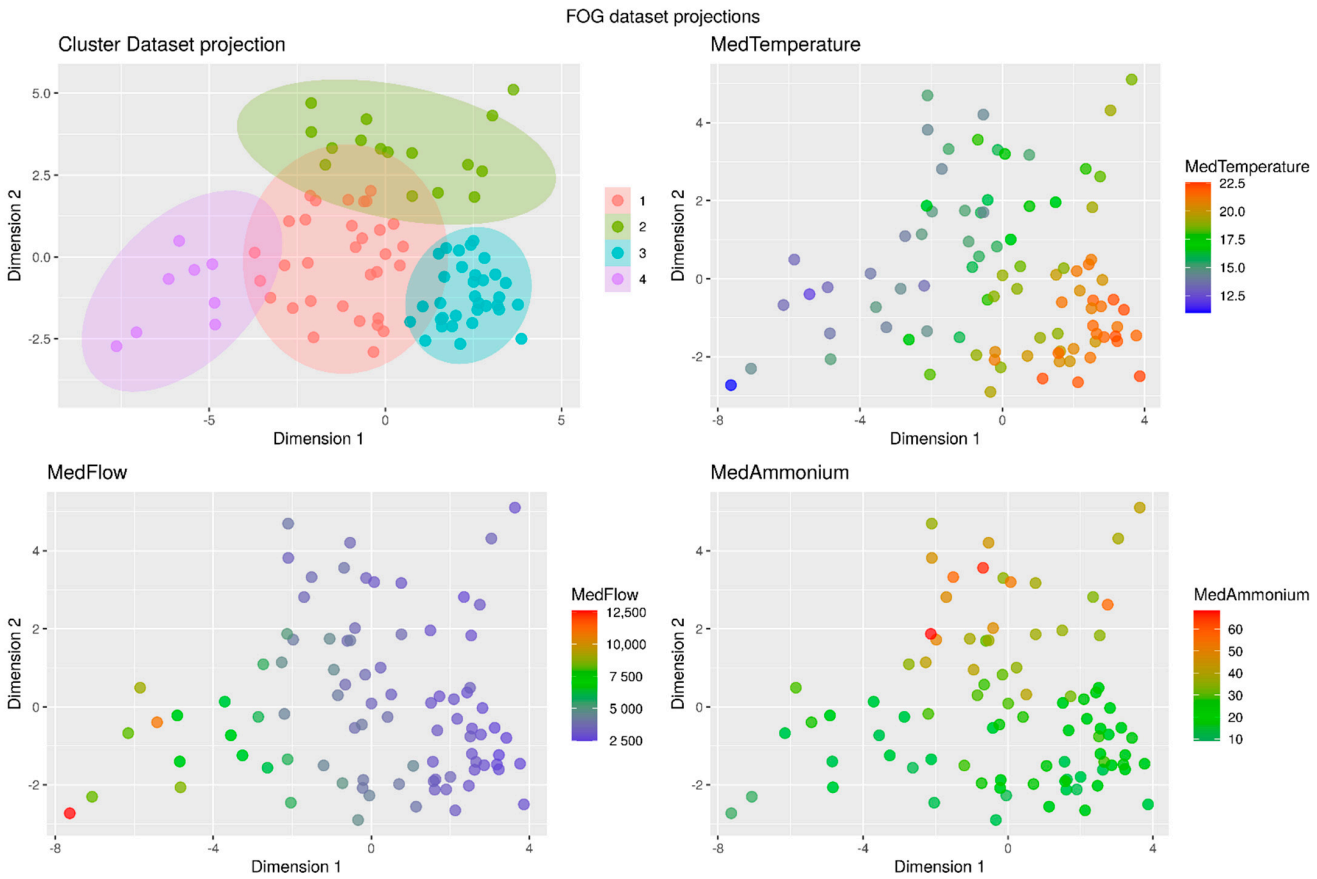


Figure 6. Comparing dataset PCA projection with clustering variables.

2.3. Methods

In this study, random forest (RF) analysis, a machine-learning approach for feature selection from highly multivariate datasets, was used to develop a forecast model of FOG content in the inlet wastewater. The RF algorithm reaches the final prediction from the majority voting of the decisions made with multiple decision trees constructed with randomly permuted features and observations via recursive partitioning [31]. RF method has been applied in a wide range of research areas due to its numerous advantages [32] and in recent years it has gained great importance in water resource-related research. Random forests have been used to address numerous research problems in WWTPs, such as:

- Estimating different parameters of water quality or processes as chemical oxygen demand (COD) [33], total suspended solids (TSS) [34], stream nitrogen (N) and phosphorus (P) concentrations [35], or influent flow of WWTPs [36];
- To monitor different treatment processes such as to make predictions of 'settleability' of activated sludge [37], or nitrogen removal systems [38];
- To generate models of energy cost [39] or pumping systems [40] in WWTPs;
- To obtain other improvements in plant control [41] or reliability of small wastewater treatment plants [42].

The main advantage of the random forest algorithm over other techniques is its great generalizability [42,43], which is why it has been used in a growing number of works related to water management [32] such as those indicated above. In addition, RF is able to provide better information compared to other methods on the importance of each input variable [36]. Good accuracy achieved by the RF models and the ability to more easily interpret the results over other methods were the main reasons for their use in this case study.

The model presented in this paper was developed using *R* [44] and the packages *caret* [45] and *randomForest* [46].

3. Results and Discussion

The representativeness of training datasets is very important to the effectiveness and overall performance of an RF model [47]. In this study, 90% of the data in the original dataset are selected randomly to generate a training dataset, while the other 10% are used to form the corresponding testing dataset in order to have a sample as representative as possible. In addition to configuring the data set, the training process requires adjusting several parameters. The number of trees (ntree) and the number of variables randomly sampled as candidates at each split (mtree) are the two most important parameters because they have a big effect on the final accuracy of an RF model [48,49]. To adjust these parameters, the cross-validation algorithm was used with a division into three folds and repeating the training ten times [50].

After the training process, different parameters to evaluate model results have been taken into consideration:

- Root mean square error (*RMSE*) is a frequently used measure of the differences between values predicted by a model and the values observed. The smaller the value, the better the model's performance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{1}$$

- Mean absolute error (*MAE*) is also a common measure to forecast a model's error.

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{2}$$

The determination coefficient (R^2) is the proportion of the variance in the dependent variable that is predictable from the independent variables and it is a statistical measure of how well a model approximates the real data points. A bigger value indicates a better fit between prediction and actual value.

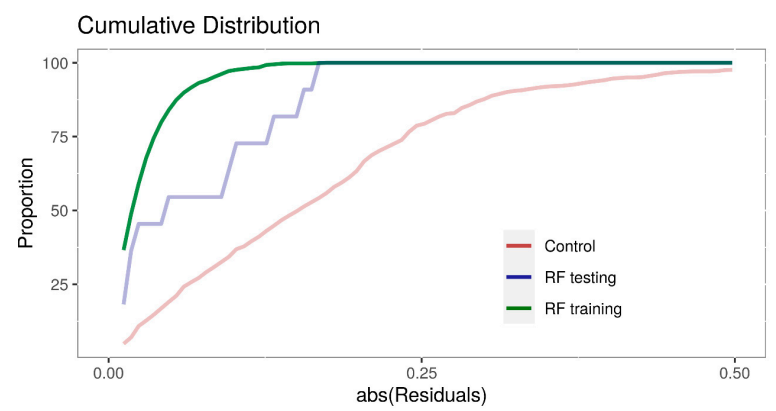
$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{3}$$

The model developed for FOG content prediction in the inlet waters to the wastewater treatment plant presents the following values (Table 3) and the three indicators show very good performance.

Table 3. Model performance indicators.

	RMSE	*MAE*	R^2
Training	0.037	0.025	0.9888
Testing	0.089	0.066	0.9348

Figure 7 compares the performance of the model in training and test with an estimate using the mean value of content in FOG. It can be seen that with a relative error of 10%, the model is capable of correctly predicting 98.45% of the cases in training and 72.73% in testing, while under these same conditions the mean FOG value would only be correct in 24.17% of the cases.



Relative Error (%)	Absolute error	Control	RF	RF Test
1%	0.012	4.86	36.57	18.18
5%	0.060	24.17	89.88	54.55
10%	0.120	41.12	98.45	72.73
14%	0.168	54.24	99.90	100.00
20%	0.240	76.65	100.00	100.00
25%	0.300	87.71	100.00	100.00

Figure 7. Model performance indicators.

Initially, 22 variables were introduced for the generation of the model, 11 of them were discarded during the training process since they were not used in any of the splits. The relative importance (Table 4) of the model variables can be calculated with samples not selected in the cross-validation sub-samples used to construct a tree [51].

Table 4. Variable relative importance.

	Standardized Overall	**Absolute Overall**
MedAmmonium	100.00	9.002
MaxAmmonium	81.046	7.606
PrecipMax	47.079	5.103
MedConductivity	23.024	3.331
MxDwR	17.963	2.958
PDwR	17.678	2.937
pH	15.501	2.777
TempExtMed	9.171	2.311
MedTemperature	6.309	2.100
MedFlow	4.074	1.935
MedPDwR	0.00	1.635

One of the most significant advantages of the RF method is its evaluation of the importance of the variables used in the training process [52]. The interpretation made of the importance of these variables in the development of the model is described below:

- In this case, the two most relevant variables are the average (MedAmmonium) and maximum (MaxAmmonium) ammonium values. This could be due to the large amount of ammonium and FOG contained in the discharges from the dairy facility served by the Villapérez WWTP as was mentioned in the case study description;
- The third most significant variable is maximum precipitation (PrecipMax). Greater precipitation implies a greater inflow into the WWTP, with more dissolved FOG, which makes it difficult to remove it in the pretreatment process;
- Urban wastewater has a steady conductivity, so it is possible to associate the variations and relevance of this variable with industrial discharges;
- The relevance of the following variables related to the number of previous days without rain (MxDwR, PDwR, and MedPDwR) can be explained in a similar way to precipitation, that is, as there is less inflow to be treated, the FOG is less dissolved and it is possible to remove it in a greater proportion;
- pH: urban wastewater has a relatively steady pH, so variations in this indicator can be associated with industrial discharges;
- The average temperature (TempExtMed) provides information on the seasonal situation at the time of analysis. A higher temperature makes it easier to emulsify the FOG and therefore its removal is more effective;
- The relevance of the average flow variable (MedFlow) can be explained in the same way as the precipitation or the number of previous days without rain mentioned above;

In Figure 8, the behavior of the training data is represented. It can be observed that the predicted data precisely fit the real ones and how the errors have a steady behavior, which reinforces the quality of the model.

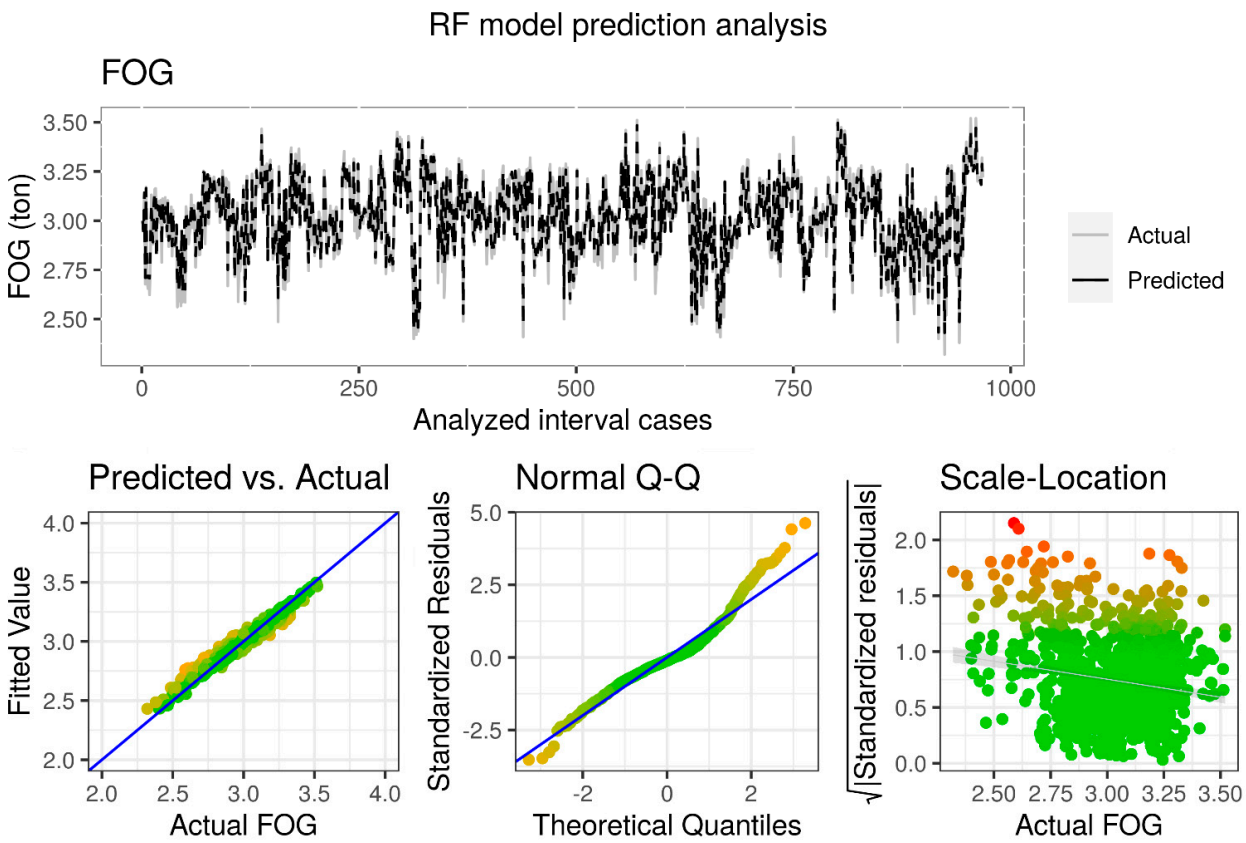


Figure 8. RF model prediction analysis (training).

Similarly, in Figure 9, it is possible to observe the performance of the RF model with the test data. The model is capable of adequately predicting the trend of the behavior of the arrival of FOG, which will provide relevant information when making decisions in plant operations.

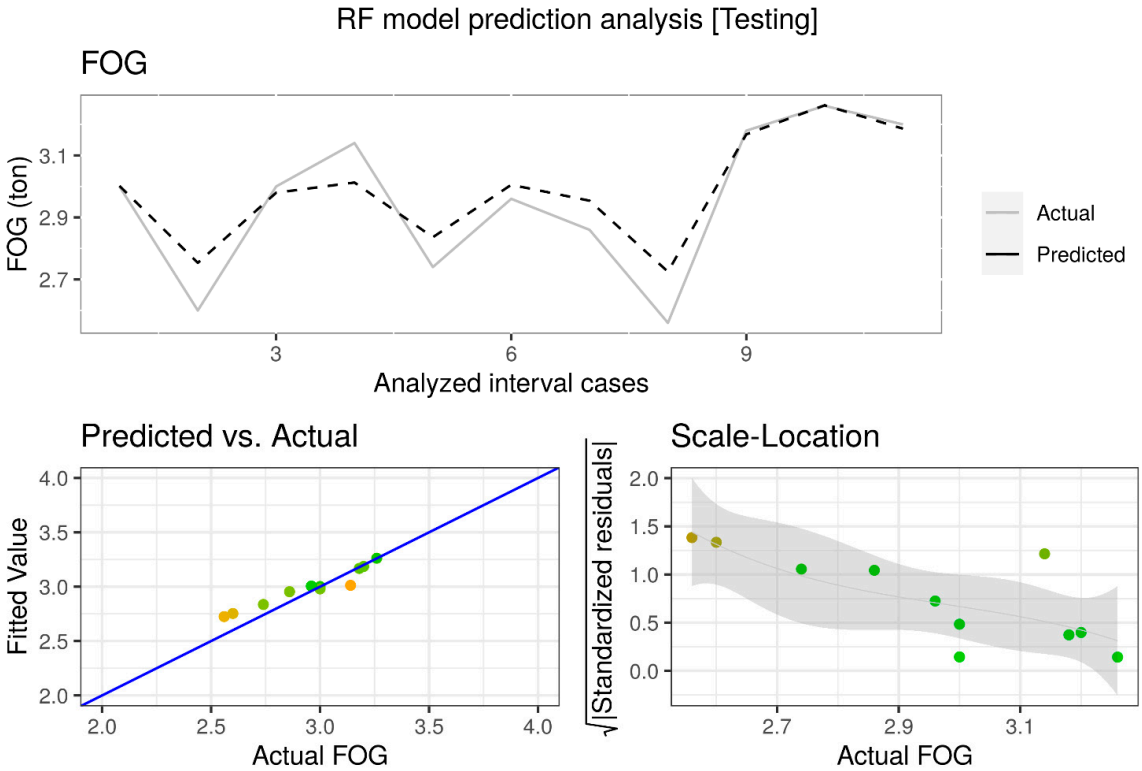


Figure 9. RF model prediction analysis (testing).

The sensitivity analysis of the FOG model developed assesses the change produced in the output in response to the variation of one (Figure 10) or two of the inputs (Figure 11). In this way, it is possible to identify from which value of a variable a trend change in the FOG content is expected.

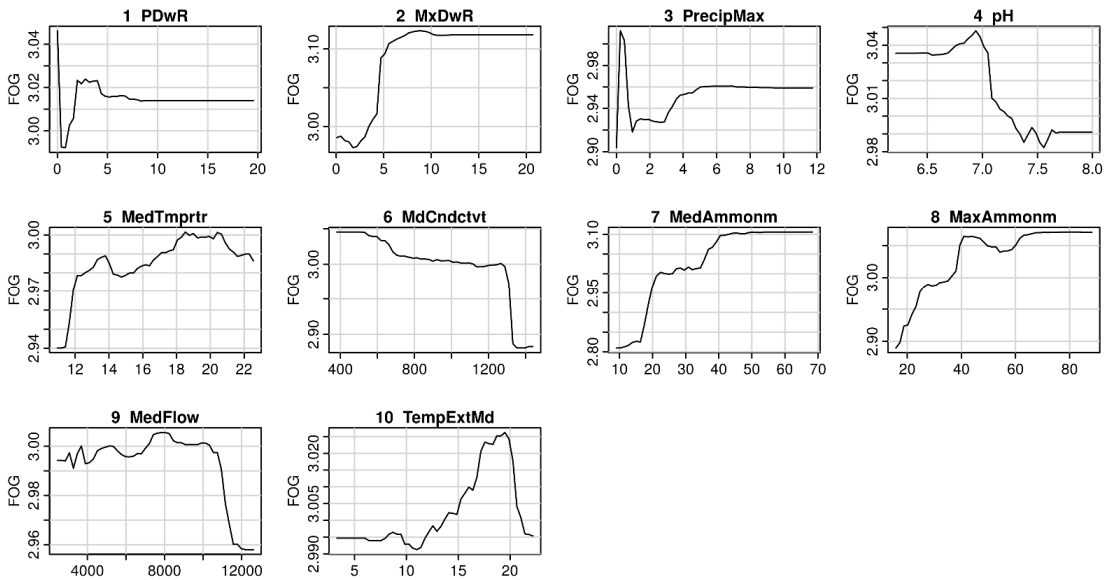


Figure 10. Sensitivity analysis (one variable).

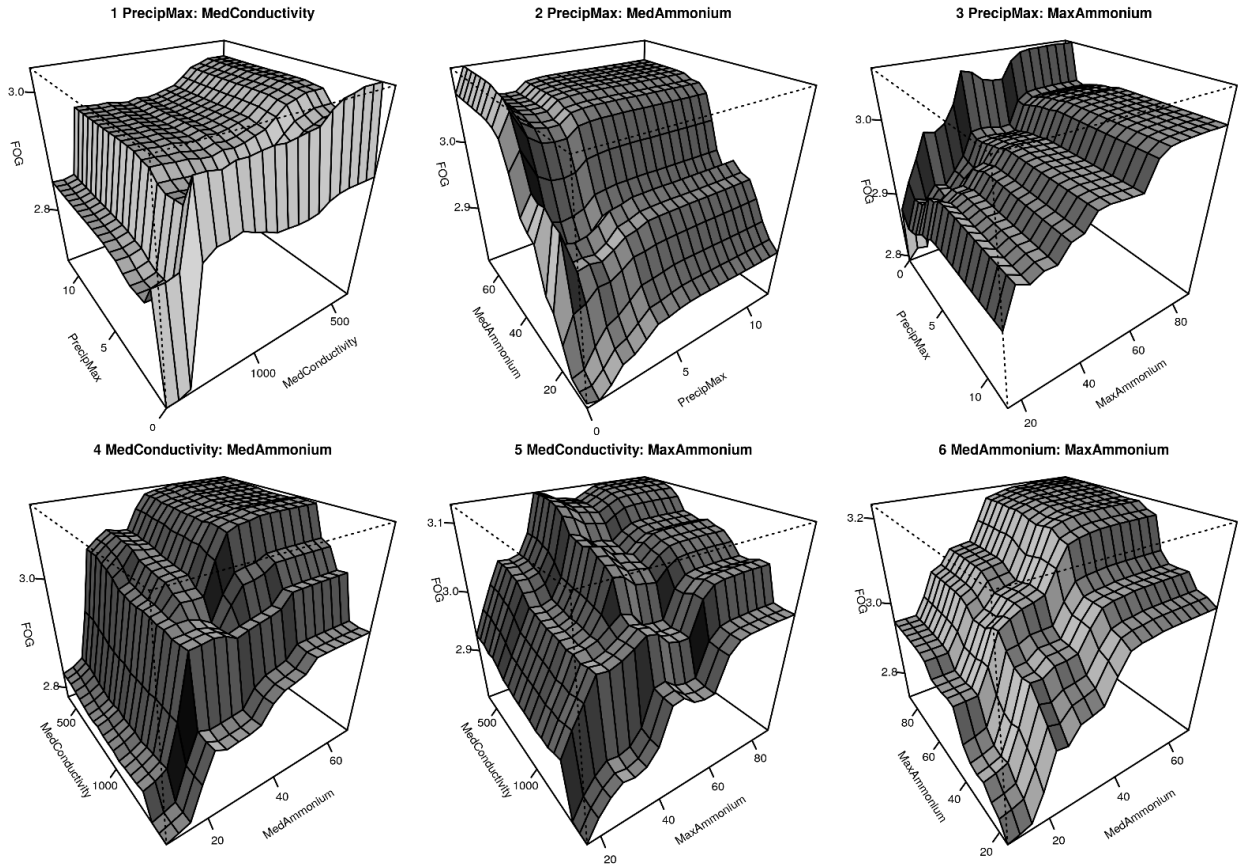


Figure 11. Sensitivity analysis (two variables).

As can be seen in Figure 10, the behavior of the variables is consistent and fits what is expected. Despite the fact that the y-axis (FOG) has small variation ranges, the expected trends can be seen. For example, it is possible to observe that increasing the average and maximum ammonium (MedAmmonium and MaxAmmonium) increases the amount of FOG (Figure 10 (7, 8)). Some studies have modeled the amount of ammonium in wastewater, indicating greater uncertainty in the estimation during periods of rain, but without referring to the content of FOG [53]. Also, when it has been raining recently, that is, a number of days without rain (PDwR) close to zero, an initial washing effect is produced in the pipes and sewers that increases the arrival of FOG while when this variable increases the amount of FOG is little influenced. This behavior, with an initial increase of all types of waste such as FOG at the beginning of the rain episodes, with a subsequent dilution, has also been observed in other research works [54]. Along with this, the changes in pH are in agreement with the results of other studies, where the pH values on rainy days are numerically higher [55].

Figure 11 shows how the variation of two variables affects the FOG content in the inlet water. As already indicated, it is confirmed that the presence of ammonium is not influenced by the variation in precipitation, since it is mainly due to discharges derived from industrial activities (Figure 11 (2, 3)). On the contrary, it can be observed how the variation of ammonium affects the conductivity values (Figure 11 (4, 5)).

Tests have been carried out with other predictive methods of regression machine learning, such as multivariate adaptive regression splines (MARS) [56] or support vector machine (SVM) [57]. However, when performing the corresponding sensitivity analyzes, it has been seen that the model generated with RF presents greater stability since it better adjusts to the behavior expected by the target variable. In this case, the other techniques extrapolate the data worse, generating anomalous values in areas where the dataset has a low information density. Many of these advantages of RF, such as the ability to identify

non-linear relationships between the predictor and the dependent variables [58], not overfitting [59], the handling of highly correlated variables [60], or the possibility of ordering the relative importance of the variables [61] have been previously identified by several authors in other fields. In addition, as other researchers indicate, the potential of this algorithm in the field of water resources has been very little exploited [32]. Even less has it been used in the field of the pretreatment stage of a WWTP which, as previously mentioned, has not received much research attention so far, which constitutes one of the novelties of this work. No other scientific publication has been found in which a similar prediction model has been presented, so it has not been possible to compare the results.

The ability to anticipate trends in incoming wastewater provided by the model will allow the pretreatment process to be adjusted to optimize FOG removal. This process does not detect if there is an increase in the FOG content, so it is not adjusted until that increase is detected in the downstream stages. For example, when large production peaks occur FOG air injection is varied to optimize emulsification. Reducing the time for the early adoption of this type of measure, thanks to the information provided by the model presented in this work, will certainly improve the removal of the FOG content and will positively affect all the treatment processes of the WWTP.

4. Conclusions

Like other fractions of urban wastewater withdrawn in the pretreatment stage of wastewater treatment plants, the optimization of FOG removal has received relatively little attention from researchers beyond its subsequent use or its influence on subsequent wastewater treatment processes. However, its influence on these later stages of wastewater treatment can be important to improve both the overall performance of WWTP and their operability. With this objective, in this work, a prediction model of the FOG content in the inlet waters of the treatment plant has been developed. The ability to provide operators with advanced information of changes in the wastewater entering the WWTP, taking into account various factors (chemical composition, meteorological changes, seasonal changes, etc.) had not been addressed so far in any other research.

The model is based on data collected for more than two years at the plant of Villapérez (Oviedo, Spain) and the well-known random forest algorithm, but which had not been used for this purpose so far. The results obtained, evaluated using several common indicators, reflect the good performance of the model both in the training ($RMSE$ = 0.037, MAE = 0.025 and R^2 = 0.9888) and test ($RMSE$ = 0.089, MAE = 0.066 and R^2 = 0.9348) stages. Thanks to the features of the RF technique, the most relevant variables used in the model have been interpreted, such as ammonia or changes in precipitation. As expected, the influence on changes in the FOG content of industrial discharges is highlighted in the case study.

Better information will enable operators to better decision-making, allowing optimization of the removal of FOG in pretreatment processes. It will result in a reduction of the content of FOG subsequent processes and a reduction of energy consumption and maintenance costs of the plant.

Future research could apply similar RF models to other WWTPs with different characteristics to verify their good performance. On the other hand, WWTPs receive other important wastes, such as gross solids or grit, whose prediction could be integrated into a more complete model of the incoming wastewater features.

Author Contributions: Conceptualization, V.M.P. and J.M.M.F.; methodology, J.M.M.F. and J.V.B.; data curation, V.M.P. and J.V.B.; writing—original draft preparation, J.M.M.F., V.M.P. and C.A.Á.; writing—review and editing, V.M.P., J.M.M.F., J.V.B. and C.A.Á. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank Aguas de las Cuencas de España (ACUAES) and the joint venture formed by Dragados S.A. and Drace Infraestructuras S.A. for their collaboration in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Husain, I.A.F.; Alkhatib, M.F.; Jammi, M.S.; Mirghani, M.E.S.; Zainudin, Z.B.; Hoda, A. Problems, Control, and Treatment of Fat, Oil, and Grease (FOG): A Review. *J. Oleo Sci.* **2014**, *63*, 747–752. [CrossRef]
2. Wallace, T.; Gibbons, D.; O'Dwyer, M.; Curran, T.P. International Evolution of Fat, Oil and Grease (FOG) Waste Management—A Review. *J. Environ. Manag.* **2017**, *187*, 424–435. [CrossRef]
3. Arthur, S.; Blanc, J. *Management and Recovery of FOG (Fats, Oils and Greases)*; CREW—Scotland's Centre of Expertise for Waters: Edinburgh, UK, 2013.
4. Salama, E.-S.; Saha, S.; Kurade, M.B.; Dev, S.; Chang, S.W.; Jeon, B.-H. Recent Trends in Anaerobic Co-Digestion: Fat, Oil, and Grease (FOG) for Enhanced Biomethanation. *Prog. Energy Combust. Sci.* **2019**, *70*, 22–42. [CrossRef]
5. Abomohra, A.E.-F.; Elsayed, M.; Esakkimuthu, S.; El-Sheekh, M.; Hanelt, D. Potential of Fat, Oil and Grease (FOG) for Biodiesel Production: A Critical Review on the Recent Progress and Future Perspectives. *Prog. Energy Combust. Sci.* **2020**, *81*, 100868. [CrossRef]
6. Mattsson, J.; Hedström, A.; Ashley, R.M.; Viklander, M. Impacts and Managerial Implications for Sewer Systems Due to Recent Changes to Inputs in Domestic Wastewater—A Review. *J. Environ. Manag.* **2015**, *161*, 188–197. [CrossRef]
7. Paraíba, O.; Tsoutsos, T.; Tournaki, S.; Antunes, D.; Lino, J.; Manning, E. Strategies for Optimization of the Domestic Used Cooking Oil to Biodiesel Chain. The European Project Recoil. In Proceedings of the 20th European Biomass Conference and Exhibition, Milan, Italy, 18–22 June 2012; pp. 18–22.
8. Kobayashi, T.; Kuramochi, H.; Xu, K.-Q. Variable Oil Properties and Biomethane Production of Grease Trap Waste Derived from Different Resources. *Int. Biodeterior. Biodegrad.* **2017**, *119*, 273–281. [CrossRef]
9. EUBIA—The European Biomass Industry Association. *Transformation of Used Cooking Oil into Biodiesel: From Waste to Resource*; Position Paper, Promotion of Used Cooking Oil Recycling for Sustainable Biodiesel Production (RecOil); The European Biomass Industry Association: Brussels, Belgium, 2015.
10. Khuntia, H.K.; Janardhana, N.; Chanakya, H.N. Fractionation of FOG (Fat, Oil, Grease), Wastewater and Particulate Solids Based on Low-Temperature Solidification and Stirring. *J. Water Process Eng.* **2020**, *34*, 101167. [CrossRef]
11. Solé-Bundó, M.; Garfí, M.; Ferrer, I. Pretreatment and Co-Digestion of Microalgae, Sludge and Fat Oil and Grease (FOG) from Microalgae-Based Wastewater Treatment Plants. *Bioresour. Technol.* **2020**, *298*, 122563. [CrossRef] [PubMed]
12. Hao, J.; de los Reyes, F.L., III; He, X. Fat, Oil, and Grease (FOG) Deposits Yield Higher Methane than FOG in Anaerobic Co-Digestion with Waste Activated Sludge. *J. Environ. Manag.* **2020**, *268*, 110708. [CrossRef] [PubMed]
13. Agabo-García, C.; Solera, R.; Pérez, M. First Approaches to Valorizate Fat, Oil and Grease (FOG) as Anaerobic Co-Substrate with Slaughterhouse Wastewater: Biomethane Potential, Settling Capacity and Microbial Dynamics. *Chemosphere* **2020**, *259*, 127474. [CrossRef]
14. Pastore, C.; Pagano, M.; Lopez, A.; Mininni, G.; Mascolo, G. Fat, Oil and Grease Waste from Municipal Wastewater: Characterization, Activation and Sustainable Conversion into Biofuel. *Water Sci. Technol.* **2015**, *71*, 1151–1157. [CrossRef] [PubMed]
15. Amha, Y.M.; Sinha, P.; Lagman, J.; Gregori, M.; Smith, A.L. Elucidating Microbial Community Adaptation to Anaerobic Co-Digestion of Fats, Oils, and Grease and Food Waste. *Water Res.* **2017**, *123*, 277–289. [CrossRef] [PubMed]
16. Bratina, B.; Šorgo, A.; Kramberger, J.; Ajdnik, U.; Zemljič, L.F.; Ekart, J.; Šafarič, R. From Municipal/Industrial Wastewater Sludge and FOG to Fertilizer: A Proposal for Economic Sustainable Sludge Management. *J. Environ. Manag.* **2016**, *183*, 1009–1025. [CrossRef] [PubMed]
17. Cheng, T.; Dairi, A.; Harrou, F.; Sun, Y.; Leiknes, T. Monitoring Influent Conditions of Wastewater Treatment Plants by Nonlinear Data-Based Techniques. *IEEE Access* **2019**, *7*, 108827–108837. [CrossRef]
18. Cheng, T.; Harrou, F.; Kadri, F.; Sun, Y.; Leiknes, T. Forecasting of Wastewater Treatment Plant Key Features Using Deep Learning-Based Models: A Case Study. *IEEE Access* **2020**, *8*, 184475–184485. [CrossRef]
19. Yuan, X.; Chen, C.; Lei, X.; Yuan, Y.; Muhammad Adnan, R. Monthly Runoff Forecasting Based on LSTM–ALO Model. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 2199–2212. [CrossRef]
20. Adnan, R.M.; Liang, Z.; Parmar, K.S.; Soni, K.; Kisi, O. Modeling Monthly Streamflow in Mountainous Basin by MARS, GMDH-NN and DENFIS Using Hydroclimatic Data. *Neural Comput. Appl.* **2021**, *33*, 2853–2871. [CrossRef]
21. Adnan, R.M.; Liang, Z.; Heddam, S.; Zounemat-Kermani, M.; Kisi, O.; Li, B. Least Square Support Vector Machine and Multivariate Adaptive Regression Splines for Streamflow Prediction in Mountainous Basin Using Hydro-Meteorological Data as Inputs. *J. Hydrol.* **2020**, *586*, 124371. [CrossRef]
22. Adnan, R.M.; Liang, Z.; Trajkovic, S.; Zounemat-Kermani, M.; Li, B.; Kisi, O. Daily Streamflow Prediction Using Optimally Pruned Extreme Learning Machine. *J. Hydrol.* **2019**, *577*, 123981. [CrossRef]
23. Sandoval, M.A.; Salazar, R. Electrochemical Treatment of Slaughterhouse and Dairy Wastewater: Toward Making a Sustainable Process. *Curr. Opin. Electrochem.* **2021**, *26*, 100662. [CrossRef]

24. Nitayapat, N.; Chitprasert, P. Characterisation of FOGs in Grease Trap Waste from the Processing of Chickens in Thailand. *Waste Manag.* **2014**, *34*, 1012–1017. [CrossRef] [PubMed]
25. Williams, T.O.; Gabel, D.; Robillard, D. FOG Waste Receiving and Processing Facility Design Considerations. *Water Pract. Technol.* **2018**, *13*, 164–171. [CrossRef]
26. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. [CrossRef]
27. Jackson, J.E. *A User's Guide to Principal Components*; John Wiley & Sons: Hoboken, NJ, USA, 2005; ISBN 978-0-471-72532-9.
28. Thorndike, R.L. Who Belongs in the Family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
29. Kaufman, L. *Finding Groups in Data*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 1990; ISBN 978-0-471-87876-6.
30. Defays, D. An Efficient Algorithm for a Complete Link Method. *Comput. J.* **1977**, *20*, 364–366. [CrossRef]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
32. Tyralis, H.; Papacharalampous, G.; Langousis, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water* **2019**, *11*, 910. [CrossRef]
33. Torregrossa, D.; Schutz, G.; Cornelissen, A.; Hernández-Sancho, F.; Hansen, J. Energy Saving in WWTP: Daily Benchmarking under Uncertainty and Data Availability Limitations. *Environ. Res.* **2016**, *148*, 330–337. [CrossRef]
34. Verma, A.; Wei, X.; Kusiak, A. Predicting the Total Suspended Solids in Wastewater: A Data-Mining Approach. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1366–1372. [CrossRef]
35. Harrison, J.W.; Lucius, M.A.; Farrell, J.L.; Eichler, L.W.; Relyea, R.A. Prediction of Stream Nitrogen and Phosphorus Concentrations from High-Frequency Sensors Using Random Forests Regression. *Sci. Total Environ.* **2021**, *763*, 143005. [CrossRef]
36. Zhou, P.; Li, Z.; Snowling, S.; Baetz, B.W.; Na, D.; Boyd, G. A Random Forest Model for Inflow Prediction at Wastewater Treatment Plants. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 1781–1792. [CrossRef]
37. Szelag, B.; Gawdzik, A.; Gawdzik, A. Application of Selected Methods of Black Box for Modelling the Settleability Process in Wastewater Treatment Plant. *Ecol. Chem. Eng. S-Chem. I Inz. Ekol. S* **2017**, *24*, 119–127. [CrossRef]
38. Song, M.J.; Choi, S.; Bae, W.B.; Lee, J.; Han, H.; Kim, D.D.; Kwon, M.; Myung, J.; Kim, Y.M.; Yoon, S. Identification of Primary Effecters of N2O Emissions from Full-Scale Biological Nitrogen Removal Systems Using Random Forest Approach. *Water Res.* **2020**, *184*, 116144. [CrossRef] [PubMed]
39. Torregrossa, D.; Leopold, U.; Hernández-Sancho, F.; Hansen, J. Machine Learning for Energy Cost Modelling in Wastewater Treatment Plants. *J. Environ. Manag.* **2018**, *223*, 1061–1067. [CrossRef] [PubMed]
40. Kusiak, A.; Zeng, Y.; Zhang, Z. Modeling and Analysis of Pumps in a Wastewater Treatment Plant: A Data-Mining Approach. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1643–1651. [CrossRef]
41. Dürrenmatt, D.J.; Gujer, W. Data-Driven Modeling Approaches to Support Wastewater Treatment Plant Operation. *Environ. Model. Softw.* **2012**, *30*, 47–56. [CrossRef]
42. Bunce, J.T.; Graham, D.W. A Simple Approach to Predicting the Reliability of Small Wastewater Treatment Plants. *Water* **2019**, *11*, 2397. [CrossRef]
43. Szeląg, B.; Bartkiewicz, L.; Studziński, J.; Barbusiński, K. Evaluation of the Impact of Explanatory Variables on the Accuracy of Prediction of Daily Inflow to the Sewage Treatment Plant by Selected Models Nonlinear. *Arch. Environ. Prot.* **2017**, *43*, 74–81. [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
45. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw. Artic.* **2008**, *28*, 1–26. [CrossRef]
46. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18–22.
47. Wang, Z.; Lai, C.; Chen, X.; Yang, B.; Zhao, S.; Bai, X. Flood Hazard Risk Assessment Model Based on Random Forest. *J. Hydrol.* **2015**, *527*, 1130–1141. [CrossRef]
48. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable Selection Using Random Forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
49. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinform.* **2008**, *9*, 307. [CrossRef] [PubMed]
50. Jiang, G.; Wang, W. Error Estimation Based on Variance Analysis of K-Fold Cross-Validation. *Pattern Recognit.* **2017**, *69*, 94–106. [CrossRef]
51. Hastie, T.; Tibshirani, R.; Friedman, J. Random Forests. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Hastie, T., Tibshirani, R., Friedman, J., Eds.; Springer Series in Statistics; Springer: New York, NY, USA, 2009; pp. 587–604. ISBN 978-0-387-84858-7.
52. Grömping, U. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]
53. Stentoft, P.A.; Munk-Nielsen, T.; Vezzaro, L.; Madsen, H.; Mikkelsen, P.S.; Møller, J.K. Towards Model Predictive Control: Online Predictions of Ammonium and Nitrate Removal by Using a Stochastic ASM. *Water Sci. Technol.* **2018**, *79*, 51–62. [CrossRef]
54. Rouleau, S.; Lessard, P.; Bellefleur, D. Behaviour of a Small Wastewater Treatment Plant during Rain Events. *Can. J. Civ. Eng.* **1997**, *24*, 790–798. [CrossRef]

55. De Oliveira, D.B.C.; Soares, W.d.A.; de Holanda, M.A.C.R. Effects of Rainwater Intrusion on an Activated Sludge Sewer Treatment System. *Rev. Ambiente Água* **2020**, *15*. [CrossRef]
56. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
57. Vapnik, V. The Support Vector Method of Function Estimation. In *Nonlinear Modeling: Advanced Black-Box Techniques*; Suykens, J.A.K., Vandewalle, J., Eds.; Springer: Boston, MA, USA, 1998; pp. 55–85. ISBN 978-1-4615-5703-6.
58. Boulesteix, A.-L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *WIREs Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [CrossRef]
59. Díaz-Uriarte, R.; de Andrés, S.A. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinform.* **2006**, *7*, 3. [CrossRef] [PubMed]
60. Ziegler, A.; König, I.R. Mining Data with Random Forests: Current Options for Real-World Applications. *WIREs Data Min. Knowl. Discov.* **2014**, *4*, 55–63. [CrossRef]
61. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. [CrossRef]

CAPITULO 5. Resultados

Si bien los resultados se encuentran descritos de forma detallada en cada uno de los artículos incluidos en esta tesis, se resumen aquí los aspectos más relevantes de cada uno de ellos.

5.1 Tratamiento de datos

A partir de la base de datos con los patrones filtrados, se realiza la preparación de los patrones que serán utilizados en el modelado. Los primeros análisis realizados buscan reducir la cantidad de variables a introducir en el modelado, además de descubrir datos que, moviéndose en intervalos viables, no son correctos.

La Figura 12 incluye los valores de correlación de todas las variables estudiadas inicialmente. La descripción de las variables se encuentra en el capítulo 3.

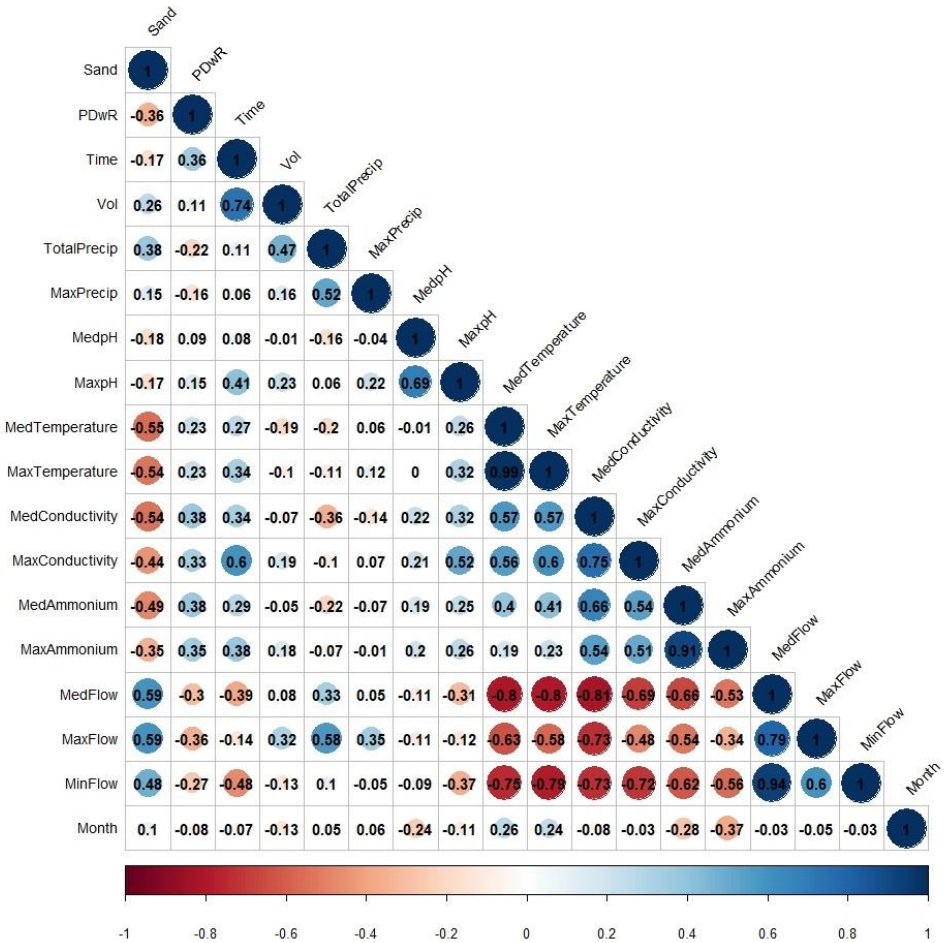


Figura 12: Matriz de correlación de variables

Es fácil deducir la alta correlación que existe entre las distintas variables de flujo así como en el caso de las variables de conductividad. Esto permite detectar errores en cada una de ellas si su distancia crece.

De forma complementaria, también a efectos de reducción de variables, se realiza un Análisis de Componentes Principales (PCA) [119]. El gráfico de la Figura 13 muestra la contribución de las distintas variables a las dimensiones de la proyección del PCA.

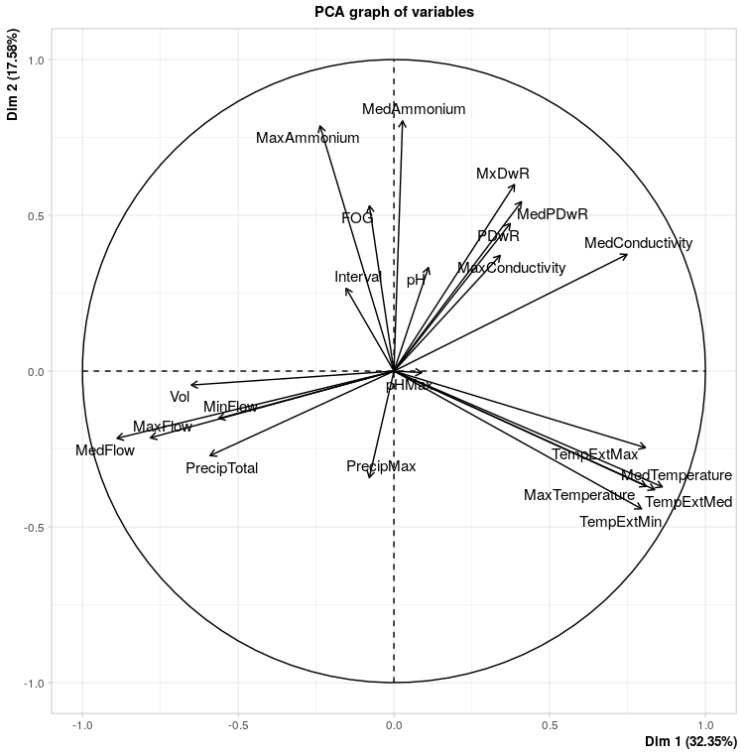


Figura 13: Representación de la proyección bidimensonal de las variables tras aplicar el PCA

Algunos aspectos que se pueden destacar de este gráfico son:

- Dos componentes principales son capaces de explicar el 50% de la información. Sobre esa proyección, se produce una separación representativa con el resto de variables orientadas en 3 ejes equidistantes.

- Las variables de temperatura (temperatura ambiente, temperatura de las aguas residuales,) aparecen agrupadas, como ya se apreciaba en las correlaciones.

- Las variables de caudal están relacionadas con el nivel de precipitación, es decir, a mayor lluvia, mayor caudal de entrada.

- La conductividad está relacionada con el número de días sin lluvia. Esto se debe a que las aguas residuales, tanto urbanas como industriales, no se diluyen con el agua de lluvia durante esos días, recibiéndose agua con más componente industrial.

- La cantidad de FOG (variable grasa) está relacionada con el amonio y, por tanto, es un parámetro importante a tener en cuenta en la modelización. Esta relación puede deberse a los vertidos industriales ya que aportan tanto grasa como nitrógeno. Sin embargo, desde un punto de vista univariante, esta relación no parece tan relevante.

Los datos son, por tanto, proyectados sobre el espacio bidimensional formado por los dos primeros vectores principales (Figura 14). Tras analizar el WCSS (*Within Clusters Summed Squares*) y utilizar el método del codo (*elbow method*, una heurística utilizada para determinar el número de grupos en un conjunto de datos tras aplicar *K-Means* [120]) se determina que el número óptimos de grupos es 4. Para la identificación de los grupos se ha elegido el *Clustering* Jerárquico [121], utilizando como método de aglomeración el *Clustering* de Enlace Completo [122].

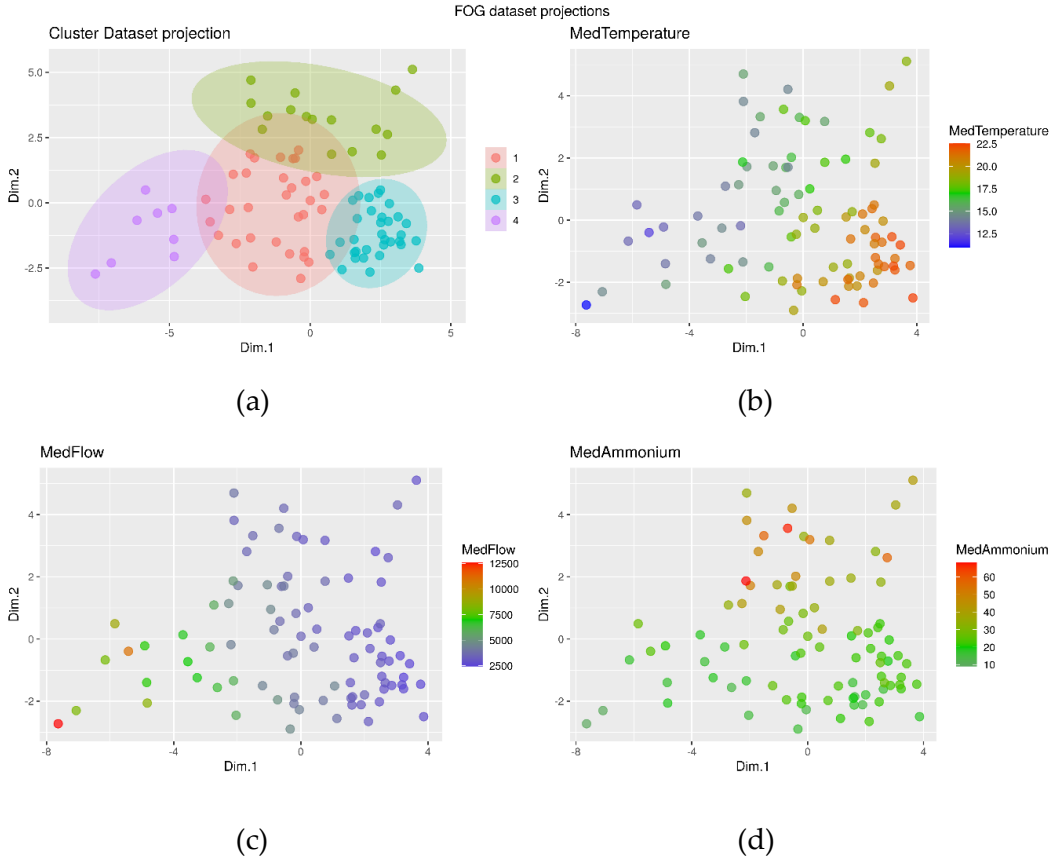


Figura 14: Determinación de clases sobre la proyección PCA

Cada grupo tiene características propias y separables. Los casos con mayor temperatura media se encuentran en la zona del grupo 3 (b). En la figura (c), se puede observar cómo los puntos con los valores más bajos de caudal medio (*MedFlow*) corresponden a los casos de los grupos 2 y 3. Por último, los puntos con los valores medios de amonio más altos (*MedAmmonium*) corresponden al grupo 2 (figura d).

Sobre los valores proyectados con reducción dimensional a los dos vectores principales, se construye un árbol que permita dividir los datos de forma automática en cada una de las categorías.

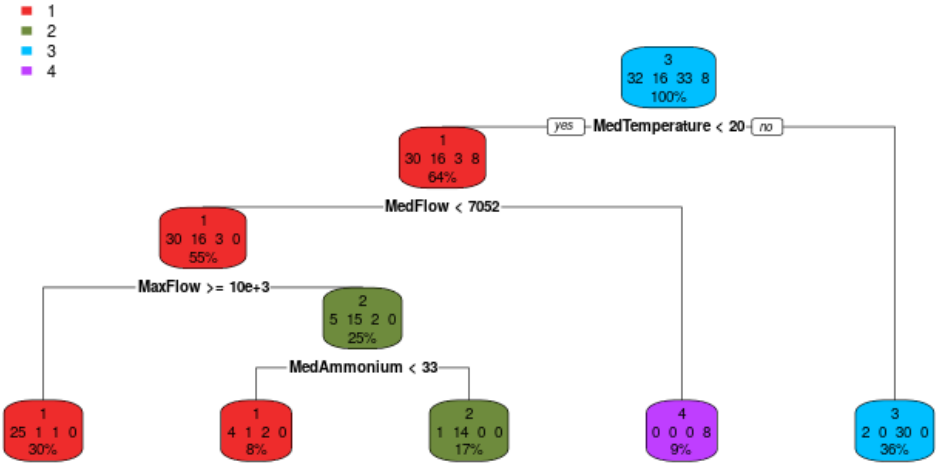


Figura 15: Formación de grupos a partir de la proyección PCA

Se dimensionan así cuatro grupos (Figura 15) con las siguientes características:

- Grupo 1: incluye los casos con un valor de caudal máximo superior a 10.000 m^3/h.
- Grupo 2: se incluye en este grupo una parte de los casos con caudal máximo superior a 10.000 m^3/h y también con amonio medio superior a 33 mg/l.
- El grupo 3 está formado por los datos en los que la temperatura media es superior a 20ºC.
- El grupo 4 se define por un caudal medio superior a 7.052 m^3/h.

El resto del análisis se realizará a partir de estos datos post-procesados, en primer lugar de forma global y posteriormente tomando cada grupo de forma independiente.

5.2 Modelización

Tras la preparación se procede a la modelización con las técnicas explicadas en el capítulo 3. Aunque todas las técnicas fueron utilizadas en los tres casos, se presentan aquí solo los datos de modelo ganador.

1. En el caso de los sólidos de desbaste el menor error se obtiene con el método *SVM* (*Support Vector Machine*), que se ha utilizado con éxito en muchos campos diferentes.
2. En el caso de la predicción de arenas el mejor método de los utilizados es *MARS* (*Multivariate adaptive regression splines*).

3. En el caso de la predicción de grasas el método con el que se obtienen mejores resultados es *RF* (*Random Forest*).

En los siguientes apartados se presentan los tratamientos y resultados para cada uno de los casos.

5.2.1 Predicción de sólidos de desbaste

La evaluación de la idoneidad del modelo de predicción se realizó mediante el coeficiente de determinación (R^2 ajustado) de los valores predichos frente al conjunto de datos reales. En este caso, aunque la precisión del modelo SVM obtenido no es muy alta es la metodología con la que se obtiene menor error, con un valor R^2=0,7093 para los datos de entrenamiento y R^2=0,6869 (Figura 16) con los datos de validación, es suficiente para predecir los cambios de tendencia en la recuperación de sólidos brutos durante las fases de pretratamiento.

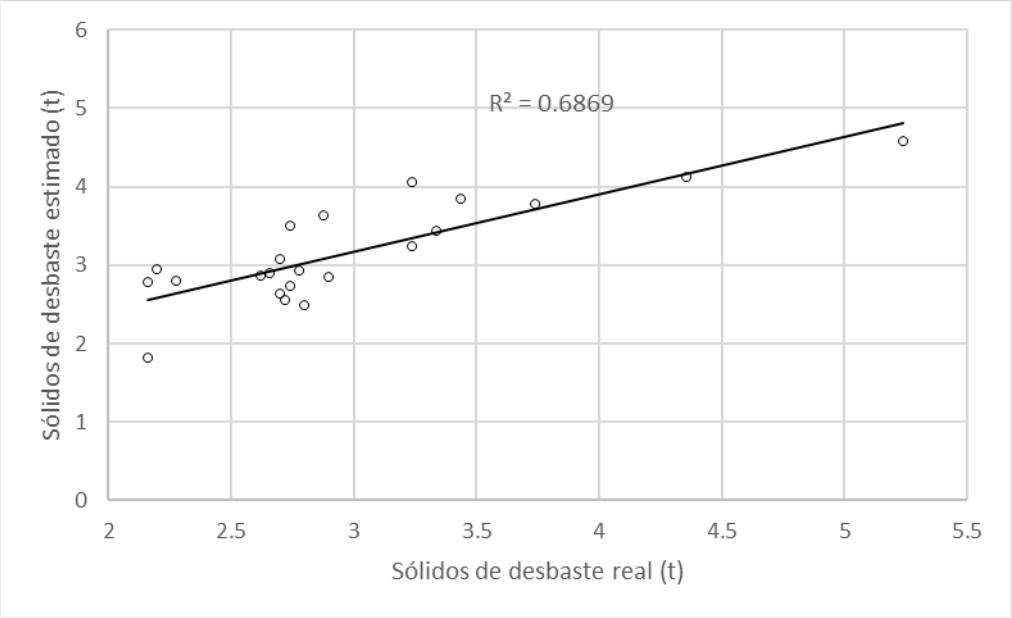


Figura 16: Resultados de predicción del modelo SVM (datos de test)

La Figura 17 muestra los valores de sólidos brutos estimados y reales a lo largo del tiempo, correspondientes al conjunto de datos de validación. Se puede observar que el modelo puede detectar cuándo se producen cambios en el contenido de sólidos brutos que llegan a la planta de tratamiento.

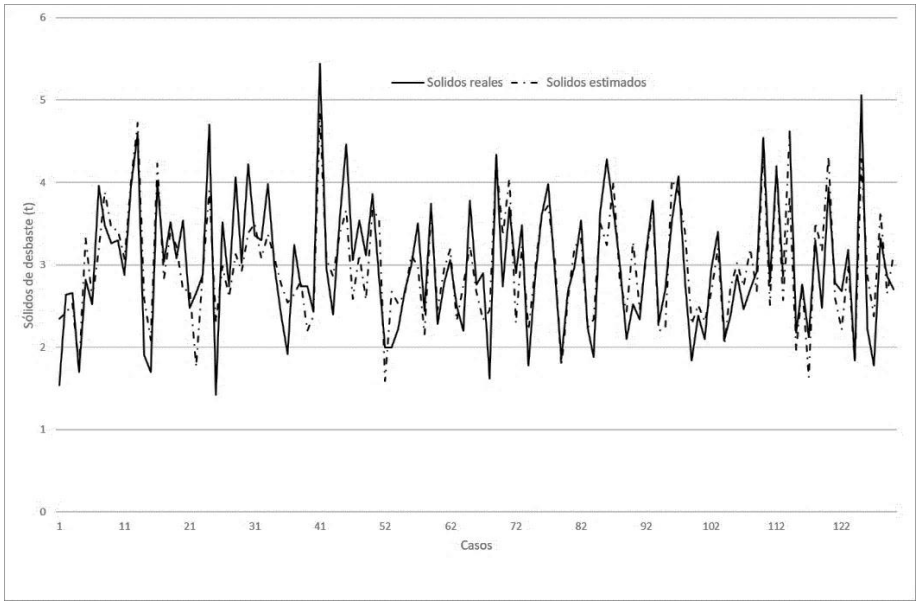


Figura 17: Comparativa entre los sólidos a la entrada de la EDAR y su valor estimado por el modelo

La Tabla 5 incluye las variables más relevantes para el modelo *SVM* a la hora de predecir la llegada de sólidos brutos a la EDAR. Las dos primeras están relacionadas con la componente estacional de esta variable. Así, un aumento de la cantidad de lluvia supone un mayor arrastre de los sólidos depositados en los colectores. Del mismo modo, el pH es un indicador de la cantidad de caudal que llega a la depuradora procedente de actividades industriales. El pH del agua procedente de las actividades domésticas es relativamente constante, mientras que el de las actividades industriales lo altera, unas veces lo sube y otras lo baja. Una de las consecuencias es el llamado "efecto fin de semana". Dado que la EDAR objeto de estudio recibe una parte importante de las aguas residuales de las instalaciones industriales y su actividad disminuye los fines de semana y los días festivos, la reducción del caudal resultante modifica el pH y, por tanto, es relevante para el modelo SVM.

Los parámetros correspondientes a la meteorología como la temperatura, la humedad media y la humedad relativa mínima, relacionadas con la posibilidad de precipitación (*MinMedRH*, *MedRH* y *TempExtMed*) son igualmente consideradas. Otro parámetro importante es el número de días anteriores sin lluvia. Los sólidos brutos se acumulan en el fondo de los colectores en los días sin lluvia y, por tanto, debería ser una variable muy relevante. Sin embargo, su influencia en la estimación del modelo es menor de lo esperado, quizá porque los periodos de tiempo son relativamente largos (media de *PDwR* = 123,4 h) y los chubascos producidos dentro de ese periodo no se detectan.

Tabla 5: Importancia de las variables en el modelo

Variable	%
Week	100
DayYear	98.87
PrecipTotal	93.84
MaxpH	79.2
MinMedRH	76.58
MedRH	63.56
TempExtMed	60.08
PDwR	59.79
MedFlow	54.75

5.2.2 Predicción de arenas

La evaluación de la idoneidad del modelo de predicción, desarrollado con el paquete *Earth* de R [93], se realizó mediante el coeficiente de determinación (R^2 ajustado) entre los valores predichos frente al conjunto de datos reales. En este caso, el modelo con el que se obtiene mejores resultados es el modelo MARS, con resultados obtenidos algo superiores al del caso de los sólidos de desbaste, R^2=0,754 (Figura 3) para los datos de entrenamiento y R^2=0,70 con los datos de validación, suficiente para predecir los cambios de tendencia en la recuperación de la arena durante las fases de pretratamiento.

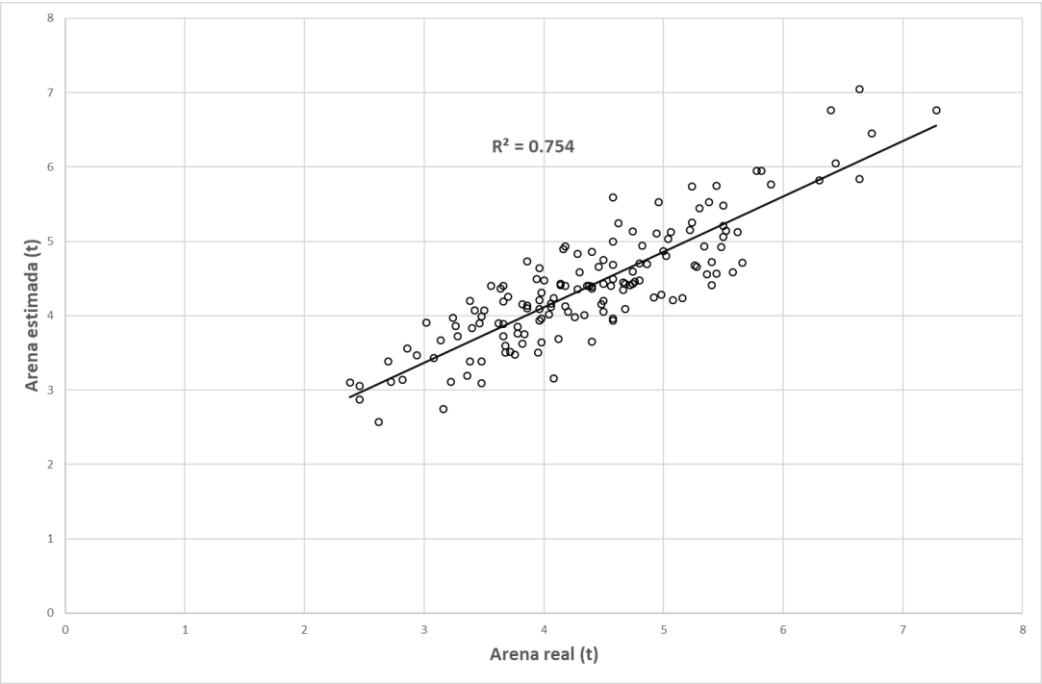


Figura 18: Resultados de predicción del modelo MARS (datos de entrenamiento)

La Figura 18 muestra los valores de arena estimados y reales a lo largo del tiempo, correspondientes al conjunto de datos de validación. Se puede observar que el

modelo es capaz de detectar cuándo se producen cambios en la cantidad de arena que llega a la depuradora con una gran homogeneidad en la calidad del ajuste.

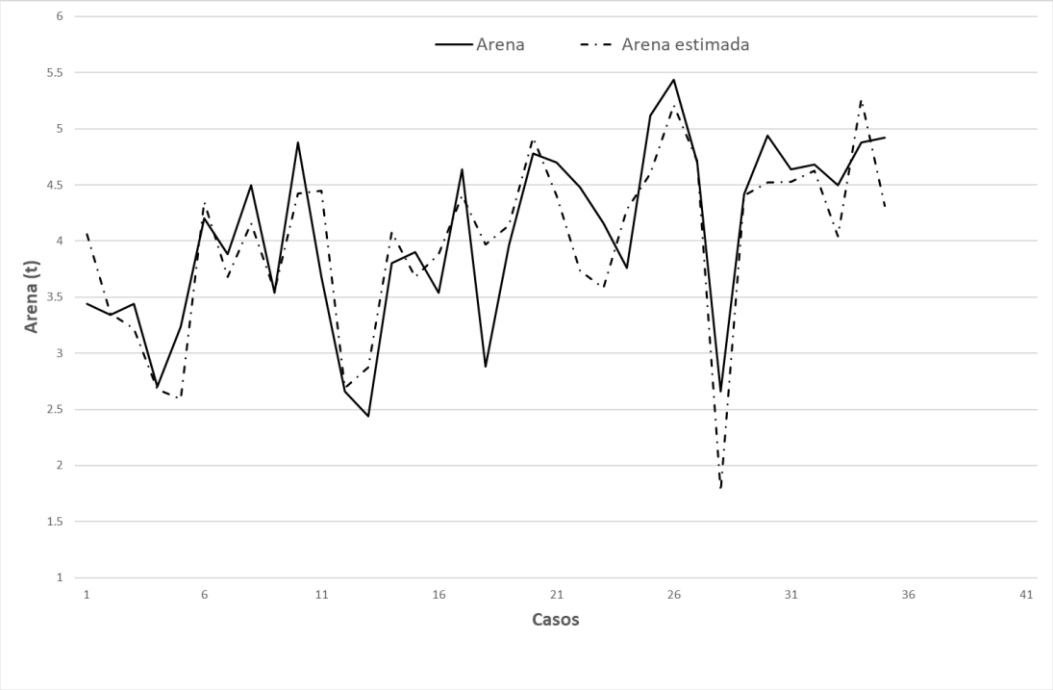


Figura 19: Arena y arena estimada (test)

Las funciones de base del modelo MARS obtenidas y sus coeficientes se muestran en la Tabla 6.

Tabla 6: Función MARS para la predicción de arenas

Elementos	Definición	Coeficiente
B_1	*1*	4.402074
B_2	*max(0, MedFlow - 4784.547)*	0.000155372
B_3	*max(0, 1.4 - TotalPrecip) * max(0, 4784.547 - MedFlow)*	- 0.0004165734
B_4	*max(0, MedpH - 6.841621) * max(0, 4784.547 - MedFlow)*	- 0.0008090167
B_5	*max(0, MedpH - 7.193376) * max(0, Month - 4)*	+ 0.9243512
B_6	*max(0, Month - 4) * max(0, 18.58 - MedAmmonium)*	+ 0.008425636
B_7	*max(0, 24.32 - Time) * max(0, MedFlow - 4784.547) * Month*	- 1.45412e-06
B_8	*max(0, 1.4 - TotalPrecip) * max(0, 4784.547 - MedFlow) * max(0, Month - 4)*	+ 5.940599e-05
B_9	*max(0, MedpH - 7.193376) * max(0, Month - 4) * max(0, MedAmmonium - 12.56)*	- 0.2440937
B_{10}	*max(0, MedpH - 7.193376) * max(0, Month - 4) * max(0, 12.56 - MedAmmonium)*	- 0.2182406
B_{11}	*max(0, MedpH - 7.193376) * max(0, Month - 4) * max(0, MedAmmonium - 16.22)*	+ 0.2774549

La Tabla 7 presenta una evaluación de la importancia de cada variable en el modelo según los siguientes criterios: el número de subconjuntos del modelo en los que se incluye cada variable (*Nsubsets*), *GVC* (validación cruzada generalizada) y *RSS* (sumas de cuadrados residuales). Estos dos últimos parámetros (*GVC* y *RSS*) se presentan en una escala de 0 a 100. El valor *GVC* que se muestra en la Tabla 7 puede entenderse como la capacidad de generalización del modelo y se analiza con los datos de prueba. Del mismo modo, el valor *RSS* representa el error que reduce una determinada variable acumulada en cada uno de los subconjuntos definidos. Por tanto, las variables que provocan una mayor reducción de ese error se consideran más significativas en el modelo.

Tabla 7: Importancia de las variables

Variable	Nsubsets	GCV	RSS
MedFlow	10	100	100
MedpH	8	85.2	85.5
Time	8	55.8	61.3
Month	7	51.4	56.1
MedAmmonium	6	45.1	49.6
TotalPrecipitation	2	24.9	26.6

Como se muestra en la Tabla 7, la principal variable que determina la cantidad de arena que llega a la depuradora es el caudal (*MedFlow*). Como era de esperar, cuanto mayor es el caudal, más arena se recoge, aunque esta relación no es directamente proporcional. El contenido de arena en las aguas residuales es relativamente bajo. Sin embargo, es mayor en las aguas pluviales, como indica la variable *TotalPrecipitación*, que también aparece como significativa. La experiencia indica que, en periodos de lluvia continua, la llegada de arena es constante y no hay efecto de lavado del colector.

El pH, incluido en este caso como valor medio (*MedpH*), es otra variable relevante ya que indica que existe una variación en la composición del agua residual debido a la presencia de agua de lluvia. Mientras que en tiempo seco la composición del agua residual es bastante constante, cuando llueve, el valor de pH registrado se modifica por el efecto de dilución.

La duración de los intervalos de tiempo (*Time*) indica la velocidad de producción de arena, mientras que la variable correspondiente al mes (*Month*) refleja la estacionalidad de la producción de arena. El comportamiento es diferente según la época del año, es decir, en las estaciones lluviosas o en los periodos con mayor probabilidad de lluvia, se produce una mayor cantidad de arena.

Al igual que el pH, el amonio (*MedAmmonium*) es una variable condicionada por la dilución del agua, pero también por la actividad industrial. Es decir, durante los días laborables, cuando hay una mayor actividad industrial, el valor medio del amonio aumenta. Sin embargo, con el pH este efecto no se detecta porque ciertas actividades industriales reducen el valor del pH mientras que otras lo aumentan.

5.2.3 Predicción de grasas

El último modelo desarrollado es el de predicción de grasas. Como en los casos anteriores, la evaluación de la idoneidad del modelo de predicción se realizó mediante el coeficiente de determinación (R^2 ajustado) entre los valores predichos y el conjunto de datos reales. La modelización se ha realizado, como en todos los casos anteriores, mediante las tres técnicas, *SVM*, *MARS* y *Random Forest*, siendo esta última la más exitosa, por lo que se selecciona y se describe. En este caso, la precisión del modelo *Random Forest* es muy alta, R^2=0,98 para los datos de entrenamiento y R^2=0,93 en los datos de validación, suficiente para predecir los cambios de tendencia en la separación de grasas durante las fases de pretratamiento (Figura 20).

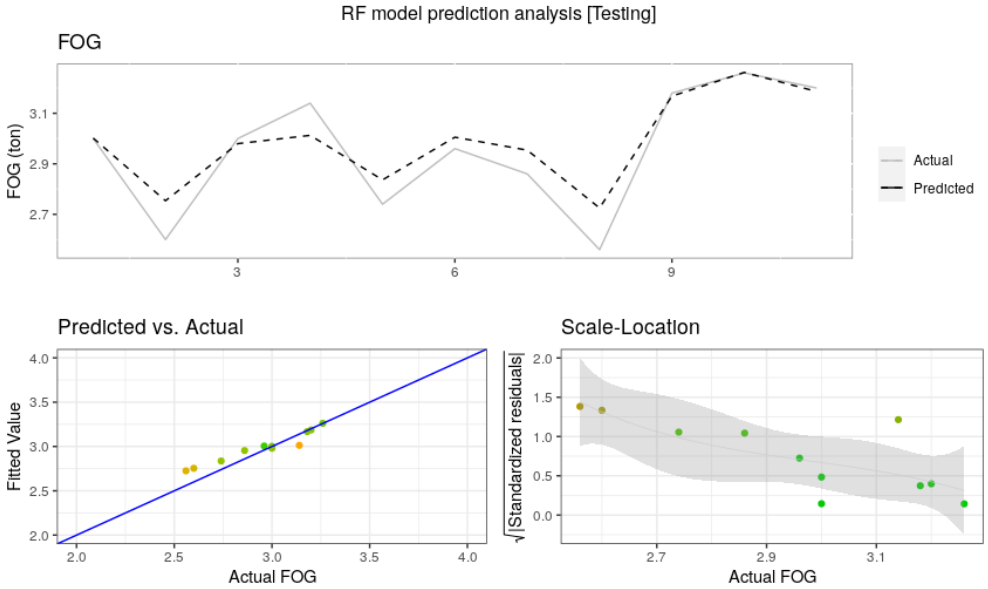


Figura 20:Análisis del modelo de predicción RF (test)

Una de las ventajas más significativas del método *Random Forest* es la evaluación de la importancia de las variables utilizadas en el proceso de entrenamiento, lo que permite conocer cuáles son las variables que aportan mayor información en la predicción. A continuación, se describe la interpretación de la importancia de estas variables en el desarrollo del modelo:

Tabla 8: Importancia de las variables RF

Variable	Importancia estandarizada
MedAmmonium	100,000
MaxAmmonium	81,046
PrecipMax	47,079
MedConductivity	23,024

Variable	Importancia estandarizada
MxDwR	17,963
PDwR	17,678
pH	15,501
TempExtMed	9,171
MedTemperature	6,309
MedFlow	4,074
MedPDwR	0,000

En este caso, las dos variables más relevantes son los valores de amonio medio (*MedAmmonium*) y máximo (*MaxAmmonium*). Esto podría deberse a la gran cantidad de amonio y grasas (*FOG*) que contienen los vertidos de la instalación láctea a la que da servicio la EDAR de Villapérez, tal y como se ha mencionado en la descripción del caso.

La tercera variable más significativa es la precipitación máxima (*PrecipMax*). Una mayor precipitación implica una mayor afluencia a la EDAR, con los *FOG* más disueltos, lo que dificulta su eliminación en el proceso de pretratamiento.

Las aguas residuales urbanas tienen una conductividad constante, por lo que es posible asociar las variaciones y relevancia de esta variable con los vertidos industriales.

La relevancia de las variables relacionadas con el número de días previos sin lluvia (*MxDwR*, *PDwR* y *MedPDwR*) se puede explicar de forma similar a la precipitación, es decir, al haber menos afluencia a tratar, los *FOG* están menos disueltos y es posible eliminarlos en mayor proporción.

Respecto al pH, las aguas residuales urbanas tienen un pH relativamente estable, por lo que las variaciones de este indicador pueden considerarse asociadas a los vertidos industriales.

La temperatura media (*TempExtMed*) proporciona información sobre la situación estacional en el momento del análisis. Una temperatura más elevada facilita la emulsión de los *FOG* y, por tanto, su eliminación es más eficaz.

Por último, la relevancia de la variable caudal medio (*MedFlow*) se explica de la misma manera que la precipitación o el número de días previos sin lluvia mencionados anteriormente.

La combinación de los tres modelos permitirá poder predecir el comportamiento global del pretratamiento adaptando su funcionamiento a la llegada de los residuos.

CAPITULO 6. Conclusiones y líneas de futuro

Aunque el pretratamiento de las aguas residuales en las EDAR ha sido objeto de, relativamente, poca investigación, es una de las etapas más afectadas por los vertidos de aguas no tratadas y de la que depende en gran medida el rendimiento y la durabilidad del resto de la planta. No completar correctamente el proceso de pretratamiento, es decir, no eliminar los residuos más grandes, la arena, los sólidos de desbaste y la grasa, genera necesariamente problemas en el resto de los tratamientos, además de dificultades operativas.

En este trabajo se aborda la predicción de cada uno de estos componentes.

6.1 Conclusiones

De los resultados obtenidos se pueden obtener las siguientes conclusiones.

1. En el caso de la arena el mejor ajuste se consigue mediante un modelo basado en datos que aplica el método *MARS*. La precisión alcanzada en las pruebas de validación (R^2=0,70), similar a la obtenida durante el entrenamiento (R^2=0,74), proporciona una nueva herramienta para una mejor gestión de las EDAR. Disponer de una estimación de la producción de arena facilita la apertura de líneas de pretratamiento en función de la predicción de incrementos significativos en la producción de arena o, por el contrario, el cierre de líneas si se predice un descenso significativo. Asimismo, en lo que respecta al llenado de contenedores, permite podrían programar avisos en el sistema *SCADA* de control de la planta, de forma que se pueda predecir cuándo será necesaria su retirada.
2. El modelo *MARS* obtenido refleja la importancia del caudal medio, pH medio y permite interpretar, en base a la experiencia de la planta, la variación de los valores de entrada como el caudal, el pH, el amoníaco, etc., que indican cambios debidos a la lluvia o a la actividad industrial, por ejemplo.
3. Los sólidos gruesos (toallitas, residuos sanitarios, hisopos, etc.), arrastrados por la lluvia a los sistemas de saneamiento, generan numerosos problemas tanto en los colectores como en las plantas de tratamiento, provocando graves atascos como se describe en múltiples referencias. Reducir esos atascos en los equipos de pretratamiento y evitar el vertido de aguas no tratadas por posibles desbordamientos es el principal objetivo de este trabajo. Cabe destacar que en estudios anteriores a este trabajo no se ha encontrado ninguna otra referencia científica que prediga un parámetro similar con la que comparar los resultados, lo que refleja su novedad, por lo que este resultado es un enorme avance dentro de la predicción de los materiales sólidos de llegada.

4. Se ha desarrollado un modelo *SVM* para predecir el contenido de sólidos brutos presentes en las aguas residuales de desbaste. La precisión alcanzada en la fase de validación es de R^2= 0,6869, ligeramente inferior a la alcanzada en el entrenamiento (R^2= 0,7093), se considera suficiente para detectar cambios de tendencia en la llegada de sólidos de desbaste a la depuradora. Disponer de esta información con antelación permitirá, como en el caso anterior, abrir líneas de pretratamiento cuando sea necesario recibir la llegada de una mayor cantidad de sólidos brutos y disponer de suficientes contenedores para su almacenamiento. Este buen funcionamiento del modelo también se ve refrendado en la comparación de la precisión del modelo con la estimación actual basada en valores medios históricos. Se puede observar que representa una mejora operativa considerable.
5. El modelo final presenta un Error Cuadrático Medio de 0,426 en el entrenamiento mientras que alcanza 0,435 en la prueba de validación. Los mayores errores del modelo se producen en los extremos, es decir, por debajo de 2 toneladas y por encima de 4 toneladas de sólidos brutos, que son valores poco habituales, por lo que no suponen un gran inconveniente y se deben a la escasa presencia de este tipo de patrones en el conjunto de datos de entrenamiento.
6. Al igual que otras fracciones de las aguas residuales urbanas retiradas en la fase de pretratamiento de las plantas de tratamiento de aguas residuales, la optimización de la eliminación de los *FOG* ha recibido relativamente poca atención por parte de los investigadores, más allá de su uso posterior o de su influencia en los procesos de tratamiento de aguas residuales posteriores. Sin embargo, su influencia en estas etapas posteriores del tratamiento de aguas residuales puede ser importante para mejorar tanto el rendimiento global de las EDAR como su operatividad. Con este objetivo, en este trabajo se ha desarrollado un modelo de predicción del contenido de *FOG* en las aguas de entrada a la depuradora.
7. El modelo se basa en los datos recogidos durante más de dos años en la planta de Villapérez (Asturias, España). Los resultados obtenidos, evaluados mediante varios indicadores comunes, reflejan el buen rendimiento del modelo tanto en la fase de entrenamiento (*MSE*=0,037, *MAE*=0,025 y R^2=0,9888) como en la de prueba (*MSE*=0,089, *MAE*=0,066 y R^2= 0,9348). Gracias a las características de la técnica de *RF*, se han interpretado las variables más relevantes utilizadas en el modelo, específicamente el amoníaco o los cambios en las precipitaciones. Como era de esperar, en el caso estudiado destaca la influencia en los cambios del contenido de *FOG* en los vertidos industriales.

Una mejor información permitirá a los operadores una mejor toma de decisiones, permitiendo la optimización de la eliminación de los *FOG* en los procesos de pretratamiento. El resultado de aplicar estos modelos será una reducción del contenido de FOG en los procesos posteriores y una reducción del consumo de energía y de los costes de mantenimiento de la planta.

6.2 Líneas de trabajo futuras

Fruto del trabajo realizado se han detectado algunos trabajos que no han podido ser abordados por ser marginales para el contenido de la tesis pero que, sin duda, podrían ser relevantes. Entre ellos destacan los siguientes:

- Los modelos desarrollados se conciben para ser utilizados en un nivel 2, es decir en ordenadores de proceso. No obstante, sería interesante integrar los 3 modelos obtenidos en un modelo único que se pudiera implementar en el *SCADA* de la planta, es decir, al nivel 1 de automatización. También sería interesante probar los modelos en otro tipo de depuradoras para analizar su validez.
- Otra línea de trabajo muy prometedora es la aplicación de los modelos fluidodinámicos de predicción de movimiento de arenas en otros ámbitos. La valoración del movimiento de arenas en cauces artificiales mediante técnicas de *Machine Learning* puede ser igualmente aplicable a sistemas naturales. Esta línea es, sin duda, muy prometedora y ya ha comenzado a desarrollarse con excelentes resultados [123]–[125].
- De modo similar, los modelos desarrollados podrían ser utilizados aguas arriba. Es conocido el problema de los conductos de saneamiento, con enormes atascos debido a las enormes cantidades de solidos que son arrojados por los usuarios. Sería de utilidad modelar los circuitos hidráulicos de modo que, en función de los caudales, condiciones geométricas y tipo de flujo se pudiese prever la posibilidad de atasco.

CAPITULO 7. Análisis del factor de impacto

Artículo 1

Título:	Sand Content Prediction in Urban WWTPs Using MARS
Autores:	Mateo Pérez V, Mesa Fernández JM, Ortega Fernández F, Morán Palacios
Revista:	Water
Año:	2020
DOI:	10.3390/w12051357
Cita:	Mateo Pérez V, Mesa Fernández JM, Ortega Fernández F, Morán Palacios H. Sand Content Prediction in Urban WWTPs Using MARS. *Water*. 2020; 12(5):1357. https://doi.org/10.3390/w12051357
Factor de Impacto (JCR):	2,544
Campo y Posición:	Q2 (Water resources)
Indicios de Calidad:	La publicación se realiza en una revista específica del sector, con el máximo nivel de especialización. Se encuentra en el primer tercil (31 de 94) de su campo. Su factor de impacto a 5 años es aún superior (2,709). Aunque al ser reciente no tiene aún citas, sólo en la página web de la editorial ya tiene más de 600 descargas, lo que demuestra el interés de la investigación realizada.

Artículo 2

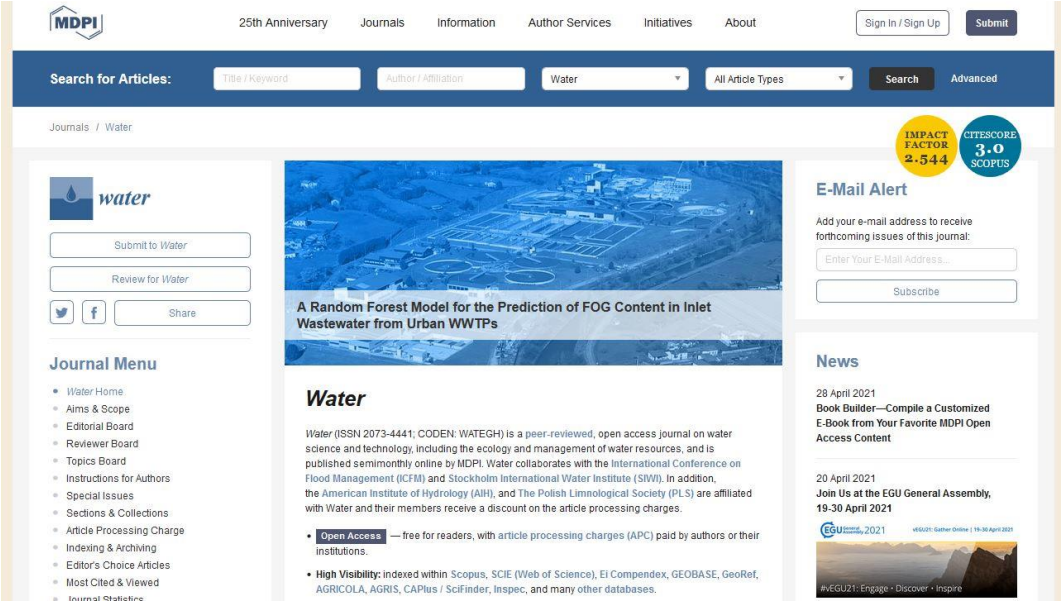
Título:	Gross Solids Content Prediction in Urban WWTPs Using SVM
Autores:	Mateo Pérez, V.; Mesa Fernández, J.M.; Ortega Fernández, F.; Villanueva Balsera, J
Revista:	Water

Año:	2021
DOI:	10.3390/w13040442
Cita:	Mateo Pérez, V.; Mesa Fernández, J.M.; Ortega Fernández, F.; Villanueva Balsera, J. Gross Solids Content Prediction in Urban WWTPs Using SVM. *Water* **2021**, *13*, 442. https://doi.org/10.3390/w13040442
Factor de Impacto (JCR):	2,544
Campo y Posición:	Q2 (Water resources)
Indicios de Calidad:	La publicación se realiza en una revista específica del sector, con el máximo nivel de especialización. Se encuentra en el primer tercil (31 de 94) de su campo. Su factor de impacto a 5 años es aún superior (2,709). Aunque al ser reciente no tiene aún citas, sólo en la página web de la editorial ya tiene más de 350 descargas, lo que demuestra el interés de la investigación realizada.

Artículo 3

Título:	A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs
Autores:	Mateo Pérez, V.; Mesa Fernández, J.M.; Villanueva Balsera, J.; Alonso Álvarez, C
Revista:	Water
Año:	2021
DOI:	10.3390/w13091237
Cita:	Mateo Pérez, V.; Mesa Fernández, J.M.; Villanueva Balsera, J.; Alonso Álvarez, C. A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs. *Water* **2021**, *13*, 1237. https://doi.org/10.3390/w13091237
Factor de Impacto (JCR):	2,544
Campo y Posición:	Q2 (Water resources)
Indicios de Calidad:	La publicación se realiza en una revista específica del sector, con el máximo nivel de especialización. Se

encuentra en el primer tercil (31 de 94) de su campo. Su factor de impacto a 5 años es aún superior (2,709). Aunque al ser reciente no tiene aún citas, sólo en la página web de la editorial ya tiene más de 180 descargas, lo que demuestra el interés de la investigación realizada. Además la publicación ha sido portada de la web de la revista el día 19/05/2021, se adjunta captura.



CAPITULO 8. Bibliografía

[1] V. Mateo Pérez, J. M. Mesa Fernández, F. Ortega Fernández, y J. Villanueva Balsera, «Gross Solids Content Prediction in Urban WWTPs Using SVM», *Water*, vol. 13, n.º 4, p. 442, 2021.

[2] V. Mateo Pérez, J. M. Mesa Fernández, F. Ortega Fernandez, y H. Morán Palacios, «Sand Content Prediction in Urban WWTPs Using MARS», *Water*, vol. 12, n.º 5, p. 1357, 2020.

[3] V. Mateo, F. Ortega Fernández, G. Martínez Huerta, y S. Andrés, «ESTIMACIÓN Y CARACTERIZACIÓN DE SÓLIDOS EN DEPURADORAS DE AGUAS RESIDUALES», 2018.

[4] V. Mateo Pérez, J. M. Mesa Fernández, J. Villanueva Balsera, y C. Alonso Álvarez, «A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs», *Water*, vol. 13, n.º 9, p. 1237, 2021.

[5] L. Metcalf, H. P. Eddy, y G. Tchobanoglous, *Wastewater engineering: treatment, disposal, and reuse*, vol. 4. McGraw-Hill New York, 1991.

[6] H. Klut, «Abwässerreinigung», *Naturwissenschaften*, vol. 1, n.º 35, pp. 831-835 Springer, ISBN: 0028-1042 1913.

[7] Pasteur, L, Koch, P, y Metchnikoff, E, *The founders of modern medicine: Pasteur, Koch, Lister*. Ayer Company Pub., 1971.

[8] Rojas, R., «Sistemas de Tratamiento de Aguas Residuales», presentado en Curso Internacional "GESTIÓN INTEGRAL DE TRATAMIENTO DE AGUAS RESIDUALES", Brasil, 2002. Accedido: feb. 17, 2020. [En línea]. Disponible en: http://files.control-ambiental5.webnode.com.co/200000093-9097e9190c/GESTION%20INTEGRAL%20DEL%20TRATAMIENTO%20AR.pdf

[9] Greenberg, A.E. y Kupka, E., «Tuberculosis transmission by waste waters-A review», *Sew. Ind. WASTES*, vol. 29, n.º 5, pp. 524-537, 1957.

[10] Imhoff, K, «The cleaning of waste water in America and England», *Z. VEREINES Dtsch. INGENIEURE*, vol. 80, pp. 1005-1006, 1936.

[11] Koop, S.H.A y van Leeuwen, C.J, «The challenges of water, waste and climate change in cities», *Environ. Dev. Sustain.*, vol. 19, n.º 2, pp. 385-418, feb. 2016, doi: 10.1007/s10668-016-9760-4.

[12] Metcalf & Eddy, INC., *Wastewater Engineering: Treatment, Disposal, Reuse*, 2ª edición. Nueva York: McGraw-Hill, 1979.

[13] M. Huang, Y. Li, y G. Gu, «Chemical composition of organic matters in domestic wastewater», *Desalination*, vol. 262, n.º 1, pp. 36-42, nov. 2010, doi: 10.1016/j.desal.2010.05.037.

[14] A. Sonune y R. Ghate, «Developments in wastewater treatment methods», *Desalination*, vol. 167, pp. 55-63, 2004.

[15] Koelling, C.P. y Rasaratnam, L., «Waste-Water-Treatmen. Desing and potimization using nonlinear search techniques», *Comput. Oper. Res.*, vol. 13, n.º 1, pp. 69-84, 1986.

[16] Grady, C.P.L, Gujer, W., Henze, M., Marais, G.V., y Matsuo, T., «A model for single-sludge waste-water treatment systems», *WATER Sci. Technol.*, vol. 18, n.º 6, pp. 47-61, 1986.

[17] Kissel, J.C., «Modeling mass-transfer in biological waste-water treatment processes», *WATER Sci. Technol.*, vol. 18, n.º 6, pp. 35-45, 1986.

[18] Henze, M. y Harremoes, P., «Proceedings of an IAWPRC specialized seminar-modeling of biological waste-water treatment - Copenhagen, Denmark, 28-30 august 1985-Preface», vol. 18, n.º 6, pp. R7-R8, 1986.

[19] Tang, C.C., Brill, E.D, y Pfeffer, J.T., «Comprehensive Model of Activated Sludge Wastewater Treatment System», *J. Environ. Eng.-ASCE*, vol. 113, n.º 5, pp. 952-969, 1987.

[20] Olsson, G., Andersson, B., Hellstrom, B.G., Holmstrom, H., Reinius, L.G., y Vopatek, P., «Measurements, Data Analysis and Control Methods in Wastewater Treatment Plants–State of the Art and Future Trends», vol. 21, n.º 10-11, pp. 1333-1345, 1989.

[21] Krovvidy, S. y Wee, W.G, «A knowledge based neural network approach for waste-water treatment system», *IJCNN Int. Jt. Conf. NEURAL Netw. VOLS 1-3*, pp. A327-A332, 1990.

[22] Studzinski, J y Bogdan, L, «Control of wastewater treatment plants using neural networks for decision making and forecasting», *ESS98 - Simul. Technol. Sci. ART*, vol. 10th European Simulation Symposium (ESS 98), pp. 633-637, 1998.

[23] Peguero Camizo, J.C., «Control de una planta de tratamiento de aguas residuales mediante redes neuronales artificiales», Universidad de Extremadura, 2003. Accedido: feb. 17, 2020. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=248

[24] Fang, F. *et al.*, «An integrated dynamic model for simulating a full-scale municipal wastewater treatment plant under fluctuating conditions», *Chem. Eng. J.*, vol. 160, n.º 2, pp. 522-529, 2010.

[25] Joong-Won L., Changwon S., Yoon-Seok T., y Hang-Sik, S., «Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network», *BIOPROCESS Biosyst. Eng.*, vol. 34, pp. 963-973, 2011.

[26] Hamed, M.M., Khalafallah, M.G., y Hassanien, E.A., «Prediction of wastewater treatment plant performance using artificial neural networks.», *Environ. Model. Softw.*, vol. 19, n.º 10, pp. 919-928, 2004.

[27] V. Hernández-Chover, L. Castellet-Viciano, y F. Hernández-Sancho, «Preventive maintenance versus cost of repairs in asset management: An efficiency analysis in wastewater treatment plants», *Process Saf. Environ. Prot.*, vol. 141, pp. 215-221, 2020.

[28] V. Hernández-Chover, Á. Bellver-Domingo, y F. Hernández-Sancho, «The influence of oversizing on maintenance cost in wastewater treatment plants», *Process Saf. Environ. Prot.*, vol. 147, pp. 734-741, 2021.

[29] S. Heo, K. J. Nam, S. Tariq, J. Y. Lim, J. Park, y C. Yoo, «A hybrid machine learning–based multi-objective supervisory control strategy for cost-effective and sustainable wastewater treatment under varying influent conditions», *J. Clean. Prod.*, p. 125853, 2021.

[30] V. M. Ortiz-Martínez, J. Martínez-Frutos, E. Hontoria, F. J. Hernández-Fernández, y J. A. Egea, «Multiplicity of solutions in model-based multiobjective optimization of wastewater treatment plants», *Optim. Eng.*, pp. 1-16, 2020.

[31] J. Pang, S. Yang, L. He, Y. Chen, y N. Ren, «Intelligent control/operational strategies in WWTPs through an integrated Q-learning algorithm with ASM2d-guided reward», *Water*, vol. 11, n.º 5, p. 927, 2019.

[32] L. Benedetti *et al.*, «Modelling and monitoring of integrated urban wastewater systems: review on status and perspectives», *Water Sci. Technol.*, vol. 68, n.º 6, pp. 1203-1215, 2013.

[33] R. Hreiz, M. A. Latifi, y N. Roche, «Optimal design and operation of activated sludge processes: State-of-the-art», *Chem. Eng. J.*, vol. 281, pp. 900-920, 2015.

[34] N. B. Borges, J. R. Campos, y J. M. Pablos, «Characterization of residual sand removed from the grit chambers of a wastewater treatment plant and its use as fine aggregate in the preparation of non-structural concrete», *Water Pract. Technol.*, vol. 10, n.º 1, pp. 164-171, 2015.

[35] L. He, T. Tan, Z. Gao, y L. Fan, «The shock effect of inorganic suspended solids in surface runoff on wastewater treatment plant performance», *Int. J. Environ. Res. Public. Health*, vol. 16, n.º 3, p. 453, 2019.

[36] G. S. do Prado y J. R. Campos, «Determinação da quantidade de areia no esgoto sanitário: metodologia e estudo de caso», *Eng. Sanit. E Ambient.*, vol. 13, n.º 3, pp. 306-312, 2008.

[37] G. S. do Prado y J. R. Campos, «O emprego da análise de imagem na determinação da distribuição de tamanho de partículas da areia presente no esgoto sanitário», *Eng. Sanit. E Ambient.*, vol. 14, n.º 3, pp. 401-409, 2009.

[38] J. M. Sidwick, «The preliminary treatment of wastewater», *J. Chem. Technol. Biotechnol.*, vol. 52, n.º 3, pp. 291-300, 1991.

[39] T. D. Collin, R. Cunningham, M. Q. Asghar, R. Villa, J. MacAdam, y B. Jefferson, «Assessing the potential of enhanced primary clarification to

manage fats, oils and grease (FOG) at wastewater treatment works», *Sci. Total Environ.*, vol. 728, p. 138415, 2020.

[40] R. Roychand, J. Li, S. De Silva, M. Saberian, D. Law, y B. K. Pramanik, «Development of zero cement composite for the protection of concrete sewage pipes from corrosion and fatbergs», *Resour. Conserv. Recycl.*, vol. 164, p. 105166, 2021.

[41] USEPA, *Primer for municipal wastewater treatment systems*. US Environmental Protection Agency Municipal Support, Division Office of …, 2004.

[42] R. M. Ashley, J.-L. Bertrand-Krajewski, T. Hvitved-Jacobsen, y M. Verbanck, *Solids in sewers*. IWA Publishing, 2004.

[43] Brown, D.M., Butler, D., Orman, N.R., y Davies, J.W., «Gross solids transport in small diameter sewers», *Water Sci. Technol.*, vol. 33, n.º 9, pp. 25-30, 1996, doi: 10.1016/0273-1223(96)00366-6.

[44] B. Eren y F. Karadagli, «Physical disintegration of toilet papers in wastewater systems: experimental analysis and mathematical modeling», *Environ. Sci. Technol.*, vol. 46, n.º 5, pp. 2870-2876, 2012.

[45] D. Butler, K. Littlewood, y N. Orman, «A model for the movement of large solids in small sewers», *Water Sci. Technol.*, vol. 52, n.º 5, pp. 69-76, 2005.

[46] C. J. Digman *et al.*, «A model to predict the temporal distribution of gross solids loading in combined sewerage systems», en *Global solutions for urban drainage*, 2002, pp. 1-13.

[47] T. Walski, B. Edwards, E. Helfer, y B. E. Whitman, «Transport of large solids in sewer pipes», *Water Environ. Res.*, vol. 81, n.º 7, pp. 709-714, 2009.

[48] T. Walski, J. Falco, M. McAloon, y B. Whitman, «Transport of large solids in unsteady flow in sewers», *Urban Water J.*, vol. 8, n.º 3, pp. 179-187, 2011.

[49] K. B. Newhart, R. W. Holloway, A. S. Hering, y T. Y. Cath, «Data-driven performance analyses of wastewater treatment plants: A review», *Water Res.*, vol. 157, pp. 498-513, 2019.

[50] C. Screens, «Wastewater Technology Fact Sheet», *U. S. Environ. Prot. Agency*, vol. Screening and Grit Removal, Accedido: feb. 17, 2020. [En línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.170.1830&rep=rep1&type=pdf

[51] I. A. Husain, an F. A. Ma, M. S. Jammi, M. E. Mirghani, Z. B. Zainudin, y A. Hoda, «Problems, control, and treatment of fat, oil, and grease (FOG): a review», *J. Oleo Sci.*, p. ess13182, 2014.

[52] T. Wallace, D. Gibbons, M. O'Dwyer, y T. P. Curran, «International evolution of fat, oil and grease (FOG) waste management–A review», *J. Environ. Manage.*, vol. 187, pp. 424-435, 2017.

[53] S. Arthur y J. Blanc, «Management and Recovery of FOG (fats, oils and greases)», *CREW Proj. CD20136 Available Online Crew Ac UkpublicationsAug 2015*, 2013.

[54] E.-S. Salama, S. Saha, M. B. Kurade, S. Dev, S. W. Chang, y B.-H. Jeon, «Recent trends in anaerobic co-digestion: fat, oil, and grease (FOG) for enhanced biomethanation», *Prog. Energy Combust. Sci.*, vol. 70, pp. 22-42, 2019.

[55] A. E.-F. Abomohra, M. Elsayed, S. Esakkimuthu, M. El-Sheekh, y D. Hanelt, «Potential of fat, oil and grease (FOG) for biodiesel production: A critical review on the recent progress and future perspectives», *Prog. Energy Combust. Sci.*, vol. 81, p. 100868, 2020.

[56] J. Mattsson, A. Hedström, R. M. Ashley, y M. Viklander, «Impacts and managerial implications for sewer systems due to recent changes to inputs in domestic wastewater–a review», *J. Environ. Manage.*, vol. 161, pp. 188-197, 2015.

[57] O. Paraíba, T. Tsoutsos, S. Tournaki, D. Antunes, J. Lino, y E. Manning, «Strategies for optimization of the domestic used cooking oil to biodiesel chain. The European project recoil», en *Proceedings of the 20th European Biomass Conference and Exhibition, Milan, Italy*, 2012, pp. 18-22.

[58] T. Kobayashi, H. Kuramochi, y K.-Q. Xu, «Variable oil properties and biomethane production of grease trap waste derived from different resources», *Int. Biodeterior. Biodegrad.*, vol. 119, pp. 273-281, 2017.

[59] E. B. I. Association, «Transformation of used cooking oil into biodiesel: From waste to resource», *UCO Biodiesel*, vol. 2030, 2015.

[60] H. K. Khuntia, N. Janardhana, y H. N. Chanakya, «Fractionation of FOG (fat, oil, grease), wastewater and particulate solids based on low-temperature solidification and stirring», *J. Water Process Eng.*, vol. 34, p. 101167, 2020.

[61] M. Solé-Bundó, M. Garfí, y I. Ferrer, «Pretreatment and co-digestion of microalgae, sludge and fat oil and grease (FOG) from microalgae-based wastewater treatment plants», *Bioresour. Technol.*, vol. 298, p. 122563, 2020.

[62] J. Hao, L. Francis III, y X. He, «Fat, oil, and grease (FOG) deposits yield higher methane than FOG in anaerobic co-digestion with waste activated sludge», *J. Environ. Manage.*, vol. 268, p. 110708, 2020.

[63] C. Agabo-García, R. Solera, y M. Pérez, «First approaches to valorizate fat, oil and grease (FOG) as anaerobic co-substrate with slaughterhouse wastewater: Biomethane potential, settling capacity and microbial dynamics.», *Chemosphere*, vol. 259, p. 127474, 2020.

[64] C. Pastore, M. Pagano, A. Lopez, G. Mininni, y G. Mascolo, «Fat, oil and grease waste from municipal wastewater: characterization, activation and sustainable conversion into biofuel», *Water Sci. Technol.*, vol. 71, n.º 8, pp. 1151-1157, 2015.

[65] Y. M. Amha, P. Sinha, J. Lagman, M. Gregori, y A. L. Smith, «Elucidating microbial community adaptation to anaerobic co-digestion of fats, oils, and grease and food waste», *Water Res.*, vol. 123, pp. 277-289, 2017.

[66] B. Bratina *et al.*, «From municipal/industrial wastewater sludge and FOG to fertilizer: A proposal for economic sustainable sludge management», *J. Environ. Manage.*, vol. 183, pp. 1009-1025, dic. 2016, doi: 10.1016/j.jenvman.2016.09.063.

[67] H. Cheng, Y. Liu, D. Huang, y B. Liu, «Optimized forecast components-SVM-based fault diagnosis with applications for wastewater treatment», *IEEE Access*, vol. 7, pp. 128534-128543, 2019.

[68] T. Cheng, A. Dairi, F. Harrou, Y. Sun, y T. Leiknes, «Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques», *IEEE Access*, vol. 7, pp. 108827-108837, 2019.

[69] X. Yuan, C. Chen, X. Lei, Y. Yuan, y R. M. Adnan, «Monthly runoff forecasting based on LSTM–ALO model», *Stoch. Environ. Res. Risk Assess.*, vol. 32, n.º 8, pp. 2199-2212, 2018.

[70] R. M. Adnan, Z. Liang, K. S. Parmar, K. Soni, y O. Kisi, «Modeling monthly streamflow in mountainous basin by MARS, GMDH-NN and DENFIS using hydroclimatic data», *Neural Comput. Appl.*, vol. 33, n.º 7, pp. 2853-2871, 2021.

[71] R. M. Adnan, Z. Liang, S. Heddam, M. Zounemat-Kermani, O. Kisi, y B. Li, «Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs», *J. Hydrol.*, vol. 586, p. 124371, 2020.

[72] R. M. Adnan, Z. Liang, S. Trajkovic, M. Zounemat-Kermani, B. Li, y O. Kisi, «Daily streamflow prediction using optimally pruned extreme learning machine», *J. Hydrol.*, vol. 577, p. 123981, 2019.

[73] M. A. Sandoval y R. Salazar, «Electrochemical treatment of slaughterhouse and dairy wastewater: toward making a sustainable process», *Curr. Opin. Electrochem.*, p. 100662, 2020.

[74] N. Nitayapat y P. Chitprasert, «Characterisation of FOGs in grease trap waste from the processing of chickens in Thailand», *Waste Manag.*, vol. 34, n.º 6, pp. 1012-1017, 2014.

[75] T. O. Williams, D. Gabel, y D. Robillard, «FOG Waste receiving and processing facility design considerations», *Water Pract. Technol.*, vol. 13, n.º 1, pp. 164-171, 2018.

[76] C. McCue, «Process models for data mining and predictive analysis», *Data Min. Predict. Anal.*, pp. 51-74, 2015.

[77] V. Vapnik, «The support vector method of function estimation», en *Nonlinear modeling*, Springer, 1998, pp. 55-85.

[78] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[79] S. M. Clarke, J. H. Griebsch, y T. W. Simpson, «Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses», *J. Mech. Des.*, vol. 127, n.º 6, pp. 1077-1087, nov. 2005, doi: 10.1115/1.1897403.

[80] V. K. Chauhan, K. Dahiya, y A. Sharma, «Problem formulations and solvers in linear SVM: a review», *Artif. Intell. Rev.*, vol. 52, n.º 2, pp. 803-855, 2019.

[81] Z. Liu y H. Xu, «Kernel parameter selection for support vector machine classification», *J. Algorithms Comput. Technol.*, vol. 8, n.º 2, pp. 163-177, 2014.

[82] Y. H. Yang, A. Guergachi, y G. Khan, «Support vector machines for environmental informatics: application to modelling the nitrogen removal processes in wastewater treatment systems», *J. Environ. Inform.*, vol. 7, n.º 1, pp. 14-25, 2006.

[83] N. M. Mahmoodi *et al.*, «Activated carbon/metal-organic framework nanocomposite: Preparation and photocatalytic dye degradation mathematical modeling from wastewater by least squares support vector machine», *J. Environ. Manage.*, vol. 233, pp. 660-672, 2019.

[84] A. S. Abobakr Yahya *et al.*, «Water quality prediction model based support vector machine model for Ungauged River catchment under dual scenarios», *Water*, vol. 11, n.º 6, p. 1231, 2019.

[85] M. Najafzadeh y M. Zeinolabedini, «Prognostication of waste water treatment plant performance using efficient soft computing models: an environmental evaluation», *Measurement*, vol. 138, pp. 690-701, 2019.

[86] M. P. Negara, E. Cornelissen, A. K. Geurkink, G. J. W. Euverink, y B. Jayawardhana, «Next generation sequencing analysis of wastewater treatment plant process via support vector regression», *IFAC-Pap.*, vol. 52, n.º 23, pp. 37-42, 2019.

[87] F. Harrou, A. Dairi, Y. Sun, y M. Senouci, «Statistical monitoring of a wastewater treatment plant: A case study», *J. Environ. Manage.*, vol. 223, pp. 807-814, 2018.

[88] C.-W. Hsu, C.-C. Chang, y C.-J. Lin, *A practical guide to support vector classification*. Taipei, 2003. Accedido: feb. 17, 2020. [En línea]. Disponible en: https://www.researchgate.net/profile/Chenghai-Yang/publication/272039161_Evaluating_unsupervised_and_supervised_image_classification_methods_for_mapping_cotton_root_rot/links/55f2c57408ae0960a3897985/Evaluating-unsupervised-and-supervised-image-classification-methods-for-mapping-cotton-root-rot.pdf

[89] C. Campbell y Y. Ying, «Learning with support vector machines», *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 5, n.º 1, pp. 1-95, 2011.

[90] K. Duan, S. S. Keerthi, y A. N. Poo, «Evaluation of simple performance measures for tuning SVM hyperparameters», *Neurocomputing*, vol. 51, pp. 41-59, 2003.

[91] F. Budiman, «SVM-RBF parameters testing optimization using cross validation and grid search to improve multiclass classification», *Научная Визуализация*, vol. 11, n.º 1, pp. 80-90, 2019.

[92] N. Kranjčić, D. Medak, R. Župan, y M. Rezo, «Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns», *Remote Sens.*, vol. 11, n.º 6, p. 655, mar. 2019, doi: 10.3390/rs11060655.

[93] R. C. Team, «R: A language and environment for statistical computing», 2013.

[94] J. H. Friedman, «Multivariate Adaptive Regression Splines», *Ann. Stat.*, vol. 19, n.º 1, pp. 1-67, 1991.

[95] D. H. Li, W. Chen, S. Li, y S. Lou, «Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS)–A case study of Hong Kong», *Energy*, vol. 186, p. 115857, 2019.

[96] G. Nalcaci, A. Özmen, y G. W. Weber, «Long-term load forecasting: models based on MARS, ANN and LR methods», *Cent. Eur. J. Oper. Res.*, vol. 27, n.º 4, pp. 1033-1049, 2019.

[97] X. Zhang, F. Fang, y J. Liu, «Weather-classification-MARS-based photovoltaic power forecasting for energy imbalance market», *IEEE Trans. Ind. Electron.*, vol. 66, n.º 11, pp. 8692-8702, 2019.

[98] Szelag, B., Bartkiewicz, L., Studzinski, J., y Barbusinski, K., «Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear», *Arch. Environ. Prot.*, vol. 43, n.º 3, pp. 74-81, 2017.

[99] O. T. Bakia, E. Arasb, U. O. Akdemirc, y B. Yilmaza, «Biochemical oxygen demand prediction in wastewater treatment plant by using different regression analysis models», *Desalin Water Treat*, vol. 157, pp. 79-89, 2019.

[100] A. Zadorojniy, S. Wasserkrug, S. Zeltyn, y V. Lipets, «Unleashing analytics to reduce costs and improve quality in wastewater treatment», *Inf. J. Appl. Anal.*, vol. 49, n.º 4, pp. 262-268, 2019.

[101] L. Breiman, «Random forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, 2001.

[102] T. G. Dietterich, «An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization», *Mach. Learn.*, vol. 32, pp. 1-22, 1998.

[103] T. G. Dietterich, «Ensemble learning. The handbook of brain theory and neural networks», *Arbib MA*, 2002.

[104] L. Breiman, J. H. Friedman, R. A. Olshen, y C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[105] H. Tyralis, G. Papacharalampous, y A. Langousis, «A brief review of random forests for water scientists and practitioners and their recent history in water resources», *Water*, vol. 11, n.º 5, p. 910, 2019.

[106] D. Torregrossa, G. Schutz, A. Cornelissen, F. Hernández-Sancho, y J. Hansen, «Energy saving in WWTP: daily benchmarking under uncertainty and data availability limitations», *Environ. Res.*, vol. 148, pp. 330-337, 2016.

[107] A. Verma, X. Wei, y A. Kusiak, «Predicting the total suspended solids in wastewater: a data-mining approach», *Eng. Appl. Artif. Intell.*, vol. 26, n.º 4, pp. 1366-1372, 2013.

[108] J. W. Harrison, M. A. Lucius, J. L. Farrell, L. W. Eichler, y R. A. Relyea, «Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression», *Sci. Total Environ.*, vol. 763, p. 143005, 2021.

[109] P. Zhou, Z. Li, S. Snowling, B. W. Baetz, D. Na, y G. Boyd, «A random forest model for inflow prediction at wastewater treatment plants», *Stoch. Environ. Res. Risk Assess.*, vol. 33, n.º 10, pp. 1781-1792, 2019.

[110] B. Szeląg, A. Gawdzik, y A. Gawdzik, «Application of selected methods of black box for modelling the settleability process in wastewater treatment plant», *Ecol. Chem. Eng.*, vol. 24, n.º 1, p. 119, 2017.

[111] M. J. Song *et al.*, «Identification of primary effecters of N2O emissions from full-scale biological nitrogen removal systems using random forest approach», *Water Res.*, vol. 184, p. 116144, 2020.

[112] D. Torregrossa, U. Leopold, F. Hernández-Sancho, y J. Hansen, «Machine learning for energy cost modelling in wastewater treatment plants», *J. Environ. Manage.*, vol. 223, pp. 1061-1067, 2018.

[113] A. Kusiak, Y. Zeng, y Z. Zhang, «Modeling and analysis of pumps in a wastewater treatment plant: A data-mining approach», *Eng. Appl. Artif. Intell.*, vol. 26, n.º 7, pp. 1643-1651, 2013.

[114] D. J. Dürrenmatt y W. Gujer, «Data-driven modeling approaches to support wastewater treatment plant operation», *Environ. Model. Softw.*, vol. 30, pp. 47-56, 2012.

[115] J. T. Bunce y D. W. Graham, «A simple approach to predicting the reliability of small wastewater treatment plants», *Water*, vol. 11, n.º 11, p. 2397, 2019.

[116] M. Kuhn, «Building predictive models in R using the caret package», *J Stat Softw*, vol. 28, n.º 5, pp. 1-26, 2008.

[117] A. Liaw y M. Wiener, «Classification and regression by randomForest», *R News*, vol. 2, n.º 3, pp. 18-22, 2002.

[118] AEMET, «Guia resumida del clima en España (1981-2010)».

[119] J. E. Jackson, *A user's guide to principal components*, vol. 587. John Wiley & Sons, 2005.

[120] R. L. Thorndike, «Who belongs in the family?», *Psychometrika*, vol. 18, n.º 4, pp. 267-276, 1953.

[121] L. Kaufman y P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[122] D. Defays, «An efficient algorithm for a complete link method», *Comput. J.*, vol. 20, n.º 4, pp. 364-366, 1977.

[123] V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y E. P. Vergara-González, «Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery», *Remote Sens.*, vol. 12, n.º 13, p. 2069, jun. 2020, doi: 10.3390/rs12132069.

[124] V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín, «Analysis of the Spatio-Temporal Evolution of Dredging from Satellite Images: A Case Study in the Principality of Asturias (Spain)», *J. Mar. Sci. Eng.*, vol. 9, n.º 3, 2021, doi: 10.3390/jmse9030267.

[125] V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín, «Determination of Water Depth in Ports Using Satellite Data Based on Machine Learning Algorithms», *Energies*, vol. 14, n.º 9, 2021, doi: 10.3390/en14092486.

Anexo I

Anexo I: Comunicaciones y artículos complementarios

Título:	ESTIMACIÓN Y CARACTERIZACIÓN DE SÓLIDOS EN DEPURADORAS DE AGUAS RESIDUALES
Autores:	Mateo, V., Ortega Fernández, F., Martínez Huerta, G., & Andrés, S.
Revista:	CIDIP 2018 (Madrid)
Año:	2018
Cita:	Mateo, V., Ortega Fernández, F., Martínez Huerta, G., & Andrés, S. (2018). ESTIMACIÓN Y CARACTERIZACIÓN DE SÓLIDOS EN DEPURADORAS DE AGUAS RESIDUALES.
	En este artículo se muestra una primera aproximación a la estimación de los sólidos no solubles de llegada a una EDAR. Sirve como un primer trabajo de análisis de las variables que influyen en la predicción de sólidos.

Título:	Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery
Autores:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y E. P. Vergara-González
Revista:	*Remote Sens*
Año:	2020
Cita:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y E. P. Vergara-González, «Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery», *Remote Sens.*, vol. 12, n.º 13, p. 2069, jun. 2020, doi: 10.3390/rs12132069.
	En este artículo lo que se refleja es la búsqueda de un algoritmo que permita relacionar los datos satélite con las batimetrías, buscando obtener estas para la realización del análisis de sedimentos

Título:	Determination of Water Depth in Ports Using Satellite Data Based on Machine Learning Algorithms
Autores:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín
Revista:	*Energies*
Año:	2021
Cita:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín, «Determination of Water Depth in Ports Using Satellite Data Based on Machine Learning Algorithms», *Energies*, vol. 14, n.º 9, 2021, doi: 10.3390/en14092486.
	En este artículo se muestra como se busca el algoritmo que arroje menor error a la hora de obtener las batimetrías.

Título:	Analysis of the Spatio-Temporal Evolution of Dredging from Satellite Images: A Case Study in the Principality of Asturias (Spain)
Autores:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín
Revista:	*JMSE*
Año:	2021
Cita:	V. Mateo-Pérez, M. Corral-Bobadilla, F. Ortega-Fernández, y V. Rodríguez-Montequín, «Analysis of the Spatio-Temporal Evolution of Dredging from Satellite Images: A Case Study in the Principality of Asturias (Spain)», *J. Mar. Sci. Eng.*, vol. 9, n.º 3, 2021, doi: 10.3390/jmse9030267.
	Obtenidos los algoritmos en este artículo se muestra como con estas técnicas se puede llegar a hacer una análisis de sedimentos, llegando a conocer su comportamiento dentro de la dinámica litoral.

04-018

ESTIMATION AND CHARACTERIZATION OF SOLIDS IN WASTEWATER TREATMENT PLANTS

Mateo, Vanesa[(1)]; Ortega Fernández, Francisco[(1)]; Martínez Huerta, Gemma[(1)]; Andrés, Sara[(1)]

[(1)]Universidad de Oviedo

UN estimates that 54% of the population lives in cities. Directive 91/271/EEC obliges urban agglomerations to adequately treat wastewater before pouring it into rivers or bodies of water, for which many treatment plants have been built for this purpose. However, practice shows that together with water one of the critical aspects for the operation of these facilities is the arrival of solid materials. These come not only from the runoff, but also from the materials themselves that the inhabitants discarded by the pipes in an inappropriate way, like wipes, chopsticks, hygienic elements, etc. These elements are extremely negative because they cause problems of operation, jams, surcharges, etc in the treatment plants and conduits as has been recently seen in several critical cases. This work analyses the existence of these problems, characterizes the materials to determine their origin, identifies their effects on the functioning of the sewage treatment plants and studies their arrival flows in order to prepare procedures for acting.

The results are obtained from two treatment plants of different characteristics, one quite urban and the other with an urban-industrial basin.

Keywords: *treatment system; sewage plants; solids; pads;*

ESTIMACIÓN Y CARACTERIZACIÓN DE SÓLIDOS EN DEPURADORAS DE AGUAS RESIDUALES

La ONU estima que un 54% de la población vive en ciudades. La directiva 91/271/CEE obliga a las aglomeraciones urbanas a tratar de forma adecuada las aguas residuales antes de realizar su vertido a los ríos o masas de agua, para lo que se han construido numerosas depuradoras que se han ido optimizando con el tiempo. No obstante, el uso demuestra que junto con el agua uno de los aspectos críticos para el funcionamiento de estas instalaciones es la llegada de materias sólidas. Estas proceden, no solo de las escorrentías, sino también de los propios materiales que los habitantes desechan por las cañerías de forma inapropiada, como las toallitas, palillos, elementos higiénicos, etc. Estos elementos son extremadamente problemáticos porque ocasionan problemas de funcionamiento, atascos, sobrecostes, etc en las depuradoras y conducciones tal como se ha visto recientemente en diversos casos críticos. Este trabajo analiza la existencia de estos problemas, caracteriza los materiales para determinar su origen, identifica sus efectos sobre el funcionamiento de las depuradoras y estudia sus flujos de llegada con el fin de preparar procedimientos de actuación.

Los resultados se obtienen de dos depuradoras de diferentes características, una absolutamente urbana y la otra con una cuenca urbano-industrial.

Palabras clave: *Depuración; EDAR; sólidos; toallitas*

Correspondencia: Francisco Ortega; fran@api.uniovi.es

1. Introducción

Aunque la captación y drenaje de agua pluviales data de tiempos antiguos, la recogida de aguas residuales no aparece como tal hasta principios del siglo XIX y su tratamiento hasta finales del siglo XIX y principios del XX. El desarrollo de la teoría del germen a cargo de Koch y Pasteur (Pasteur, L et al., 1971), en la segunda mitad del siglo XIX, marcó el inicio de una nueva era en el campo del saneamiento. Hasta ese momento se había profundizado poco en la relación entre contaminación y enfermedades, y no se había aplicado al tratamiento de aguas la bacteriología, disciplina entonces en sus inicios (Metcalf & Eddy, INC., 1995). Las aguas se vertían directamente a los ríos sin tratar o se aplicaban directamente al terreno, los brotes de enfermedades, como un brote de cólera producido en Inglaterra en la segunda mitad del siglo XIX, hace que se empiece a avanzar en el tratamiento de las aguas residuales (Rojas, R., 2002).

A la hora de diseñar un tratamiento del agua residual es necesario estimar los caudales de diseño y la composición del agua residual que se separa en dos componentes: aguas negras y aguas pluviales.

Para la estimación de las aguas pluviales se suele utilizar el método racional, aplicado desde hace más de 100 años hasta tiempos recientes (incluso en determinados diseños como en los drenajes de carretas todavía se encuentra en uso). Este método se basa en los principios de hidrología para determinar el caudal de la cuenca que puede llegar a los colectores teniendo en cuenta las características meteorológicas de la zona, de acuerdo a la siguiente ecuación:

$$Q = \frac{C \times i \times A}{360} \tag{1}$$

Siendo,

Q: caudal de diseño, correspondiente al periodo de diseño seleccionado (m3/s)

C: coeficiente de escorrentía

i: intensidad de la lluvia de diseño (mm/h)

A: área de la cuenca (ha)

Desafortunadamente, este sistema no refleja de forma suficientemente adecuada la Predicción de valores de entrada

En cuanto al caudal de agua residual negra que llega a una depuradora, normalmente este se determina a partir del consumo de agua potable de la población cuya agua residual se quiere tratar, aplicándole un coeficiente de pérdidas, generalmente alrededor de 0,8 (Hernandez Muñoz, A., 2001). Debido a que no siempre se dispone de datos de consumo de agua potable se aplican tablas tipo como las de (Metcalf & Eddy, INC., 1995), en la que se incorporan los caudales en función de los tipos de negocio y viviendas.

Del mismo modo, para estimar la composición del agua residual se utilizan tablas de características tipo en función de la distribución de la población como las de (Rojas, R., 2002). El objetivo tradicionalmente es determinar el caudal y las cargas a tratar, caracterizadas como demanda química y biológica de oxígeno. Su determinación es básica para garantizar un funcionamiento correcto de la instalación. A estos elementos tradicionales se debe añadir la presencia de sólidos, tanto arenas como otros sólidos, que afectan muy negativamente al proceso y deben ser previstos.

Desafortunadamente esta aproximación introduce mucha incertidumbre a la hora de

determinar caudales y cargas, puesto que se utilizan tablas globales que no tienen en consideración las características de la zona, la posible componente industrial, posibles infiltraciones, características específicas de la población, zona geográfica, nivel de desarrollo, etc.

2. Estimación de caudales y cargas

Las instalaciones de depuración de aguas recogen cuencas de gran dimensión y diversidad, en las que se mezclan aguas que proceden de diversas escorrentías, aguas de tipo urbano, comercial e industrial. Esta diversidad, la constante evolución de las redes y su dispersión no permiten calcular de modo real las capacidades de cuenca y realizar los estudios hidrológicos sin enormes esfuerzos que deberían ser continuados en el tiempo. Por ello, desde hace tiempo se planteó la necesidad de modelar cada caso a partir de datos reales.

La disponibilidad de datos derivada de la mayor automatización con llevó el inicio de la modelización que busca predecir los parámetros de entrada en una depuradora (Gadzala, Z and Scarlik, K, 1983). Los estudios iniciales se reducían en general en el análisis de tendencias a partir de históricos de datos amplios (Jeng-gang L., 1998).

La tendencia cambia con la aparición de las redes neuronales artificiales. Una red neuronal artificial es una red de muchos procesadores simples (nodos), cada uno con una pequeña memoria local. Los nodos están conectados por canales de comunicación (conexiones) que generalmente contienen datos numéricos codificados de varias maneras. Los nodos operan sobre sus propios datos y entradas que reciben de las conexiones (Warren, S, 2005).

Dentro del campo de las depuradoras, las primeras simulaciones con redes neuronales artificiales se dirigieron a la verificación de su aplicabilidad en sus ecuaciones básicas de funcionamiento, usando el modelo matemático tradicional utilizado en el tratamiento de las aguas residuales (ecuaciones incluidas en el apartado anterior) (Studzinski, J, 1998). Los modelos fueron haciéndose progresivamente más complejos, como en el caso de Peguero que, en su tesis doctoral (Peguero Camizo, J.C., 2003), desarrolla un modelo de predicción de funcionamiento de una depuradora basado en redes neuronales artificiales, mediante la simulación del comportamiento del reactor biológico en función de la variabilidad de los parámetros de entrada y actuando sobre la recirculación de fangos.

De la misma forma, (Joong-Won L. et al., 2011) desarrolla 4 modelos de comportamiento de DBO, DQO, SS y nitrógeno mediante redes neuronales combinadas con algoritmos genéticos.

Pronto el éxito de la técnica dio lugar a su aplicación a modelos predictivos, como el desarrollado por El-DIn (El-Din, AG, 2002) para predecir el caudal de llegada a una depuradora en tiempo de lluvia a partir de los datos de precipitación. (Wei, X., 2015) buscan la relación entre la climatología y el caudal que llega a planta mediante ANN, consiguiendo un modelo con una precisión bastante alta pero en una serie temporal corta (300 minutos).

De la misma forma, (Bartkiewicz, L. et al., 2016) estudian la predicción de caudal de llega a planta a partir del registro de precipitaciones mediante redes neuronales artificiales, utilizan las precipitaciones y el caudal de entrada a planta durante 3 años. Para evaluar el impacto de las variables exógenas en la calidad del modelo, se aplicó el método de regresión logística a partir de la tasa de precipitación y el flujo diario de aguas residuales, que se retrasaron adecuadamente en relación con los valores de flujo de entrada pronosticados.

(Bing, Q. and Bang, L., 2015) añaden a la predicción del caudal la caracterización de sus propiedades, estudiando la predicción de DQO que llega a una depuradora mediante redes neuronales artificiales, comparando los resultados con los sistemas estándar de modelado.

(Szelag, B. et al., 2016) desarrollaron modelos de predicción de parámetros a partir de datos

de tasa de flujo utilizando algoritmos de árbol tipo Random Forest. Posteriormente los mismos autores (Szelag, B. et al., 2017) comparan estos resultados con los obtenidos mediante técnicas tradicionales de regresión lineal. Utilizan métodos SVM, Random Forest, K-NN y sistemas de regresión de Kernel.

La existencia de todas estas publicaciones muestra la idoneidad del uso de técnicas de tratamiento de datos, Data Mining o Big Data en la estimación tanto del caudal de entrada a la planta como de las características más relevantes de este desde el punto de vista de diseño biológico, es decir, DBO5, nitrógeno, fósforo, sólidos en suspensión, etc.

Sin embargo, el funcionamiento de una planta no pasa exclusivamente por la determinación de las características biológicas y químicas de su caudal líquido de entrada. Existen multitud de referencias en los últimos años que dejan en evidencia un problema creciente de enormes consecuencias económicas y medioambientales: los sólidos transportados en el flujo de entrada.

Los sólidos de desbaste (toallitas, residuos sanitarios, bastoncillos, etc.), arrojados por la población en sus saneamientos, generan problemas tanto en las depuradoras como en los propios colectores, llegando a provocar severos atascos tal como se muestra en múltiples referencias. Su llegada, dadas sus características, no es directamente proporcional a la precipitación. Por otra parte, las arenas depositadas en la entrada requieren un control, limpieza y retirada que conlleva un esfuerzo operativo que debe ser previsto.

Los estudios realizados hasta el momento sobre el tema se refieren a su degradación durante el transporte por los colectores, concretamente en el único elemento parcialmente degradable: el papel higiénico (Beytullah E., 2012). También se ha desarrollado una modelización del movimiento de los sólidos gruesos o de desbaste en colectores, aunque centrado en los elementos de mayor peso (Peen, R et al., 2014). No existen referencias a la determinación de los caudales de sólidos esperados a la entrada de la depuradora, a pesar de su importancia. En este trabajo se realiza una aproximación a uno de estos componentes sólidos: las arenas.

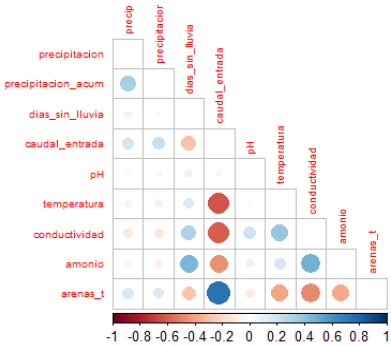
3. Caudal de arenas

Con el fin de determinar la evolución de las arenas en la entrada de la depuradora se toman datos de dos depuradoras, una de carácter básicamente urbano y otra que tiene un fuerte componente industrial. Los datos se muestrean cada 9 minutos durante un año completo y se complementan con variables meteorológicas.

A partir de esta información se pretende predecir la cantidad de arena que llegue en cada momento con la precisión suficiente como para optimizar el funcionamiento de la instalación, previendo la puesta en marcha de dobles líneas y destinando al personal necesario en función de las cantidades previstas.

Entre las 26 variables detectadas se seleccionan aquellas que se consideran más relevantes, incluyendo la precipitación acumulada, el número de días sin lluvia, el caudal de entrada, la temperatura o la conductividad y pH del agua. Un análisis de correlaciones muestra la enorme relación existente entre el caudal de entrada y la arena que, en primera aproximación puede permitir asumirlos como proporcionales, aunque el ajuste fino del sistema pueda requerir correcciones posteriores.

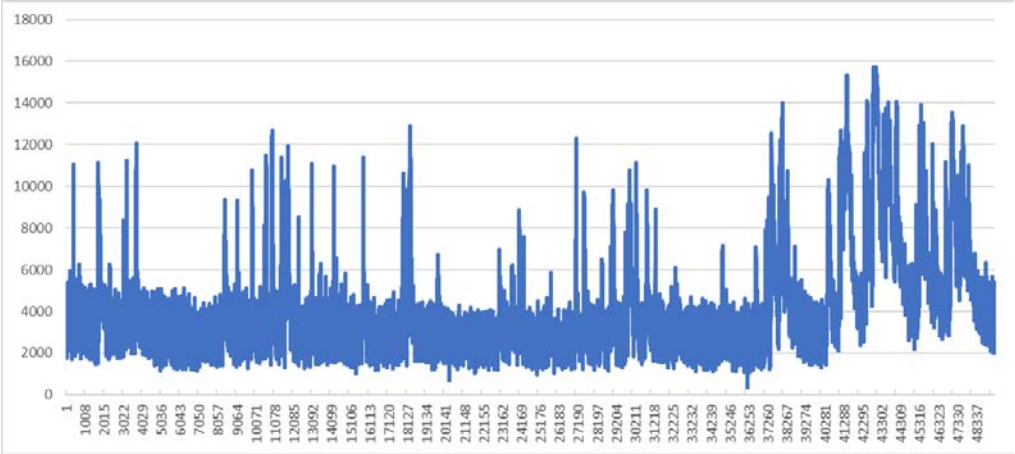
Figura 1: Correlación entre variables implicadas en la depuradora.



Por tanto, la predicción de la cantidad de arena se deducirá directamente de la predicción del caudal de llegada.

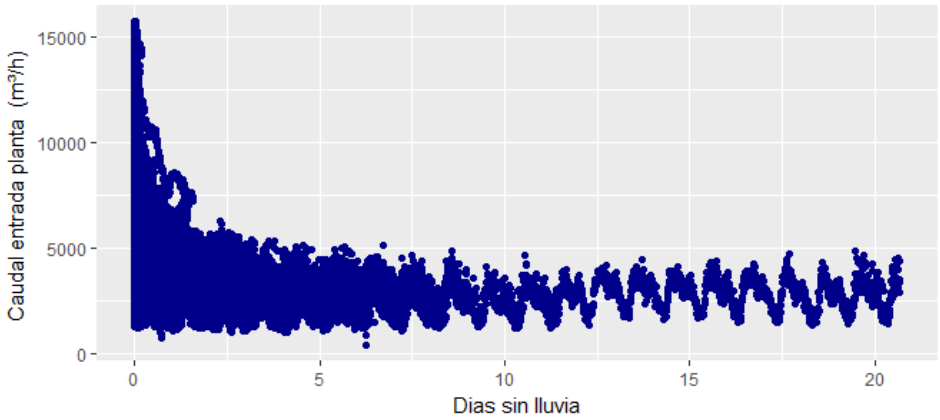
El comportamiento del caudal frente al tiempo es el siguiente:

Figura 2: Evolución del caudal frente al tiempo.



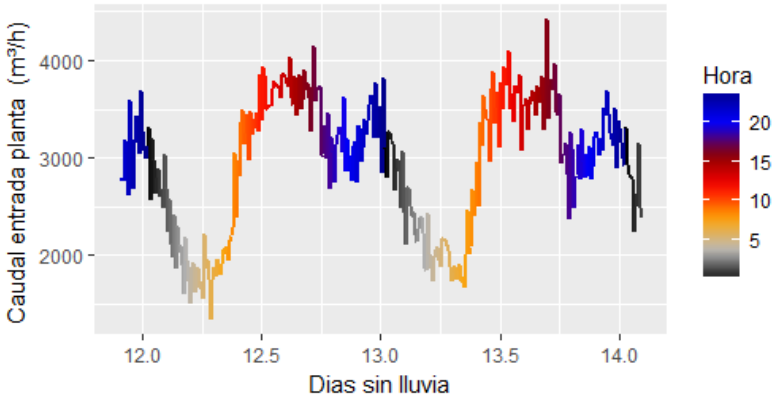
No se aprecia en sí misma una serie temporal que parezca introducir componente cíclica. Sin embargo, la introducción de una nueva variable, denominada días sin lluvia, permite observar efectos importantes.

Figura 3: Caudal de entrada en la planta frente a los días sin lluvia.



Es fácil distinguir en ella una zona con un comportamiento repetitivo y un caudal relativamente estable que supone el caudal de fondo procedente de los días sin aportación de pluvial. Estudiando la zona más estable, se puede observar esa ciclicidad que resulta ser diaria, con bajos caudales nocturnos y la existencia de dos picos máximos, uno situado en el mediodía y el otro en el anochecer.

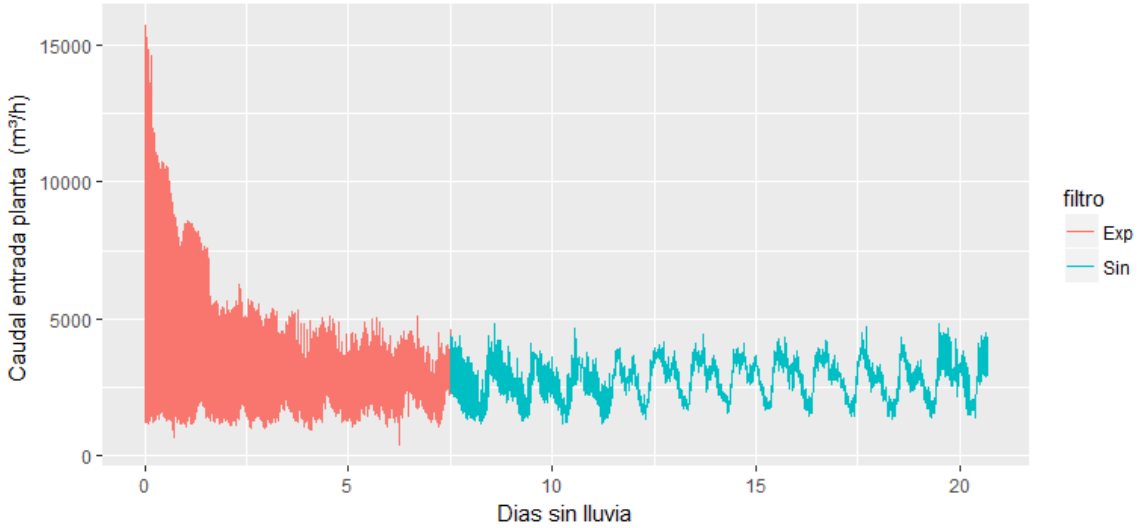
Figura 4: Datos diarios.



Esta curva se repite fundamentalmente en datos posteriores al día 7, siendo mucho menos perceptible al principio, asumiendo que ese es el punto de estabilidad. Por tanto, se decide construir un modelo por tramos, con dos zonas y dos funciones combinadas:

- Una exponencial que represente la fuerte subida en los días de lluvia y
- Una sinusoidal que incorpora los flujos diarios

Figura 5: Zona sobre los que se realiza la regresión sinusoidal.



Los días de lluvia mayores de 7,5 reflejan la forma del caudal en función de la hora, independientemente de mes, y los días de lluvia.

Las ondas sinusoidales con amplitud A y fase φ, A sin(x+ φ), se pueden escribir como la forma lineal:

$$a \sin(x) + b \cos(x) \tag{2}$$

donde a y b son tales que

$$A = \sqrt{a^2 + b^2} \tag{3}$$

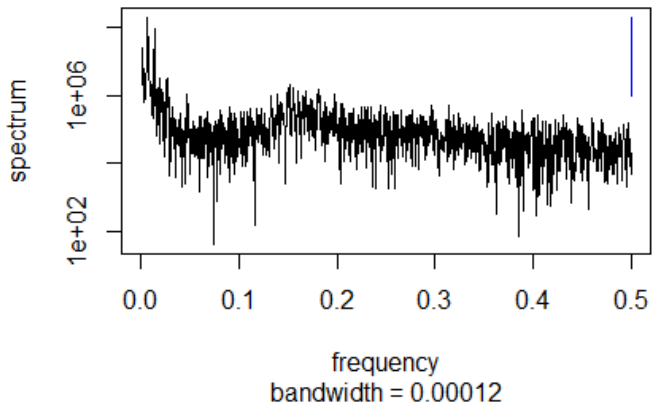
$$\sin \varphi = \frac{b}{\sqrt{a^2 + b^2}} \tag{4}$$

Se puede ver que son equivalentes dado que

$$\begin{aligned} a \sin(x) + b\, cos(x) &= \sqrt{a^2 + b^2}\left(\frac{ba}{\sqrt{a^2 + b^2}}\sin(x) + \frac{b}{\sqrt{a^2 + b^2}}\cos(x)\right) \\ &= A[\sin(x)\cos(\varphi) + \cos(x)\sin\varphi] = A\sin(x + \varphi) \end{aligned} \tag{5}$$

Se realiza una transformación al dominio de la frecuencia con el fin de buscar las frecuencias dominantes obteniendo un periodo de 150 que es casi el número medio de muestreos por día.

Figura 6: Espectro de frecuencias de la serie caudal



La ecuación resultante de la optimización por minimitos cuadrados es:

caudal_entrada = 2826.7

-701.3*sin(2*pi*dias_sin_lluvia) -	132.2*cos(2*pi*dias_sin_lluvia)	
+ 276.3*sin(4*pi*dias_sin_lluvia)	+	344.0*cos(4*pi*dias_sin_lluvia)
-8.4*sin(6*pi*dias_sin_lluvia)	+	7.1*cos(6*pi*dias_sin_lluvia)
-47.9*sin(8*pi*dias_sin_lluvia)	+	82.9*cos(8*pi*dias_sin_lluvia)

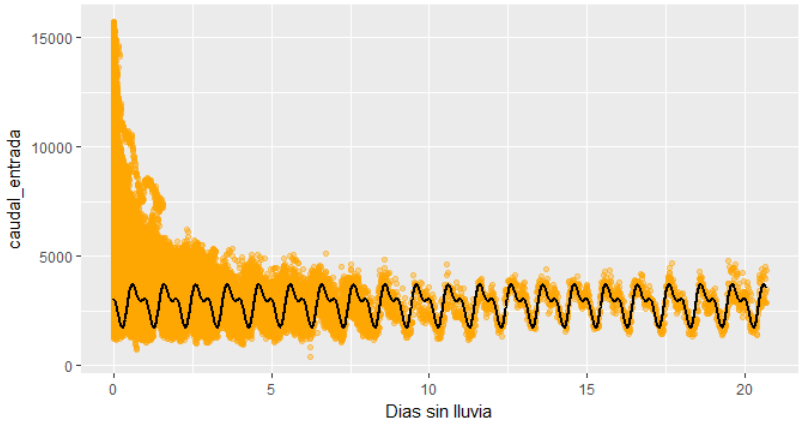
```
Residuals:
    Min      1Q  Median      3Q     Max
-1189.13 -237.70   10.08  229.04 1600.38

Coefficients:
                                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                     2826.683     9.556 295.802  < 2e-16 ***
sin(2 * pi * dias_sin_lluvia) -701.332      13.549 -51.761  < 2e-16 ***
cos(2 * pi * dias_sin_lluvia) -132.168      13.479  -9.806  < 2e-16 ***
sin(4 * pi * dias_sin_lluvia)  276.290      13.483  20.492  < 2e-16 ***
cos(4 * pi * dias_sin_lluvia)  344.021      13.545  25.398  < 2e-16 ***
sin(6 * pi * dias_sin_lluvia)   -8.430      13.515  -0.624 0.532864
cos(6 * pi * dias_sin_lluvia)    7.147      13.511   0.529 0.596924
sin(8 * pi * dias_sin_lluvia)  -47.873      13.509  -3.544 0.000406 ***
cos(8 * pi * dias_sin_lluvia)   82.864      13.508   6.134 1.08e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 381.9 on 1591 degrees of freedom
Multiple R-squared:  0.7151, Adjusted R-squared:  0.7136
F-statistic: 499.1 on 8 and 1591 DF,  p-value: < 2.2e-16
```

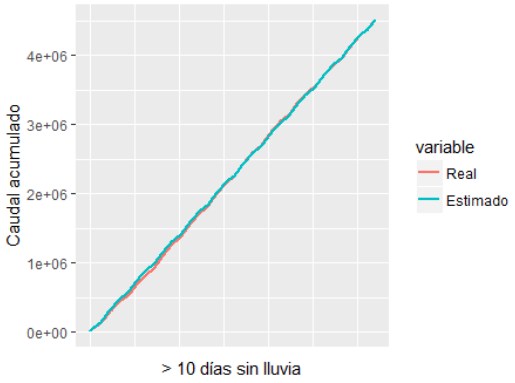
En este caso, el tercer armónico no sería estadísticamente significativo, por lo que se podría eliminar, consiguiendo una mejora del Adjusted R-squared: 0.7139. El comportamiento de la aproximación resulta excelente en la zona aplicada (días sin lluvia superior a 7,5). Por debajo de este valor existen otros factores que hacen que, aunque se deduce un comportamiento similar, la cantidad de caudal sea considerablemente mayor.

Figura 7: Comportamiento de la ecuación sinusoidal en todo el grupo de datos.



El comportamiento del modelo se muestra mediante confrontación del valor acumulado de los caudales real y estimado.

Figura 8: Resultados de la aproximación sinusoidal.



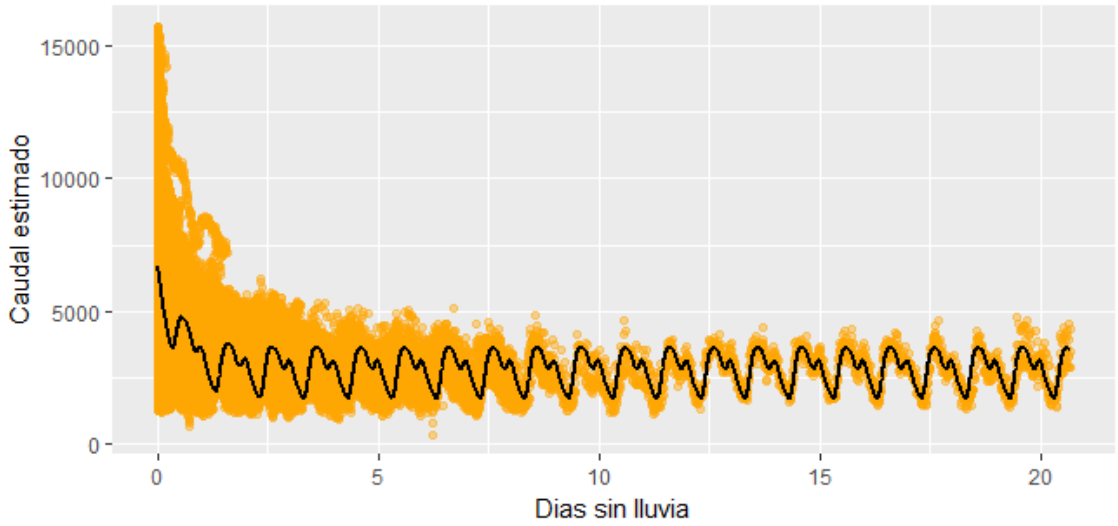
Para representar la zona inicial, se realiza una regresión exponencial.

$$Caudal = a * e^{b*Días\_sin\_lluvia} \tag{6}$$

Aplicando el mismo concepto anterior y una optimización por mínimos cuadrados, la ecuación resultante es:

$$Caudal = 3657 * e^{-2.65*Días\_sin\_lluvia} \tag{7}$$

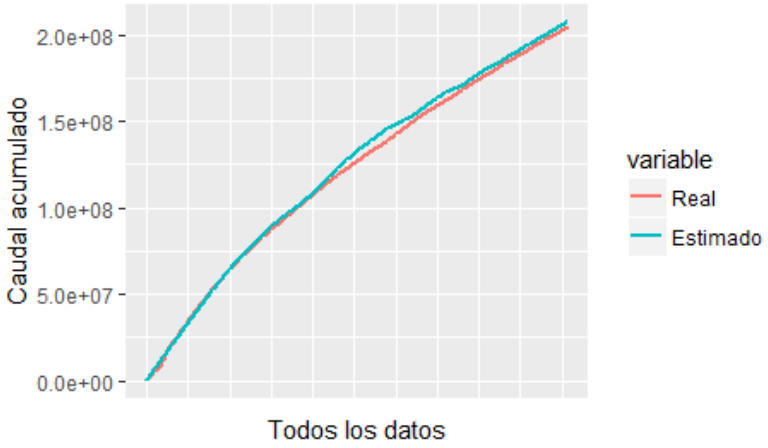
Figura 9: Caudal real frente a estimado con la composición exponencial-sinusoidal.



Aparentemente el resultado muestra que se puede aproximar el caudal a partir de 7 días sin lluvia con una precisión más que suficiente para lo esperado, aunque en la parte inicial la representatividad sea aparentemente menor. Sin embargo, desde un punto de vista

operativo el caudal instantáneo no resulta relevante puesto que las recogidas y tratamientos se realizan siempre en zonas de acumulación. El resultado del modelo con caudales agregados muestra, sin embargo, que los resultados son óptimos si comparamos caudales recogidos, es decir el caudal acumulado hasta ese momento.

Figura 10: Caudal acumulado real y estimado.



Es, por tanto, posible determinar el caudal acumulado con una precisión muy elevada incluso en los casos en los que el efecto del agua pluvial y de escorrentía es importante, cuando el número de días sin lluvia es bajo. En estos últimos la precisión es más limitada porque el caudal viene afectado mucho más por los volúmenes de lluvia precipitados, aspecto que en este modelo no se considera.

Dada la elevada correlación, superior a 0,87, existente entre caudal y arena, es posible también realizar esta predicción de la llegada de arena con una precisión suficiente para lo que requieren las prácticas operativas.

4. Conclusiones

Uno de los problemas crecientes en el ámbito del mantenimiento y operación de las depuradoras de aguas residuales es la presencia de sólidos gruesos no biodegradables. A día de hoy no existen modelos de predicción de la llegada de estos sólidos a las depuradoras. El poder determinar cuándo se van a producir los episodios de afluencia de sólidos antes de que pasen permitiría tomar las medidas necesarias en las depuradoras para minimizar el impacto que generan estos sólidos.

El estado del arte muestra la existencia de modelos de caudal de entrada, pero para fluidos. Se puede además deducir que la tendencia de los modelos predictivos apunta al uso de redes neuronales artificiales, o sistemas de Soft Computing en general, como herramienta para la predicción de parámetros en las depuradoras.

La predicción de las arenas recibidas a la entrada de la depuradora puede, en primera instancia basarse en el caudal. La aproximación sinusoidal-exponencial aporta muy buenos resultados en términos acumulados, suficiente para la planificación de las labores operativas en planta, especialmente en periodos más secos. No obstante, estas predicciones son mejorables introduciendo factores de variación, fundamentalmente precipitación acumulada, día de la semana, mes, etc.

5. Referencias

Bartkiewicz, L., Szelag, B., and Studzinski, J (2016). Impact Assessment of Input Variables and ANN Model Structure on Forecasting Wastewater Inflow into Sewage Treatment Plants. OCHRONA SRODOWISKA *38*, 29–36.

Beytullah E., F.K. (2012). Physical Disintegration of Toilet Papers in Wastewater Systems: Experimental Analysis and Mathematical Modeling. Enviromental Science & Technology 2870–2876.

Bing, Q., and Bang, L. (2015). Prediction of Sewage Wastewater Quality Based on PSO-LIBSVM. PROCEEDINGS OF THE 2015 INTERNATIONAL SYMPOSIUM ON COMPUTERS & INFORMATICS *13*, 280–286.

El-Din, AG, S., DW (2002). A neural network model to predict the wastewater inflow incorporating rainfall events. WATER RESEARCH *36*, 1115–1126.

Gadzala, Z, and Scarlik, K (1983). Forecast for the chemicals demand, necessary for domestic water sewage systems operation in years 1985-1990. Przemysl Chemiczny *62*, 597–600.

Hernandez Muñoz, A. (2001). Depuración y desinfección de aguas residuales (Colegio de Caminos, Canales y Puertos).

Jeng-gang L. (1998). Feasibility study on prediction of properties of municipal solid waste with time series models. JOURNAL OF HAZARDOUS MATERIALS *58*, 47–57.

Joong-Won L., Changwon S., Yoon-Seok T., and Hang-Sik, S. (2011). Sequential modelling of a full-scale wastewater treatment plant using an artificial neural network. BIOPROCESS AND BIOSYSTEMS ENGINEERING *34*, 963–973.

Metcalf & Eddy, INC. (1995). Ingeniería de aguas residuales. Tratamiento, vertido y reutilización. (McGraw-Hill).

Pasteur, L, Koch, P, and Metchnikoff, E (1971). The founders of modern medicine: Pasteur, Koch, Lister.

Peen, R, Shütze, M, and Friendler, E (2014). Assessment of the effects of greywater reuse on gross solids movement in sewer systems. Water Science & Technology *99*, 99–105.

Peguero Camizo, J.C. (2003). Control de una planta de tratamiento de aguas residuales mediante redes neuronales artificiales. Universidad de Extremadura.

Rojas, R. (2002). Sistemas de Tratamiento de Aguas Residuales. (Brasil), p.

Studzinski, J, B., L. (1998). Control of wastewater treatment plants using neural networks for decision making and forecasting. ESS'98 - SIMULATION TECHNOLOGY: SCIENCE AND ART *10th European Simulation Symposium (ESS 98)*, 633–637.

Szelag, B., Bartkiewicz, L., and Studzinski, J. (2016). Black-box Forecasting of Selected Indicator Values for Influent Wastewater Quality in Municipal Treatment Plant. OCHRONA

SRODOWISKA *38*, 39–46.

Szelag, B., Bartkiewicz, L., Studzinski, J., and Barbusinski, K. (2017). Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear. ARCHIVES OF ENVIRONMENTAL PROTECTION *43*, 74–81.

Warren, S (2005). Neural Network FAQ.

Wei, X., K., A. (2015). Short-term prediction of influent flow in wastewater treatment plant. STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT *29*, 241–249.

remote sensing

MDPI

Article

Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery

Vanesa Mateo-Pérez [1], Marina Corral-Bobadilla [2,*], Francisco Ortega-Fernández [1] and Eliseo P. Vergara-González [1]

1 Project Engineering Department, University of Oviedo, 33004 Oviedo, Principality of Asturias, Spain; mateovanesa@uniovi.es (V.M.-P.); fdeasis@uniovi.es (F.O.-F.); vergaraeliseo@uniovi.es (E.P.V.-G.)
2 Department of Mechanical Engineering, University of La Rioja, 26004 Logroño, La Rioja, Spain
* Correspondence: marina.corral@unirioja.es; Tel.: +34-941-299-274

Received: 13 May 2020; Accepted: 25 June 2020; Published: 27 June 2020

Abstract: Knowledge of the free draft of ports is essential for the adequate management of ports. To maintain these drafts, it is necessary to carry out dredging periodically, and to conduct bathymetries using traditional techniques, such as echo sounding. However, an echo sounder is very expensive and its accuracy is subject to weather conditions. Thus, the use of recent advancements in remote sensing techniques provide a better solution for mapping and estimating the evolution of the seabed in these areas. This paper presents a cost-effective and practical method for estimating satellite-derived bathymetry for highly polluted and turbid waters at two different ports in the cities of Luarca and Candás in the Principality of Asturias (Spain). The method involves the use of the support vector machine (SVM) technique and open Sentinel-2 satellite imagery, which the European Space Agency has supplied. Models were compared to the bathymetries that were obtained from the in situ data collected by a single beam echo sounder that the Port Service of the Principality of Asturias provided. The most accurate values of the training and testing dataset in Candás, were R^2 = 0.911 and RMSE = 0.3694 m, and R^2 = 0.8553 and RMSE = 0.4370 m, respectively. The accuracies of the training and testing dataset values in Luarca were R^2 = 0.976 and RMSE = 0.4409 m, and R^2 = 0.9731 and RMSE = 0.4640 m, respectively. The regression analysis results of the training and testing dataset were consistent. The approaches that have been developed in this work may be included in the monitoring of future dredging activities in ports, especially where the water is polluted, muddy and highly turbid.

Keywords: bathymetry; support vector machine (SVM); shallow water; satellite; Sentinel-2

1. Introduction

A bathymetric model is an essential source of information for an understanding of the marine environment. It should be the starting point for any project that is based on marine cartography. Accurate bathymetry data also facilitate the habitat assessment [1,2], classification and detailed representation of the seabed [3] and an understanding of the morphology of the area. Water depth information is also essential for hydrodynamic and wave modelling, sediment transport and environmental exploration, facilitates simulation of the impact of construction and dredging activities, etc. [4]. Similarly, accurate bathymetry data are of utmost importance for the stabilization of beaches and, hence, the security of buildings that are located near the sea. Further, it is essential in scientific research and modeling sea floor relief. These are also necessary for the exploration, exploitation, conservation and administration of natural resources and, especially, for a coastal environment impact assessment and protection. [5,6].

The ports are viewed as buffer zones or protected areas, in which ships can find protection from the action of waves. These configurations of the ports aid the dissipation of the waves, but not the action of the coastal dynamics. The coastal dynamics mobilize the sediments, which are then deposited in the calm water areas of the ports. Consequently, dredging operations are required to empty the navigation channel of these sediments and avoid serious disaster [7]. Dredging recovers the seabed that is currently full of sand and thus facilitates access by ships to the ports by means of the navigation channel. Dredging is also required to ensure that the minimum draft that is necessary for ships to navigate and maneuver within the ports is maintained. It is also necessary for the development of port infrastructure. Bathymetric mapping is used in, and is essential to, the management of such port operations. In any event, whether bathymetric work is used as an aid to navigation or as part of a process or activity in maritime engineering, its high economic significance is clear. Even small variations in vertical measurement would have great economic repercussions if the level of error if not evaluated. However, on some occasions, knowledge of errors is necessary. On other occasions, it is helpful to reduce any errors and, therefore, increase the reliability of the information. Bathymetry studies are conducted by the use of various techniques. Each technique can give a different result depending on the precision that it provides. Conventional techniques, such as airborne-, ground- and ship-borne-based surveying provide very accurate measurements [8]. Among the most commonly used techniques is the use of echo sounders. Techniques that involve vessel-based single beam echo sounding, however, are limited due to problems in accuracy and precision, as well as the difficulty to access shallow coastal waters. Currently, multibeam echo sounders and light detection and ranging (LiDAR) are commonly used for high-resolution bathymetry retrieval in shallow areas [9,10]. Nevertheless, these techniques are better employed in small areas and are limited by high costs [8–12]. These factors cause other techniques, such as remote sensing, to become competitive and attractive methods of providing reliable depth estimates at a much lower cost [13,14].

Remotely sensed technology is considered a low-cost, time-effective and widely adopted solution for satellite-derived bathymetry (SDB), which can be considered as a promising alternative tool to map bottom depths in areas with highly dynamic seabed characteristics [15]. SDB methods can be classified according to a physics-based approach or empirical approach: the first simulates the light that interacts through the water column, and the second develops regressions between spectral radiation and in situ calibration data [16]. Bathymetric information from shallow areas is key to managing coastal environments, however, there is still incomplete and spatially limited coverage, especially in optically shallow areas because the water clarity has a significant and variable impact on SDB accuracy.

Machine learning methods have been used to estimate water depth from remote optical images. One of the initial attempts used a combination of multispectral data and radiometric techniques [17]. With the arrival of Landsat images, bathymetry monitoring methods were improved and applied effectively to optical satellite images [18,19]. The advancement of remote sensing technology has enabled numerous researchers to expand the use of these techniques with improved spectral resolution [20–22]. In recent years, thanks to the technological advances and improvements that satellites are bringing to the study of the marine environment, topographic surveys are conducted by the use of satellites and remote sensing technology. This technology permits the use of high-resolution satellite images to determine depth ranges based on the wavelength of the spectral bands of the image. The main obstacles that are encountered when applying these technologies are the turbidity of the water and the reflectance penetration [23]. The suspended particulate matter, which is the main contributor to turbidity, introduces a confusing reflectance of light that the satellite detects. The waters of different turbidity levels disperse the incoming radiation differently, which implies greater complexity and development problems in the highly dynamic coastal regions and anthropogenic ecosystems of the ports [24,25].

In this study, regular optical satellite images were used. More specifically, the ESA Sentinel-2 constellation (two satellites) was used to obtain bathymetric mapping by the use of a support vector machine (SVM), a machine learning technique. These satellites' images are freely available, have a resolution of 10 to 60 m and a revisit interval of five days. Sentinel-2 has six land monitoring bands,

each of which can be compared to Landsat-8. In addition, the satellite has three other bands, thereby covering the red-edge spectrum [26]. At present, these data that this satellite provides and the use of advanced computational techniques for bathymetry estimations represent an important advancement in this field [27–29].

To obtain deep water inversion from optical sensors, investigators have employed regression models that are based on machine learning techniques. Liu et al. [30], for example, investigated the performance of two artificial neural network methods—general regression neural networks (GRNN) and multilayer perceptron (MLP)—as methods for possible use in bathymetry studies. The results showed that artificial neural networks are more useful and accurate than the inversion model and regression tree. Other researchers have used an artificial neural network (ANN) [31,32] for estimating the depths of shallow waters. More recently, other authors used the machine learning technique of SVM to estimate shallow water depths, for instance, [25] applied the non-linear machine learning technique of SVM to Landsat images in Saint Maarten Island and the Ameland Inlet in the Dutch Wadden Sea and experienced an overall error of 8.26% and 14.43%, respectively. At Kauai Island in Hawaii, [33] proposed a spatially distributed SVM system to use in estimating the bathymetry of shallow water by the use of optical satellite images, as well as SVMs that were locally trained with spatially weighted votes for the prediction. In the present case, the experimental results indicated that the localized model gave a 60% lower bathymetry estimation error than from the root mean squared error (RMSE). In recent years, with the use of remote optical observation, many studies have been undertaken to estimate bathymetry in shallow water [25,34]. However, because the underlying surfaces of the harbors are submerged and frequently obscured by turbid and mud, it can be difficult to estimate changes in depths of the bottoms of harbors. Only recently, on the coast of Misano, Muzirafuti et al. [35] published a comparative quantitative analysis of the log band ratio and the optimal band ratio methods of analysis, which are employed regularly in bathymetry. The study considered the potential application of these methods in the multispectral satellite imaging of a coastal area to determine the spectral band ratio that would provide water depth information with the most accuracy, particularly in shallow turbid water. This methodology implies having great knowledge of data processing. A simple system is sought in this work in order to apply it to the management of ports without a need to resort to the bathymetry conducted in situ using an echo sounder.

This study presents optimal satellite-based bathymetry derivation models that have been developed for use with highly turbid waters at two different ports in the cities of Luarca and Candás in the Principality of Asturias (Spain). The research used data that were provided by the Sentinel-2 satellite, as well as the logarithmic relationship and analytical approaches using SVM. The work has sought to provide an efficient method to derive bathymetric data from Sentinel-2 images. Satellite-derived bathymetry maps can be used to provide low-cost and high-density data for later use in numerical models in port research. It is hoped that the approach that was developed in this work is included in the monitoring of the bathymetry of the Candás and Luarca ports as a fast and economical alternative to conventional bathymetry to evaluate the need for dredging. This technique can be used also in studying coastal management.

2. Materials and Methods

2.1. Study Sites and Data Sites

The first study site is the port of Candás (Figure 1). Candás is a coastal town in Asturias Principality in one of the northernmost points of the Iberian Peninsula. The town has a population of approximately 7000 inhabitants. Candás has a medium-sized port where fishing boats and nautical sports boats coexist, although the latter outnumber the former. The port has 188 fixed moorings and 48 temporary moorings for use by nautical sports boats. The minimum drafts of the port range between 3.5 m in the navigation channel and 1 m at the interior docks, where the boats are small. Fish production in 2019 was 112 tons.

Figure 1. Study area (**a**) Candás port; (**b**) bathymetry using echo-sounding measurements. The colors denote the depth of the water in meters.

Bathymetric data were obtained from the Port Service of the Principality of Asturias, which conducts accurate bathymetric studies that concern modeling water quality and morphological changes in ports. These bathymetries are conducted to ensure good management of the exploitation of the ports. Knowledge of the seabed facilitates the appropriate management and exploitation of the ports. Therefore, the bathymetric studies are conducted as part of its conservation and maintenance work for the planning of dredging. In this study, the only three available bathymetric charts for 2016, 2018 and 2019 were used to adjust the satellite images at various depths. This bathymetric information was collected by use of an echo sounder single beam Navisound 210 (Reson, Inc.; Slangerup, Denmark) that has a variable frequency acoustic profiler (201 kHz/33 kHz). Its position was determined by using GPS. The second study site was Luarca port (Figure 2), a coastal town in the western area of the Principality of Asturias in Spain. The town has a population of approximately 5200 inhabitants. It has a medium-sized port that is home to fishing boats and nautical sports, although the fishing sector predominates. Fish production in 2019 was 499t. As the boats that use the port are generally small vessels, the minimum drafts range between three meters in the navigation canal and two meters in the docking area. Bathymetric data also were collected by use of an echo sounder single beam Navisound 210 dual frequency (190–235 kHz), and a 1 cm vertical resolution.

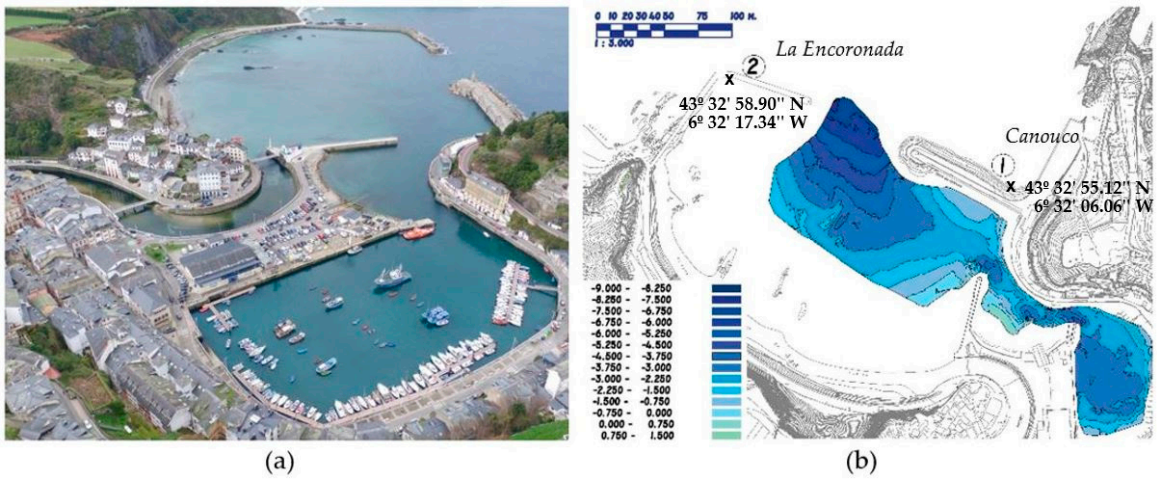


Figure 2. Study area (**a**) Luarca port; (**b**) bathymetry using echo-sounding measurements. The colors denote the depth of the water in meters, and the number denotes the dikes of Canouco and La Encoronada.

The study's echo sounding data for this area were determined by reference to survey positions of the UTM/WGS84 ZONA 30N, on 16 October 2016, 12 March 2018 and 29 April 2019 in the case of Candás,

as well as 28 June 2016, 10 May 2018 and 28 May 2019 for Luarca (see Table 1). The port data points for the in situ measurements at Candás were taken by several passes of the vessel. The distribution of the passes was made using a 10 × 20 m mesh. In Luarca, the data points were taken by passes, the distribution of which was made with a 10 × 30 m mesh in the sheltered area (between the Canouco and La Encoronada dikes, see Figure 2b) and 30 × 30 m in external waters.

2.2. Satellite Data

The data from the Sentinel-2A and -2B twin polar-orbiting satellites were used to estimate the study sites' depth of water. Sentinel-2 imagery data can be obtained from the European Space Agency's sci-hub-portal (ESA) [36,37]. The satellites Sentinel-2A and -2B were launched on June 23, 2015, and March 2017, respectively [38]. In order to evaluate the capabilities of Sentinel-2's coastal capabilities, it is necessary to possess a thorough understanding of revisit intervals and statistics regarding the world's coastlines. Sentinel-2 collects data during its orbital ground swaths. These data are subsequently interpolated on military grid reference system (MGRS) zones that can be accessed by the public. The Sentinel-2 images were collected on 3 November 2016, 16 March 2018 and 22 April 2019 in Candás port. In the study of Luarca, the satellite images were collected on 29 June 2016, 10 May 2018 and 30 May 2019 (see Table 1). The Sentinel-2 satellite is equipped with a single multispectral instrument (MSI) that has thirteen spectral bands. The bands use a push broom sensor. The latter collects rows of image data during the orbital swath and uses the satellite's forward motion along its path to generate new rows for acquisition [39]. Bands have a spatial resolution of 10 to 60 m. For example, the resolution of B2, B3, B4 and B8 is 10 m, whereas that of B5, B6, B7, B8A, B11 and B12 is 20 m. The resolution of the remaining bands is 60 m (Table 2). The dataset from Sentinel-2 was chosen due to its temporal proximity to the echo-sounding bathymetry dates and the availability of cloud free data.

Table 1. Dates of acquisition of Sentinel-2 data and in situ measurements data.

	Dates of Acquisition	
	In Situ	**Sentinel-2**
Candás	16 October 2016	03 November 2016
	12 March 2018	16 March 2018
	29 April 2019	22 April 2019
Luarca	28 June 2016	29 June 2016
	10 May 2018	10 May 2018
	28 May 2019	30 May 2019

Table 2. Sentinel-2 bands.

Band	Spectral Region	Resolution [m]	Central Wavelength [nm]
B1	Coastal aerosol	60	443
B2	Blue	10	490
B3	Green	10	560
B4	Red	10	665
B5	Vegetation red edge	20	705
B6	Vegetation red edge	20	740
B7	Vegetation red edge	20	783
B8	NIR	10	842
B8A	Narrow NIR	20	865
B9	Red Edge	60	940
B10	Water vapor	60	1375
B11	SWIR–Cirrus	20	1610
B12	SWIR	20	2190

2.3. Methodology

2.3.1. Pre-Processing of Satellite Images

The Sentinels Application Platform (SNAP) was used to view and export data. It is software offered at no charge by the European Space Agency [37] to process and analyze satellite images from the Sentinel satellite fleet. This program has a repertoire of tools (called Sentinel Toolboxes) that are specific to working with the images, whether they are Sentinel-1 radar images or the optical Sentinel-2 and Sentinel-3 multiband images. In any case, the SNAP tools can be used to manage multispectral images from missions, such as Envisat, Landsat, MODIS or SPOT. What was obtained were data from bands with different resolutions. The first step in using these data is to transform all bands to the same resolution. All spectral bands of the Sentinel-2 image were resampled to a 10 m resolution [40,41]. Resampling of the downloaded images was conducted with the software ESA SNAP (v7.0.1) [42] using the S2 Resampling Processor. As a result, a set of data is obtained that is not georeferenced, because we have only the longitude and latitude of the corners of the study portion. To determine the positioning of the reference points, the geographical location of each point is defined by its longitude and latitude using the SNAP program. From the longitude and latitude, a coordinate projection is made using the WGS84 ellipsoid, obtaining the coordinates in ETRS89. This is the same system with which the positions obtained by the echo sounder are projected. Position average errors in the ellipsoid projections are of the order of 1 cm.

2.3.2. Pre-Processing of Data. Generation of Comparison Bathymetry Grid

The data obtained were compared to the bathymetry projected. For this, coordinates were projected using a geodetic calculator, since the bathymetry uses ETRS89 coordinates, based on the same ellipsoid WGS84. Then, the data of the bands that are associated with UTM x-y coordinates could be obtained. To assign the z coordinate, the annual bathymetries of the study ports are used. These are undertaken by a single beam probe that is mounted on a vessel. From these z data that are obtained every 10 cm, a surface of the port's bottom is generated. This surface is obtained using digital models of the terrain type triangulated irregular network (TIN). In this case, triangulation is conducted using linear interpolation. The error generated in this process is small since the surfaces of the seabed are smooth surfaces and without great irregularities. The z coordinates are referenced to zero of the port itself (the minimum level that has been measured at the highest low tide of the last 15 years). Each pixel is assigned the corresponding dimensions based on the former's x-y location (Figure 3).

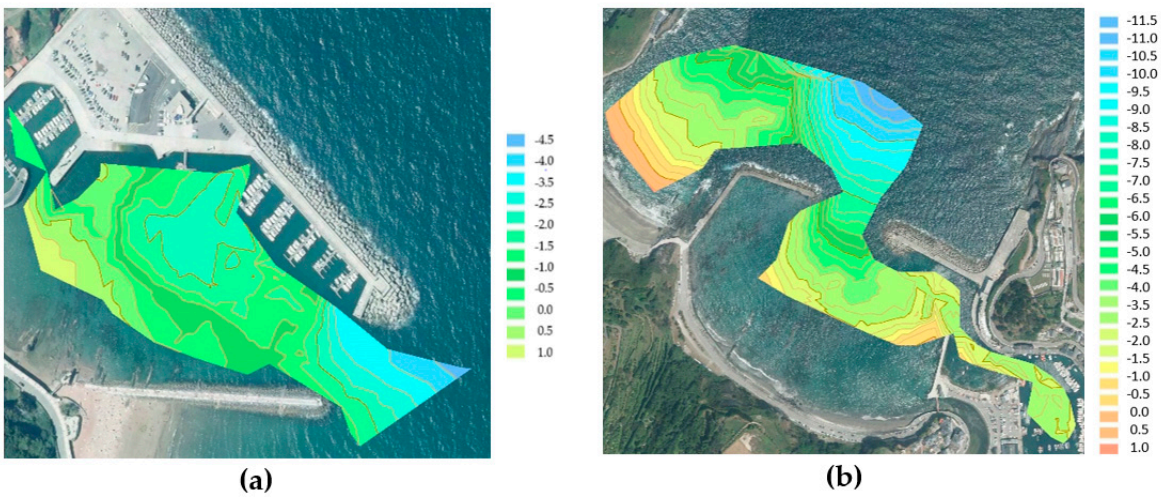


Figure 3. Bathymetry maps obtained using echo sounder in situ measurements. (**a**) Candás; (**b**) Luarca. The colors denote the depth of the water in meters.

2.3.3. Bathymetry Estimation Based on Support Vector Machines

This study applied support vector algorithms to derive bathymetry from water reflectivity. Stumpf et al. [43] suggested a linear model, despite the fact that, in many ways, it did not always result in a relationship between the water depths and the ratio that was linear. Thus, it is best to obtain it by exploring the relationship of a non-linear function (f) to map water depth (Z) (Equation (1)).

$$Z = f\frac{\ln[nR_w(\lambda_i)]}{\ln\left[nR_w\left(\lambda_j\right)\right]} \tag{1}$$

where n is a fixed value and R_w is the reflectance that has been observed for the wavelength (λ) of bands i and j.

Support vector machine is one of the many non-linear regression techniques. It has been studied extensively and finds use as a universal approximation [44–46]. SVM has some advantages over other methods. Its classification accuracy is relatively higher when the inputs are correctly selected. This method is based on a kernel-based algorithm. Its new input estimations depend on an evaluation by the kernel function of a subcategory of events in a training stage. The task to use this method is to identify a function that will minimize Equation (2)'s final error.

$$y(x) = w^T \cdot \phi(x) + b \tag{2}$$

where $y(x)$ is the predicted value, b represents the value of the bias and $\phi(x)$ maintains the feature space transformation. An ϵ insensitive error function (Equation (4)) replaces the error function in the linear regression (Equation (3)) in this method. Equation (4) assigns a zero to values if ϵ is greater than the difference between the predicted and target value. If the difference is greater than, or equal to ϵ, the value of the error function does not change. To minimize Equation (5), the difference between the predicted and targeted values is also assigned a cost (C).

$$\frac{1}{2}\sum_{n=1}^{n}[y_n - t_n]^2 + \frac{\lambda}{2}\|w\|^2 \tag{3}$$

$$E_\epsilon(y(x) - t) = \begin{cases} 0, & if\ \left|y(x) - t\right| < \epsilon \\ \left|y(x) - t\right| - \epsilon, & otherwise \end{cases} \tag{4}$$

$$C\sum_{n=1}^{n} E_\epsilon(y(x_n) - t_n) + \frac{1}{2}\|w\|^2 \tag{5}$$

where $y(x)$ is the value that is predicted by Equation (2), t denotes the searched target function, ϵ is the margin when the function fails to penalize and C denotes the penalty. The process is optimized, although the initial function (Equation (3)) becomes more complex (Equation (6)).

$$y(x) = \sum_{n=1}^{n}(\alpha_i - \alpha_i^*) < x_i \cdot x > +b \tag{6}$$

where α is one solution to the optimization problem that can occur with the Lagrangian theory. The data are changed by the function to data of a feature space. This improves the non-linear problem's accuracy. As a result, the final function resembles Equation (7).

$$y(x) = \sum_{n=1}^{n}(\alpha_i - \alpha_i^*)k(x_i,\ x) + b \tag{7}$$

For the purpose of classification, the best kernel is generally the Gaussian radial basis function (RBF). It provides the greatest overall accuracy and kappa [47]. This study used this RBF function (Equation (8)).

$$k(x_i,\ x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} \tag{8}$$

To program the methodology that was proposed, the R statistical software was selected [48].

2.3.4. Data Processing

In this work, the dataset was used for the support vector machine studies in the two study areas of Candás and Luarca. Model training used 80% of the data points and testing accounted for the remaining 20% (Table 3). This ratio is common in machine learning studies [49,50]. The tests were conducted with several random data segmentations a total of 20 times. The additional results were eliminated and the average of the remaining data was used for the studies.

Table 3. Number of data points of the training and testing data sets for the study areas.

Port	Training	Testing
Candás	1092	284
Luarca	1593	388

Checking and applying a valid methodology is necessary for the implementation of the remote sensing measurements of the bathymetry of a specific area from optical images. In doing so, it was necessary to discriminate and select only the most appropriate bands and to determine the corrections of the images. The input variables that were used to model the bathymetry using SVM techniques were correlated using Equation (1). If all possible combinations between bands were considered, the number of input variables would be 68. Considering that this number of input variables was excessive, and in order to optimize them, the correlation of the different bands to each other was analyzed. To conduct the correlation analysis of the various bands, Pearson's correlation coefficient was used. Pearson's correlation coefficient (R) appears in Table 4 as a measure of the linear correlation of each study's corresponding band pairs. A 1.0 coefficient of correlation indicates that two variables are correlated perfectly, whereas a coefficient of 0.0 indicates the absence of a linear relationship [51]. Most bands, except B9 and B1, were highly correlated with each other. To determine which bands were unnecessary, the pairs of bands that had a correlation greater than 0.9 were plotted. Figure 4 provides the scatter diagrams that show the relationship between pairs of bands that seem to supply visually the same information from Sentinel-2. Perfect agreement between two bands was indicated in each diagram by a 1-to-1 line.

Table 4. R values between main bands of Sentinel-2.

Band	B1	B2	B3	B4	B5	B6	B7	B8	B8A	B9	B11	B12
B1												
B2	0.96											
B3	0.89	0.98										
B4	0.85	0.95	0.99									
B5	0.82	0.9	0.95	0.97								
B6	0.83	0.91	0.96	0.97	1							
B7	0.82	0.9	0.95	0.97	0.99	1						
B8	0.81	0.92	0.97	0.99	0.97	0.98	0.98					
B8A	0.81	0.9	0.95	0.97	0.99	1	1	0.98				
B9	−0.04	0.06	0.22	0.27	0.36	0.34	0.33	0.31	0.34			
B11	0.74	0.84	0.9	0.94	0.98	0.98	0.98	0.96	0.98	0.38		
B12	0.72	0.82	0.89	0.93	0.97	0.97	0.98	0.95	0.99	0.39	1	

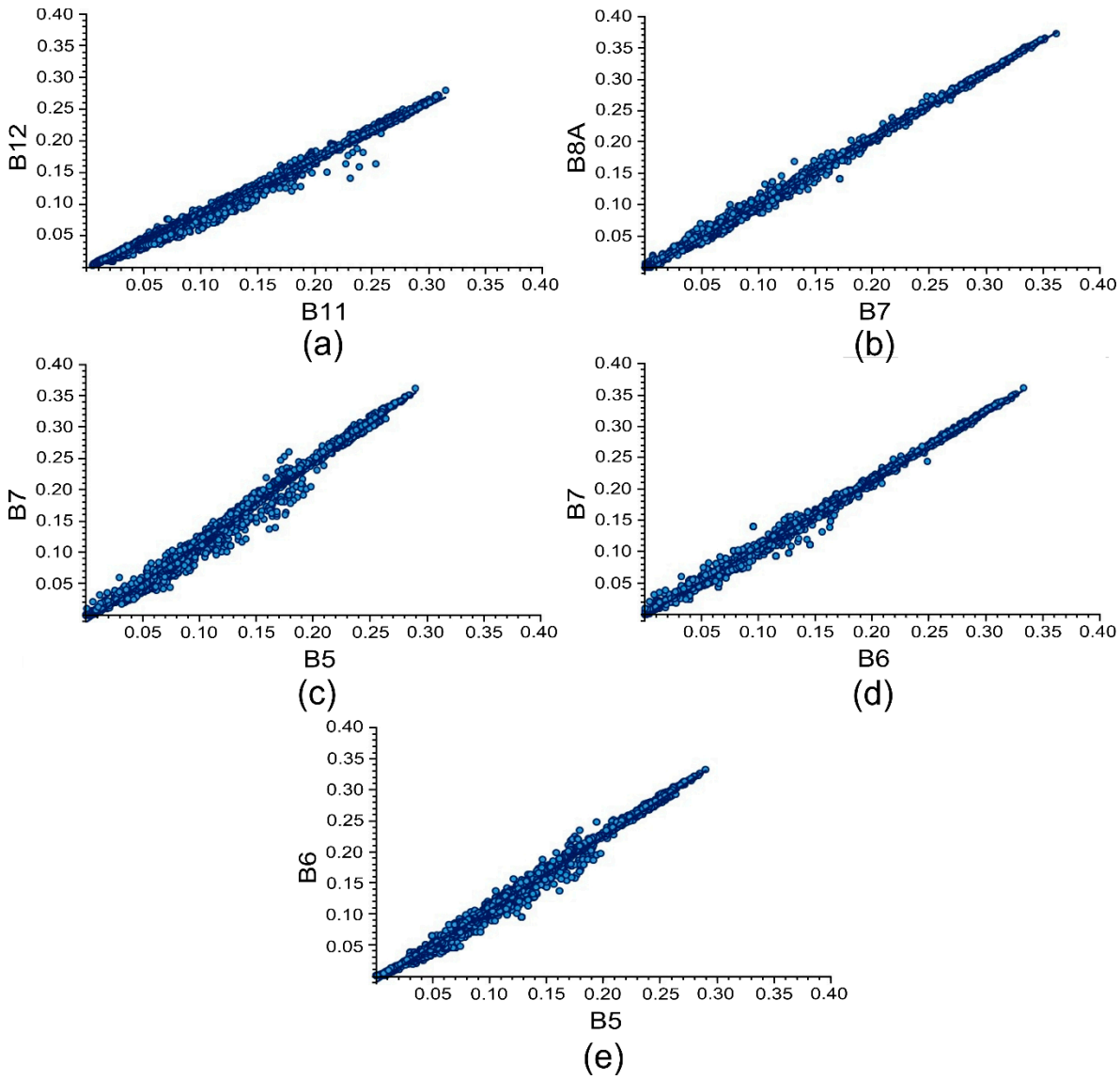


Figure 4. Scatter diagram of the relationship of the pairs of bands that appear to provide visually the same information from Sentinel-2.A (**a**) B11-B12; (**b**) B7-B8A; (**c**) B5-B7; (**d**) B7-B6; (**e**) B5-B6.

Figure 4 indicates the high correlation between the bands. For this reason, some of the bands represented (B5, B6, B8A and B11) were eliminated for modeling with SVM. Under the same conditions, the bands with the highest resolution were maintained, since they have no calculated data due to interpolation. Subsequently, the correlation of these 28 variables was analyzed again, and one variable was eliminated. The bands that were chosen for data modeling were B1, B2, B3, B4, B7, B8, B9 and B12. Finally, the status of the tide at the time of the orthophoto was added as an input variable to the model.

3. Results and Discussion

A comparison was made of the bathymetry that the Sentinel-2 satellite data made possible and what the echo sounder in situ measurements provided for the ports of Candás and Luarca. The mean absolute error (MAE) and the root mean squared error (RMSE) were analyzed to determine the generalization capacities of the regression models. They can be calculated by equations 9 and 10.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Z_{SVM} - Z_{echo}| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_{SVM} - Z_{echo})^2} \quad (10)$$

In this case, Z_{SVM} is the SVM predicted depths from Sentinel-2 images and Z_{echo} is the echo sounder depths from the field data points. Further, the coefficient of determination is provided, it (R^2) indicates the regression model's "goodness of fit." On the other hand, the adjusted R^2 penalizes the R^2 value for each predictor variable in the regression model (in this case, the input variable).

Table 5 shows the depth characteristics of the study areas. It shows negative values for the minimum depth (−5.0149 m in Candás and −11.9601 m in Luarca). The reason for this is that zero, "0", refers to the minimum level that has been registered in that port. In Spain, Royal Decree 1071/2007 established that the surface of the sea in the city of Alicante (Spain) has a zero value. Therefore, a bathymetric measurement is normally considered to be the distance between the bottom to the zero value, zero. In this study, it is considered to be the lowest value of the surface of the water at each port (i.e., the minimum level that has been measured at the highest tide in the last 15 years).

Table 5. Depth characteristics of the study areas.

Port	Max Depth [m]	Min Depth [m]	Mean Depth [m]
Candás	1.3461	−5.0149	−1.5519
Luarca	1.5979	−11.9601	−4.0694

The coefficient of determination or R^2, MAE and RMSE that appear in Table 6 were obtained to determine the ability to predict and generalize of the SVM regression model that was obtained using the training dataset. It can be seen that the values of both R^2 coefficients for Candás (R^2 = 0.911) and Luarca (R^2 = 0.976) are very close to 1, which is high. This indicates a high correlation between the observed and estimated values. The table also shows that the MAE and RMSE values are small and similar for Candás (MAE = 0.2779 m and RMSE = 0.3274 m) and Luarca (MAE = 0.3694 m and RMSE = 0.4409). Thus, the adjustment of the regression model is accurate.

Table 6. Results for R^2, mean absolute error (MAE), root mean squared error (RMSE) and relative error in a comparison of support vector machine (SVM) predicted values and in situ measurement values of depths using the training dataset.

Port	R^2	MAE [m]	RMSE [m]
Candás	0.911	0.2779	0.3694
Luarca	0.976	0.3274	0.4409

After determining the R^2, MAE and RMSE errors, scatter diagrams of the SVM predicted depth from the Sentinel-2 images (Z_{SVM}) vs. the echo sounder depth (Z_{echo}) from the training dataset for Candás (see Figure 5) and Luarca (see Figure 6) were created. The points that are closest to the diagonal line have the highest correlation to the regression models. In this case, both the Candás and Luarca training dataset points are very close to the diagonal line.

In addition, Table 7 shows the R^2, MAE and RMSE that were obtained using the testing dataset. Similarly, it can be seen that, for the training dataset, the values of coefficients R^2 for both Candás (R^2 = 0.8553) and Luarca (R^2 = 0.9731) are very close to 1.0. This, also, is very high. It indicates that the estimated values and the observed values are highly correlated. Further, the values of MAE and RMSE that are shown in this table are small and similar for Candás (MAE = 0.3421 m and RMSE = 0.4370 m) and Luarca (MAE = 0.3678 m and RMSE = 0.4640). The regression analysis results for the training and testing dates were consistent.

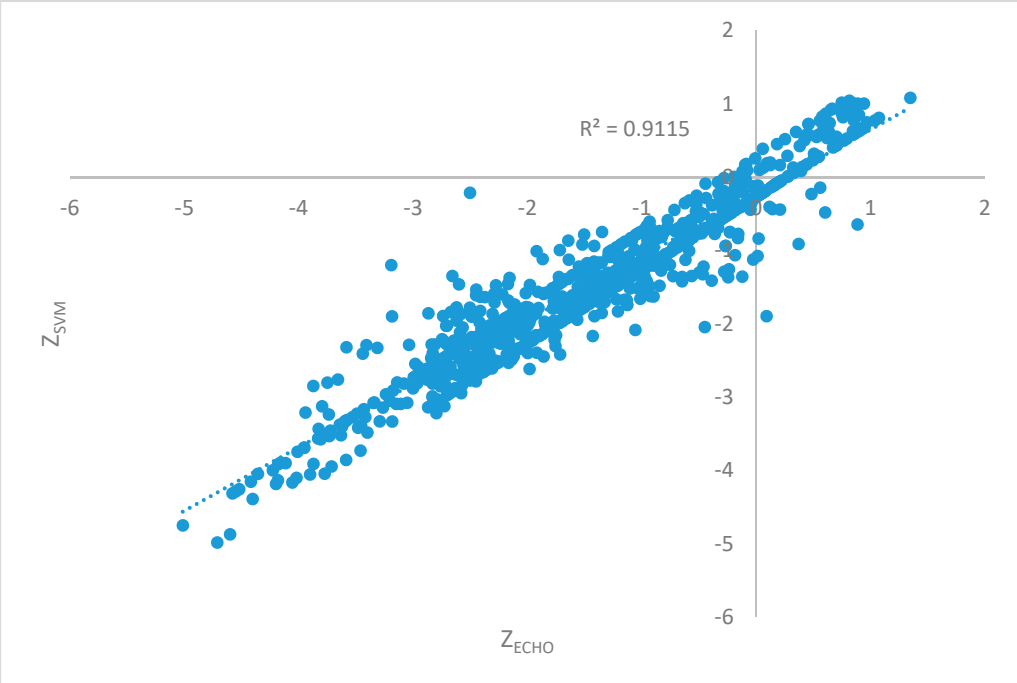


Figure 5. Scatter diagram of training data showing SVM predicted depth (m) from Sentinel-2 images (Z_{SVM}) vs. echo sounder depth (Z_{echo}) from field data points for Candás.

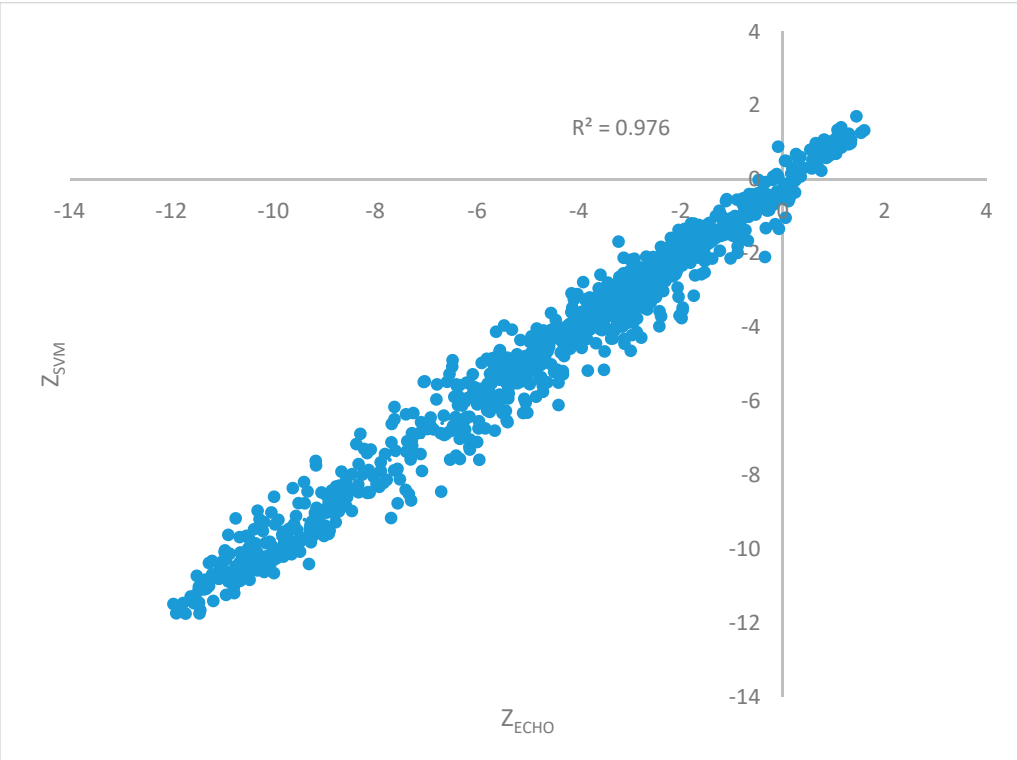


Figure 6. Scatter diagram of training data showing SVM predicted depth (m) from Sentinel-2 images (Z_{SVM}) vs. echo sounder depth (Z_{echo}) from field data points for Luarca.

After determining the R^2, MAE and RMSE errors, a scatter diagram of the SVM predicted depth from the Sentinel-2 images (Z_{SVM}) vs. the echo sounder depth (Z_{echo}) from the testing dataset for Candás (See Figure 7) and Luarca (See Figure 8) were created. As in Figures 5 and 6, points that are

closest to the diagonal line mean that the correlation with the regression models is greater. In this case, both the Candás and Luarca testing dataset points are also very close to the diagonal line.

Table 7. Results of the R^2, MAE, RMSE and relative error when comparing SVM predicted and in situ measurement value of depths using the testing dataset.

Port	R^2	MAE [m]	RMSE [m]
Candás	0.8553	0.3421	0.4370
Luarca	0.9731	0.3678	0.4640

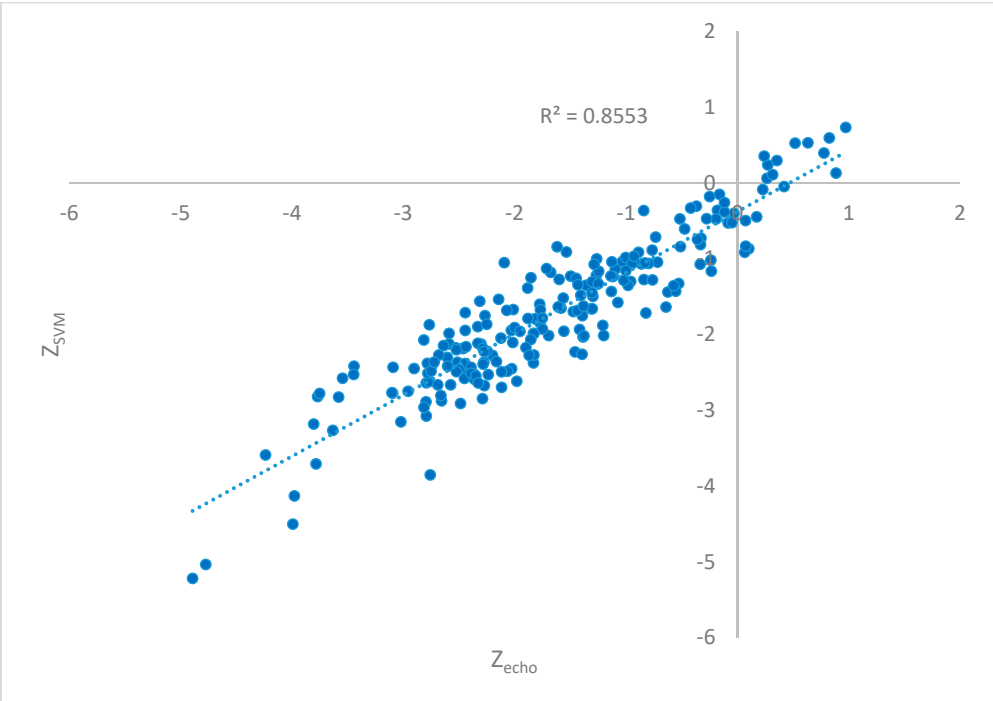


Figure 7. Scatter diagram of testing data showing SVM predicted depths (m) from Sentinel-2 images (Z_{SVM}) vs. echo sounder depth (Z_{echo}) from field data points for Candás.

In addition, the relative error of the average depth at the port was calculated using the training and testing dataset (Table 8). The relative error values were 22.05% and 9.04% for Candás and Luarca, respectively. It is greater in Candás because the average depth there is less. In order to obtain a more representative average error for depth, also the error in the average range of depths was calculated. In this case, the errors were 10.76% in Candás and 5.43% in Luarca for the training dataset, and 8.74% in Candás and 4.83% in Luarca for the testing dataset. Despite high relative errors, it is important to note that the objective of the work is that the surface generated by means of the dimensions that were obtained should reflect reality and detect conflictive areas that are in need of dredging. Therefore, although the error is important, it is less so than the general behavior of the generated surface.

Table 8. Results of the relative error when comparing the SVM predicted and in situ measurement values of depth.

	Training Dataset		Testing Dataset	
Port	Relative Error [%]	Relative Error Range Depth [%]	Relative Error [%]	Relative Error Range Depth [%]
Candás	17.90%	8.74%	22.05%	10.76%
Luarca	8.05%	4.83%	9.04%	5.43%

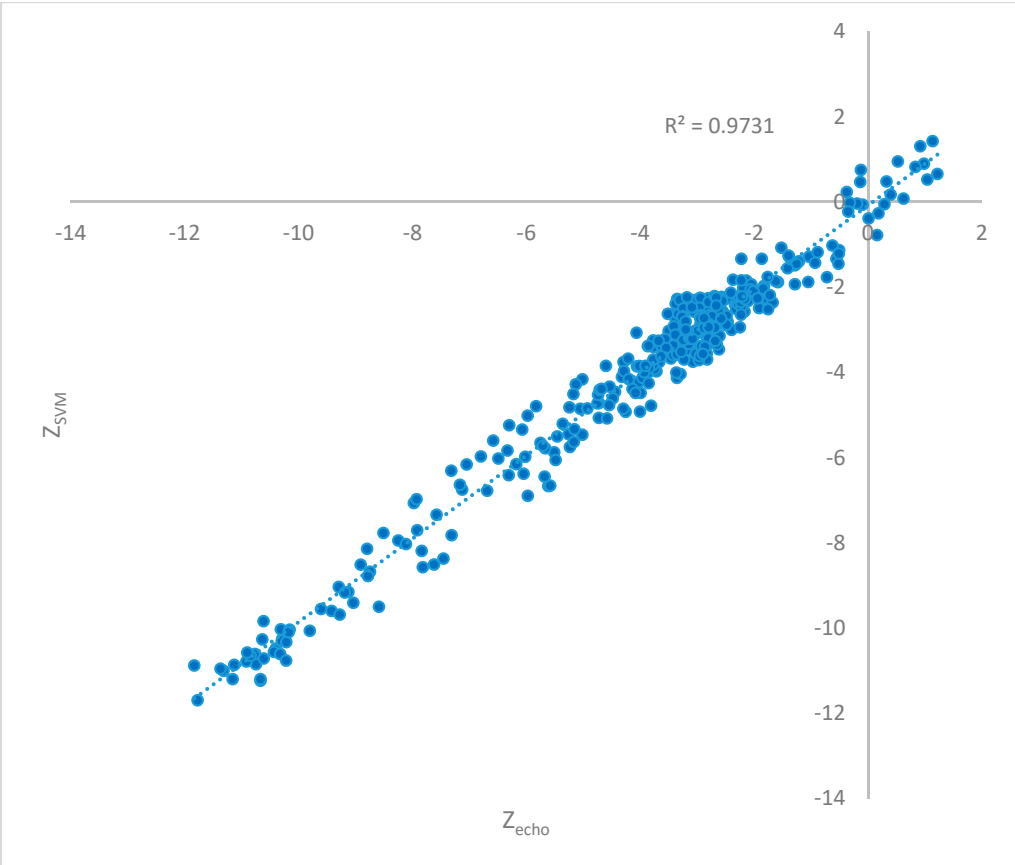


Figure 8. Scatter diagram of testing data showing SVM predicted depths (m) from Sentinel-2 images (Z_{SVM}) vs. echo sounder depth (Z_{echo}) from field data points for Luarca.

Figures 9 and 10 represent the bathymetry maps of Candás and Luarca that were obtained using the echo sounder in situ measurements and SVM algorithms.

In the estimated bathymetries that are shown in Figures 9a and 10a, one can detect the transitions and changes in depth in the ports that were studied using this methodology. This is very interesting, because in bathymetry, maps are more helpful in detecting the transitions and average behavior of the bottom than the elevation at a specific point. Figures 9 and 10 show that, although the errors are slightly higher in Luarca (Figure 10a), the general behavior of the seabed has been determined in two ports, both in the deepest areas towards the open sea and in the shallowest and sheltered areas. In addition. Figures 9a and 10a show how the algorithm determined correctly the areas of greatest depth and those of lowest depth with smooth transitions, as well as the contour lines of the coasts. In addition, complex areas of very shallow depth, such as the interior dock of the port of Candás (Figure 9a), are also detected correctly.

To complete the study, a detailed image of the behavior of the contour lines obtained from the bathymetry is generated. In the first phase, shallower areas are compared (Figure 11a,b and Figure 12a,b), and in the second phase, zones of greater water depth are compared (Figure 11c,d and Figure 12c,d).

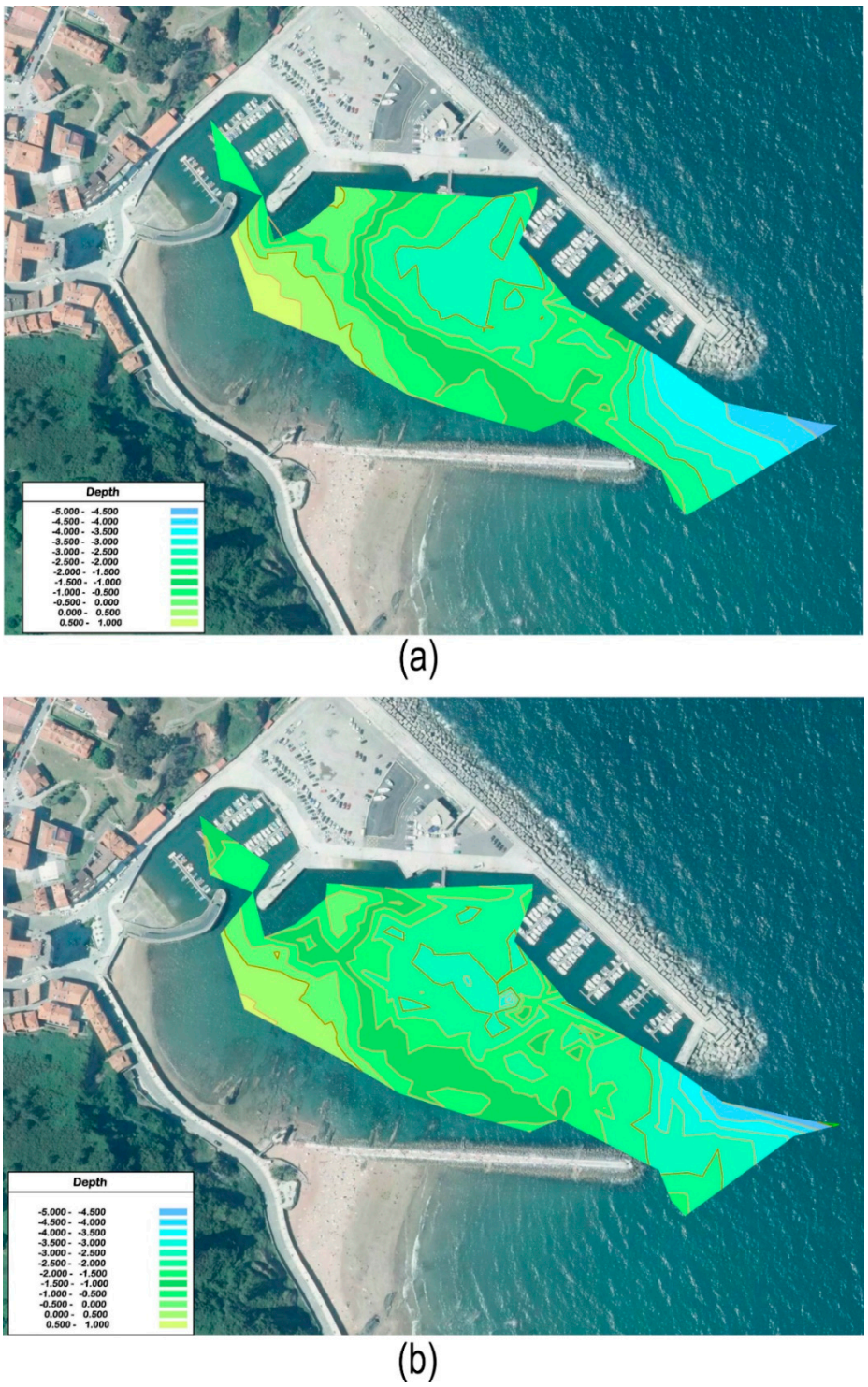


Figure 9. Bathymetry maps of Candás obtained using (**a**) echo sounder in situ measurements (Z_{echo}); (**b**) SVM (Z_{SVM}). The colors denote the depth of the water in meters.

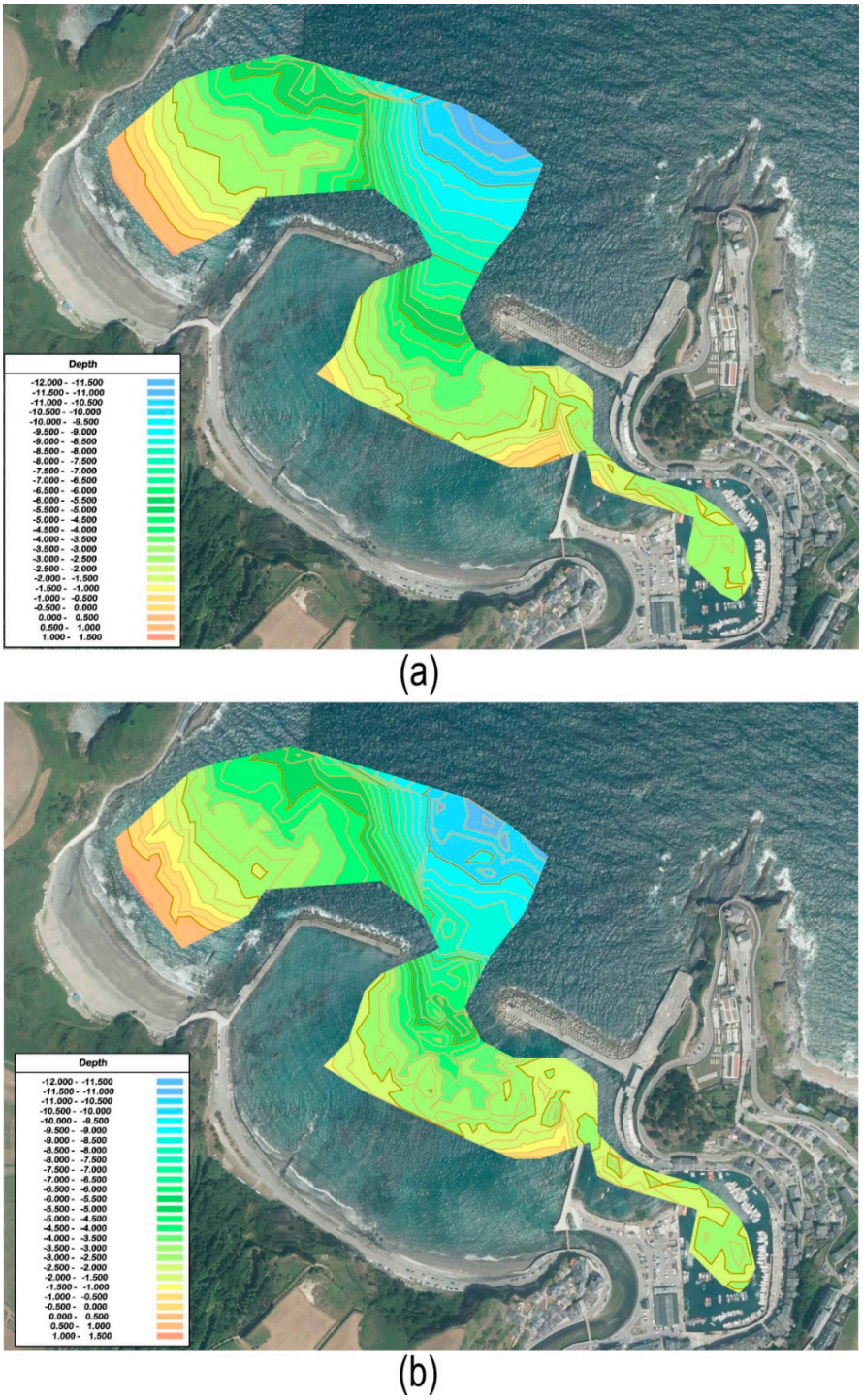


Figure 10. Bathymetry maps of Luarca obtained using (**a**) echo sounder in situ measurements (Z_{echo}); (**b**) SVM (Z_{SVM}). The colors denote the depth of the water in meters.

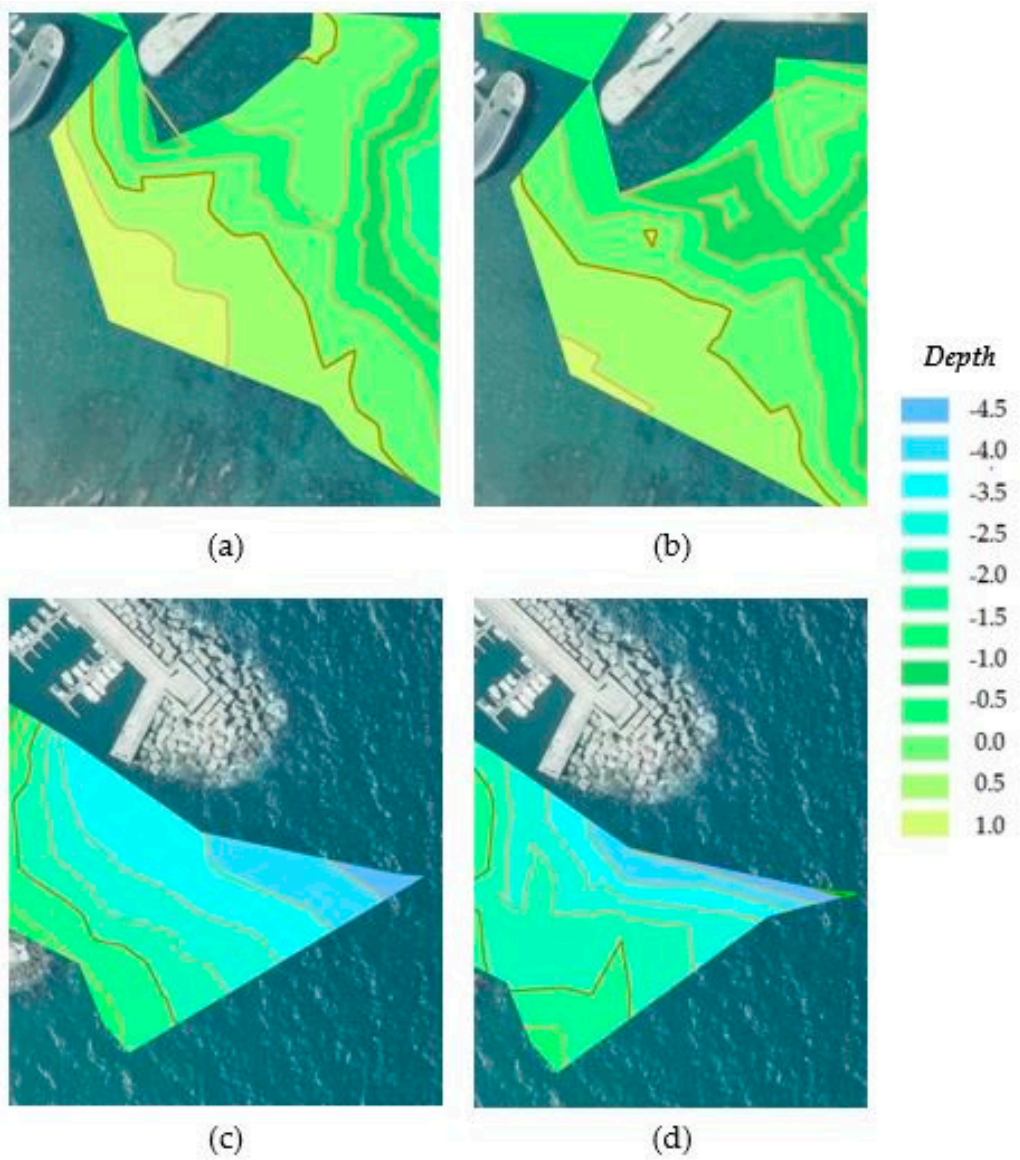


Figure 11. Details of zoomed areas in bathymetry maps in Candás. (**a**,**c**) from echo sounder in situ measurements; (**b**,**d**) of SVM depth estimates. The colors denote the depth of the water in meters.

Figure 11 shows some details of zoomed areas to compare the bathymetry maps from the echo sounder in situ measurements to the SVM depth estimates for Candás. In Figure 11a,b, it is revealed that the contour lines have substantially similar shapes and also identify a zone of less depth and, therefore, an area that may need dredging. From Figure 11b,d, it is shown that the representation of the curves is also similar, although the result using an echo sounder indicates a smoother surface. Further, a zone of greater depth is detected in the port. As expected, it was located at the entrance.

In the same way, in Figure 12, they were compared to some details of zoomed areas in Luarca from the echo sounder in situ measurements (Figure 12a,c) and from the SVM depth estimates (Figure 12a,c). In Figure 12a,b, it can be seen that the contour lines have shapes that are substantially similar and parallel to the coastline. In the echo sounder bathymetry, the traces are smoother and better adjusted to the type of bottom with a smooth slope. These results, which were predicted by SVM, are considered to be valid, since they detect correctly the shallowest areas. Finally, in comparing Figure 12c,d, it is seen that the representation of the echo sounder is also smoother. Although the contour lines of the model do not conform precisely to reality, the areas of greater depth are detected.

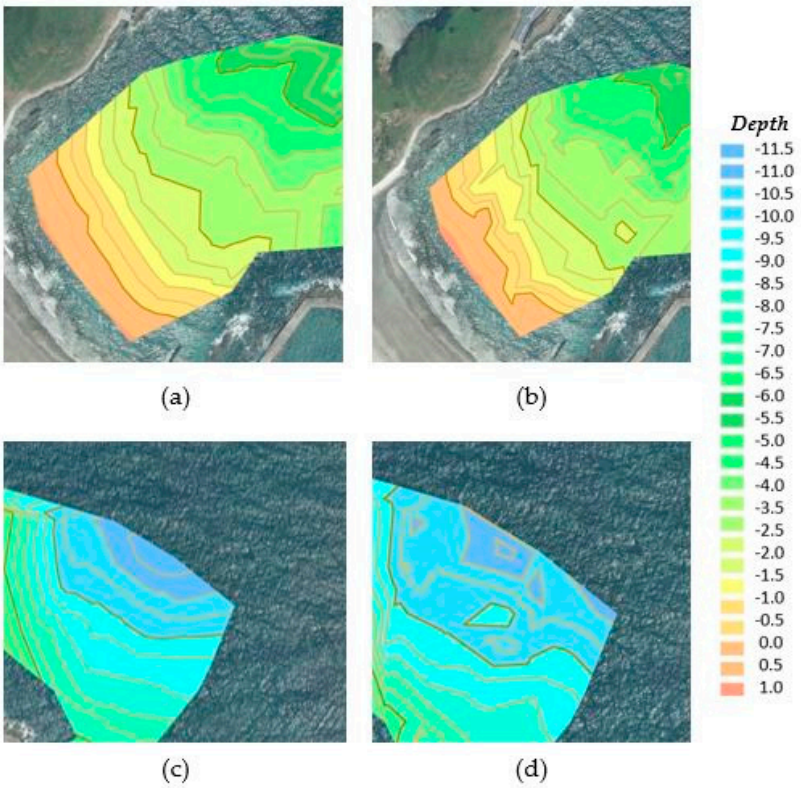


Figure 12. Details of zoomed areas in the bathymetry maps in Luarca. (**a**,**c**) from echo sounder in situ measurements; (**b**,**d**) from the SVM depth estimates. The colors denote the depth of the water in meters.

4. Conclusions

This work was an examination of a technique of remote sensing bathymetry that is based on support vector machine techniques and Sentinel-2 imagery. It was used to estimate the depth in the turbid water of two different ports in the cities of Luarca and Candás in the Principality of Asturias, Spain. The satellite-estimated depth provides an alternative method with which to respond to the increasing demand for coastal topography and bathymetry information in shallow water areas [52]. This approach brings a new perspective to remotely sensed estimated bathymetry. It also provides a high level of accuracy and a cost-effective and efficient solution to the turbid areas of ports, which must be dredged periodically to maintain the free draft and for adequate port management. The visual and statistical results of the study of the ports of Luarca and Candás demonstrate the capacities of the SVM techniques for the prediction of depths from satellite images. The proposed method achieved a greater accuracy for the training dataset in Candás with a mean absolute error of 0.228 m, a root mean squared error of 0.369 m and a coefficient of determination or R^2 value of 0.911. The overall errors that were experienced using the testing dataset were a mean absolute error of 0.368 m, a root mean squared error of 0.463 m and a R^2 value of 0.855. In the case of Luarca, the SVM method produces depth estimates for the training dataset with a mean absolute error of 0.3274 m, a root mean squared error of 0.441 m and an R^2 value of 0.976. The errors in using the testing dataset were a mean absolute error of 0.378 m, a root mean squared error of 0.464 m and an R^2 value of 0.973. The low error values obtained in training and testing for both study ports highlight the precision of the bathymetries that were obtained. However, these values are higher than those that other authors obtained [9,23,43,53]. This may be due to the color and turbidity, since the bottoms, which were contaminated and had a muddy composition, have higher light absorption than the sandy bottoms with clear waters that have been analyzed in most studies. Another factor that may have affected these results was the use of a free satellite that has a lower resolution than the satellites that other authors employed. In the future,

different artificial neural network techniques should be studied for the estimation of depths with high accuracy from open Sentinel-2 images.

Author Contributions: Conceptualization, V.M.-P. and F.O.-F.; software and validation, V.M.-P. and F.O.-F.; formal analysis, M.C.-B.; data curation, V.M.-P.; writing—original draft preparation, M.C.-B. and V.M.-P.; writing—review and editing, V.M.-P., M.C.-B., and E.P.V.-G.; supervision, F.O.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science, Technology and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is part-funded by the European Regional Development Fund (ERDF).

Acknowledgments: The authors wish to thank to the Port Service and Transport Infrastructures of the Principality of Asturias for the bathymetries from the in situ data collected.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strayer, D.L.; Malcom, H.M.; Bell, R.E.; Carbotte, S.M.; Nitsche, F.O. Using geophysical information to define benthic habitats in a large river. *Freshw. Biol.* **2006**, *51*, 25–38. [CrossRef]
2. Brown, C.J.; Blondel, P. Developments in the application of multibeam sonar backscatter for seafloor habitat mapping. *Appl. Acoust.* **2009**, *70*, 1242–1247. [CrossRef]
3. Ferretti, R.; Fumagalli, E.; Caccia, M.; Bruzzone, G. Seabed classification using a single beam echosounder. In Proceedings of the OCEANS 2015—Genova, Genoa, Italy, 18–21 May 2015.
4. O'Hara Murray, R.B.; Gallego, A.G. Data review and the development of realistic tidal andwave energy scenarios for numerical modelling of Orkney Islands waters, Scotland. *Ocean Coast. Manag.* **2017**, *147*, 6–20. [CrossRef]
5. KhaledSeif, A.; Kuroiwa, M.; Abualtayef, M.; Mase, H.; Matsubara, Y. A hydrodynamic model of nearshore waves and wave-induced currents. *Int. J. Nav. Archit. Ocean. Eng.* **2011**, *3*, 216–224. [CrossRef]
6. Clementi, E.; Oddo, P.; Drudi, M.; Pinardi, N.; Korres, G.; Grandi, A. Coupling hydrodynamic and wave models: First step and sensitivity experiments in the Mediterranean Sea. *Ocean Dynam.* **2017**, *67*, 1293–1312. [CrossRef]
7. Tang, K.; Pradhan, B. Converting digital number into bathymetric depth: A case study over coastal and shallow Water of Langkawi Island, Malaysia. In Proceedings of the FIG Working Week, Athens, Greece, 24 January 2015.
8. Mason, D.C.; Gurney, C.; Kennett, M. Beach Topography Mapping—A Comparison of Techniques. *J. Coast. Conserv.* **2000**, *6*, 113–124. [CrossRef]
9. Janowski, L.; Trzcinska, K.; Tegowski, J.; Kruss, A.; Rucinska-Zjadacz, M.; Pocwiardowski, P. Nearshore Benthic Habitat Mapping Based on Multi-Frequency, Multibeam Echosounder Data Using a Combined Object-Based Approach: A Case Study from the Rowy Site in the Southern Baltic Sea. *Remote Sens.* **2018**, *10*, 1983. [CrossRef]
10. Madricardo, F.; Foglini, F.; Kruss, A.; Ferrarin, C.; Pizzeghello, N.M.; Murri, C.; Rossi, M.; Bajo, M.; Bellafiore, D.; Campiani, E.; et al. High Resolution Multibeam and Hydrodynamic Datasets of Tidal Channels and Inlets of the Venice Lagoon. *Sci. Data* **2017**, *4*, 170121. [CrossRef]
11. Horritt, M.S.; Bates, P.D.; Mattinson, M.J. Efects of mesh resolution and topographic representation in 2D finite volume models of shallow water fluvial flow. *J. Hydrol.* **2006**, *329*, 306–314. [CrossRef]
12. Coggins, L.X.; Ghadouani, A. High-resolution bathymetry mapping of water bodies: Development and implementation. *Front. Earth Sci.* **2019**, *7*, 330. [CrossRef]
13. Lyzenga, D.R.; Malinas, N.P.; Tanis, F.J. Multispectral bathymetry using a simple physically based algorithm. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2251–2259. [CrossRef]
14. Gao, J. Bathymetric mapping by means of remote sensing: Methods, accuracy and limitations. *Prog. Phys. Geogr.* **2009**, *33*, 103–116. [CrossRef]
15. Jégat, V.; Pe'eri, S.; Freire, R.; Klemm, A.; Nyberg, J. Satellite-derived bathymetry: Performance and production. In Proceedings of the Canadian Hydrographic Conference, Halifax, NS, Cannada, 16–19 May 2016.
16. Hodúl, M.; Bird, S.; Knudby, A.; Chénier, R. Satellite derived photogrammetric bathymetry. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 268–277. [CrossRef]

17. Lyzenga, D. Passive remote sensing techniques for mapping water depth and bottom features. *Appl. Opt.* **1978**, *17*, 379–383. [CrossRef]
18. Lyzenga, D. Remote sensing of bottom reflectance and ater attenuation parameters in shallow water using aircraft and Landsat data. *Int. J. Remote Sens.* **1981**, *2*, 71–82. [CrossRef]
19. Van Hengel, W.; Spitzer, D. Multi-temporal water depth mapping by means of Lansat TM. *Int. J. Remote Sens.* **1991**, *12*, 703–712. [CrossRef]
20. Mishra, D.; Narumalani, S.; Lawson, M.; Rundquist, D. Bathymetric mapping using IKONOS multispectral data. GISci. *Remote Sens.* **2004**, *41*, 301–321.
21. Su, H.; Liu, H.; Heyman, W. Automatic derivation for bathymetric information for multispectral satellite imagery using a non-linear inversion model. *Mar. Geod.* **2008**, *31*, 281–298. [CrossRef]
22. Lyons, M.; Phinn, S.; Roelfsema, C. Inegrating Quickbird multi-spectral satellite and field data: Mapping bathymetry, Seagrass Cover, Seagrass species and change in Moreton bay, Australia in 2004–2007. *Remote Sens.* **2011**, *3*, 42–64. [CrossRef]
23. Misra, A.; Vojinovic, Z.; Ramakrishnan, B.; Luijendijk, A.; Ranasinghe, R. Shallow water bathymetry mapping using Support Vector Machine (SVM) technique and multispectral imagery. *Int. J. Remote Sens.* **2018**, *39*, 4431–4450. [CrossRef]
24. Caballero, I. Assessment of a multi-scene approach with sentinel-2A/B imagery to estimate satellite-derived Bathymetry over moderately turbid regions. In Proceedings of the Poster Presented at the Living Planet Symposium, Milan, Italy, 13–17 May 2019.
25. Traganos, D.; Poursanidis, D.; Aggarwal, B.; Chrysoulakis, N.; Einartz, P. Estimating satellite-derived bathymetry (SDB) with the google earth engine and sentinel-2. *Remote Sens.* **2018**, *10*, 859. [CrossRef]
26. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
27. Caballero, I.; Stumpf, R.P.; Meredith, A. Preliminary assessment of turbidity and chlorophyll impact on bathymetry derived from sentinel-2a and sentinel-3a satellites in south florida. *Remote Sens.* **2019**, *11*, 645. [CrossRef]
28. Almar, R.; Kestenare, E.; Reyns, J.; Jouanno, J.; Anthony, E.; Laibi, R.; Hemer, M.; Du, Y.; Ranasinghe, R. Response of the Bight of Benin (Gulf of Guinea, West Africa) coastline to anthropogenic and natural forcing, part1: Wave climate variability and impacts on the longshore sediment transport. *Cont. Shelf Res.* **2015**, *110*, 48–59. [CrossRef]
29. Bergsma, E.W.; Almar, R.; de Almeida, L.P.M.; Sall, M. On the operational use of uavs for video-derived bathymetry. *Coast. Eng.* **2019**, *152*, 103527. [CrossRef]
30. Liu, S.; Gao, Y.; Zheng, W.; Xiaolu, L. Performance of Two Neural Network Models in Bathymetry. *Remote Sens. Lett.* **2015**, *6*, 321–330. [CrossRef]
31. Wang, Y.; Zhang, P.; Dong, W.; Zhang, Y. Study on Remote Sensing of Water Depths Based on BP Artificial Neural Network. *Mar. Sci. Bull.* **2007**, *9*, 26–35.
32. Ceyhun, Ö.; Yalçın, A. Remote sensing of water depths in shallow waters via artificial neural networks. Estuar. *Coast. Shelf. Sci.* **2010**, *89*, 89–96. [CrossRef]
33. El-Mewafi, M.; Salah, M.; Fawzi, B. Assessment of Optical Satellite Images for Bathymetry Estimation in Shallow Areas Using Artificial Neural Network Model. Am. *J. Geogr. Inf. Syst.* **2018**, *7*, 99–106.
34. Wang, L.; Liu, H.; Su, H.; Wang, J. Bathymetry retrieval from optical images with spatially distributed support vector machines. *GISci. Remote Sens.* **2019**, *56*, 323–337. [CrossRef]
35. Muzirafuti, E.-A.; Barreca, G.; Crupi, A.; Faina, G.; Paltrinieri, D.; Lanza, S.; Randazzo, G. The Contribution of Multispectral Satellite Image to Shallow Water Bathymetry Mapping on the Coast of Misano Adriatico, Italy. *J. Mar. Sci. Eng.* **2016**, *8*, 126. [CrossRef]
36. European Space Agency. European Space Agency, 2019b. ESA Sentinel 2 Orbit Description. Available online: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/satellite-description/orbit (accessed on 29 September 2019).
37. European Space Agency. Sentinel-2 MSI Technical Guide. Available online: https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi (accessed on 10 September 2019).
38. Nowakowski, T. *Arianespace Successfully Launches Europe's Sentinel-2A Earth Observation Satellite*; Spaceflight insider: Kourou, French Guiana, 2015.

39. Kaplan, G.; Avdan, U. Object-based water body extraction model using Sentinel-2 satellite imagery. *Eur. J. Remote Sens.* **2017**, *50*, 137–143. [CrossRef]
40. Poursanidis, D.; Traganos, D.; Reinartz, P.; Chrysoulakis, N. On the use of Sentinel-2 for coastal habitat mapping and satellite-derived bathymetry estimation using downscaled coastal aerosol band. *Int. J. Appl. Earth Obs.* **2019**, *80*, 58–70. [CrossRef]
41. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm.* **2018**, *146*, 305–319. [CrossRef]
42. ESA SNAP. Available online: https://step.esa.int/main/toolboxes/snap (accessed on 12 October 2019).
43. Stumpf, R.P.; Holderied, K.; Sinclair, M. Determination of Water Depth with High- Resolution Satellite Imagery over Variable Bottom Types. *Limnol. Oceanogr.* **2003**, *48*, 547–556. [CrossRef]
44. Vapnik, V.; Golowich, S.E.; Smola, A. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; pp. 281–287.
45. Clarke, S.M.; Griebsch, J.H.; Simpson, T.W. Analysis of support vector regression for approximation of complex engineering analyses. *J. Mech. Des. Trans. ASME* **2005**, *12*, 1077–1087. [CrossRef]
46. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
47. Kranjčić, N.; Medak, D.; Župan, R.; Rezo, M. Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns. *Remote Sens.* **2019**, *11*, 655.
48. Kuhn, M. Classification and regression training. In *R Package Version 6.0–24*; 2014; Available online: https://ui.adsabs.harvard.edu/abs/2015ascl.soft05003K/abstract (accessed on 26 June 2020).
49. Parameswaran, S.; Weinberger, K.Q. Large margin multi-task metric learning. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.57: Red Hook, NY, USA, 2010; pp. 1867–1875.
50. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139. [CrossRef]
51. Cowan, G. *Statistical Data Analysis*; Oxford University Press: Oxford, UK, 1998.
52. Benveniste, J.; Cazenave, A.; Vignudelli, S.; Fenoglio-Marc, L.; Shah, R.; Almar, R.; Andersen, O.; Birol, F.; Bonnefond, P.; Bouffard, J.; et al. Requirements for a Coastal Hazards Observing System. *Front. Mar. Sci.* **2019**, *6*, 348. [CrossRef]
53. Geyman, E.C.; Maloof, A.C. A simple method for extracting water depth from multispectral satellite imagery in regions of variable bottom type. *Earth Space Sci.* **2019**, *6*, 527–537. [CrossRef]

energies

MDPI

Article

Determination of Water Depth in Ports Using Satellite Data Based on Machine Learning Algorithms

Vanesa Mateo-Pérez [1], Marina Corral-Bobadilla [2,*], Francisco Ortega-Fernández [1] and Vicente Rodríguez-Montequín [1]

[1] Project Engineering Department, University of Oviedo, 33004 Oviedo, Spain; mateovanesa@uniovi.es (V.M.-P.); fdeasis@uniovi.es (F.O.-F.); montequi@uniovi.es (V.R.-M.)
[2] Department of Mechanical Engineering, University of La Rioja, 26004 Logroño, Spain
* Correspondence: marina.corral@unirioja.es; Tel.: +34-941-299-274

Abstract: One of the fundamental maintenance tasks of ports is the periodic dredging of them. This is necessary to guarantee a minimum draft that will enable ships to access ports safely. The determination of bathymetries is the instrument that determines the need for dredging and permits an analysis of the behavior of the port bottom over time, in order to achieve adequate water depth. Satellite data processing to predict environmental parameters is used increasingly. Based on satellite data and using different machine learning algorithm techniques, this study has sought to estimate the seabed in ports, taking into account the fact that the port areas are strongly anthropized areas. The algorithms that were used were Support Vector Machine (SVM), Random Forest (RF) and the Multi-Adaptive Regression Splines (MARS). The study was carried out in the ports of Candás and Luarca in the Principality of Asturias. In order to validate the results obtained, data was acquired in situ by using a single beam provided. The results show that this type of methodology can be used to estimate coastal bathymetry. However, when deciding which system was best, priority was given to simplicity and robustness. The results of the SVM and RF algorithms outperform those of the MARS. RF performs better in Candás with a mean absolute error (MAE) of 0.27 cm, whereas SVM performs better in Luarca with a mean absolute error of 0.37 cm. It is suggested that this approach is suitable as a simpler and more cost-effective rough resolution alternative, for estimating the depth of turbid water in ports, than single-beam sonar, which is labor-intensive and polluting.

Keywords: environmental monitoring; satellite remote sensing; Sentinel-2; SVM; RF; MARS; polluted water; bathymetry

check for updates

Citation: Mateo-Pérez, V.; Corral-Bobadilla, M.; Ortega-Fernández, F.; Rodríguez-Montequín, V. Determination of Water Depth in Ports Using Satellite Data Based on Machine Learning Algorithms. *Energies* **2021**, *14*, 2486. https://doi.org/10.3390/en14092486

Academic Editors: Sandro Nizetic, Gwanggil Jeon and William Holderbaum

Received: 3 March 2021
Accepted: 25 April 2021
Published: 27 April 2021

1. Introduction

The bathymetry of coastal zones is important for many applications. These include navigation, infrastructure maintenance, dredging planning, managing the environment, hydrographic applications and coastal engineering sciences [1–4]. Sediment deposition and erosion in these shallow areas occurs frequently due to tides, wave propagation and intensive human activities [5]. Management of and planning for these areas of endeavor require updated and accurate information. In turn, this requires efficient technologies to record these never-ending changes. Although detailed information of the seabed is essential worldwide for the management of coastal environments, there are still economical and logistical constraints. In the specific actions program for ports and port facilities, dredging maintenance of docks and navigation channels is undertaken as actions that are necessary to guarantee navigation within ports and operation of their infrastructure and facilities. Most ports have dredging channels that experience sedimentation and reduce the depth of water that is available for navigation [6]. Ports operate with a minimum draft that is necessary to accommodate the ships. Most ports need maintenance dredging at some point to improve and facilitate navigation and for the development and maintenance of infrastructures in the marine and fluvial environment [7,8]. It is necessary to update the

bathymetries periodical. In order to minimize unnecessary or excessive dredging and the associated expense [9], it is important to determine and model seafloor levels accurately. Conventional methods of bathymetric data acquisition generally provide accurate depth profiles or point measurements along transects. However, they are limited by their logistical expenses and inefficiency in use. Further, they are difficult to apply in remote areas. Echo sounders are normally used to measure depths [10,11]. Single beam echo sounders are most commonly used for port studies. The measurement range of depths of these systems is from 0 to 5 m, but it can perform measurements greater than 5 m, a value well above what dredging operations require. The echo sounder measures the time for the signal from the transducer to reach the receiver, after being reflected by the background. By this means, it is possible to measure the depth of both the seabed and any object that is below the surface of the sea. This tool provides valid results in port studies. It will continue to do so, if the studies are well planned and executed. However, problems in precision and accuracy limit its use, as does the difficulty of using the tool for shallow coastal waters. Its operating costs are also high and its use requires many safety precautions [11]. These aspects increase the attractiveness of such alternative techniques as remote sensing, to provide reliable lower costing depth estimates [12–15]. Combining echo sounders' and satellite data provides another means to improve bathymetric estimation. Many topographic studies are carried out by remote sensing technology. This solves various problems that require studies of the depths on temporal and spatial scales that are impossible to achieve [16–19]. Some authors have developed simple methods that use optical images to estimate the depth of water. They include the use of linear regression logarithms [20–22].

Machine Learning (ML) techniques have become popular in recent years to estimate bathymetries using optical sensors. This is due to the increasing availability of new satellites and rapid advances in algorithm development and data availability [23]. For example, Neural Networks (NN) is a potential tool that numerous researchers have used recently in a wide variety of remote sensing applications [24–26]. Other researchers have been using Support Vector Machine (SVM) [27] as an alternative to NN to improve the performance of bathymetric recovery algorithms. It works well for nonlinear classification, time series prediction and regression [28]. Another non-linear regression algorithm, Random Forest (RF), is suitable for the construction of regression models that involve satellite images for bathymetry data [29–31]. Recently, Multi-Adaptive Regression Splines (MARS) has been used as relatively novel method for modelling and approximating nonlinear bathymetry measurements in shallow coastal areas [32]. These data-driven models are generally regarded as offering greater flexibility and accuracy in using satellite images to estimate water depths [33].

The literature reveals areas of possible future use of remotely sensed data in studies of water depth in clear shallow water [34,35]. However, the inherent conditions of ports that have highly polluted and turbid waters have often compromised the results that have been obtained. This may be related directly to the water's inherent optical properties (attenuation coefficient, dispersion, absorption . . .). This paper proposes a comparison of three different approaches for bathymetry estimation at two ports located at Candás and Luarca (Spain). The water depth estimation models were created by use of the Support Vector Machine, Random Forest and Multi-Adaptive Regression Splines methods. These proposed bathymetry methodologies were applied to Sentinel-2 images and compared to echo sounder depth data of the two study ports.

Previous studies have investigated the use of SVM techniques to estimate water depth in ports [36], the main advantages of machine learning methods include their reproducibility and their potential for continuous updating. In the current study, three machine learning techniques, SVM, MARS and RF, were compared to construct bathymetry maps using geographic information systems and remote sensing techniques, in order to provide the most efficient and simplest depth estimation model based on the accuracy of the resulting models for the ports. These three models were used with the main objective of using satellite data sets rather than extensive field studies. The novelty of this work is

the application of the methodology proposed in port areas. The fundamental difference between the port areas that have been analyzed in this work and other areas in which similar studies have been carried out, are the characteristics of the bottom, polluted and darker areas compared to light and sandy areas. In addition, these bathymetric maps in port areas can identify areas with accumulation of sediments, so that areas in need of dredging can be easily detected as well as being applied in the future in different ports. The intention is to provide a fast, operational and low cost alternative to traditional bathymetry with which to assess the need for port maintenance dredging.

2. Materials

2.1. Areas of Study and Field Measurements

The sites that were studied were the Port of Candás (43°35′25″ N to 5°45′43″ W) and the Port of Luarca (43°32′45″ N to 6°32′1″ W), located on the Cantabrian Sea coast (Bay of Biscay) on Spain's northern coast. The Port of Candás (Figure 1) has been the object of extensive rebuilding and extensions, although mainly since the 18th century. In the early 1950s, the dock was expanded leading to a gradual silting. This resulted in a gradual decrease in the draft of the port. Recently, work has been undertaken to improve this situation with new extensions to the port's levees. The port's traffic today consists of cargo vessels and recreational boats. However, the minimum draft in the port's operating varies from 1 m near the docks for small boats to 3.5 m for the navigation channel. The water in the access area reaches a depth of up to 5 m.

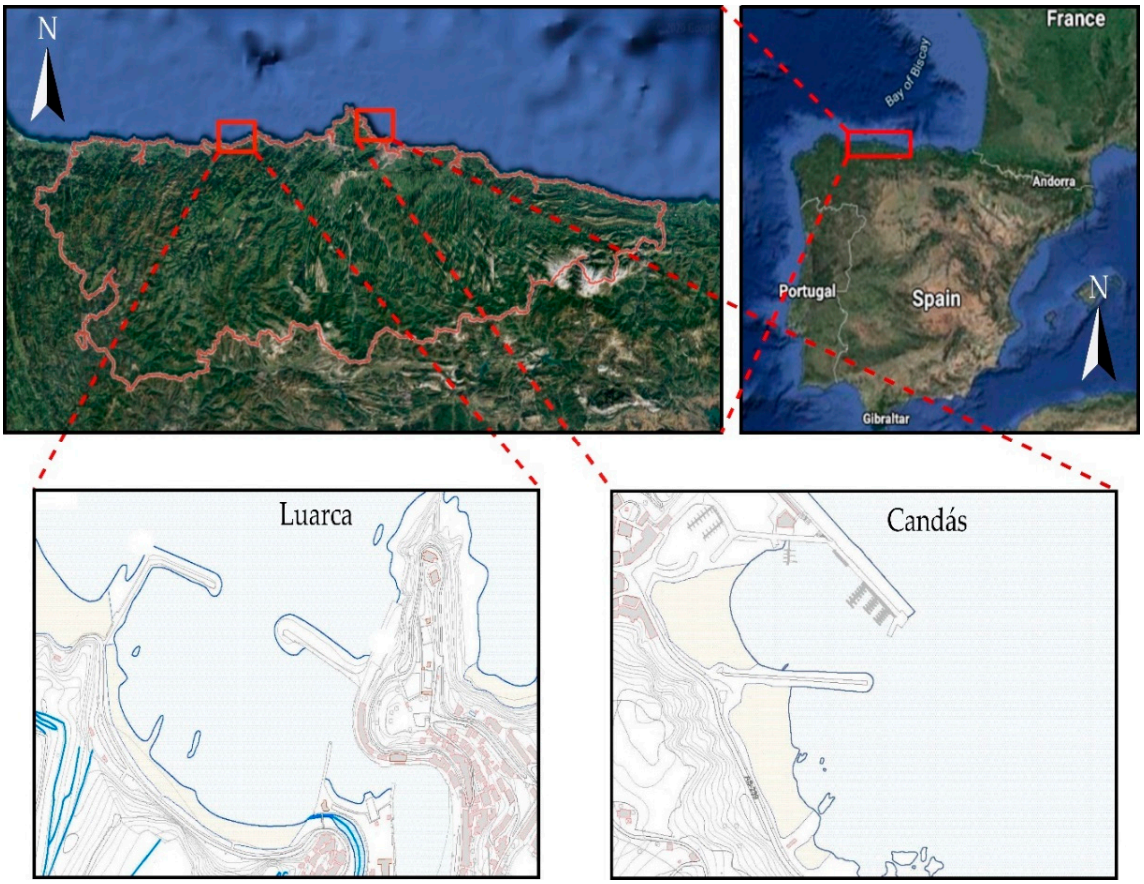


Figure 1. Study site locations in Spain coastal waters. Port of Candás (43°35′25″ N to 5°45′43″ W) and Port of Luarca (43°32′45″ N to 6°32′1″ W).

The second study site was the Port of Luarca (Figure 1). From its origin in the 10th century, Luarca has been linked to maritime activity and a fishing enclave. However, it

was not until the 20th century that Luarca's outer breakwaters were built. Because only small boats use the port, the minimum draft ranges from two meters in the docking area to three meters in the navigation canal and as much as twelve meters in the access area. The docking area and the navigation channel are dredged every year to maintain the necessary depth of water.

The bathymetry data of the Candás and Luarca ports were provided by the port service. The latter determines the quality of the water in the port each year and, also, any morphological changes. For both ports, the measurements were carried out using a Navisound 210 sounder (Reson, Inc.; Slangerup, Denmark) single beam echo, with a variable frequency acoustic profiler between 201 kHz and 33 kHz, and a 1 cm vertical precision. A Differential Global Positioning System (D-GPS) determined the position. Because the depth measurement data is affected by the tide, the depth measurement referred to the mean sea level. Table 1 provides the depth characteristics of the Port of Candás and Port of Luarca. They show a medium depth of −1.5519 m for Candás and −4.0694 m for Luarca. These negative values are in relation to the port's minimum level during the highest tide in the last 15 years. That level was recorded as zero.

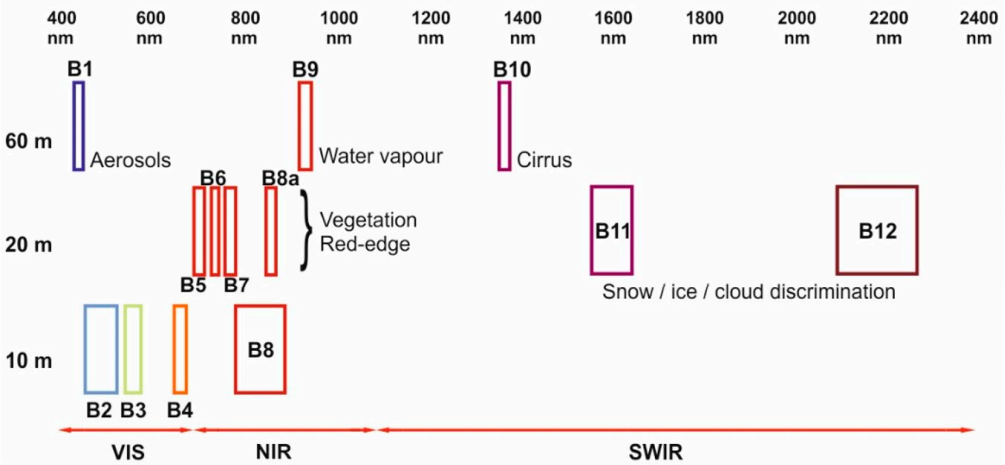
Table 1. Properties depth characteristics of Candás and Luarca ports.

Port	Max Depth (m)	Min Depth (m)	Mean Depth (m)
Candás	1.3461	−5.0149	−1.5519
Luarca	1.5979	−11.9601	−4.0694

The bathymetry elevations in this zone were referenced to the UTM/WGS84 projection ZONE 30N. The images were acquired on 16 October 2016; 12 March and 29 April 2019 for Candás and on 28 June 2016; 10 May 2018 and 18 May 2019 for Luarca. All images were acquired during calm weather. They were selected as the reference data set and compared to the satellite-delivered bathymetry products for each study area.

2.2. Satellite Data Acquisition

Sentinel-2 is the latest generation of the European Space Agency (ESA) [37]. The Copernicus program is an ambitious program for Earth observation than has been designed to obtain current and accurate information that can be accessed easily. The data from the Sentinel-2 satellite were used to predict the water depth of the study ports. The Sentinel-2 satellite conducted measurements in 13 spectral bands. Its spatial resolutions extended from 10 to 60. The spectral channels of the satellite include four bands of a 10 m spatial resolution. They were B2 (blue), B3 (green), B4 (red) and B8 (near infrared). There are also six bands of 20 m spatial resolution, four of which are used for the characterization of the vegetation in the bands B5, B6, B7 and B8a (red-edge). The two other bands are used for applications, such as the detection of clouds, snow or ice (B11 and B12). Finally, three bands of 60 m spatial resolution were used for atmospheric corrections and cloud screening. These were aerosols (B1), water vapor (B9) and cirrus detection (B10). The characteristics of specified spectral bands and their resolutions are shown in Figure 2. The satellite data were selected based on the proximity to the date of the in situ bathymetry and the least amount of clouds at the time of data acquisition (Figure 2).



	Dates of acquisition
Candás	03 November 2016, 16 March 2018 and 22 April 2019
Luarca	29 June 2016, 10 May 2018 and 30 May, 2019

Figure 2. Sentinel-2 spectral bands vs. spatial resolution, and dates of acquisition.

3. Methods

3.1. Pre-Processing of Satellite Images

Data captured by Sentinel-2A satellite, which currently is in orbit, is available without charge under an open license through portals, such as the Copernicus Open Hub. To visualize and preprocess Sentinel-2A data (10 m resolution), SNAP (Sentinel Application Platform) software (v7.0.1) was used [38]. This is an open source architecture that combines all toolboxes from the ESA. The data from the Sentinel-2A satellite reflectance bands (1, 2, 3, 4, 5, 6, 7, 8, 8A, 9, 11 and 12) were used to predict the depth of the water at the study ports. A resampling of all the spectral bands of the satellite images was performed, transforming the resolution of all of them to a resolution of 10 × 10 m [39,40]. This was done with SNAP software using the S2 Resampling Processor. As a result, a dataset without georeferencing was obtained, and to determine the positioning of the reference points, the geographical location of each point was defined by its longitude and latitude using the SNAP program. Then, using the WGS84 ellipsoid, a coordinate projection was created to, obtain the coordinates in ETRS89. This system was also used to project the positions that the echo sounder provides. The ellipsoid projections have an average position error of 1 cm. The data that was obtained was compared to the bathymetry that was projected. To accomplish this, a geodesic calculator was used to project the coordinates. This enabled the authors to obtain the data for bands that are associated with Universal Transverse Mercator (UTM). The annual in-situ bathymetries of the study ports that were provided for the Principality of Asturias Port Service were used to assign the z coordinate. Bathymetries were carried out by means of a single beam echo sounder located on a boat. Data are measured every 10 cm in each data acquisition beam. From the points, the surfaces were obtained using digital terrain models. Linear interpolation was used in the triangulation in this case. The error that was incurred in this process is not great due to the seabed's smooth surface and lack of significant irregularities. The z coordinates are related to the port's zero. The dimensions of each pixel are assigned on the basis of its former x–y location. The method of analysis and pre-processing the Sentinel-2 images is shown schematically in Figure 3.

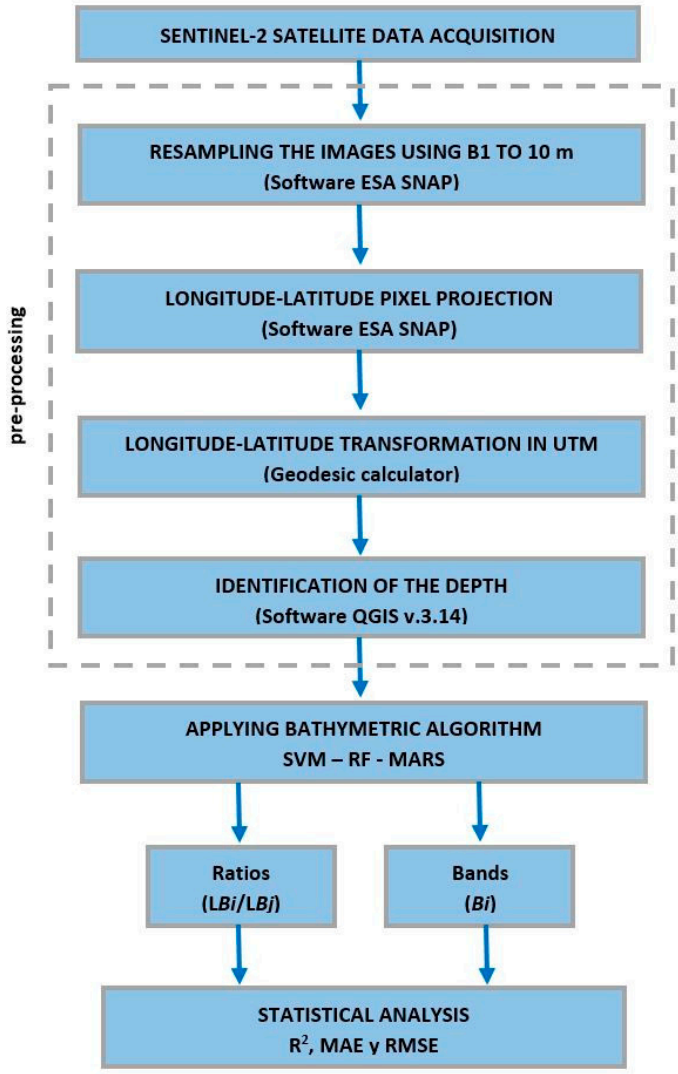


Figure 3. Methodological workflow for obtaining bathymetric maps from Sentinel-2 images.

3.2. Proposed Algorithms for Bathymetry Mapping

3.2.1. Support Vector Machines

The support vector machine is a widely used linear regression technique [41–44]. This technique provides higher accuracy when inputs are properly selected. SVM uses a kernel-based algorithm. Its new input estimations require the kernel function's evaluation of a subcategory of events during a training stage. The challenge with this method is in identifying a function to minimize Equation (1)'s final error.

$$y(x) = w^T \cdot \phi(x) + b \tag{1}$$

where $y(x)$ is the predicted value, w is a vector with parameters that the model defines, b is the value of the bias and $\phi(x)$ denotes the feature-space transformation. In this case, the error function i in the linear regression (Equation (1)) is replaced by an ϵ insensitive error function (Equation (2)). Equation (3) assigns a zero to value if ϵ exceeds the difference between the predicted value and target value. If the difference is equal to, or exceeds, ϵ, the error function's value remains unchanged. Equation (4) can be minimized by assigning a cost (C) to the difference between the predicted value and the targeted value.

$$\frac{1}{2}\sum_{n=1}^{n}[y_n - t_n]^2 + \frac{\lambda}{2}\|w\|^2 \tag{2}$$

$$E_\epsilon(y(x)-t) = \begin{cases} 0, & if\ |y(x)-t| < \epsilon \\ |y(x)-t|-\epsilon, & otherwise \end{cases} \tag{3}$$

$$C\sum_{n=1}^{n} E_\epsilon(y(x_n)-t_n) + \frac{1}{2}\|w\|^2 \tag{4}$$

where, ϵ is the margin if the function fails to impose a penalty, t represents the searched target function, C is the penalty and $y(x)$ is the value that Equation (1) predicted. The final function resembles Equation (5).

$$y(x) = \sum_{n=1}^{n}(\alpha_i - \alpha_i{}^*)k(x_i,\ x) + b \tag{5}$$

where α is a solution for the occasionally encountered optimization problem with the Lagrangian Theory.

The Gaussian Radial Basis Function (RBF) is generally the best kernel. It ensures the highest overall accuracy and Kappa [45]. The RBF function was used in this study (Equation (6)).

$$k(x_i,\ x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}} \tag{6}$$

The SVM was conducted in the R statistical computing environment using the "e1071" package (version 1.7-1, The R Foundation for Statistical Computing, Vienna, Austria) [46].

3.2.2. Random Forest

Random Forest refers to a model that was developed by Breiman [47]. It provides classification and regression. That creates model numerous classification trees by use of a randomized subset of predictors [48]. The algorithm grows many of these trees. Each tree begins as a sample of training data. The tree building process involves a random subset of predictor variables, which are used at each fork in the process. Thus, each tree is unique. The basic principle employed is that each tree is a poor predictor, but any pair of trees provide very different responses, thereby aggregating the predictions of uncorrelated trees. This reduces the prediction variance and improves accuracy [49–51]. The number of trees in this work was set at 100. Fifteen randomly selected variables were assigned to ted for each node. The minimum size of nodes was set at the default values. The RF algorithm was implemented by the Random Forest (v 4.6-2, The R Foundation for Statistical Computing, Vienna, Austria) [52] R package [46], to predicted bathymetry maps.

3.2.3. Multi-Adaptive Regression Splines

The Multi-Adaptive Regression Splines algorithm [53] is a nonparametric multiple regression method that uses adaptively selected spline functions [54]. Although based on linear relationships, it identifies and simulates a model with coefficients that change with changes in the predictor variable's level [45]. Depth water models were constructed by use of the Earth package [55] under R environment [46].

The MARS principle is based on the linear basis functions of Equations (7) and (8).

$$|x-c|_+ = max(0, x-t) = \begin{cases} x-c & x > c \\ 0 & x \le c \end{cases} \tag{7}$$

$$|c-x|_+ = max(0, c-x) = \begin{cases} c-x & x < c \\ 0 & x \ge c \end{cases} \tag{8}$$

where c is the connecting knot or intersection between successive splines.

The MARS model takes the form of a linear combination of these basis functions, as Equation (9) [56].

$$y = \beta_o + \sum_{i=1}^{n} \beta_i\ B_i(x) \tag{9}$$

where $B_i(x)$ is the basis functions, β_o is the bias, and β_i is the coefficients of basic functions that are calculated by a least square method. The letter n is number of terms in the model and is calculated after two successive steps.

3.3. Data Processing

In data mining, data preparation is one of the essential steps for modeling. In this study the collected dataset from the two ports was randomly divided into two parts, in order to properly validate the models, 80% of the data were used for training and the remaining 20% for testing. The data selection was carried out randomly and was tested with 5 different data sets, the mean of the 5 tests was taken as error. The data points for the port of Candás were divided into 1092 model generation points and 284 validation points. The data points that were used for the Luarca model were 1593 and 388 for training and testing respectively. The training and testing data set was randomly selected, five different groups of random data were generated, validating that the dataset was homogeneous. In this case, this validation system was used, although other authors use cross-validation systems [57,58]. It is necessary to use a valid method for implementing the remote sensing bathymetric measurements of an area from optical images. Thus, to choose the appropriate variables and to decide on corrections to make to the images, a Principal Component Analysis (PCA) was conducted. PCA is based on the component substitution of the original data for spectral transformation [59]. It is used in this work to minimize repetitive information within strongly correlated Sentinel-2 bands and to produce a set of linearly uncorrelated variable values, which are known as principal components (PC). PC1 is considered to contain the greatest amount of information from an original multispectral image with the greatest variance (74.2% in this case), whereas PC2 explains 14.0% of the total variance.

Figure 4 represents the projected data by using the principal components, PC1 and PC2. It is apparent that there are two clearly differentiated data groups, a group of data from the port of Luarca (blue points) and a group with data from both ports. The reason for this is that the depth range in Candás is much less than in Luarca. Therefore, there is a relationship between the points in shallower areas and the B9 band and the tide. Also, there is an area of greater depths that is more related to the remaining bands.

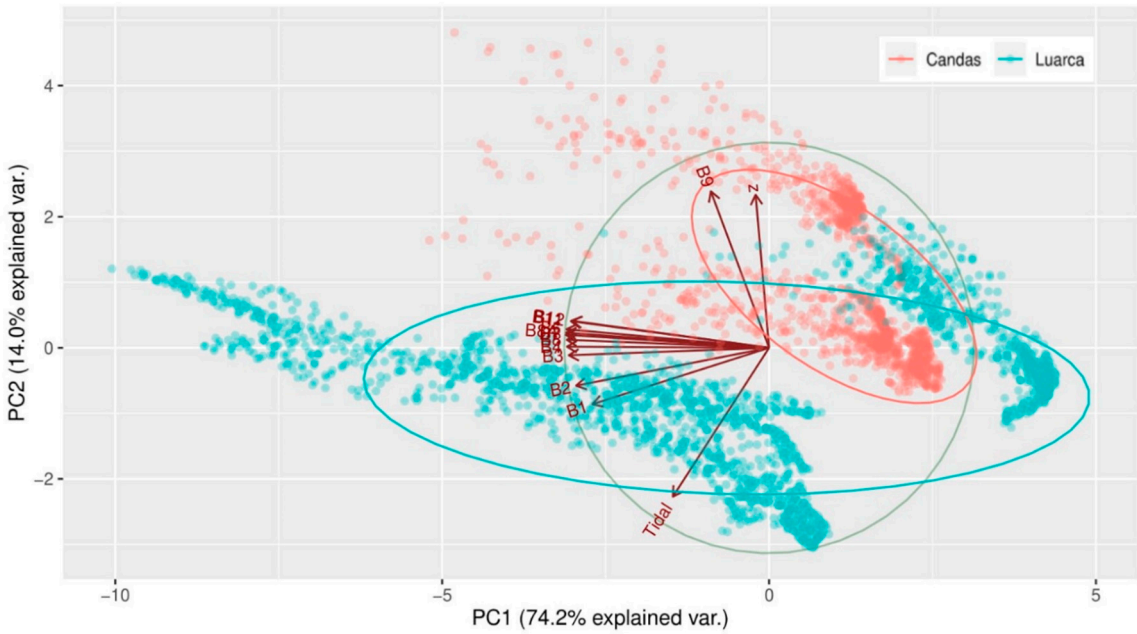


Figure 4. Principal component analysis of the variables.

In addition, Figure 4 shows a strong relationship between all bands, except for band B9. In remote sensing, the adjacent Sentinel-2 bands are correlated to each other. The correlation coefficients (R^2) between all variables studied appear in Figure 5. They indicate a strong statistical relationship among Sentinel-2 bands. A coefficient of correlation of 1.0 indicates a perfect correlation between the two variables. In contrast, a coefficient of 0.0 indicates that the two variables are not correlated at all [60]. Thus, the bands with the highest coefficient of correlation, were chosen for data modeling.

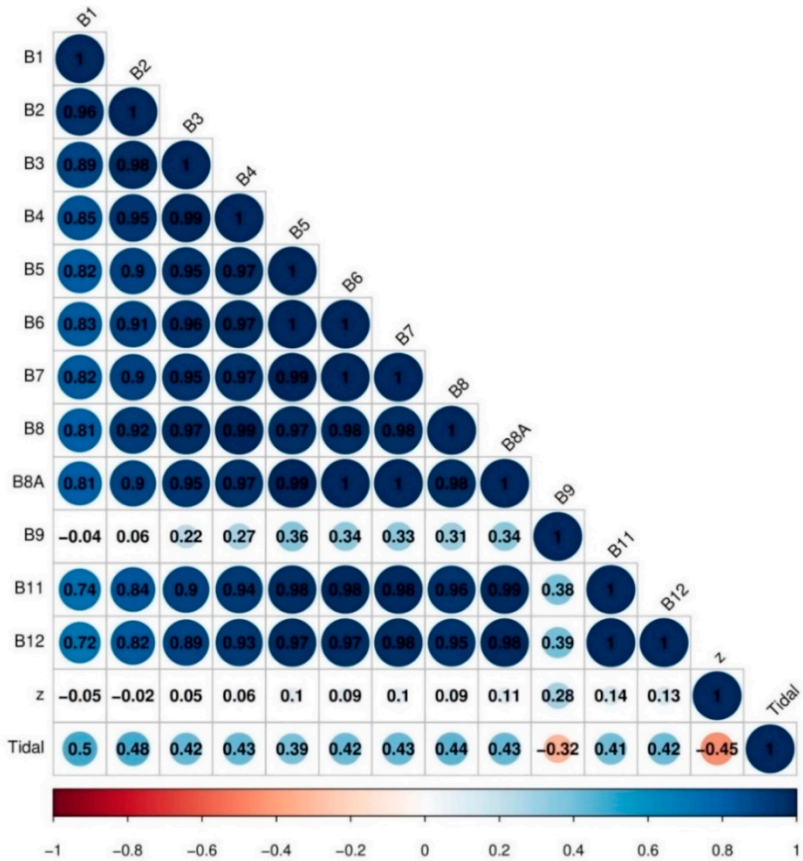


Figure 5. Correlation between Sentinel-2 bands.

Lyzenga [20] used two bands to offset the disadvantages of using a single-band linear correlation of reflectance ($R(\lambda_i)$) and water depths (Z). It is assumed that the column of water was uniform and the bottom's surface was homogenous (Equation (10)).

$$Z = b \log R(\lambda_i) + c \tag{10}$$

The ratio algorithm (Equation (11)) estimates depth without a need for bottom reflectance [61]

$$Z = m \frac{\log R(\lambda_i)}{\log R(\lambda_j)} + c \tag{11}$$

where $R(\lambda_i)$ and $R(\lambda_j)$ are reflectance in bands i and j. As in the case of a linear algorithm, information from any bands in the satellite image can be transformed into a multiple linear regression by Equation (12) [61].

$$Z = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} \frac{log\ R(\lambda_i)}{log\ R(\lambda_j)} + c_{ij} \tag{12}$$

Stumpf et al. [61] suggested a linear model, although it did not always reveal a linear relationship between water depth and the dates of satellites. It is better to obtain this by examining the relationship between a non-linear function and depth (Z), (Equation (13)).

$$Z = f\frac{ln[nR_w(\lambda_i)]}{ln[nR_w(\lambda_j)]} \tag{13}$$

where R_w is the observed reflectance of the wave length (λ) of bands i and j, and n is a fixed value.

Each algorithm (SVM, RF and MARS) was trained by using two different options, in which the input variables vary. In the first option, the (B_i) bands were used as input variables to obtain the bathymetry maps by analyzing water-leaving reflectance. This option had been used previously by several authors [30,31,42]. In the second option, a band ratio method was used to estimate water depths by using, as input variables, two radiance bands through the relationship that had been observed in Equation (14). This technique has been used by many researchers [27,62]. To date, the results that have been obtained by the two previously proposed options have not been compared. The analysis of the results will enable selection of the option that is best in selecting the input variables for each implemented algorithm and obtaining the simplest model.

4. Results

Three statistical metrics were used to compare the accuracies of the SVM, RF and MARS models. They were the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the correlation coefficient (adjusted R^2). These are calculated by the following Equations (14)–(16).

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Z_{Sentinel} - Z_{echo}| \tag{14}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_{Sentinel} - Z_{echo})^2} \tag{15}$$

$$R^2 = \frac{\sum_{i=1}^{N}(Z_{echo} - \overline{Z_{echo}})(Z_{Sentinel} - \overline{Z_{Sentinel}})}{\sqrt{\sum_{i=1}^{N}(Z_{echo} - \overline{Z_{echo}})^2(Z_{Sentinel} - \overline{Z_{Sentinel}})^2}} \tag{16}$$

where $Z_{Sentimel}$ are the depths that were predicted by the three proposed methodologies (SVM, RF and MARS) from satellite images. Z_{echo} is the in-situ echo sounding depths and N is the number of data.

In case of Candás Port, results of the testing data set appear in Table 2 (the best results appear in bolded fonts). The RF and SVM methods achieved good predictive performance. Thus, predictions of depths by the two models improved greatly there in comparison to MARS. All models were implemented. Their R^2, MAE and RMSE values were analyzed for both options, using the bands as input variables (Bands (B_i)) and using spectral band pairs with a high coefficient of determination estimated by Equation (13) as input variables (Ratios (LB_i/LB_j)).

Table 2 shows that the most consistent method, according to the correlation coefficients, was RF-Bands (0.92). This was followed closely by RF-Ratios (0.87), and SVM (0.85 and 0.74 for Ratios and Bands respectively). The MARS (0.62 for Ratios and 0.69 for Bands) gave the poorest predictive performance. The best performance for MAE in each of the proposed algorithms was achieved by using RF-Bands (0.27 m), followed by SVM-Ratios (0.34 m) and MARS-Ratios (0.50 m). According to RMSE, random forest models provided the most robust methodology, especially for RF-Bands to excellent performance (0.33 m). It was followed closely by SVM-Ratios (0.44 m). MARS (0.59 m) varied the most.

Table 2. Error statistics reported in meters for the SVM, RF and MARS algorithms and the Candás Port validation dataset.

Algorithm		R^2	MAE (m)	RMSE (m)
SVM (RBF kernel)	Ratios (LB_i/LB_j)	0.85	0.34	0.44
	Bands (B_i)	0.74	0.43	0.52
RF	Ratios (LB_i/LB_j)	0.87	0.32	0.39
	Bands (B_i)	0.92	0.27	0.33
MARS	Ratios (LB_i/LB_j)	0.62	0.51	0.60
	Bands (B_i)	0.69	0.50	0.59

Figure 6 shows the bathymetry map of Candás that was created by using the echo sounder measurements, and Figure 7 represents the bathymetry maps of Candás in a comparison and evaluation of the best performance of the three proposed methodologies that were obtained with ratio bands (LB_i/LB_j as input variables) and the SVM algorithm (Figure 7a), the RF algorithm (Figure 7b) and MARS algorithm (Figure 7c).

Figure 8 represents the water depth maps of Candás that were produced by using bands (B_i) as input variables and the SVM algorithm (Figure 8a), the RF algorithm (Figure 8b) and the MARS algorithm (Figure 8c). Figures 7 and 8b show that RF-Bands algorithm is very effective in prediction depths from satellite images. This algorithm produced the fewest errors. Figure 7b indicates that there are fewer areas of low points or high points. This corresponds to reality, as see in the bathymetry by an echo sounder (Figure 6), the transitions and slopes are smooth, which corresponds to the actual seabed. It can be concluded that the best results in the graphic representation are provided by Figures 7b and 8b and associated with Random Forest. However, it is the latter (Figure 8b) that is associated with modeling with the use of the bands, with similarity to the results of using the echo sounder measurements (Figure 6).

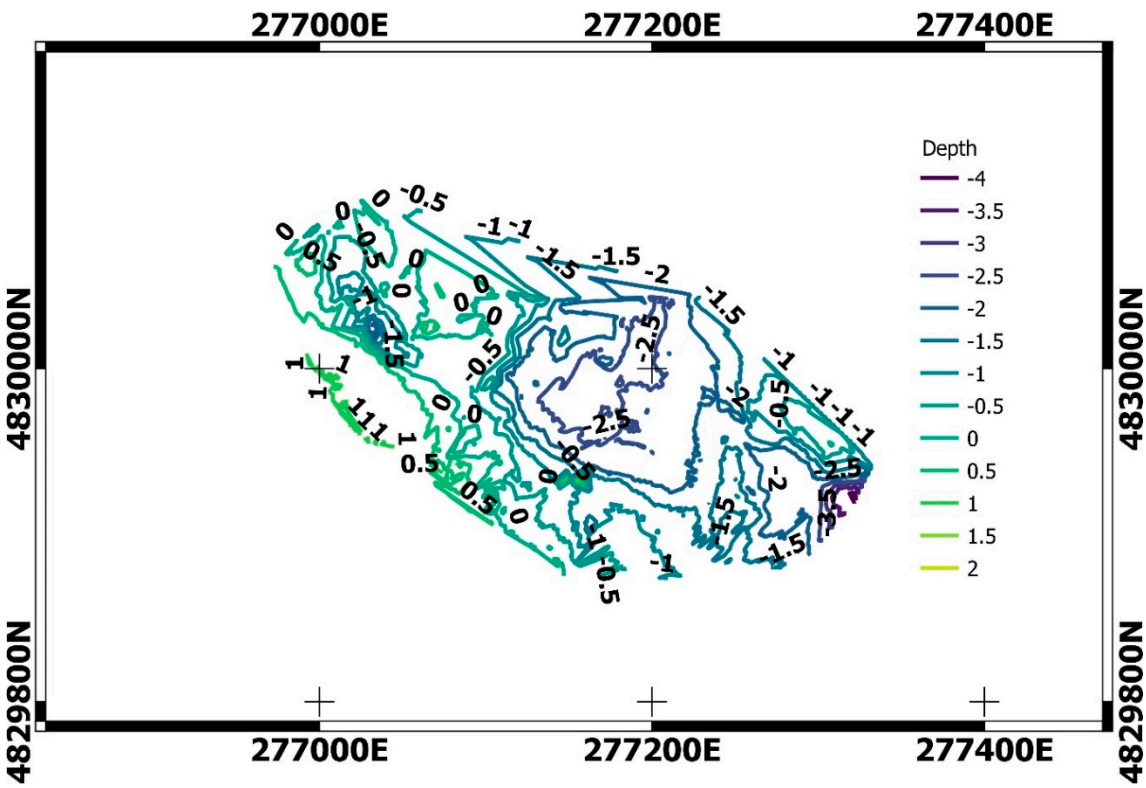


Figure 6. Water depth map obtained using echo sounder measurements in Candás. The color indicates the depth in meters.

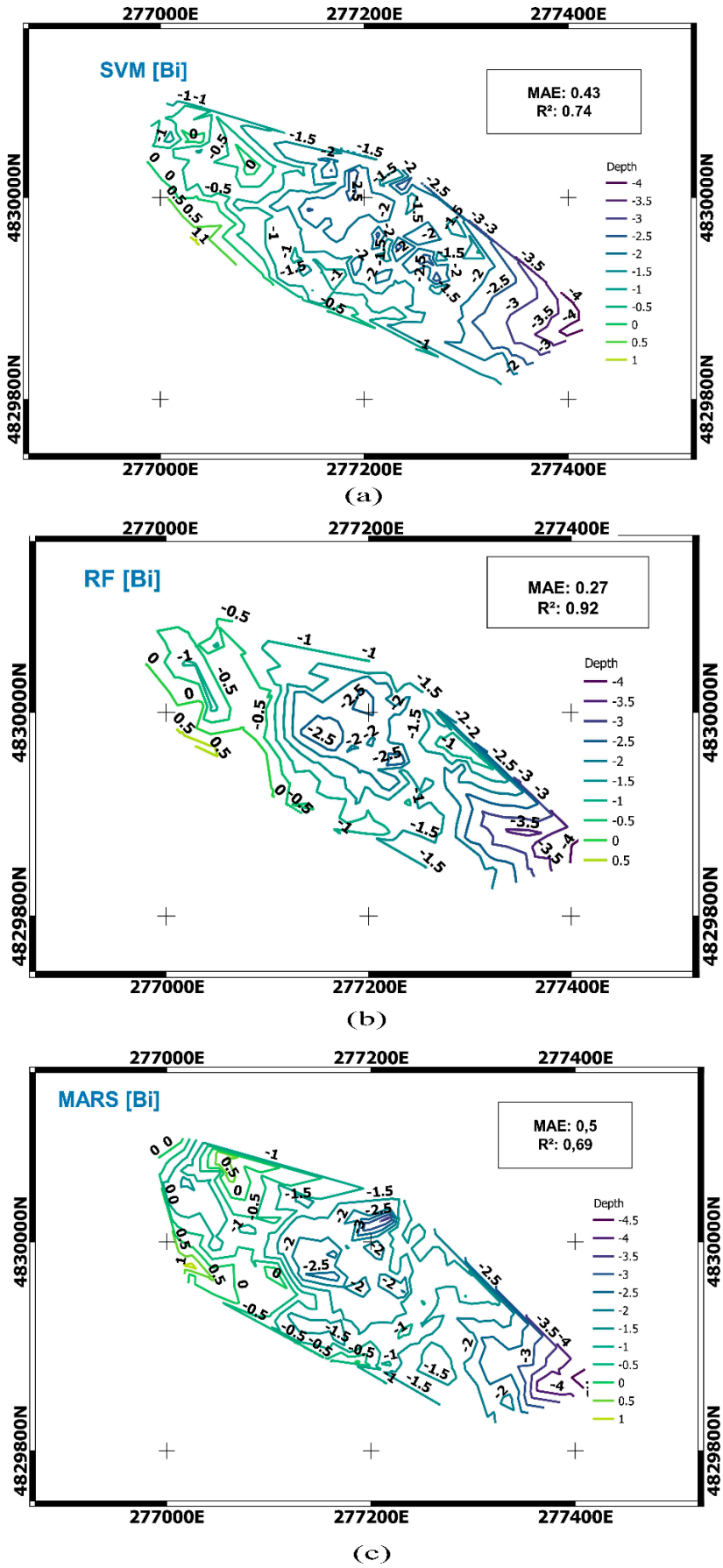


Figure 7. Comparison of water depth maps of Candás obtained using different algorithms (**a**) SVM, (**b**) RF and (**c**) MARS, using ratios bands as input variables. The color indicates the depth in meters.

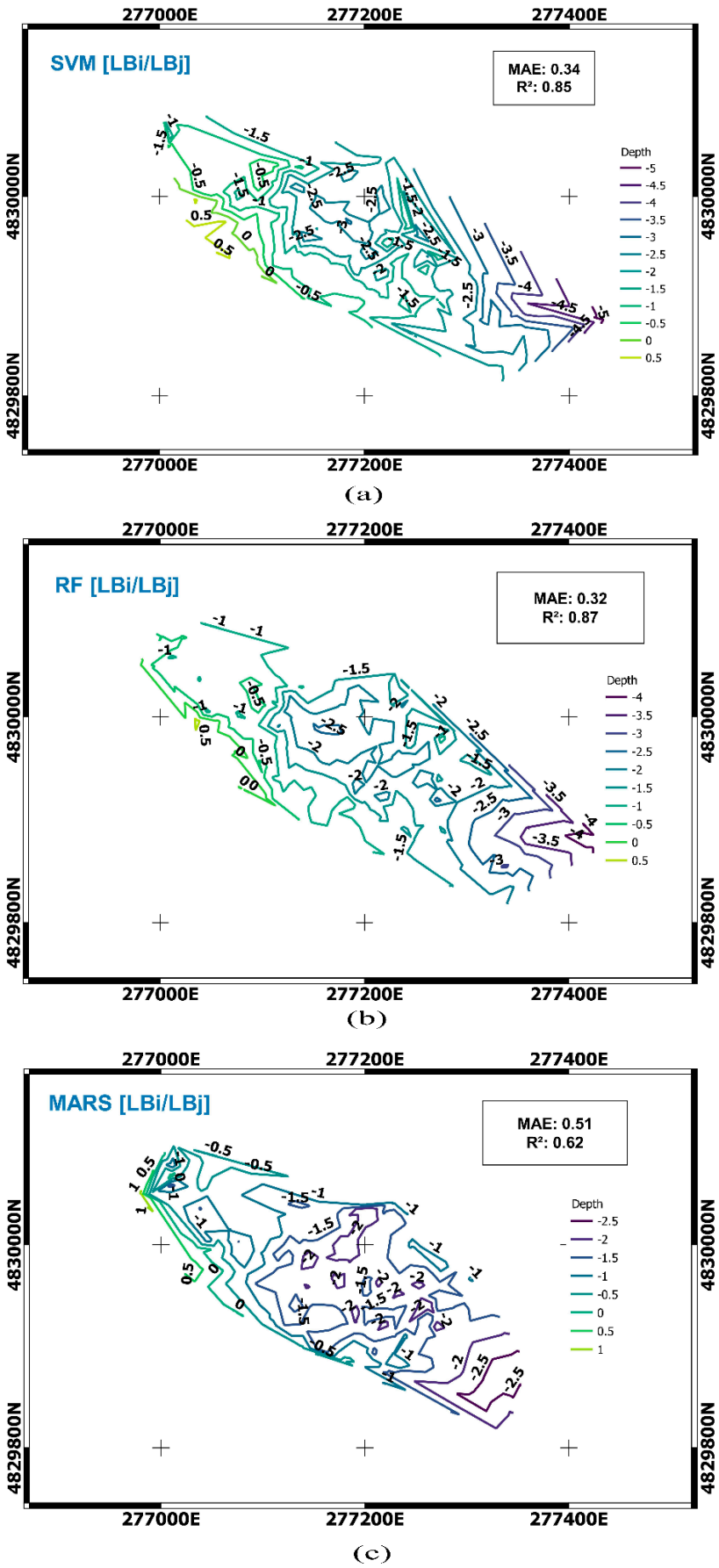


Figure 8. Comparison of water depth maps of Candás obtained using different algorithms (**a**) SVM, (**b**) RF and (**c**) MARS, using bands as input variables. The color indicates the depth in meters.

Testing dataset results for Port of Luarca appear in Table 3 (the best results are bolded). As for Candás, the RF and SVM methods provided good predictive performance in comparison to MARS. Table 3 shows that the most consistent method according to the correlation coefficients was RF-Bands (0.974). This was followed closely by SVM-Ratios (0.973), RF-Ratios (0.96) and SVM-Bands (0.96), whereas MARS-Ratios (0.95) gave the poorest predictive performance. The best performance for MAE in each of the proposed algorithms was achieved when using RF-Bands (0.37 m) and SVM-Ratios (0.37 m). This was followed by MARS-Bands (0.48 m). According to RMSE, SVM models had the most robust methodology, especially for Ratios, which gave an excellent performance (0.46 m), followed closely by RF-Bands (0.47 m). The performance of MARS (0.59 m) varied the most.

Table 3. Error statistics in meters for the SVM, RF and MARS algorithms and the Luarca Port testing dataset.

Algorithm		R^2	MAE (m)	RMSE (m)
SVM (RBF kernel)	Ratios (LB_i/LB_j)	0.973	0.37	0.46
	Bands	0.96	0.45	0.58
RF	Ratios (LB_i/LB_j)	0.96	0.41	0.56
	Bands	0.974	0.37	0.47
MARS	Ratios (LB_i/LB_j)	0.95	0.53	0.65
	Bands	0.96	0.48	0.59

Figure 9 shows the bathymetry map of Luarca that was created by using the echo sounder measurements, and Figure 10 represents the bathymetry maps of Luarca in a comparison and evaluation of the best performance of the three proposed methodologies that were created using ratios bands (LB_i/LB_j) as input variables, and the SVM algorithm (Figure 10a), the RF algorithm (Figure 10b) and the MARS algorithm (Figure 10c). Figure 11 shows the bathymetry maps of Luarca using bands (B_i) as input variables and the SVM algorithm (Figure 11a), the RF algorithm (Figure 11b) and the MARS algorithm (Figure 11c). From Figures 9–11, it can be concluded that the SVM and RF algorithms are a very effective predictors of depths from satellite images. In the methodologies (RF and SVM) that generated the best results, smooth curves can be seen that are substantially parallel to the beach.

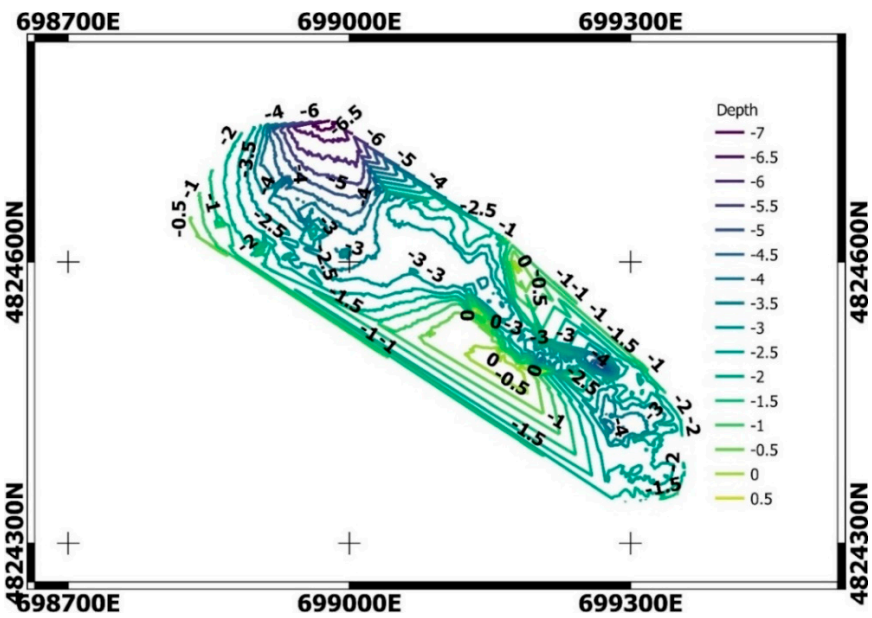


Figure 9. Water depth map obtained by using echo sounder measurements for Luarca. The color indicates the depth in meters.

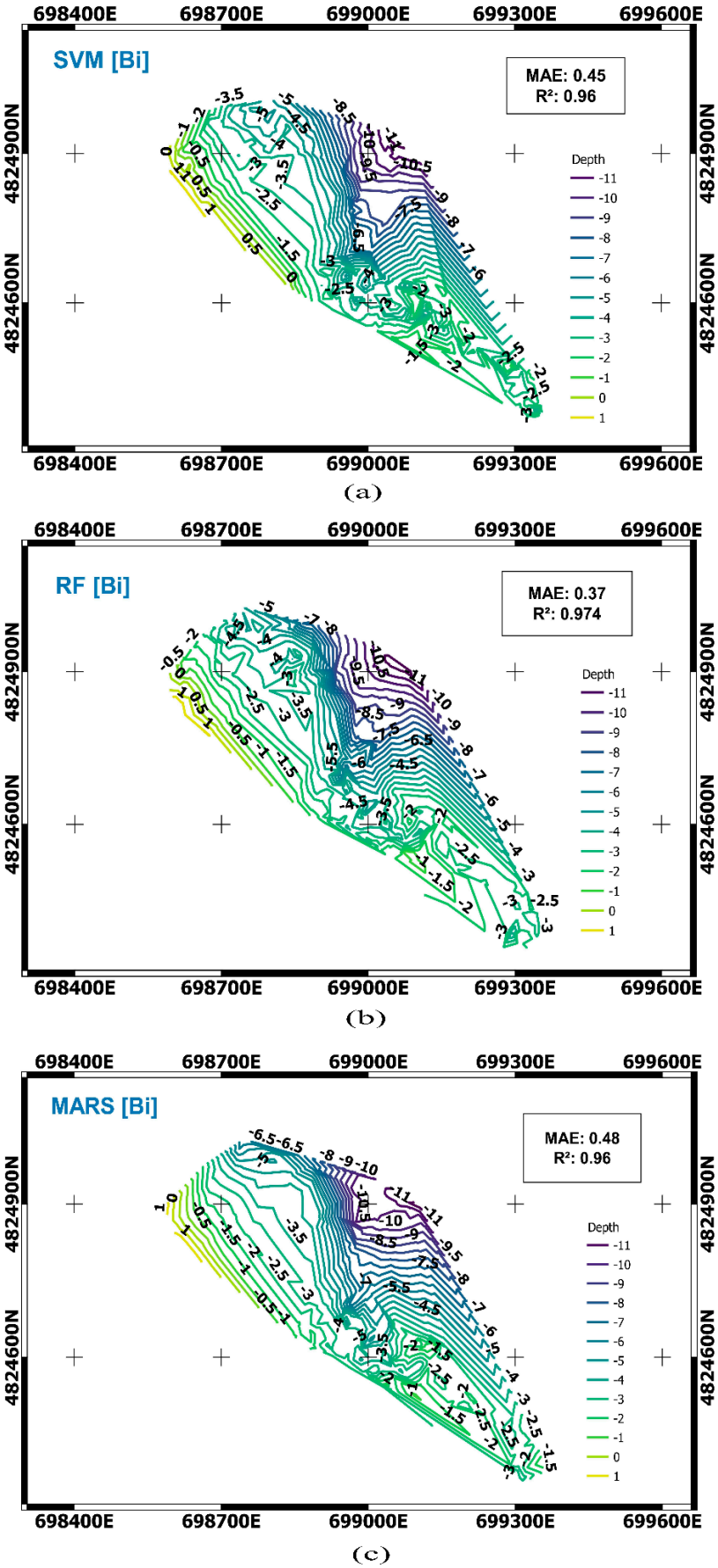


Figure 10. Comparison of water depth maps of Luarca obtained using different algorithms (**a**) SVM, (**b**) RF and (**c**) MARS, using ratios bands as input variables. The color indicates the depth in meters.

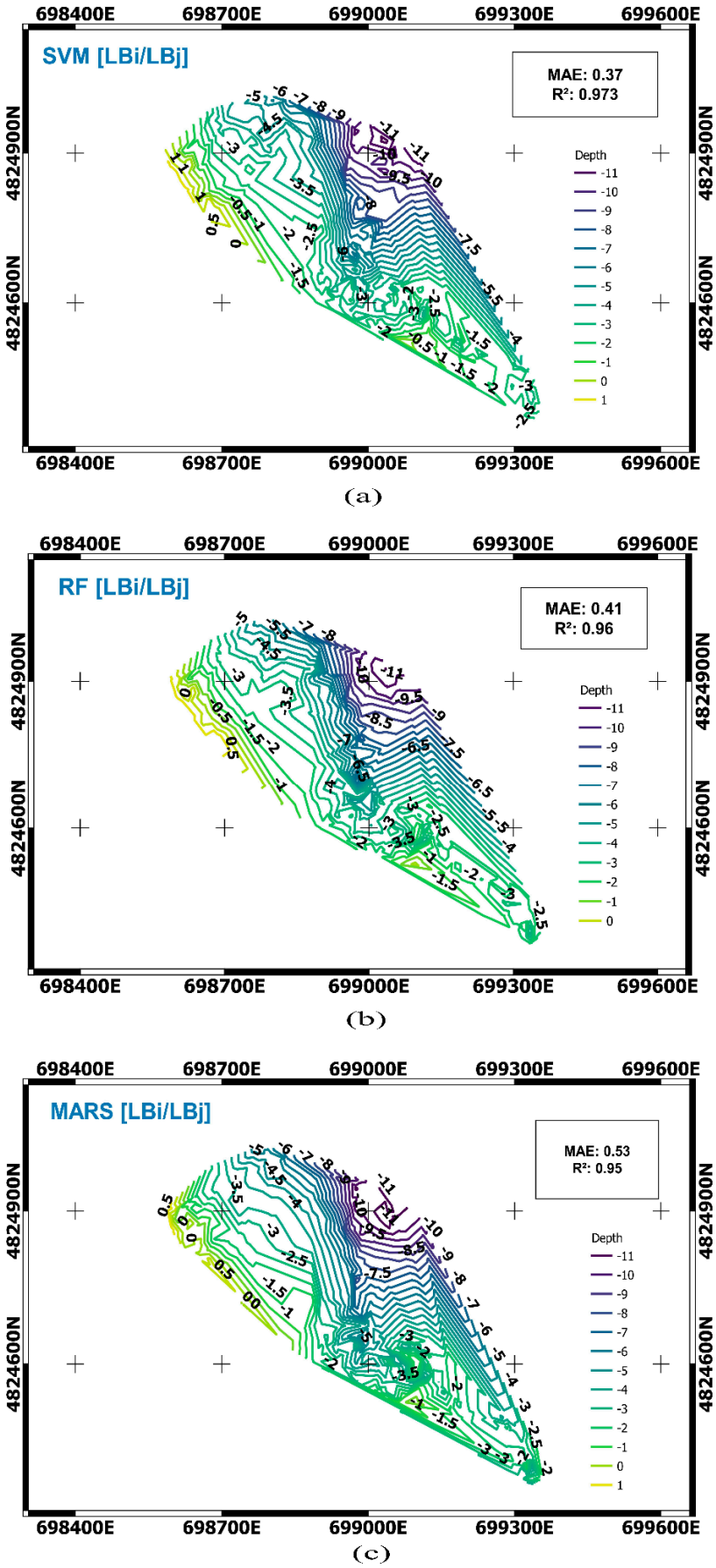


Figure 11. Comparison of water depth maps of Luarca obtained using different algorithms (**a**) SVM, (**b**) RF and (**c**) MARS, using bands as input variables. The color indicates the depth in meters.

The research that is presented in this work suggests that RF models detect depths in port waters well and can replace echo-sounded bathymetry measurements to measure port depths. The studies demonstrated that all methods provided good predictive performance of models. In Tables 2 and 3, it can be observed that the values of correlation coefficients using the three proposed algorithms are very close to 1.0 (which is very high). The worst R^2 was obtained when using the MARS algorithm for Candás. The results that were obtained with MARS provided the greatest error, although the latter is the algorithm that best adapts to the generation of surfaces. The reason is that the position of the points was not used as an input variable, due to a desire to obtain a model that is as general as possible. Thus, it is so generic that a single model has been created for two ports with different characteristics. In all of the bathymetric maps for all models (Figures 7, 8, 10 and 11), the algorithms identified correctly the deepest areas, areas where the depth is lowest with smooth transitions and coastal contour lines. Complex areas that are very shallow also were identified correctly.

In order to analyze the factors that may influence the errors, their relationship with depth was analyzed. The global errors of Candás and Luarca have been analyzed jointly. The results appear in Table 4. This table shows how in the studied models the error did not increase with increasing depth.

Table 4. A comparison of MAE error in the use of RF and SVM techniques.

	MAE (m)	
Depth Interval	**RF**	**SVM**
2 m to 0 m	0.32	0.4
0 m to −2 m	0.26	0.34
−2 m to −4 m	0.36	0.34
−4 m to −6 m	0.61	0.39
−6 m to −8 m	0.61	0.58
−8 m to −10 m	0.23	0.4
−10 m to −12 m	0.26	0.23

Figure 12 provides the histogram for the relationship between MAE errors and the depth interval in the port area for the algorithms of greatest accuracy; RF-bands and SVM-Ratios.

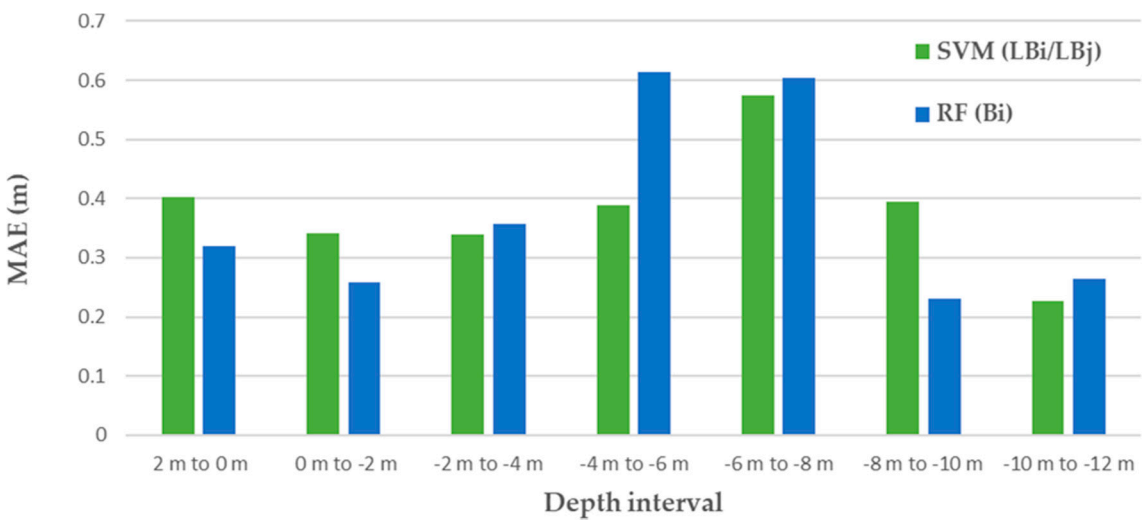


Figure 12. Variation of MAE errors versus depth interval in the port area for the two selected algorithms: RF-bands and SVM-Ratios.

As can be seen in Figure 12 and Table 4, the RF-Band technique gives better results in the depths between +2 m to −2 m and −8 m to −10 m. SVM-Ratio algorithms have better

results at between −4 m and −8 m and −10 to −12 m. In the range between −2 m and −4 m, the MAE results for the two algorithms are very similar.

5. Discussion

In machine learning, there is no single algorithm or solution that adapts to all the analyzed data, so it is quite common to work with several algorithms to find the best or most adjusted solutions. To our knowledge, no study previously has compared the use of SVM, RF, and MARS to study bathymetric mapping for the determination of depth in anthropized water areas, including ports that experience contamination and processes of accretion. Bathymetric mapping requires is highly precise design characteristics. The research that is presented in this work suggests that RF models detect depths in port waters well and can replace echo-sounded bathymetry measurements to measure port depths. The models that are presented in this study produced fewer errors in Candás than in Luarca. It is also necessary to consider that the bottoms of the port of Candás is sandy. In Luarca, there are areas of greater depth, although there are more rocky outcrops. The bathymetry obtained permits an understanding of the state of the port areas and where there is less draft and, therefore, a greater deposit of sediments. The latter makes it important to have this system of obtaining the bathymetry to reveal the areas where dredging is necessary for proper operation of the port. Thus, the main advantage of the technique that was implemented in this study is acquiring a greater understanding of the topography of the seabed at ports. The technique offers a high level of precision, applicability in areas of turbid and shallow water, rapidity of use and flexibility. Finally, the proposed method is economical. Thus, the results of this work offer a valuable contribution to the provision of useful information for the management of port maintenance dredging.

To compare the results that have been obtained with previous results from other authors studies, many variables, such as depth range, nature of bottom, image quality and water-quality should be considered [63]. Many authors [39,56,64] have applied SVM and RF methods to estimate the bathymetry in shallow water using satellite imagery. The errors that those have found are greater than those in this study. The cause could be the color and turbidity, because the bottoms were muddy and contaminated. They absorbed more light than did the sandy bottoms below waters of high transparency that many studies have examined. However, the errors obtained in this work are lower than those obtained in [32], where the authors proposed empirical approaches of bathymetry estimations in different locations with a silt-sand bottom water area, and a high-turbidity, clay bottom area. Using free and open-access satellite data that does not provide a resolution that is less than what other authors experiences and the satellites that they used may also have affected these results. However, unlike the studies cited above, it is important to note that the proposed methodology is applied to study anthropized water areas, not coastal areas whose bottoms are cleaner and the waters are clearer. In fact, it is in these types of areas where this system is most useful since it would allow the analysis of the sedimentation process that occurs in ports.

This study confirms the viability of machine learning models using Sentinel-2 images, we have proposed a methodology to build the best performance model that could be applied to different anthropized water areas. Sentinel-2 images can be used effectively to determine bathymetry in the study area, and this methodology could be extended to different ports in the Principality of Asturias that have similar characteristics for water. The use of this methodology could also be extended to other ports in the world. In future work, it could be interesting to analyze the composition of the seabed looking for different algorithms based on that composition.

6. Conclusions

This approach brings a new perspective in the subject of determination of water depth in ports using remote sensing technology, this technology is considered a time-effective, low-cost, and wide-coverage solution. It is also a supplement and improvement

to traditional bathymetric measurement methods and techniques. This study compared SVM, RF and MARS methods of bathymetry prediction using Sentinel-2 images, in order to propose a simple and robust model for bathymetry mapping in anthropized water areas. This depth estimation is needed for the dredging processes, especially for maintaining the free draft and adequate port management, also the analysis of the behavior of the bottom of the ports provides valuable information the knowledge of their behavior in the face of littoral dynamics. The algorithms were applied in two different ports—in Candás and Luarca (Asturias, Spain)—with different numbers of available data points. The depths that were determined were compared to those that were produced by a depth sounder in-situ measurements. The three proposed approaches used bands and bands logarithms ratios as input variables. The errors obtained were admissible since the oscillations in the background due to the storms have an order of magnitude greater than the errors obtained in the models.

At Candás, the RF method provided the best bathymetry predictions. Further, its results were most consistent, according to the correlation coefficients RF-Bands (0.92). The method's best performance was achieved by the use of the RF-Bands (0.27 m). The Random Forest model was the most robust methodology for RMSE, especially for the Bands option where it gave an excellent performance (0.33 m). In case of Luarca, the coefficient of determination that was obtained was very strong in case of RF-Bands (0.974). It is closely followed by SVM-Ratios (0.973). For MAE, the highest performance was achieved using RF-Bands and SVM-Ratios (0.37 m). For RMSE, the SVM models had the most robust methodology, especially for Ratios, with an excellent performance (0.46 m), closely followed by RF-Bands (0.47 m).

The results that the RF and SVM algorithms provided exceeded those of the MARS algorithm. In addition, the RF method produced results that were more accurate in Candás, and the SVM method in Luarca. The difference in results between the two models is very small. It should be noted that the best RF result was obtained from the Bands. For SVM, the best result came from the use of the Ratios. Therefore, in order to choose a single model, RF is considered best due to its simplicity and its need for fewer input variables. Validation method used in this work use randomly chosen training and testing sets, which ignore spatial autocorrelation (SAC) in data. This may lead to overoptimistic assessment of model predictive power [58]. Our intention is to address SAC properly in future studies. Also, in future work, a very interesting option could be to use both techniques, applying one or the other depending on the previous depth result.

Author Contributions: Conceptualization, V.M.-P. and F.O.-F.; software and validation, V.M.-P.; methodology, M.C.-B.; data curation, V.M.-P.; writing—original draft preparation, M.C.-B.; writing—review and editing, F.O.-F., and V.R.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science, Technology and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is part-funded by the European Regional Development Fund (ERDF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the Port Service and Transport Infrastructures of the Principality of Asturias for their collaboration in this work.

Conflicts of Interest: The authors declare there are no conflict of interest.

References

1. Caballero, I.; Stumpf, R.P. Retrieval of nearshore bathymetry from Sentinel-2A and 2B satellites in South Florida coastal waters. *Estuar. Coast. Shelf Sci.* **2019**, *226*, 106277. [CrossRef]
2. Clementi, E.; Oddo, P.; Drudi, M.; Pinardi, N.; Korres, G.; Grandi, A. Coupling hydrodynamic and wave models: First step and sensitivity experiments in the Mediterranean Sea. *Ocean Dyn.* **2017**, *67*, 1293–1312. [CrossRef]

3. Murray, R.B.O.; Gallego, A. Data review and the development of realistic tidal and wave energy scenarios for numerical modelling of Orkney Islands waters, Scotland. *Ocean Coast. Manag.* **2017**, *147*, 6–20. [CrossRef]
4. Tronvig, K.A. Near-shore bathymetry. *Hydro Int.* **2005**, *9*, 24–25.
5. Ceyhun, Ö.; Yalçın, A. Remote sensing of water depths in shallow waters via artificial neural networks. *Estuar. Coast. Shelf Sci.* **2010**, *89*, 89–96. [CrossRef]
6. Norén, A.; Fedje, K.K.; Strömvall, A.-M.; Rauch, S.; Andersson-Sköld, Y. Integrated assessment of management strategies for metal-contaminated dredged sediments—What are the best approaches for ports, marinas and waterways? *Sci. Total Environ.* **2020**, *716*, 135510. [CrossRef] [PubMed]
7. Wang, L.; Chen, L.; Tsang, D.C.W.; Li, J.-S.; Baek, K.; Hou, D.; Ding, S.; Poon, C.-S. Recycling dredged sediment into fill materials, partition blocks, and paving blocks: Technical and economic assessment. *J. Clean. Prod.* **2018**, *199*, 69–76. [CrossRef]
8. Cecchi, G.; Vagge, G.; Cutroneo, L.; Greco, G.; Di Piazza, S.; Faga, M.; Zotti, M.; Capello, M. Fungi as potential tool for polluted port sediment remediation. *Environ. Sci. Pollut. Res.* **2019**, *26*, 35602–35609. [CrossRef] [PubMed]
9. EL-Hattab, A.I. Single beam bathymetric data modelling techniques for accurate maintenance dredging. *Egypt. J. Remote Sens. Space Sci.* **2014**, *17*, 189–195. [CrossRef]
10. Kearns, T.A.; Breman, J. Bathymetry-The art and science of seafloor modeling for modern applications. *Ocean Globe* **2010**, 1–36.
11. Coggins, L.X.; Ghadouani, A. High-Resolution Bathymetry Mapping of Water Bodies: Development and Implementation. *Front. Earth Sci.* **2019**, *7*, 330. [CrossRef]
12. Lyzenga, D.R.; Malinas, N.P.; Tanis, F.J. Multispectral bathymetry using a simple physically based algorithm. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2251–2259. [CrossRef]
13. Gao, J. Bathymetric mapping by means of remote sensing: Methods, accuracy and limitations. *Prog. Phys. Geogr. Earth Environ.* **2009**, *33*, 103–116. [CrossRef]
14. Sánchez-Carnero, N.; Ojeda-Zujar, J.; Rodríguez-Pérez, D.; Marquez-Perez, J. Assessment of different models for bathymetry calculation using SPOT multispectral images in a high-turbidity area: The mouth of the Guadiana Estuary. *Int. J. Remote Sens.* **2014**, *35*, 493–514. [CrossRef]
15. Chust, G.; Grande, M.; Galparsoro, I.; Uriarte, A.; Borja, Á. Capabilities of the bathymetric Hawk Eye LiDAR for coastal habitat mapping: A case study within a Basque estuary. *Estuar. Coast. Shelf Sci.* **2010**, *89*, 200–213. [CrossRef]
16. Flener, C.; Lotsari, E.; Alho, P.; Käyhkö, J. Comparison of empirical and theoretical remote sensing based bathymetry models in river environments. *River Res. Appl.* **2012**, *28*, 118–133. [CrossRef]
17. Giardino, C.; Bresciani, M.; Matta, E.; Brando, V.E. Imaging Spectrometry of Inland Water Quality in Italy Using MIVIS: An Overview. In *Advances in Watershed Science and Assessment*; Younos, T., Parece, T.E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 33, pp. 61–83. ISBN 978-3-319-14211-1.
18. Jawak, S.D.; Vadlamani, S.S.; Luis, A.J. A Synoptic Review on Deriving Bathymetry Information Using Remote Sensing Technologies: Models, Methods and Comparisons. *Adv. Remote Sens.* **2015**, *04*, 147–162. [CrossRef]
19. Hedley, J.; Roelfsema, C.; Chollett, I.; Harborne, A.; Heron, S.; Weeks, S.; Skirving, W.; Strong, A.; Eakin, C.; Christensen, T.; et al. Remote Sensing of Coral Reefs for Monitoring and Management: A Review. *Remote Sens.* **2016**, *8*, 118. [CrossRef]
20. Lyzenga, D.R. Passive remote sensing techniques for mapping water depth and bottom features. *Appl. Opt.* **1978**, *17*, 379. [CrossRef]
21. Bierwirth, P.N.; Lee, T.J.; Burne, R.V. [Marine S.R.Corp., Washington, DC (United States) Michigan Environmental Research Inst.,. Ann Arbor (United States)]. Shallow sea-floor reflectance and water depth derived by unmixing multispectral imagery. *Photogramm. Eng. Remote Sens. USA* **1993**, *59*, 331–338.
22. Provost, J.-N.; Collet, C.; Rostaing, P.; Pérez, P.; Bouthemy, P. Hierarchical Markovian segmentation of multispectral images for the reconstruction of water depth maps. *Comput. Vis. Image Underst.* **2004**, *93*, 155–174. [CrossRef]
23. Peterson, K.; Sagan, V.; Sidike, P.; Cox, A.; Martinez, M. Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. *Remote Sens.* **2018**, *10*, 1503. [CrossRef]
24. Pahlevan, N.; Smith, B.; Schalles, J.; Binding, C.; Cao, Z.; Ma, R.; Alikas, K.; Kangro, K.; Gurlin, D.; Hà, N.; et al. Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sens. Environ.* **2020**, *240*, 111604. [CrossRef]
25. Liu, S.; Gao, Y.; Zheng, W.; Li, X. Performance of two neural network models in bathymetry. *Remote Sens. Lett.* **2015**, *6*, 321–330. [CrossRef]
26. El-Mewafi, M.; Salah, M.; Fawzi, B. Assessment of Optical Satellite Images for Bathymetry Estimation in Shallow Areas Using Artificial Neural Network Model. *Am. J. Geogr. Inf. Syst.* **2018**, *7*, 99–106.
27. Wang, L.; Liu, H.; Su, H.; Wang, J. Bathymetry retrieval from optical images with spatially distributed support vector machines. *GIScience Remote Sens.* **2019**, *56*, 323–337. [CrossRef]
28. Wang, W.; Men, C.; Lu, W. Online prediction model based on support vector machine. *Neurocomputing* **2008**, *71*, 550–558. [CrossRef]
29. Sagawa, T.; Yamashita, Y.; Okumura, T.; Yamanokuchi, T. Satellite Derived Bathymetry Using Machine Learning and Multi-Temporal Satellite Images. *Remote Sens.* **2019**, *11*, 1155. [CrossRef]

30. Yunus, A.P.; Dou, J.; Song, X.; Avtar, R. Improved Bathymetric Mapping of Coastal and Lake Environments Using Sentinel-2 and Landsat-8 Images. *Sensors* **2019**, *19*, 2788. [CrossRef]
31. Kogut, T.; Weistock, M. Classifying airborne bathymetry data using the Random Forest algorithm. *Remote Sens. Lett.* **2019**, *10*, 874–882. [CrossRef]
32. Mohamed, H.; Nadaoka, K.; Nakamura, T. Assessment of Machine Learning Algorithms for Automatic Benthic Cover Monitoring and Mapping Using Towed Underwater Video Camera and High-Resolution Satellite Images. *Remote Sens.* **2018**, *10*, 773. [CrossRef]
33. Dou, J.; Yunus, A.P.; Tien Bui, D.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.-W.; Khosravi, K.; Yang, Y.; Pham, B.T. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [CrossRef]
34. Eugenio, F.; Marcello, J.; Martin, J. High-Resolution Maps of Bathymetry and Benthic Habitats in Shallow-Water Environments Using Multispectral Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3539–3549. [CrossRef]
35. Evagorou, E.; Mettas, C.; Agapiou, A.; Themistocleous, K.; Hadjimitsis, D. Bathymetric maps from multi-temporal analysis of Sentinel-2 data: The case study of Limassol, Cyprus. *Adv. Geosci.* **2019**, *45*, 397–407. [CrossRef]
36. Mateo-Pérez, V.; Corral-Bobadilla, M.; Ortega-Fernández, F.; Vergara-González, E.P. Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery. *Remote Sens.* **2020**, *12*, 2069. [CrossRef]
37. European Space Agency. ESA Sentinel 2 Orbit Description. 2019. Available online: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/satellite-description/orbit (accessed on 29 September 2019).
38. SNAP. Available online: http://step.esa.int/main/toolboxes/snap (accessed on 12 July 2019).
39. Poursanidis, D.; Traganos, D.; Reinartz, P.; Chrysoulakis, N. On the use of Sentinel-2 for coastal habitat mapping and satellite-derived bathymetry estimation using downscaled coastal aerosol band. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *80*, 58–70. [CrossRef]
40. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [CrossRef]
41. Geyman, E.C.; Maloof, A.C. A Simple Method for Extracting Water Depth From Multispectral Satellite Imagery in Regions of Variable Bottom Type. *Earth Space Sci.* **2019**, *6*, 527–537. [CrossRef]
42. Vapnik, V.; Golowich, S.E.; Smola, A.J. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In *Advances in Neural Information Processing Systems 9*; Mozer, M.C., Jordan, M.I., Petsche, T., Eds.; MIT Press: Boston, MA, USA, 1997; pp. 281–287.
43. Clarke, S.M.; Griebsch, J.H.; Simpson, T.W. Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses. *J. Mech. Des.* **2005**, *127*, 1077–1087. [CrossRef]
44. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Cham, Switzerland, 2006; ISBN 1-4939-3843-6.
45. Kranjčić, N.; Medak, D.; Župan, R.; Rezo, M. Support Vector Machine Accuracy Assessment for Extracting Green Urban Areas in Towns. *Remote Sens.* **2019**, *11*, 655. [CrossRef]
46. Kuhn, M. Classification and Regression Training. R Package Version 2014. Available online: https://ui.adsabs.harvard.edu/abs/2015ascl.soft05003K/abstract (accessed on 27 April 2021).
47. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Reiss, H.; Cunze, S.; König, K.; Neumann, H.; Kröncke, I. Species distribution modelling of marine benthos: A North Sea case study. *Mar. Ecol. Prog. Ser.* **2011**, *442*, 71–86. [CrossRef]
49. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Cham, Switzerland, 2013; Volume 112.
50. Prasad, A.M.; Iverson, L.R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181–199. [CrossRef]
51. Peters, J.; Baets, B.D.; Verhoest, N.E.C.; Samson, R.; Degroeve, S.; Becker, P.D.; Huybrechts, W. Random forests as a tool for ecohydrological distribution modelling. *Ecol. Model.* **2007**, *207*, 304–318. [CrossRef]
52. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
53. Friedman, J.H. Multivariate Adaptive Regression Splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
54. Hansen, M.H.; Kooperberg, C. Spline Adaptation in Extended Linear Models with comments and a rejoinder by the authors. *Stat. Sci.* **2002**, *17*, 2–51. [CrossRef]
55. Milborrow, S.; Hastie, T.; Tibshirani, R. Earth: Multivariate Adaptive Regression Spline Models. R Package Version 2014. Available online: https://rdrr.io/cran/earth/ (accessed on 27 April 2021).
56. Emamgolizadeh, S.; Bateni, S.M.; Shahsavani, D.; Ashrafi, T.; Ghorbani, H. Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). *J. Hydrol.* **2015**, *529*, 1590–1600. [CrossRef]
57. Pohjankukka, J.; Pahikkala, T.; Nevalainen, P.; Heikkonen, J. Estimating the Prediction Performance of Spatial Models via Spatial K-Fold Cross Validation. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2001–2019. [CrossRef]
58. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models. *Nat. Commun.* **2020**, *11*, 4540. [CrossRef] [PubMed]

59. Deng, J.S.; Wang, K.; Deng, Y.H.; Qi, G.J. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. *Int. J. Remote Sens.* **2008**, *29*, 4823–4838. [CrossRef]
60. Cowan, G. *Statistical Data Analysis*; Oxford University Press: Oxford, UK, 1998; ISBN 0-19-850156-0.
61. Stumpf, R.P.; Holderied, K.; Sinclair, M. Determination of water depth with high-resolution satellite imagery over variable bottom types. *Limnol. Oceanogr.* **2003**, *48*, 547–556. [CrossRef]
62. Legleiter, C.J.; Roberts, D.A.; Lawrence, R.L. Spectrally based remote sensing of river bathymetry. *Earth Surf. Process. Landf.* **2009**, *34*, 1039–1059. [CrossRef]
63. Pan, Z.; Glennie, C.; Fernandez-Diaz, J.C.; Starek, M. Comparison of bathymetry and seagrass mapping with hyperspectral imagery and airborne bathymetric lidar in a shallow estuarine environment. *Int. J. Remote Sens.* **2016**, *37*, 516–536. [CrossRef]
64. Misra, A.; Vojinovic, Z.; Ramakrishnan, B.; Luijendijk, A.; Ranasinghe, R. Shallow water bathymetry mapping using Support Vector Machine (SVM) technique and multispectral imagery. *Int. J. Remote Sens.* **2018**, *39*, 4431–4450. [CrossRef]

Journal of
Marine Science and Engineering

MDPI

Article

Analysis of the Spatio-Temporal Evolution of Dredging from Satellite Images: A Case Study in the Principality of Asturias (Spain)

Vanesa Mateo-Pérez [1], Marina Corral-Bobadilla [2,*], Francisco Ortega-Fernández [1] and Vicente Rodríguez-Montequín [1]

[1] Project Engineering Department, University of Oviedo, 33004 Oviedo, Principality of Asturias, Spain; mateovanesa@uniovi.es (V.M.-P.); fdeasis@uniovi.es (F.O.-F.); montequi@uniovi.es (V.R.-M.)
[2] Department of Mechanical Engineering, University of La Rioja, 26004 Logroño, La Rioja, Spain
* Correspondence: marina.corral@unirioja.es; Tel.: +34-941-299-274

Abstract: One of the fundamental tasks in the maintenance of port operations is periodic dredging. These dredging operations facilitate the elimination of sediments that the coastal dynamics introduce. Dredging operations are increasingly restrictive and costly due to environmental requirements. Understanding the condition of the seabed before and after dredging is essential. In addition, determining how the seabed has behaved in recent years is important to consider when planning future dredging operations. In order to analyze the behavior of sediment transport and the changes to the seabed due to sedimentation, studies of littoral dynamics are conducted to model the deposition of sediments. Another methodology that could be used to analyze the real behavior of sediments would be to study and compare port bathymetries collected periodically. The problem with this methodology is that it requires numerous bathymetric surveys to produce a sufficiently significant analysis. This study provides an effective solution for obtaining a dense time series of bathymetry mapping using satellite data, and enables the past behavior of the seabed to be examined. The methodology proposed in this work uses Sentinel-2A (10 m resolution) satellite images to obtain historical bathymetric series by the development of a random forest algorithm. From these historical bathymetric series, it is possible to determine how the seabed has behaved and how the entry of sediments into the study area occurs. This methodology is applied in the Port of Luarca (Principality of Asturias), obtaining satellite images and extracting successive bathymetry mapping utilizing the random forest algorithm. This work reveals how once the dock was dredged, the sediments were redeposited and the seabed recovered its level prior to dredging in less than 2 months.

Keywords: Luarca port; random forest; dredging activities; sediment transport; satellite images

check for updates

Citation: Mateo-Pérez, V.; Corral-Bobadilla, M.; Ortega-Fernández, F.; Rodríguez-Montequín, V. Analysis of the Spatio-Temporal Evolution of Dredging from Satellite Images: A Case Study in the Principality of Asturias (Spain). *J. Mar. Sci. Eng.* **2021**, *9*, 267. https://doi.org/10.3390/jmse9030267

Academic Editor: Rodger Tomlinson

Received: 5 February 2021
Accepted: 25 February 2021
Published: 2 March 2021

1. Introduction

Dredging activities are commonly used in coastal areas to maintain the designed depth of navigation channels or basins and to remove deposited sediments. These activities include not only the processes of removing sediment from the bottom, but also their subsequent transport to another location. Dredging of ports improves access and exit conditions for cargo and passenger ships [1,2]. This is important for marine safety, removing sediment that has accumulated in channels in order to maintain the designed depths of the existing facilities [3,4]. However, despite its importance in maritime works and its effect on economic and social development, and the current advances in dredging techniques, the siltation of ports continues to be one of the least understood branches of coastal engineering. It is of great importance to ports, in maintaining and improving its depth and, also, in developing new facilities or creating new ports [5]. Thus, it is particularly important to study the sedimentation that takes place in the basins and entrance channels of ports. Excessive sedimentation can impair the functioning of the port and lead to a decrease in

economic activities [6–8]. The construction of dikes and other coastal protection works have a great influence on hydrodynamic forces and movement of sediment in the offshore area [9]. Hydrodynamic and sedimentary conditions can cause sedimentation of the port channel, which must be removed by maintenance dredging [10].

The Port of Luarca is located in the Principality of Asturias and is one of the most important ports in northern Spain. In recent years there have been significant changes in the marine dynamics of the port, as well as in the erosion–sedimentation processes that operate there. To maintain navigation, dredging is undertaken once a year. Technical specifications include the estimation of siltation volume, location, duration and phases of dredging works to be performed, the dredging method and the location of suitable sediment dumping locations that will not affect the environment adversely [11]. At Luarca, sandy bottoms prevail, although several areas are presently covered by mud [12]. The purpose of dredging the dock area is to guarantee the functionality of the port, which is conditioned by the depth of the dock. The location of the Port of Luarca at the mouth of the Rio Negro favors the accumulation of sedimentary materials of fluvial origin. The dock's location forestalls the tidal dynamics that are necessary to wash away the deposited sediments. For this reason, periodic dredging is necessary to maintain the dock. This enables optimal levels of use of the different port areas to be achieved. Throughout its operational life, the port has faced significant sedimentation problems and a consequent decrease in the depth of the draft in its entrance channel. The problem of sedimentation is increasing due to the environmental problems that are associated with dredging and the restrictive regulations imposed as a result. This sedimentation problem may be related to the positioning and layout of the port basin and its entry location [13,14]. Due to high sedimentation rates, the basin and entrance channel of the Port of Luarca must be dredged regularly for the port to continue its operation.

Many different methodologies have been used in an effort to understand the behavior of sediments. These have ranged from the use of probabilistic design methods [15–18] to direct numerical simulations [19–21]. Direct numerical simulations provide a rough understanding of the behavior of sediments. However, it is a costly methodology, because it requires the survey of the entire area in order to understand its 3D geometry. Another technique that could be used to study the behavior of sediments is to conduct an analysis of the bathymetries that were obtained in the study area. If the evolution of the seabed is known, it is possible to know where the sediment deposits are produced and their sedimentation rate. This technique would also help to determine the need to undertake dredging, as well as its periodicity. The main limitation when analyzing the behavior of the seabed is the availability of bathymetries. Bathymetries are normally conducted before and after dredging, but not frequently enough to understand the evolution of the seabed, due to their high cost and the difficulty of carrying them out. This clearly limits their use in analyzing the behavior of the seabed. Therefore, for this purpose, bathymetries are obtained from satellite images using the simplest techniques that offer a high level of precision. In addition, the techniques must be applicable in areas of turbid and shallow water and provide rapidity and flexibility of use.

Conventional bathymetric methods usually provide depth profiles or point measurements of acceptable accuracy. However, their use is costly and inefficient. Bathymetric estimation can be improved by combining echo sounding and satellite data. Remote sensing technology is used in many topographical studies. This is very useful for situations in which the depths of water are required on temporal and spatial scales, but are difficult to obtain [22–25]. Simple methods that use optical images in estimating depths are used by some investigator methods, as well as linear regression algorithms to estimate the depth of water [26]. In recent years, the use of machine learning techniques and optical sensors to estimate bathymetries has gained popularity. This can be attributed to advances in the development of algorithms, as well as greater availability of data and new satellites. Many authors have sought to determine the depths of water depths from optical sensors and the use of regression models that were derived from machine learning techniques [27–32]. The

random forest algorithm was used in this study to obtain the bathymetries. This algorithm has a number of advantages, including, low-cost, simplicity, time-effectiveness, and its wide-coverage for shallow water. Thus, the random forest algorithm has been found to be applicable to construct regression models that rely on bathymetry data from satellite images [33–41].

Many studies have concentrated on migration of sediments over the long term and how this affects water quality and the ecosystem condition. They model movement of material that is removed during the dredging operation [42]. However, optimizing the dredging operation has not been proposed. The main objective of this work is to determine if the behavior of sedimentation differs with the depth of dredging. In other words, is there a stable mean depth outside of which erosion and sedimentation occur more quickly? This is possible due to the capacity of the methodology, which enables data from the past with a frequency of 1–2 months to be analyzed. This provides great information of the seabed's behavior.

2. Study Site

The Port of Luarca is situated in northwestern Spain on the coast of the Cantabrian Sea at a longitude of 6°32′1″ W and a latitude of 43°32′45″ N (Figure 1a). Luarca has been associated with maritime activity since it began as a fishing enclave in the 10th century. Several changes have occurred throughout past years in Luarca Port. In 1910, the Canouco dike (Figure 1b), 40 m in length, was built to shelter the outer basin of the port. In 1940 a new dike, the La Encoronada dike, was built (Figure 1b), 124 m in length on the western side [43]. It is necessary to conduct dredging of the dock annually. The lack of draft means that the boats cannot enter the port, as many currents are generated in the entrance channel. Thus, navigation in this area is endangered by the risk that the boats will be stranded, especially during storms. The port is used generally by small boats. Thus, the minimum drafts in the navigation channel range from 2 to 3 m, with an average tidal run of 2 m. The draft in the docking area is somewhat less—2 m. Although dredging is conducted in the inlet channels, it also affects the inner basin. Of the two areas, it is the inner basin that receives the most interventions and is most limited due to draft. Nalona, a dredger boat (Suction Hopper Dredger, IMO 9047453) [44], has been used to dredge the bed surrounding the port. After the sand and mud have been collected from the bottom of the Port of Luarca, the sand is deposited in the area of Punta Muyeres, in front of the third beach of the village, in an effort to renew this area. However, it is not very clear whether this discharge location benefits the port, since a large amount of the sand that is dumped there will be moved by the currents and re-enter the dikes. The mud, on the other hand, is deposited at another location that is farther from the beaches.

The dates on which the dredging is carried out in the inner basin are provided by the Port Service of the Principality of Asturias. Dredging operations were carried out in 2017 and 2018. In 2017, the dredging began in October, but was stopped because of a problem with the dredge. It resumed in November and ended in December, 2017. In 2018, dredging began on November 15 and continued until the end of the year. In 2019, the inlet channel was dredged, but little work was done in the inner basin.

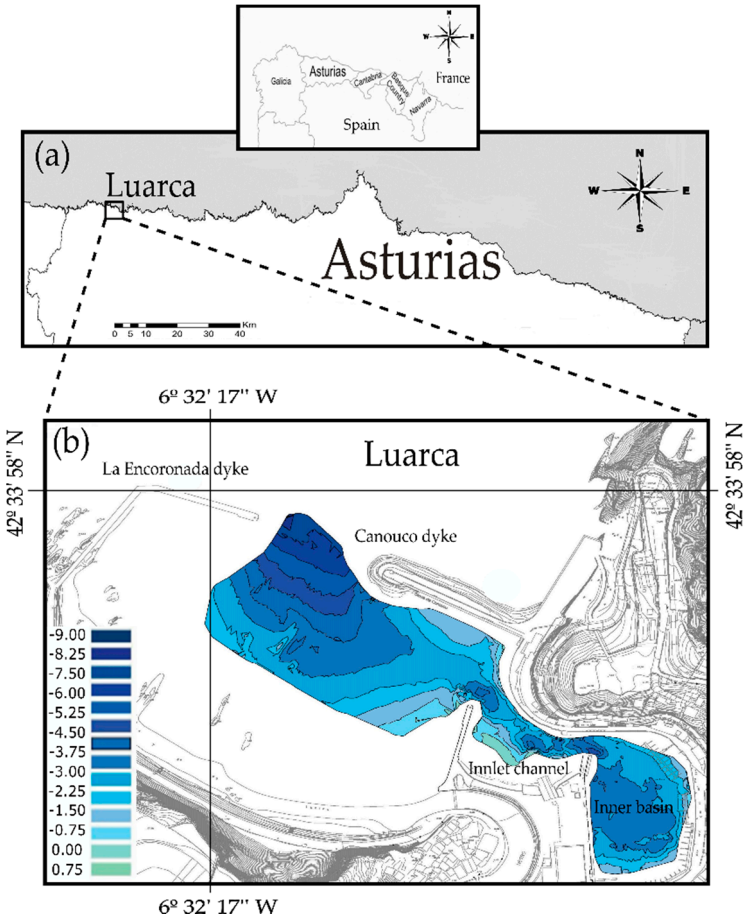


Figure 1. The Port of Luarca: (**a**) position of the port on the Cantabrian Sea coast; (**b**) the study site location.

3. Data and Methodology

3.1. Data In Situ Measurements and Satellite Data

Field measurements are generally necessary to study the sedimentation in the port, including the hydrodynamic parameters of the area, in this case only bathymetries will be used as field measurements, and these bathymetries are provided annually by the Port Service of the Principality Asturias. The bathymetries were conducted with the use of a Navisound 210 single beam echo sounder, with a 1 cm vertical resolution and dual frequency (Reson, Inc.: Slangerup, Denmark). The specifications of echo sounder Navisound 210 are shown in Table 1.

Table 1. Characteristics of echo sounder Navisound 210.

Navisound 210/400	Specifications
Frequency	190–235 kHz
Potency of transmission	300 W
Impedance	100 Ohm
Echo approval length	100 µs–210 kHz
Depth range	0.5–100/400/1200 m depending on the frequency
Resolution	1 cm
Accuracy	1 cm at 210 kHz (1 σ) assuming correct sound velocity, transducer depth etc.

The study's echo sounding data for Luarca were determined, on 28 June 2016, 10 May 2018 and 28 May 2019. For the present work, XY positions from the survey data were

projected using UTM Zone 30N. As they are usually costly, they are normally used only for specific dates. The satellite Sentinel-2 provided the data with which to predict the bathymetry of the water depth at the port (Table 2).

Table 2. Dates of acquisition of bathymetries.

Year			
2017	**2018**	**2019**	**2020**
		05/01/2019	25/01/2020
	24/02/2018	24/02/2019	19/02/2020
	16/03/2018		10/03/2020
	10/04/2018	20/04/2019	
		30/05/2019	
	24/06/2018	14/06/2019	
04/07/2017	29/07/2018	24/07/2019	
13/08/2017	18/08/2018	23/08/2019	
02/09/2017	02/09/2018	12/09/2019	
02/10/2017	02/10/2018	22/10/2019	
21/11/2017	16/11/2018	21/11/2019	
21/12/2017	31/12/2018		

3.2. Methodology

The process to obtain the bathymetries is shown schematically in Figure 2.

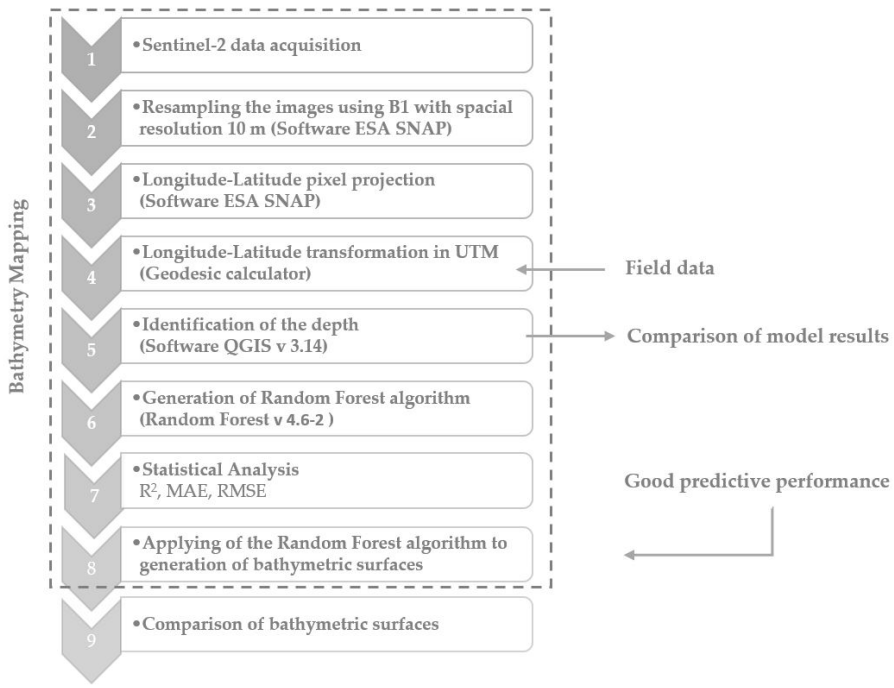


Figure 2. Methodological workflow for development of bathymetric maps.

3.2.1. Processing Satellite Images

The processing step concerns using Sentinel-2A data as vector of variables for training and fitting of the random forest model (Figure 2). Sentinel-2A satellite data was retrieved from ESA Copernicus Open Access Hub. SNAP (Sentinel Application Platform) software was used to visualize and preprocess Sentinel-2A data (10 m resolution) [45]. The data from the Sentinel-2A satellite reflectance bands (B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, and B12) were used to predict the depth of the water at the study port. All spectral bands in the Sentinel-2A images were resampled to achieve a resolution of 10 m using the SNAP S2 Resampling Processor [15,46,47]. As a result, a dataset without georeferencing

was obtained, and to determine the positioning of the reference points, the geographical location of each point was defined by its longitude and latitude using the SNAP program. Then, using the WGS84 ellipsoid, a coordinate projection was created in order to obtain the coordinates in ETRS89 [48]. This system was also used to project the positions that the echo sounder provides. The ellipsoid projections had an average position error of 1 cm. The data that was obtained was compared to the bathymetry that had been projected. To accomplish this, a geodesic calculator was used to project the coordinates. This enabled the authors to obtain the data for bands that are associated with UTM x-y coordinates. The random forest algorithm was used to assign the z coordinate. From these points and using the QGIS software (version 3.14), the surface was obtained by digital models of the TIN (Triangulated Irregular Network) terrain type. Triangulation was undertaken using linear interpolation [49]. The error that this process involved was small due to the absence of any great irregularities in the smooth surfaces of the seabed. The port's zero serves as the reference point for the z coordinates. The former is the minimum measured level of the surface of the water during the highest low tide in the last 15 years. Pixels were assigned dimensions on the basis of their x-y locations.

3.2.2. The Random Forest Algorithm

The random forest algorithm [50] facilitates classification and regression by use of a randomized subset of predictors that enable it to create a variety of classification trees [51]. Many of the trees start as bootstrapped training data samples. A random subset of predictor variables (z coordinate) is used at each fork in the process. This causes each tree to be different. Although each tree is a poor predictor, each pair of them provides a different response. This aggregates the predictions of uncorrelated trees, reducing prediction variance and increasing accuracy [52–54]. This work involved 100 trees and 13 variables (Sentinel-2A satellite bands B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, and B12, and tidal). The default value was adopted as the minimum size of nodes. The random forest algorithm for prediction of bathymetry was provided by the Random Forest (v 4.6-2) R package [55].

3.2.3. Training and Testing Dataset

In order to evaluate the accuracy of the model, the dataset that was obtained from Sentinel-2 was divided into two groups. Training the random forest algorithm required 80% of the dataset (1593 data points), and testing the model used the remaining 20% (388 data points). Data from echo sounding measurements were used to calibrate the model without any kind of elaboration [56]. To test the model, the data obtained with the model were compared with field measurements for 20% of the points. As the tide affected the depth measurement results, the measured depths were the mean sea levels (MSL) that served as references. This was accomplished by deducting the measured depth during high tide from the MSL. The tidal data were provided by the nearest tidal station [34].

Finally, in order to determine the accuracy of the model's estimates of depth, satellite derived bathymetry maps were compared with that of the field measurements obtained by using the echo sounder. The residual statistic between the satellite derived depth (*Zsatellite*) and the echo sounding measurements (*Zecho-sounder*) were reported along three metrics. They were the mean absolute error (MAE), the root mean squared error (RMSE) and the correlation coefficient (adjusted R^2). They are calculated by the equations below.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Z_{Satellite} - Z_{echo-sounder}\right|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_{Satellite} - Z_{echo-sounder}\right)^2}$$

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}\left(Z_{echo-sounder} - \overline{Z_{echo-sounder}}\right)\left(Z_{Satellite} - \overline{Z_{Satellite}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Z_{echo-sounder} - \overline{Z_{echo-sounder}}\right)^2\sum_{i=1}^{n}\left(Z_{echo-sounder} - \overline{Z_{echo-sounder}}\right)^2}}$$

where $Z_{Satellite}$ are the depths that the random forest methodology predicted from Sentinel-2 images. $Z_{echo\text{-}sounder}$ denote the in situ echo sounding depths and n is the number of data.

4. Results

The random forest algorithm achieved a good predictive performance, with an MAE of 0.37, an RMSE of 0.47 and an R^2 of 0.974. The testing data set results appear in Table 3.

Table 3. Error statistics reported in meters produced by the random forest algorithm.

Algorithm	MAE (m)	RMSE (m)	R^2
Random forest	0.37	0.47	0.974

Figure 3 provides a histogram of the relationship between mean absolute error (MAE) and the depth of the port of Luarca that was obtained by the random forest algorithm. As can be seen in Figure 3, the maximum error (60 cm) occurred between −6 and −8 m and the minimum (24 cm) occurred at −10 m. This error is acceptable for the purpose of this study, which is to analyze the behavior of the seabed with respect to time.

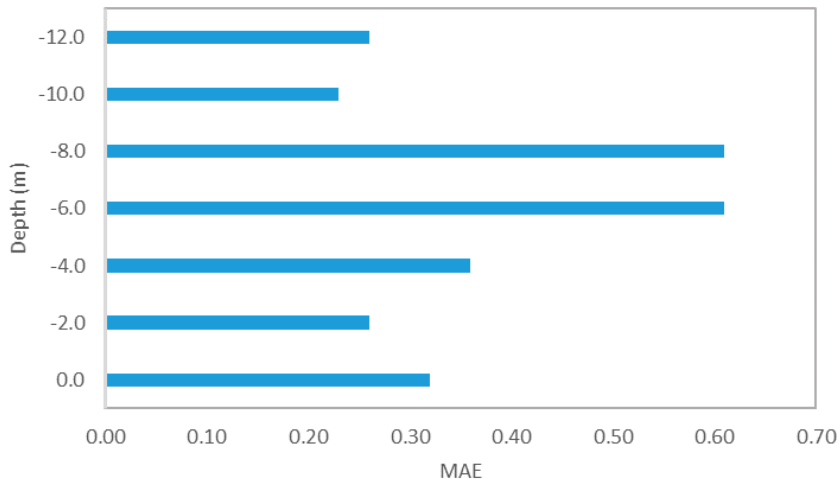


Figure 3. Variation of MAE errors versus depth (m).

As an example of the results that were obtained, the bathymetries before and after the 2017 dredging are represented in 3D and 2D (Figures 4 and 5). Figure 4 represents the 3D evolution of the seabed for 21/11/17 (Figure 4a), 21/12/17 (Figure 4b) and 24/02/18 (Figure 4c). These show that the elevation of the seabed was maintained in Figure 4a,b, but had declined in Figure 4c due to dredging.

Figure 5 represents one of the dredging episodes by a series of bathymetries that were analyzed in the plane and profile view. These indicate the effect of the dredging that was undertaken. First, when dredging is carried out, there is a decrease in the elevation of the basin and the inlet channel (Figure 5a,c). The elevation recovers when the dredging stops (Figure 5b,d). This effect is seen in both the inner dock (section AA′) and in the inlet channel (section BB′). In both representations (Figures 4 and 5), the areas of greatest variation are the inner basin and the inlet channel. These are the areas where dredging takes place. Additionally, in Figures 4 and 5, the areas in which the greatest changes occur are in the lower right corner (inner dock area and inlet channel) and in the upper left corner that coincides with the beach area. Both are areas that undergo many changes in topography due to coastal dynamics.

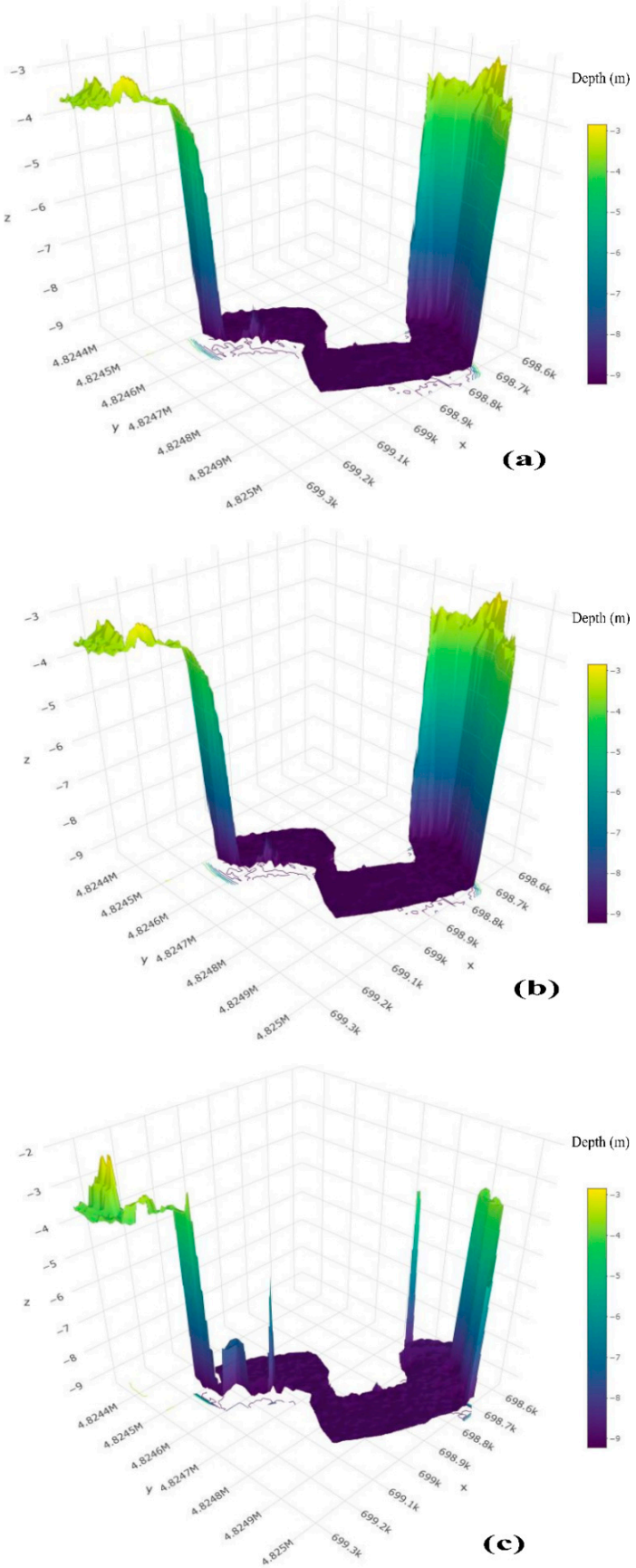


Figure 4. 3D bathymetries. Dates: (**a**) 21/11/17, (**b**) 21/12/17 and (**c**) 24/02/18.

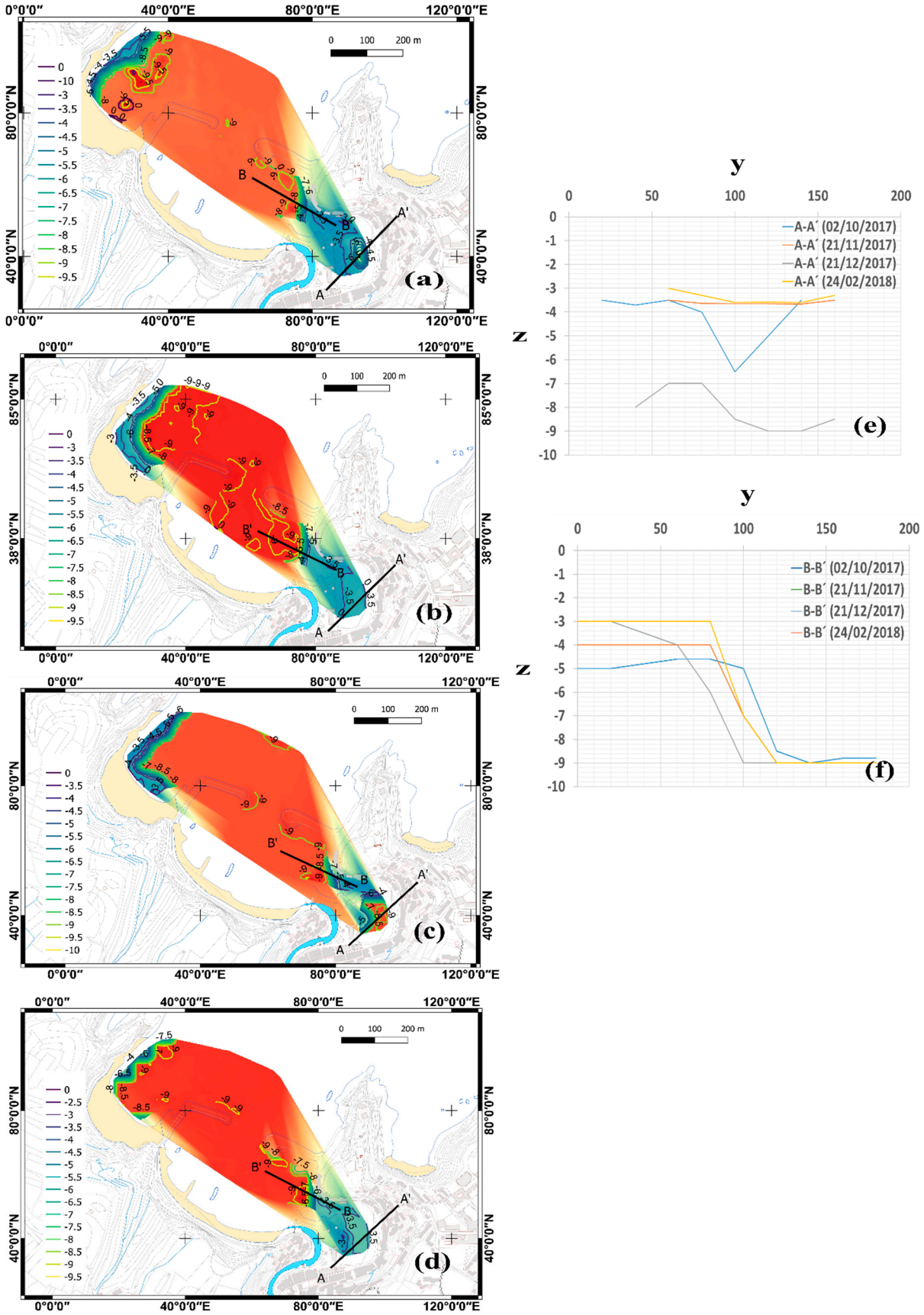


Figure 5. A representation of the behavior of the interior basin the dredging on: (**a**) 02/10/17, (**b**) 21/11/17, (**c**) 21/12/17 and (**d**) 24/02/18; and cross-sectional profile (**e**) AA′ of the inner dock, and (**f**) BB′ of the inlet channel.

After the surfaces were generated, they were compared—bathymetry to bathymetry—sequentially in time in order to examine the erosion and the sedimentation that occurred. This is done by comparing the difference between the surfaces that are generated for each bathymetry that is analyzed. To study the evolution of sediment in the basin of the Port of Luarca, the surfaces for each of the bathymetries obtained from the use of satellite images and the random forest algorithm were generated. Once the surfaces were obtained, the differences between them were analyzed. Comparing these surfaces, the erosion and sedimentation volumes were obtained.

In the results that appear in Table 4, the first and second columns provide the dates on which the bathymetry comparison was made. The third column provides the duration or time interval between bathymetries. The fourth column shows the volume of the main activity (erosion or sedimentation) by day that was generated. The fifth column shows the erosion or decrease in volume between the first and second bathymetry (sediment that exited the basin area), whereas the sixth column provides the sedimentation or volume increase between the first and second bathymetry (sediment that entered the basin area). The seventh column provides the total value of the difference between sedimentation and erosion. The eighth column shows an error assessment analysis in terms of volume change. The last column of the table shows when the dredging was undertaken.

Table 4. Evolution of sedimentation in the Port of Luarca.

Starting Date	Finishing Date	Time (Day)	Rate (m^3/Day)	Erosion Volume (m^3)	Sedimentation Volume (m^3)	Total (m^3)	Total Error (m^3)	
04/07/2017	13/08/2017	40	29.26	74.88	1245.11	1170.23	±110.31	
13/08/2017	02/09/2017	20	402.46	0.00	8049.16	8049.16	±878.69	
02/09/2017	02/10/2017	30	−548.80	16,463.97	0.00	−16,463.97	±1571.27	Start Dredge
02/10/2017	21/11/2017	50	99.10	263.30	5218.13	4954.82	±337.91	Stop Dredge
21/11/2017	21/12/2017	30	−1231.98	37,030.08	70.80	−36,959.29	±3343.98	Dredge
21/12/2017	24/02/2018	65	589.90	80.24	38,423.49	38,343.25	±3502.55	
24/02/2018	16/03/2018	20	−310.62	6215.43	2.93	−6212.50	±432.64	
16/03/2018	10/04/2018	25	169.35	24.94	4258.79	4233.86	±287.25	
10/04/2018	20/05/2018	40	120.86	0.00	4834.25	4834.25	±370.84	
20/05/2018	24/06/2018	35	19.80	518.85	1211.81	692.96	±68.85	
24/06/2018	29/07/2018	35	−48.44	2084.56	389.03	−1695.53	±165.76	
29/07/2018	18/08/2018	20	46.14	159.70	1082.48	922.78	±89.14	
18/08/2018	02/09/2018	15	−104.60	1668.90	99.85	−1569.05	±150.34	
02/09/2018	02/10/2018	30	−57.35	1792.12	71.55	−1720.57	±157.61	
02/10/2018	16/11/2018	45	14.65	532.99	1192.31	659.32	±59.24	
16/11/2018	31/12/2018	45	−597.85	26,910.49	7.34	−26,903.15	±2752.28	Dredge
31/12/2018	05/01/2019	5	3556.35	12.93	17,794.68	17,781.76	±1655.70	
05/01/2019	24/02/2019	50	239.28	33.00	11,997.16	11,964.16	±832.46	
24/02/2019	26/03/2019	30	−1.03	438.94	408.12	−30.82	±3.04	
26/03/2019	20/04/2019	25	697.02	0.00	17,425.54	17,425.54	±1971.12	
20/04/2019	05/05/2019	15	−592.93	8893.90	0.00	−8893.90	±1227.16	
05/05/2019	30/05/2019	25	−311.02	7775.56	0.00	−7775.56	±894.89	
30/05/2019	14/06/2019	15	594.02	0.00	8910.36	8910.36	±1046.90	
14/06/2019	24/07/2019	40	−236.76	9470.57	0.00	−9470.57	±1101.37	
24/07/2019	23/08/2019	30	−20.47	697.96	83.96	−614.01	±60.14	
23/08/2019	12/09/2019	20	2.99	267.98	327.82	59.84	±5.82	
12/09/2019	22/10/2019	40	−39.67	1632.29	45.49	−1586.80	±150.54	
22/10/2019	21/11/2019	30	−66.90	2007.10	0.00	−2007.10	±181.28	
21/11/2019	26/12/2019	35	−132.28	4742.30	112.66	−4629.64	±319.05	
26/12/2019	25/01/2020	30	287.98	0.00	8639.33	8639.33	±630.13	
25/01/2020	19/02/2020	25	46.14	120.48	1273.91	1153.43	±115.83	
19/02/2020	10/03/2020	20	−118.73	2374.51	0.00	−2374.51	±234.22	
10/03/2020	04/04/2020	25	52.91	137.30	1459.93	1322.63	±128.26	

Table 4 shows the increase in the rate of sedimentation, notably after the dredging work of 589.90 m^3/day after the 2017 dredging and 3556.50 m^3/day after the 2018 dredging. In the 2017 dredging, 16,463.97 m^3 were extracted and, after stopping the work, the amount of sediment that was released was 37,030.08 m^3. In the following period, 38,423.49 m^3 sediment were extracted, which is more than was dredged in the previous month. In fact, 5218.13 m^3 of sedimentation occurred in the breakdown period.

In the 2018 dredging, 26,910.49 m^3 were extracted, 17,794.68 m^3 of which were extracted in just 5 days. In order to analyze more precisely the operation of the inner basin of the Port of Luarca, the variations in elevation during the dredging work were represented, as well as after the completion of the dredging in 2017 and 2018 (Figures 6 and 7).

In Figure 6a it can be seen that, after stopping at the end of October 2017, the areas where the dredging was conducted during the September–October period were immediately filled with sand again. The same is seen in Figure 6b,c. It can be seen that the elevation dropped substantially while the dredging was being conducted (Figure 6b), but had recovered in the next bathymetry (Figure 6c) to the original elevation.

The bathymetries in Figure 6 also show how dredging was carried out. In the specific way that the suction dredger is used, it acts on an area of the surface, substantially lowering the elevation. Later, the coastal dynamics fill these areas, thereby lowering the level of the entire dock. In reality, this low point is not filled with the adjacent material in the basin itself. Instead, it is filled by external sediment. Thus, the dredging effect does not last very long. This same analysis was conducted for the 2018 dredging (Figure 7).

Figure 7 represents the variation of the bathymetries of the bottom in the Port of Luarca. Figure 7a shows that the bottom falls due to the dredging carried out, achieving a total decrease in height of 5.5 m at the lowest point. In the bathymetry of Figure 7b, practically everything that has been extracted by dredging has been replaced, increasing the elevation by 5 m where the increase had been greatest. In this case, it is important to note that the variation that is reflected in the bathymetry occurs only on 5 days. In Figure 7c it can be seen that the bottom continued to recover the sediments throughout the following months until it had returned to its original level.

To analyze the behavior of the basin bottom elevation for periods not analyzed above, the average depth of the basin bottom of the port was analyzed (Figure 8).

Figure 8 shows the average depth of the bottom at the dock in the Port of Luarca. It can be seen that the dredging was done at the end of 2017. It began in October 2017, but became impossible to continue during November due to problems with the dredge. Thus, this dredging ended in December 2017. Figure 8 shows that the recovery of elevation was almost immediate. Additionally, as can be seen in Figure 8, the same phenomenon occurred in 2018, with dredging concluding at the end of the year, but the port's equilibrium level recovering in the beginning of 2019.

Aside from the phenomena of dredging and rapid recovery of the average level, the behavior of the bottom was quite stable, except for some sediment entry points. It is important to note that even these sediment entry peaks are regulated and without external intervention, with the bottom returning to its average level. In this case, a mean elevation or equilibrium profile of between −3 and −4 m was verified.

To determine why there are points that experienced an entry of sand in the docking area that exceeded the balance level, the wave height was analyzed (Figure 9). The data was supplied by a multiparametric buoy that was deployed in the area surrounding Gijón Harbor [42].

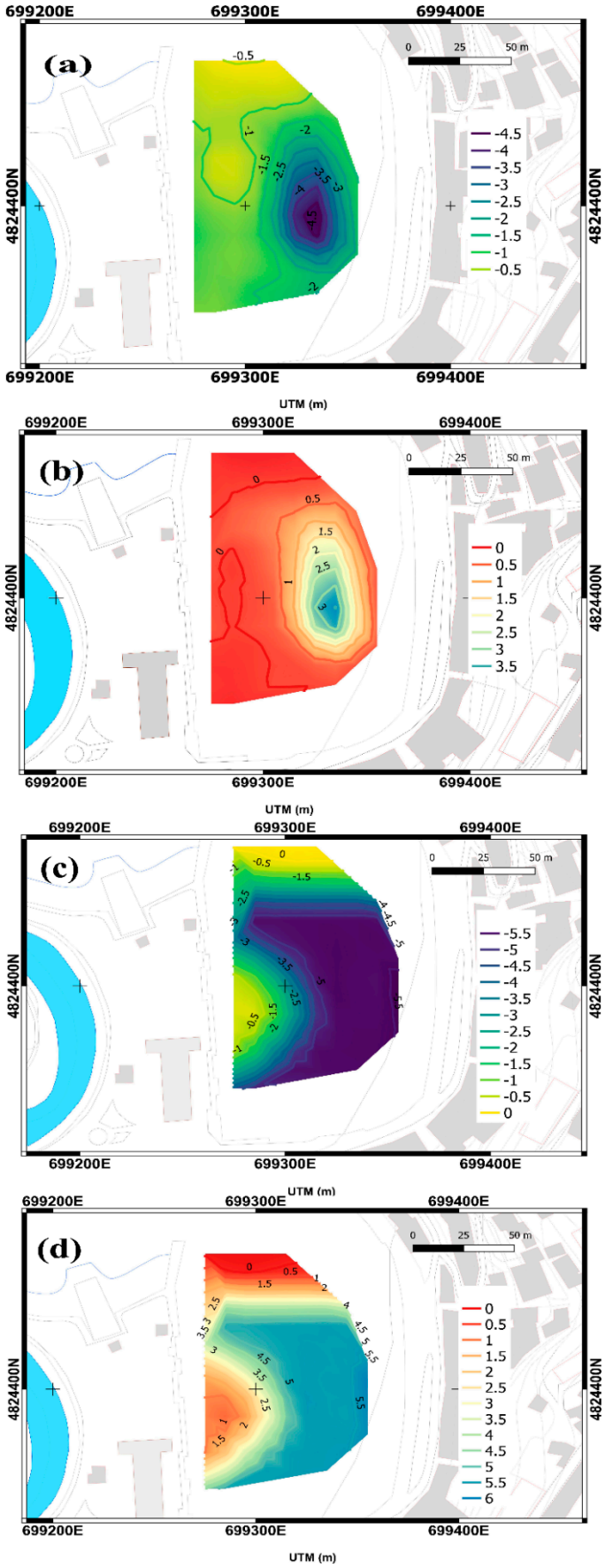


Figure 6. Difference in elevation of the bathymetries in the Port of Luarca between (**a**) 02/09/17 and 20/10/17, (**b**) 02/10/17 and 02/11/17, (**c**) 21/11/17 and 21/12/17, and (**d**) 21/12/17 and 24/02/18.

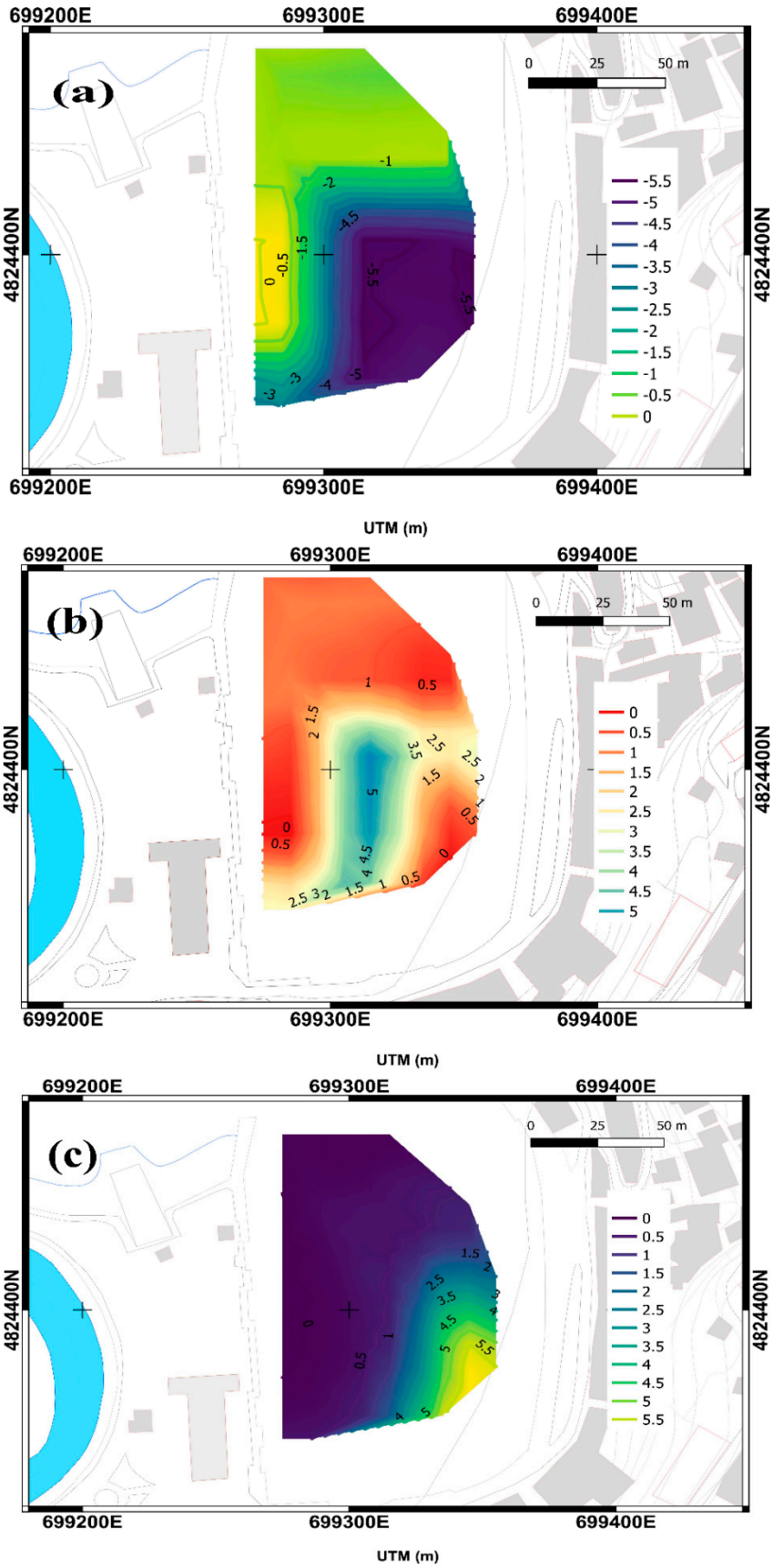


Figure 7. Difference in elevation of the bathymetries in the Port of Luarca between (**a**) 16/11/18 and 31/12/18, (**b**) 31/12/18 and 05/01/19, and (**c**) 05/01/19 and 24/02/19.

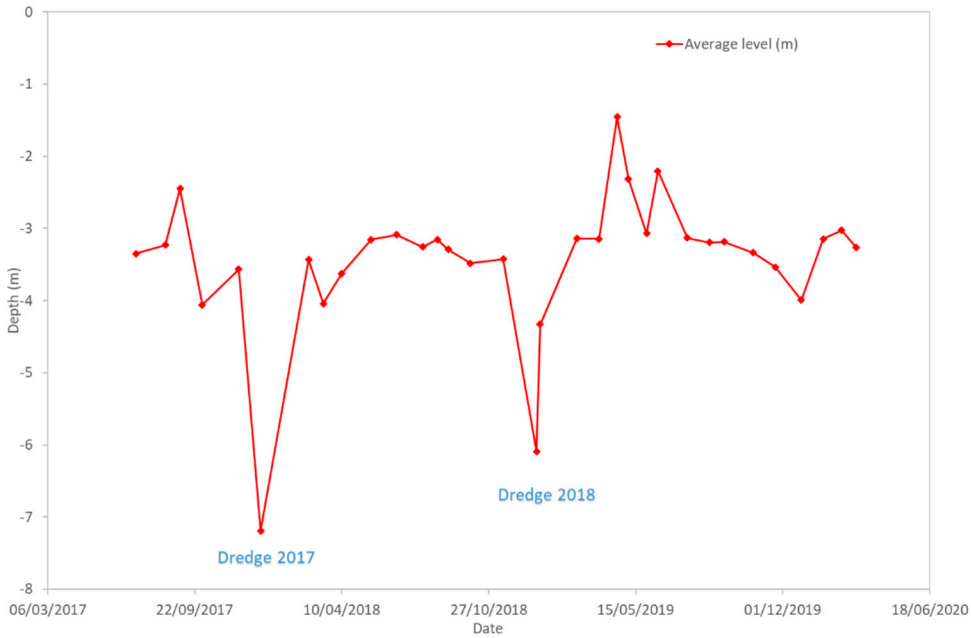


Figure 8. Average depth of the bottom of the dock in the Port of Luarca.

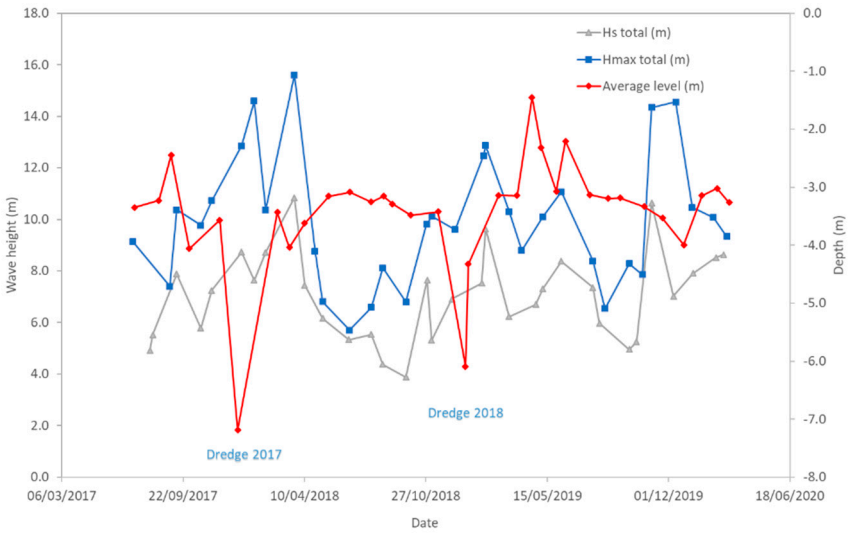


Figure 9. The effect of the wave height (H) on the average level of the dock in the Port of Luarca.

Figure 9 shows the effect of waves on the average level of the dock in the Port of Luarca. It shows that there were three important wave points. The one that followed the 2017 dredging produced a slight rebound of the mean elevation of the basin. The second important wave point occurred after the second dredging, which generated a more significant increase in the mean elevation. Finally, the third wave point of importance marked the end of the period that was being studied. However, as at the end of 2019, no dredging was conducted, and the elevation was stable. Thus, the impact of the waves was barely noticed. There appeared to be a relationship between the oscillation of the seabed levels and the waves. However, in order to confirm this, it would be necessary to carry out a more detailed study to measure the waves in the port itself.

The sand entry peaks, but much less pronounced, correspond to storms. As can be seen in Figure 9, these sand entries also disappeared, recovering the equilibrium point. The Negro River is another entry point for sedimentation. This river is a short coastal river in the north of Spain. It runs through the western part of the Asturias and flows into the

Cantabrian Sea, on the beach of Luarca. There is an Automatic Hydrological Information System (SAIH) in this river. It was from this control point that the data were obtained [57].

To study the possible influence of the river on the contribution of sediments in the basin, the mean level of the basin was compared to the level of the river (Figure 10).

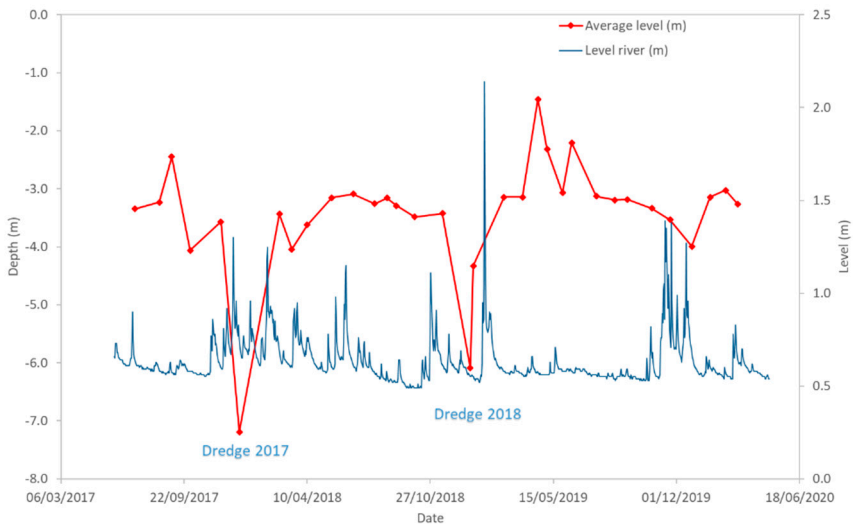


Figure 10. Effect of the Negro River on the average level of the dock in the Port of Luarca.

Figure 10 shows that a significant elevation of the river level does not imply an entry of sediments. The highest elevation of the Negro river level occurred on 01/23/2019 and the lowest level of the dock was reached at the end of March 2019 and ended 1 month later.

5. Conclusions and Discussion

One of the most important tasks for port maintenance is dredging. The objective of this is to maintain the draft within values that permit vessel to enter and exit from the port. This work examined how the bottom of the interior basin of the Port of Luarca behaves during dredging and subsequent periods by analyzing the data of 4 years. In order to study the variations in the surface of the seafloor, it was necessary to obtain bathymetries of it. Bathymetries are conducted annually in this port, but there is insufficient data for an analysis of these variations. These bathymetries must include the depth of the seabed at several points that are spaced appropriately, in addition to the planimetric position of these points and measurements of the variations in the mean level of the surface of the sea. Due to the ongoing modeling of seafloor level and the numerous bathymetric survey data sets and use of remote sensing technology, it is possible to estimate the volume of sedimentation and, therefore, how much sediment has accumulated in a navigation channel. The layers of sediment to remove in maintenance dredging are generally not very thick. Thus, in order to avoid unnecessary dredging-related expense, seafloor levels should be determined and modeled accurately.

The purpose of his study was to find a practicable method to obtain a dense time series of bathymetries using satellite data. This would enable the past behavior of the seabed to be analyzed. Bathymetries could also be obtained by continuous conventional measurements using echo sounders. However, there are issues of accuracy and precision at present. Additionally, shallow coastal waters are difficult to access. Hence, it is not possible at present to analyze past behavior. It is also necessary to wait until a sufficiently large data series can be obtained before the behavior of the seabed can be described with certainty.

To facilitate this work, several bathymetries were obtained on different dates from Sentinel-2 satellite images and the use of a random forest-type algorithm with an MAE of 0.37, an RMSE of 0.47, and an R^2 of 0.974. From these bathymetric data, the daily dredging excavation rates were determined. These were 1231.98 m^3/day in 2017 and 597.85 m^3/day in 2018. The analysis of the bathymetries indicates that, after this dredging, the bottom

recovered almost immediately (in the dredging of 2018 in 5 days) with filling rates of 589.90 and 3556.35 m^3/day, for 2017 and 2018, respectively. It also can be stated that the sporadic arrival of sediment above the mean depth also disappeared without external intervention, keeping the mean depth between −3 and −4 m. As a result, it can be concluded that lowering the elevation by dredging does not seem adequate, unless the dredging depth is close to the average elevation that is maintained naturally.

For future work, it might be useful to extend this investigation to other ports. Before determining the need for dredging, it would be useful to know if there is a depth beyond which it is not appropriate to conduct further dredging due to the rate at which sediment returns. The analysis of the behavior of the bottom of the ports in recent years provides valuable information for the planning of maintenance tasks and the knowledge of their behavior in the face of littoral dynamics.

Author Contributions: Conceptualization, V.M.-P. and F.O.-F.; software and validation, V.M.-P.; methodology, M.C.-B.; data curation, V.M.-P.; writing—original draft preparation, M.C.-B.; writing—review and editing, F.O.-F., and V.R.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science, Technology and Innovation Plan of the Principality of Asturias (Spain) Ref: FC-GRUPIN-IDI/2018/000225, which is partly funded by the European Regional Development Fund (ERDF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the Port Service and Transport Infrastructures of the Principality of Asturias for their collaboration in this work.

Conflicts of Interest: The authors declare there are no conflict of interest.

References

1. Quang Tri, D.; Kandasamy, J.; Cao Don, N. Quantitative Assessment of the Environmental Impacts of Dredging and Dumping Activities at Sea. *Appl. Sci.* **2019**, *9*, 1703. [CrossRef]
2. Bolam, S.G.; Rees, H.L. Minimizing Impacts of Maintenance Dredged Material Disposal in the Coastal Environment: A Habitat Approach. *Environ. Manag.* **2003**, *32*, 171–188. [CrossRef]
3. Norén, A.; Fedje, K.K.; Strömvall, A.-M.; Rauch, S.; Andersson-Sköld, Y. Integrated Assessment of Management Strategies for Metal-Contaminated Dredged Sediments—What Are the Best Approaches for Ports, Marinas and Waterways? *Sci. Total Environ.* **2020**, *716*, 135510. [CrossRef]
4. Wang, W.; Men, C.; Lu, W. Online Prediction Model Based on Support Vector Machine. *Neurocomputing* **2008**, *71*, 550–558. [CrossRef]
5. Khorram, S.; Khalegh, M.A. A Novel Hybrid MCDM Approach to Evaluate Ports' Dredging Project Criteria Based on Intuitionistic Fuzzy DEMATEL and GOWPA. *WMU J. Marit. Aff.* **2020**, *19*, 95–124. [CrossRef]
6. Cáceres, R.A.; Zyserman, J.A.; Perillo, G.M.E. Analysis of Sedimentation Problems at the Entrance to Mar Del Plata Harbor. *J. Coast. Res.* **2016**, *32*, 301–314. [CrossRef]
7. Feola, A.; Lisi, I.; Salmeri, A.; Venti, F.; Pedroncini, A.; Gabellini, M.; Romano, E. Platform of Integrated Tools to Support Environmental Studies and Management of Dredging Activities. *J. Environ. Manag.* **2016**, *166*, 357–373. [CrossRef]
8. Mahmoodi, A.; Lashteh Neshaei, M.A.; Mansouri, A.; Shafai Bejestan, M. Study of Current- and Wave-Induced Sediment Transport in the Nowshahr Port Entrance Channel by Using Numerical Modeling and Field Measurements. *J. Mar. Sci. Eng.* **2020**, *8*, 284. [CrossRef]
9. Chen, B.; Wang, K. Suspended Sediment Transport in the Offshore near Yangtze Estuary* *Project Supported by the National Natural Science Foundation of China (Grant No.40576017), the National Basic Research Program of China (973, Program, Grant No. 2007CB411804). *J. Hydrodyn. Ser. B* **2008**, *20*, 373–381. [CrossRef]
10. Zuo, S.; Xie, H.; Ying, X.; Cui, C.; Huang, Y.; Li, H.; Xie, M. Seabed Deposition and Erosion Change and Influence Factors in the Yangshan Deepwater Port over the Years. *Acta Oceanol. Sin.* **2019**, *38*, 96–106. [CrossRef]
11. Erftemeijer, P.L.A.; Robin Lewis, R.R. Environmental Impacts of Dredging on Seagrasses: A Review. *Mar. Pollut. Bull.* **2006**, *52*, 1553–1572. [CrossRef]
12. Flor, G.; del Busto, J.A.; Blanco, G.F. Morphological and Sedimentary Patterns of Ports of the Asturian Region (NW Spain). *J. Coast. Res.* **2006**, *48*, 35–40.

13. Stock, F.; Knipping, M.; Pint, A.; Ladstätter, S.; Delile, H.; Heiss, A.G.; Laermanns, H.; Mitchell, P.D.; Ployer, R.; Steskal, M.; et al. Human Impact on Holocene Sediment Dynamics in the Eastern Mediterranean—the Example of the Roman Harbour of Ephesus. *Earth Surf. Process. Landf.* **2016**, *41*, 980–996. [CrossRef]
14. Sharaan, M.; Ibrahim, M.G.; Iskander, M.; Masria, A.; Nadaoka, K. Analysis of Sedimentation at the Fishing Harbor Entrance: Case Study of El-Burullus, Egypt. *J. Coast. Conserv.* **2018**, *22*, 1143–1156. [CrossRef]
15. Poursanidis, D.; Traganos, D.; Reinartz, P.; Chrysoulakis, N. On the Use of Sentinel-2 for Coastal Habitat Mapping and Satellite-Derived Bathymetry Estimation Using Downscaled Coastal Aerosol Band. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *80*, 58–70. [CrossRef]
16. Vittori, G.; Blondeaux, P.; Mazzuoli, M.; Simeonov, J.; Calantoni, J. Sediment Transport under Oscillatory Flows. *Int. J. Multiph. Flow* **2020**, *133*, 103454. [CrossRef]
17. Finn, J.R.; Li, M.; Apte, S.V. Particle Based Modelling and Simulation of Natural Sand Dynamics in the Wave Bottom Boundary Layer. *J. Fluid Mech.* **2016**, *796*, 340–385. [CrossRef]
18. Finn, J.R.; Li, M. Regimes of Sediment-Turbulence Interaction and Guidelines for Simulating the Multiphase Bottom Boundary Layer. *Int. J. Multiph. Flow* **2016**, *85*, 278–283. [CrossRef]
19. Kidanemariam, A.G.; Uhlmann, M. Direct Numerical Simulation of Pattern Formation in Subaqueous Sediment. *J. Fluid Mech.* **2014**, *750*, 1–13. [CrossRef]
20. Leont'yev, I.O.; Akivis, T.M. Modeling of Coastal Dynamics of the Anapa Bay-Bar. *Oceanology* **2020**, *60*, 279–285. [CrossRef]
21. Armanini, A.; Cavedon, V.; Righetti, M. A Probabilistic/Deterministic Approach for the Prediction of the Sediment Transport Rate. *Adv. Water Resour.* **2015**, *81*, 10–18. [CrossRef]
22. Flener, C.; Lotsari, E.; Alho, P.; Käyhkö, J. Comparison of Empirical and Theoretical Remote Sensing Based Bathymetry Models in River Environments. *River Res. Appl.* **2012**, *28*, 118–133. [CrossRef]
23. Giardino, C.; Bresciani, M.; Matta, E.; Brando, V.E. Imaging Spectrometry of Inland Water Quality in Italy Using MIVIS: An Overview. In *Advances in Watershed Science and Assessment*; Younos, T., Parece, T.E., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 33, pp. 61–83. ISBN 978-3-319-14211-1.
24. Jawak, S.D.; Vadlamani, S.S.; Luis, A.J. A Synoptic Review on Deriving Bathymetry Information Using Remote Sensing Technologies: Models, Methods and Comparisons. *Adv. Remote Sens.* **2015**, *4*, 147–162. [CrossRef]
25. Hedley, J.; Roelfsema, C.; Chollett, I.; Harborne, A.; Heron, S.; Weeks, S.; Skirving, W.; Strong, A.; Eakin, C.; Christensen, T.; et al. Remote Sensing of Coral Reefs for Monitoring and Management: A Review. *Remote Sens.* **2016**, *8*, 118. [CrossRef]
26. Lyzenga, D.R. Passive Remote Sensing Techniques for Mapping Water Depth and Bottom Features. *Appl. Opt.* **1978**, *17*, 379. [CrossRef]
27. Cheng, N.-S.; Chiew, Y.-M. Pickup Probability for Sediment Entrainment. *J. Hydraul. Eng.* **1998**, *124*, 232–235. [CrossRef]
28. Liu, S.; Gao, Y.; Zheng, W.; Li, X. Performance of Two Neural Network Models in Bathymetry. *Remote Sens. Lett.* **2015**, *6*, 321–330. [CrossRef]
29. El-Mewafi, M.; Salah, M.; Fawzi, B. Assessment of Optical Satellite Images for Bathymetry Estimation in Shallow Areas Using Artificial Neural Network Model. *Am. J. Geogr. Inf. Syst.* **2018**, *7*, 99–106.
30. Obelcz, J.; Wood, W.T.; Phrampus, B.J.; Lee, T.R. Machine Learning Augmented Time-Lapse Bathymetric Surveys: A Case Study from the Mississippi River Delta Front. *Geophys. Res. Lett.* **2020**, *47*, e2020GL087857. [CrossRef]
31. Tonion, F.; Pirotti, F.; Faina, G.; Paltrinieri, D. A Machine Learning Approach to Multispectral Satellite Derived Bathymetry. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *3*, 565–570. [CrossRef]
32. Mavraeidopoulos, A.K.; Oikonomou, E.; Palikaris, A.; Poulos, S. A Hybrid Bio-Optical Transformation for Satellite Bathymetry Modeling Using Sentinel-2 Imagery. *Remote Sens.* **2019**, *11*, 2746. [CrossRef]
33. Sagawa, T.; Yamashita, Y.; Okumura, T.; Yamanokuchi, T. Satellite Derived Bathymetry Using Machine Learning and Multi-Temporal Satellite Images. *Remote Sens.* **2019**, *11*, 1155. [CrossRef]
34. Manessa, M.D.M.; Kanno, A.; Sekine, M.; Haidar, M.; Yamamoto, K.; Imai, T.; Higuchi, T. Satellite-Derived Bathymetry Using Random Forest Algorithm and Worldview-2 Imagery. *Geoplanning J. Geomat. Plan.* **2016**, *3*, 117–126. [CrossRef]
35. Kogut, T.; Weistock, M. Classifying Airborne Bathymetry Data Using the Random Forest Algorithm. *Remote Sens. Lett.* **2019**, *10*, 874–882. [CrossRef]
36. Yunus, A.P.; Dou, J.; Song, X.; Avtar, R. Improved Bathymetric Mapping of Coastal and Lake Environments Using Sentinel-2 and Landsat-8 Images. *Sensors* **2019**, *19*, 2788. [CrossRef] [PubMed]
37. Bures, L.; Sychova, P.; Maca, P.; Roub, R.; Marval, S. River Bathymetry Model Based on Floodplain Topography. *Water* **2019**, *11*, 1287. [CrossRef]
38. Setiawan, K.T.; Suwargana, N.; Ginting, D.N.B.; Manessa, M.D.M.; Anggraini, N.; Adawiah, S.W.; Julzarika, A.; Surahman, S.; Rosid, S.; Supardjo, A.H. Bathymetry extraction from spot 7 satellite imagery using random forest methods. *Int. J. Remote Sens. Earth Sci. IJReSES* **2019**, *16*, 23–30. [CrossRef]
39. Moeinkhah, A.; Shakiba, A.; Azarakhsh, Z. Assessment of Regression and Classification Methods Using Remote Sensing Technology for Detection of Coastal Depth (Case Study of Bushehr Port and Kharg Island). *J. Indian Soc. Remote Sens.* **2019**, *47*, 1019–1029. [CrossRef]

40. Ha, N.T.; Manley-Harris, M.; Pham, T.D.; Hawes, I. A Comparative Assessment of Ensemble-Based Machine Learning and Maximum Likelihood Methods for Mapping Seagrass Using Sentinel-2 Imagery in Tauranga Harbor, New Zealand. *Remote Sens.* **2020**, *12*, 355. [CrossRef]
41. Misra, A.; Ramakrishnan, B. Assessment of Coastal Geomorphological Changes Using Multi-Temporal Satellite-Derived Bathymetry. *Cont. Shelf Res.* **2020**, *207*, 104213. [CrossRef]
42. Huber, M.E.; Zigic, S.; Gilbert, R.; Smith, D.; Edison, K.; Goudkamp, K.; Langtry, S.; Burling, M. Improved Dredge Material Management for the Great Barrier Reef Region. In Proceedings of the Australasian Port and Harbour Conference; Engineers Australia: Barton, Australia, 2013; p. 400.
43. Lara, J.L.; Lucio, D.; Tomas, A.; Di Paolo, B.; Losada, I.J. High-Resolution Time-Dependent Probabilistic Assessment of the Hydraulic Performance for Historic Coastal Structures: Application to Luarca Breakwater. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **2019**, *377*, 20190016. [CrossRef]
44. Nalona (Suction Dredger) Registered in—Vessel Details, Current Position and Voyage Information—IMO 9047453 | AIS Marine Traffic. Available online: https://www.marinetraffic.com (accessed on 18 September 2019).
45. Sentinel Application Platform. ESA Toolboxes, 2009. SNAP. Available online: http://step.esa.int/main/toolboxes/snap (accessed on 12 January 2020).
46. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-Resolution of Sentinel-2 Images: Learning a Globally Applicable Deep Neural Network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [CrossRef]
47. Rumora, L.; Miler, M.; Medak, D. Impact of Various Atmospheric Corrections on Sentinel-2 Land Cover Classification Accuracy Using Machine Learning Classifiers. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 277. [CrossRef]
48. Son, N.-T.; Chen, C.-F.; Chen, C.-R.; Guo, H.-Y. Classification of Multitemporal Sentinel-2 Data for Field-Level Monitoring of Rice Cropping Practices in Taiwan. *Adv. Space Res.* **2020**, *65*, 1910–1921. [CrossRef]
49. Alwhaely, U.; Hussein, M.A.; AL-Kaaby, L.F. Using GIS and Remote Sensing Satellite Data to Mapping and Monitoring Shatt Al-Arab Estuary (out Bar Area) and Nearby Coastline Southern Iraq. *Al-Qadisiyah J. Pure Sci.* **2020**, *25*, 1–21. [CrossRef]
50. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
51. Reiss, H.; Cunze, S.; König, K.; Neumann, H.; Kröncke, I. Species Distribution Modelling of Marine Benthos: A North Sea Case Study. *Mar. Ecol. Prog. Ser.* **2011**, *442*, 71–86. [CrossRef]
52. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Cham, Switzerland, 2013; Volume 112.
53. Prasad, A.M.; Iverson, L.R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181–199. [CrossRef]
54. Peters, J.; Baets, B.D.; Verhoest, N.E.C.; Samson, R.; Degroeve, S.; Becker, P.D.; Huybrechts, W. Random Forests as a Tool for Ecohydrological Distribution Modelling. *Ecol. Model.* **2007**, *207*, 304–318. [CrossRef]
55. Kuhn, M. Caret: Classification and Regression Training. R Package Version. Available online: https://www.R-project.org (accessed on 5 February 2021).
56. Mateo-Pérez, V.; Corral-Bobadilla, M.; Ortega-Fernández, F.; Vergara-González, E.P. Port Bathymetry Mapping Using Support Vector Machine Technique and Sentinel-2 Satellite Imagery. *Remote Sens.* **2020**, *12*, 2069. [CrossRef]
57. Confederación Hidrográfica del Cantábrico. Available online: https://www.chcantabrico.es/las-cuencas-cantabricas/marco-fisico/hidrologia/rios/negro (accessed on 12 November 2019).