




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



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

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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.

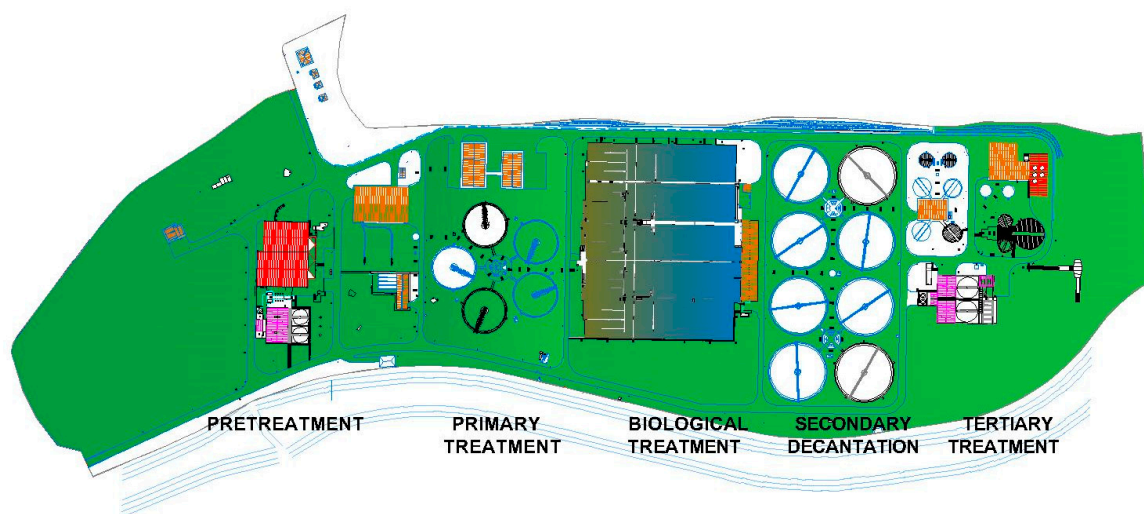


Figure 1. Plan view of the Villapérez wastewater treatment plant (WWTP) (Asturias, Spain).

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 m³/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 m³/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 m³. 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 m³/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.

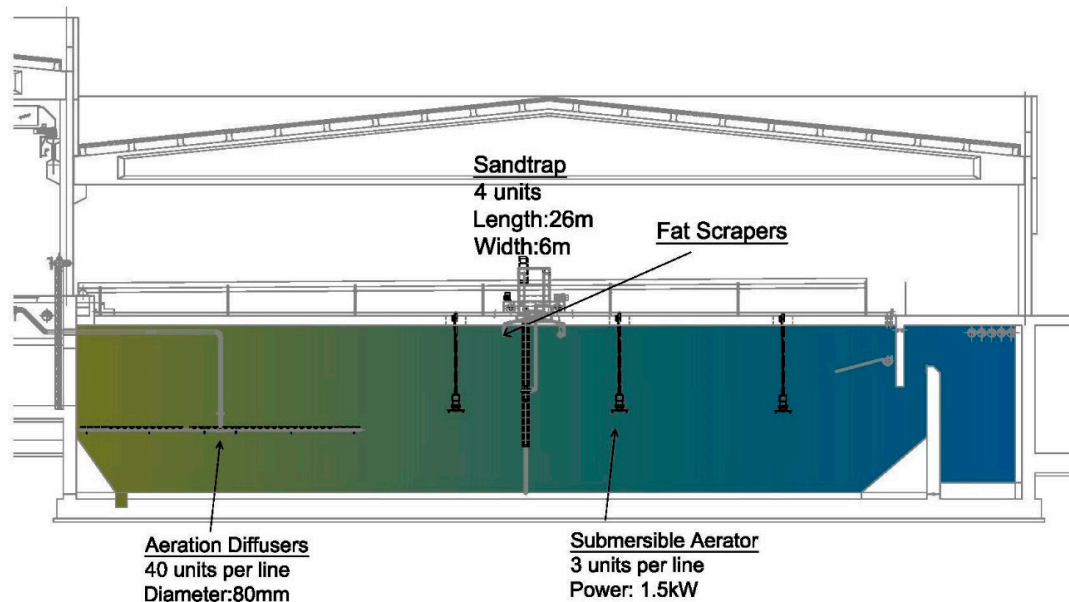


Figure 2. A sectional view of a FOG concentrator.

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 (BOD ₅)	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	Maximum flow	m ³ /h	9216.91	3958.30	3446.37	17,885.11
MinFlow	Minimum 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	Maximum Ambient Temperature	°C	17.50	5.12	4.60	28.20
TempExtMin	Minimum 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.

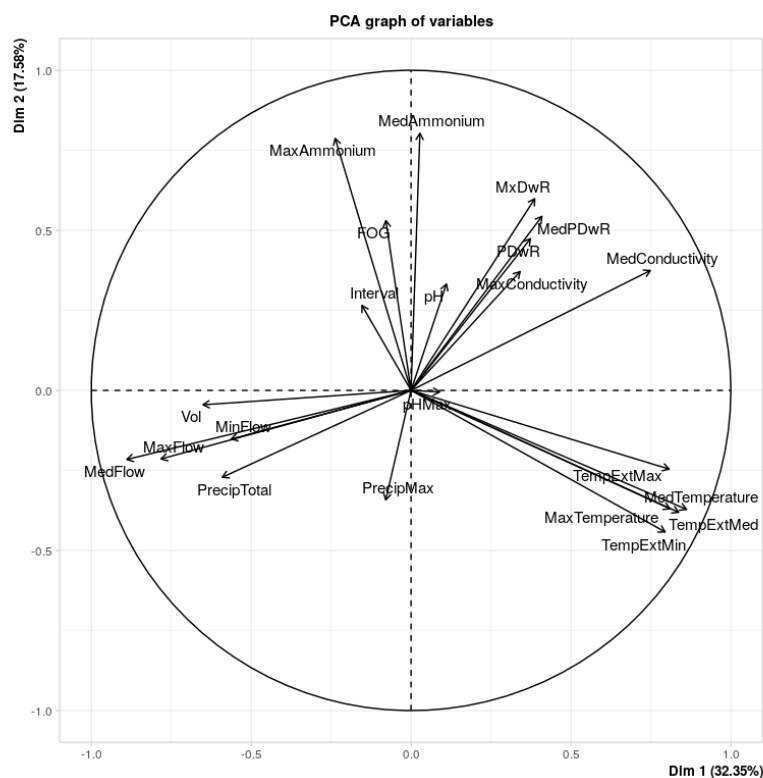


Figure 3. PCA graph of variables.

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.

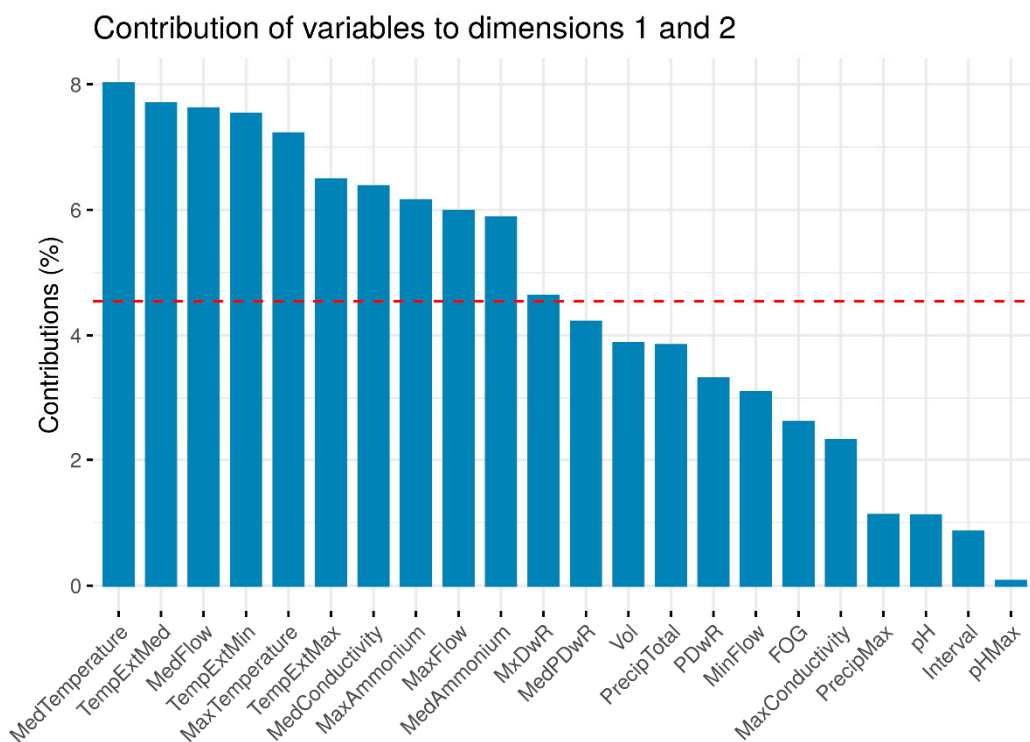


Figure 4. Contribution of variables to Dimensions 1 and 2.

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.

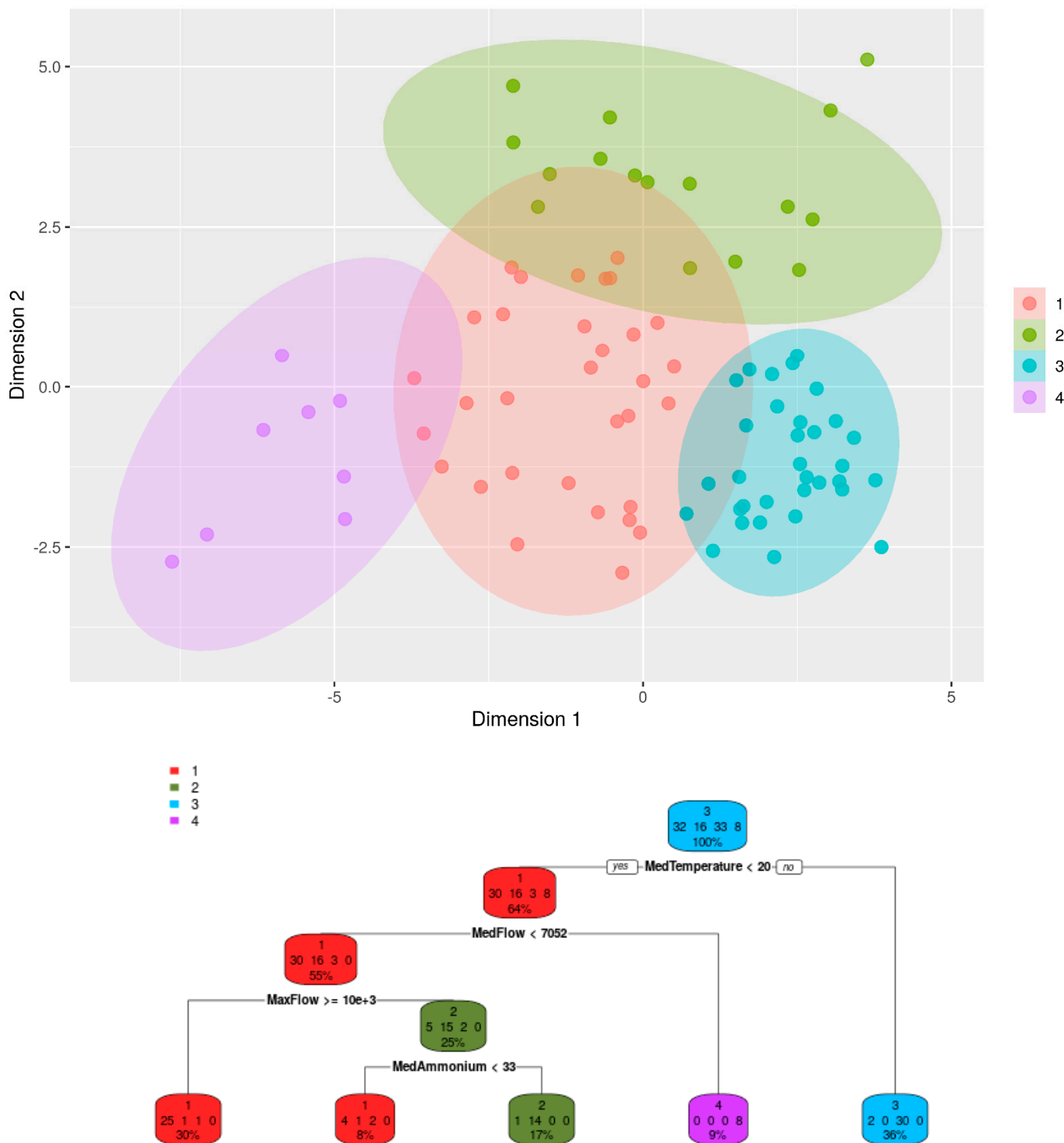


Figure 5. FOG Clusters of PCA projection.

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 m³/h;
- Cluster 2 part of the cases with a maximum flow greater than 10,000 m³/h and also with average ammonia 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 m³/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 ammonia values (MedAmmonium) correspond to cluster 2.

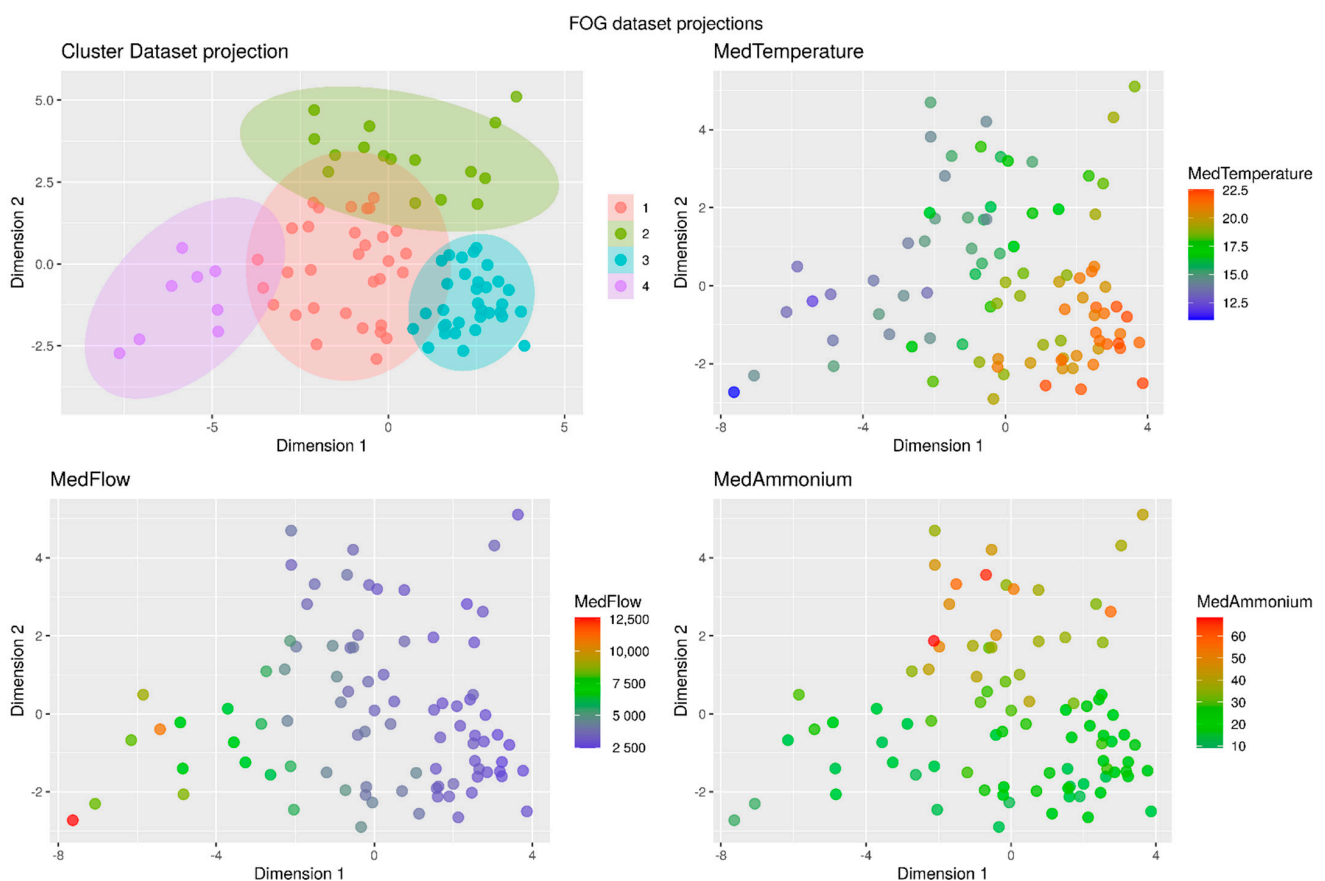


Figure 6. Comparing dataset PCA projection with clustering variables.

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 (mtry) 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.

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

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

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

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.

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

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.

	<i>RMSE</i>	<i>MAE</i>	<i>R</i> 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.

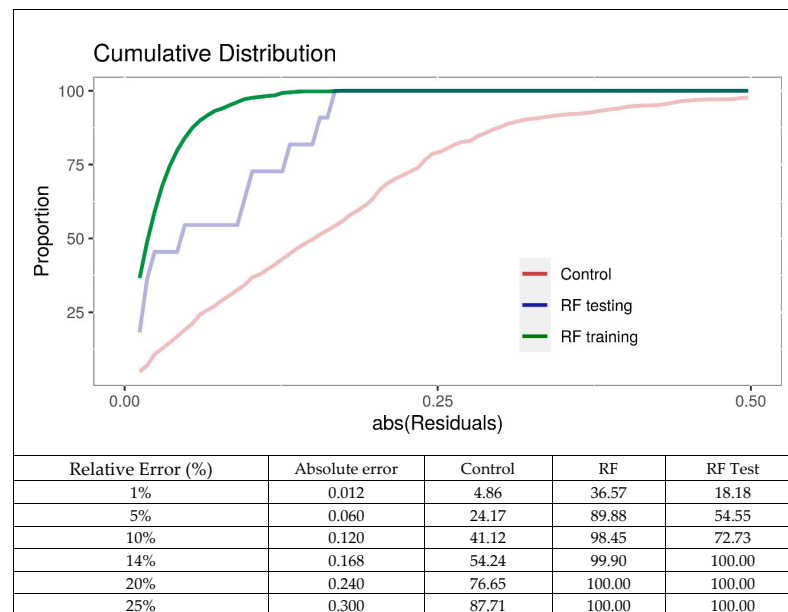


Figure 7. Model performance indicators.

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.

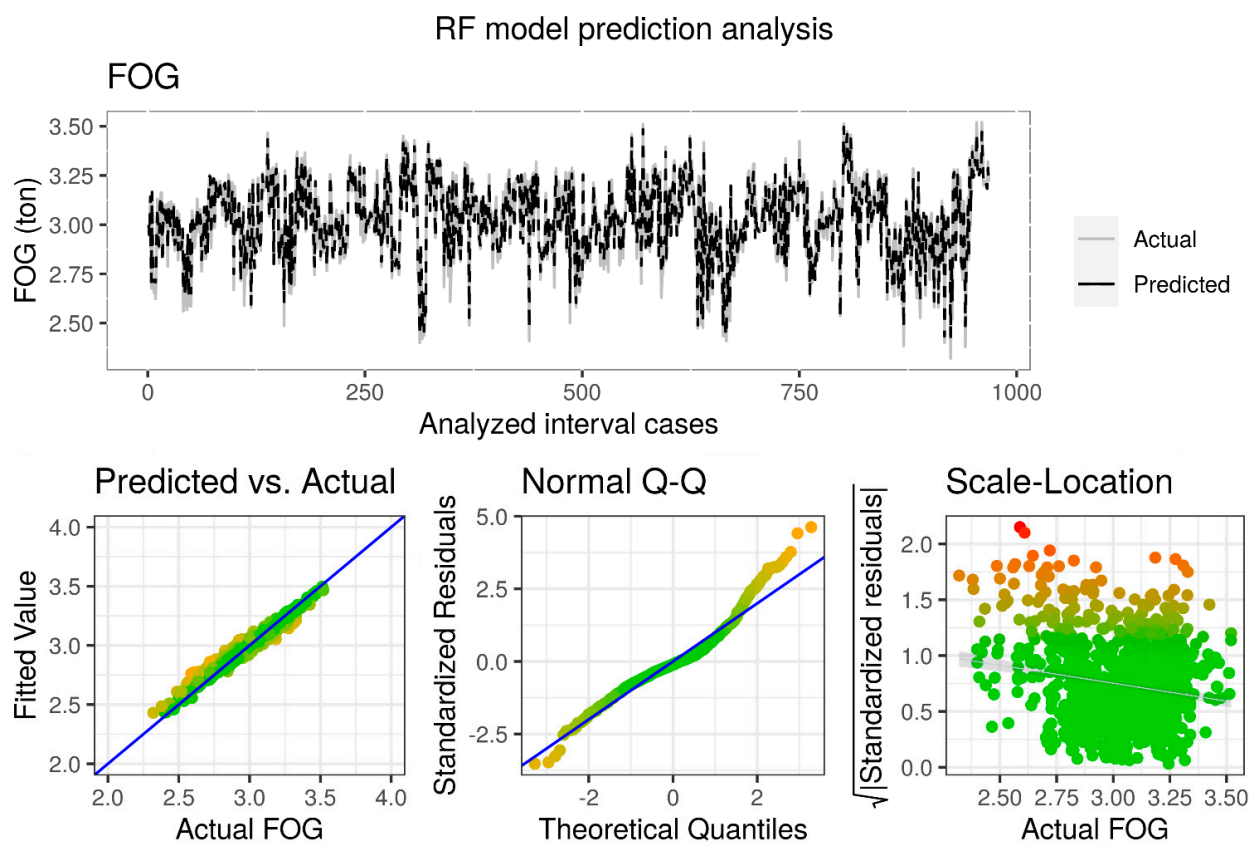


Figure 8. RF model prediction analysis (training).

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.

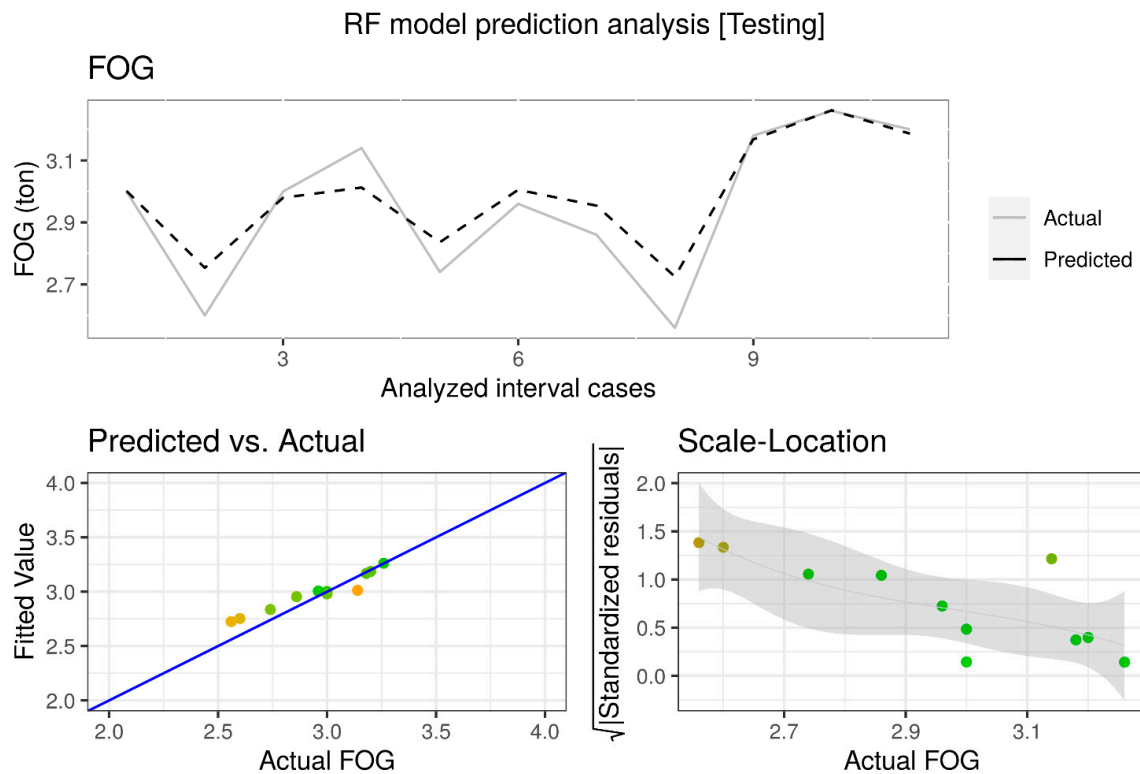


Figure 9. RF model prediction analysis (testing).

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.

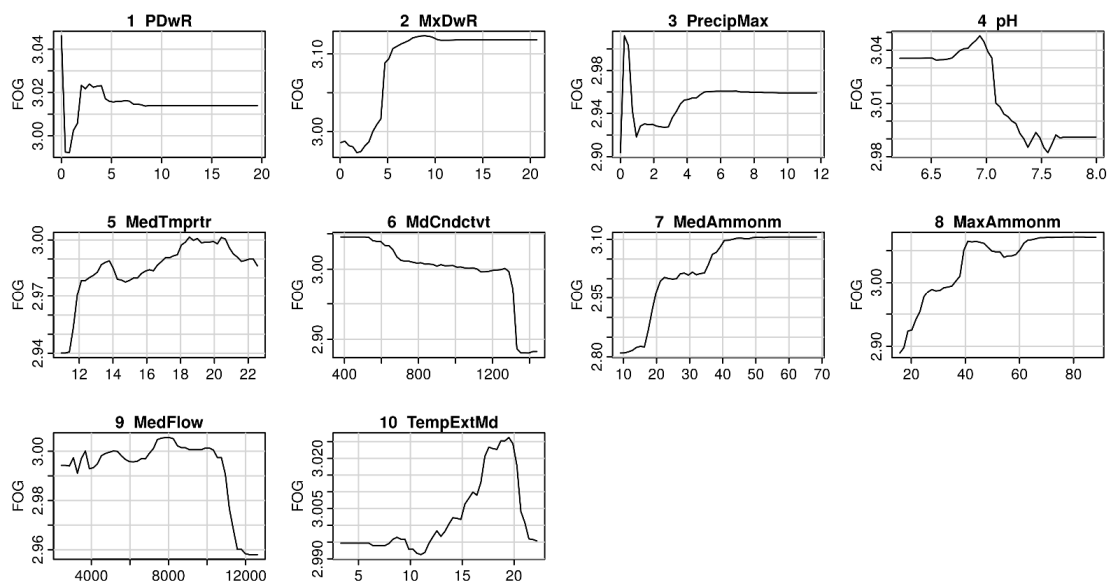


Figure 10. Sensitivity analysis (one variable).

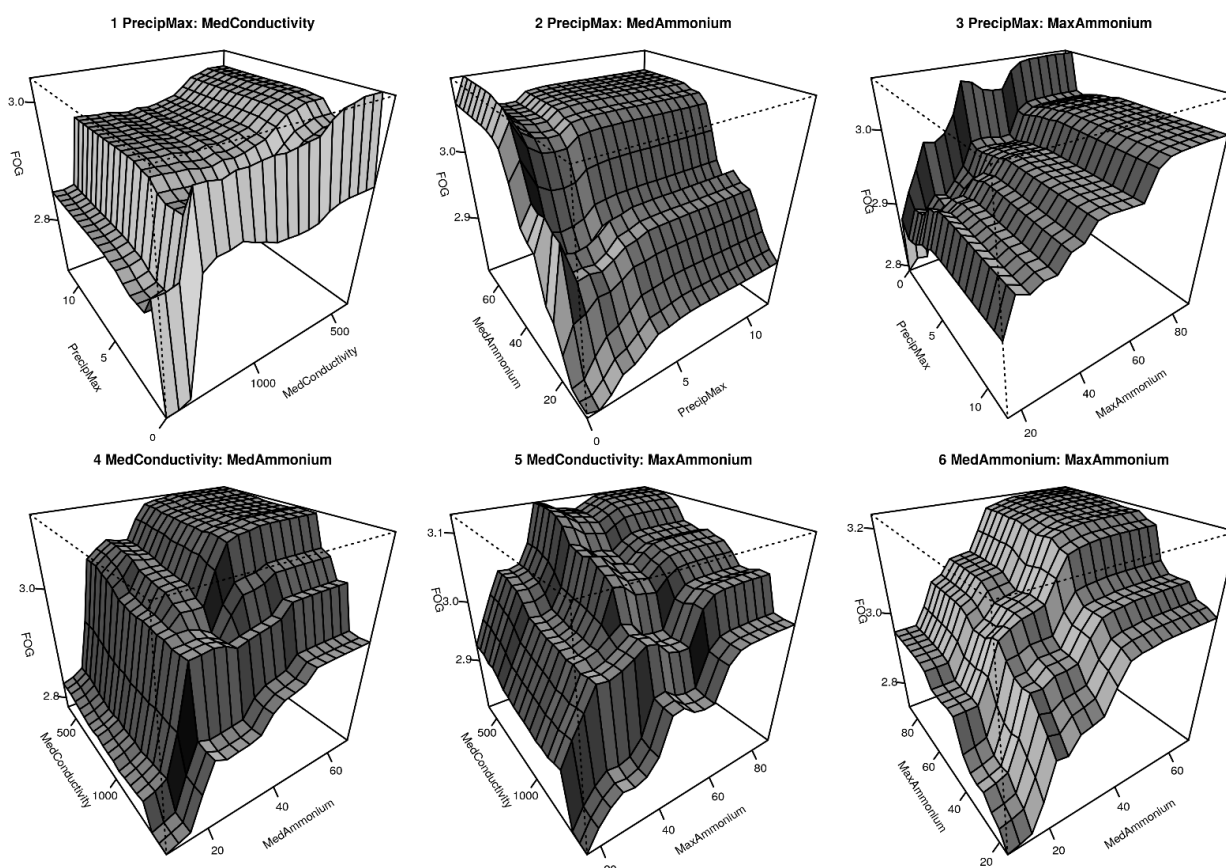


Figure 11. Sensitivity analysis (two variables).

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

- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Arthur, S.; Blanc, J. *Management and Recovery of FOG (Fats, Oils and Greases)*; CREW—Scotland's Centre of Expertise for Waters: Edinburgh, UK, 2013.
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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.
- 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. [\[CrossRef\]](#)
- 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.
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Agabo-García, C.; Solera, R.; Pérez, M. First Approaches to Valorize Fat, Oil and Grease (FOG) as Anaerobic Co-Substrate with Slaughterhouse Wastewater: Biomethane Potential, Settling Capacity and Microbial Dynamics. *Chemosphere* **2020**, *259*, 127474. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bratina, B.; Šorgo, A.; Kramberger, J.; Ajdnik, U.; Zemljic, 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- 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 Settability 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 Effectors 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. Szelag, 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. Stenftoft, 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](#)]