Jul 13th, 2:30 PM - 2:50 PM

A Multi-disciplinary Procedure to Ascertain Biofilm Formation in Drinking Water Pipes

Eva Ramos-Martinez
FluIng-IMM Universitat Politècnica de València, evarama@upv.es

Manuel Herrera
University of Bath, amhf20@bath.ac.uk

Joaquín Izquierdo
FluIng-IMM Universitat Politècnica de València, jizquier@upv.es

Rafael Pérez-García
FluIng-IMM Universitat Politècnica de València, rperez@upv.es

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

A Multi-disciplinary Procedure to Ascertain Biofilm Formation in Drinking Water Pipes

Eva Ramos-Martínez[a], Manuel Herrera[b], Joaquín Izquierdo[a], Rafael Pérez-García[a]
[a]FluIng-IMM Universitat Politècnica de València. Camino de Vera s/n, 46022 Valencia, Spain.
[b]EDEn - ACE Dept. – University of Bath, Bath, UK
e-mail: {evarama, jizquier, rperez}@upv.es, amhf20@bath.ac.uk

Abstract: Biofilm development in drinking water distribution systems (DWDSs) is a real problem negatively affecting service and water quality, and, thus, the satisfaction of the final consumers. It is the direct and indirect responsible for many of the DWDSs' problems, and a lot of resources are invested to mitigate its effects. Addressing this problem has been a concern of researchers and DWDS managers for years. However, it is only recently that both technology and data have been available to support the new approach presented in this work. Our proposal is based on the combination of various existing data sets from similar studies to conduct a meta-data analysis of biofilm development. The approach lies on an intensive data pre-processing. Having a complete and extensive database on biofilm development in DWDSs allows applying Machine Learning techniques to develop a practical model. It is based on a multidisciplinary research vision to formulate effective biofilm control strategies. This work presents the basis for the development of a useful decision-making tool to assist in DWDS management. The negative effects on service and consumers caused by biofilm would be mitigated maintaining it at the lowest level. The performance of the suggested models is tested with data coming from two different case-studies: the DWDSs of the city of Thessaloniki (Greece) and the Pennine Water Group experimental facility (UK). The results obtained validate this methodology as an excellent approach to studying biofilm development in DWDSs.

***Keywords**: biofilm; drinking water distribution system; pre-processing; random forests; regression trees*.

1 INTRODUCTION

The most important factor in planning and operating a water distribution system is satisfying consumer demand. This means continually providing users with quality water in adequate volumes at reasonable pressure, and so ensuring a reliable water distribution system. While there is a plethora of different approaches for analysing and predicting water demand [Herrera et al., 2010; Brentan et al., 2016], studies on the water quality supplied to the final customer are not so frequent. In this regard, most of the reported problems by end-users to water utilities are aesthetic deterioration of water (resulting in colour, odour and taste degradation) [Vreeburg and Boxall, 2007], while operational problems caused by pipe biocorrosion are also of concern [Lopes et al., 2009]. Both problems usually have a common origin in communities of microorganisms growing within the inner pipe walls in contact with water, also known as biofilm. Biofilm also reduces flow speed and pipe capacity of circulation [Cowle et al., 2014]. Most importantly, biofilm formation can be involved in health issues deriving from both its associated disinfectant decay and its role as a pathogen shelter [Adhikari et al., 2012; Ashbolt 2015].

This paper proposes a multidisciplinary approach aiming at formulating effective biofilm control strategies. Numerous studies have been carried out in relation to the influence that pipes and water flow characteristics have on biofilm development. However, the considered features affecting biofilm are studied individually or at most in pairs [Shaw et al., 2014; Wu et al., 2015]. One of the main novelties introduced in this work is the investigation of a larger number of conditions that ease biofilm development in pipes, since it is currently accepted that biofilm formation depends on complex interactions among such aspects as water quality, infrastructure, and operational factors associated with distribution systems. Obtaining biofilm samples, that are representative of the spatial, temporal and physicochemical variation of real drinking water distribution systems (DWDSs), is highly

challenging, since they are live, functioning systems comprised of buried infrastructure [Fish et al., 2015]. Consequently, much of the current understanding about DWDS biofilms is based upon from pilot or bench-top scale experimental models of drinking water systems [Fish et al., 2015], under simplified conditions. To cope with these drawbacks, this work proposes using both data collected in a thoroughgoing, state-of-the-art review, and data obtained in two case studies.

The first source of data used in this work is derived from a combination of information found in literature and data directly provided by other researchers thanks to their collaboration in this proposal. All the information is pre-processed and homogenised through suitable approaches to outlier detection, selection of variables, and handling missing data. An Exploratory Data Analysis is done over the now clean and complete synthetic database. This provides useful insights to carry out further statistical analysis to understand biofilm formation by Machine Learning (ML) methods. In this case, Regression Trees and Random Forest algorithms have been applied. This methodology is validated when observing the good results obtained when testing the performance of the models with data from two different case-studies: the DWDS of the city of Thessaloniki (Greece) and the Pennine Water Group experimental facility (UK).

2 DATA SCIENCE BASED APPROACH

When studying biofilm development in DWDSs, getting field data is an arduous task which requires high workload and time, while developing experimental laboratory studies is still very complex and highly qualified staff and equipment are needed. In both cases, time tends to be too long and the amount of data obtained scarce. These difficulties, along with the complexity of the communities and environment under study, result in studies' simplification [Ramos-Martínez, 2016]. Generally, no more than one or two factors in relation to biofilm development are studied and/or simplified growth devices are used.

Currently, we have at our disposal technology and data of great quality to support new research approaches. In this context, the undertaken work and the acquired knowledge over the last years in the field of the study of biofilm development in DWDSs have been used. We propose the collection and pre-processing of the data obtained during these years of research to overcome the difficulties found when acquiring DWDSs' biofilm data. That is, the combination of multiple datasets on similar studies to carry out an analysis of these meta-data and study the biofilm development in DWDSs through partial views of the problem.

2.1 Data collection

The first step in data collection is to deeply understand the treated subject. When studying the biofilm development in DWDSs it is important to know that microbiological aspects are not the only explanation aspects. Biofilm also depends on a complex interaction between water quality, infrastructure and operational factors of the system itself.

Biofilm data have been collected from previous research works on biofilm development in DWDSs (the selected papers, published from 1998 to 2013, cannot be included here for space reasons; we refer the reader to [Ramos-Martínez, 2016] for an exhaustive list). The journal papers analysed for the study have been obtained from various scientific search engines such as Web of Science, Google Scholar, IEEE Xplore Digital Library and ScienceDirect, among others. They are all search engines for scientific and academic research that search directly for articles in peer-reviewed and well-regarded publications. The main searched keywords have been *biofilm*, *drinking water distribution systems*, *HPC/cm*^{2} and *R2A*, and the various combinations among them. (*HPC* stands for heterotrophic plate count). The papers found under these criteria have been studied to be included in the data compilation. All the measurements associated with HPC/cm^2 biofilm data have also been compiled.

To begin with, some of the main criteria used to exclude data from the study are:

1. Studies based on cultured communities seeded with investigator-selected species or developed using an inoculum.
2. Biofilm developed on unrepresentative materials for DWDSs. The use of glass coupons within annular reactors is very common.

3. Cases where synthetic water is used or the quality of the water is modified, turning away from the common drinking water conditions (e.g.: increasing the concentration of an element over its natural range in normal conditions). If applicable, just the data obtained under control conditions have been selected.
4. The data obtained when a product different to chlorine, or none, is used as secondary disinfectant. Disinfection practices vary widely in European countries, being the previously mentioned the two mainly used.
5. Biofilm data not obtained under R2A long incubation conditions, i.e., 5 to 7 days of incubation between 20 or 28ºC [Reasoner, 2004].

2.2 Data pre-processing

Knowledge is often scattered in a bunch of different sources and in different forms that must be synthesized and turned into clean processed data before any serious analysis. Getting and pre-processing data means transforming raw data into clean data ready for analysis. In fact, pre-processing often ends up being the most important component of the data analysis in terms of effect on the downstream data, and so, it is critically important [Gibert et al., 2008].

The data has been collected in a typical data format, into a rectangular array with one row per experimental subject and one column for each subject identifier, outcome variable, and/or explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable. The compiled variables can be classified in four groups attending to their nature: physical characteristics, hydraulic characteristics, sampling and incubation, and physico-chemical characteristics of water. The target variable, the biofilm data, has been called “hpc”.

At this point, we have a data set with 409 cases and more than 60 variables. However, most of these variables have a lot of missing values. Since we work with multiple data sources we found a huge variability of measurements resulting in a great number of missing values. For example, such a key element as water content of organic carbon is measured in almost all the papers. However, different organic carbon fractions are measured, resulting in not comparable data and great number of missing values. The data cleansing procedure in summarized in Figure 1. For outlier detection an unsupervised anomaly detection algorithm is applied, the local outlier factor (LOF) algorithm. After cleaning, the data set is formed by 284 examples and 15 attributes.

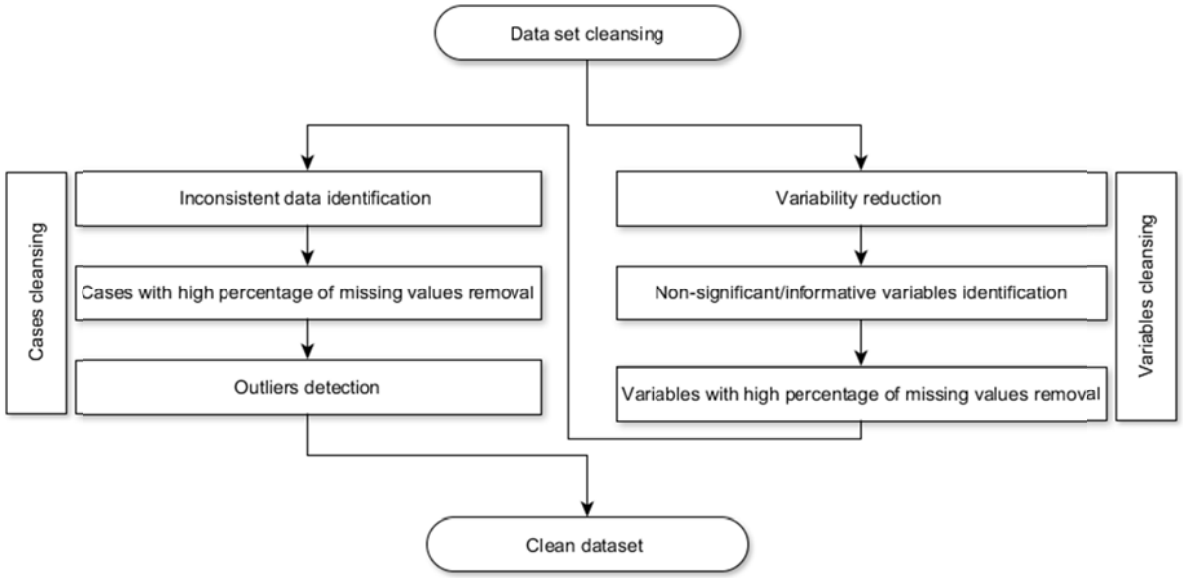

Figure 1. Data cleansing process

The database presents 80% of complete cases (228), since 4 of the 15 variables (*Pipe material*, *Water Source*, *Water Temperature* and *Free Chlorine*) have missing values. The variables *Water Temperature* and *Free Chlorine* are the ones with highest number of missing values, while the number of missings in the variables *Pipe material* and *Water Source* are very scarce (1 and 6, respectively).

The presence of missing values is a common problem in data analysis. In the case where removing variables or observations with missing data is not an option it must be resorted to fill in or "impute" missing values. Imputation methods keep the full sample size, which can be advantageous for bias and precision. To solve this problem we have applied Multivariate Imputation by Chained Equations (MICE) that has emerged as a principled method of dealing with missing data [Azur et al., 2011]. After data set reconstruction the final data set is formed by 284 complete cases with 14 attributes and a target variable (*hpc*). Finally, to be sure that there are not hidden correlations a survey plot has been carried out on the complete cases. At this point, the pre-processing step is finished. The variables and categories of the resulting data set are presented below.

1. Physical characteristics of the system
   - Device (*device*): propella reactor *PR,* flow cell system *FC*, annular reactor *AR*, Robbins device *RD*, Pedersen device *PE*, direct *D* (samples obtained directly from real DWDSs), pipe *P* (samples obtained from pilot scale system)
   - Tested material (*material): t*hermoplastic polymers *TP*, iron based *I*, steel based *S*, cement based *C*
   - Duct's shape (*pipe_like*): Yes *Y,* No *N*
2. Hydraulic characteristics of the system
   - Circulation type (*c_type*): Single pass *SP* (water flowing past the device does not return), continuous *C* (there is some recirculation of water), no continuous *NC* (water is constantly recirculating; there is no renewal)
   - Constant circulation (*c_constant*): Yes *Y,* No *N*
3. Sampling and Incubation
   - Removal technique (*removal*): low *L*, medium *M*, strong *S*
   - Type of insert (*insert*): slide *S*, coupon *C,* direct *D* (samples are directly taken from the pipe wall or from inserts that do not stand above the pipe wall and respect the curvature of the pipe, simulating the real conditions of DWDS pipes)
   - Incubation time (*inc_time*)
   - Incubation temperature (*inc_temp*)
   - Plating method (*culture*): spread plate *S*, pour plate *P*
4. Physico-chemical characteristics of water
   - Water itinerary (*itinerary*): from the tap *T*, from the water treatment plant *TR*
   - Water source (*w_source*): groundwater *G*, superficial water *S*
   - Water temperature (*w_temp*)
   - Residual free chlorine concentration (*freeCl*)
6. Biofilm
   - Log of R2A cultivable cell per cm^2 in biofilm (*hpc*)

2.3 Model development

Nowadays, data is not only becoming more accessible but also more understandable to computers and analysts. Data driven solutions are rapidly advancing and becoming very valuable tools. ML methods have a leading role in this transformation of data into valid and useful knowledge.

2.3.1 Regression Trees

Due to the nature of our synthetic database, there are incidental or inherent dependencies that make metadata present a trend towards a natural hierarchical structure. Applying the Regression Tree (RT) methodology to the complete database obtained allows us to develop a valid model to explore.

RTs are ML methods for constructing non-linear prediction models from data. The models are obtained by recursively partitioning the data space and fitting a simple prediction model within each partition [Loh, 2011]. As a result, the partitioning can be represented graphically as a decision tree. Each of the terminal nodes, or leaves, of the tree represents a cell of the partition, and has attached a simple model which applies in that cell only. RTs are a piecewise-constant models. There are several advantages associated to this approach [Shalizi, 2006]:

1. Making predictions is fast.
2. It is easy to understand what variables are important in making the prediction.

3. Because the algorithm asks a sequence of hierarchical Boolean questions, it is relatively simple to understand and interpret the results.
4. If some data is missing, we might not be able to go all the way down the tree to a leaf, but we can still make a prediction by averaging all the leaves in the sub-tree we do reach.
5. The model gives a jagged response, so it can work when the true regression surface is not smooth. If it is smooth, though, the piecewise-constant surface can approximate it arbitrarily closely (under the assumption of having enough leaves).
6. There are fast, reliable algorithms to learn these trees.

The RT analysis has been implemented through the R package “rpart” version 4.1-10 [Therneau et al., 2015].

2.3.2 Random Forests

Random Forest (RF) algorithms are ensemble learning algorithms. As a result, they can be more accurate and robust to noise than single classifiers [Rodríguez-Galiano et al., 2012]. An RF [Breiman, 2001] is an ensemble classifier consisting of many decision trees, where the final predicted class for a test example is obtained by combining the predictions of all individual trees. Each tree contributes with a single vote for the assignment of the most frequent class to the input data [Rodríguez-Galiano et al., 2012]. An RF uses a random feature selection, a random subset of input features or predictive variables in the division of every node, instead of using the best variables, which reduces the generalization error. Additionally, to increase the diversity of the trees, an RF uses bootstrap aggregation (bagging) to make the trees grow from different training data subsets [Gray et al., 2013]. In summary, an RF is an all-purpose model that performs well on most problems, can handle noisy data, categorical or continuous features, and selects only the most important features [Lantz, 2013].

The RF algorithm used has been implemented through the R package “randomForest”, version 4.6-12 [Breiman et al., 2015]. The regression type of random forest has been used. An ensemble of 500 trees has been created and the number of variables tried at each split has been set to 5. The goal of using a large number of trees is to train enough so that each feature has a chance to appear in several models.

3 MODEL PERFORMANCE

It is important to note that not all the variables have been used in the RT construction. Actually, just *culture, device, freecl, inc_temp, itinerary, material, removal* and *w_temp* variables have been used. This means that the variables that have not been used (*pipe_like, c_type, c_constant, insert, inc_temp* and *w_source*) have been considered not relevant for the construction of the model. The tree is split in the first place by the *device* variable. The devices P, D and AR are grouped together, therein suggesting that have a similar behaviour. That is, the cylinder devices that are more similar to the real pipe conditions have been separated from the rest of the devices that do not resemble a pipe. These are PE, RD and FC. The branch of the P, D and AR devices is just split by the *removal* variable, thus suggesting that it is an important issue to take into account when sampling. It can influence the obtained results and, thus, the possible comparisons among different studies.

In the RF, since there is not a graphical representation, we focus on %IncMSE. It increases with importance of the variable. We observe that *inc_temp* is especially important. This variable has been pointed as one of the most important in the previous RT. However, the most relevant in the previous case was the *device* variable, which in the RF is third in importance. In the second place, with a value very similar to the *device* variable, we find the *culture* variable. It enhances its already known importance [Van Soestbergen and Lee, 1969] when comparing HPC results.

Prior to applying the algorithms to the synthetic database, a stratified sampling has been carried out to keep a representative amount of the model data to be, subsequently, used to test the performance of the final model. The number of data kept for test is 20; thus, the analysis has been performed with the 265 remaining data.

The models have been tested with the metadata kept from the model (n = 20) and with the real data obtained from two case studies (n = 9). Data obtained directly from the DWDS of the city of

Thessaloniki in Greece (n = 7) and data from the Pennine Water Group experimental facility of the University of Sheffield in UK (n = 2). The performance of the models has been measured by the Pearson correlation coefficient [Benesty et al., 2009].

When the models are tested with the data extracted from the database (Figure 2), a Pearson correlation coefficient of almost 0.9 is obtained for both models. Both values are quite high, and, although similar, they show that the RF performance is slightly better.

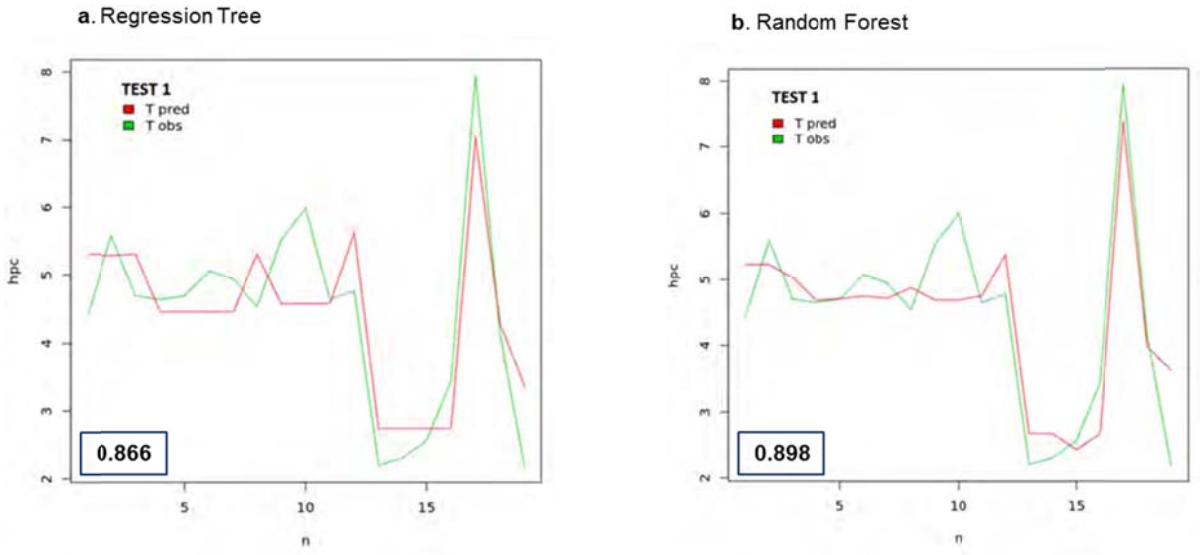

Figure 2. Performance of the models after Test 1

In the second test (Figure 3), the data from the case studies are used. The obtained results still being good, however, provide lower values than in the first test (Figure 2). In this second test all the data are from real DWDSs, where the environmental variability is greater than in lab scale or bench top models. There are also missing values (*Water Temperature* and *Free Chlorine* variables) in data 3, 4, 5 and 6. Both reasons are probably affecting the performance of the models (Figure 3). In both tests, the RF model fits better to the data. In the second test, the performance of the RF model is clearly better than RT model's.

The performance of the models decreases when comparing it with the results obtained in Test 1 (Figure 2). It is, probably, because in the second test all the data are from real DWDSs, where the environmental variability is greater than in lab scale or bench top models. The presence of missing values (*Water Temperature* and *Free Chlorine* variables) in data 3, 4, 5 and 6 is probably affecting the performance of the models. In both tests, the RF model fits better to the data. The good performance of the ensemble techniques on this approach has been already observed when applying them to biofilm metadata [Ramos-Martínez, et al. 2014]. For example, it can be observed that the behaviour of last two points (Sheffield data) improve with the RF model (Figure 3). In this case, the RF model takes properly into account the variability in disinfectant concentration observed in the Sheffield cases. It assigns more biofilm development to the case with less chlorine concentration, reducing the error, while the RT gives the same value to both cases. In general, Figure 3 shows how the RF model adapts better to the tested data.

4 CONCLUSIONS

This work offers a comprehensive knowledge discovery in databases (KDD) framework to accelerate biofilm analysis using technological advances and data science. We propose data pre-processing techniques to compile the currently available information of the DWDS conditions that affect biofilm development in order to be able to study the effect that the joint influence of these characteristics has in biofilm development. Our proposal is to achieve pre-processing of all the work already developed in this field, preparing a case study database to do inferences by posterior ML analyses. The implementation of this family of techniques in the study of biofilm development in DWDSs opens a vast field to explore with promising results.

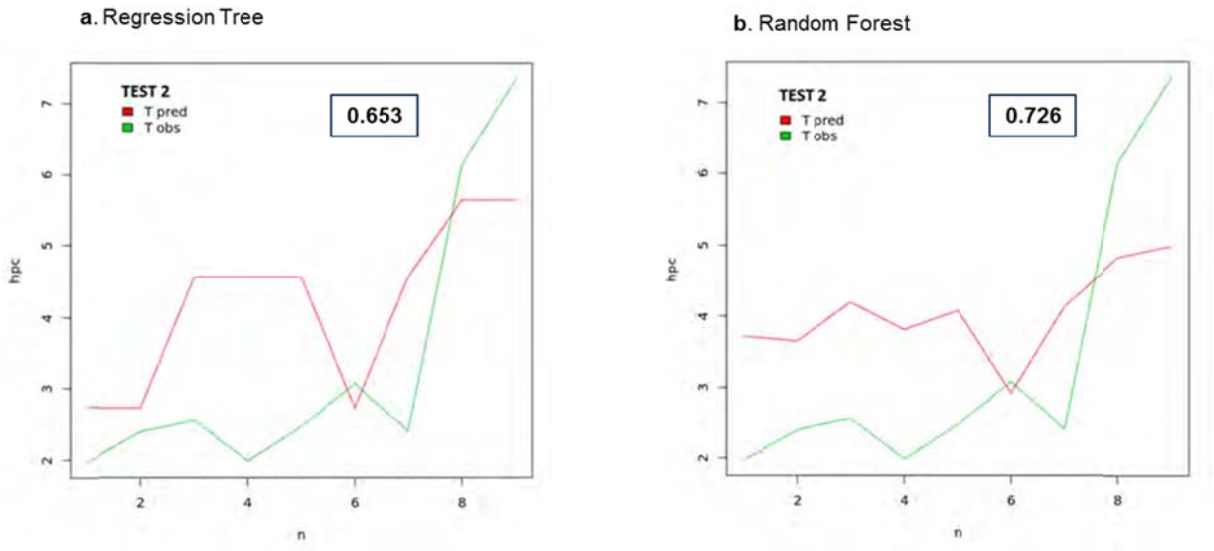

Figure 3. Performance of the models after Test 2

Although, unlike the RT, the RF model is not easily interpretable and may require some work to tune the model to the data, its performance has demonstrated to be better in this case. The fact that RF is an ensemble learning algorithm confer to it very valuable properties that make it more robust and proper for our study. The variability found in the prediction of the different data points could suggest that there are some cases that are better or worse represented in the database making their prediction more robust or weaker. Other possible explanation could be related to the microbial ecology of the biofilm. The cases better predicted may correspond to those situations where biofilm development is mainly influenced by the studied variables, so its behaviour is well performed by the model. In contrast, the prediction may be less accurate in those cases that other factors, not taken into account in the model, are more influential. In both cases, it is suggested that adding new data and increasing the database size would help create a more robust model.

According to the RF obtained results there are some variables that are, clearly, more influential in the model prediction, namely: *inc_temp*, *device*, *culture* and *freecl.* The fact that three of the four more influential variables are related with the research methodology and not with the environment where the biofilm has grown enhances the importance of developing a standard protocol for the study of biofilm in DWDSs. It could allow faster progression in DWDS biofilm research, achieving more practical and implementable results.

Biofilm development in DWDSs is a real problem negatively affecting the service and water quality offered by water utilities, and, thus, satisfaction of the final consumers. Addressing this problem has been a concern of researchers and DWDS managers for years, but it is now that technology and data have been available to support the new approach that we present it in this work. We have developed a scalable and interesting set of tools to understand biofilm behaviour with respect to its environment, and develop models that can be used as decision-making tools in DWDS management to mitigate biofilm negative effects on the service by: improving flushing location and frequency, identifying chlorination point location, or helping with the design of DWDSs or new sections of them, among others.

REFERENCES

Adhikari, R.A., Sathasivan, A., Bal Krishna, K.C., 2012. Effect of biofilms grown at various chloramine residuals on chloramine decay. Water Sci. Tech.: Water Supply, 12(4), 463-469.

Ashbolt, N.J., 2015. Microbial contamination of drinking water and human health from community water systems. Current Environ. Health Reports, 2(1), 31-47.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: What is it and how does it work? Int. J. Meth. Psych. Res., 20(1), 40-49.

Benesty, J., Chen, J., Huang, Y., Cohen, I. 2009. Springer Topics in Signal Processing. Ch. Pearson Correlation Coefficient, 1-4, Springer.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5-32.

Breiman, L. 2015. Fortran original Report by Liaw, A., Cutler, A., Wiener, M. Breiman and Cutler's Random Forests for Classification and Regression, Version 4.6-12, 2015.

Brentan, B.M., Luvizotto Jr, E., Herrera, M., Izquierdo, J., Pérez-García, R., 2016. Hybrid regression model for near real-time urban water demand forecasting. J. Comput. Appl. Math., (in press).

Cowle, M.W., Babatunde, A.O., Rauen, W.B., Bockelmann-Evans, B.N., Barton, A.F. 2014. Biofilm development in water distribution and drainage systems: dynamics and implications for hydraulic efficiency. Environ. Tech. Rev., 3(3), 3147.

Demsar, J., Zupan, B., 2004. Orange: from experimental machine learning to interactive data mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

Fish, K.E., Collins, R., Green, N.H., Sharpe. R.L., Douterelo, I., Osborn, A.M., Boxall, J.B., 2015. Characterization of the physical composition and microbial community structure of biofilms within a model full-scale drinking water distribution system. PloSone, 10(2):e0115824.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A. 2013. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, 65:167-175.

Gibert, K., Izquierdo, J., Holmes, G., Athanasiadis, I., Comas, J., Sànchez-Marrè, M., 2008. On the role of pre and post-processing in environmental data mining, Proceedings of iEMSs 2008 International Congress on Environmental Modelling and Software, 1937-1958.

Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. J. Hydrol., 387(1), 141-150.

Lantz. B., 2013. Machine Learning with R. PACKT Publishing.

Loh, W. Y. 2011. Classification and regression trees, John Wiley and Sons, Inc.,

Lopes, F., Moprin, P., Oliveira, R., Melo, L., 2009. Impact of biofilms in simulated drinking water and urban heat supply systems. Int. J. Environ. Eng., 1(3), 276-294.

Ramos-Martínez, E. Assessing biofilm development in drinking water distribution systems by Machine learning methods. Ph.D. thesis, Universitat Politècnica de València, Spain.

Ramos-Martínez, E., Herrera, M., Izquierdo, J., Pérez-García, R. 2014. Ensemble of naïve bayesian approaches for the study of biofilm development in drinking water distribution systems. Int. J. Comput. Math., 99 (1), 135-146.

Reasoner, D.J., 2004. Heterotrophic plate count methodology in the United States. Int. J. Food Microb., 92(3), 307-315.

Rodríguez-Galiano, V.F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P.M., Jeganathan, C., 2012. Random forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sensing Environ., 121, 93107.

Shalizi, C., 2006. Regression Trees. Department of Statistics. Carnegie Mellon University.

Shaw, J.L., Monis, P., Fabris, R., Ho, L., Braun, K., Drikas, M., Cooper, A., 2014. Assessing impact of water treatment on bacterial biofilms in drinking water distribution systrems using high-throughput DNA sequencing. Chemospehere, 117, 185-192.

Therneau, T., Atkinson, B., Ripley, B. 2015. Recursive Partitioning and Regression Trees. R Package, Version 4.1-10. Available from http://cran.R-project.org.

Vreeburg, I.J., Boxall, J., 2007. Discolouration in potable water distribution systems: A review. Water Res., 41(3), 519-529.

Van Soestbergen, A. A., Lee, C. H., 1969. Pour plates or streak plates? Appl. Microb., 18(6), 1092-1093.

Wu, H., Zhang, J., Mi, Z., Xie, S., Chen, C., Zhang, X., 2015. Biofilm bacterial communities in urban drinking water distribution systems transporting waters with different purification strategies. Appl. Microb. Biotech., 99(4), 1947-1955.