Water Utility Journal 7: 3-11, 2014.

Pre-processing and visualization of biofilm development in drinking water distribution systems

E. Ramos-Martínez [1*], M. Herrera [2], J. Izquierdo [1] and R. Pérez-García [1]
[1] *FluIng-IMM, Universitat Politècnica de València, Spain*
[2] *InfraSense - Dept. of Civil and Environmental Eng., Imperial College London, United Kingdom*
[*] *e-mail: evarama@upv.es*

Abstract: Biofilm develops in drinking water distribution systems (DWDSs) as complex microorganism communities covered by an extracellular polysaccharide which provides them structure, protection against disinfectants and helps retain food. Biofilm poses a serious risk to DWDSs. The presence of substantial and active attached biomass in the inner wall of pipes can protect pathogenic microorganism, lead to the formation of high biocorrosion zones, and consume residual disinfectant. Various studies have been performed in relation to the influence that the characteristics of DWDSs have in biofilm development. Nevertheless, their joint influence has scarcely been studied, due to the complexity of the community and the environment under study. In this paper, an innovative work is carried out with the introduction of pre-processing methodologies as new tools that allow using the knowledge gained on the development of biofilm in DWDSs in a practical and efficient way. We compile currently available information of the physical and hydraulic conditions of the DWDSs that affect biofilm development to study the effect that the joint influence of these characteristics has in biofilm development. This article proposes pre-processing by Machine Learning approaches, preparing a case-study database to get inference results by posterior analyses. This helps to develop a scalable and interesting set of tools to understand biofilm behaviour with respect to their physical and hydraulic environment. Finally, intelligent data visualization techniques are applied to carry out an exploratory analysis of the obtained metadata and to identify possible interesting patterns and groups related to biofilm development in DWDSs.

Keywords: biofilm, pre-processing, visualization, distance based clustering, radial basis function interpolation

1. INTRODUCTION

Biofilm develops in DWDSs as complex communities of microorganisms bound by an extracellular polysaccharide polymer, the glycocalyx, which provides them structure, protection and helps retain food. These communities of organisms form spontaneously in the presence of moisture and bind strongly against the initial repulsion at the inner pipe wall and modify it as they capture more nutrients and new bacteria. Developed biofilm is very strong and poses a serious problem when a clean and disinfected environment is needed. Apart from the health risk, due to the role of biofilm as microbial pathogens reservoir (Momba *et al.*, 2000), biofilm is also responsible for many other DWDSs problems. For example: aesthetical deterioration of water, proliferation of higher organisms, operational problems, biocorrosion (Videla & Herrera, 2005), and consumption of disinfectant (de Beer *et al.*, 1994), among others. Although in most countries regulated quantities of residual disinfectant are present in DWDSs these are not enough to avoid the presence of biofilm in these systems. So, nowadays, biofilm represents a paradigm in the management of water quality in all DWDSs.

Numerous researches have been carried out on biofilm ecology and on how the various characteristics of DWDSs affect them. However, most of the studies have focused on just one or two of these aspects, and almost none has studied their joint influence on biofilm development. This is because, as known, obtaining data in the ecology field is hard and not very productive work.

When we focus on biofilm development in DWDSs the limitation to obtain data increases. Access and sampling in these systems are quite difficult, since in addition, water utility managers usually impose restrictions in sampling and in divulging data. This makes experimental approaches the most common way to get data. However, due to its intricacy, most studies are simplified, focusing just on one aspect related to biofilm development, not taking into account the joint influence of the whole system variables.

This work aims to approach this problem. We have compiled biofilm data from different sources and resorted to Machine Learning approaches to pre-process the data, preparing a case-study meta-database to do inferences by posterior analyses. This develops a scalable and interesting set of tools to understand biofilm behaviour with respect to their environment, thus, increasing the effectiveness of DWDSs management and the quality of the distributed water.

The roadmap of the paper is as follows. Section 2 introduces the background techniques used, and a number of techniques for pre-processing and visualization of metadata are explained. Among them, we will consider processes such as distance based clustering for extreme values detection or radial basis functions to manage missing data. Next, Section 3 presents the obtained results in our experimental study. Conclusions, recommendations, and future challenges close this paper in Section 4.

2. BACKGROUND TECHNIQUES USED

With the aim of identifying the combination of DWDSs' physical and hydraulics conditions that favour high biofilm development, we have resorted to compile the currently available data of various researches. These works have studied some physical and/or hydraulic conditions of DWDSs that affect biofilm development. Finally, we have applied data mining approaches to get a complete and broad meta-database to enable inference tasks by posterior analysis.

To achieve this purpose measurements of biofilm growth in DWDSs have been collected from various literature sources. In this case the sources of information have been scientific papers studying the development of these communities in DWDSs, either in real or in simulated systems in laboratory. Besides measurements of biofilms, information of the associated conditions was also collected. The total number of cases gathered was 303. As expected, this compilation was not achieved in a straightforward manner and we have to take into account that the complexity especially increases when we analyse data provided by very different studies and information sources (Aubrecht *et al*., 2003). Thus, we had to manage heterogeneity in data measurements, multi-scalarity, important presence of missing data, and different codifications, among other drawbacks.

Besides the classical statistical techniques for data cleaning, inductive techniques are an alternative, when analytical/traditional methods fail, are too slow, or simply do not exist. That is why we have considered data mining techniques to approach this issue covering the important part of pre-processing in data analysis (Gibert *et al*., 2008).

2.1. Statistical tools in the pre-processing

Pre-processing ranges from the simplest descriptive techniques to the more sophisticated data analysis methods, depending on the nature of data and the goals of the analysis itself. In general, most of the operations performed in pre-processing steps can be reduced to two main families of techniques, namely, detection techniques and transforming techniques. Detection techniques are those oriented to detect imperfections in data sets or to verify the accomplishment of required assumptions for a particular analysis. Transforming techniques are those oriented to perform transformations in the data set in order to correct the imperfections detected before, or to achieve the necessary technical conditions to apply a certain analysis technique.

2.1.1. Detection techniques: Distance based clustering

Clustering is a popular technique used to assemble similar data points or objects in groups or clusters (Jain & Dubes, 1988). Clustering methods are mainly divided into two categories based on the way they produce results, namely, partitional clustering and hierarchical clustering methods. Partitional clustering methods create a single clustering (flat clustering) of a dataset. Partitional methods can be further categorized into two classes based on the criteria used either distance based and/or density based methods. Distance based methods (Kriegel *et al*., 2011) optimize a global criteria based on the distance between patterns (Patra *et al*., 2011). This makes these methods suitable algorithms to build outlier detection techniques. Most of these techniques rely on the key assumption that normal objects belong to large and dense clusters, while outliers form small-size clusters (Loureiro *et al*., 2004; Niu *et al*., 2007).

In biofilm metadata problems we usually need to manage data sets containing inputs of several types. This requires for a clustering algorithm to be scalable and capable of handling different attribute types. Classical methods are not the answer: for example, PAM (Partitioning Around Medoids) algorithm can handle various attribute types but is not efficient with large datasets. *K*-means algorithms (Hartigan & Wong, 1979; Likas *et al.,* 2003) can handle large data sets but deal with only data sets formed from interval-scaled variables. CLARA (Clustering Large Applications) algorithm is a combination of sampling approach and the PAM algorithm. Instead of finding medoids, each of which is the most representative object in a cluster for the entire data set, CLARA draws a sample from the data set and uses the PAM algorithm to select an optimal set of medoids from the sample (Wei *et al.,* 2003). To alleviate sampling bias, CLARA repeats the sampling and clustering processes multiple times and selects the best set of medoids to define the final clustering.

2.1.2. Transforming techniques: Radial basis functions

The use of radial basis functions (RBFs) as tools for interpolation fits perfectly in the philosophy of working with metadata, restoring automatically the necessary data to achieve the posterior analysis and creation of 'data-driven' models (Herrera *et al*., 2011). Radial basis functions (RBFs) have been developed for the interpolation of scattered multivariate data (Press *et al*., 2007; Wright *et al.,* 2003). The method uses linear combinations of N_{RBF} radially symmetric functions, $h_i(x)$; based on the Euclidean distance or other different metric, to approximate response functions such as Equation 1 shows.

$$f_p(x) = \sum_{i=1}^{N_{RBF}} w_i h_i(x) + \varepsilon_i \quad (1)$$

where w_i represents the coefficients of the linear combinations, $h_i(x)$ the radial basis functions, and ε_i are independent errors with variance σ^2. A typical radial function is the Gaussian function (with centre c and radio δ), as is expressed in (Eq.2).

$$h_i(x) = \exp\left(-\frac{(x-c)^2}{\delta^2}\right) \quad (2)$$

The basic idea behind RBFs can be generalized to consider alternative loss functions and basis functions, in a scheme known as kernel-based regression (Buhmann, 2003).

2.2. Visualization

There are numerous techniques that can be used to visualize multidimensional data. In this case, we have chosen *scatterplot* and *RadViz*.

- Scatterplot is a very useful tool for data analysis. Let us suppose that (x_i, y_i), $i = 1, ..., n$, are paired measurements of two variables, x and y. A scatterplot of y_i against x_i gives plenty of information about the amount of association between *x* and *y*, the dependence of *y* on *x* if *y* is a response and *x* is a factor, clusters of points, outliers, and a host of other things (Cleveland *et al.,* 1984).

- RadViz: The Radial Coordinate visualization is a neat non-linear multi-dimensional visualization technique in which *n*-dimensional data points are laid out as points equally spaced around the perimeter of a circle. One of the ends of *n* springs is anchored to these *n* perimeter points. The other ends of the springs are attached to a data point. The spring constant K_i equals the value of the *i*-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. All the data point values are usually normalized to have values between 0 and 1 (Nováková *et al*., 2009).

Finding interesting projections can be a difficult and time consuming task for the analyst, since the number of possible projections increases exponentially with the number of concurrently visualized attributes (Leban *et al.,* 2006). For this reason, VizRank algorithm is applied to the results of the previous analyses. VizRank, is able to automatically rank visual projections of classified data by their success in showing different class values well separated (Demsar *et al.,* 2005). Projections that provide perfect class separation (there is no overlap between classes) receive value 100, while less informative projections receive correspondingly lower values (Leban *et al*., 2006). After chosen it, in the case of the RadViz visualization, FreeViz algorithm is also implemented. This algorithm, using optimization procedures, chooses the projection that best separates instances of different classes relaxing the constraints of placement of feature anchors.

3. EXPERIMENTAL STUDY

In our case, we compiled biofilm development data regarding physical and hydraulic characteristics of DWDSs from previous research works. Due to the difficulty associated with the study of biofilm in DWDSs the data obtained was incomplete and not easy to compare. With the aim of having a broad and complete database, which provides robustness to posterior analyses, a pre-processing methodology was applied. After reaching our objective successfully, various techniques of intelligent data visualization were used to make an exploratory analysis of the metadata obtained.

Firstly, we identified the physical and hydraulic characteristics of DWDSs that have been studied individually and are known to influence biofilm development. These are: pipe material (metallic, plastic or cement), roughness, hydraulic regime, flow velocity, and hydraulic retention time (HRT).

In the process of data collection we found that almost there were no measurements of pipe roughness in relation to biofilm development. Knowing that the accumulation of corrosion products and dissolved substances can increase the roughness of the pipes (Christensen, 2009), and the fact that older deposits may have greater biomass and contain more bacteria (Chowdhury, 2011), we decided to use the pipe age as a variable instead of using pipe roughness.

In the case of HRT measurements we found a similar problem. This time, in order to increase the number of data available for the pre-processing analysis, we created a synthetic index, called "water age". We used the HRT (h) and the distance to the point of chlorination (km) since they increase with the age of the water in the system. With the aim of normalization, each variable, HRT and distance to the point chlorination, was scaled. The minimum value was subtracted from the current value and divided by the difference between the maximum and the minimum values. We wanted to merge two variables into one. In order not to bias the study we used the inverse proportion existing in the original data. HRT was multiplied by a factor of 0.3, while the distance to the disinfection point is weighted with a factor of 0.7: since there was 2.5 times more data of HRT than data of distances to the disinfection point, HRT data was multiplied by a factor nearly 2.5 times smaller than the factor that multiplied the distance to the disinfection point. Accordingly, the two variables had a comparable influence on the index generation. Finally a re-scale is done. So, 'water age' is an index between 0 and 1, which increases with the age of the water. Values close to one correspond to older water.

It relation to biofilm measurements it is worth to say that the Heterotrophic Plate Count (HPC/cm^2) was chosen as biofilm quantification method because it is the most widely used, being more available data measured using it. This enables us to have a much larger number of data and, thus increasing the reliability of our results. Finally, the studied attributes were pipe material, pipe age, hydraulic regime, flow velocity, water age and biofilm amount.

Applying the pre-processing techniques to the 303 cases with incomplete data studied, we managed to complete and make useful for our purpose 210 cases.

Once having completed the database we proceeded to make a discretization based on bibliography and expert knowledge. It is worth to note that all the pre-processing tasks were performed with continuous data in all the attributes for which it was possible. That way, on the one hand, we use all the information available to the pre-processing, not losing information, and on the other, we reduce the uncertainty associated with the pre-processing. The resulting categories, after discretization, are shown in Table 1.

Table 1. Variables and categories.

P.Material (years)	**P.Age (years)**	**Biofilm (HPC/cm2)**	**Flow Velocity (m/s)**	**Water Age**	**Hydraulic Regime**
Metallic (M)	Old (O) [≥ 31]	High (H) [$\geq 10^7$]	High (H) [1.8-3.5]	High (H) [0.7-1]	Laminar (L)
Cement (C)	Medium (M) [11-30]	Medium (M) [10^4-10^6]	Medium (M) [0.8-1.7]	Medium (M) [0.4-0.6]	Turbulent (T)
Plastic (P)	Young (Y) [0-10]	Low (L) [0-10^3]	Low (L) [0-0.7]	Low (L) [0-0.3]	-

Next, we proceeded to make an exploratory analysis of the metadata obtained. We decided to use tools of intelligent data visualization because of their usefulness to identifying patterns and clusters in multidimensional data. We focused on two types of techniques, namely RadViz and scatterplot.

The RadViz visualization made it possible to map our 5-dimensional data onto a plane (Figure 1). In our case, biofilm is treated as a hidden attribute and the other attributes are anchored on the boundary of the circle. The green dots represent the cases with low biofilm development, the yellow dots the ones with medium development, and the red ones are those with high biofilm development. Blurred zones represent the influence areas. Then, the VizRank algorithm was applied to the obtained results. The best projection (Figure 1) had a score of 75.31, and the number of attributes was reduced to 4. Pipe material variable was thus removed. This fact may be interpreted in the sense that the evolution of the pipe material is partially represented through the pipe age. It can be said that, in most DWDSs, metal pipes tend to be the oldest and plastic ones the newest. Anyway, the second best configuration maintains all the variables, including the pipe material, and has a score of 75.27, very close to the best one. This high score indicates that the chosen physical and hydraulic characteristics of DWDSs, as well as the pre-processing carried out, are appropriate as they allow a fairly clear distinction between the different groups formed in relation to biofilm development.

Finally, to improve as much as possible the visualization of the obtained graph, the FreeViz algorithm was used (Figure 1). After its implementation, an improvement in the clustering view is observed. Although there is some overlap, a gradual change from low biofilm development to medium development and from this to high biofilm development can be clearly observed (Figure 1 right). Most of the cases are located in the area corresponding to medium biofilm development (yellow area). High biofilm development (red area) seems to be mostly influenced by flow velocity. All the cases in this area have turbulent hydraulic regime, low water age, medium pipe age and low flow velocity, except one that has medium flow velocity. In the same way, the big zone of low biofilm development (green area) seems to be mainly influenced by the pipe age variable. In this case, all the instances located in this area correspond with young pipes that, according to the bibliography, tend to develop less biofilm than older pipes (Figure 1, right).

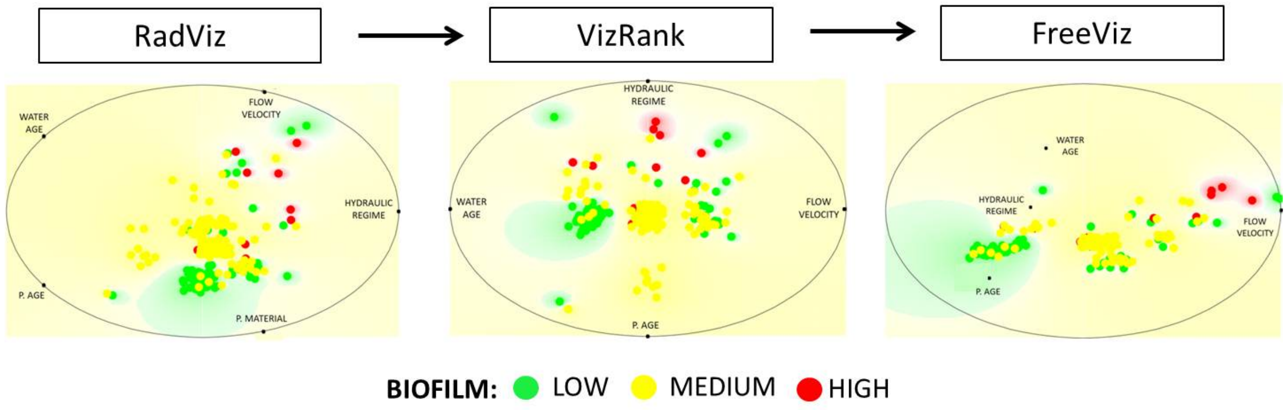

Figure 1. Resulting RadViz, VizRank And FreeViz graphs

We also used the VizRank optimizing algorithm in the scatterplot visualizations. The results with highest scores are shown in Figure 2.

In the scatterplot (1, 1) we find, as expected, that the probability of high biofilm development is higher when the pipe is medium age or old (Chowdhury, 2011). Another interesting fact to highlight is that when the pipe is young and the flow velocity is high the probability of low biofilm development is very high. We would expect, as the theory says, an increase in biofilm development with high flow velocity as seems to occur when the pipe age is medium or high. This may be explained because pipes increase their roughness and deposits content with age and rough surfaces have greater potential for biofilm development (Chowdury, 2011). Bigger roughness and deposits provide more colonization surface and refuge to biofilm (Lehtola *et al.,* 2006), protecting it from disinfectant and water drag forces. So, biofilm growing in new pipes may easily detach as water drag forces increase with the flow velocity.

In the scatterplot (1, 2), something similar is found. As said before, more biofilm development is expected when flow velocity increases (Lehtola *et al.,* 2006). Yes, this does not happen in the case of plastic pipes. This phenomenon may be explained by the roughness of the materials. The roughness of plastics is very low and biofilm may not be able to attach strongly to the inner pipe surface and, as a result, it is easily detached with high flow velocities. We also see that metallic pipes, as is found in bibliography (Niquette *et al*., 2000), tend to have higher biofilm development than non-metallic pipes. In our case, almost no cases are found with low biofilm development in the case of the metallic pipes.

High probability of low biofilm development is also found in the scatterplot (2, 1) when flow velocity is high and the hydraulic regime is turbulent. This does not happen when the flow is laminar. This can be explained by the fact that turbulence regimes induce highest shear forces than laminar regimes thus favouring biofilm detachment. It is also shown that no cases with high biofilm development are found in laminar regime. This agrees with the literature, since biofilm in turbulent flow tends to have increased cell density than biofilms in laminar flow (Simoes *et al.,* 2007). Anyway, it must be noticed that the number of cases that correspond with laminar hydraulic regime are much less than the cases corresponding to turbulent flow. So we have to warn that in this regard the results may be biased.

The last scatterplot (2, 2) shows that, contrarily to what is expected according to the bibliography (EPA, 2002), the cases with highest biofilm development are found when the water age is low, especially when the flow velocity is low too. In this case, although the probability of low biofilm development is still high when the flow velocity is high, cases with high biofilm development and high flow velocity are found.

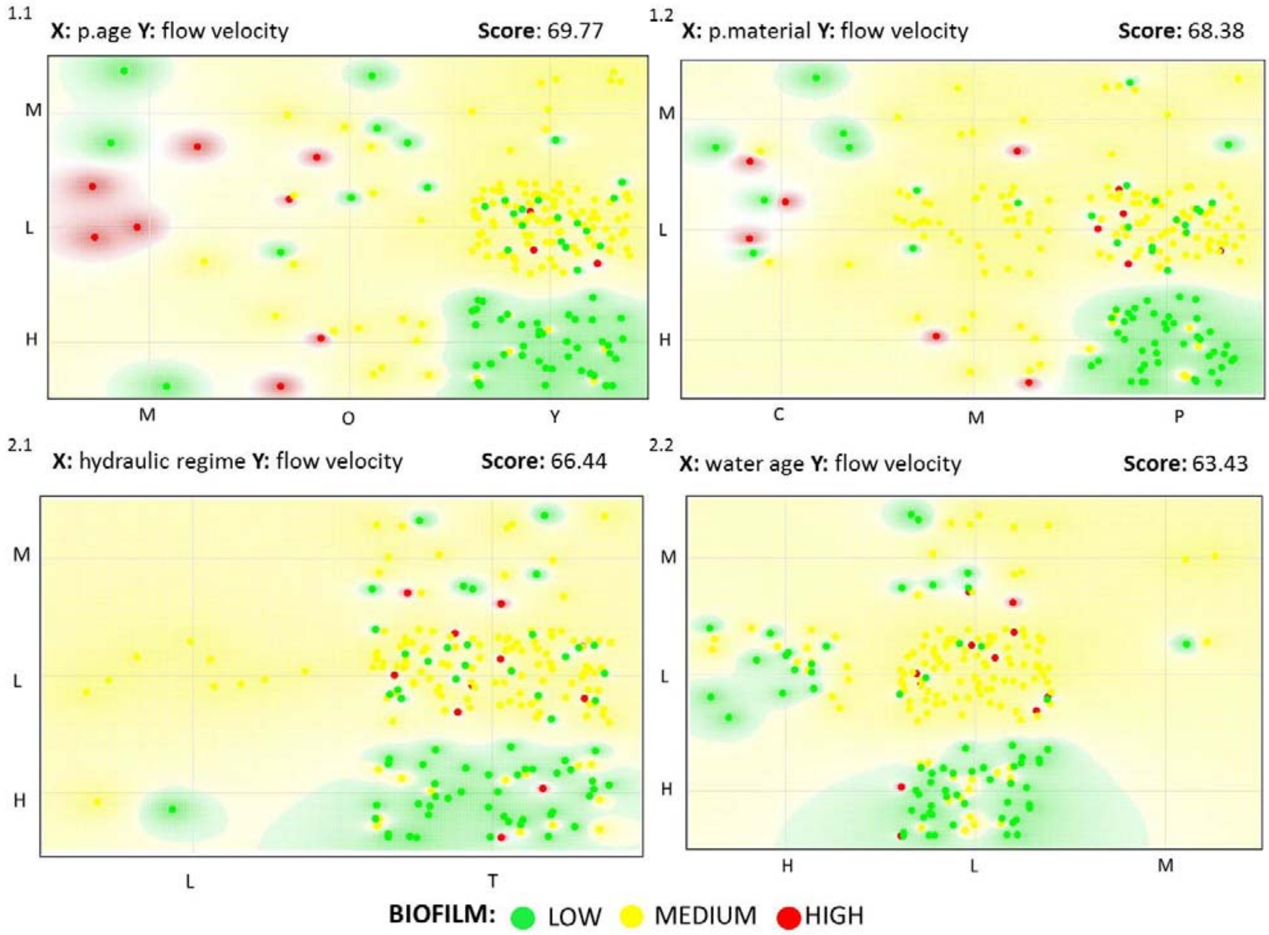

Figure 2. VizRank application to the scatterplot visualization

Finally, it is worth to enhance that the flow velocity attribute is found in all the scatterplots with the highest VizRank scores. It seems to have special influence on biofilm development, specially, when the flow velocity is high. It may mean that biofilm development is very sensible to flow velocity.

All the analyses were made with Orange Canvas 2.0b software (Demsar *et al*., 2004).

4. DISCUSSION

This study provides an overview of an innovative work with the introduction of pre-processing methodologies as new tools that allow using the level of knowledge gained on the development of biofilm in DWDSs in a practical and efficient way. Pre-processing methodologies can become very important tools in the ecology field in the future. They enable efficient use of the knowledge reached in the research field, are not time consuming, and cope with scarce and incomplete data normally found in the ecology field. In addition, this methodology does not require highly qualified equipment, and allows better implementation and interpretation, as well as easier application and better understanding of the biofilm processes and interactions that occur in DWDSs.

Besides, to the knowledge of the authors, most of the research carried out in DWDSs biofilm development, due to the great difficulty of study of the whole system effect, have focused on the effect of just one or two attributes. On the contrary, the methodology carried out in this paper has allowed us to provide an approach to study and understand the combined effect of all the physical and hydraulic characteristics of DWDSs on the biofilm development. Although in the present paper just an exploratory analysis of the metadata has been done some interesting aspects have been found. It is worth to note the different effect of high flow velocity depending on pipe material, pipe age and the hydraulic regime. Observing the amount of information obtained just with an exploratory visualization of the data, it is reasonable to think that a deeper study of the database obtained would provide new information. The data pre-processing methodology presented in this paper has demonstrated to be a very useful tool for the exploratory study of the combined effect of DWDSs physical and hydraulic characteristics on biofilm development.

In the future, apart from performing further analyses on the obtained database, it is also planned to include the physical and chemical characteristics of water in the pre-processing in order to get a more complete database and have a better idea of biofilm development in DWDSs, in general, and in the various areas of a given DWDS, in particular.

REFERENCES

Aubrecht, J., Mikosovski, P. & Kouba, Z., 2003. Metadata Driven Data Preprocessing for Data Mining. J. Pokorny, V. Snasel (Eds.) Technical University of Ostrava; 55-62.

de Beer D., Srinivansa R. & Stewart P. S., 1994. Direct Measurement of chlorine penetration into biofilms during disinfection. Applied Environmental Microbiology; 60(3): 4339.

Buhmann, Martin D. 2003. Radial Basis Functions: Theory and Implementations, Cambridge University Press.

Chowdhury, S.,2011. Heterotrophic bacteria in drinking water distribution system: a review. Environmental Monitoring and Assesment; 2407-2415.

Christensen, R., 2009. Age effects on iron-based pipes in water distribution systems. Master's thesis.

Cleveland, W.S. & McGill, R. 1984.The many faces of a scatterplot. Journal of the American statistical association; 79(338): 807-822.

Demsar, J., Zupan, B., leban, G.,& Tomaz, C. 2004. Orange: from experimental machine learning to interactive data mining. Lecture Notes in Computer Science; 3202: 537-539

Demsar, J. and Leban, G. & Zupan, B. 2005. FreeViz - An Intelligent Visualization Approach for Class-Labeled Multidimensional Data Sets.

EPA, 2002. Effects of water age on distribution system water quality.

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

Hartigan, J. A. & Wong, M. A., 1979. A k-means clustering algorithm. Applied Statistics; 28: 100-108.

Herrera, M., García-Díaz, J.C., Izquierdo, J., Pérez-García, R., 2011. Municipal water demand forecasting: tools for intervention time series, Stochastic Analysis and Applications; 29: 998-1007.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice-Hall.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2006. Data preprocessing for supervised learning. International Journal of Computer Science; 1 (2): 111-117.

Kriegel, H.P., Kröger, P., Sander, J., Zimek, A. 2011. Density-based Clustering. WIREs Data Mining and Knowledge Discovery; 1(3): 231-240.

Kurgan, L., Musilek, P., 2006. A survey of knowledge discovery and data mining process models. Knowledge Engineering Review; 21 (1): 1-24.

Leban, G. , Zupan, B., Vidmar, G. & Bratko, I. 2006. VizRank: Data Visuallization Guided by Machine Learning. Data Mining and knowledge discover; 13: 119-136.

Lehtola, M.J., Laxander, M., Miettinen, I.T., Hirvonen, A., Vartiainen, T. & Martikainen, P.J. 2006. The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system consisting of copper or polyethylene pipes. Water Res.; 40(11): 2151-2160.

Likas, A., Vlassis, N., Vebeek, J. L., 2003. The global k-means clustering algorithm. Pattern Recognition; 28: 451-461.

Loureiro, A., Torgo, L., Soares, C., 2004. Outlier detection using clustering methods: a data cleaning application. In: Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector.

Momba, M.N.B. & Binda M.A. 2002. Combining chlorination and chloramination processes for the inhibition of biofilm formation in drinking surface water system models. Journal of Applied Microbiology; 92: 641-648.

Ndiongue, S., Huck, P.M. & Slawson, R. M. 2005. Effects on temperature and biodegradable organic matter on control of biofilms by free chlorine in a model drinking water distribution system. Water Res.; 39: 953-96.

Niquette, P., Servais, P. & Savoir, R. 2000. Impacts of pipe material on densities of fixed bacterial biomass in a drinking water distribution system. Water Res.; 34(6): 1952-1956.

Niu, K., C. Huang, S., Zhang, Chen, J., 2007. Outlier detection using distance distribution clustering. In: Lecture Notes in Artificial Intelligence (LNAI) 4819. T. Washio et al. (Eds.) - Springer Verlag; 332-343.

Nováková, L. & Stepankova O., 2009. RadViz and identification of cluster in multidimensional data. 13th International Conference Information Visualisation.

Patra, B.K., Nandi S., Viswanathan, P. 2011. A distance based clustering method for arbitrary shaped clusters in large datasets. Pattern Recognition; 44: 2862-2870.

Press, W.H., Teukolsky, S.A.; Vetterling, WT; Flannery, BP, 2007. Radial Basis Function Interpolation", Numerical Recipes: The Art of Scientic Computing (3rd ed.), New York: Cambridge University Press.

Simoes, M., Pereira, M., Vieira, M., 2007. The role of hydrodynamic stress on the phenotypic characteristics of single and binary biofilms of pseudomonas fluorescens. Water Science and Tecnology; 55 (8-9): 437-445.

Videla H. A. & Herrera L. K., 2005. Microbiologically influenced corrosion: looking to the future, International Microbiology; 8: 169-180.

Wright, G.B., 2003. Radial Basis Functions interpolation: Numerical and analytical developments, PhD. thesis

Wei, C., Lee, Y. & Hsu, C., 2003. Empirical comparison of fast partitioning- based clustering algorithms for large data sets. Experts Systems with Applications; 24: 351-363.