Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
AGR2013-0773

Drinking Water Distribution Systems Characteristics on Biofilm Development: A Kernel based Approach

Eva Ramos-Martínez
FluIng-IMM, Universitat Politècnica de València
Spain

Manuel Herrera
BATir, Université libre de Bruxelles
Belgium

Joaquín Izquierdo
FluIng-IMM, Universitat Politècnica de València
Spain

Rafael Pérez-García
FluIng-IMM, Universitat Politècnica de València
Spain

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209
Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

Printed in Athens, Greece by the Athens Institute for Education and Research.

ISSN **2241-2891**
18/12/2013

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Ramos-Martínez, E., Herrera, M., Izquierdo, J. and Pérez-García, R. (2013) **"Drinking Water Distribution Systems Characteristics on Biofilm Development: A Kernel based Approach"** Athens: ATINER'S Conference Paper Series, No: **AGR2013-0773**.

Drinking Water Distribution Systems Characteristics on Biofilm Development: A Kernel based Approach

Eva Ramos-Martínez
FluIng-IMM, Universitat Politècnica de València
Spain

Manuel Herrera
BATir, Université libre de Bruxelles
Belgium

Joaquín Izquierdo
FluIng-IMM, Universitat Politècnica de València
Spain

Rafael Pérez-García
FluIng-IMM, Universitat Politècnica de València
Spain

Abstract

Biofilm develops in drinking water distribution systems (DWDSs) as layers of microorganisms bound by an organic matrix and attached to pipe walls. The presence of substantial and active attached biomass can lead to a decrease in water quality by generating bad tastes and odours, operational problems, biocorrosion, and residual chlorine consume, among others problems. Recently, it has also become evident that biofilm can serve as an environmental reservoir for pathogenic microorganisms, resulting in a potential health risk for humans if left unnoticed. Various studies have been performed in relation to the influence that a number of characteristics of the DWDSs have in biofilm development. Nevertheless, their joint influence, apart from few exceptions, has been scarcely studied, due to the complexity of the community and the environment under study. This research aims to study the effect that the interaction of the physical and hydraulic conditions of the DWDSs has on biofilm development. To achieve this goal we apply Kernel methods for the study of biofilm behaviour. They give a systematic and principled approach to training learning machines. Their accuracy and simplicity to approach complex problems has been a decisive factor when choosing this form of addressing the study of biofilm behaviour in DWDSs. As a result, we claim that deeper understanding of the consequences that the interaction of the relevant hydraulic and physical factors of DWDSs have on biofilm development may be obtained. Thus, the effectiveness of the DWDSs management and the quality of the distributed water would increase.

Keywords: Drinking water quality, biofilm, drinking water distribution systems, Kernel methods, support vector clustering.

Aknowledgment: This work has been supported by project IDAWAS, DPI2009-11591, of the Dirección General de Investigación of the Ministerio de Ciencia e Innovación of Spain, and ACOMP/2011/188 of the Conselleria d'Educació of the Generalitat Valenciana.

Corresponding Author:

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 to 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 modified it as capture more nutrients and new bacteria. A developed biofilm is very strong and a problem when a clean and disinfected environment is needed. Apart from the health risk that biofilm in DWDS may be, due to its role as microbial pathogens reservoir (Momba *et al*., 2000), biofilm is responsible for many other DWDSs problems. For example: aesthetical deterioration of the water, proliferation of higher organisms, operational problems, biocorrosion (Videla and Herrera, 2005), consumption of disinfectant (de Beer *et al*., 1994), among others. Although in most countries regulated quantities of residual disinfectant are present in the 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 DWDS.

Survival and regrowth of microorganisms in DWDSs is affected by biological aspects, as well as the interaction of various other factors (Yu *et al.*, 2010). Numerous studies have examined the influence that a number of characteristics of DWDSs have in biofilm development (Silhan *et al.*, 2006; Zhou *et al.*, 2009). Nevertheless, their joint influence, apart from a few exceptions (Simoes *et al.*, 2006), has been scarcely studied, due to the complexity of the community and the environment under study. Nonetheless, by compiling data from different studies and using machine learning techniques we have generated a comprehensive and extensive enough database to do inference by posterior analysis (Ramos-Martínez *et al.*, 2013). This paper, particularly, focuses on studying the effect that the interaction of the relevant hydraulic and physical characteristics of DWDSs has on biofilm development resorting to Kernel methods. They give a systematic and principled approach to training learning machines and are able to manage different types of data, detect data out of range, and achieve a good generalisation performance. Their accuracy and simplicity to approach complex problems has been a decisive factor when choosing this form of addressing the study of biofilm behaviour in DWDSs. As a result, we are going to achieve a deeper understanding of the consequences that the interaction of the relevant hydraulic and physical factors of DWDSs has on biofilm development. Thus, the effectiveness of the DWDSs management and the quality of the distributed water would increase.

The roadmap of the paper is as follows. Section 2 introduces the material and methods use in this paper, introducing the study case and Kernel methods. Next Section 3 presents the results and discussion obtained after applying these methods in our database. Conclusions and recommendations close this paper in Section 4.

Material and Methods

This paper aims to study the joint influence that the relevant hydraulic and physical characteristics of DWDSs have on biofilm development. To carry out this work, we gathered biofilm data from experts, preprocessed the data using machine learning approaches, and prepared a case-study database. This compilation was not at all straightforward. Data provided by different studies and information sources was ambiguous, difficult to compare, and incomplete. Among other difficulties, we had to manage heterogeneity in data measurements, multiscalarity, important missing data, and differing codifications. Thus, the data was preprocessed to generate a complete database. This preprocessing involved a detection stage, where outliers were found and removed using clustering techniques, and a transformation stage, where lost data was reconstructed by suitable imputation using mainly artificial neural networks (Ramos-Martínez *et al.*, 2013). This resulted in a database of 210 complete cases. The variables of the database were found relevant to biofilm development in DWDSs when individually studied by various researchers. The variables used in this study are described below.

(i) Flow velocity. The nutrient mass transfer increases with flow velocity as this favours biofilm development (Lehtola *et al.*, 2006). Nevertheless, specific velocities of between 3-4 m/s may favour its release (Cloete *et al.*, 2003).

(ii) Hydraulic regime. This may be turbulent or laminar (Table 1). Biofilm is likely to be more active in turbulent flow, having more mass per cm^2, increased cell density, and distinct morphology, than biofilm in laminar flow (Simoes *et al.*, 2007).

(iii) Pipe material. Pipe material may be metal, plastic, or cement (Table 1). In general, metal pipes tend to develop more biofilm than cement pipes, and these more than the plastic pipes (Niquette *et al.*, 2000). This is because pipes with a rough surface have greater potential for biofilm growth (Chowdhury, 2011). Rough surfaces provide more area for biofilm growth and protect growth from water shear forces.

(iv) Pipe age. The accumulation of corrosion and dissolved substances in older pipes can increase their roughness (Christensen, 2009) thus favouring biofilm development. In addition, older deposits may have greater biomass and bacteria content (Chowdhury, 2011). We divide the pipes into young, medium, and old (Table 1).

(v) Water age. The longer the time that the water is in the system, the greater the residual disinfectant decay, sediment deposition, and temperature increase (EPA, 2002). All are factors that favour biofilm development. So, 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 the age of the water in the system increases too. With the aim to normalize them, 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 are those with older water.

(vi) Biofilm. We chose the heterotrophic plate count (HPC/cm^2) as the biofilm quantification method. Although there are other methods, this is the most commonly used, and so more data is available. Based on the observed biofilm data distribution and expert criteria, this data was divided into low, medium and high biofilm development (Table 1).

Table 1. *Variables and categories of the database*

P.MATERIAL	P.AGE (years)
metallic	high [≥31]
cement	medium [11-30]
plastic	low [0-10]
BIOFILM (HPC/cm^2)	FLOW VELOCITY (m/s)
high [≥10^7]	high [1.8-3.5]
medium [10^4-10^6]	medium [0.8-1.7]
low[0-10^3]	low[0-0.7]
HYDRAULIC REGIME	WATER AGE
laminar	high [0.7-1]
turbulent	medium [0.4-0.6]
-	low[0-0.3]

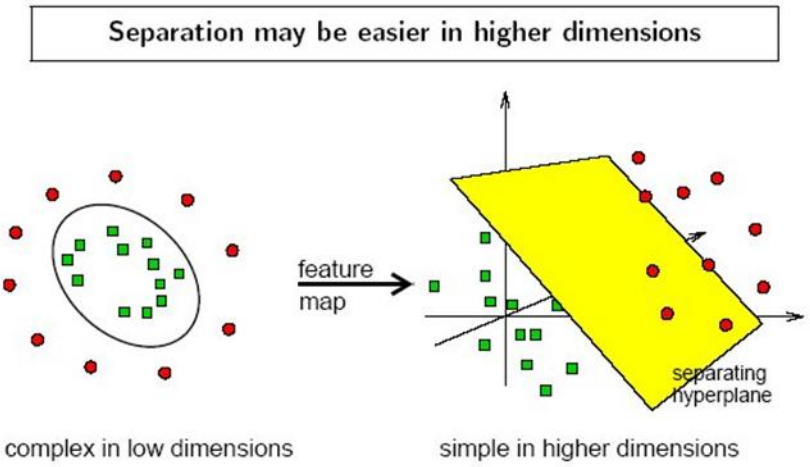
So, the variables of the database can be divided into two continuous inputs: *flow velocity* and *water age*, and three categorical inputs: *hydraulic regime*, *pipe material*, and *pipe age*. After that, with the aim to find the best way to perform our analysis we discretize the continuous variables of the database, to normalize the database. So, finally, we have two databases, one with mixed variables, continuous and discrete, which conserve all the available information and the other that has been normalized discretising all the variables. The resulting variables and categories of the databases are shown in Table 1.

Kernel Methods

The corresponding analyses are going to be carried out with Kernel methods. Their accuracy and simplicity to approach complex problems has been a decisive factor when choosing this form of addressing the study of biofilm behaviour in DWDSs. Kernel-based learning methods are a class of algorithms for pattern analysis (Shawe-Taylor and Cristianini, 2006). They provide a powerful way of detecting nonlinear relations using well-understood linear algorithms in an appropriate space. Thus, each problem is approached by mapping the data into (a high dimensional) feature space, defined by a *kernel* *function* such as in (Scholkopf and Smola, 2002; Karatzoglou, 2006; Hofmann et al. 2008). The benefit of this process is that for nonlinear feature maps we are able to produce nonlinear learning functions based on a linear approach. Furthermore, these methods enable us to work by inner products. This is computationally simpler than explicitly working in the feature space. The function *K* is called the *kernel function*, and the above described approach is the so-called *kernel trick*.

Support Vector Machines (SVMs) work with kernel methods (Figure 1). Training a SVM for classification, regression or novelty detection involves solving a quadratic optimization problem. Using a standard quadratic problem solver for training an SVM would involve solving a big QP problem even for a moderate sized data set.

Figure 1. *Support Vector Machines*


Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

In classification, SVMs separate the different classes of data by a hyperplane. It can be shown that the optimal, in terms of classification performance, hyperplane (Vapnik 1998) is the one with the maximal margin of separation between the two classes. It can be constructed by solving a constrained quadratic optimization problem whose solution in terms of a subset of training patterns that lie on the margin. Several approaches have been suggested in the case of more than two classes in the output. Nevertheless, two are the most popular among them: (1) “one against many” (*o-a-m*) where each category is split out and all of the other categories are merged; and, (2) “one against one” (*o-a-o*) where *k*(*k*-1)/2 models are constructed where k is the number of categories (Hsu and Lin, 2002).

Results and Discussion

The database under study is formed by 210 pipes with their corresponding physical and hydraulic characteristics. The output is the quantity of *biofilm* in these pipes that is measured into three categories: *High*, *Medium*, and *Low*.

Thus, we work with *o-a-o* methods to expand the classical SVM searching a classification by SVM methods should be adapted to these three possible outputs.

The different kernels compared in this work are shown in Table 2. In Table 3 we can see their performances respect to the nature of the database proposed. All of them are results of using a third part of the database for test and the other two parts are for training and validating. The SVM parameters: *C* and *gamma*, are tuned by a grid-search algorithm between the limits [1, 100] and [0,1], respectively. The best values are *C* = 10 and *gamma* = 0.1 for the mixed database and *C* = 1, *gamma* = 0.1 for the discretized one.

Table 2. *Most popular kernels chosen for the comparisons*

Kernel	Expresion
RBF	$k(x,y) = \exp\left(-\frac{\|\|x-y\|\|^2}{2\sigma^2}\right)$
Linear	$k(x,y) = x^T y + c$
Polynomial	$k(x,y) = (\alpha x^T y + c)^d$
Sigmoid	$k(x,y) = \tanh\,(\alpha x^T y + c)$

In this Table 3, *Diag.* calculates the percentage of data points in the main diagonal of each confusion matrix for each test and *Kappa* is a correction of *Diag*. that represents an index of agreement (between estimated and observed classification) for each solution. Kappa determines to what extent the observed agreement is higher than the one that it is expected by chance (Landis and Koch, 1977). According to them, the best result is obtained with the RBF function in the case of the discrete database and with the linear function in the case of the mixed database. If we compare the results of the both databases we realize that the mixed database obtains a better performance (Table 3). We observe that working with a discretized database (by the continuous input conversion into categorical variables) does not improve in this case the accuracy of the classification.

Table 3. *Kernel methods performance respect to the nature of the database*

Kernel methods		RBF	Linear	Polynomial	Sigmoid
Mixed database	Diag.	0.8000	**0.8428**	0.5857	0.7714
	Kappa	0.5654	**0.6819**	0.2042	0.4966
Discretized database	Diag.	**0.7571**	0.7571	0.6142	0.6857
	Kappa	**0.5339**	0.5128	0.2837	0.3596

Looking the results we resort to one of the last challenges in kernel methods, working with multiple-kernel in case to deal with multiple type of data too (Gonen and Alpaydin, 2011). Multiple-kernel in SVM is ideally adapted to the problem of data integration as it enables distinct types of data to be converted into a common usable format by a weighted combination of as many different kernels as data types are in the database. When applying multiple-kernel (Table 4) we see that obtains better results than in the others cases (Figure 2). This is because it uses a suitable combination of kernels for the mixed nature of the database: an RBF kernel for the continuous input and a Linear kernel for the categorical one. The confusion matrix for the test dataset approached by the multiple-kernel is represented in the Table 5. The accuracy on the high biofilm prediction is significant for the aim of our study. Being able to identify the most problematic pipes, the ones than accordingly to their hydraulic and physical characteristics are more prone to support an elevate biofilm development, can be a meaningful improve in the DWDSs management. Knowing these pipes of the network the efforts (flushing, chlorination, etc.) can be directed to reduce biofilm formation and mitigate, in this way, more efficiently the problems associated with it in DWDSs.

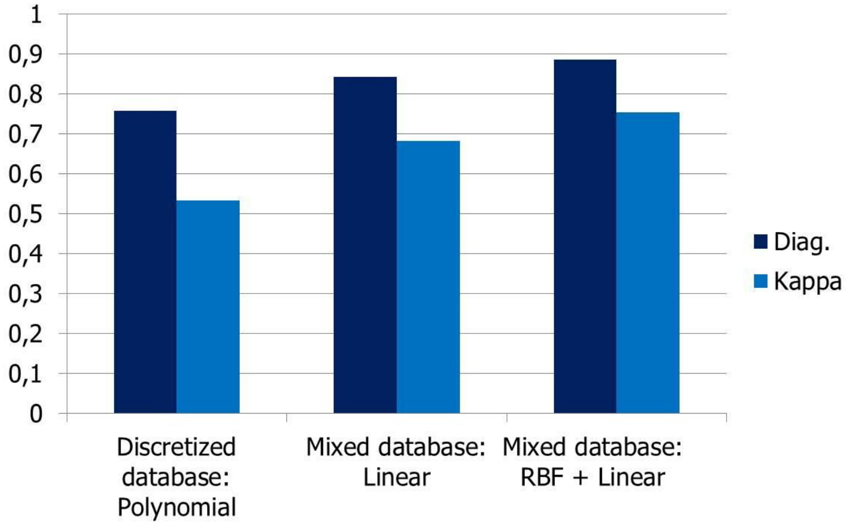
Table 4. *Multiple-kernel performance*

Kernel methods		Multiple
Mixed database	Diag.	0.8857
	Kappa	0.7548

Table 5. *Confusion matrix for the test dataset approached by the multiple-kernel*

Pred \ Observed	H	L	M
H	2	0	1
L	0	17	2
M	1	4	43

Figure 2. *The obtained improvement in the performance*



Conclusions

This study provides an overview of an innovative work that uses Kernel methods as interesting tools that enable the use of knowledge gained in the development of biofilm in DWDSs in a practical and efficient manner. In addition, the effect that the interaction among the hydraulic and physical characteristics of the DWDSs, relevant in biofilm development, has been introduced in this proposal.

The complexity of the community and the environment under study is the reason why there is a lack of works that study the influence that the whole set of characteristics of the DWDSs has on biofilm development. We have chosen Kernel methods to approach this problem because their ability to collect information in an efficient and appropriate way, besides, the fact that its adaptation is simple, in contrast with other machine learning methods. Multiple-kernel has demonstrated to be the best approach to our aim. The combination of linear and RBF methods allows to use all the available knowledge, without losing information discretizing data.

The present work represents the base of a more complex tool of decision support system in DWDSs. It is an advance in the study of biofilm development in DWDSs as it is the starting point to foreseeing the degree of these communities development in the DWDSs' pipes.

References

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

Chowdhury, S. (2011). Heterotrophic bacteria in drinking water distribution system: a review, *Environmental Monitoring and Assesment*, 2407–2415.

Christensen, R.T. (2009). Age Effects on Iron-Based Pipes in Water Distribution Systems, Utah State University.

Cloete, T.E. and Westard, D. and van Vuuren, S.J. (2003). Dynamic response of biofilm to pipe surface and fluid velocity, *Water Science and Technology* 45, 57–59.

Gonen, M., Alpaydin, E. (2011). Multiple kernel learning algorithms*, Journal of Machine Learning Research* 12, 2211-2268.

Hofmann, T., Scholkopf, B., Smola, A. (2008). Kernel methods in Machine Learning. *Annals of Statistics 36* (3), 1171–1220.

Hsu, C.-W., Lin, C.-J. (2002). A comparison of methods for muti-class support vector machines, *IEE Transactions on Neural Networks* 13, 415-425.

Karatzoglou, A. (2006) Kernel methods software, algorithms and applications. Ph.D. thesis.

Landis J. and Koch, G. (1977). The measurement of observed agreement for categorical data, *Biometrics* 33 pp.159–174.

Lehtola, M.J. and Laxandera, M. and Miettinena, I.T. and Hirvonec, A. and Vartiainenb, T. and Martikainenc, P.J. (2006). The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system, *Water Research* 40, 2151–2160.

Momba, M.N.B., Kfir, R., Venter, S.N. & Cloete, T.E. (2000). Overview of biofilm formation in distribution systems and its impact on the deterioration of water quality. *Water SA*, Vol. 26(1), pp 59-66.

Niquette, P. M. and Servais, P. and Savoir, R. (2000). The role of hydrodynamic stress on the phenotypic characteristics of single and binary biofilms of Pseudomonas fluorescens, *Water Resources* 64.

Ramos-Martínez, E., Herrera, M, Izquierdo, J., Pérez-García, R. (2013). Pre-processing meta-data on biofilm development in drinking water distribution systems, *Hydroinformatics*, submitted.

Scholkopf, B., Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Shawe-Taylor, J., Cristianini, N. (2006). *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Silhan, J. and Corfitzen, C.B. and Albrechtsen H.J. (2006). Effect of temperature and pipe material on biofilm formation and survival of Escherichia coli in used drinking water pipes: a laboratory-based study, *Water Science and Technology* 54, 49–56.

Simoes, M. and Simoes, L.C. and Machado, I. and Pereira, M.O. and Vieira, M.J., (2006), Control of flow-generated biofilms with surfactants evidence of Resistance and Recovery, *Food and Bioproducts Processing* 84, 338–345.

Simoes, M. and Pereira, M.O. and Vieira, M.J. (2007). The role of hydrodynamic stress on the phenotypic characteristics of single and binary biofilms of Pseudomonas fluorescens, *Water Science and Technology* 55, 437–445.

United States Environmental Protection Agency, (2002). *Effects of water age on distribution system water quality*. Paper Issue.

Vapnik, V. N. (1998). The nature of statistical learning theory. Spinger-Verlag.

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

Yu, J. and Kin, D. and Lee, T. (2010). Microbial diversity in biofilms on water distribution pipes of different material, *Water Science and Technology* 61, 163–171.

Zhou, L.L. and Zhang, Y.L. and Li, G.B. (2009). Effect of pipe material and low level disinfectants on biofilm development in a simulated drinking water distribution system, *Zhejiang University* 10, 725–731.