

UNIVERSITAT POLITÈCNICA DE VALÈNCIA



DOCTORAL THESIS

Assessing Biofilm Development in Drinking Water Distribution Systems by Machine Learning Methods

Supervisors:

Prof. Dr. Rafael PÉREZ GARCÍA

Prof. Dr. Joaquín IZQUIERDO SEBASTIÁN

Universitat Politècnica de València

Dr. Manuel HERRERA FERNÁNDEZ

University of Bath

Author:

Eva RAMOS MARTÍNEZ

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Water and Environmental Engineering

in the

FluIng Multidisciplinary Research Group

Institute for Multidisciplinary Mathematics

Department of Hydraulic and Environmental Engineering

18th April 2016

Declaration of Authorship

I, Eva RAMOS MARTÍNEZ, declare that this thesis titled, 'Assessing Biofilm Development in Drinking Water Distribution Systems by Machine Learning Methods' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“It’s not the critic who counts. It’s not the person who points out how the person who’s actually doing things is doing them wrong or messing up. It’s the person who’s actually trying to get things done, even when there are obstacles in the way.”

Teddy Roosevelt

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Abstract

FluIng Multidisciplinary Research Group

Institute for Multidisciplinary Mathematics

Department of Hydraulic and Environmental Engineering

Doctor of Philosophy

in

Water and Environmental Engineering

**Assessing Biofilm Development in Drinking Water Distribution Systems by
Machine Learning Methods**

by

Eva RAMOS MARTÍNEZ

Abstract

One of the main challenges of drinking water utilities is to ensure high quality supply, in particular, in chemical and microbiological terms. However, biofilms invariably develop in all drinking water distribution systems (DWDSs), despite the presence of residual disinfectant. As a result, water utilities are not able to ensure total bacteriological control. Currently biofilms represent a real paradigm in water quality management for all DWDSs. Biofilms are complex communities of microorganisms attached to a surface and to each other by an extracellular polymer that provides them with structure, protection from toxics and helps retain food. Besides the health risk that biofilms involve, due to their role as a pathogen shelter, a number of additional problems associated with biofilm development in DWDSs can be identified. Among others, aesthetic deterioration of water, biocorrosion and disinfectant decay are universally recognized. A large amount of research has been conducted on this field since the earliest 80's. However, due to the complex environment and the community studied most of the studies have been carried out under simplified conditions. We refer the research previously done while acquiring new knowledge on biofilm growth in DWDSs to change the common approaches of these studies. Our proposal is based on arduous preprocessing and posterior analysis of the physico-chemical and microbiological data by Machine Learning approaches. A multi-disciplinary procedure is undertaken, helping as a practical approach to develop a decision-making tool to inform DWDS management to maintain minimum biofilm growth and mitigating its negative effects on water supply. A methodology to detect the more susceptible areas to biofilm development in DWDSs is proposed. Knowing the location of these hot-spots in the network will aid specific and localised mitigation strategies, thus saving resources and money. Also, prevention programs could be developed, to allow for acting before the consequences of biofilm are noticed by the consumers. Consequently, the economic cost would be reduced and the service quality would improve, increasing consumers' satisfaction.

Resumen

Uno de los principales objetivos de las empresas encargadas de la gestión de los sistemas de distribución de agua potable (DWDSs, del inglés Drinking Water Distribution Systems) es asegurar una alta calidad del agua en su abastecimiento, tanto química como microbiológica. Sin embargo, la existencia de biofilms en todos ellos, a pesar de la presencia de desinfectante residual, hace que no se pueda asegurar un control bacteriológico total, por lo que, hoy en día, los biofilms representan un paradigma en la gestión de la calidad del agua en los DWDSs. Los biofilms son comunidades complejas de microorganismos recubiertas de un polímero extracelular que les da estructura y les ayuda a retener el alimento y a protegerse de agentes tóxicos. Además del riesgo sanitario que suponen por su papel como refugio de patógenos, existen muchos otros problemas asociados al desarrollo de biofilms en los DWDSs, como deterioro estético del agua, biocorrosión y consumo de desinfectante, entre otros. Una gran cantidad de investigaciones se han realizado en este campo desde los primeros años 80. Sin embargo, debido a la complejidad del entorno y la comunidad estudiada la mayoría de estos estudios se han llevado a cabo bajo ciertas simplificaciones. En nuestro caso, recurrimos a estos trabajos ya realizados y al conocimiento adquirido sobre el desarrollo del biofilm en los DWDSs para cambiar el enfoque en el que normalmente se enmarcan estos estudios. Nuestra propuesta se basa en un intenso pre-proceso y posterior análisis con técnicas de aprendizaje automático. Se implementa un proceso multidisciplinar que ayuda a la realización de un enfoque práctico para el desarrollo de una herramienta de ayuda a la toma de decisiones que ayude a la gestión de los DWDSs, manteniendo, en lo posible, el biofilm en los niveles más bajos, y mitigando sus efectos negativos sobre el servicio de agua. Se propone una metodología para detectar las áreas más susceptibles al desarrollo del biofilm en los DWDSs. Conocer la ubicación de estos puntos calientes de biofilm en la red permitirá llevar a cabo acciones de mitigación de manera localizada, ahorrando recursos y dinero, y asimismo, podrán desarrollarse programas de prevención, actuando antes de que las consecuencias derivadas del desarrollo de biofilm sean percibidas por

los consumidores. De esta manera, el coste económico se verá reducido y la calidad del servicio mejorará, aumentando, finalmente, la satisfacción de los usuarios.

Resum

Un dels principals reptes dels serveis d'aigua potable és garantir el subministrament d'alta qualitat, en particular, en termes químics i microbiològics. No obstant això, els biofilms desenvolupen invariablement en tots els sistemes de distribució d'aigua potable (DWDSs, de l'anglès, Drinking Water Distribution Systems), tot i la presència de desinfectant residual. Com a resultat, les empreses d'aigua no són capaces de garantir un control bacteriològic total. Actualment el biofilms representen un veritable paradigma en la gestió de la qualitat de l'aigua per a tots les DWDSs. Els biofilms són comunitats complexes de microorganismes vinculats per un polímer extracel·lular que els proporciona estructura, protecció contra els tòxics i ajuda a retenir els aliments. A més del risc de salut que impliquen els biofilms, com a causa del seu paper com a refugi de patògens, una sèrie de problemes addicionals associats amb el desenvolupament del biofilm en els DWDSs pot ser identificat. Entre altres, deteriorament estètic d'aigua, biocorrosió i decadència de desinfectant són universalment reconeguts. Una gran quantitat d'investigació s'ha realitzat en aquest camp des dels primers anys de la dècada del 80. No obstant això, a causa de la complexitat de l'entorn i la comunitat estudiada, la major part dels estudis s'han desenvolupat sota certes simplificacions. Recorrem a aquest treball ja realitzat i a aquest coneixement adquirit en el creixement de biofilms en els DWDSs per canviar el punt de vista clàssic del biofilm en estudis en els DWDSs, que inclouen les següents característiques. La nostra proposta es basa en l'ardu procés previ i posterior anàlisi mitjanant enfocaments d'aprenentatge automàtic. Es va dur a terme un procediment multidisciplinari, ajudant com un enfocament pràctic per desenvolupar una eina de presa de decisions per ajudar a la gestió dels DWDS a mantenir, en la mesura possible, els biofilm en els nivells més baixos, i la mitigació dels seus efectes negatius sobre el servei. Es proposa una metodologia per detectar les àrees més susceptibles al desenvolupament de biofilms en els DWDSs. En conèixer la ubicació d'aquests punts calents de la xarxa, les accions de mitigació podrien centrar-se més específicament, estalviant recursos i diners. A més, els programes de prevenció es podrien desenvolupar, actuant abans que les conseqüències del biofilm es noten pels consumidors.

D'aquesta manera, el cost econòmic seria reduït i la qualitat del servei podria millorar, finalment augmentant la satisfacció dels consumidors.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisors Rafael Pérez García, Joaquín Izquierdo Sebastián and Manuel Herrera Fernández for believing in me for this project and guiding me during these years. I would like to specially thank them for the positivism with which they have always surrounded me. A special mention also deserve the rest of the members of the FluIng research group, extraordinary people that made this journey much easier.

I would also like to express my gratitude for the Ph.D. grant of the Spanish Ministry Economy (Ref.: BES-2010-039-045); as well as for the travel assistance of this Ministry (Ref.: EEBB-I-13-06371 and Ref.:EEBB-I-14-09035) and of the Greek Ministry of Education (Ref.: 17472) that enabled me to carry out part of this work at the Aristotle University of Thessaloniki (Greece), supervised by the Prof. E. Darakas, and at the University of Sheffield (UK), supervised by the Prof. J. Boxal. It was a privilege to work with so talented and supportive researchers. A significant part of this thesis is due to the help I have received from them and their valuable teams.

Of course, more than thanks to my parents, for their unconditional support and patience, and to all my family for being always there for me. This thesis is specially for my intelligent and brave grandmother, Maria Angeles Pirón. Finally, thanks to my friends for being around even when I am miles away.

Contents

Declaration of Authorship	i
Abstract	iii
Acknowledgements	ix
List of Figures	xiv
List of Tables	xvii
1 Introduction	1
1.1 Objectives of the research	4
1.2 Outline of the thesis	5
1.3 Contributions	6
1.4 Statement of Originality	7
1.5 Publications	7
1.5.1 Publications in scientific Journals and Books	8
1.5.2 Works presented in conferences at national or international level	9
2 Biofilm in drinking water distribution systems	13
2.1 Biofilm overview	14
2.2 Biofilm development in DWDSs	15
2.3 Problems associated with biofilm development in DWDSs	17
2.3.1 Health risk	17
2.3.2 Aesthetic deterioration of water	19
2.3.3 Proliferation of higher organisms	20
2.3.4 Disinfectant decay	22
2.3.5 Biocorrosion	23
2.3.6 Operational problems	26
2.4 Biofilm control in DWDSs	27
3 Current approaches to study biofilm development in DWDSs	32
3.1 Biofilm growth devices	32
3.2 <i>In situ</i> biofilm sampling	36
3.2.1 Pipe cut-out sampling	37

3.2.2	Pipe device sampling	38
3.3	Microbial quantification in DWDS biofilm	39
3.4	Heterotrophic plate count (HPC)	42
4	Case Studies	47
4.1	Selection of case studies	48
4.2	Case study 1. Drinking water distribution system of Thessaloniki, Greece	49
4.2.1	Sampling protocol	52
4.2.2	Descriptive analysis	56
4.3	Case study 2. Pennine Water Group pilot distribution system in Sheffield, United Kingdom	57
4.3.1	Pennine Water Group's experimental facility and operating con- ditions	62
4.3.2	Biofilm sampling	65
5	Getting and pre-processing data	70
5.1	Data collection	71
5.2	Data pre-processing	74
5.2.1	Data unification	75
5.2.1.1	Variables design	75
5.2.1.1.1	Physical characteristics of the system	75
5.2.1.1.2	Hydraulic characteristics of the system	77
5.2.1.1.3	Sampling and incubation	80
5.2.1.1.4	Physico-chemical characteristics of water	84
5.2.1.1.5	Biofilm	86
5.2.2	Data cleansing	87
5.2.2.1	Variables cleansing	87
5.2.2.1.1	Variability reduction	87
5.2.2.1.2	Non-significant/informative variables identifi- cation	88
5.2.2.1.3	Removal of variables with high percentage of missing values	89
5.2.2.2	Cases cleansing	90
5.2.2.2.1	Inconsistent data identification	90
5.2.2.2.2	Removal of cases with high percentage of miss- ing values	90
5.2.2.2.3	Outlier detection	90
5.2.2.3	Clean data set	92
5.2.3	Data set reconstruction	94
5.2.3.1	Imputation of missing values	96
5.2.3.2	Complete data set	97
6	Data set: Exploratory Data Analysis	99
6.1	Descriptive data analysis	99
6.1.1	Target attribute	100
6.1.2	Categorical attributes	100
6.1.3	Continuous attributes	105
6.2	Exploratory data analysis	108

6.2.1	Categorical attributes	108
6.2.2	Continuous attributes	110
6.2.2.1	Data set clustering	111
7	Model development	118
7.1	Regression Trees	118
7.2	Regression Tree implementation	120
7.2.1	Testing the Regression Tree model	122
7.3	Random Forests	125
7.4	Random Forests implementation	126
7.4.1	Testing the Random Forest model	128
7.5	Conclusions	129
8	From pipe to network	131
8.1	Multi-agent systems	131
8.2	Discriminant Analysis via Label Propagation	133
8.3	Graph Theory Measurements to Assess the Importance of the Edges	134
8.3.1	Edge betweenness centrality	134
8.4	Case Study	135
8.5	Further application: Biofilm susceptibility as criteria for rehabilitation actions in DWDSs	138
8.6	Conclusions	139
9	Conclusions and Future Work	143
9.1	Merits of the new approach	144
9.2	Practical implications	145
9.3	Future perspectives	147
9.4	Final conclusions	149
A	Compiled variables with less than the 15% of data	150
B	Extract of the first 50 elements of the synthetic database	152
C	Ensemble of naïve Bayesian approaches for the study of biofilm development in drinking water distribution systems	154
C.1	Naïve Bayesian approaches	154
C.1.1	Augmented Bayesian Classifiers	155
C.1.2	A combined approach: bagging naïve bayes	156
C.1.3	A hybrid approach: Bagging leafs of naïve Bayesian trees	157
C.1.4	Summary of the results and conclusions	158
D	Modelling the Biofilm Development Process within pipes with Multi-agent systems	160
D.1	Modelling the Biofilm Development Process	160
E	Presentation of the web page sections	163
E.1	Presentation of the web page sections	163

Bibliography

166

List of Figures

2.1	Steps in DWDSs biofilm development.	17
2.2	Generalized trophic interactions in DWDSs	21
2.3	Bioelectrochemical interpretation of the role of biofilm in pipe biocorrosion [1]	24
2.4	Annual cost of corrosion in the utilities category in U.S. [2]	25
2.5	The distribution system as a biofilm growth reactor [3]	29
2.6	Metallic pipe with tubercles from the drinking water distribution system of Thessaloniki	30
2.7	Cleaning pig in a pipeline	31
3.1	Propella reactor [4]	33
3.2	Semicircular duct flow cell system [4]	34
3.3	Annular reactor with coupons/slides in the outer cylinder [5]	35
3.4	Cross-section of a Robbins device demonstrating the arrangement of the mounted slides [6]	35
3.5	General scheme of a Pedersen device. Eva Ramos Martínez ©.	36
3.6	General steps in pipe cut-out biofilm sampling	44
3.7	<i>In situ</i> biofilm sampling in DWDSs.	45
3.8	Left: the MRD developed by the Griffith University, Queensland. Right: MRD developed by the University of New South Wales/CRC for Water Quality and Treatment. Figures obtained from [7]	45
3.9	The Pennine Water Group coupon mounting within a pipe section [8]	46
4.1	The water supply network of the Universitat Politècnica de València - UPV	48
4.2	Thessaloniki, Greece.	50
4.3	Thessaloniki urban and metropolitan areas map. Licensed under CC BY-SA 3.0 via Commons	51
4.4	Aliakmon Dam. Figures obtained from EYATH	51
4.5	Thessaloniki's main water treatment plant. Figures obtained from Special Service for Water Supply and Sewerage of Thessaloniki (E.Y.D.E. Thessalonikis)	52
4.6	Thessaloniki's water treatment process.	52
4.7	Water supply network reservoirs. Figure obtained from Special Service for Water Supply and Sewerage of Thessaloniki (E.Y.D.E. Thessalonikis)	53
4.8	Detail of the sampled area in a plastic pipe	54
4.9	Biofilm sampling in Thessaloniki drinking water distribution system	55
4.10	Data obtained in each replicate and sampling point	58
4.11	Scatter-plots of the biofilm data obtained in the DWDS of Thessaloniki	59

4.12	Location of Sheffield city in UK	60
4.13	The geographical area covered by Loxley 2004 Water Supply Zone	61
4.14	Pennine Water Group’s experimental facility. Images borrowed from Dr. Katherine Fish, University of Sheffield	62
4.15	Schematic of each pipe loop. Figure obtained from [9]	63
4.16	Pennine Water Group coupon showing outer coupon (surface area 224 mm ²) with l insert (surface area 90 mm ²). Figure obtained from [8]	63
4.17	Coupons location in the pipe loop. Image borrowed from Dr. Katherine Fish, University of Sheffield	64
4.18	The three different hydraulic regimes based on daily patterns observed in real DWDS in the UK. Figures obtained from [10]	65
4.19	Biofilm sampling in PWG experimental facility. Images borrowed from Dr. Katherine Fish, University of Sheffield	66
4.20	Bacteria growth in the R2A agar plates	67
4.21	Isolated bacteria under UV light	68
4.22	PCR products visualized by agarose gel electrophoresis	69
5.1	Data cleansing process	87
5.2	Outliers in a 2-dimensional data set	91
5.3	Survey plot	92
5.4	Proportion and combination of the missing data in the database	95
5.5	Imputed data (in red) for the variables <i>w_temp</i> and <i>freecl</i> in each MICE imputation process	97
5.6	Survey plot	98
6.1	Biofilm attribute (<i>hpc</i>) statistics	101
6.2	Device attribute	101
6.3	Pipe material attribute	102
6.4	Duct shape attribute	102
6.5	Circulation type attribute	103
6.6	Constant circulation attribute	103
6.7	Removal technique attribute	103
6.8	Insert type attribute	104
6.9	Incubation time attribute	104
6.10	Plating method attribute	105
6.11	Itinerary attribute	105
6.12	Water source attribute	105
6.13	Incubation temperature attribute	106
6.14	Water temperature attribute	106
6.15	Free chlorine concentration attribute	107
6.16	Boxplots of the target attribute, <i>hpc</i> , grouped by the classes of the categorical attributes	114
6.17	Scatterplot of the water temperature attribute	115
6.18	Scatterplot of the free chlorine attribute	115
6.19	Scatterplot of the incubation temperature attribute	116
6.20	Average Silhouette width for 11 clusters	116
6.21	Data set partitioning	117

7.1	The obtained Regression Tree	121
7.2	Cross validation of the Regression Tree	123
7.3	The performance of the Regression Tree when testing it with metadata (Test 1) and study cases data (Test 2)	124
7.4	A Random Forest execution	125
7.5	The performance of the Random Forest when testing it with metadata (Test 1) and case study data (Test 2)	128
8.1	A multi-agent system.	132
8.2	Areas based on pipe average age used to design the network.	136
8.3	Results of the discriminant analysis via label propagation.	136
8.4	Results of the <i>edge betweenness</i> score.	137
8.5	Pipes susceptible to be replaced.	139
8.6	Biofilm susceptibility after progressive pipe replacement.	141
8.7	Evolution of biofilm susceptibility when replacing pipes.	142
9.1	QR code of the web page	147
9.2	The NetLogo model embedded in the web page	148
C.1	Bagging naïve Bayes process.	157
C.2	Kappa statistic value and RMSE for TAN, BNB, NBT and B-NBT.	158
C.3	Error percentages of the confusion matrix.	159
D.1	Modelling biofilm development within a pipe	162
E.1	The appearance of the web page	163
E.2	The “Biofilm for All” project presentation in the web page	164
E.3	The “Contact us” section in the web page	164
E.4	The “Already done” section of the web page	165

List of Tables

3.1	Main advantages and limitations of the presented devices. Extended from [11].	37
3.2	R2A media formula [12].	43
4.1	Data from the sampling in the DWDS of Thessaloniki (CI: Cast iron; PVC: Polyvinyl chloride; AC: Asbestos cement).	57
4.2	Main characteristics of the data attributes.	57
4.3	Data from PWG experimental facility.	67
5.1	Contacts made during the personal and institutional networking.	71
5.2	Journal papers used as data sources.	72
5.3	Removed papers.	73
5.4	Main variables of the physical characteristics group of the dataset.	76
5.5	Main variables of the group of hydraulic characteristics of the dataset.	77
5.6	Main variables of the sampling and incubation group of the dataset.	81
5.7	Main variables of the physico-chemical characteristics of the water group of the data set.	84
6.1	A general description of the data set.	100
7.1	Test data from the case studies.	124
7.2	Variable importance in Random Forest implementation.	127
8.1	Method for label propagation in practice.	133
8.2	Range of ages and materials of the pipe materials.	136
A.1	Compiled variables with less than the 15% of data.	151
B.1	Extract of the first 50 elements of the synthetic database.	153

Kenbeo kenmaro. . .

Chapter 1

Introduction

During the last years, a number of aspects have led water supply managers' interest in improving drinking water quality protection and control to focus on distribution – not only treatment. The new analytical techniques, particularly bacteria counting techniques, along with the increasingly demanding drinking water quality regulations and the fact that, nowadays, consumers are much more informed and aware about these issues, have raised the expectations on the quality of served water. To satisfy these demands water utility managers are attempting to increase infrastructure efficiency, as well as a cost-effective balance in drinking water distribution systems (DWDSs). However, this is not an easy task.

Most of the research on DWDS has been conducted at treatment level, contributing to improved treated water quality. However, there is a broad consensus [13] that the final goal of water utilities should be to offer good quality drinking water at the customers' taps rather than only at the treatment plant. It is known that the design of distribution systems inevitably causes water quality decay [14]. However, generally, the system design, by itself, does not explain this decay to its full extent. The reasons for high deterioration of water quality in distribution systems are not entirely clear, but it is known that one of the main actors involved in this decline is biofilm development on the pipe inner walls [14]. Biofilm is a complex structure of microbial communities that develops in the presence of water adhered to surfaces and coated by a protective layer segregated by

themselves [15]. Thus, microbial biofilms are capable of withstanding biocides and antibiotics more effectively than free-living microorganisms, thus supporting significantly higher doses of antimicrobial [16]. In addition to the fact that the presence of biofilm poses a constant threat to health, biofilm development brings in additional costs. It has an impact on energy requirements of a system, cause corrosion, and increase frictional drag [17], among others.

The study of biofilm formation in DWDSs is not new. The prevalence and significance of biofilm in DWDSs have encouraged extensive research from more than 30 years, since the first approaches carried out by Characklis in the 80's. However, operating pipes are not readily accessible and *in situ* data are very hard to acquire. Collecting samples from operational systems can represent an important challenge [18]. Aside from the obvious technical difficulties, access and sampling in these systems represent a major cost. This makes experimental approaches the most accessible way of acquiring data. Most of biofilm research in DWDSs is based on studying the influence of the hydraulic conditions, the physical characteristics of the distribution system and the physico-chemical characteristics of the circulating water, among others, on biofilm development, besides from the interactions among microorganisms [19]. Although numerous studies have been carried out in relation to the influence that a number of characteristics of the DWDSs have on biofilm development [10, 20–22], in most cases, these characteristics are studied individually or in pairs, at most.

Distribution systems are the major components of water utilities. Complex processes, including, physical, chemical and biological reactions take place inside. The complex and diverse environments inside the pipes favours the development of heterogeneous habitats in DWDSs, along time and space, making biofilm to exist at different levels of abundance within distribution systems. A better understanding of the microbial ecology of distribution systems is necessary to design innovative and effective control strategies that will ensure safe and high-quality drinking water at the end tap [23]. To understand its microbiology, it is important to understand the micro-environments available in DWDSs and how these environments affect biofilm development. A necessary and intensive work is being carried out in this direction through simplified approaches in order to study

the effect that each factor has on biofilm development. The methodological approach presented here, combines previous research and new acquired data on biofilm growth in DWDS. This approach will aid water utilities to develop new decision tools to control and minimised biofilm growth in DWDSs. We propose a methodology to detect the more susceptible areas to biofilm development in DWDSs. These hot-spots within the network are where mitigation actions should be enforced, thus saving resources and money, and aiding to develop prevention programs to be carried out before the consequences of biofilm development are noticed by the consumers. In this way, the economic cost would be reduced and the service quality improved, increasing users' satisfaction.

With this aim, we present innovative methodologies, changing the classical point of view of biofilm studies in DWDSs. Our proposal is based on arduous preprocessing and posterior analysis by Machine Learning approaches. The currently available data of different researches that have studied biofilm development in DWDSs under different conditions have been collected, pre-processed and analysed. The final experimental study is built up from the resulting preprocessing (unification, cleansing and reconstruction) of the previous databases extracted from the literature. Given the moderate size of the information collected, the process to combine all the data is based on easy looped Structured Query Language (SQL) queries. However, this data assimilation is not achieved in a straightforward manner since the aim is to acquire novel and useful knowledge after analysing the data. Complexity specially increases when the data is provided by various studies and information sources [24]. Different methodologies and biofilm growth reactors are used in the biofilm studies, making data collection ambiguous and difficult to compare, if not incomplete. Thus, we manage heterogeneity in data measurements, multi-scalarity, important presence of missing data, and different codifications, among other drawbacks. In summary, to avoid the classical limitations found when studying biofilm development in DWDSs, innovative tools are applied to the current resources to improve knowledge on biofilm in DWDSs. A multi-disciplinary, even transdisciplinary, procedure is undertaken, with the final aim of integrating the biofilm paradigm in the near future urban water supply system management.

1.1 Objectives of the research

The final aim of this research is to develop a practical and nowadays implementable model based on a multidisciplinary research vision to formulate effective biofilm control strategies. The crucial role of multidisciplinary in advancing biofilm research is demonstrated [25]. Drinking water distribution systems should act as protective barriers and with the relevant knowledge should be operated and maintained to prevent contamination and growth of microorganisms as the treated water travels to the consumer.

The way to achieve this main objective involves satisfying various partial aims, among which the following ones are considered:

- Compilation and updating of the existing information regarding biofilm development in DWDSs, covering both, microbiological and hydraulic aspects of the systems. This process is carried out to incorporate previous knowledge in relation to the behaviour of biofilm in these systems, and to unify and to determine what are the main limitations when studying biofilm development in DWDSs. The main features to take into account when studying biofilm and the current applied techniques for biofilm quantification are also reviewed.
- Generation of a complete data set by collecting data from different sources (quantification studies of biofilm in real or simulated laboratory DWDSs and bench top devices), in a first stage, and the subsequent application of pre-processing techniques to improve the quality of the database. Various processes such as outlier detection, elimination of incorrect data, and reconstruction of missing data, among others, have been performed. Finally, an exploratory study of the obtained data set has been carried out to check if there is any inconsistency in the database and establish its reliability.
- Development of a new methodology and approach to the study of biofilm in DWDSs through the application of machine learning techniques in order to extract valid and practical information from the generated data set.

- Testing the performance of the developed model to check its applicability and integration in current DWDS management with data obtained from the case studies in this work.

1.2 Outline of the thesis

This thesis has eight chapters:

- **Chapter 1** corresponds to this Introduction.
- **Chapter 2** describes the basics of microbial aspects of biofilm. It focus on biofilm development in drinking water distribution systems. The main problems caused in these systems due to the presence of these microbial communities are explained. The current biofilm control strategies in DWDSs are introduced and the necessity for more effective solutions is demonstrated.
- **Chapter 3** reviews the current situation of biofilm development in DWDSs. The various devices used at lab scale and bench top experimental approaches are described, together with their advantages and disadvantages. The traditional and the most up-to-date techniques of biofilm quantification are introduced. The heterotrophic plate count technique is further explained since it is the technique selected for this work.
- **Chapter 4** presents the two case studies used in this thesis. Both are described in detail. The sampling protocols are reported and a picture of the data obtained in each case study is shown.
- **Chapter 5** develops the study approach carried out in this work. The data compilation and pre-processing procedures are explained in detail. This last step is divided into data unification, cleansing and reconstruction, which are further explained.
- **Chapter 6** presents the obtained data set. It is deeply described and an exploratory analysis is also carried out. First the variables are individually studied regarding biofilm development, and secondly, the whole data set is considered.

- **Chapter 7** undertakes the presentation of the proposed algorithms. A theoretical introduction is made. Finally, the algorithms are tested and the results are discussed.
- **Chapter 8** proposes a multi-agent based methodology to develop a network scale biofilm development model. A possible practical implementation of this model is presented.
- **Chapter 9** summarizes the main conclusions and contributions of this work and offers ideas for further extend this research.

1.3 Contributions

The main contributions of this thesis are divided into two parts. One is based on the new methodological approach undertaken in the study field of biofilm development in DWDSs. The other is about the method to improve current DWDS management by integrating biofilm into the system.

This proposal, which we consider very innovative in this field, lays the foundations for the implementation, in operating water utilities, of a tool to identify and predict the DWDSs areas that are more or less prone to harbor biofilm. Thus, the managers of water services would have a complementary tool to aid decision making. It would increase the efficiency in the water services management and help to improve the quality of the service and of the tap water.

However, in this work, in addition to these main contributions, other achievements of interest have been achieved.

- The factors that determine the development of biofilms in DWDSs have been studied in detail and the problems associated with their presence in these systems. This has been achieved by collecting and updating the existing information regarding the development of biofilms in DWDSs, covering microbiological and DWDSs related aspects.

- In a first stage, a complete and extensive database has been generated by collecting data from different sources (quantification studies of biofilms in real or simulated laboratory DWDSs), the subsequent application of pre-processing techniques has been managed to obtain a complete database on the development of biofilms in DWDSs and supply systems' characteristics.
- We have identified patterns of biofilm behaviour in DWDSs depending on the studied characteristics by applying various Machine Learning techniques on metadata.
- The effectiveness and usefulness of the data science and Machine Learning techniques in this field has been also highlighted . These are very useful techniques to find interesting patterns and analyzing the particular influence of variables. These methods when used properly can be implemented in the study of biofilm development in DWDSs.
- A novel algorithm based on multi-agent systems has been developed in order to detect the most susceptible areas to biofilm development in DWDSs and to be used as a decision-making tool in DWDSs management.

1.4 Statement of Originality

This is to certify that, to the best of my knowledge the content of this thesis is my own work. This thesis has not been submitted, either whole or in part, for any degree or diploma at any higher education institution, except where acknowledged in the text. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

1.5 Publications

The main published contributions related to this thesis are given below:

1.5.1 Publications in scientific Journals and Books

- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Multi-Agent Approach to Biofilm Development in Water Supply Systems. Athens: ATINER'S Conference Paper Series, WAT2015, No. 1693, 2015 [26].
- M. Herrera; E. RAMOS-MARTINEZ; J. Izquierdo; R. Pérez-García. Graph constrained label propagation on water supply networks. *AI Communications* 28, 47-53, 2015 [27].
- E. RAMOS-MARTINEZ; M. Herrera; J. Gutiérrez-Pérez; J. Izquierdo; R. Pérez-García. Rehabilitation Actions in Water Supply Systems: Effects on Biofilm Susceptibility. *Procedia Engineering* 89, 225-231, 2014 [28].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Pre-processing and visualization of biofilm development in drinking water distribution systems. *Water Utility Journal* 7, 3-11, 2014 [29].
- E. RAMOS-MARTINEZ; J. A. Gutiérrez-Pérez; M. Herrera; J. Izquierdo; R. Pérez-García. Pipe database analysis transduction to assess the spatial vulnerability to biofilm development in drinking water distribution systems. Ch. 7 in: J.C. Cortés, L. Jódar Snchez and R.J. Villanueva (eds.), *Mathematical Modeling in Engineering & Social Sciences*, Nova Science Publishers, Hauppauge, NY, pp. 71-80, 2014 [30].
- M. Herrera; E. RAMOS-MARTINEZ; J. A. Gutiérrez-Pérez; J. Izquierdo; R. Pérez-García. On Kernel spectral clustering for identifying areas of biofilm development in water distribution systems. Ch. 8 in: J.C. Cortés, L. Jódar Snchez and R.J. Villanueva (eds.), *Mathematical Modeling in Engineering & Social Sciences*, Nova Science Publishers, Hauppauge, NY, , pp. 81-89, 2014 [31].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Ensemble of naïve Bayesian approaches for the study of biofilm development in drinking water distribution systems. *International Journal of Computational Mathematics* 91(1), 135-146, 2014 [32].

- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Drinking water distribution systems characteristics on biofilm development: a Kernel based approach. ATINER'S Conference Paper Series, AGR2013, No. 0773, 2013 [33].
- J. A. Gutiérrez-Pérez; M. Herrera; R. Pérez-García; E. RAMOS-MARTINEZ. Application of graph spectral methods in the vulnerability assessment of water supply networks. *Mathematical and Computer Modelling* 57, 1853-1859, 2013 [34].

1.5.2 Works presented in conferences at national or international level

- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Multi-Agent Approach to Biofilm Development in Water Supply Systems. *3rd Annual International Conference on Water*. Athens, Greece. 13-16/07/2015 [26].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Evaluación de la distribución espacial del biofilm en los sistemas de abastecimiento de agua. *XXVI Congreso Latinoamericano de Hidráulica*. Santiago de Chile, Chile. 25-29/08/2014 [35].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Rehabilitation action in water supply systems effects on biofilm development. *Water Distribution System Analysis (WDSA)*. Bari, Italia. 14-17/07/2014 [28].
- E. RAMOS-MARTINEZ; J.A. Gutiérrez Pérez, M. Herrera; J. Izquierdo; R. Pérez-García. Biofilm susceptibility in a drinking water distribution system regarding 24 hours curve demand. *7th International Congress on Environmental Modelling and Software (iEMSs)*. San Diego, California. 15-19/06/2014 [36].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Biofilm: influencia del diseño y operación de los sistemas de abastecimiento de agua. *IX SELASI. Seminario Euro Latinoamericanos de Sistemas de Ingeniería*. La Victoria, Venezuela. 12-15/11/2013 [37].

- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Métodos Kernel para el estudio del desarrollo de biofilm en los sistemas de distribución de agua potable. XII Simposio Iberoamericano sobre planificación de sistemas de abastecimiento y drenaje. Buenos Aires, Argentina. 11-15/11/2013 [38].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Estudio de la influencia relativa en el desarrollo de biofilm de las características físicas e hidráulicas de los sistemas de distribución de agua potable. III Jornadas de Ingeniería del Agua. Valencia, Spain. 23-24/10/2013 [39].
- E. RAMOS-MARTINEZ; J.A. Gutiérrez Pérez, M. Herrera; J. Izquierdo; R. Pérez-García. Metadata on biofilm development in drinking water distribution systems. XVI International Congress of the Catalan association for Artificial Intelligence. Vic, Spain. 23-25/10/2013 [40].
- M. Herrera; E. RAMOS-MARTINEZ; J. Izquierdo; R. Pérez-García. Graph constrained label propagation on water supply networks. XVI International Congress of the Catalan association for Artificial Intelligence. Vic, Spain. 23-25/10/2013 [41].
- J. A. Gutiérrez-Pérez; E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Graph spectral method to assess biofilm development in drinking water distribution systems. Mathematical Modelling in Engineering & Human Behaviour 2013. Valencia, Spain. 04-06/09/2013 [42].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Drinking water distribution systems characteristics on biofilm development a kernel based approach. Annual International Forum on Water 2013. Athens, Greece. 15-18/07/2013 [43].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Pre-processing and visualization on biofilm development in drinking water distribution systems. 8th International Conference of European Water Resources Association. Oporto, Portugal. 26-19/06/2013 [44].

- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Biofilms en los sistemas de distribución de agua potable. Aproximación basada en sistemas multi-agente. XI Seminario Euro Latinoamericano de sistemas de ingeniería. La Habana, Cuba. 26-30/11/2012 [45].
- J. A. Gutiérrez-Pérez; M. Herrera; J. Izquierdo; R. Pérez-García; E. RAMOS-MARTINEZ. Enfoque multi-agente para la identificación de elementos vulnerables en una red de abastecimiento. XI Seminario euro latinoamericano de sistemas de ingeniería. La Habana, Cuba. 26-30/11/2012. [46].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Ensemble of multiple data mining approaches to biofilm development in drinking water distribution systems Mathematical Modelling in Engineering & Human Behaviour 2012. Valencia, Spain. 04-07/09/2012. [47].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Estudio del desarrollo de biofilm en tuberías mediante redes bayesianas con variables mixtas. XXV Congreso Latinoamericano de Hidráulica. San José, Costa Rica. 09-12/09/2012. [48].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Evaluación de las características físicas e hidráulicas de los sistemas de distribución de agua que determinan el desarrollo de biofilms. XI Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje. Coimbra, Portugal. 02-04/07/2012. [49].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Modelling biofilm formation and evolution in drinking water distribution systems using a multi-agent approach. 2nd Meeting of Young Researchers Modelling Biological Processes. Granada, Spain. 02-06/07/2012 [50].
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Assessing variations in biofilm development in a drinking water distribution system by an object oriented Bayesian network approach. 6th International Congress on Environmental Modelling and Software (iEMSs). Leipzig, Alemania. 01-05/06/2012. [51].

-
- E. RAMOS-MARTINEZ; M. Herrera; J. Izquierdo; R. Pérez-García. Evaluación del desarrollo de biofilms en los sistemas de abastecimiento de agua mediante redes bayesianas. X Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje. Valencia, Spain. 30/11/2011. [52].
 - J. A. Gutiérrez-Pérez; M. Herrera; R. Pérez-García; E. RAMOS-MARTINEZ. La vulnerabilidad de los sistemas de abastecimiento de agua. X Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje. Morelia, Mexico. 30/11/2011 [53].

Chapter 2

Biofilm in drinking water distribution systems

Nowadays, every country is facing formidable challenges in meeting the rising socio-economical demands necessary to keep up developing in an increasingly competitive world. Improving drinking water access and quality brings large benefits to the development of countries at every economic level, through improvements in health outcomes and economy [54].

When compared with other sectors, particularly such other major social sectors as education and health, sanitation and drinking-water sectors receive relatively low priority for both official development assistance and domestic allocations [55]. However, DWDSs provide a basic service for citizens. Several problems associated with the management of these systems must be traced and addressed to improve the quality of the served water and of the supply, thus, to improve the efficiency of DWDS management. Biofilm is, from a microbiological perspective, one of the main sources of problems reported on DWDSs [11].

2.1 Biofilm overview

In the last years, a change in the microbiological paradigm of free floating bacteria has occurred affecting a wide range of areas (medicine, environment, and industry, among others). A wide variety of microorganisms live in communities associated to surfaces in contact with water. They are called biofilms [56]. Such biofilms are represented by structured consortia of sessile microorganisms characterized by surface attachment, self-produced exopolymeric structure or matrix (EPS), structural, functional and metabolic heterogeneity, capable of having intercellular communication by quorum-sensing and plurispecific composition [57]. These multispecies microconsortia can result from an association between metabolically cooperative organisms. Their proximity facilitates interspecies substrate exchange and the removal or distribution of metabolic products [58]. From an ecological point of view, populations of bacteria arise from individual cells, and form metabolically similar populations. These carry out interdependent physiological processes. In essence, biofilm represents an interdependent microbial community [58].

Specific extracellular signals regulate the activation of the metabolic pathways that lead to biofilm formation. These external signals come from diverse sources [59]. Biofilm formation can be considered a mechanism to protect microbial community from adverse environmental conditions. Microorganisms experience certain level of shelter and homeostasis when residing within a biofilm, thanks to the surrounding EPS [58]. In addition, it prevents antimicrobial agents to accessing biofilm by acting as an ion exchanger [60]. Biofilm polymers act as a sorptive sponge which binds and concentrates organic molecules and ions close to cells. EPS provides protection from a variety of environmental stressors, such as UV radiation, pH shifts, osmotic shock, and desiccation [58].

It has been calculated that between 70% and 99% of biofilm composition is water. Microorganisms just represent between 10% and 50% of biofilm total volume. EPS accounts for 50% to 90% of the total organic carbon of biofilms [61]. Concerning biofilm structure, it does not follow a standard rule. In DWDSs, biofilm may cover the entire surface [62] or form dispersal aggregates [63]. It depends on numerous environmental

factors and on the fact that every microbial community is unique. Biofilm structure can be influenced by several factors, such as surface and interface properties [64], nutrient availability, microbial community composition, and hydrodynamics, making the actual structure of any biofilm probably a sole feature of the environment in which it develops [61]. The highly permeable water channels interspersed throughout biofilm in the areas surrounding the microcolonies provide an effective nutrient and a metabolite interchange flow system with the bulk aqueous phase, enhancing nutrient availability as well as removal of potentially toxic metabolites [58].

The microbiological characteristics of the source water and sediments affect the structure and composition of the biofilm developed in DWDSs. The microorganisms involved in biofilm formation in DWDSs are those who have been released directly in the treatment plant or have been introduced into the distribution system at some point downstream of the treatment plant. Microorganisms penetrate into DWDSs by crossing the filtration line of the treatment plant in association with turbidity particles or coalesced to fine particles of the activated carbon used in filtration [65] or by intrusion, due to external contamination events in different steps of water treatment, storage and transportation: cross connections, backflow events, pipe breaks, negative pressure and due to improper flushing and disinfection procedures [57]. These microorganisms that successfully penetrate into distribution systems multiply in bulk water or in biofilms.

2.2 Biofilm development in DWDSs

Biofilm formation at molecular level varies greatly among different bacteria. However, there are some general features that are recognized in biofilm formation [59]. The sequence of general events leading to formation of a biofilm on a pipe surface is presented below in the following bullet points.

1. Conditioning Stage - Any surface immersed in water attracts within seconds organic and inorganic molecules from the overlying water forming a conditioning

film. The conditioning film is formed mainly by organic molecules creating a relatively nutrient-rich local environment in a nutrient-depleted environment, as is drinking water [17].

2. Initial Cell Attachment - The conditioning film neutralizes the surface charge, provide nutrients and polarize the forces between the film and the microorganisms, thus, the primary colonising bacteria adhere to pipe surfaces [17]. This initial attachment is based on electrostatic attraction and physical forces, not on chemical attachments. Some of the attached cells may permanently adhere to the surface [66]. While initial attachment depends on the physiological state of the cell prior to adhesion, significant changes in gene/protein expression occur upon irreversible adhesion and further sessile cell division to colonize the site of adhesion [67].
3. Slim formation - The primary colonising bacteria multiply and secret EPS, forming stronger bonds, which cement the compacted cell matrix to the pipe surface [17]. The EPS also act as an ion-exchange system for trapping and concentrating trace nutrients from the water.
4. Main Development - Further colonization is promoted by the EPS through physical restraint and electrostatic interaction. These secondary colonizers metabolize secreted products from the primary colonizers as well as produce their own secretion products which other cells then use in turn [66]. Over time bacteria multiplication and growth occurs, resulting in a thicker and denser structure.

The biofilm structure grows in the direction perpendicular to the wall; different parts of the biofilm are subjected to different conditions, which become gradually more hostile as the distance from the wall increases. This stage of development continues until a point of equilibrium is reached between the favourable and adverse growth conditions [17].
5. “Steady State” - The mature, fully functioning biofilm is like a living tissue on the pipe surface [66]. Shear forces exerted by flowing water impact on the mechanical stability of biofilms causing continuous erosion of surface layers and population succession [7]. Biofilm tends to reach a pseudo-steady state (Figure 2.1).

When biofilm achieves certain critical density, sloughing phenomena occur. This happens as cells and/or whole cell clusters are ‘sloughed off’ the surface and are carried by the flow as floating microbial communities which then settle downstream inducing colonization in other areas [17].

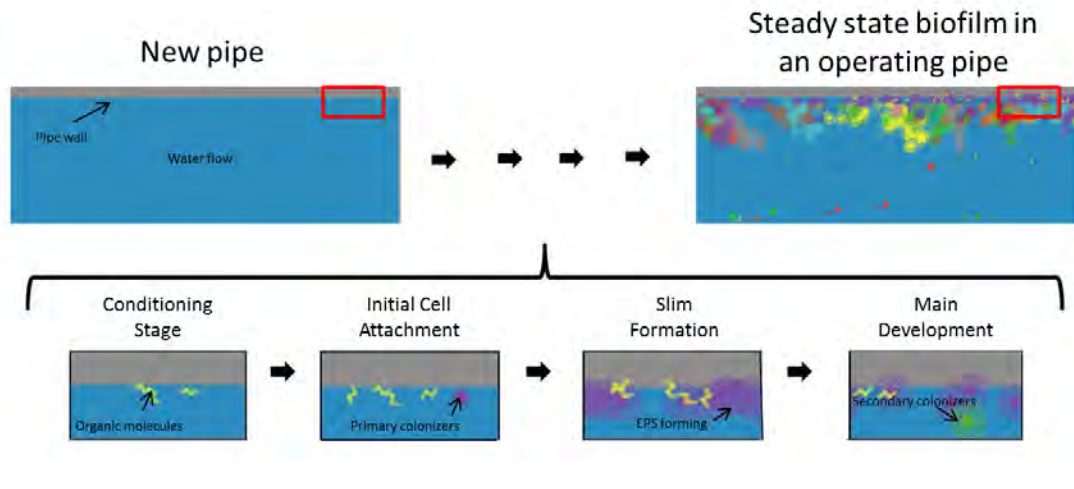


FIGURE 2.1: Steps in DWDSs biofilm development [3].

2.3 Problems associated with biofilm development in DWDSs

Biofilm development in DWDSs is lengthy. However, biofilm eventually has an impact on manifold aspects, from the quality of the served water to the hydraulic performance of the distribution network. In this section, the main problems related to the presence of biofilm in DWDSs are presented.

2.3.1 Health risk

A common feature of water-based pathogens is their ability to grow to problematic concentrations within biofilms on pipe walls and sediments, particularly during periods of water stagnation and warmer conditions; therefore, control above some critical concentration is necessary to manage pathogens. [68]. The most alarming consequences are multiplication and dispersion into water of bacterial pathogens, opportunistic pathogens [69], parasitic protozoa, viruses and toxins release by fungi and algae [57]. Biofilm offers

these microorganisms the necessary electrochemical and nutritional micro-environments for their survival and development [70].

Pathogens in water, even below detection limits, can accidentally attach to biofilm which act as their environmental reservoir and represent a potential water contamination source [71]. Various pathogenic bacteria have been identified in DWDS biofilms [72]. However, it is not clear how long they survive within them, since it depends on the species' biology and ecology, and the system's conditions [71].

Among the nuisance bacteria regularly found in drinking water and biofilms, species that are not common in aquatic environments may appear due to contamination events, with major impacts upon human health. Enteric bacteria [73] such as *Escherichia coli*, *Klebsiella pneumoniae*, *K. oxytoca*, *Enterobacter cloacae*, *E. agglomerans*, *Helicobacter pylori*, *Campylobacter* spp., *Shigella* spp., *Salmonella* spp., *Clostridium perfringens*, *Enterococcus faecalis*, *E. faecium*, as well as environmental bacteria becoming opportunistic pathogens [69] *Legionella pneumophila*, *Pseudomonas aeruginosa*, *P. fluorescens*, *Aeromonas hydrophila*, *A. caviae*, *Mycobacterium avium*, *M. xenopi*, together with other waterborne agents have been detected in drinking water biofilms [57].

Pathogenic bacteria can differentiate into primary and secondary. Primaries are those which can produce disease on the host (in this case, humans) by themselves. Among the pathogens found in DWDSs, these bacteria are often a minority. The secondary pathogenic (or also known as opportunistic pathogens, since they affect immunocompromised patients [69]) bacteria cannot set themselves any disease because they do not resist the host's defence mechanisms. They only achieve colonization when these mechanisms are depressed for various reasons. For example, physical agents, such as extreme temperatures or humidity, and chemical agents, such as corticosteroids and primary pathogen infections. The opportunistic pathogens are particularly important in DWDSs because they can cause disease in people with a weakened immune system. For example, elderly people, infants, cancer patients receiving chemotherapy or radiation therapy, people infected with HIV and patients with burns or transplants in hospitals are particularly susceptible to infections by opportunistic bacteria [74] [75].

Although bacteria are most common in DWDS biofilms, other types of potentially pathogen organisms have been identified, such as viruses and fungi. Most viruses found in DWDSs with impact on human health, are called enteric viruses and cause gastrointestinal diseases. In biofilms, viruses can accumulate, but not reproduce. However, it is known that in the presence of chlorine there are 10 times more viruses in biofilms than in the water flow, and in the absence of chlorine, 20 times more [65]. Fungi in drinking water are involved in health problems, originating from mycotoxins, animal pathogens and allergies. Water contaminated with fungi is of importance to hospitals, where immunocompromised patients undergo treatment: *Aspergillus* accounts for the majority of infections, with *Aspergillus fumigatus* accounting for 90% of cases [76].

In addition to act as shelters and reservoirs for microorganisms biofilm can enhance the health risk for humans indirectly. Some pathogens can grow and persist in DWDSs using metabolic products produced by non pathogenic members of biofilms [77]. This has particular relevance to organisms such as *Legionella*. Research has revealed that *Legionella* thrive in biofilms, and interaction with other organisms in biofilms is important for their survival and proliferation in water [78]. Moreover, relationships with certain algae and bacteria in biofilms also foster the growth of *Legionella* [78]. In the same way, although most of the bacteria in DWDS biofilms are not pathogenic [79], it is important to note that in some cases a prolonged treatment can select chlorine resistant-bacteria sub-populations [80].

2.3.2 Aesthetic deterioration of water

Aesthetic and organoleptic characteristics of water may be affected by a series of chemical substances, resulting in colour, odour and taste degradation. These chemical compounds are usually attributed to microbial biofilms associated to drinking water processing and distribution [57].

Biofilms represent small ecosystems within DWDSs and, although bacteria are the most abundant, there are other types of organisms associated with these ecosystems. In fact, the main cause of this last problem are fungi found in biofilms. This is because many of

the products and by-products of these organisms' metabolism have the ability to infuse to the treated water taste and smell, which affects directly to consumers [65].

It has also been found that algae that have the ability to grow in the distribution systems in the absence of light and housed in biofilms, can also impair the organoleptic properties of DWDS water. These algae can proliferate in the dark due to its ability to develop heterotrophic metabolism, using carbon as an energy source, and develop within biofilms [81]. Other biofilm organisms can also provide tastes and smells generating substances like worms and amoebae [82].

Water discolouration is caused by the detachment of iron and magnesium salts due to the corrosion of the inner walls of the pipes. Corrosion processes may be favoured by the metabolism of some DWDS biofilm microorganisms, as the sulphate-reducing bacteria (SRB). This phenomenon is particularly important in U.S. and in some European countries where cast iron was the first material used in DWDSs. The problem of discolouration has a great impact to the consumer, generating a lot of complaints from the users due to the so-called red water events. The problems associated with episodes of discoloured waters are mainly aesthetic. However, a more important consequence may be the loss of residual disinfectant and the consequent increase in biofilm development [83].

2.3.3 Proliferation of higher organisms

Biofilm in DWDSs may serve as a base of the food chain for fungi, protozoa, worms and crustaceans, among others. These organisms may be present in DWDSs even in the presence of residual disinfectant [84]. Food chains in distribution systems are relatively short and most animal species belong to the same trophic level. The majority are grazers and detritivore, and although some species of small carnivores have been even found, larger carnivores are scarce or non-existent. The existing general trophic interactions in DWDSs are summarized in Figure 2.2.

In the literature there are reports of the presence of small animals in DWDSs of North America, Africa and South and East Asia since the late XIX century (before the use of

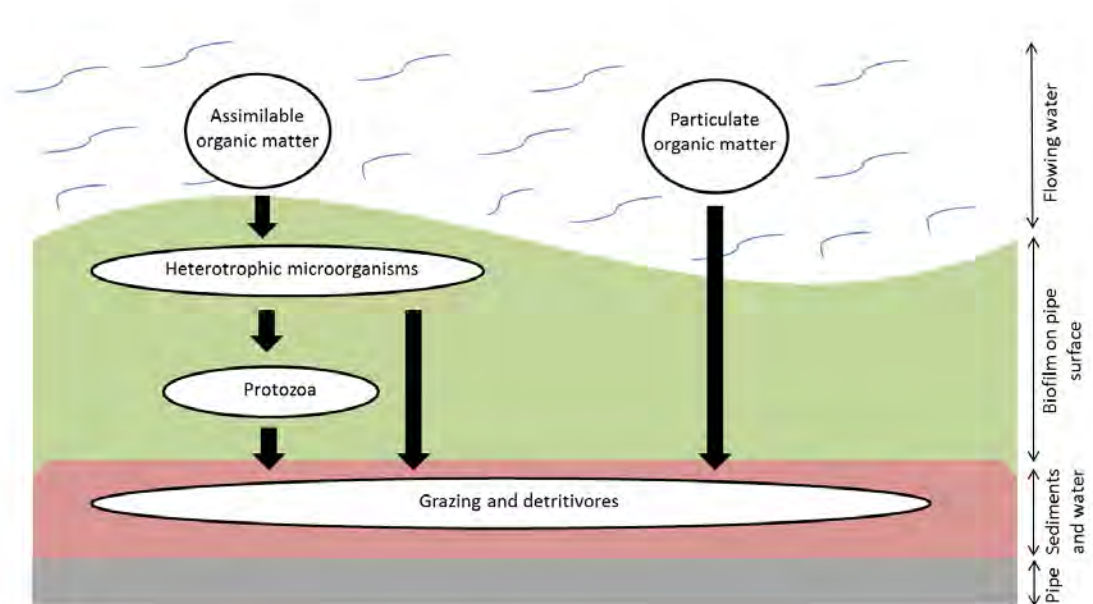


FIGURE 2.2: Generalized trophic interactions in DWDSs. Eva Ramos Martínez ©.

filtration and disinfection to be extended) to the XXI century. For example, DWDSs small animals population was studied in the UK during the 60's and 70's of the last century: 50 DWDSs were sampled and animals were found in all of them, although the water utility managers and their customers often were unaware of their presence [85].

Animals may be present in DWDSs due to various causes:

- They pass the treatment processes or they have colonized part of the treatment plant.
- They enter through defects in the integrity of distribution system, as bad covered reservoirs.
- They form breeding populations in distribution systems.

Their presence represents a double threat to human health, being also related to amoeba-resisting bacteria, such as *Legionella* spp. and *Mycobacterium* spp., which proliferate in protozoa thus increasing the probability of causing diseases in humans [57].

Moreover, the presence of animals in DWDSs is associated with discoloured water problems and can generate consumers' complaints. They can be both cause and effect. It

has been observed that animals grow especially in the low flow points as pipe dead spots where sediments accumulate. When examining samples of discoloured water it has been found that part of the particulate matter correspond to fragments of animals like empty shells, which are coloured with iron. Moreover, animal decomposition and faeces can create taste and odour problems in water. In contrast, since these animals feed on particulate organic matter, they limit the growth capacity of microorganisms such as Actinomycetes, which may cause smell and taste problems in water. However, both observations are assumptions, and it is unknown exactly the importance of these since the biomass of microorganisms in DWDSs is much higher than that of metazoa [85].

In tropical and subtropical countries some species of aquatic invertebrates act as intermediate hosts for parasites. However, in temperate countries, there is no evidence that any animal, found in DWDSs, is directly harmful to humans [85]. It is known that some pathogenic bacteria, such as *Legionella*, can grow and survive within certain amoeba (protozoa) in DWDSs [86] that exist in these systems feeding on biofilms (Figure 2.2).

2.3.4 Disinfectant decay

At the inlet of the distribution system, residual disinfectant is applied to water in an attempt to maintain during the time spent in the system the quality levels acquired in the treatment plant.

Although the main characteristic of the residual disinfectants is to remain in the water to prevent contamination up to the point of consumption, disinfectants are consumed in distribution systems due to reactions that occur, both, in the circulating water mass and in the pipe wall of the distribution systems. Deposits, corrosion products, microorganisms, organic impurities, ammonium and metal compounds (such as ferrous and manganese ions), are some of the water constituents which react with and consume the residual disinfectants [87]. Factors associated with the pipe wall that have been demonstrated to influence the chlorine decay are the material and diameter of the pipe, the initial concentration of chlorine, deposits and corrosion products, and biofilms [88].

Although disinfectant decay due to the presence of biofilm is a process that is still poorly understood, it is known that biofilm reacts with disinfectant consuming it. Biofilm chlorine concentrations being 20% or less than the concentration in the circulating water have been reported [89]. It has been shown that chlorine can diffuse into biofilms and be consumed. The limited penetration of chlorine into biofilm is not only due to a transient diffusion, but it is also caused by the neutralization of chlorine in the biofilm matrix [89]. Similarly, chloramine decay in a distribution system accelerates with the presence of biofilms and sediments [90]. It has also been observed that, under comparable conditions, there is a variability in the chlorine penetration rate into the biofilm. This suggests the existence of local differences in biofilm with respect to resistance to chlorine efficiency [89]. Areas with high resistance to chlorine can have greater capacity for reduction of chlorine than areas with faster penetration of chlorine. This may be due to the existence of higher cell density, sub-populations with higher reducing power per cell or higher density or reducing power of extracellular polymeric substances. The biofilm EPS is involved in the interaction with disinfectant, which can be associated to partial consumption. Cell bound EPS consumes disinfectant, retard bacterial membrane permeabilization, and thus decrease the susceptibility of bacteria [91]. In addition, EPS provides adsorption sites for dissolved organic matter (DOM) as a result of their composition and influence on biofilm structure. This DOM, which it is not sufficiently removed during conventional treatment processes, can exert a disinfectant demand [92].

Biofilm development on the pipe walls in DWDSs is directly involved in disinfectant consumption and has been proposed as a factor to be consider in the chlorine decay in DWDS modelling [93].

2.3.5 Biocorrosion

Biocorrosion or Microbiological Influenced Corrosion (MIC) refers to the microbial influence on the kinetics processes of metal corrosion. The MIC is caused by microorganisms adhered to the interface, i.e., biofilms [94] (Figure 2.3). Biocorrosion processes are associated with microbial activity, the products of their metabolic activity, including enzymes,

exopolymers, organic and inorganic acids and volatile components such as ammonia and hydrogen sulfide [95].

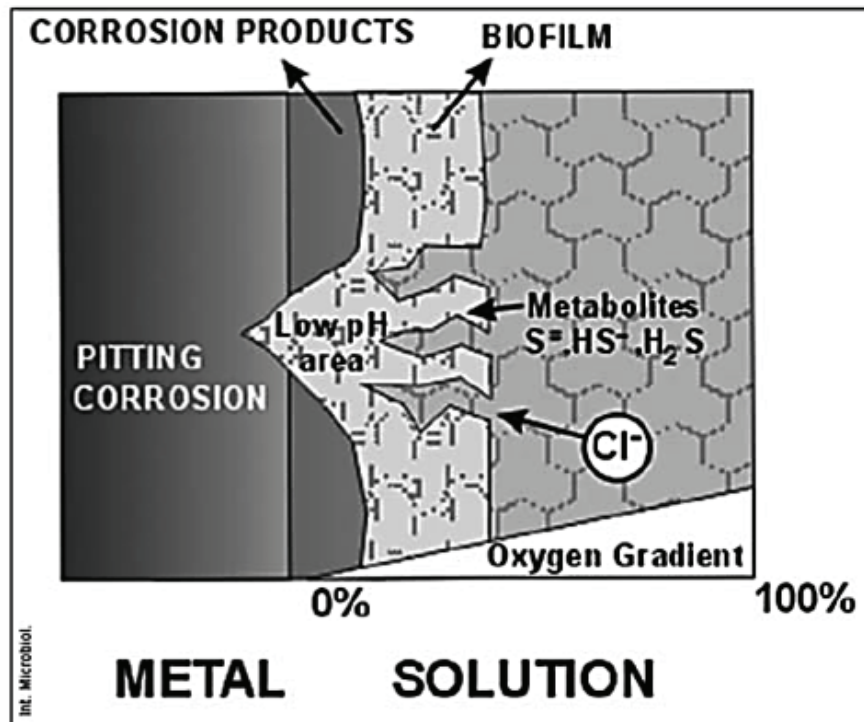


FIGURE 2.3: Bioelectrochemical interpretation of the role of biofilm in pipe biocorrosion. Figure obtained from [1].

The microorganisms involved in the MIC of metals such as iron, copper and aluminium, and its alloys are physiologically different. From an electrochemical point of view, corrosion is a chemical reaction where electrons are transferred from a zero valence metal to an external electron acceptor, causing the release of metal ions into the surrounding medium [96]. The ability of many bacteria to replace the oxygen for other oxidized compound as final electron acceptor in respiration allows them to be active in a wide range of conditions. The ability to produce a wide range of corrosive metabolic products in a wide range of environmental conditions makes microorganisms a real threat to the stability of the metals, including those that are designed to resist corrosion.

The basic conditions for MIC are present in DWDSs: bacteria and metal surfaces in contact with biofilm formation on the pipe walls [96]. The main group of bacteria associated with corrosion failures in metallic structures are the sulphate reducing bacteria (SRB), although there are many other groups capable of carrying out the MIC

in DWDSs (Figure 2.4). They all coexist in biofilms, often forming communities able to affect the electrochemical processes through cooperative metabolism, which these species individually are unable to start [94]. Microbial colonization of metal surfaces produces important changes in the type and concentration of ions, pH and redox potential, altering the passive or active behaviour of the metal substrate and the corrosion rates.

The cost of corrosion and prevention strategies in the United States were estimated at about 276 billion dollars a year, representing 3.1% of the gross domestic product (GDP). Other studies in the UK, Japan, Australia and Germany considered that the costs associated with corrosion corresponds to between 1% and 5% of the GDP of the respective countries [97]. The cost of the MIC is being estimated to represent 20% of the total cost associated with corrosion. Although there are no official data about the cost generated by the MIC in DWDSs, an idea of its impact can be made by observing the magnitude of the costs associated with corrosion in water distribution systems (Figure 2.4).

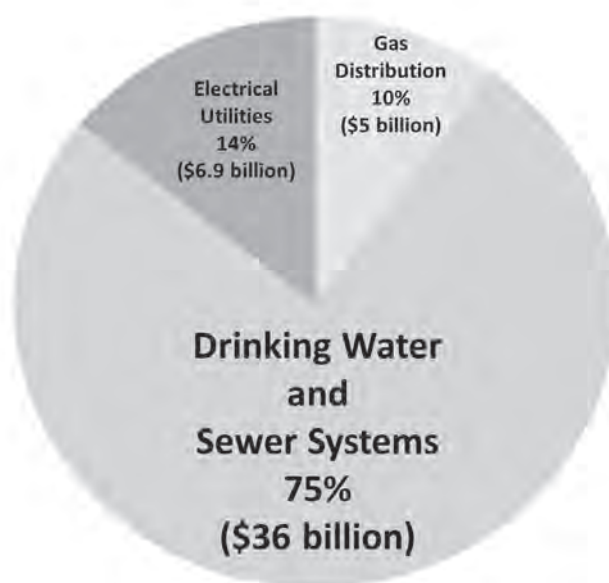


FIGURE 2.4: Annual cost of corrosion in the utilities category in U.S. [2].

The energy companies monitor the cost generated by the MIC, due to the associated

great damage cost. In the nuclear power generation plant operated by Ontario Hydro (Canada) cooling pipes were damaged by MIC and the estimated cost associated with the replacement of these pipes was of 300,000\$ per day and tube [98]. Although there are no actual data on the economic impact of MIC in DWDSs, it is not difficult to get an idea of the economic importance that must have in these systems. Among the costs attributable to this process we can mention the expenses associated with anti-corrosion treatments, replacement of pipes and damaged structures and costs associated with breakdowns or leaks, among others.

Aside from the economic cost associated with the MIC, it is important to note that DWDS infrastructures deterioration is a major cause of water quality and quantity decline [83]. Failures in distribution systems can cause, both, economic and social damage, among which damage to adjacent properties and businesses, traffic delays and other public nuisances must be added [99]; including the bad image that is given to the consumer when water is discoloured due to the presence of corrosion products [96].

2.3.6 Operational problems

Biofilm formation in drinking water pipes reduces the speed of the flow and the pipes' capacity of circulation. Its potential impact on pipeline performance can be significant [17].

Friction along the solid-liquid interface tends to increase with increasing surface roughness and interface instabilities [100]. It is the primary cause of energy losses and thus, of the flow capacity reduction within pipelines [17]. The operating point of a DWDS with a pumping system can be detrimentally affected. In the same way, maintenance costs [101] and operational efficiency may be unfavourably affected. This usually has significant implications for the cost and carbon footprint of pumping operations [17]. Thus, biofilm formation in DWDSs can affect the energetic efficiency of the system, having economical and environmental consequences.

For pipe engineered materials surface roughness is a function solely of roughness height. However, the change in surface roughness dynamics when biofilm develops in the inner

pipe wall depends on various parameters, such as the physical and structural nature (i.e. gelatinous or filamentous) of the biofilm [102]. Therefore, the effective roughness of a biofouled surface can be significantly higher than the one predicted based upon the roughness height and wall similarity hypothesis (turbulence outside the roughness sublayer is not affected by surface condition at sufficiently high Reynolds numbers, if the roughness height is small compared to the thickness of the boundary layer [103]) alone [17]. Biofilm development in pipe walls causes frictional drag, pressure drops and a need for increased pumping power, having an impact on energy requirements of DWDSs.

As a biofilm proliferates in a pipe, it reduces the pipe diameter causing an increase in the pipe frictional resistance [104], particularly in long pipe runs [17]. In [104] an increase of 33% in flow resistance is observed after the development of a biofilm layer of 160 μm . This increase goes up to 68% when the biofilm layer mean thickness was 350 μm . This results in reduced efficiency and increased costs [96] [105].

The presence of biofilm within a pipeline is operationally unavoidable, and consequently, its surface dynamics, as opposed to the characteristics of the underlying ‘clean’ engineered surface, should represent the ‘true’ effective roughness of all pipelines in service [17]. However, the current design approaches are outdated and the knowledge is limited to undergo these approaches.

2.4 Biofilm control in DWDSs

Ideally, preventing biofilm formation and not treating it after development would be the best option to avoid its negative consequences in DWDSs. However, currently there is no known technique able to successfully prevent biofilm development without causing adverse side effects [106].

Nowadays, two conventional anti-biofilm approaches are applied to try to control the microbial growth in DWDSs.

- Provide disinfectant residuals

The most common approach to limit potential unwanted microbial growth in DWDSs is the addition of disinfectants such as chlorine, chlorine dioxide or monochloramine after treatment [107] [108]. Chloramines, seem to be more stable, remaining for a longer time in the system, and are more effective in penetrating the biofilms [109]. However, it is well known that biofilm organisms display much higher tolerance to biocidal agents than their freely suspended counterparts [110]. Even if the biofilm viable population is reduced to less than 1% of the total population it can reseed biofilm development [106].

In this way, a preventive method of biofilm development in DWDSs could be to avoid, as far as possible, the existence of dead points in distribution systems or to increase control over them. In these points disinfectant consumption is very high and the presence of biofilms increases.

- Produce biologically stable water.

It refers to the inability of drinking water to support microbial growth. Biological stability is a function of biologically available organic carbon. Instability is measured as an increase of biomass and a concomitant decrease of substrate [111]. However, inhibition of biofilm formation by limitation of the carbon source is a virtually impossible procedure, as ultra-pure water systems have been found to support the formation of biofilms [112].

As it is virtually impossible to keep a DWDS completely sterile, microorganisms on surfaces will always be present, waiting for traces of nutrients. Even if 99.99% of all bacteria are eliminated by pre-treatment (e.g. microfiltration or biocide application), a few surviving cells will enter the system, adhere to surfaces, and multiply at the expense of biodegradable substances dissolved in the bulk aqueous phase.

The literature demonstrates that there is no strategy with absolute biofilm control efficiency [106]. More than 95% of the DWDS pipe biomass is located on the walls and less than 5% in the water phase [113]. It can be assumed that DWDS pipes resemble biofilm growth reactors, with a set of complex components and reactions (Figure 2.5).

Thus, a multidisciplinary approach to the treatment and management of biofilms seems to be the best approach.

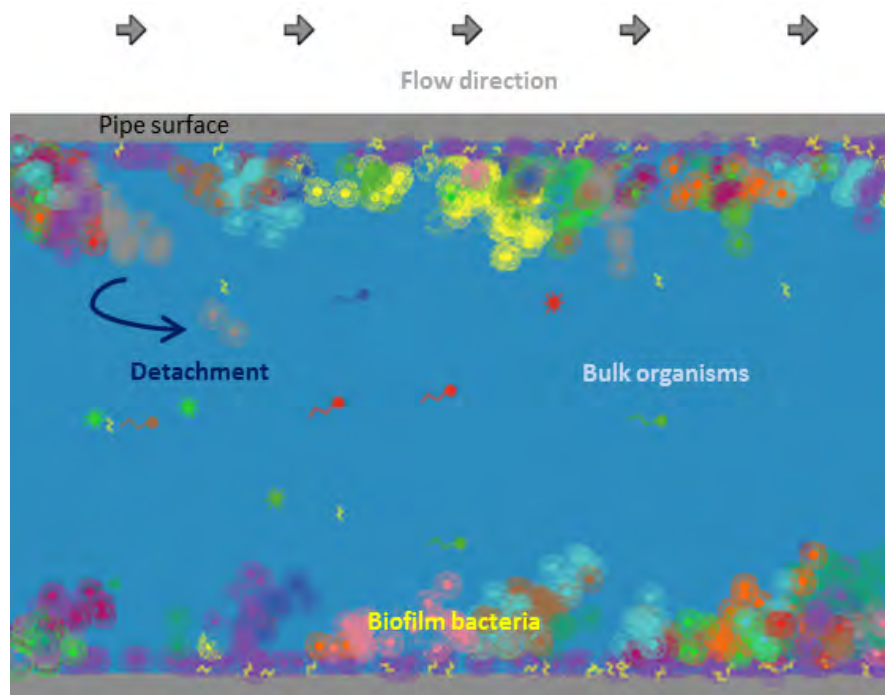


FIGURE 2.5: The distribution system as a biofilm growth reactor [3].

No approach has shown complete success in eliminating biofilms in DWDS pipe surfaces. A good programme for controlling biofilm development in DWDSs must incorporate multiple approaches. Once water is already in the distribution system, besides the chemicals control (residual disinfectants) a regular maintenance programme should also be carried out by hydraulic washes (flushing) and physical mechanisms (pigging).

Cleaning is an important issue in biofilm management. For cleaning, cohesion of the biofilm and adhesion to surfaces have to be overcome, which are both aspects of the mechanical stability of biofilms [114]. These help to redistribute the residual disinfectant to all sections of the system and eliminate the existing biofilms and sediments [115]. Flushing is defined as opening hydrants in a specific area, for a certain time, until the coming out water is of the desired quality [116]. Flushing has no lasting effect and the process should be repeated periodically.

Currently, the flushing programs are established as corrective measures in response to

users complaints after installations or repairs have been made, to remove the contaminants that inadvertently have been introduced into the system. Flushing programs can also be used as preventive maintenance practice. Some pipes have corrosion tubercles requiring the use of mechanical techniques to be removed, pigging. These tubercles (see Figure 2.6) are clusters of oxidized material which consume disinfectant and provide shelter for microorganisms, thus facilitating the formation of biofilms. The pigging technique uses pigs (pipeline inspection gauges) (see Figure 2.6) to perform various maintenance operations on a pipeline. This is done without stopping the flow in the pipeline. However, neither flushing nor pigging solutions are permanent and may not be sufficient to control a well-established biofilm. In some cases, replacement of the pipe is the most sensible option [117].



FIGURE 2.6: Metallic pipe with tubercles from the drinking water distribution system of Thessaloniki. Eva Ramos Martínez ©.

New trends are being studied to avoid biofilm development, such as the development of surfaces to which biofilms do not attach strongly. However, such materials and polymer coatings have neither been tested nor applied on DWDSs [118] [119]. Electric fields have also been used for both prevention of microbial adhesion and inhibition of biofilm growth [120]. Practical observation, however, has shown that all kinds of electrodes immersed into water can be colonized by biofilms [114].

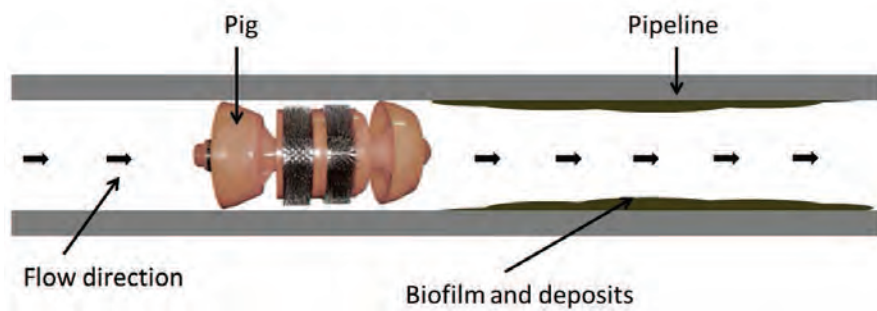


FIGURE 2.7: Cleaning pig in a pipeline. Eva Ramos Martínez ©.

Biofilms are present, to a greater or lesser extent, in all DWDSs. Knowing how different factors contribute to the growth of biofilms in these systems and ways to control these parameters is devised as the best prevention. Any step towards better understanding biofilm growth and properties will expand the possibilities for a flexible, effective and environmentally suitable response to biofilm development in DWDSs [114].

Chapter 3

Current approaches to study biofilm development in DWDSs

Biofilm research is a key component in DWDS microbial studies. However, pipes are not readily accessible since they are functioning systems comprised of buried infrastructure. Thus, obtaining representative biofilm samples of the spatial, temporal and physico-chemical variation of real DWDSs is highly challenging [121]. Sampling just the flowing water is not an option. It does not give information about the biofilm location, extent or composition and, moreover, it generally underestimates by several orders of magnitude the true microbial (surface) burden of a system [114]. Thus, collecting biofilm samples is necessary for its study. However, sampling biofilm from real systems is an enormous challenge [18].

3.1 Biofilm growth devices

To facilitate the study of various abiotic factors that might influence biofilm formation bench-top laboratory biofilm reactors are, habitually, used. Indeed, much of the current understanding about DWDS biofilm is based on studies from pilot or bench-top scale experimental models of drinking water systems [121].

The most commonly used devices in laboratory experiments are:

- Propella reactor. It is formed by two concentric cylinders. The propeller pushes the liquid down through the inner tube and then up through the annular section between both cylinders (Figure 3.1). The rotation speed of the propeller controls the fluid velocity, hydraulic residence time, and the flow rate. It is a perfectly mixed reactor. Normally, the coupons (removable surfaces for the growth of biofilm, used for quantification or identification of microorganisms) are located in the outer tube, although there are exceptions where coupons are located in the inner tube. The location of the coupons resembles better a real pipe situation and facilitates the sampling process. Generally, the removal of coupons does not change the flow conditions [11].

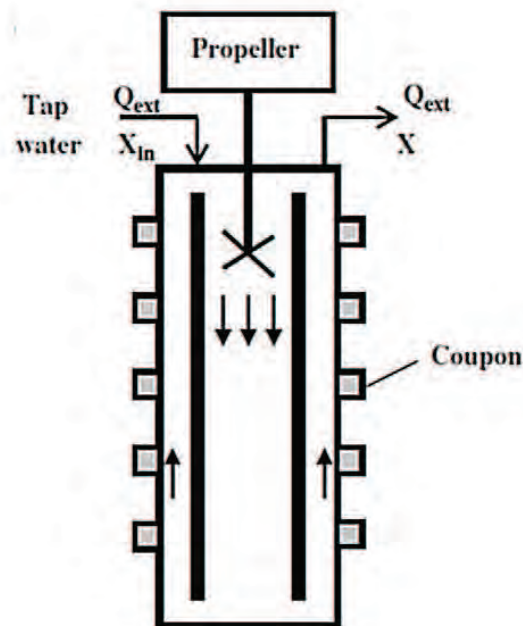


FIGURE 3.1: Propella reactor. Figure obtained from [4].

- Flow cell system. It consists of a duct segment. The coupons are removable and are inserted in the inner wall allowing for biofilm sampling over time. However, this system may present different configurations [11].
 - Semicircular duct flow cell reactor. The coupons are located on the flat wall and the flow passes through the duct from bottom to top (Figure 3.2).

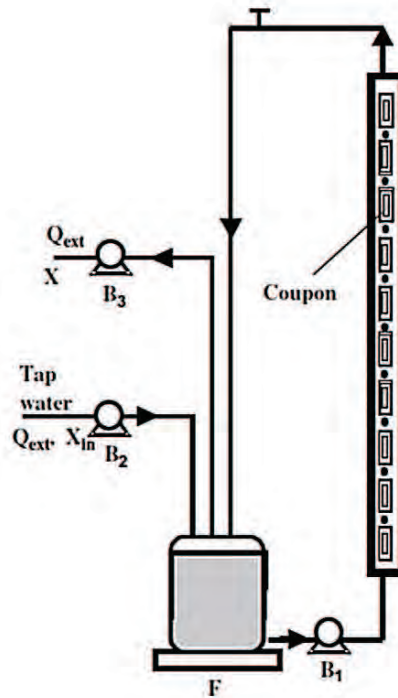


FIGURE 3.2: Semicircular duct flow cell system. Figure obtained from [4].

- Parallel plate flow cell reactor. A rectangular flow channel with small removable coupons inside.
- Annular reactor. It is also known as Rotatorque. It is constituted by two cylinders. One static external cylinder and other rotating internal cylinder. The speed is controlled by a motor in order to define the desired shear stress. The reactor can operate as an open/continuous system. Normally, the inner cylinder supports the coupons. However, in some cases, the coupons are located in the outer cylinder (Figure 3.3).
- Robbins device. It is a pipe with several threaded holes. Screws with slides are mounted on the front side and placed in these holes (Figure 3.4). They are aligned parallel to the water flow and can be removed independently [11].

Since the slides produce significant changes of the water flow, in some cases modifications were developed to avoid the flow characteristics perturbation. Modifications have also been applied to provide a large number of sample surfaces.

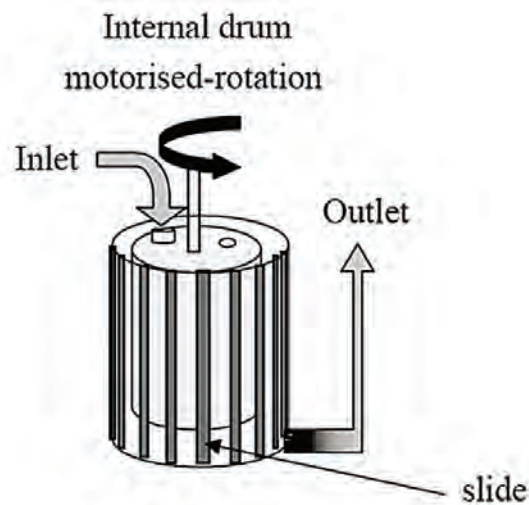


FIGURE 3.3: Annular reactor with coupons/slides in the outer cylinder. Figure obtained from [5].

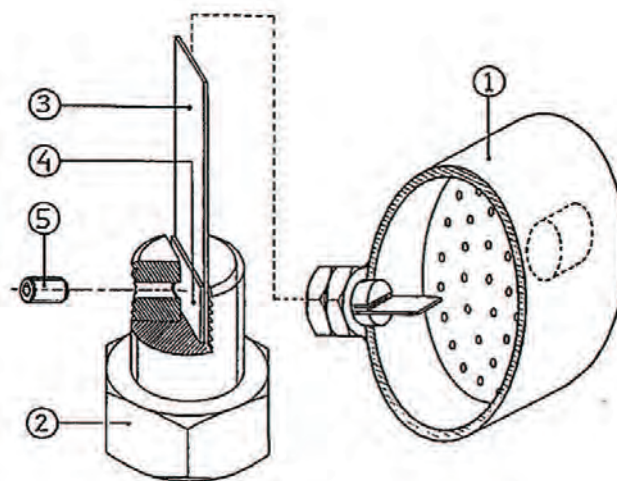


FIGURE 3.4: (1) Cross-section of a Robbins device demonstrating the arrangement of the mounted slides (3) into the cleft of screws (2) fastened by a plate (4) pressed by a countersunk screw (5). Figure obtained from [6].

The Robbins device is a very used device to study biofilm behaviour in pilot scale and also in real DWDSs [11].

- Pedersen device. It was named after its originator, Pedersen, in 1982 [122]. It was used to study biofilms in flowing-water systems. It consists of microscope cover slips fitted into acrylic plastic holders forming two parallel test piles, each with space for 19 slips (Figure 3.5).

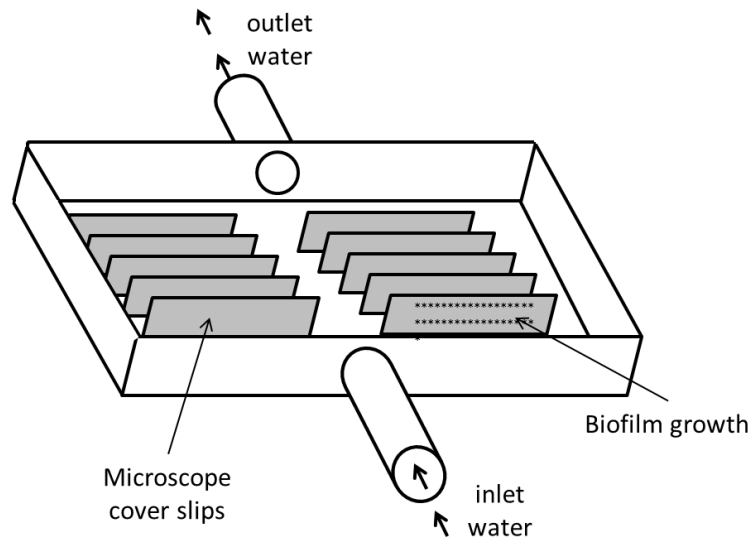


FIGURE 3.5: General scheme of a Pedersen device. Eva Ramos Martinez ©.

The sampling process in this device is done at fixed times. Normally, one sample consists of two slips, one from each of the two parallel piles. The sampled slips are replaced with new ones in order to maintain the flow conditions [122].

The main advantages and constrains of the devices presented above are summarized in Table 3.1.

3.2 *In situ* biofilm sampling

Bench scale systems are more often used for research due to their smaller size, better manipulation and lesser cost [11]. However, it is known that these systems do not exactly replicate the conditions of real pipe networks [8].

Currently two main different approaches exist for studying biofilms *in situ* in real DWDSs. One involves cut-outs of pipes; the other one relies on devices inserted into the pipe [18].

TABLE 3.1: Main advantages and limitations of the presented devices. Extended from [11].

Growth device	Advantages	Limitations
Propella	Easy control of the flow conditions; residence time controlled independently from the flowing process; flow conditions very similar to DWDSs; allows simultaneous study of different materials; allows periodical sampling	Changes in the flow caused by coupons; lack of sufficient sampling surface area
Flow cell reactor	Flow conditions similar to DWDSs; independent sampling at the desired time without changing or stopping the flow; allows study of different materials at the same time; easy to control environmental conditions	Flow changed by the coupons; biofilms are formed on a flat surface; lack of sufficient sampling surface area
Annular reactor	Allows study of different materials at the same time; interesting to assess the role of hydrodynamic conditions on biofilms; high surface area; easy sampling process; shear stress control independent from the fluid flow	The coupons can change the flow patterns; non-ideal mixing; non-uniform biofilm formation
Robbins device	Can be applied to real DWDSs with operational conditions very similar to reality; allows study of different materials simultaneously	The flow characteristics are changed with the presence of the coupons; the operational conditions cannot be effectively controlled when used in real DWDSs; lack of sufficient sampling surface area
Pedersen device	Possibility to study different materials; easy to control of operational conditions	The flow changes in the boundaries of the coupons; the biofilm is formed on a flat surface; lack of sufficient sampling surface area

3.2.1 Pipe cut-out sampling

Pipe cut-out sampling protocols are labour-intensive and expensive. Furthermore, the excavation and cutting processes often lead to concerns with contamination and representative sampling [18]. There is not a standard protocol to be followed when sampling

in situ biofilm on the internal surface of DWDS pipes. However, some general steps are recommended in order to assure a minimum quality of the samples (see Figure 3.6).

During sampling, the pipe cut (Figure 3.7) must be done as quickly as possible and any joining sections unbolted rather than cut to minimise disturbance. Any water drained from the main must be pumped away before it could re-enter the pipe and cause contamination.

It is recommended to take samples from a variety of locations as far into the pipe and in as representative an area as possible where risk of contamination is minimised [123]. Removed biofilm must be re-suspended in an appropriate buffer solution (for example: 1/4 strength Ringers solution, phosphate buffered saline (PBS) or sterile (chlorine-free) tap-water). The scraping tool or swab should be rinsed into the biofilm sample suspension, taking into account the volume used to enable calculation of the concentration per unit of area [7]. Samples must be transported immediately to the laboratory in insulated cold boxes.

It is highly challenging to acquire biofilm samples that are representative of the spatial, temporal and physico-chemical variation of real DWDSs since communities comprised in the functioning of buried infrastructures are alive [121]. In the limited cases that biofilm *in situ* samplings are carried out, due to the cost and complexity of the process, sampling is normally based on availability, due to maintenance operations of the network, and not as part of a structured survey.

3.2.2 Pipe device sampling

Pipe device sampling can be carried out either within a pilot-scale test facility or in an operational DWDS. It allows the study of biofilm dynamics over time in relation to changing abiotic and biotic factors *in situ*.

Commonly, coupons that can be deployed repeatedly are used. Devices, as the explained before modified Robbins device (MRD), are also installed (Figure 3.8). They can be run at system pressure. They are intended to fit coupons flush to the device's wall to minimise the disruption of hydraulic conditions [7]. Generally, the main limitation of

these devices is that they distort hydraulic conditions in pipes. In most cases, shear stress and turbulence regimes are different from those expected in real pipes, artificially influencing the way biofilms develop [18].

New devices are being developed in order to avoid these drawbacks. In this context, the Pennine Water Group coupon (PWG Coupon) goes a step further, since the coupon is curved and therefore sits flush with the pipe wall thus reducing the distortion of hydraulic conditions (Figure 3.9) [8]. Another advantage is that the coupon comprises two parts; a removable ‘insert’, which allows the analysis of biofilms *in situ* and an outer part that can be used to extract nucleic acids for further characterisation of microbial communities [18].

3.3 Microbial quantification in DWDS biofilm

There is no single accepted method to quantify biofilm cells [121]. The most commonly used biofilm quantification techniques are presented below.

- Culture-dependent techniques - Heterotrophic plate count (HPC)

Culturing bacteria on a solid, non-selective growth medium is a simple and widely accepted method to isolate and enumerate bacteria from biofilms, as well as from the bulk water [7].

Specific media is used to select a given metabolic group. Thus, plate counts do not provide an accurate understanding of the diversity of the microbial population present in a water sample. In the same way, the viable but non-culturable fraction of organisms are not accounted for with these techniques. Despite the limitations of culture-dependent methodologies, they are the current regulatory requirement used by water companies and analytical laboratories to routinely monitor microbial quality of drinking water [18].

The reference method used for routine bacteriological monitoring in drinking water is heterotrophic plate count measurements. It assesses only heterotrophic bacteria able to form colonies on a solid medium at a specific temperature in a specific time.

The low cost, relative simplicity, wide acceptance and long history of the method makes HPC a convenient tool for water utilities [18] and thus, for the scope of this research. This methodology is further described below in Section 3.4.

- Epifluorescence microscopy - Total cell count (TCC)

Epifluorescence microscopy based methods offer a faster alternative for monitoring the quality of drinking water than traditional plate counts, which have long incubation times [18]. Bacteria are concentrated using membrane filtration, then bacteria are stained with a fluorescent dye and the total cell count are determined by microscopic counting [124]. For optimal accuracy, it is recommended to count about 400 cells on multiple filters. Different fluorescent dyes can be used. Some of the most commonly used dyes to quantify microorganisms in biofilm and water samples are acridine orange (AO) [125], 4,6-di-amino-2 phenylindole (DAPI) [126] and 5-cyano-2,3 Dytolyl Tetrazolium Chloride (CTC) [127]. The reported TCC results are 10^5 to 10^7 cells/ml in bulk water, 10^5 to 10^7 cells/cm² on pipe wall biofilm [128].

Total counts provide no information on whether biofilm bacteria are alive or dead. The Live/Dead BacLight stain (Molecular Probes) uses a combination of two nucleic acid stains (SYTO 9TM dye (green-fluorescent) and propidium iodide (PI) (red-fluorescent)) to discriminate between live and dead bacteria. The SYTO 9TM dye penetrates all membranes while PI can only penetrate cells with damaged membranes. Thus, cells with undamaged membranes will stain green while cells with compromised membranes will stain red [129].

The flow cytometry method (FCM) for a total bacteria count has been found to be rapid and simple [130]. It has been recently introduced in drinking water research. It can count both the cultivable and non-cultivable cells with high sensitivity and accuracy [128]. Direct count and flow cytometry methods have been compared and FCM was reported to be more accurate [131]. However, the use of FCM in the field of drinking water distribution system research is an emerging technique [128]. There are, to date, few FCM results of pipe wall biofilm bacteria available.

- Biomass estimation - Adenosine triphosphate (ATP) assay

The adenosine triphosphate assay estimates biomass and bacterial growth through quantification of ATP. It is a rapid approach with low detection limits [128]. Adenosine triphosphate (ATP) is the major form of cellular energy. As such, the concentration of ATP reflects both the concentration of cells in biofilm as well as their metabolic state [7]. Briefly, cellular ATP reacts with a luciferin-luciferase complex, the luminescence produced in this reaction is proportional to the concentration of ATP, which is then correlated to the quantity of biomass in the sample [111]. The method requires a luminometer to measure the light production. Nowadays, ATP can be easily assessed using the BacTiter-Glo Microbial Cell Viability Assay (Promega, UK), which allows quantification of several samples simultaneously using a microplate reader [18].

The ATP assay is fast, of low cost and easy to perform, thus it is an ideal tool for monitoring purposes [18]. The use of ATP is well-established in drinking water-related research and is used as a reliable method to estimate microbial activity [111]. ATP has been successfully used to quantify active biomass in water treatment processes, distributed water, and pipe wall biofilm bacteria in drinking water distribution systems [128]. However, it must be noted that ATP measurement gives no indication of activity *in situ* [7].

It should be noted that all mentioned methods are designed for water samples. For measurements conducted with DWDSs biofilm samples, in most cases, pretreatment is necessary to detach the microbes into suspension for further analysis [128]. At this point it is important to note that, although various cell detachment methods (scraping, swabbing, sonication, stomaching) may not directly interfere with the analysis, differences in terms of the effectiveness of biofilm removal methods from the coupons will affect the obtained results. With any method, it is required the removal of biofilm from a test surface [7].

3.4 Heterotrophic plate count (HPC)

Heterotrophic plate count has a long history as a water quality indicator. HPC has been used since 1894 to determine bacterial concentrations in distribution of drinking water [132]. Over the decades, interpretation of HPC results has shifted from indicating drinking-water safety to a role in determining drinking-water quality [133]. The basis for water quality assessment is outlined in several national and international standards, and even though regulations differ slightly, requirements generally include monitoring of microbial parameters such as faecal indicators (coliforms, *Escherichia coli*, *Enterococcus* spp.), the opportunistic pathogen *Pseudomonas aeruginosa*, and determining the HPC [134].

HPC is the primary parameter for assessment of the general microbiological quality of drinking water [84], and many studies have applied HPC analysis for enumerating biofilm cells per surface area of material supporting the biofilm [11]. It measures the concentration of viable heterotrophic bacteria able to reproduce under established test conditions. HPC is a subset of the total live cell count. The most critical limitation of HPC is that it only counts media-cultivable bacteria, thus the fraction of cells which are not cultivable under standard culture conditions is not recognized [11]. The microorganisms that have complex nutritional requirements and viable but non-culturable (VBNC) bacteria will not be detected [7]. HPC guarantees that all cells counted are viable. Information on viability may be more useful than total cell counts in situations where a large portion of cells may be dead or unable to replicate to produce an infection. However, HPC does not provide an indication of pathogenicity since it cannot distinguish the majority of harmless bacteria from the few of clinical relevance. HPC determinations can be a useful tool to monitor efficacy of drinking water treatment processes and undesirable changes in bacterial water quality during storage and distribution, but not health associated risks [135].

There are several standardized HPC methods but not a generally approved standard operating procedure [18], different nutrients, culture media, temperature, and incubation periods are applied in HPC methods resulting in significant differences in HPC

enumeration [134]. Thus, given the difficulties in comparing results from different studies, [128] the thesis focuses on long term incubation R2A agar (7 days, 20-28°C). It is a low nutrient agar that gives higher counts than high-nutrient formulations [12], and it is recommended in the US [136]. R2A agar was proposed in 1979 by Reasoner and Geldenreich [137] and a few years later accepted by the American Public Health Association (APHA) as an alternative medium for the enumeration of stressed cells in treated potable water [138]. By the use of a medium like R2A in combination with a low temperature and long incubation time it is possible to induce the recovery of these damaged cells [138].

TABLE 3.2: R2A media formula [12].

Proteose peptone Number 3 or polypeptone	0.5 g.
Casamino acids	0.5 g.
Yeast extract	0.5 g.
Glucose	0.5 g.
Soluble starch	0.5 g.
Sodium pyruvate	0.3 g.
Di-potassium hydrogen phosphate (K_2HPO_4)	0.3 g.
Magnesium sulfate heptahydrate ($MgSO_4 \cdot 7H_2O$)	0.05 g.
Agar	15.0 g.
Reagent grade water	1l.
Final pH 7,2 0,2 at 25°C	

In the R2A agar the source of nitrogen is the peptone and the yeast extract supplies the vitamins and growth factors. The source of carbon is the glucose and magnesium sulphate and potassium phosphate maintains the osmotic pressure. Di-potassium Phosphate is used to balance the pH, and Magnesium Sulfate Heptahydrate is a source of divalent cations and sulfate. Sodium Pyruvate increases the recovery of stressed cells. The starch is a detoxifier and sodium piruvate increases the recuperation of stressed cells. The agar acts as a gelling agent [138] (see Table 3.2).

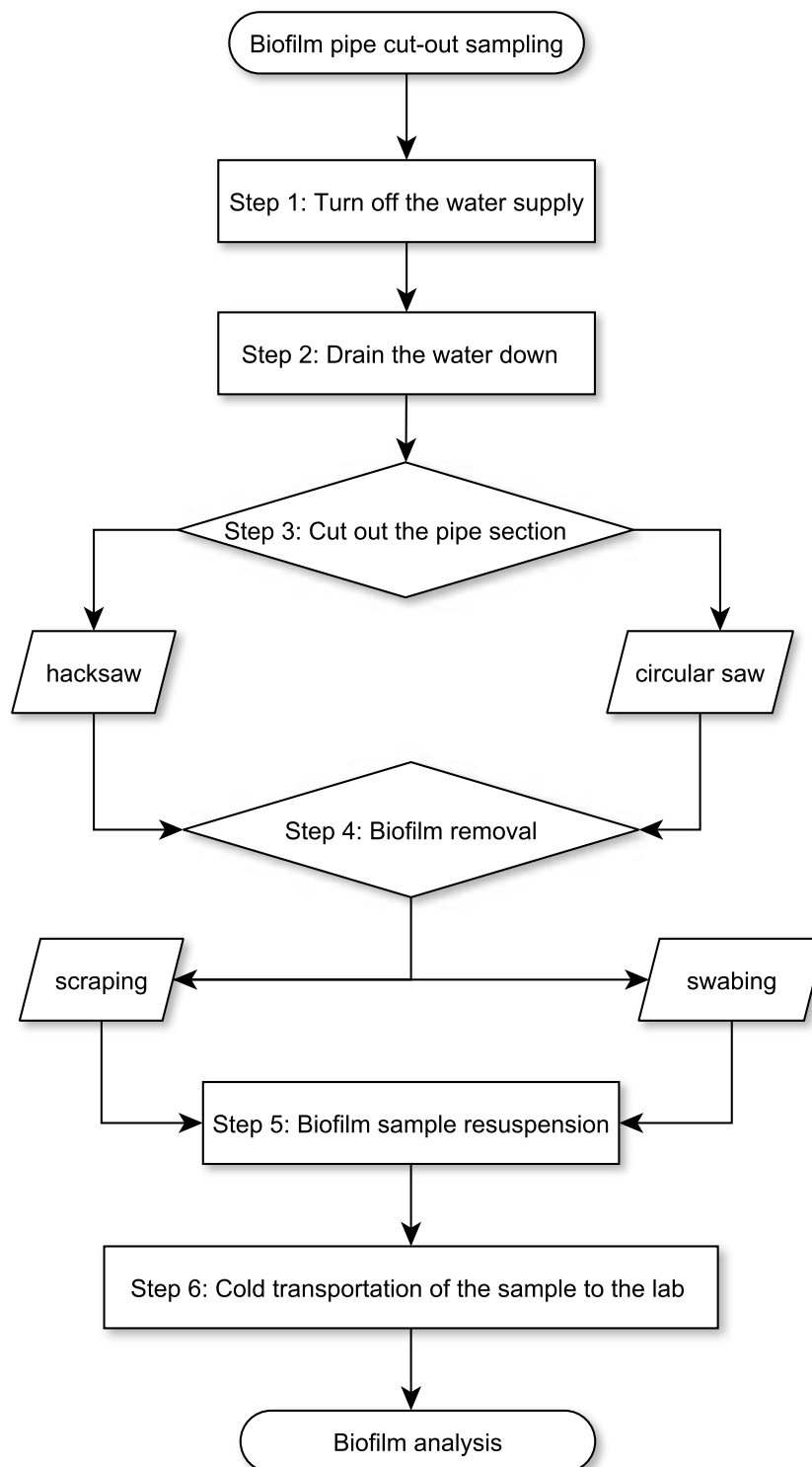


FIGURE 3.6: General steps in pipe cut-out biofilm sampling.



FIGURE 3.7: *In situ* biofilm sampling in DWDSs. Eva Ramos Martínez ©.

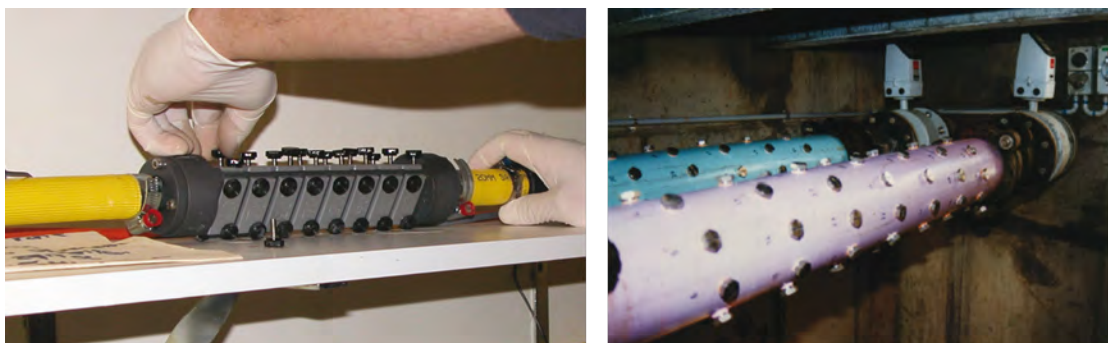


FIGURE 3.8: Left: the MRD developed by the Griffith University, Queensland. Right: the MRD developed by the University of New South Wales/CRC for Water Quality and Treatment. Figures obtained from [7].



FIGURE 3.9: The Pennine Water Group coupon mounting within a pipe section. Figure obtained from [8].

Chapter 4

Case Studies

Deep understanding of the interactions among the large spectrum of DWDS characteristics and how they globally affect biofilm development is needed. In this context, the main objective of this research is to predict the cultivable bacteria attached to the inner walls of DWDS pipes, based on as many as possible characteristics, using for the analyses the maximum amount of information to be handled by Machine Learning methods. To this purpose, we have sampled biofilm in operational DWDSs, where all the parameters do interact.

A case study is an empirical inquiry where a contemporary phenomenon within its real-life context is investigated, especially when boundaries between phenomenon and context are not clearly evident, and in which multiple sources of evidence are generally used [139]. Although case studies are very time consuming, and can be difficult to carry out and analyse, also have some very beneficial advantages. They help to build upon or enhance a body of knowledge, compare specific aspects across other case studies and dig into specific situations and extract ideas that can be generalized into principles for others to apply [140].

4.1 Selection of case studies

The water supply network of the Universitat Politècnica de València - UPV was first selected to collect biofilm samples. After checking the map and characteristics of the network (Figure 4.1) it was confirmed that it had enough variability to be a suitable network for our aim. Besides, its size, complexity and proximity made it the best option to start with. The Vice Chancellor of Infrastructure of the UPV agreed to our proposal therefore we proceeded to select the location of the minimum sampling points needed to carry out our study, taking into account the physical and hydraulic characteristics of the supply system. Unfortunately, when asked for a budget to carry out the work needed to access the pipes it was found that it was more difficult than expected and so, more expensive than predicted. There were no possibilities to afford the final cost of the sampling.

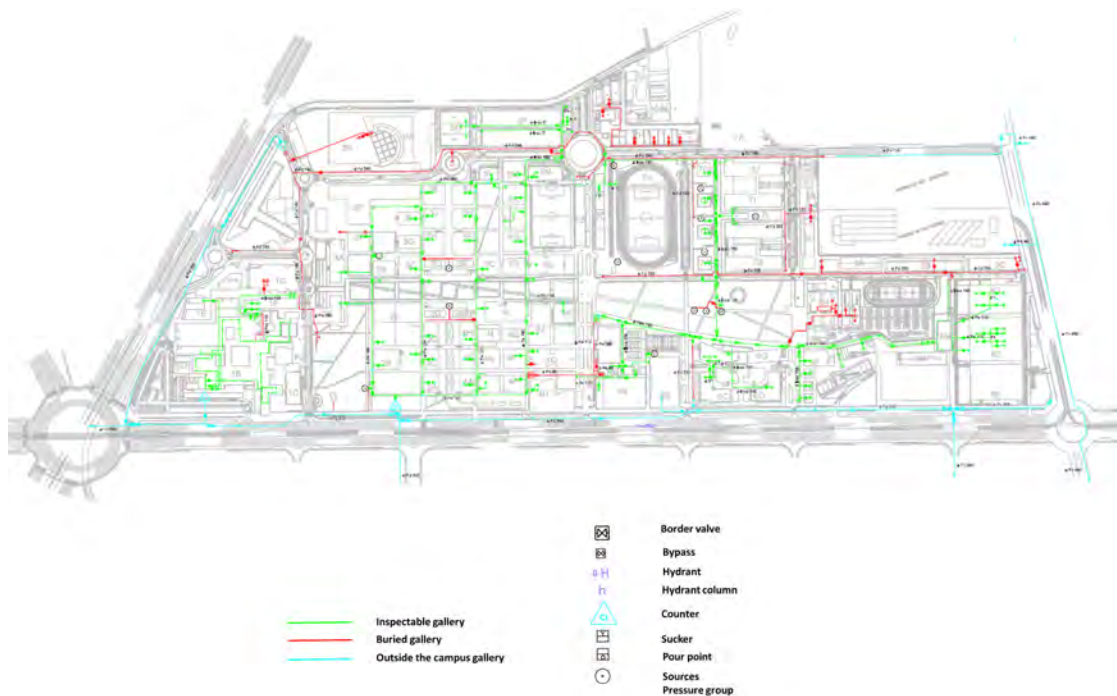


FIGURE 4.1: The water supply network of the Universitat Politècnica de València - UPV.

After realizing that digging just for samplings was not an option it was decided to get

in contact with the water supply network managers in order to convince them to let us sampling while the daily maintenance works of buried pipes was carried out. A report was written to explain the effects that biofilm development causes in DWDSs and the benefits that would be obtained by understanding how the interaction of the different characteristics of these systems affects biofilm development. The main benefit is to minimize biofilm growth, and therefore, its negative consequences in the quality of water and the service of the water utilities. Since part of the water companies are also managed by the councils, some councillors from different councils were also contacted in order to find out if they were willing to collaborate. The initial round of contacts resulted in various meetings. Although the interest and need of the study were not questioned, due to issues regarding confidentiality of the data and results, no agreement was achieved.

At this point, Professors with experience in working and collaborating with water utilities for academic purposes were contacted. Through Prof. Konstantinos Katsifarakis, we got in contact with Prof. Efthymios Darakas of the Aristotle University of Thessaloniki - AUTH (Greece), with experience in working with the Thessaloniki Water and Sewage Company - EYATH, that accepted to collaborate in the project. Thanks to his help, the Laboratory of Environmental Engineering and Planning of the AUTH, the EYATH and the grants awarded by the Ministry of Economy and Competitiveness of Spain (Ref.: EEBB-I-2013-06371) and the Hellenic Republic State Scholarships Foundation - IKY (Ref.: 16754) a protocol for sampling biofilm inside the pipes of the city of Thessaloniki was performed.

4.2 Case study 1. Drinking water distribution system of Thessaloniki, Greece

Located on the Aegean Sea in north-eastern Greece, Thessaloniki is the country's second largest city (Figure 4.2).

The Thessaloniki Urban Area is formed by six self-governing municipalities (see Figure 4.3), where, by far the largest municipality is the municipality of Thessaloniki (the city



FIGURE 4.2: Thessaloniki, Greece.

centre). In the 2011 Greek census, the municipalities of the urban area had a combined population of 790,824 inhabitants, while their combined land area was 111,703 km².

Management of water resources and collection and treatment of urban and industrial sewage in the broader area of Thessaloniki is carried out by the semi-private utility Thessaloniki Water Supply & Sewerage Co. S.A. (EYATH) [141]. The main raw water is obtained from surface (Aliakmonas river), although there are also groundwater reserves. The work begins at Barbares (Aliakmon Dam), roughly 40 km before the estuary of the river (Figure 4.4). The water flows by gravity in a 50 km linking canal and is transported up to the Axios River. It then passes Axios River via a 1.5 km long Axios siphon, flowing to the pump room of Sindos, through an 8.5 km closed conduit. From there, it is pumped up to the installations of water treatment (Refineries), through a 4.7 km pressurized pipe.

The process followed in the treatment plant (Figure 4.5) is shown in Figure 4.6.



FIGURE 4.3: Thessaloniki urban and metropolitan areas map. Licensed under CC BY-SA 3.0 via Commons.



FIGURE 4.4: Aliakmon Dam. Figures obtained from EYATH.

The ozonation process, which aims to break down organic compounds and facilitate their adsorption on the activated carbon, also serves as a first disinfection step. Final disinfection with chlorine is applied for residual effectiveness [142].

Clean potable water flows to a reservoir with a capacity of 75,000 m³, from where it is distributed, via various conduits adding up to 36 km long. The existing water supply reservoirs are located in Diavata, Eyosmos, Polixni, Neapoli, Vlatades, Toympa and Kalamaria (Figure 4.7).

The water distribution network has 2,200 km of length, 48 pumping stations, SCADA



FIGURE 4.5: Thessaloniki's main water treatment plant. Figures obtained from Special Service for Water Supply and Sewerage of Thessaloniki (E.Y.D.E. Thessalonikis).

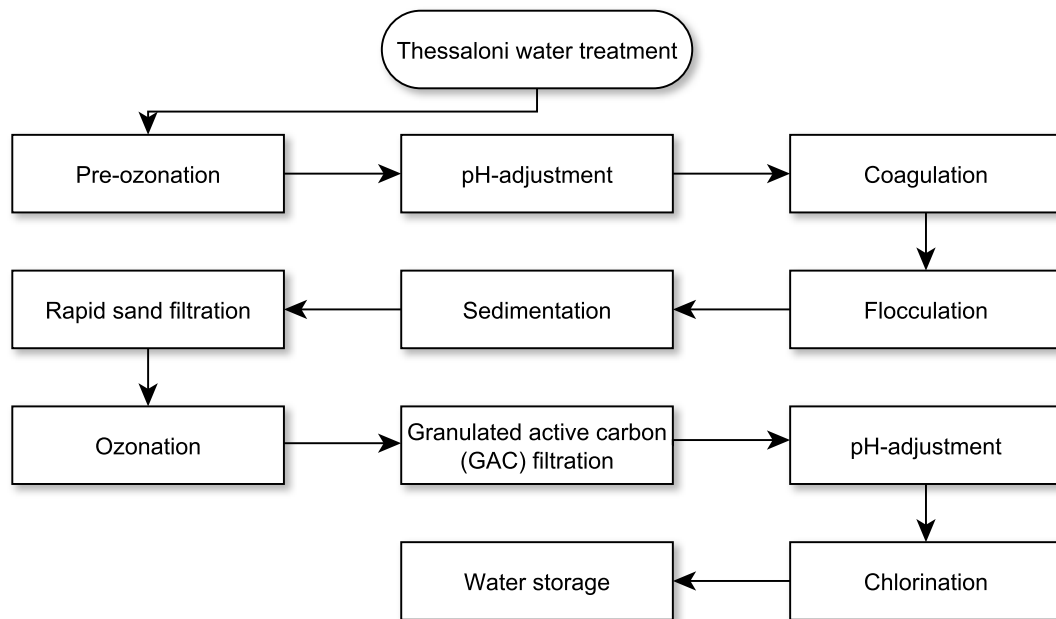


FIGURE 4.6: Thessaloniki's water treatment process.

surveillance system, and 510,000 water supply connections. The public “asset” company (EYATH Fixed Assets) owns the infrastructure for water abstraction works, pumping stations and wells and conveyance networks [141].

4.2.1 Sampling protocol

To get biofilm samples from the DWDS of Thessaloniki we closely worked with the EYATH operators.

To this end we accompanied them every time they changed a pipe, either because of leakage or renewal purposes. Thus, the sampling has been based on availability and not



FIGURE 4.7: Water supply network reservoirs. Figure obtained from Special Service for Water Supply and Sewerage of Thessaloniki (E.Y.D.E. Thessalonikis).

on a structured survey. The working schedule developed is based on the pipe cut-out sampling procedure explained previously in Subsection 3.2.1 (Figure 3.6). However, prior to start the sampling some tests were performed in order to decide the best protocol for the sampling.

- Grid area: after a literature review two areas were described of 30 [143] and 4 [123] cm^2 . After testing the different areas the grid of 4 cm^2 was chosen since the amount of collected sample from different pipe materials were suitable to our purpose.
- Effectiveness of swabs removal method: after testing the removal of biofilm from different pipes of different materials it was observed that five is the best number of swabs to be used in a 4 cm^2 grid to exhaust the sample (Figure 4.8).
- Solution volume: a volume, between 10 and 15 ml of sterile water, was used to place the swabs after sampling. It was observed that that volume allowed to obtain a good number of colonies in the plates without too many dilutions.



FIGURE 4.8: Detail of the sampled area in a plastic pipe. Eva Ramos Martínez ©.

During the sampling process, after removing a pipe section of 60 cm length approximately, we used sterilized gloves and materials to avoid as much as possible potential external contamination (Figure 4.9). The sampling grid has been laid on the inner pipe wall, trying to place it as far as possible from the sides of the pipe section to avoid contamination due to the manipulation of the pipe. When swabbing the same movements were carried out each time in order to maintain a constant swabbing effort (covering the area with the swabs 5 times in each direction). The swabs were placed in a flask with 10 - 15 ml of sterile water and transported in a cold box to the laboratory. The samples were processed as soon as possible after the collection to minimize changes in bacterial population. In no case the recommended maximum elapsed time between collection and examination of samples of 8h (maximum transit time 6h; maximum processing time 2h) [12] was exceeded.

In the laboratory, the flasks with the samples were vortexed during 3 minutes at maximum speed. After that, using serial dilutions in sterile water, the samples were cultured by the spread plate method in previously prepared R2A agar plates (20 ml). An aliquot of 0.5 ml of the sample was then added onto the surface of the pre-dyed R2A agar plate and spread until the inoculum was completely absorbed by the medium. Each dilution examined was tested in duplicate and colonies were counted after incubation at 25°C for 7 days. Counting was carried out by manual counting with appropriated illumination. Only plates having 30 to 300 colonies/plate were considered [12]. If there was no plate with 30 to 300 colonies, the plate having a nearest count was used. If the number of



FIGURE 4.9: Biofilm sampling in Thessaloniki drinking water distribution system. Eva Ramos Martínez ©.

colonies per plate exceeded 300, the plate was divided with a marker in equal parts (4 parts) having representative colony distribution and the number of colonies just in one of the sections was counted, multiplying the number of colonies obtained by the number of sections made.

To compute the heterotrophic plate count per unit surface area (CFU/cm^2), the average number (duplicate plates of the same dilution) of CFU per plate is multiplied by the reciprocal of the dilution used and divided by the volume of the aliquot pipetted to get the CFU/ml (Equation 4.1). Then, it is multiplied by the volume of the solution and divided by the area of the grid (Equation 4.2).

$$CFU/ml = \frac{CFU/plate \cdot \text{Dilution factor}}{\text{Aliquot (ml)}} \quad (4.1)$$

$$CFU/cm^2 = \frac{CFU/ml \cdot \text{Solution volume (ml)}}{\text{Grid area (cm}^2\text{)}} \quad (4.2)$$

Two sampling campaigns were carried out. The first in the summer of 2013 and the second in the winter of 2014. During this time we contacted the EYATH, prepared and

optimized the sampling protocol, carried out the sampling and dealt with the setbacks.

Since the Thessaloniki's hydraulic model is not fully developed yet it was not possible to know the actual hydraulic conditions of the pipes. However, during the sampling it was tried to gather the maximum amount of information (Table 4.1), specifically regarding pipe material and age, two parameters that have been found to be relevant in biofilm development [14]. The rough surfaces protect the biofilm from detachment and provide greater area for protection and colonization [84]. Roughness varies within pipe materials and pipe age. Since accumulation of corrosion products and dissolved substances in the pipes increase with time, also pipe roughness does [144].

When it was possible, samples of the circulating water were also obtained and several physico-chemical parameters measured (see Table 4.1). Chlorine concentration and water temperature are the main parameters that we focused on. It is known that a low concentration of disinfectant reduces stress on biofilm and temperature favours bacterial growth [145].

4.2.2 Descriptive analysis

Generally, data analyses start describing the data, and then move to the exploratory, inferential, predictive, causal, and mechanistic analysis, thus increasing difficulty and complexity. If the data set (Table 4.1) has few observations the combination of the input values is limited. This biases the probability of finding relationships among these values. Taking into account the characteristics of the data set and the fact that the significance of a relation depends on the sample size, only a descriptive analysis was performed on the data (Table 4.1).

In Table 4.2 the main characteristics of the data set variables are shown. In Figure 4.10 the values of the two replicates made for of each sample are presented. As we expected, no big differences are observed between them.

When comparing the biofilm data obtained with some of the variables (Figure 4.11) it is observed that the highest biofilm development corresponds to the asbestos cement (AC) pipe, which is also the one with larger diameter. Despite of this, no notable differences

TABLE 4.1: Data from the sampling in the DWDS of Thessaloniki (CI: Cast iron; PVC: Polyvinyl chloride; AC: Asbestos cement).

Sample	HPC/cm ² average	HPC/cm ² SD	Diameter (mm)	Pipe material	Pipe age (years)
A	9.25E+01	12.021	200	CI	30
B	2.49E+02	1.414	100	CI	30
C	3.62E+02	76.986	110	PVC	10
D	9.60E+01	8.485	110	PVC	10
E	2.95E+02	192.510	110	PVC	10
F	1.21E+03	331.368	300	AC	30
G	2.54E+02	89.873	160	PVC	10

Sample	Location	Total Cl (mgCl/l)	Free Cl (mgCl/l)	pH	Temperature (°C)
A	Thessaloniki	0.29	0.20	7.54	26.65
B	Thessaloniki	0.22	0.10	7.52	25
C	Pavlos Melas	NA	NA	NA	NA
D	Kordelio-Evosmos	NA	NA	NA	NA
E	Pylea	NA	NA	NA	NA
F	Kordelio-Evosmos	0.35	0.19	8.2	21.1
G	Thessaloniki	NA	NA	NA	NA

TABLE 4.2: Main characteristics of the data attributes.

Sample	HPC/cm ²	HPC/cm ² SD	Diameter (mm)	Pipe material	Pipe age (years)
A:1	Min. : 92	Min. : 1	Min. :100	AC :1	Min. :10.0
B:1	1st Qu.: 172	1st Qu.: 10	1st Qu.:110	CI :2	1st Qu.:10.0
C:1	Median : 254	Median : 77	Median :110	PVC:4	Median :10.0
D:1	Mean : 366	Mean :102	Mean :156		Mean :18.6
E:1	3rd Qu.: 328	3rd Qu.:141	3rd Qu.:180		3rd Qu.:30.0
F:1	Max. :1215	Max. :331	Max. :300		Max. :30.0
G:1	NA's :0	NA's :0	NA's :0	NA's :0	NA's :0

Sample	Location	Cl total (mgCl/l)	Cl free (mgCl/l)	pH	Temperature (°C)
A:1	Kordelio-Evosmos:2	Min. :0.2	Min. :0.1	Min. :7.5	Min. :21.1
B:1	Pavlos Melas :1	1st Qu.:0.3	1st Qu.:0.1	1st Qu.:7.5	1st Qu.:24.1
C:1	Pylea :1	Median :0.3	Median :0.2	Median :7.5	Median :27.0
D:1	Thessaloniki :3	Mean :0.3	Mean :0.2	Mean :7.8	Mean :25.6
E:1		3rd Qu.:0.3	3rd Qu.:0.2	3rd Qu.:7.9	3rd Qu.:27.8
F:1		Max. :0.4	Max. :0.2	Max. :8.2	Max. :28.6
G:1	NA's :0	NA's :4	NA's :4	NA's :4	NA's :4

are observed in biofilm development regarding the other pipe materials, diameters and ages. The same was observed when focusing on the sampling location.

4.3 Case study 2. Pennine Water Group pilot distribution system in Sheffield, United Kingdom

Keeping in mind the main objective of this work and to avoid all of the issues found when sampling biofilm in real DWDSs we sampled in the Pennine Water Group pilot

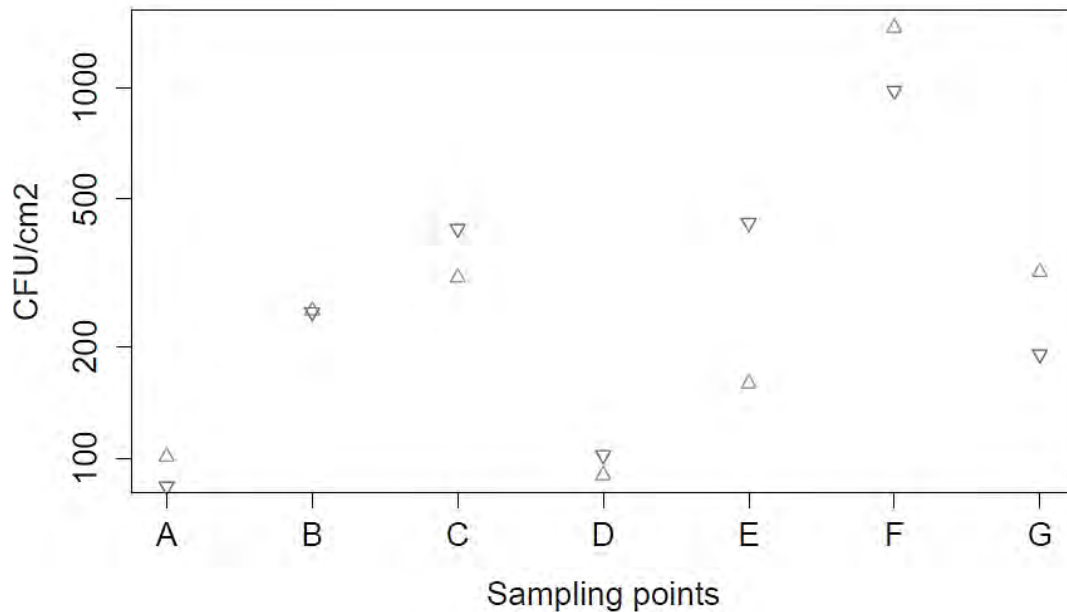


FIGURE 4.10: Data obtained in each replicate and sampling point.

distribution system.

After a literature review, we found that the Pennine Water Group (PWG) experimental facility satisfied all our requirements. Thanks to the inestimable help of Prof. Joby Boxall and his team, and the scholarship granted by the Spanish Ministry of Economy and Competitiveness (Ref.: EEBB-I-14-09135) we conducted a biofilm sampling protocol in the University of Sheffield during the summer of 2014.

Sheffield is a city located in South Yorkshire, England, UK (Figure 4.12). In 2011, Sheffield had a population of 551,800 inhabitants, approximately. It is part of the wider Sheffield urban area, which has a population of 640,720 inhabitants.

Water treatment and supply are run by the Yorkshire Water Services (YWS). YWS manages the collection, treatment and distribution of water in Yorkshire. It is a big company that provides 1.24 billion litres of drinking water every day across Yorkshire. It operates more than 700 water and sewage treatment works and 120 reservoirs. The University of Sheffield is located within the Loxley 2004 Water Supply Zone (Figure 4.13). The water supplied to the zone is classified as being soft to moderately soft

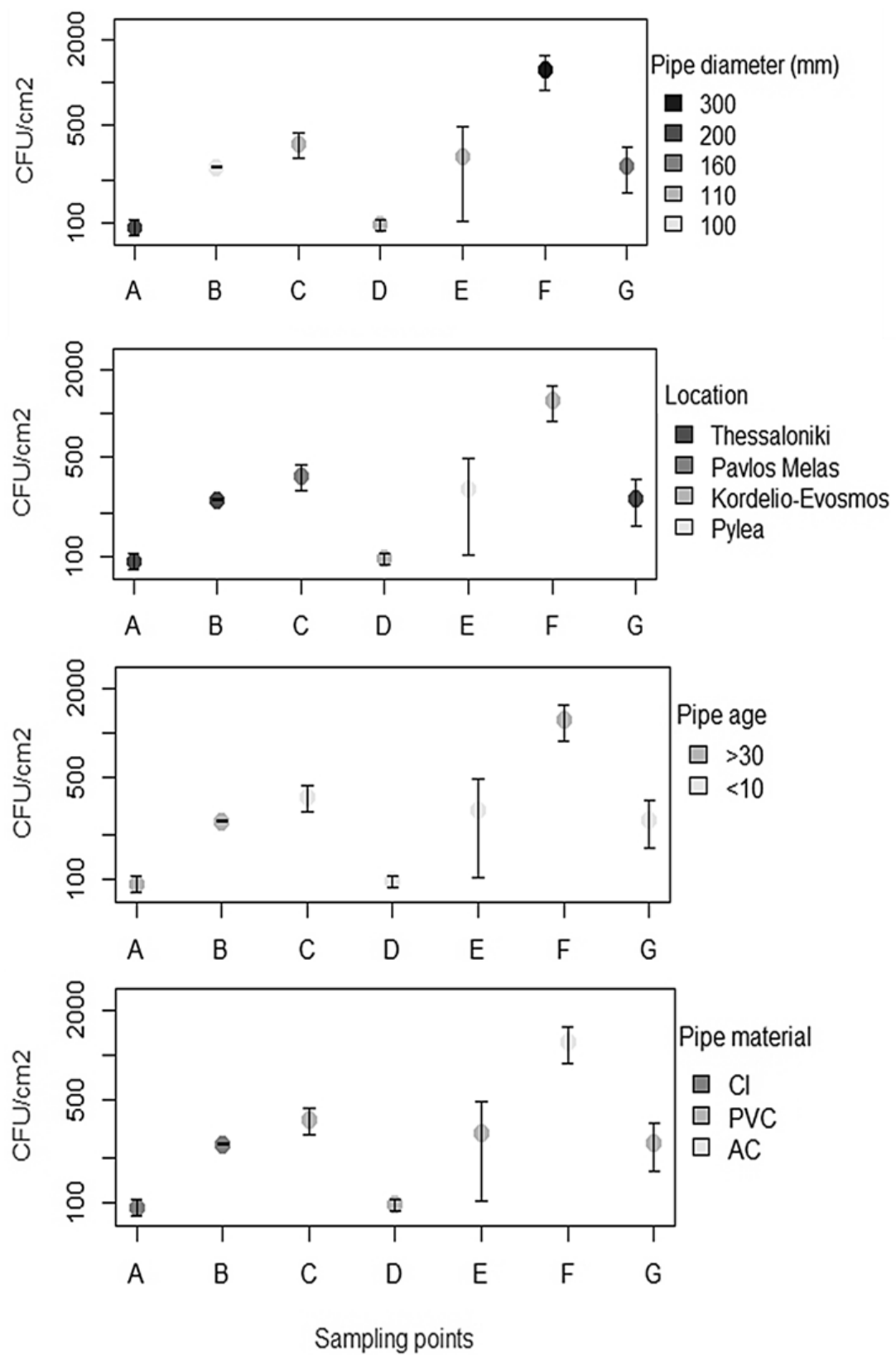


FIGURE 4.11: Scatter-plots of the biofilm data obtained in the DWDS of Thessaloniki.



FIGURE 4.12: Location of Sheffield city, in South Yorkshire, England, UK.

water, which is river/reservoir derived. The zone is predominantly fed from Loxley Water Treatment Works, although sometimes can also be fed from Ewden Water Treatment Works or Rivelin Water Treatment Works. Below we give a process overview of the water treatment process at Loxley Water Treatment Works.

1. Clarification process that includes dissolved air flotation. This process uses ferric sulphate ($Fe_2(SO_4)_3$) as the coagulant chemical and lime for pH adjustment.
2. Rapid gravity filtration with lime for pH adjustment.
3. Addition of monosodium dihydrogenphosphate (MSP, NaH_2PO_4) for plumbosolvency control.
4. Secondary filtration through manganese contactors with addition of chlorine and lime for pH adjustment.
5. Final treatment with chlorine addition

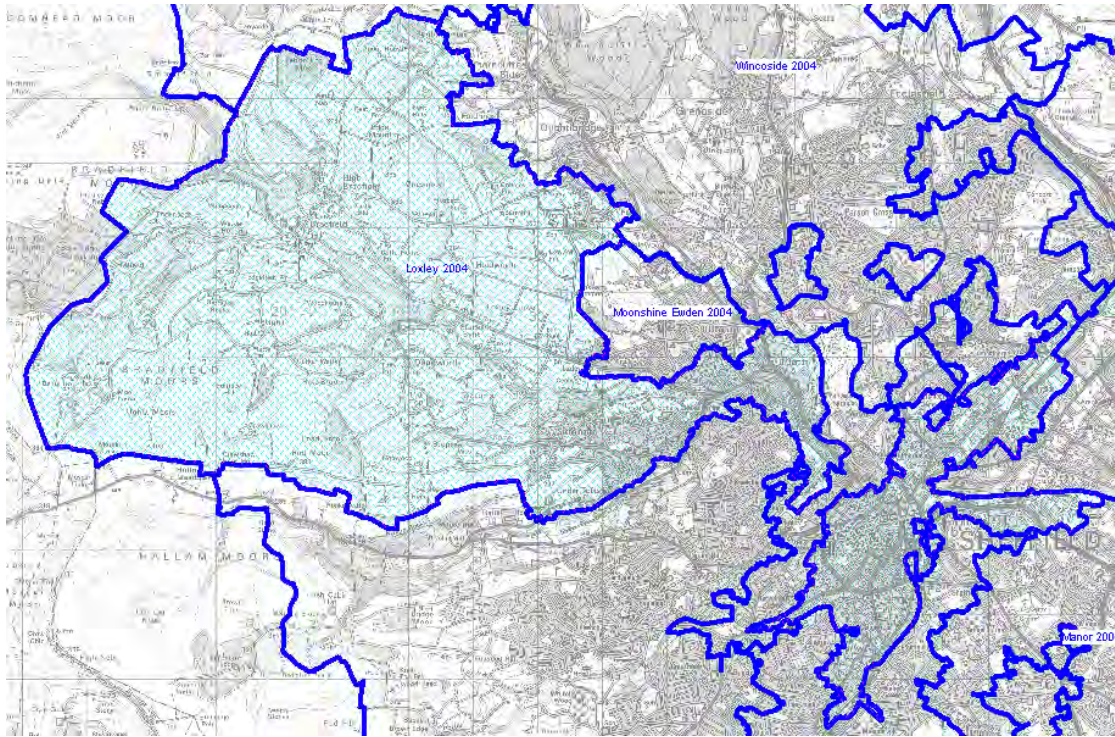


FIGURE 4.13: The geographical area covered by Loxley 2004 Water Supply Zone. COPYRIGHT STATEMENTS: Based upon Ordnance Survey map data with the permission of the Controller of Her Majesty's Stationery Office, ©Crown Copyright. Licence No. 100019559. Copyright for additional WSZ boundary data shown on this map rests with Yorkshire Water.

In April 2005 Yorkshire Water Services established a strategic 5 year partnership with the Pennine Water Group (PWG) at the University of Sheffield, to meet YWS research needs associated with all aspects of buried infrastructure assets. The PWG is an Engineering and Physical Sciences Research Council (EPSRC) funded Platform Grant centre dedicated to research into water and wastewater. It is headed by a Management Team including Prof. Boxall.

They apply a multidisciplinary approach to study DWDSs, from fundamental to applied aspects of drinking water infrastructure, with a particular emphasis on full scale representative laboratory based facilities. Their work is complemented with field experiments in operational systems with water company partners. Their research into water quality focuses on discolouration phenomena, chlorination regimes and biofilm physical and community structure.

4.3.1 Pennine Water Group's experimental facility and operating conditions

The experimental facility consists of three independent recirculating loops (Figure 4.14). Each loop is fed by a pump and returns to a reservoir (Figure 4.15). Thus, flow in each loop is individually controlled to generate different hydraulic regimes. Unlike bench scale experiments the full scale pipe surface area of the test loop facility enables fully realistic exchange processes and interactions between the bulk fluid and the pipe wall to occur, thus replicating realistic conditions in typical DWDSs [146].

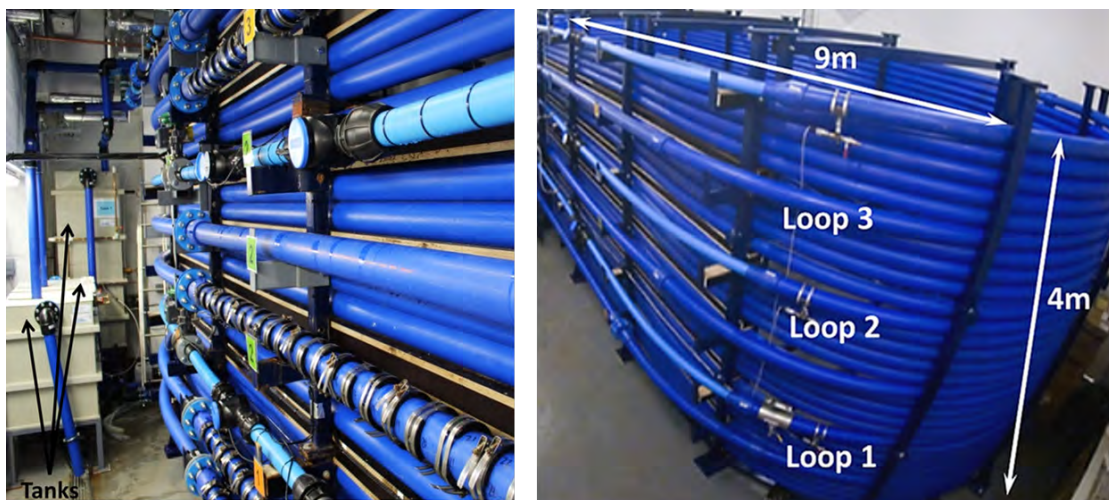


FIGURE 4.14: Pennine Water Group's experimental facility. Images borrowed from Dr. Katherine Fish, PWG, University of Sheffield.

Loops are made of High-Density Polyethylene (HDPE) pipe. Polyethylene pipe was selected as it is a prevalent and representative current material used in distribution systems world-wide. In order to provide representative water quality the facility is fitted with a trickle feed (and drain) from the local water distribution system [10].

Before experiments commenced, the facility was disinfected with 20 mg/l of RODOLITE H (RODOL Ltd, Liverpool, UK), which is a solution of sodium hypochlorite with less than 16% free available chlorine. The system was flushed for 3 turnovers at the maximum attainable flow rate (7 l/s) and left standing for 24 h. After that period the system was flushed again at the maximum flow rate with fresh water until the levels of chlorine were similar to those of the local tap water. The PWG coupon design (Figure 4.16)

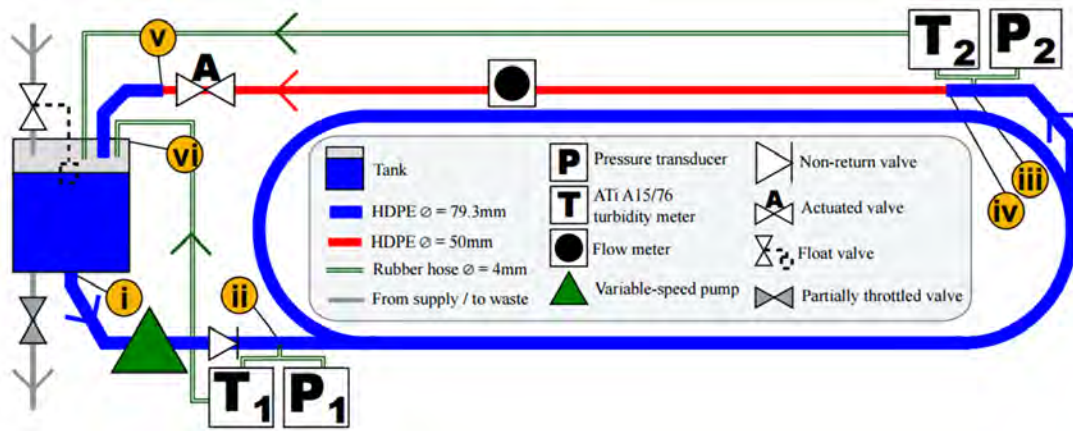


FIGURE 4.15: Schematic of each pipe loop. Figure obtained from [9].

allows direct insertion and close alignment with the internal pipe surface minimizing the distortion of boundary layer conditions that influence biofilm formation, such as boundary shear stress and turbulent driven exchange with the bulk water body. The facility thus allows the formation, growth, and detachment of biofilms to be captured under controlled but fully realistic conditions [10].

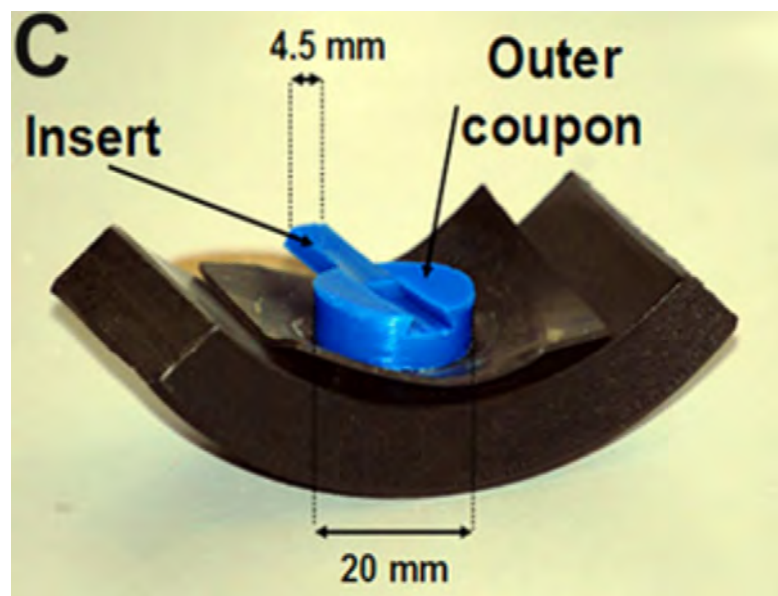


FIGURE 4.16: Pennine Water Group coupon showing outer coupon (surface area 224 mm²) with 1 insert (surface area 90 mm²). Figure obtained from [8].

Investigating the microbiological component of the pipe wall material was achieved by arbitrary fitting along and around the sample length of each pipe loop with PWG coupons,

as described in [8]. Coupons were fitted along the length of each pipe loop to facilitate the pipe wall microbiological studies (Figure 4.17).

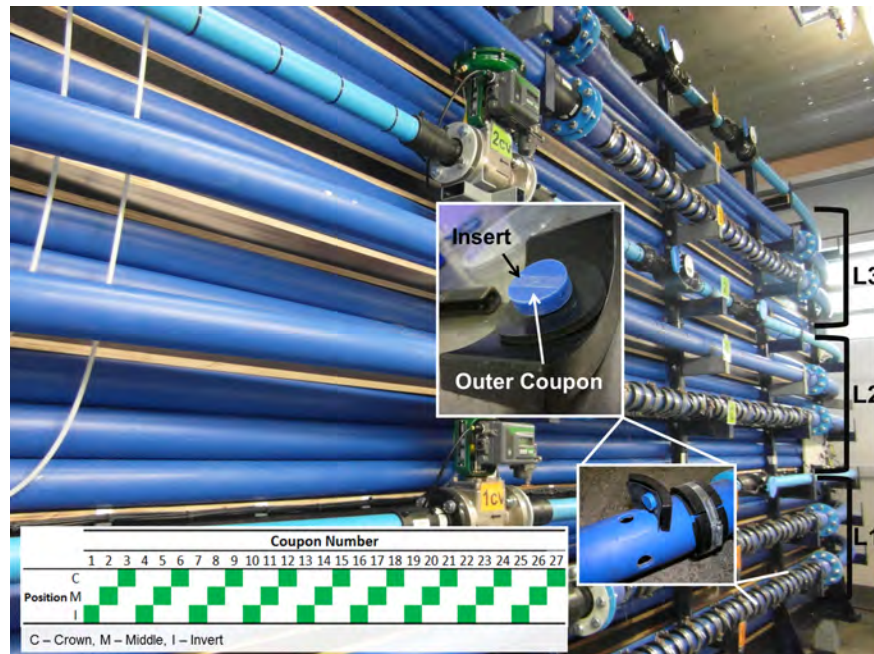


FIGURE 4.17: Coupons location in the pipe loop. Image borrowed from Dr. Katherine Fish, PWG, University of Sheffield.

During the growth phase, water partially recirculated within each system but the outflow to waste from each tank and the inflow from supply were controlled to give a 24-hour hydraulic residence time. This prevented stagnation and provided constant renewal [9].

The room temperature was controlled at 16°C, as fluctuations in temperature can have an important effect on bacterial growth. This temperature is representative of average spring and summer temperatures in UK DWDSs, thus accurate for real systems and providing maximum representative levels of microbial activity [10]. System monitoring, control and data logging were all automated. Flow is modulated using a valve. During the experiment a low varied flow (LVF), ranging from 0.2 to 0.5 l/s, was applied based on daily patterns observed in real DWDSs in the UK [147]. This LVF daily regime (Figure 4.18) was repeated for a growth phase of 28 days in loop 2 (L2), 29 days in loop 3 (L3), and 30 days in loop 1 (L1). The one day difference between growth phases is due to technical issues and limitation of time to process the coupons.

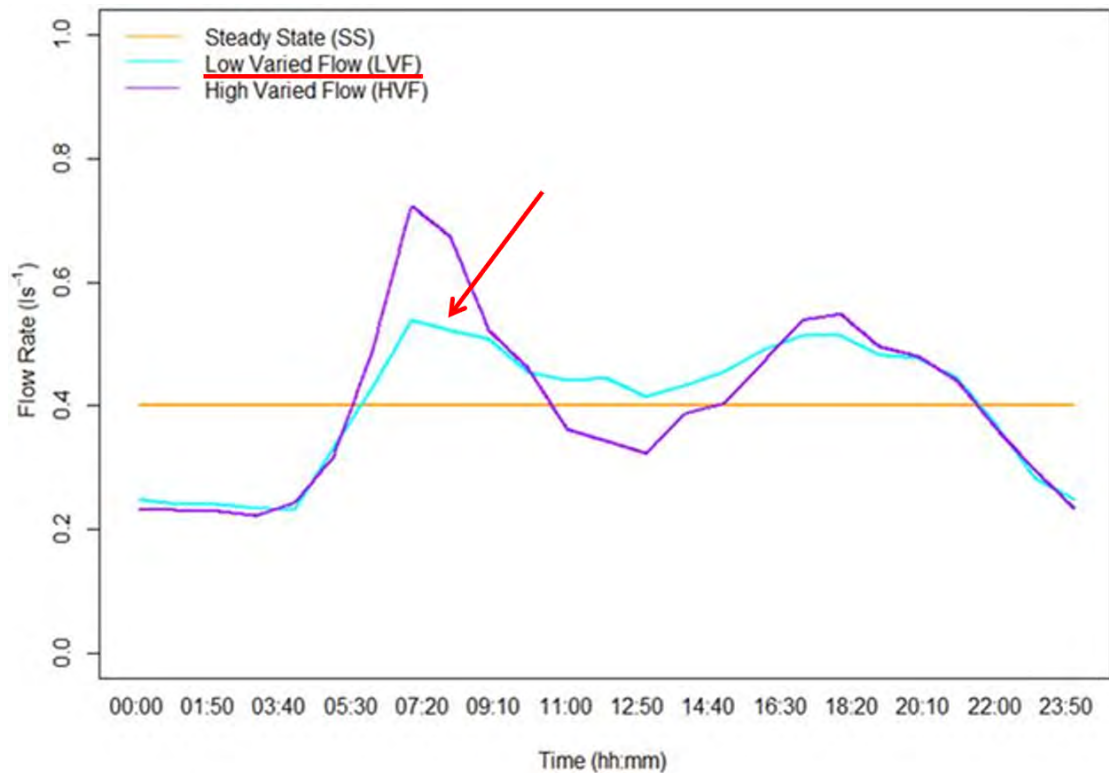


FIGURE 4.18: The three different hydraulic regimes based on daily patterns observed in real DWDS in the UK. Figure obtained from [10].

During the experiment the effect of chlorine concentration on biofilm development was considered also interesting to be tested. L1 was the control loop. The water flowing within it was not altered. The L2 was the loop where a chlorine boost was promoted by adding sodium hypochlorite to the local water distribution system before arriving to tank 2. In L3, instead, dechlorination was promoted by adding sodium ascorbate.

4.3.2 Biofilm sampling

After 28 days coupons were sampled ($n = 9$, three per loop, one from each position). All biofilm samples were collected without draining the system, limiting the impact of sampling upon biofilm accumulation, and replacement sterile coupons were immediately inserted into the pipe [121]. Inserts and outer coupons were separated aseptically (Figure 4.19). Biofilm was removed from all outer coupons ($n = 9$) by brushing (30 horizontal/vertical strokes) into 30 ml sterile PBS [121]. The resulting suspensions were processed immediately after sampling. They were vortexed for 3 minutes at 24x100

rpm. After making the serial dilutions, the examined dilutions were cultured in triplicate. After 7 days at 22°C the colonies forming units (CFU) were counted. The number of heterotrophic viable bacteria found in the samples is reported.

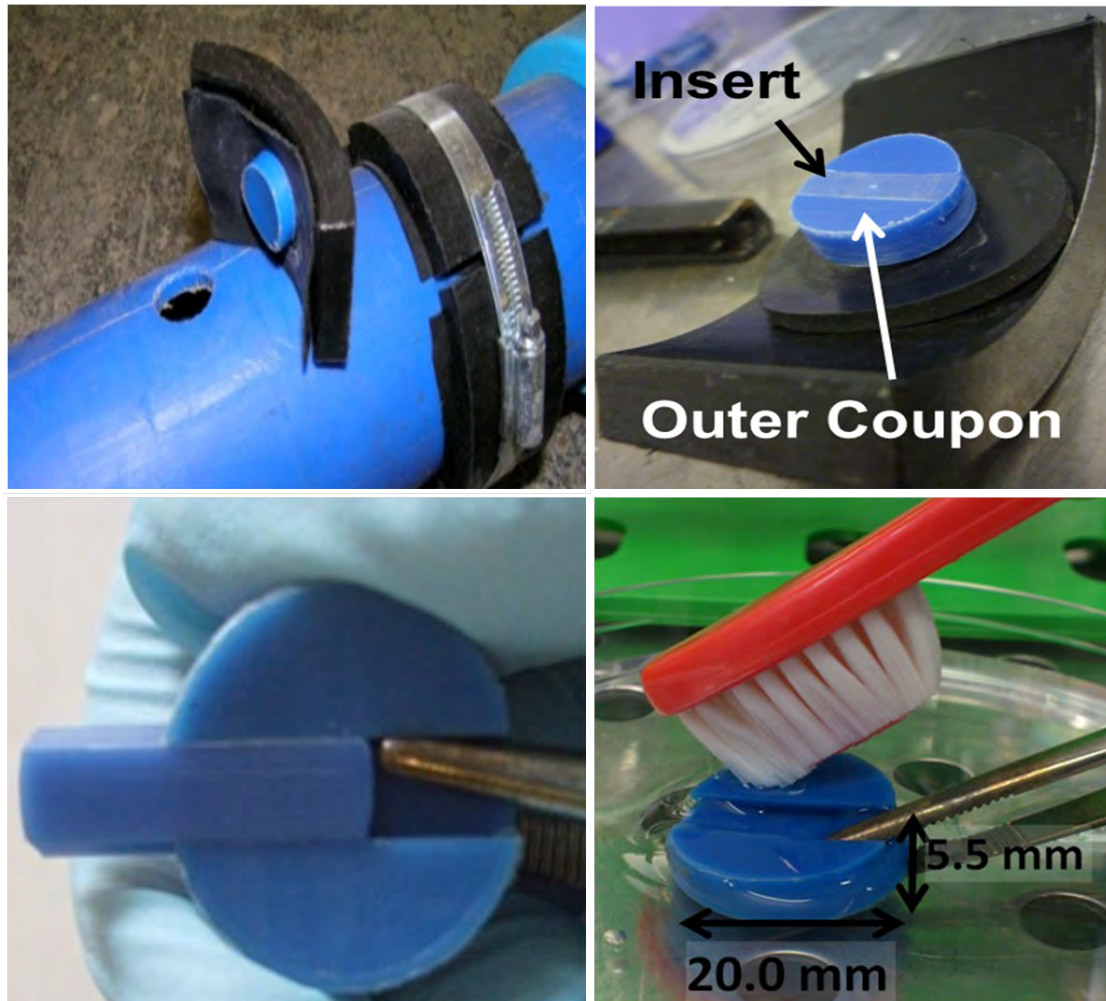


FIGURE 4.19: Biofilm sampling in PWG experimental facility. Images borrowed from Dr. Katherine Fish, PWG, University of Sheffield.

During the experiment, samples from the circulating water were taken from each loop. Free chlorine, total chlorine, water temperature and pH were measured in each loop three times per week, while the concentration of iron (Fe), phosphorus (P) and total organic carbon (TOC), only once a week. The obtained values are presented in Table 4.3.

There are no CFU/cm² values from L2 due to a strange bacteria proliferation growing like a slim that made the plates uncountable (Figure 4.20). This issue was also found

TABLE 4.3: Data from PWG experimental facility.

	CFU/cm ²		Free Cl (mgCl/l)		Total Cl (mgCl/l)		Temperature (°C)	
	Average	SD	Average	SD	Average	SD	Average	SD
L1	1.38E+06	5.66E+05	0.31	0.07	0.44	0.09	14.67	0.26
L2	NA	NA	0.70	0.23	0.79	0.24	14.65	0.24
L3	2.23E+07	7.73E+06	0.06	0.05	0.15	0.10	14.59	0.58

	pH		Fe (µg/l)		P (µg/l)		TOC (mg/l)	
	Average	SD	Average	SD	Average	SD	Average	SD
L1	6.90	0.89	46.59	17.25	1251.33	34.37	1.67	0.34
L2	6.95	0.75	60.79	41.26	1248.67	29.59	1.56	0.06
L3	7.16	1.36	53.47	24.42	1253.33	27.39	1.61	0.08

when culturing the inlet water from the deposits.

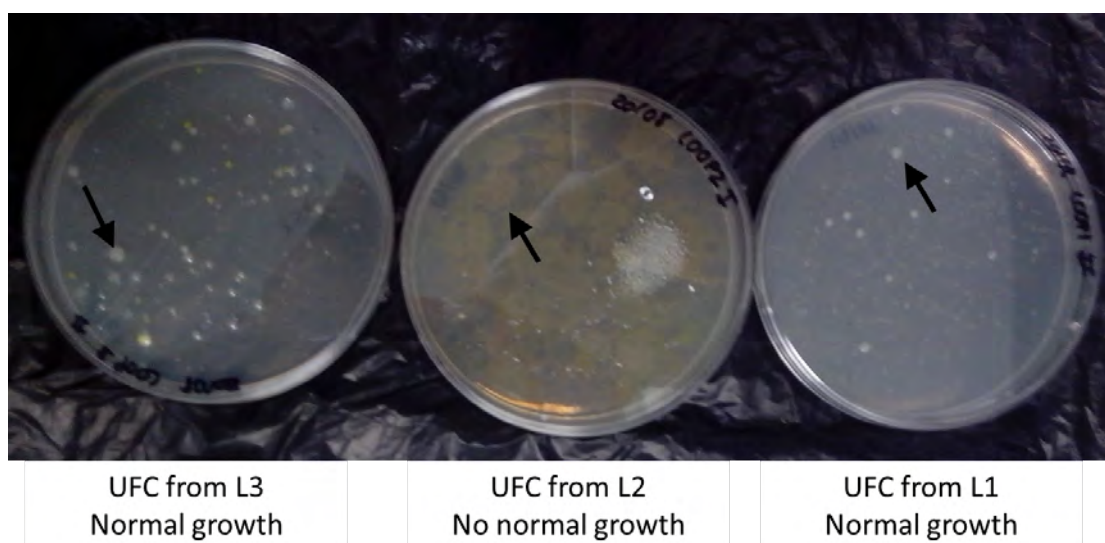


FIGURE 4.20: Bacteria growth in the R2A agar plates. Eva Ramos Martínez ©.

This problem seemed to be due to a bad odour generated by mucilaginous bacteria. This bacteria's high proliferation did not allow the formation of single colonies, therefore, no CFU could be reported. We isolated it for identification. It was observed that it was blue-green fluorescent under ultraviolet (UV) light (Figure 4.21) but Gram staining [148] did not result in conclusive results. Thus, the bacterial 16S ribosomal RNA (rRNA) gene was amplified in order to identify it. Sequence analysis of the 16S rRNA gene is a powerful mechanism for identifying new pathogens. Sequencing of the 16S rRNA gene serves as an important tool for determining phylogenetic relationships among bacteria.

The features of this molecular target make it a useful phylogenetic tool, for bacterial detection and identification [149].

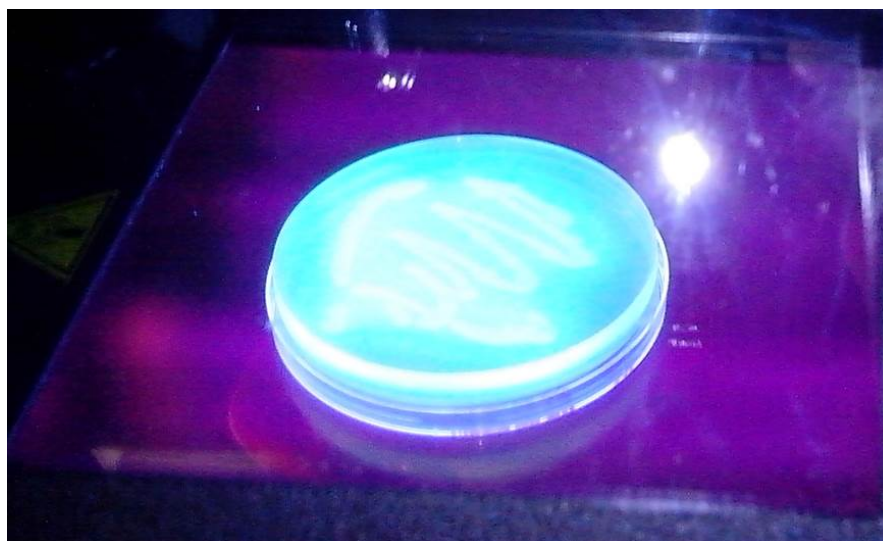


FIGURE 4.21: Isolated bacteria under UV light. Eva Ramos Martínez ©.

Polymerase chain reaction (PCR) amplification was performed using the 63F and 1387R primers [150]. The thermal cycler was programmed to perform 30 cycles consisting of 95°C for 1 min, 55°C for 1 min, and 72°C for 1.5 min followed by a final extension step of 5 min at 72°C. PCR products were visualized by agarose gel electrophoresis (Figure 4.22).

To achieve the minimum PCR DNA concentration needed to send the samples for sequencing (50 ng/ μ l), the DNA was concentrated using the Colum Filter Kit (Cat- No : BIO 52058 Bionline). For Sanger sequencing of bacterial 16S rRNA genes the extracted DNA was sent to the Core Genomic Facility at the Medical School of the University of Sheffield that utilises the Applied Biosystems 3730 DNA Analyser.

Sequences were visualized using the FinchTV software version 1.4.0 (<http://www.geospiza.com/finchtv.html>) and aligned with Clone Manager 9.0 (Sci-Ed Software, Cary, North Carolina, USA). The similarity search, to identify the sequence based on sequence homologous, was performed using the blast program of the National Center for Biotechnological Information (NCBI) (<http://blast.ncbi.nlm.nih.gov.com>). The bacteria was identified as *Pseudomonas* genus.

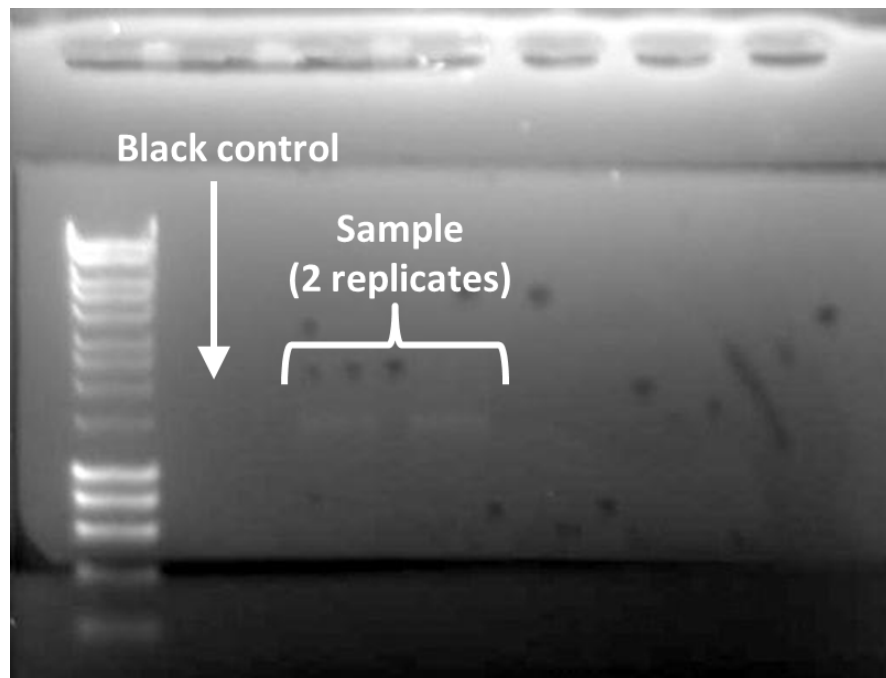


FIGURE 4.22: PCR products visualized by agarose gel electrophoresis. Eva Ramos Martínez ©.

In a previous work it has been observed that the genus *Pseudomonas* was the predominant genus in biofilm composition. Particularly, at LVF conditions, with a relative abundance up to 65% [10]. This suggests that species belonging to genera *Pseudomonas* have an enhanced ability to express extracellular polymeric substances to adhere to surfaces and to favour co-aggregation between cells. It was observed that the percentage of the bacterial genera changed between hydraulic conditions but not clear variation trend has been found [10]. The factors that promoted the high development of this bacteria observed in the samples remain unknown.

Chapter 5

Getting and pre-processing data

Getting field data is an arduous task which requires high workload and time, while developing experimental laboratory studies is still very complex and requires highly qualified staff and equipment. In both cases the time needed tends to be too long and the amount of data obtained scarce. This constraint on obtaining data is a handicap for researches. It slows down the process of obtaining results, reducing the competitiveness compared to other research fields.

Given the high difficulty of studying the whole system influence on biofilm development and the fact that the real operating conditions of the pipes are rarely known and the hydraulic conditions at biofilm scale are still being discussed, we apply an innovative approach. We change the commonly used approaches in DWDSs' biofilm studies towards the implementation of data science techniques, innovative discipline in this field, in order to develop a practical tool for DWDS managers. The combination of various existing data sets on similar studies to conduct a meta-data analysis of biofilm development is proposed in order to cover the study of the environment influence through partial views of the problem.

5.1 Data collection

Currently, we have technology and data of great quality to support new research approaches. Data acquisition has been carried out through an exhaustive search and an intensive personal and institutional networking. Some of the contacted professionals and institutions are listed in Table 5.1.

TABLE 5.1: Contacts made during the personal and institutional networking.

Expert	Institution	Country
Prof. Jean Claude Block	Nancy University	France
Prof. Laurence Mathieu	University of Lorraine	France
Prof. Joby Boxall	University of Sheffield	United Kingdom
Prof. Efthymios Darakas	Aristotle University of Thessaloniki	Greece
PhD Sean McKenna	IBM	Ireland
PhD Noel Muñoz Soto	University of Valle Cinara Institute	Colombia
PhD Sharon A. Waller	Northwestern University	United States
MsC Maria Ximena Trujillo Gomez	University of los Andes	Colombia

Biofilm data have been collected from previous research works of biofilm development in DWDSs (Table 5.2). The journal papers included in 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 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/cm²” and “R2A”, and the various combinations among them. The papers found under these criteria have been studied to be included in the data compilation.

All the measurements associated with HPC/cm² biofilm data have also been compiled. A letter has been assigned to each studied paper and a number to each reported case in order to create a key attribute. At the beginning just the following cases have been discarded.

1. Studies based on cultured communities seeded with investigator-selected species or developed using an inoculum.

TABLE 5.2: Journal papers used as data sources.

Id	Cases	Year	Author	Journal
A	A1-A8	2007	Manuel et al.	Water Research
B	B1-B6	2003	Batt et al.	Water Research
C	C1-C24	2002	Momba & Binda	Journal of Applied Microbiology
D	D1-D16	2005	Ndiongue et al.	Water Research
E	E1-E12	2008	Sylvestry-Rodriguez et al.	Applied and Environmental Microbiology
F	F1-F16	1999	Volk & LeChevallier	Applied and Environmental Microbiology
G	G1-G18	2004	Wingender & Flemming	Water Science & Technology
H	H1-H44	2000	Zacheus et al.	Water Resources
I	I1-I18	2005	Chu et al.	Journal of Environmental Management
J	J1-J15	2004	Lehtola et al.	Water Research
K	K1-K69	2004b	Lehtola et al.	Journal of Industrial Microbiology & Biotechnology
L	L1-L109	2006	Tsvetanova	Chemicals as Intentional & Accid. Global Env. Threats
M	M1-M6	2003	Schwartz et al.	Journal of Applied Microbiology
N	N1-N18	1998	Percival et al.	Water Resources
O	O1-O12	2011	Jang et al.	Microbiological Biotechnology
P	P1-P20	1998	Momba et al.	Water Science Technology
Q	Q1-Q29	2003	Ollos et al.	Journal AWWA
R	R1-R37	2002	Boe-Hansen et al.	Water Supply Research and Technology
S	S1-S38	2001	Hallam et al.	Water Research
T	T1-T30	1998	Ollos	Ph.D. dissertation, University of Waterloo, Ontario
U	U1-U30	2005	Gagnon et al.	Water Research
V	V1-V2	2013	Gosselin et al.	Water Research
W	W1-W17	2012	Jang et al.	The Journal of Microbiology
X	X1-X30	1999	Percival et al.	Industrial Microbiology and Biotechnology
Y	Y1-X16	2007	N. Muñoz Soto	Ph.D. dissertation, University of Valle Cinara Institute

2. Biofilm developed on unrepresentative materials for DWDSs. The use of glass coupons within reactors is very common.
3. Cases where the quality of the water was 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, was used as secondary disinfectant. This restriction has been applied since the European Union has issued standards for drinking water, and these standards do not require disinfection [151]. Disinfection practices vary widely in European countries, being the previously mentioned the two mainly used.

In this first step, nearly 600 data of biofilm, with their associated variables, have been compiled from 25 different works that study biofilm development in DWDSs. After the literature compilation, the obtained data and their source have been carefully checked. At this point, the framework of the compilation has been reduced.

Since the aim is to have an idea of the global conditions in DWDSs to predict their effect on biofilm development, it has been decided to remove the cases where synthetic water has been used. That is, manipulated drinking water, where some chemical elements are removed from the water and afterwards artificially added. After carefully studying this procedure, it was decided to eliminate these papers from the data set.

In this way, it is assured that the results are representative of the complex, multi-species communities that develop naturally in DWDSs. The rest of the papers have been discarded for reasons related with the methodology used to assess the HPC/cm², which differ from the recommendations suggested in [152] for R2A agar long incubation, i.e., 5 to 7 days of incubation between 20 or 28 °C. The papers removed are listed in Table 5.3.

TABLE 5.3: Removed papers.

Id	Cases	Year	Author	Reason
B	B1-B6	2003	Batt et al.	Synthetic water
D	D1-D16	2005	Ndiongue et al.	Synthetic water
E	E1-E12	2008	Sylvestry-Rodriguez et al.	Synthetic water
I	I1-I18	2005	Chu et al.	Synthetic water
Q	Q1-Q29	2003	Ollos et al.	Synthetic water
R	R1-R37	2002	Boe-Hansen et al.	Incubation at 15°C
T	T1-T30	1998	Ollos	Synthetic water
U	U1-U30	2005	Gagnon et al.	Synthetic water
V	V1-V2	2013	Gosselin et al.	Incubation during 14 days
Y	Y1-Y16	2007	M. Muñoz Soto	No R2A agar

Bacteria in DWDSs transit from planktonic growth to the stage of irreversible attachment, from irreversibly attached cells to the stage of mature biofilms, and the transition from mature-stage biofilm to the dispersion stage. These processes are not necessarily synchronized throughout the entire biofilm [153]. Due to the scope of this work we are only interested in “mature” biofilm. Well developed biofilm increases the cell adhesion rate [154], while individual microcolonies may detach from the surface or may give rise to planktonic revertants that swim or float away from these matrix-enclosed structures, leaving hollow remnants of micro-colonies or empty spaces that become parts of the biofilm water channels [153]. Biofilm is a dynamic structure and there is not established age threshold to determine this issue. The environmental conditions can influence the

time taken to build up a mature biofilm. Biofilm formation and development will depend on the organisms involved, the nature of the surface being colonized, and the physical and chemical conditions of the environment. It has been observed that in oligotrophic environments, as drinking water, biofilms can take over 10 days to reach structural maturity, based on microscopically measured physical dimensions and visual comparison [153]. In other cases, biofilm growth is considered to take from 2 weeks to 1 month [155]. When biofilm formation in drinking water have been studied, in some cases 48h old biofilms have been considered mature [156], while in others cases, in constant hydraulic conditions, several months or more than a year have been considered [157]. In our case, taken into account the nature of the data sets, the data availability and the information found in the literature, biofilm data ≥ 20 days have been considered while the rest have been disregarded.

Despite the amount of data has been reduced, we claim that their quality for our purpose has been clearly improved.

5.2 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 [158].

Pre-processing involves reading data from a very large number of different sources, merging it together, sub-setting it, reshaping it, transforming it, summarizing it, and then finding some data sources that can be used to augment the available data and getting data ready to actually perform useful analysis on it. Pre-processing is a very complex task and sometimes is opened to criticism when innovative resources are used. However, it must be kept in mind that while accurate prediction heavily depends on measuring

the right variables, it is also clearly known that more data and simpler models tends to work better.

5.2.1 Data unification

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.

5.2.1.1 Variables design

The compiled variables can be classified in four groups attending to their nature: physical characteristics (Table 5.4), hydraulic characteristics (Table 5.5), sampling and incubation (Table 5.6), and physico-chemical characteristics of water (Table 5.7). The nature of the variables and categories is further explained below. The target variable has been called *hpc*. The variables with no more than 15% of the cases are not presented but can be found in Appendix A and were kept in the data set for the posterior cleaning process.

5.2.1.1.1 Physical characteristics of the system

In this group we represent the variables related with the physical characteristics of the systems where biofilm has grown (Table 5.4).

- Device: The complexity of DWDS micro-environments have led in most cases to use different growth devices to study them (See Table 3.1). The different categories found are:
 - Propella reactor.
 - Flow cell system.
 - Annular reactor.
 - Robbins device.

TABLE 5.4: Main variables of the physical characteristics group of the dataset.

Physical characteristics of the system	
Origin	Devices
	Real DWDSs
Device	Propella reactor
	Flow cell system
	Annular reactor
	Robbins device
	Pedersen device
	Direct
	Pipe
Tested material	Thermoplastic polymers
	Metals
	Cement based
Duct's shape	Yes
	No

- Pedersen device.
 - Direct. Cases where the biofilm sample has been obtained directly from the pipe of the DWDS, without using any intermediary device.
 - Pipe. This category refers to the cases where a pilot scale water distribution system or sections of pipes have been used as biofilm growth devices.
- Tested material. The ability of pipe materials to support drinking water biofilm varies dramatically from plastic to metal pipes as they exhibit different degrees of surface roughness and chemical activity [21]. Three main groups of materials have been specified.
 - Thermoplastic polymers. In this category the materials found are polyvinyl chloride, polypropylene, polyethylene and polycarbonate. This last one, is not commonly used in DWDSs due to its expensive cost, but there are a lot of reactor studies carried out using polycarbonate slides. It is autoclavable, robust to frequent use, cheap and suitable for microscopy as well as for comparison purposes with previous studies. For this reason, the supply companies of the reactors use it as a basic material for the slides.

- Metals. Free chlorine is known to preferentially react with ferrous iron to produce the insoluble ferric hydroxide quenching their antimicrobial effects [159]. Thus, it has been divided into steel based and iron based pipes.
 - * Steel based. Stainless steel, mild steel, carbon steel, galvanized mild steel and galvanized steel.
 - * Iron based. Cast iron and ductile iron.
- Cement based. The categories found are cement, reinforced concrete and cement lined pipes.
- Duct shape. This variable has been introduced since different growth devices are used in the study of biofilm development. It is a binary variable, Yes or No. Two conditions must be satisfied to consider that the duct shape is like a pipe shape. Firstly, the duct must be a cylinder and, secondly, in the cases of annular reactors, which are formed by concentric cylinders, the biofilm coupons must be located in the outer cylinder, resembling as much as possible real pipe conditions.

5.2.1.1.2 Hydraulic characteristics of the system

The tested hydraulic conditions are included in (Table 5.5).

TABLE 5.5: Main variables of the group of hydraulic characteristics of the dataset.

Hydraulic characteristics of the system	
Hydraulic regime	Dimensionless
	Laminar
Hydraulic diameter	Transient
	Turbulent
Flow velocity	Expressed in mm
Flow rate	Expressed in l/h
	Single pass
Circulation type	Continuous
	No continuous
	Yes
Constant circulation	No

- Hydraulic regime. In fluid mechanics, Reynolds number (Re) is a dimensionless number defined as the ratio of inertial forces to viscous forces and is used to describe

the flow conditions of a fluid. Hydraulic regime may be laminar, transition and turbulent flow. It is useful to help predict similar flow patterns in different fluid flow situations. Its calculation is dependent of the reactor flow geometry. The Reynolds number for the flow in a pipe or tube can be defined by equations 5.1 and 5.3, where D_h is the hydraulic diameter of the pipe (m), ρ is the fluid density (kg/m^{-3}), v is the flow velocity (m s^{-1}), μ is the dynamic viscosity of fluid (N s m^{-2}).

$$Re_{pipe} = \frac{\rho v D_h}{\mu} \quad (5.1)$$

For stirred tanks, as is the case of the annular reactors, the Reynolds number is defined by Equation 5.2, where N is the rotational velocity and D is the diameter of agitator [160].

$$Re_{stirredtank} = \frac{ND^2\rho}{\mu} \quad (5.2)$$

However, it is not possible to calculate this value for all the devices. Moreover, the definition of laminar and turbulent flow regimes varies according to the system used [11]. That is, in cylindrical pipes, $Re \leq 2300$, $2300 < Re < 4000$, $Re \geq 4000$, correspond to laminar, transition and turbulent flow conditions, respectively. In a stirred tank these values are $Re \leq 10$, $10 < Re < 10^4$, $Re \geq 10^4$. Re values are not comparable among different devices that is why a hydraulic regime variable is used instead of Re.

- Hydraulic diameter (D_h). This term is commonly used when handling flow in non-circular tubes and channels. It allows to calculate many variables in the same way as for a round tube, and it is defined as

$$D_h = \frac{4A}{P}, \quad (5.3)$$

where A is the cross sectional area and P is the wetted perimeter of the cross-section.

For a circular tube, it is calculated as:

$$D_h = \frac{4 \pi D^2}{\pi D} = D. \quad (5.4)$$

For an annular duct (with and inner shaft tube, as is the cases of the annular reactors and Propella) the D_h is defined as:

$$D_h = \frac{4 \cdot 0.25 \pi (D_{outer}^2 - D_{inner}^2)}{\pi (D_{outer} + D_{inner})} = D_{outer} - D_{inner}. \quad (5.5)$$

In this case the D_h values are expressed in mm.

In the cases where the dimensions of the biofilm growth device are not specified, but a reference or the model of the annular reactor or Propella is indicated, more bibliography search has been conducted in order to get that information. In most of the cases no productive results have been obtained.

- Flow velocity. It is measured in m/s. In the cases where this value was not given, but flow rate and D_h (m) were available, and the biofilm device was a pipe of a DWDS, the velocity was calculated based on the equation

$$V = \frac{Q}{S}, \quad (5.6)$$

where Q (m^3/s) is the flow rate and A (m^2/s) is the area of the cross section of the pipe, that is

$$A = \frac{\pi D^2}{4}, \quad (5.7)$$

where for circular tubes $D_h = D$ (see Equation 5.4).

- Flow rate. It is measured in l/h. The same process explained above for flow velocity has been applied for the cases of circular tubes where flow velocity and D_h data were given with not flow rate information.
- Circulation type. Three categories have been described.

- Single pass. Water flowing past the device does not return.
 - Continuous. There is some recirculation of water.
 - No continuous. Water is constantly recirculating; there is no renewal.
- Constant circulation. The categories are Yes or No. The circulation is constant if there are not variations in the flow velocity during the study. This variable is important since in real DWDSs the hydraulic conditions change along the days regarding water demand, changing the environmental conditions of biofilm. The steady state hydraulic conditions *versus* those that simulate/include some form of diurnal cycle (as seen in real systems) is important for biofilm layers development, but very few experimental systems replicate this situation.

5.2.1.1.3 Sampling and incubation

There is not established protocol for biofilm sampling and processing. In this category all the variables related with these processes are included, in the case these diverse procedures are somehow affecting the final results (Table 5.6).

- Type of insert. The inserts are the removable coupons where biofilm grows to be subsequently analysed. There are three main types of inserts.
 - Coupon. It is the type of insert that is located in the wall of the device trying not to perturb the water flow but that does not respect the curvature of the wall.
 - Slide. This inserts clearly interrupt the water flow and do not try to reproduce the wall conditions.
 - Direct. This refers to the cases where the samples are directly taken from the pipe wall or inserts that do not stand above the pipe wall and respect the curvature of the pipe, simulating the real conditions of DWDS pipes.
- Scraping technique. The scraping technique refers to the technique used to detach biofilm from the surface where it has developed.

TABLE 5.6: Main variables of the sampling and incubation group of the dataset.

Sampling and Incubation	
Type of insert	Coupon
	Slide
	Direct
Scraping technique	Scraper
	Glass beads
	Swabs
	Scraper and swab combination
	None
Re-suspension method	Shaking
	Vortex
	Sonication
	Combined sonication and vortex
	Mechanical homogenization
Re-suspension solution	None
	No treated water
	Distilled water
	Deionized water
	Saline solution
Plating method	Buffer
	Spread plate
Incubation time	Pour plate
	Expressed in days
Incubation temperature	Expressed in °C

- Scraper. Biofilm is manually detached using a scraper. In this category devices as lab spatula, lifter, scalpel or brushes are included.
- Glass beads. Glass beads of any specific diameter are used.
- Swabs. Cotton swabs are used to swabbing the surface where biofilm has grown.
- Scraper and swab combination. First the biofilm surface is scraped and after that it is swabbed.
- None. No scraping technique is used. The sample is directly re-suspended.
- Re-suspension method. It refers to the technique used to disintegrate biofilm to obtain individual bacteria in the solution.
 - Shaking. Manually or with a shaker.

- Vortex. It is carried out in a vortex mixer. It basically consists of an electric motor that transmits the motion to the liquid of the sample and a vortex is created.
 - Sonication. Is the process of applying sound energy to agitate particles in a liquid sample. In this category ultrasonic frequencies (>20 kHz) are included.
 - Combined sonication and vortex. Both techniques are used; previous to vortex, the sample has been sonicated.
 - Mechanical homogenization. Many different models have been developed using various physical technologies for disruption. The most commonly used is the stomacher, a laboratory paddle blender.
 - None. No physical disruption technique is used, the sample is just scraped.
- Re-suspension solution. In order to culture biofilm samples, these are diluted in a sterile re-suspension solution. This can be:
 - No treated water. It refers to water that has been just sterilized. Water is free from all forms of microbial life (such as fungi, bacteria, viruses, spore forms, etc.). Normally, autoclaves are used to this aim. High pressures enable steam to reach high temperatures, thus increasing its heat content and killing power. Minimum times are usually 15 minutes at 121°C .
 - Distilled water. When water is taken from other sources (that is, other than rain water or snow) it may contain salts and other dissolved solids and it is called impure water. Distillation is carried out to remove these dissolved solids. Distilled water is basically water that has been purified through evaporation. After evaporation, steam is re-captured through condensation.
 - Deionized water. Deionized water is deeply de-mineralized, ultra-pure water with a pH close to 7 at delivery and the electrolytic conductivity about $0.055 \mu\text{S}/\text{cm}$. In order to obtain high quality pure deionized water, a multi-stage water purification process can be used. After pre-cleaning, water is supplied to the reverse osmosis membrane, and then water is filtered through a special de-ionization medium, which removes the rest of the ions in the water. The

purity of deionized water can exceed the purity of distilled water. In this category Milli-Q water is included. Milli-Q is a trademark created by Millipore Corporation to describe ultra-pure water of Type 1 [161]. Milli-Q water purifiers use resin filters and de-ionization to purify water.

- Saline solution. It is used to create an isotonic medium for bacteria. Normally it refers to a sterile solution of 0.9% w/v of sodium chloride (*NaCl*).
- Buffer. It is a solution used to prevent changes in pH. It consists in an aqueous solution of a mixture of a weak acid and its conjugate base, or vice versa. The most commonly used is the phosphate-buffered saline (*PBS*).
- Plating method. Spread plate method or pour plate are the most used techniques when counting colony forming units.
 - Pour plate. The pour plate method involves adding a small volume of sample (0.1-2.0 ml) to melted agar (44-46 °C) and then pouring the mixture into Petri dishes and allowing it to solidify. The plates are then inverted, to prevent condensation on the covers, and incubated [162] [12].
 - Spread plate. The spread plate method has the advantage of using solidified agar, thereby eliminating the possibility of heat shock. The sample (0.1-0.5 ml) is spread on the surface of the agar with a clean sterile spreading rod until being absorbed into the agar surface. Then the plates are incubated [162] [12].

Pour plate method is neither as accurate nor as precise as the spread plate procedure for HPC enumeration [163]. It has been also found that pour plate method procedures yield lower bacteria count than spread plate methods [164]. This may be due to some of the drawbacks that pour plate methods present. For example, the media must be sufficiently cooled before adding the sample, otherwise some bacteria may be killed, but not too much cool to be solidified. Due to these issues, spread plate is more commonly used.

- Incubation time. It is measured in days and oscillates between 7 and 5 days in the R2A long incubation method [12].

- Incubation temperature. It is measured in °C and goes from 20°C to 28°C [12]. In the cases where ranges of temperature were given, the average temperature was selected. In no case this variation was greater than 2°C.

5.2.1.1.4 Physico-chemical characteristics of water

The measurements that describe the flowing water characteristics are included in this group (Table 5.7). The average values have been included. In the cases where geometric means were given these were not taken into account.

TABLE 5.7: Main variables of the physico-chemical characteristics of the water group of the data set.

Physico-chemical characteristics of water	
Water itinerary	Tap Water treatment plant
Disinfectant type	None Chlorine
Total residual chlorine	Expressed in mg/l
Residual free chlorine	Expressed in mg/l
Water temperature	Expressed in °C
pH	Dimensionless
Total organic carbon	Expressed in mg C/l
Assimilable organic carbon	Expressed in μ C/l

- Water itinerary. In this case, it is differentiated between the cases where the studied water is obtained from the tap or directly from the waterworks. This variable has been designed with the aim to observe if the fact that the water has been distributed through the DWDS affects somehow biofilm development.
 - Tap
 - Water treatment plant
- Disinfectant type. Since we have limited our search to systems with no secondary disinfectant or just chlorine, the possible categories are two. In synthetic water chlorine is removed, mainly with GAC/BAC filters but this cases are not included.
 - None

- Chlorine
- Total residual chlorine. When chlorine is added to water, some of the chlorine reacts first with organic materials and metals in the water and it is not available for disinfection, which is known as chlorine demand of water. The remaining chlorine after chlorine demand is called total chlorine [165]. It is expressed in mg/l.
- Residual free chlorine (*freeCl*). Total chlorine is further divided into:
 - Combined chlorine. The amount of chlorine that has reacted with nitrates and is unavailable for disinfection.
 - Free chlorine. The most important type of chlorine. The chlorine available to inactivate disease-causing organisms, and thus a measure to determine the potability of water [165].

Although the DPD (N,N-diethyl-p-phenylenediamine) method is internationally recognized as the standard method of testing chlorine in water, the method used for measuring chlorine concentration was also checked [166]. In all the cases in which the method was reported, the DPD method was used. The DPD test is easy to perform, requires little apparatus, is inexpensive, and adapts well to field test situations.

In the cases where just “residual chlorine” was given, not indicating if it was free or total, with no other clue, no values were incorporated to the database.

- Water temperature. It is measured in °C. Temperature is widely recognized as an important controlling factor in influencing bacterial growth [167].
- pH. The pH of water is a measure of the acid/base equilibrium and, in most natural waters, is controlled by the carbon dioxide-bicarbonate-carbonate equilibrium system. Although pH usually has no direct impact on water consumers, it is one of the most important operational water-quality parameters. Careful attention to pH control is necessary at all stages of water treatment to ensure satisfactory water clarification and disinfection. For effective disinfection with chlorine, the

pH should preferably be less than 8. The optimum pH will vary in different supplies according to the composition of the water and the nature of the construction materials used in the distribution system, but is often in the range 6.5-8.5 [168].

- Total organic carbon. It is measured in mg C/l. The organic carbon in water is composed of a variety of organic compounds in various oxidation states. Some of these carbon compounds can be further oxidized by biological or chemical processes. Total organic carbon (TOC) is a more convenient and direct expression of total organic content than other compounds but does not provide the same kind of information [169]. In drinking water TOC ranges from less than 100 $\mu\text{g C/l}$ to more than 25,000 $\mu\text{g C/l}$ [169].
- Assimilable organic carbon (AOC). It is measured in $\mu\text{g C/l}$. The AOC concentration in a water sample is governed by simple monod type kinetics; it is proportional to the density of the organisms that can grow in it [170]. It is accepted as an indicator of bacterial regrowth [171].

Since AOC concentration can be measured in various ways the method used for AOC determination in the remaining papers was checked. In all the cases it was a modification [172] of Van der Kooij *et al.*'s method (1982) [173]. The same was done with TOC measurements. This time the applied methodology was specified just in one paper. Since TOC is an automatized measurement, normally just the model of the TOC analyzer is specified.

5.2.1.1.5 Biofilm

The number of colonies forming units per cm^2 found in the biofilm samples are presented. This is the objective variable.

- R2A cultivable cell in biofilm (*hpc*). Measured in $\log\text{CFU}/\text{cm}^2$. Further explained in Section 3.4.

5.2.2 Data cleansing

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 is summarized in Figure 5.1 and explained in detail below. Some of the subsections have been further developed due to their importance. They required further discussion as a result of their relevance or scientific debate (as they have also been more studied and developed in the literature).

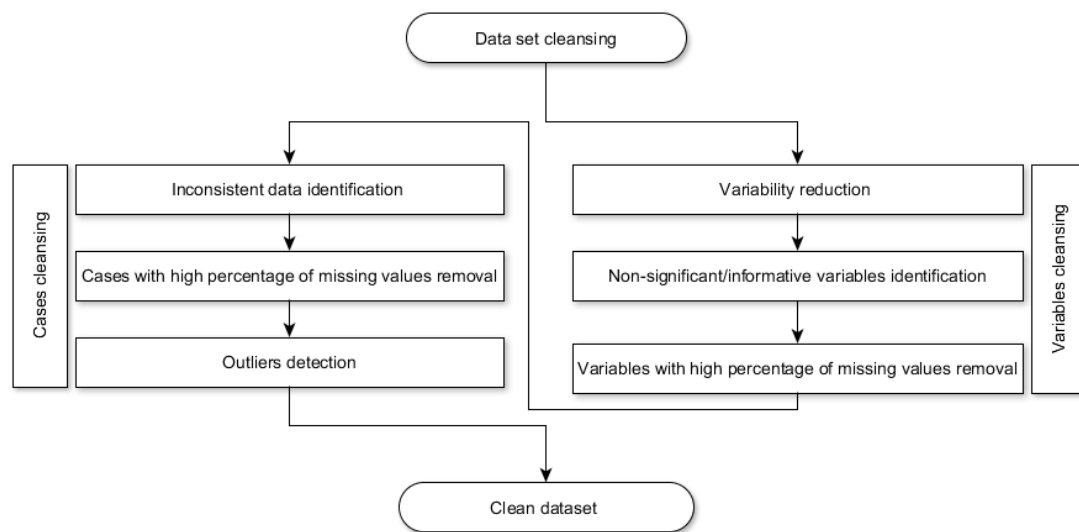


FIGURE 5.1: Data cleansing process.

5.2.2.1 Variables cleansing

5.2.2.1.1 Variability reduction

When comparing various removal and re-suspension techniques an analysis of variance revealed that the removal step was more significant at the 5% level to the recovery of biofilm cells [174]. Differences in culturable and total bacteria have been observed when

comparing different removal methods. However, the variability of the *Re-suspension method* and *Scraping technique* variables are not well represented in the database, creating a bias risk. In order to keep this information in the database, its variability has been reduced by combining these two variables into one. This variable has been called *Removal* and its categories are *Low*, *Medium* and *Strong*. *Low* corresponds to the cases where no disruption method has been used. *Medium* refers to the cases that scraping, swabbing or glass beads have been used followed by shaking or vortexing. It has been observed that the manual methods are reasonably comparable with each other, scraping and swabbing generated similar bacterial numbers [174]. *Strong* are the cases where sonication has been applied since it has been observed that automated procedures tend to be more effective than the manual ones [174].

5.2.2.1.2 Non-significant/informative variables identification

The removal of non informative variables reduces the redundancy of the database, enhancing the principle of parsimony (or simplicity), which is influential in problems of statistical inference, since it explicitly favours simpler models.

The effect of the re-suspension media has been already tested when determining HPC from biofilm samples on to R2A agar and no significant differences have been found [174]. Thus, the variable *Re-suspension media* has been removed.

The variables *Flow rate* and *Reynolds number* are dependent of the *Velocity* and *Dh* variables for ducts. To avoid redundant information in the database and to simplify it as much as possible, just the *Hydraulic regime* variable is kept since it has been demonstrated the influence of the hydraulic regime in biofilm development [10], so that it cannot be omitted, and it is the only variable directly comparable among the different devices.

Since the free chlorine concentration variable has been kept, the variable *disinfectant type* is no longer necessary and it is removed.

5.2.2.1.3 Removal of variables with high percentage of missing values

There is not established yet any cut-off value in the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences. The proportion of missing data is directly related to the quality of statistical inferences [175]. Accordingly, it was decided to remove the variables with less than 70% of data. After this step, most of the physico-chemical parameters of the water have been removed. This is not an important issue if the main characteristics that are known that affect biofilm development are kept. These are chlorine concentration, water temperature and organic carbon content. However, all the variables related with this last issue have been removed. In general, microorganisms need a C : N : P (carbon, nitrogen and phosphorous) ratio of 100 : 10 : 1, where carbon is the growth-limiting nutrient in most DWDSs [159]. Thus, more or less carbon concentration can limit microorganism growth. To solve this setback a new variable has been added to provide an indirect idea of the water carbon content, *water source*. It is known that groundwater tends to have lower concentration of organic materials than surface water and may affect biofilm development. Thus, a difference between surface water (*S*) and groundwater (*G*) is made. In the cases were this issue was not specified, but the name of the supplier waterworks was given, a web search was undertaken to get this information. When both type of waters were mixed the main source has been specified.

Most of the hydraulic characteristics have also been discarded mainly because different devices are used not allowing easy hydraulic comparison among them and the conditions are variable in operating systems, not allowing to specify a value for these variables. However, since the *device* variable has been maintained this provides an indirect idea (approximation) of the hydraulic characteristics of the system.

5.2.2.2 Cases cleansing

5.2.2.2.1 Inconsistent data identification

In some cases ($K1$, $K3$, $K5$ and $K24$) the variable hpc takes the value $\text{HPC}/\text{cm}^2 = 1$. If plates from all dilutions of any sample have no colonies, the count is reported as less than one (< 1) times the reciprocal of the corresponding lowest dilution. Since the recommended aliquot for HPC in R2A is 0.1 - 0.5 ml, $\text{HPC}/\text{cm}^2 = 1$ can be misunderstood. Thus, these cases have been removed.

5.2.2.2.2 Removal of cases with high percentage of missing values

As stated above there is not an acceptable percentage of missing data, thus, as in the case of the variables, cases with more than the 70% of missing values have been excluded from the analysis. In particular, these are all the cases from the paper [176].

5.2.2.2.3 Outlier detection

The presence of outliers in data sets can lead to model miss-specification, biased parameter estimation and incorrect results. It is therefore important to identify them prior to modelling and analysis [177].

Figure 5.2 illustrates outliers in a simple 2-dimensional data set. The data has a ‘normal’ region, G , since most observations lie in this area. Points that are sufficiently far away from the regions, points O_1 and O_2 , are outliers [178].

Outlier detection techniques can operate in one of these three modes, based on principled and systematic techniques:

- Supervised outlier detection: There is prior information about the abnormalities in the data. The training data set has labelled instances for normal data, as well as outlier data [179].

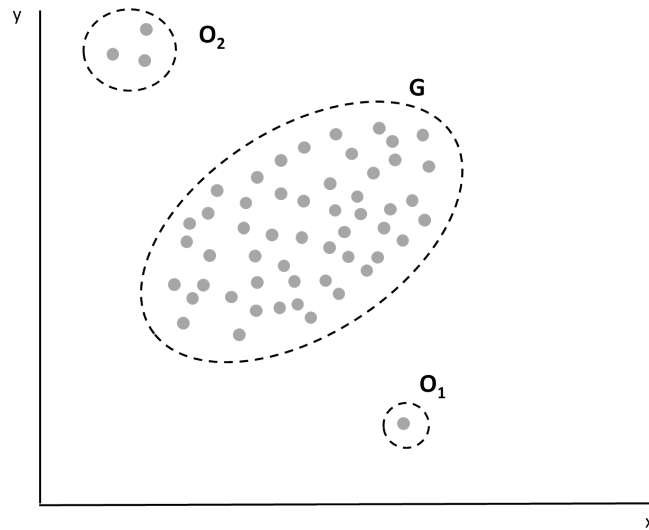


FIGURE 5.2: A simple example of outliers in a 2-dimensional data set.

- Semi-Supervised outlier detection: It uses both unlabeled and labelled data. Semi-supervised outlier detection can improve the accuracy of unsupervised outlier detection using supervision of some labelled data while reducing the need for expensive labelled data required in supervised methods [180].
- Unsupervised outlier detection: It does not require training data. Outliers are detected based on the assumption that they statistically deviate from normal data. Unsupervised outlier detection is most widely applied [181].

An unsupervised anomaly detection algorithm is applied in this case, the local outlier factor (LOF) algorithm [182]. The LOF assigns to each object a degree of being an outlier, by quantifying how outlined an object is. This degree is called the local outlier factor of an object. It is ‘local’ in that the degree depends on how isolated the object is with respect to the surrounding neighbourhood. LOF is able to find outliers which appear to be meaningful, but can otherwise not be identified with existing approaches [182]. This algorithm has been implemented through the R Package ‘DMwR’, version: 0.4.1 [183]. This function, given a data set, produces a vector of local outlier factors for each case. A sensitivity analysis was carried out, and it was determined that 12 (around 5% of the entire database) were outliers to be removed. In Figure 5.3 it is shown a principal component plot where the 12 resulted outliers can be observed.

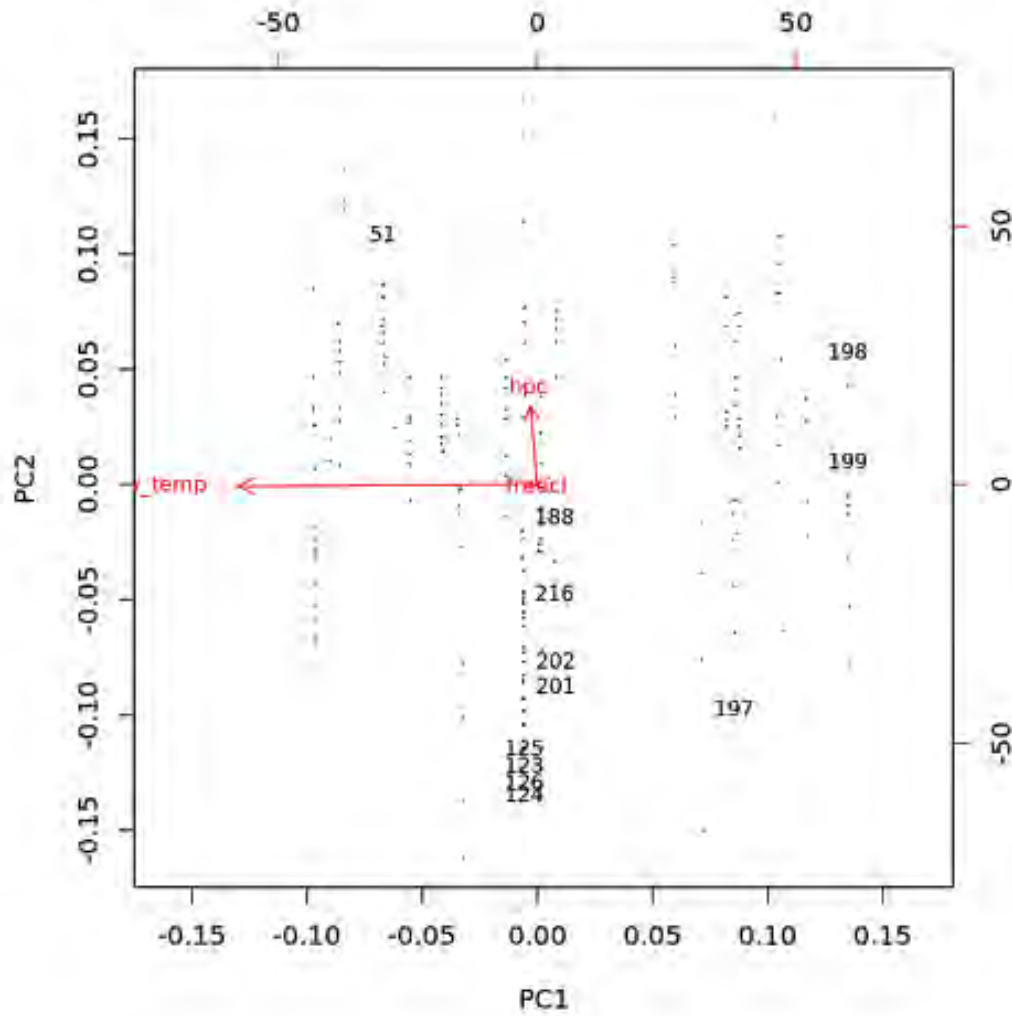


FIGURE 5.3: Detected outliers after the implementation of the local outlier factor (LOF) algorithm to the dataset.

The survey plot has been carried out with the software Orange Canvas 2.7.8 [184].

5.2.2.3 Clean data set

After pre-processing the data set is formed by 284 examples, 15 attributes and 1 meta attribute (*Id*). The variables and categories of the resulting data set are presented below.

1. Physical characteristics

- Device

- Propella reactor (*PR*)
- Flow cell system (*FC*)
- Annular reactor (*AR*)
- Robbins device (*RD*)
- Pedersen device (*PE*)
- Direct (*D*)
- Pipe (*P*)
- Tested material (*material*)
 - Thermoplastic polymers (*TP*)
 - Iron based (*I*)
 - Steel based (*S*)
 - Cement based (*C*)
- Duct's shape (*pipe_like*)
 - Yes (*Y*)
 - No (*N*)

2. Hydraulic characteristics

- Circulation type (*c_type*)
 - Single pass (*SP*)
 - Continuous (*C*)
 - No continuous (*NC*)
- 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)
- Incubation time (*inc_time*)
- Incubation temperature (*inc_temp*)
- Plating method (*plating*)
 - 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*).

5. Biofilm

- R2A cultivable cell in biofilm (*hpc*)

5.2.3 Data set reconstruction

The database presents 80% of complete cases (228), since 4 of the 15 variables (*material*, *wat_source*, *wat_temp* and *free_Cl*) have missing values. The variables *wat_temp* and

free_Cl are the ones with highest number of missing values, while the number of missing data in the variables *material* and *wat_source* are very scarce (1 and 6, respectively). The proportion and combination of these missing values in the database are shown in Figure 5.4. The upper figure shows the amount of missing values in each variable, while the other represents all the existing combinations of missing and non-missing values. Cases with 1 and 2 missing values are observed in the synthetic database. There are no cases with more than 2 missing values.

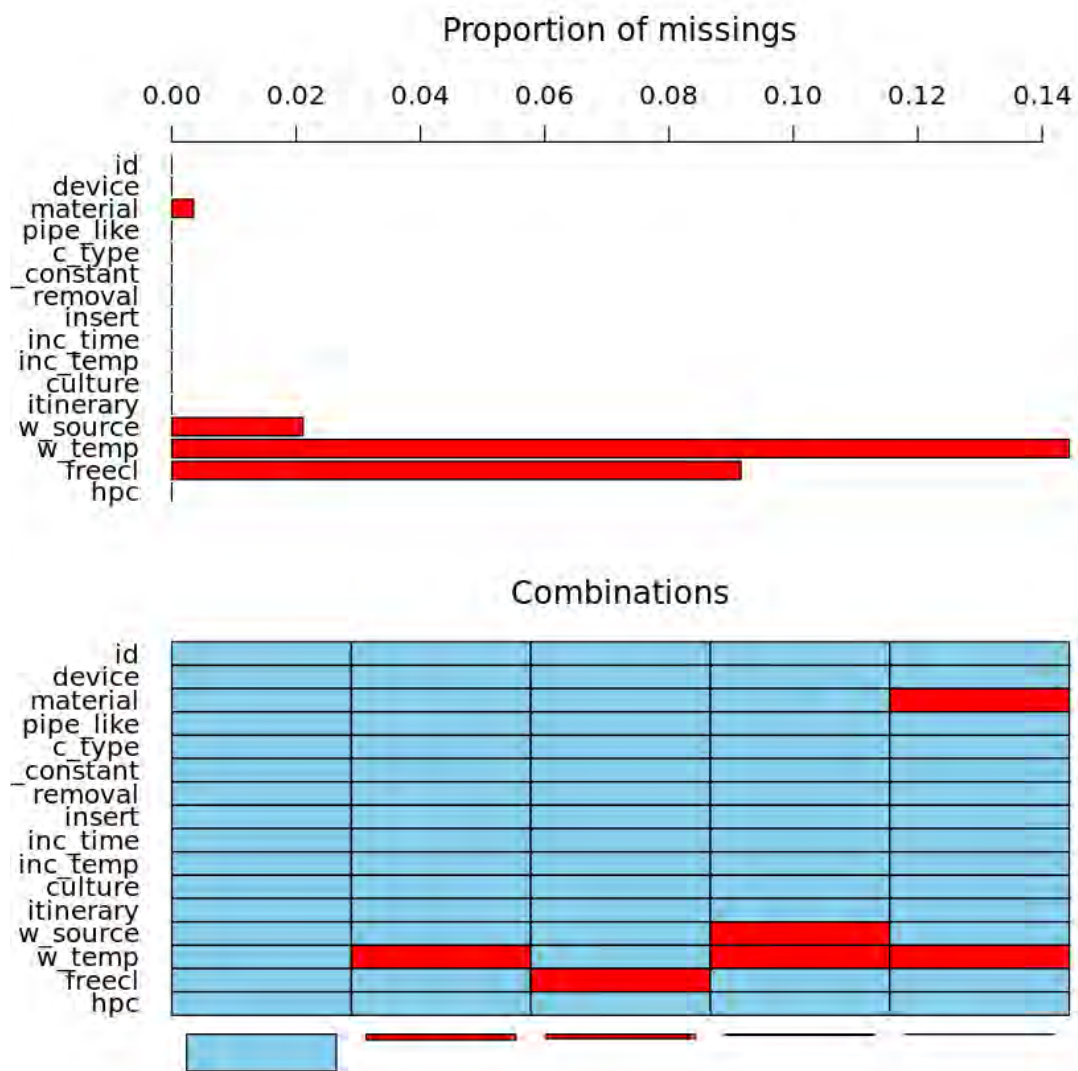


FIGURE 5.4: Proportion and combination of the missing data in the database. Missing data are represented in red colour.

5.2.3.1 Imputation of missing values

The presence of missing data in the data set is an issue that must be seriously taken into account, if the aim is to perform more complex and sophisticated analysis requiring high quality input data. The presence of missing values is a common problem in data analysis. In the cases where removing variables or observations with missing data is not an option these missing values must be fill in or “impute” missing values. Imputation methods keep the full sample size, which can be advantageous for bias and precision. Replacing each missing value with the mean of the observed values for such variable is perhaps the easiest way of imputation. However, this method can severely distort the distribution for this variable. It can lead to complications with summary measures including, notably, underestimation of the standard deviation [185]. The missing value problem is fundamental in data sets [186] and many works have contributed in this field. 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 [187]. In the MICE procedure a series of regression models are run whereby each variable with missing data is modelled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution, with, for example, binary variables modeled using logistic regression and continuous variables modeled using linear regression [187]. To implement this algorithm we have used the R package ‘mice’ [188] that imputes incomplete multivariate data by chained equations.

In MICE the entire imputation process is repeated to generate multiple imputed datasets. The observed data is the same across the imputed datasets; only the values that had originally been missing will differ [187]. In our case, taking into account the size of our dataset and the reduced number of missing values, the number of repetitions has been set in the default value 5, as suggested. In Figure 5.5 the values imputed in each repetition for the two principal variables with missing values, residual free chlorine (*freecl*) and water temperature (*w_temp*) can be observed. Once the data have been imputed, each imputed dataset is ‘complete’ in the sense that it has no missing values. An iterative process is launched through the results of the imputation error calculated through correlations between the involved variables. This iterative process learn from

those mistakes to minimize them. It evolves heuristically to the slightest mistake by an appropriate combination of these imputed values in every step.

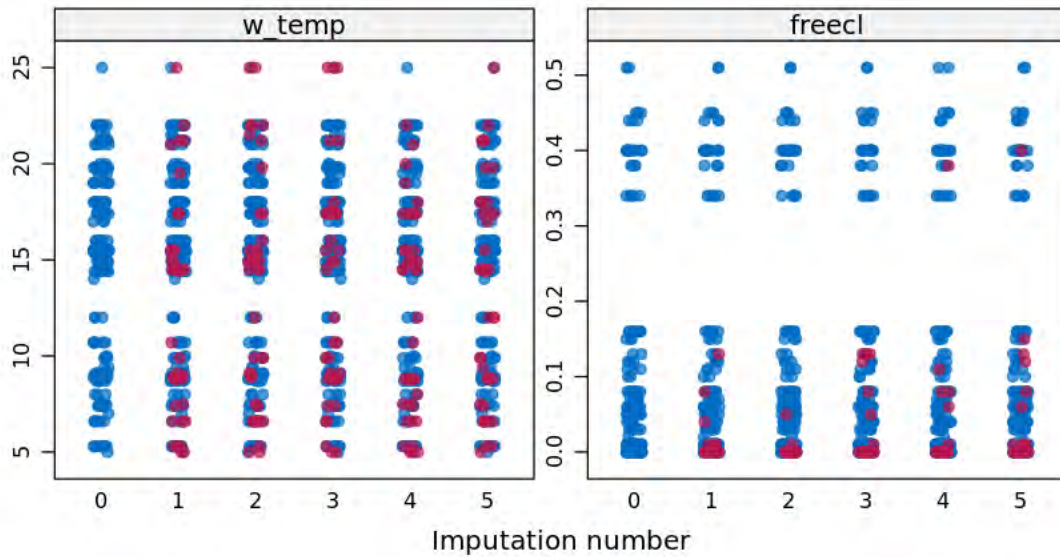


FIGURE 5.5: Imputed data (in red) for the variables *w_temp* and *freecl* in each MICE imputation process.

5.2.3.2 Complete data set

After data set reconstruction the final data set is formed by 284 complete cases with 15 attributes. Finally, to be sure that no hidden correlations occur a survey plot has been carried out on the complete cases (Figure 5.6). A survey plot is a simple multi-attribute visualization technique that can help to spot correlations between any two variables. It has been carried out with the software Orange Canvas 2.7.8 [184]. The data on a specific attribute is shown in a single column, where the length of the line corresponds to the dimensional value. It is observed that no variables have the same shape indicating no correlation among them. At this point, the pre-processing step is finished.

An extract of the database can be found in Appendix B.

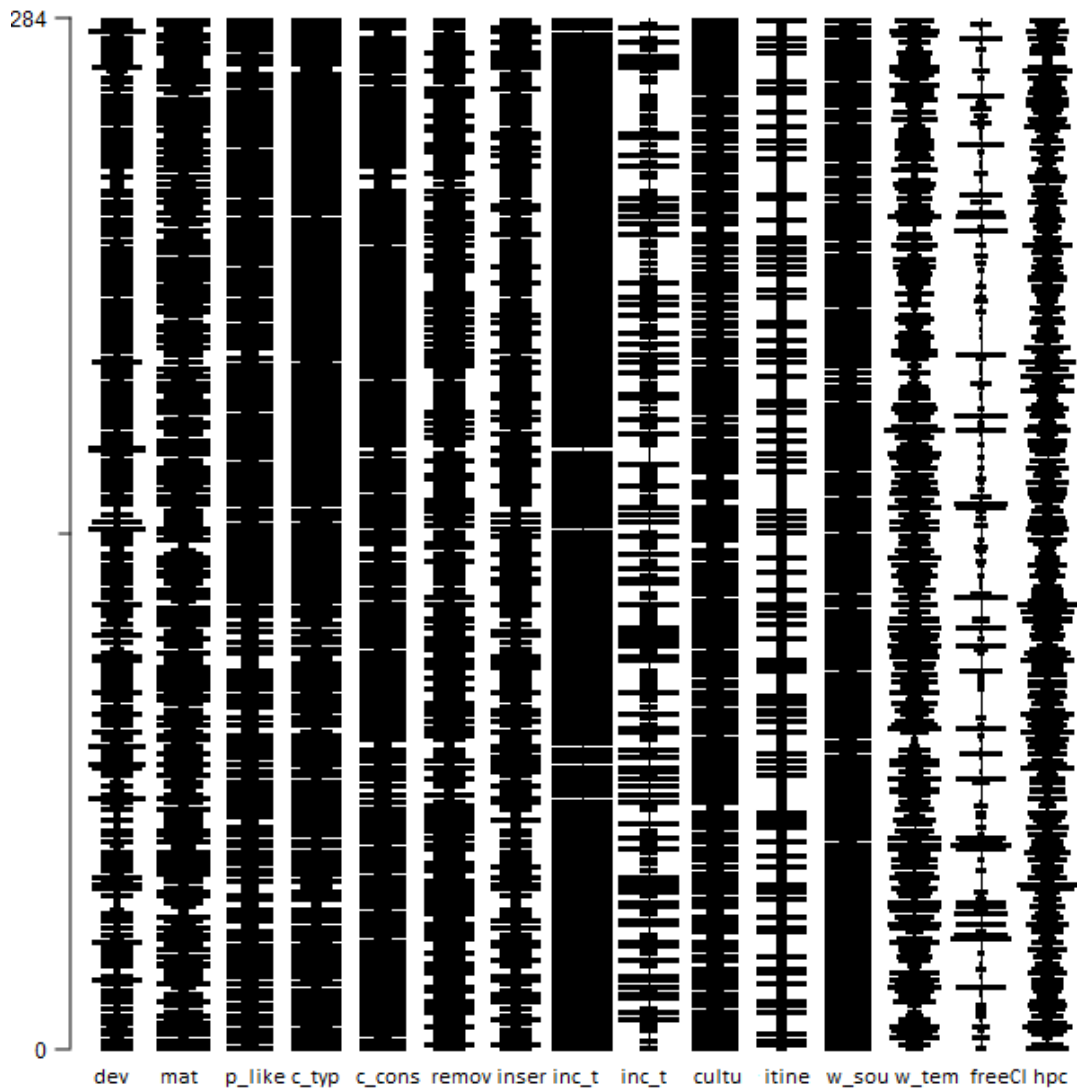


FIGURE 5.6: Survey plot of the final data set.

Chapter 6

Data set: Exploratory Data Analysis

Exploratory data analysis (EDA) is a well-established statistical methodology that provides conceptual and computational tools for discovering basic patterns to foster posterior refinement leading to develop hypothesis of interest and more advanced analysis. It is mainly used for error detection, checking of assumptions, preliminary selection of appropriate models, determining relationships among the explanatory variables, and assessing the direction and rough size of relationships between explanatory and outcome variables.

EDA does not include formal statistical modelling. However, EDA of the data set can help understand the data set and select subsets of the data for further investigation. Moreover, one can get a prior notion of the relationships among the variables. EDA can lead us to have an idea of the main variables that are influencing biofilm development in DWDSs, and focus our attention on most relevant aspects of the problem.

6.1 Descriptive data analysis

Prior to performing exploratory data analysis, a descriptive analysis has been undertaken. The descriptive analysis remains in the mere graphical representation and a basic analytical summary of the data, while, EDA provides a step forward suggesting

basic patterns for further analysis. The descriptive analysis goal is just to describe the set of data, helping to show and summarize the data. The studied data set raw report is provided in Table 6.1. The descriptive analysis of the attributes has been carried out with the software Orange Canvas 2.7.8 [184].

TABLE 6.1: A general description of the data set.

Cases	284
Attributes	14 (<i>device, material, pipe_like, c_type, c_constant, removal, insert, inc_time, inc_temp, plating, itinerary, w_source, wat_temp, and freeCl</i>)
Meta attributes	1 (<i>Id</i>)
Target attribute	<i>hpc</i>

The meta attribute *Id* refers to the key variable. It allows to follow up the data source from which every data set register has been obtained.

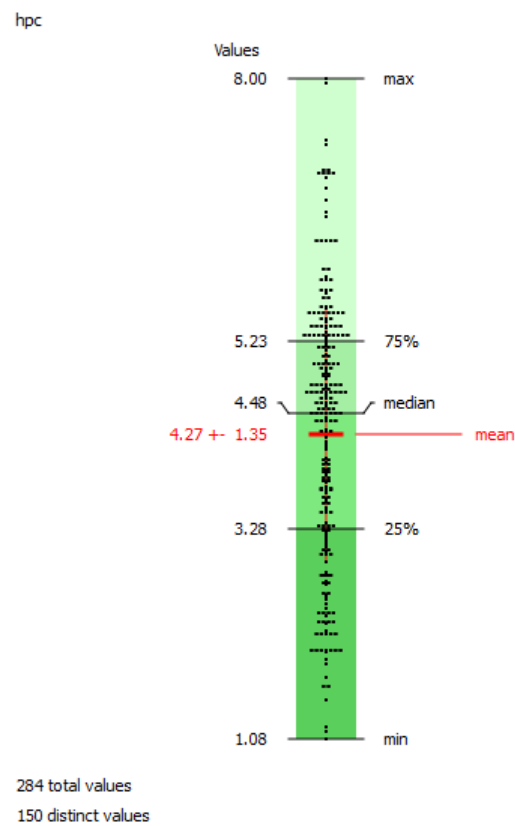
6.1.1 Target attribute

The target attribute *hpc* (Figure 6.1) is the outcome that we would like to predict.

The observed range of the *hpc* attribute is well characterized in the data set. Mean and median values coincide. The mean is affected by a single change in any of the data values. As a consequence, it is more sensitive than the median to outliers or data corresponding to large values. The median is robust since it is computed through the central point of the ordered sample instead of being affected by any individual value. The differences between mean and median are useful to detect asymmetries in the data distribution or the presence of the previously mentioned outliers. The symmetric distribution of the variable *hpc* indicates that our target variable is well represented in the data set, although higher values are less common than the lower ones as observed in the 75th percentile.

6.1.2 Categorical attributes

The categorical attributes of the data set are eleven. Their categories and distribution are presented below.

FIGURE 6.1: Biofilm attribute (*hpc*) statistics.

- Device type

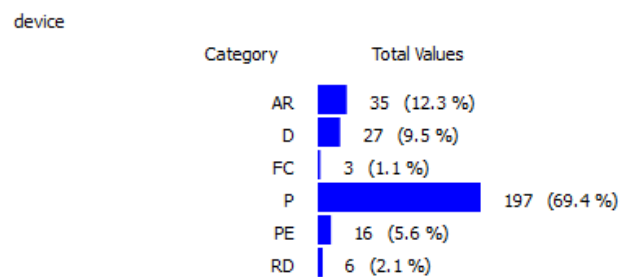


FIGURE 6.2: Device attribute: AR, annular reactor; D, direct; FC, flow cell; P, pipe; PE, Pedersen; RD, Robbins; PR, propella.

It can be observed (Figure 6.2) that the data set is mainly formed by data obtained from pilot DWDSs, which are the most similar devices to real DWDSs. Generally, their main difference is that steady state hydraulic conditions are performed in

this pilot systems, while variable flow usually occurs in real DWDSs. However, the data obtained from operating DWDSs are the third most represented class in the data set. The devices that less resemble real DWDSs are the ones that are less represented, but their presence is important since they can contribute with an interesting combination of variables from which interesting information can be extracted.

- Pipe material

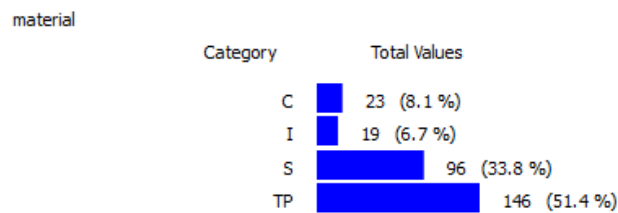


FIGURE 6.3: Pipe material attribute: C, cement; I, iron; S, steel; TP, thermoplastic.

Thermoplastic polymers are the ones that are more represented in the data set (Figure 6.3). Metallic pipes are divided into two categories. Iron based pipes are less represented than the rest pipes. Altogether metallic pipes present more cases than cement pipes.

- Duct shape

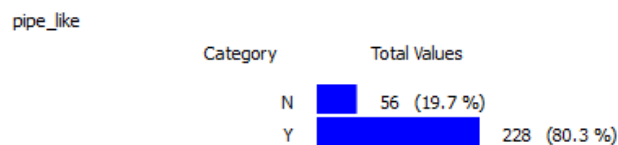


FIGURE 6.4: Duct shape (pipe-like) attribute: N, No; Y, Yes.

Since the data set is mainly formed by data obtained from pilot systems, ARs and DWDSs, as observed above (Figure 6.2), the 80% of the data has been obtained from pipe-like ducts (Figure 6.4).

- Circulation type

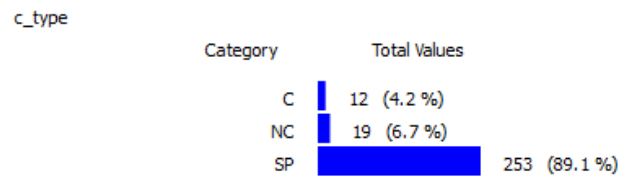


FIGURE 6.5: Circulation type attribute: C, continuous flow; NC, non-continuous flow; SP, single pass.

Almost the 90% of the data have been obtained from single pass systems as DWDSs are. Continuous and non-continuous flows represent the 4% and 6% of the data respectively (Figure 6.5).

- Constant circulation

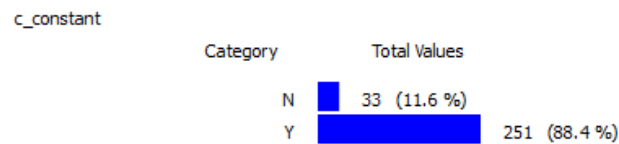


FIGURE 6.6: Constant circulation attribute: N, No; Y, Yes.

In this case, most of the instances correspond with constant circulation (Figure 6.6), opposite to the variable flow that is found in operating DWDSs. This is due to the fact that the vast majority of the pilot scale systems are run at steady state conditions due to technical and financial limitations. Data obtained from non constant flow comes mainly from operating DWDSs.

- Removal technique

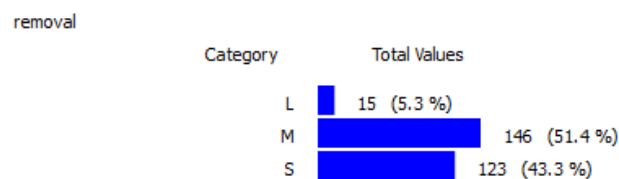


FIGURE 6.7: Removal technique attribute: L, low; M, medium; S, strong.

The theoretically more strong techniques are more represented than the "weaker" ones (Figure 6.7). The former also represents the researchers' preferred choice when analysing biofilm samples.

- Insert type

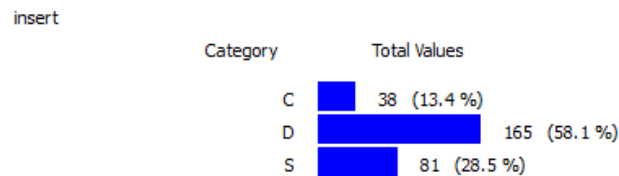


FIGURE 6.8: Insert type attribute: S, slide; C, coupon; D, direct.

Most of the data have been obtained directly from pipe walls. This is, no slide or coupon that could distort the hydraulic conditions have been used. Coupons are less represented (Figure 6.8).

- Incubation time

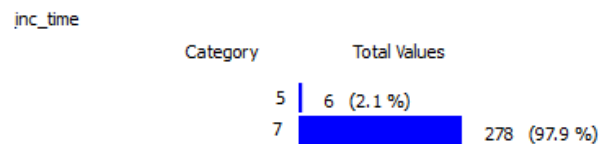


FIGURE 6.9: Incubation time attribute (days).

According to the long time R2A incubation protocol, samples can be incubated from 5 to 7 days [152]. Thus, the categories for this variable would be theoretically three, although just data for 5 and 7 days of incubation have been found, most of them for 7 days (Figure 6.9).

- Plating method

The spread plate method is the most commonly used plating method as observed in Figure 6.10. It may be due to the possible drawbacks that can be found when using the pour plate method and which have been explained in Chapter 5.

- Itinerary

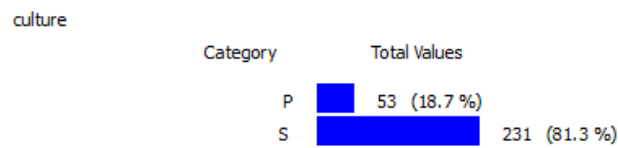


FIGURE 6.10: Plating method attribute: P, pour plate; S, spread plate.

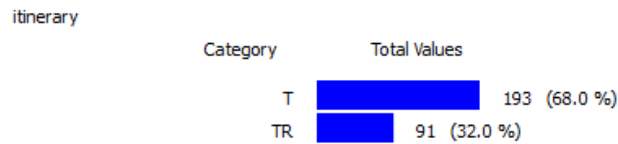


FIGURE 6.11: Itinerary attribute: T, tap; TR, treatment plant.

Most of the cases have been performed with tap water, after being distributed. Around the 30% of the data have been obtained from works performed with water directly obtained from the waterworks (Figure 6.11).

- Water source

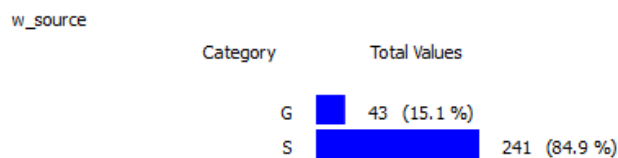


FIGURE 6.12: Water source attribute: G, groundwater; S, superficial.

Groundwater is the less represented category with 15% of the total (Figure 6.12).

6.1.3 Continuous attributes

Three out of the 14 variables of the data set are continuous attributes. Their distributions are presented below.

- Incubation temperature

The incubation temperature for R2A long incubation can range from 20°C to 28°C [152]. As observed in Figure 6.13 it seems that the lowest or highest temperatures

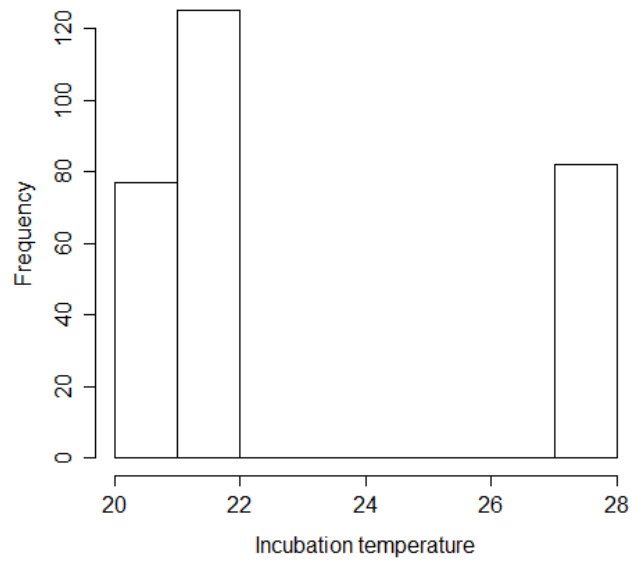


FIGURE 6.13: Incubation temperature attribute.

within this range are preferred when conducting the experiments. The lower values are 20°C and 22°C, while the higher values correspond to 28°C

- Water temperature

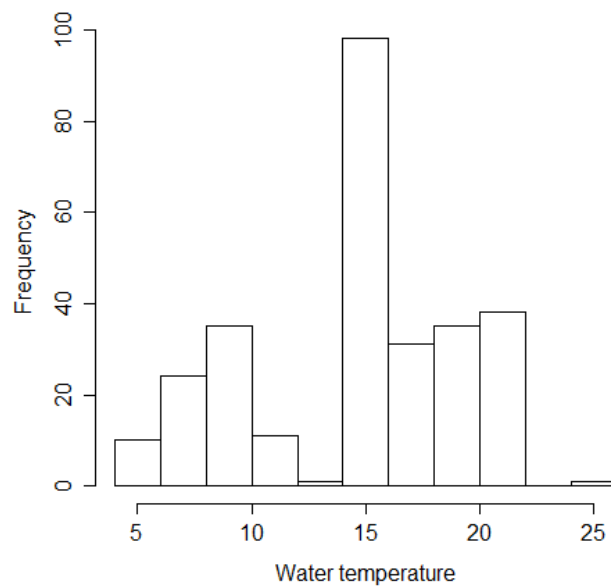


FIGURE 6.14: Water temperature attribute.

The values of water temperature found in the data set range from 5.3°C to 22°C. Almost in a symmetric way, all the values in the range are represented in the data set (Figure 6.14). Since water has a high specific heat index (the amount of heat per unit mass required to raise the temperature by 1°C) its temperature changes, over day and night, and over seasons, are more gradual and no so extreme as for the environmental temperature. However, the differences on water temperature found in different regions can be very pronounced.

- Free chlorine concentration

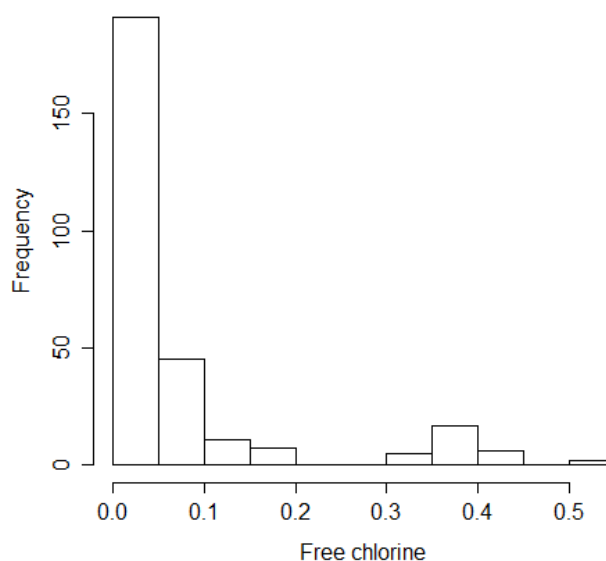


FIGURE 6.15: Free chlorine concentration attribute.

The range of values found in the *freeCl* variable goes from 0 to 0.51 (Figure 6.15). The lower values of free chlorine are commonly found at the dead-end points of DWDSs. As explained before in this section, these problematic areas are a key issue in water quality management in DWDSs. Thus, these points represent specially interesting cases for the study of biofilm development in DWDSs due to their vulnerability to bacterial growth. In fact, secondary chlorination dose rates are generally determined by trying to achieve a free chlorine residual of >0.1 mg/l at the network extremes [189].

An ideal system supplies free chlorine at a concentration of 0.3-0.5 mg/l [190]. However, for example, free chlorine concentrations in most Canadian drinking water distribution systems range from 0.04 to 2.0 mg/l [191]. Chlorine is rapidly consumed and high values are extremely rare in the distribution pipes. In fact, this is probably the reason why the lower values of free chlorine are more represented in the data set.

6.2 Exploratory data analysis

Exploratory data analysis is a good way to discover new connections. Connections are useful to define future data science projects, and to confirm the exploration performed. However, it is important to notice that they are not the final answer on any particular problem, and they should not be used for generalizing or predicting.

6.2.1 Categorical attributes

In Figure 6.16 the target attribute, *hpc*, is grouped by the classes of the categorical attributes of the data set. The results obtained for each variable are explained below.

- Device type. Regarding the average biofilm found in each device, in the data set, the devices can be divided in two groups. The devices *AR*, *D* and *P* belong to the group with lower biofilm, less than 5 logUFC/cm², while the devices *FC*, *PE* and *PR* present more than 5 logUFC/cm². In a rough way, it could be said that devices that physically less resemble pipes have higher biofilm development.
- Pipe material. No big differences are observed among the different materials. However, the biofilm average values found in iron based pipes tend to be the highest ones [14], although the values in thermoplastic pipes are also high.
- Duct shape. Although both categories, yes (*Y*) and no (*N*), present similar values, the average value of the category *N* is higher than the *Y* category. This is in agreement with the trend found when analysing the *device* attribute.

- Circulation type. The non-continuous circulation (*NC*) clearly presents higher values of biofilm than the rest of the categories. The categories single pass (*SP*) and continuous circulation (*C*) present very similar values. This differentiation may be observed due to the fact that normally *NC* circulation is more used in bench top devices than in pilot scale systems. As observed above, it seems that these bench top devices tend to support higher biofilm development than pilot scale systems (*P*) or operating DWDSs (*D*), that usually have *C* and *SP* circulation, respectively.
- Constant circulation. Both categories have similar values, however the category *Y* seems to present higher values. In our data set the data with no constant circulation correspond mainly with the data obtained directly from operating DWDSs that, until now, seemed to have a trend to develop lower biofilm development.
- Removal technique. In this case, contrary to what is expected, the low removal techniques (*L*) present higher values than the strong (*S*) and medium (*M*) removal techniques. In contrast to the observed, in literature it is found that automated procedures tend to be more effective than the manual ones [174].
- Insert type. All the cases present similar values, however the lowest values of biofilm are observed when the type of insert is a slide (*S*). The other two categories present similar values.
- Incubation time. In the boxplot, five days incubation seems to present more CFU than 7 days incubation. However, 5 days incubation is represented by low number of cases in the data set (Fig. 6.9). Thus, this result can be biased and affected by other variables.
- Plating method. The cases that pour plate method (*P*) was used present lower quantity of biofilm development than those where the spread plate (*S*) plating method was applied. Similar conclusions have been found in the literature [167].
- Itinerary. According to the observed biofilm, development seems to be higher when the flowing water is obtained directly from the waterworks (*TR*) than when it corresponds to tap water (*T*). This does not agree with what it was expected.

Water directly obtained from the waterworks is of better quality than that from the tap. However when observed the *TR* instances in the data set, it is found that most of them have no disinfectant residual. This fact could be affecting the observed results.

- Water source. Both categories present similar distributions. However, opposite to what is expected, groundwater (*G*) presents a higher average than surface water (*S*). However, the *G* category is less represented than the *S* in the data set, and this result can be influenced by this fact.

6.2.2 Continuous attributes

In the case of the continuous attributes, a scatterplot has been applied to each one in order to study individually its relation with the variable *hpc*. Scatterplots are graphical representation of the relationship between two quantitative variables plotted along two axes. They are very useful as visualization tools. They help to identify the possible relationship between two variables that are plotted in pairs.

Data visualization is an essential tool in data analysis since it enables to visually detect complex structures and patterns in the data. The most natural way to identify clusters is by using data visualization because human mind excels in prompt interpretation of visual information [192]. It plays a crucial role in identifying interesting patterns in exploratory data analysis [193].

In this case, a linear regression line has been added to represent the trend of the relationship between the two variables. In this way we are able to have a visual idea of the strength of the direct relationship between the variables and if this relation is positive or negative. A second line, has also been added, the LOWESS (Locally Weighted Scatterplot Smoothing) line [194]. It is a non-parametric regression that creates a smooth line through the scatterplot to facilitate the visualization of any possible relationship between variables. These analyses have been implemented through the ‘car’ R package, version 2.1-0 [195].

In the case of water temperature (Figure 6.17), it can be observed a slightly increasing trend in both lines. That is, more biofilm development is associated with higher water temperature. Temperature is known to be an important factor in biofilm development [196]. High temperature favour a growing rate of bacteria, if these are in the tolerance range of the studied bacteria. However, the LOWESS line shows that this relationship strength decreases when the water temperature is around 15°C.

When testing the free chlorine residual (Figure 6.18) the linear regression line does not present any clear slope. However, when focusing on the LOWESS line it is found that, as expected [83], a trend toward lower biofilm development when increasing the free chlorine concentration is observed.

In this case (Figure 6.19), a clear negative slope is found in both lines, opposite to the expected, since as mentioned before, it is well known that temperature favours bacterial growth [196]. However, the data presents a clear non-homogeneous distribution and it can be affecting the results.

6.2.2.1 Data set clustering

Agglomerative hierarchical clustering has been applied to the dataset. The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis [197]. In hierarchical clustering a dendrogram is created. The algorithm begins with each point in its own cluster and progressively joints the closest cluster to reduce the number of clusters to 1 [198]. Subsequently, data is continually fused one-by-one in order of highest similarity and, eventually, all data are contained in the final cluster at similarity 0.0.

A Gower's distance matrix has been used since we work with a mixed data set including categorical and continuous variables. The Gower's distance matrix has been produced using the function named *daisy()* of the package *cluster* v. 2.0.2 of R [199]. It computes all the pairwise dissimilarities (distances) between observations in the dataset. The main feature of Gowers distance [200] is its ability to handle different variable types

(e.g. nominal, ordinal, (a)symmetric binary) even when different types occur in the same data set. Each variable is first standardized by dividing each entry by the range of the corresponding variable, after subtracting the minimum value; consequently the rescaled variable has range $[0,1]$. The *hclust()* function performs the hierarchical cluster analysis from the dissimilarity matrix calculated previously.

The number of clusters have been chosen through the silhouette method. That is, each cluster is represented by a so-called silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The entire clustering is displayed by combining the silhouettes into a single plot, allowing an appreciation of the relative quality of the clusters and an overview of the data configuration. The average silhouette width provides an evaluation of clustering validity, and might be used to select an appropriate number of clusters [201]. In this case, the biggest average silhouette width has been obtained when the number of clusters reached the number 12 (it was tried from $n = 2$ to $n = 20$). Thus, twelve clusters were selected. They obtained an average Silhouette width of 0.55, which means that a reasonable structure has been found 6.20. Finally a partitioning has been applied using the *clusplot()* function [202] of the Flexible Procedures for Clustering-fpc R package version 2.1-10 [203] to visualize these groups (Figure 6.21). A bivariate plot has been created to visualize a partition (clustering) of the data. All observations are represented by points in the plot, using principal component or multidimensional scaling. In our case, these two components explain 43.27% of the point variability. Around each cluster an ellipse is drawn.

The clusters found, somehow, represent the variability of scenarios found in the data set. There are three main clusters. The biggest one is mainly formed by steel based pipes from pilot scales systems, with low concentrations of free chlorine and water temperature around 15°C . All the cases are from surface water. The second one, is represented by thermoplastic pipes from single pass pilot scale systems. It presents a high variability in the rest of variables. The third big cluster is mainly formed by thermoplastic and cement pipes, also from single pass pilot scale systems. It is characterized by the fact that all the cases were sampled by low removal technique and are from surface water.

The rest of medium/small size are mainly characterized by the type of devices that have been used, suggesting that it is an influential factor to take into account.

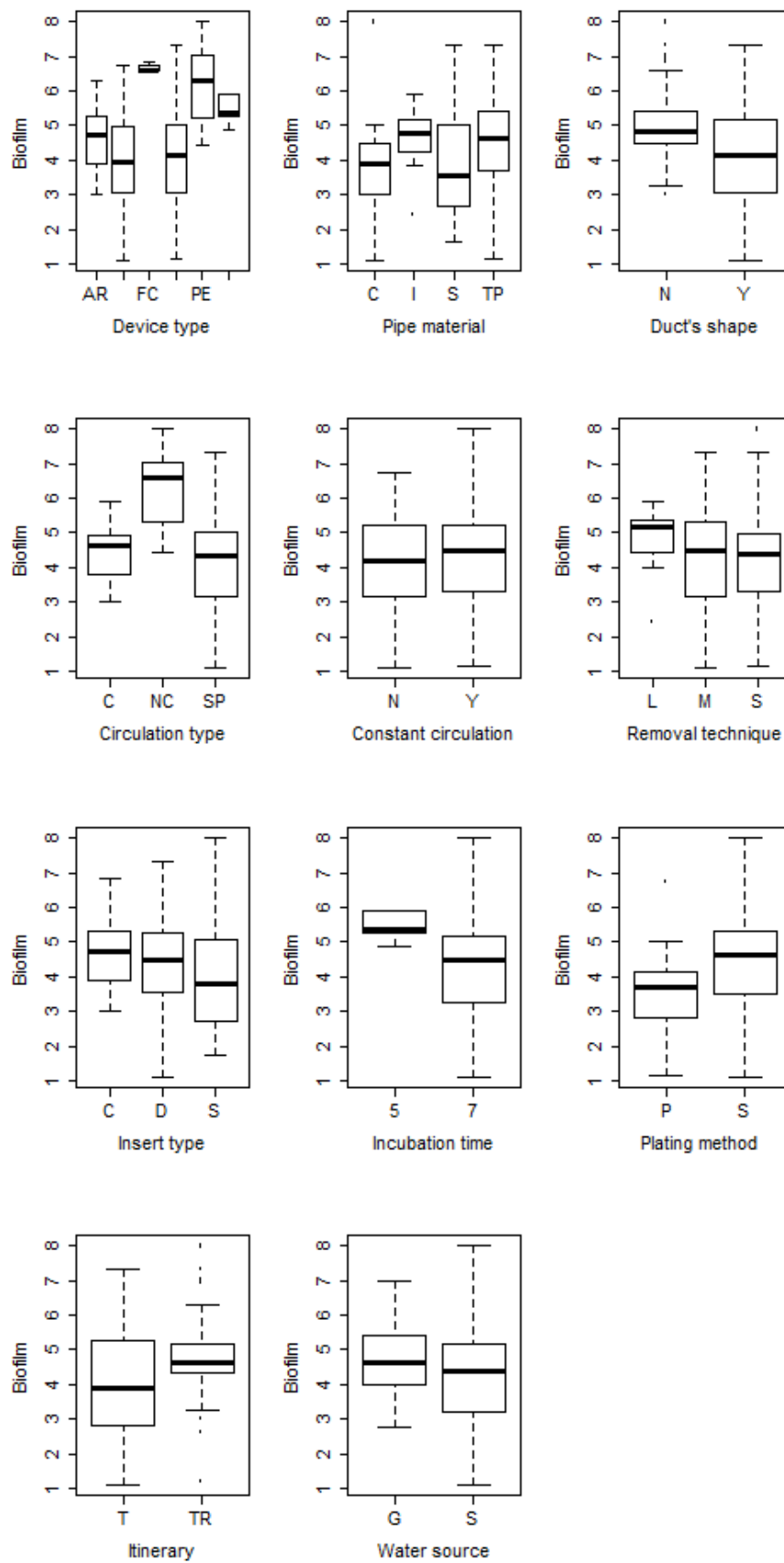


FIGURE 6.16: Boxplots of the target attribute biofilm grouped by the classes of the categorical attributes.

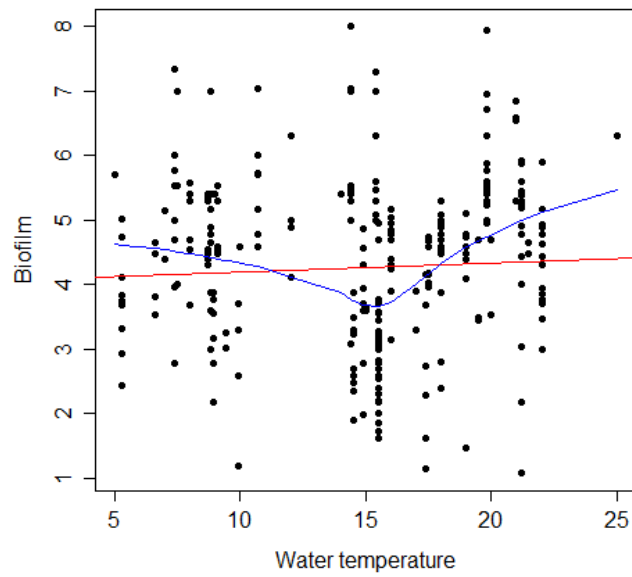


FIGURE 6.17: Scatterplot of the water temperature attribute. The red line represents the linear regression line and the blue one the LOWESS line.

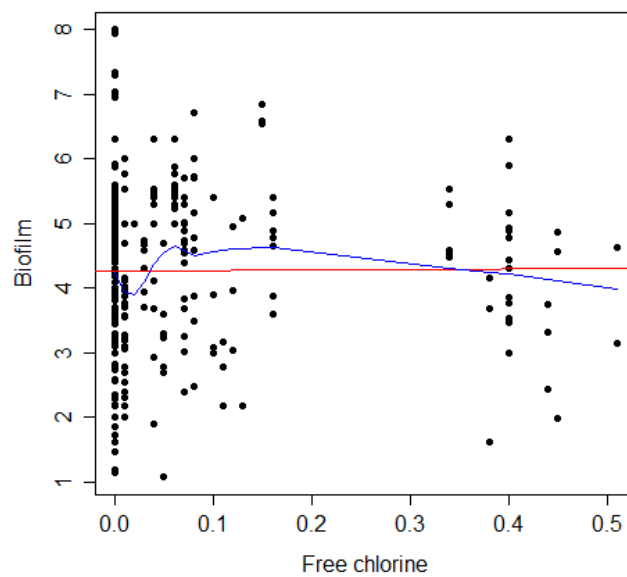


FIGURE 6.18: Scatterplot of the free chlorine attribute. The red line represents the linear regression line and the blue one the LOWESS line.

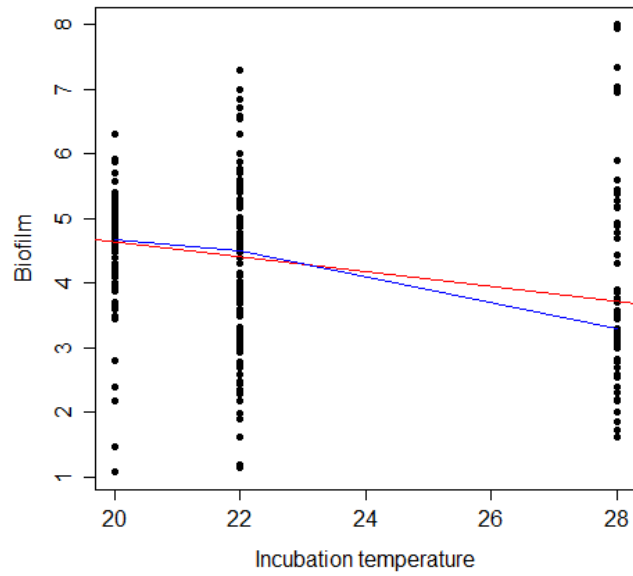


FIGURE 6.19: Scatterplot of the incubation temperature attribute. The red line represents the linear regression line and the blue one the LOWESS line.

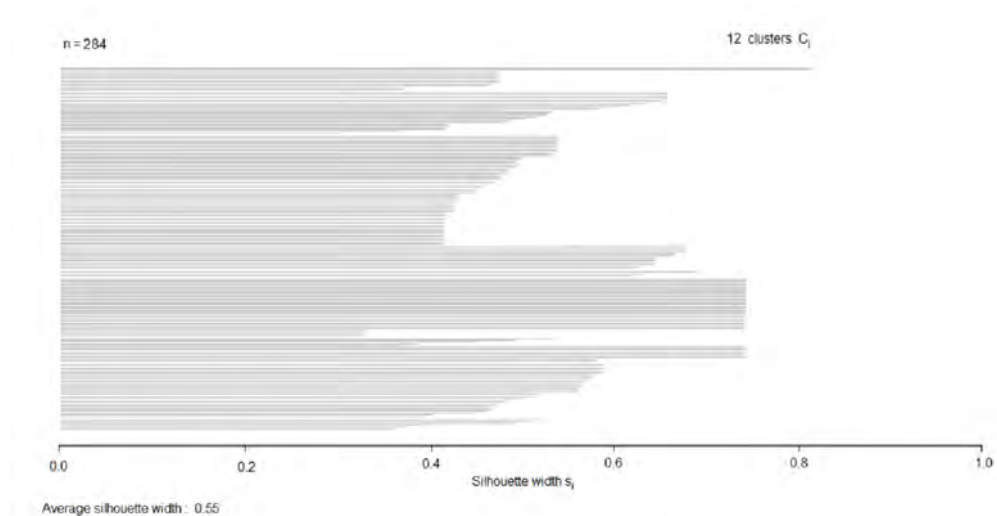


FIGURE 6.20: Average Silhouette width for 11 clusters.

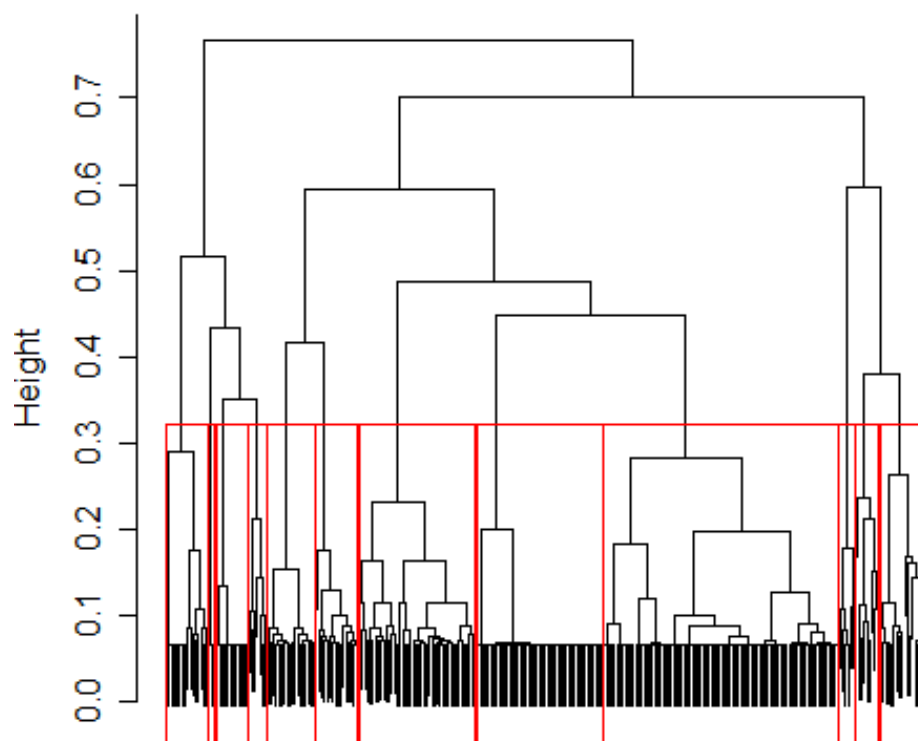


FIGURE 6.21: Agglomerative hierarchical clustering applied to the dataset. The formed clusters are grouped by the red line.

Chapter 7

Model development

Traditionally, transforming data into knowledge has been, and still is, in many situations, a matter of analysis and interpretation performed manually. This approach is slow, expensive and highly subjective, since many important decisions have to be made, not on the amount of data available, but following the intuition of the user, who does not have the necessary knowledge [204]. Nowadays, in a data-rich world, data is not only becoming more available but also more understandable to computers and analysts. Data driven solutions are rapidly advancing and becoming very valuable tools. Machine Learning (ML) methods have a leading role in this transformation of data into valid and useful knowledge. In ML, patterns and models are automatically extracted from the information provided in the databases. It is the system, not the user, that finds the hypothesis and checks its validity.

7.1 Regression Trees

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

Regression trees are machine-learning 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 [205]. The recursive partitioning algorithm is the key to the non-parametric statistical method of classification and regression trees (CART) [206]. As a result, the partitioning can be represented graphically as a decision tree. Prediction trees use the tree to represent the recursive partition. Each of the terminal nodes, or leaves, of the tree represents a cell of the partition, and has attached to it a simple model which applies in that cell only. A point x belongs to a leaf if x falls in the corresponding cell of the partition. To figure out which cell we are in, we start at the root node of the tree, and ask a sequence of questions about the involved features. The intermediate nodes are labelled with questions, and the edges or branches between them labelled with the answers [207]. Regression trees are suitable for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values [205].

For classical regression trees, the model in each cell is just a constant estimate of Y , the target vector. That is, let the points $(x_1, y_1), (x_2, y_2), \dots, (x_c, y_c)$ be all the samples belonging to the leaf-node l . Then our model for l is just $y = \frac{1}{c} \sum_{i=1}^c y_i$, the sample mean of the dependent variable in that cell. This is a piecewise-constant model [207]. There are several advantages associated to this approach [207]:

- Making predictions is fast.
- It is easy to understand what variables are important in making the prediction. Because the algorithm asks a sequence of hierarchical Boolean questions, it is relatively simple to understand and interpret the results.
- 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.

- 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).
- There are fast, reliable algorithms to learn these trees.

7.2 Regression Tree implementation

The RT analysis has been implemented through the R package ‘rpart’, version 4.1-10 [208]. It applies a recursive partitioning for regression trees [206]. The variables have been split according with their nature, by class in the case of the categorical variables and by Anova splitting the continuous ones.

It must be noticed that prior to applying the algorithm to the synthetic database a stratified sampling has been carried out in order to keep out of the model a representative amount of the data to be, subsequently, used to test the performance of the final model. The sampling has been performed with the Orange Canvas software [184] with a high random seed. The number of data kept for test are 20, thus, the analysis has been performed in the 265 remaining data. The obtained RT is presented in Figure 7.1.

The variables actually used in the tree construction have been *culture*, *device*, *freecl*, *inc-temp*, *itinerary*, *material*, *removal* and *w-temp*. That 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 pipes 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. According to the

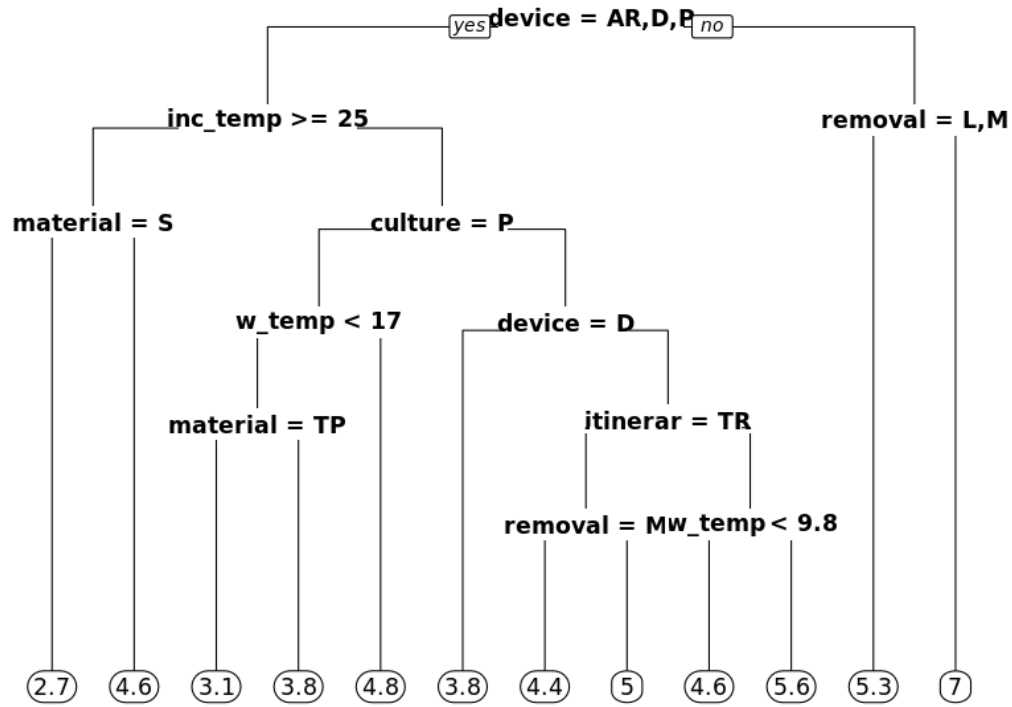


FIGURE 7.1: The obtained Regression Tree.

results in these cases, strong removal (S) leads to higher counts of biofilm than medium (M) and low (L) removal techniques.

The branch of the P , D and AR devices is further split into incubation temperature above or below 25°C . In the cases that the temperature is 25°C or more the branch finishes with one more division. It distinguishes between steel based (S) pipe materials and the rest, and assigns less biofilm development to the first type.

For the case pipe-like with incubation temperature above 25°C the next split is related with the culture technique. It has been already reported [164] that differences are found in HPC counts between the studied culture techniques (*Pour plate* and *Spread plate*).

The branch of the cases where the Spread plate technique has been used is further split. One branch includes the cases where the biofilm sample has been obtained from real DWDSs. This distinction remarks that at this point there are evident differences when obtaining the samples from operating DWDSs. When the samples have not been obtained from real DWDSs the branch is additionally split by the *itinerary* variable. The cases that are from the treatment plant (*TR*), that have not been in contact with DWDS pipes, are subsequently divided into those obtained with a medium strength (*M*) removal technique or not. If yes its value is lower. Since a similar case have been already observed in other branch, this split maybe represent the differentiation between medium (*M*) and low (*L*) strength removal techniques and the strong techniques, but there were no cases that represent the *L* cases in this branch and that is why it is not represented in the split. That is, the results are influenced by the variability of the data in the database. The cases obtained from tap water are divided between those with water temperature below 9.8°C and those above that temperature. Temperature is widely recognized as an important controlling factor in bacterial growth [167]. Thus, it is normal to observe higher biofilm development in the cases with higher water temperature.

In Figure 7.2 it can be observed how the error decreases with the size of the tree. The algorithm stops when this error do not decreases any more. This error is calculated by taking each time 10 items out of the tree and testing the regression with them.

7.2.1 Testing the Regression Tree model

The model has been tested with the metadata kept out of the model (Test 1) and with the data (Table 7.1) obtained from the study cases analysed in this work (see Chapter 4). Taking into account the design of the PWG coupons [8], specially designed to avoid any hydraulic disturbance, a *D* value has been given to the data obtained from the PWG rig in Sheffield in *insert* variable.

The performance of the model has been measured by the Pearson correlation coefficient [209]. A correlation value of $r = 0.866$ has been obtained in Test 1 (Figure 7.3) and a correlation of $r = 0.653$ in Test 2 (Figure 7.3). Although both values are good the performance is better in the first test, probably due to the fact that in the second test

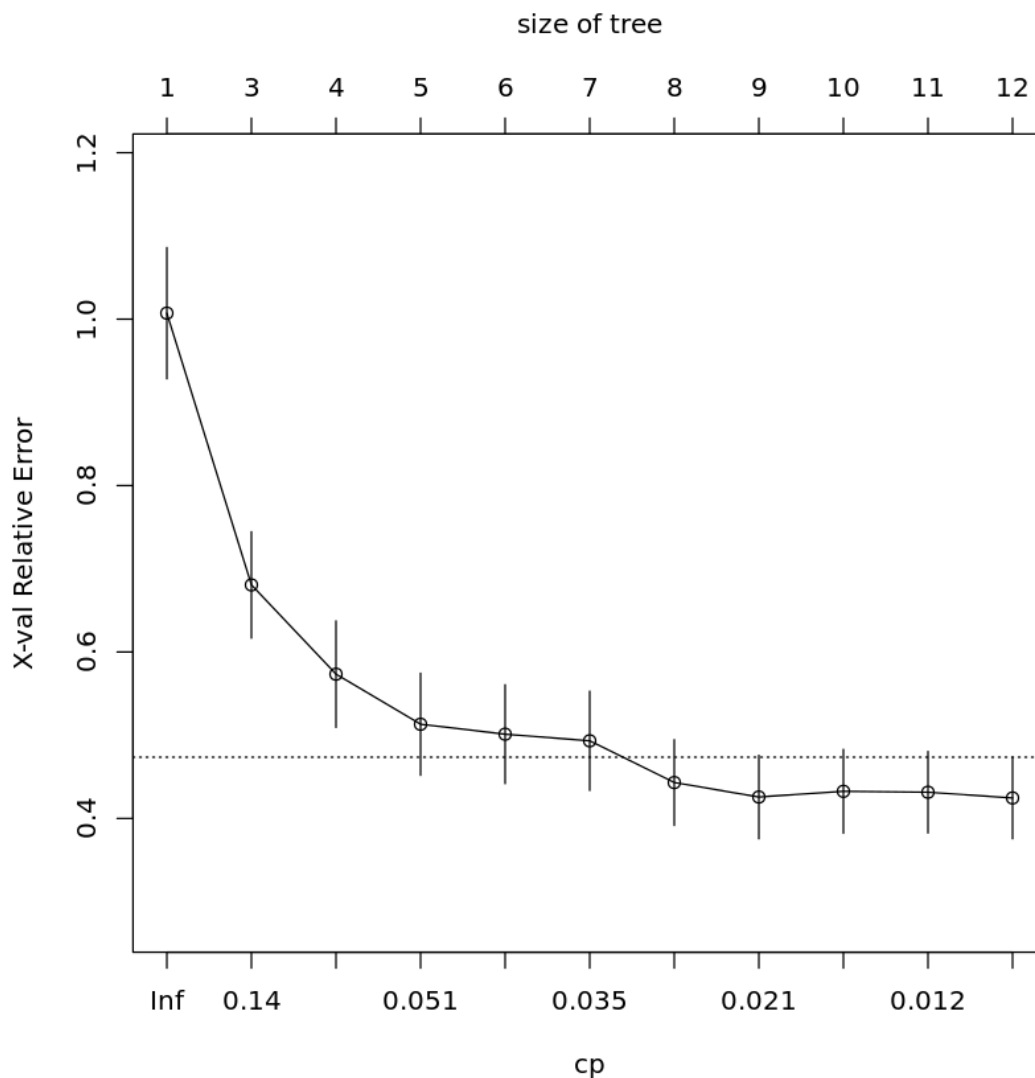


FIGURE 7.2: Cross validation of the Regression Tree.

all the cases are from real DWDSs where the variability of the conditions is bigger than in lab scale or bench top models. The good behaviour of the model can be graphically observed in Figure 7.3.

In Test 2 (Figure 7.3), the worst behaviour of the model seems to be in the 3rd, 4th, 5th and 9th values. The first three values correspond to cases with missing values; this issue may be affecting the good performance of the model. The last one corresponds to the cases with higher values of biofilm, which, although also give high prediction values, they do not reach the observations. For both cases, coming from Sheffield, the same prediction is made. However, one of them presents much lesser concentration of free

TABLE 7.1: Test data from the case studies.

Study case	Id	device	material	pipe.like	c_type	c_constant	removal	insert
Thessaloniki	T1	D	M	Y	SP	N	M	D
Thessaloniki	T2	D	M	Y	SP	N	M	D
Thessaloniki	T3	D	TP	Y	SP	N	M	D
Thessaloniki	T4	D	TP	Y	SP	N	M	D
Thessaloniki	T5	D	TP	Y	SP	N	M	D
Thessaloniki	T6	D	C	Y	SP	N	M	D
Thessaloniki	T7	D	TP	Y	SP	N	M	D
Sheffield	S1	P	TP	Y	SP	N	M	D
Sheffield	S2	P	TP	Y	SP	N	M	D

Study case	inc_time	inc_temp	culture	itinerary	w_source	w_temp	freecl	hpc
Thessaloniki	7	25	S	T	S	26.65	0.2	1.96
Thessaloniki	7	25	S	T	S	25	0.1	2.39
Thessaloniki	7	25	S	T	S	NA	NA	2.55
Thessaloniki	7	25	S	T	S	NA	NA	1.98
Thessaloniki	7	25	S	T	S	NA	NA	2.46
Thessaloniki	7	25	S	T	S	NA	NA	3.08
Thessaloniki	7	25	S	T	S	21.1	0.19	2.4
Sheffield	7	22	S	T	S	14.67	0.31	6.13
Sheffield	7	22	S	T	S	14.59	0.06	7.34

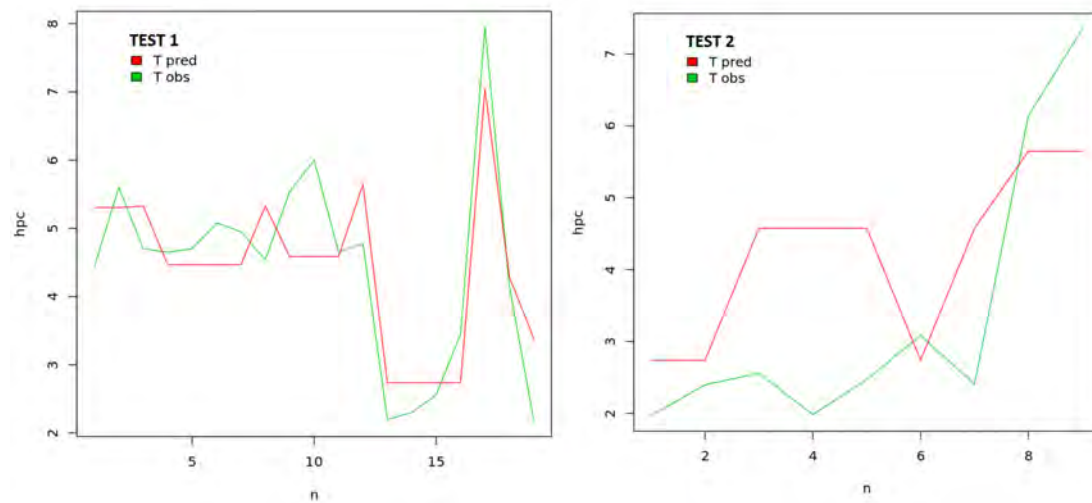


FIGURE 7.3: The performance of the Regression Tree when testing it with metadata (Test 1) and study cases data (Test 2).

residual. It seems that the model do not take this fact into account.

7.3 Random Forests

In order to try to improve the performance of the RT we have applied Random Forest (RF) algorithms. RFs are ensemble learning algorithms, meaning that they can be more accurate and robust to noise than single classifiers [210]. A random forest [211] 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 (Figure 7.4). Each tree contributes with a single vote for the assignment of the most frequent class to the input data [210]. An RF algorithm 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 [212].

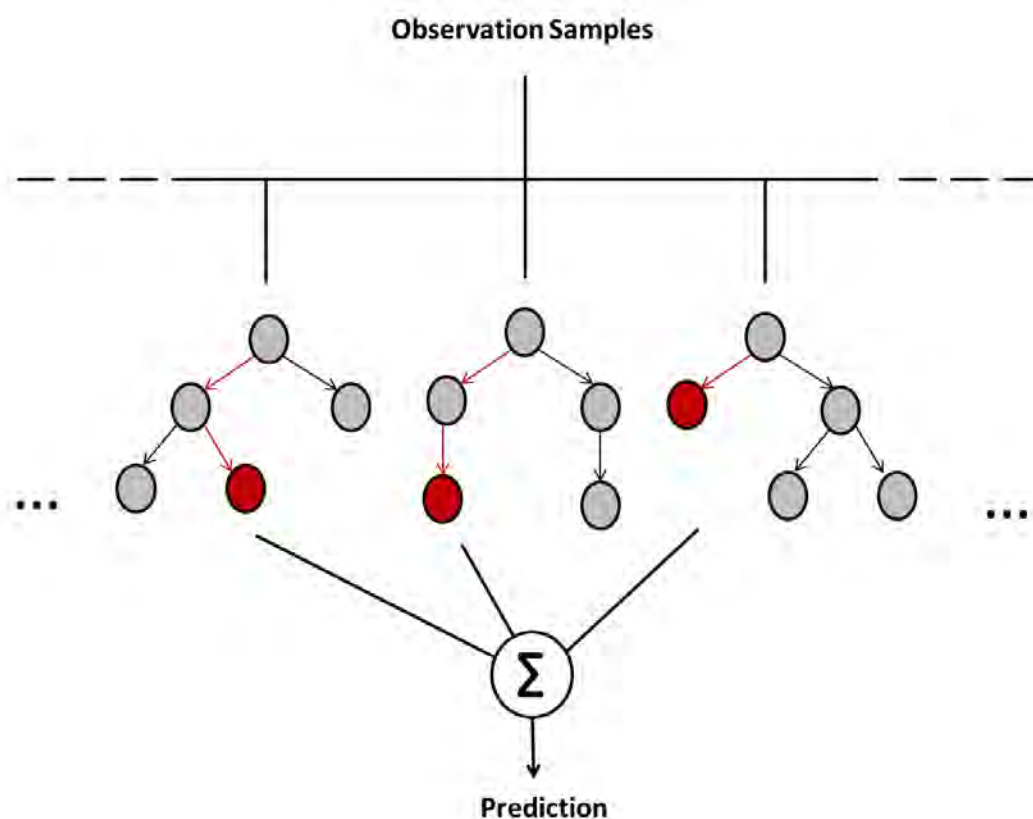


FIGURE 7.4: A Random Forest execution.

The training set for each individual tree in a random forest is constructed by sampling N examples at random with replacement from the N available examples in the dataset. This is known as bootstrap sampling. Bagging describes the aggregation of predictions from the resulting collection of trees. As a result of the bootstrap sampling procedure, approximately one third of the available N examples are not present in the training set of each tree [212]. These are referred to as the ‘out-of-bag’ data (OOB) of the tree, for which internal test predictions can be made. Note that a different OOB subset is formed for every tree of the ensemble, from the non-selected elements, by the bootstrapping process. These OOB elements, which are not considered for the training of the tree, can be classified by the tree to evaluate its performance. The proportion between the misclassifications and the total number of OOB elements contributes an unbiased internal estimation of the generalization error of the RF [210].

In summary, an RF algorithm is an all-purpose model that performs well on most problems, can handle noisy data, uses categorical or continuous features, and selects only the most important features [213].

7.4 Random Forests implementation

The Random Forest algorithm used has been implemented through the R package ‘randomForest’, version 4.6-12 [214]. 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 in 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. The obtained mean of squared residuals has been 0.561, explaining 68.96% of the variance.

%IncMSE (Table 7.2), is the increase in mean squared error (MSE) of predictions as a result of variable j being permuted. The importance of the variable increases the %IncMSE value. When looking at %IncMSE (Table 7.2), we observe that *inc_temp* is specially 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, that 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 [164] when comparing HPC results. The variable *freecl* also takes similar values to the previously mentioned variables. The free chlorine role inactivating microorganisms is well known [165]. Other quite influential variables are *itinerary*, *material*, *w_temp*, *removal* and *c_type*. Except for *itinerary*, the rest of variables are attributes that are normally studied in biofilm development in DWDS researches. The fact that experiments made in waterworks may not be generalized to the behaviour of biofilm in DWDSs is an important issue to take into account and to be further studied. The less influential studied variables are *pipe_like* and *inc_time*. The low influence of the *pipe_like* variable may be because it is partially represented through the *device* attribute. The incubation time of the samples (5 or 7 days), although influential, seems not to be very deciding.

In the same way, more useful variables achieve higher increases in node purities. This refers to splits with a high inter-node ‘variance’ and a small intra-node ‘variance’. The values of IncNodePurity (Table 7.2) can be biased. Thus, they must to be carefully treated. However, in general, similar trends to the ones described in %IncMSE are observed.

TABLE 7.2: Variable importance in Random Forest implementation.

	%IncMSE	IncNodePurity
device	25.16	55.73
material	20.81	39.99
pipe_like	9.22	9.19
c_type	17.06	34.82
c_constant	10.51	6.96
removal	18.91	24.77
insert	12.93	14.09
inc_time	4.45	0.73
inc_temp	30.58	59.59
culture	26.85	33.35
itinerary	22.25	21.33
w_temp	20.69	62.33
freecl	24.87	29.77

7.4.1 Testing the Random Forest model

The results obtained when testing the RF with the metadata kept out of the model are shown in Figure 7.5 (Test 1). A correlation value of $r = 0.898$ has been achieved, very similar to that obtained with the RT.

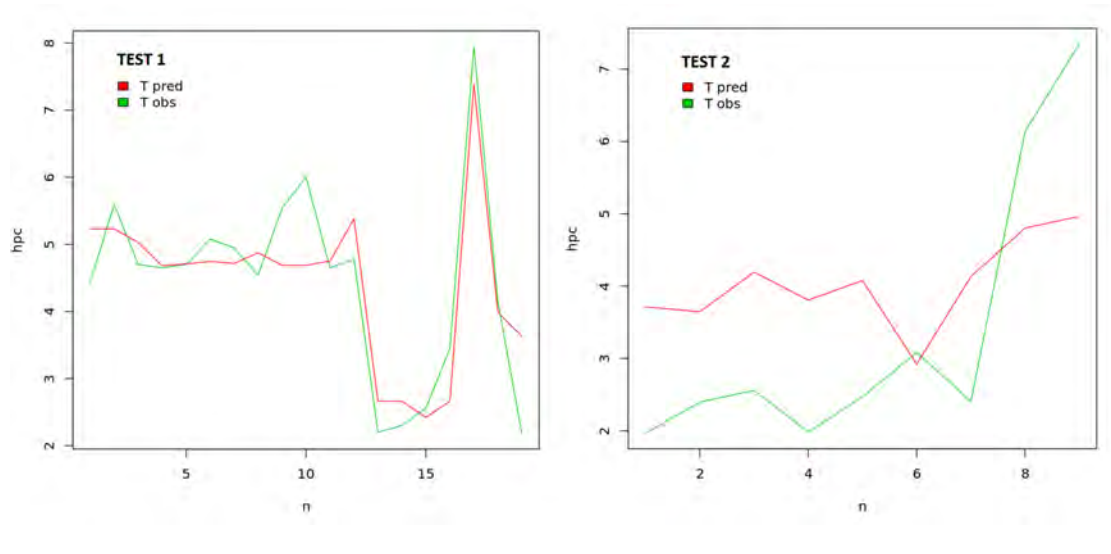


FIGURE 7.5: The performance of the Random Forest when testing it with metadata (Test 1) and case study data (Test 2).

When testing the data from the case studies (Table 7.1) the correlation value is 0.726 (Test 2 in Figure 7.5). This is a good value and higher than the one obtained with the RT. The good performance of the ensemble techniques on this approach has been already observed when applying them to biofilm metadata [32] (This work has been published as a journal paper and a summarized version is presented in Appendix C). In this case, it can be observed that the behaviour of the problematic points observed in the RT model (Figure 7.3) has improved with the RF model (Figure 7.5). In this case, the 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 thus reducing the error. In general, Figure 7.5 shows how the RF model adapts better to the tested data.

7.5 Conclusions

Although, unlike the regression tree, 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 confers to it very valuable properties that make it more robust and proper for our study. RFs perform well on the smallest datasets because re-sampling methods are inherently part of its designs [213]. They also have the ability to incorporate evidence from multiple types of learners. That is, these models divide the task into smaller portions, so they are more likely to more accurately capture subtle patterns, which a single global model might miss. Besides, since the opinions of several learners/trees are incorporated into a single final prediction, no single bias is able to dominate. This reduces the chance of over-fitting to a learning task [213]. All these facts have made that RF could get the good performance shown.

When observing the RF and RT results it seems that the cases best and worst predicted are the same in both RT and RF. This phenomenon could suggest that there are some cases that are best or worst represented in the database making their prediction more robust or, contrarily, weaker. Other possible explanation could be related to the microbial ecology of biofilm. The cases best predicted may correspond with those situations where biofilm development is mainly influenced by the studied variables, so its behaviour is well described by the model. In contrast, the prediction may be less accurate in those cases in which other factors, not taken into account in the model, are more influential. In both cases, it can be 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 *freec1*. 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.

Chapter 8

From pipe to network

Since now all the developed work has been carried out at pipe level. At this point we jump to network scale in order to be able to identify, regarding the studied variables, the most susceptible areas of the DWDSs to support higher biofilm development. This chapter provides an overview of an innovative perspective in the study of biofilm development in DWDSs. It has been applied a label negotiation, via discriminant analysis and label propagation. A multi-agent system (MAS) has been the selected tool to apply this methodology.

8.1 Multi-agent systems

A multi-agent system (MAS) consists of a population of autonomous entities (agents) situated in a shared structured framework (environment) [215]. These agents operate independently but are also able to interact with their environment, coordinating themselves with other agents (Figure 8.1) [26]. This coordination may imply cooperation if the agent society works synergically. Thus, in a cooperative community, agents have usually individual capabilities which, combined, will lead them to solving the entire problem. But cooperation is not always possible and there are instances where agents are competitive, having divergent goals. In this later case, the agents also should take into account the actions of others. However, even if the agents are able to act and achieve their goals by themselves, it may be beneficial to partially cooperate to improve performance,

thereby forming coalitions. Turning on to coordinating activities, either in a cooperative or a competitive environment, is one basic way to solve the potential conflicts that may arise among agents. These coordinating activities take place through negotiation, interactions based on communication and reasoning regarding the state and intentions of other agents [26]. There are some properties which agents should satisfy [216]: reactivity, perceiving their environment; pro-activeness, being able to take initiative; and social ability, interacting with other agents. Besides, the agents are computationally efficient because concurrency of computation is exploited as long as communication is kept minimal. We deploy agents with redundant characteristics, which offer system reliability [217]. Since the agent modularity allows handling their properties locally, this system is easy to maintain. Agents solve different problems adapting their activity on different environments by organizing themselves. The environment, which is the place where agents live, structures the multi-agent system as a whole; and manages resources and services, maintaining ongoing activities in the system and defining concrete means for the agents to communicate [26].

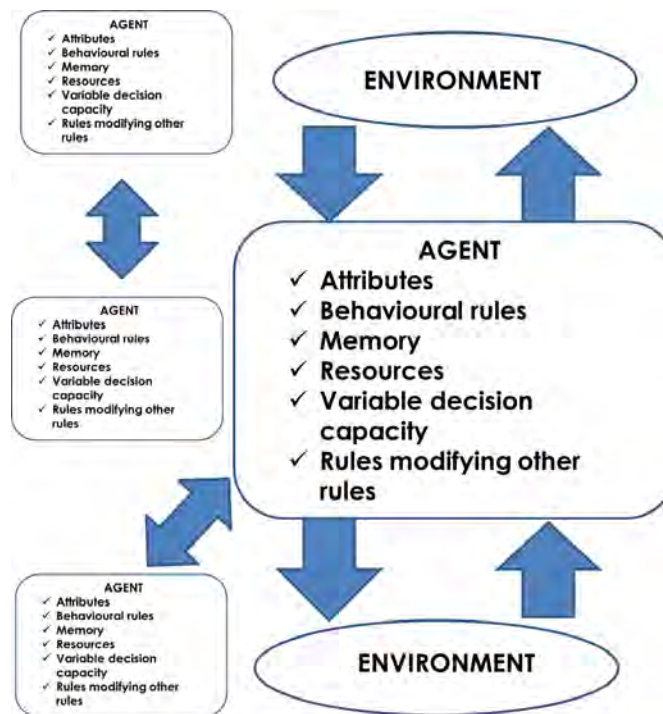


FIGURE 8.1: A multi-agent system.

Once agents have been defined and their relationships established, a schedule of combined actions on these objects defines the processes to occur, in our case, the assessment of

the vulnerability level to biofilm development [218].

8.2 Discriminant Analysis via Label Propagation

The label propagation associated with discriminant analysis clustering is used to approach a discriminant analysis in a practical case-study. Thus, pipes of a given DWDS can be classified depending on the similarities of the constructed database. Once the DWDS pipes have been classified by the aforementioned discriminant analysis, an agent-based method is launched. So, in this case, pipes properties are inherited by the nodes and node membership to the clusters are renegotiated [216, 218]. Thus, this process can be understood as a *label propagation method* methodology. Table 8.1 summarizes the process.

TABLE 8.1: Method for label propagation in practice.

MAS method for label propagation
1. Discriminant analysis based on theoretical database clustering
2. Membership negotiation
2.1. Facilitate sharing the same label by neighboring pipes for continuous variables such that: <ul style="list-style-type: none"> - have more similar <i>variable 1</i> than the average of their current cluster. - have more similar <i>variable 2</i> than the average of their current cluster. - have more similar <i>variable ...</i> than the average of their current cluster.
2.2. Facilitate sharing the same label by neighboring pipes for discrete variables such that: <ul style="list-style-type: none"> - have more similar <i>variable 1</i> than their neighboring pipes. - have more similar <i>variable 2</i> than their neighboring pipes. - have more similar <i>variable ...</i> than their neighboring pipes.
3. <i>If</i> there are not changes in last iteration <i>then</i> stop. <i>Otherwise</i> go to 2.

The agent-based model performs a mixture of individual and collective actions. It can explore good network sectorization layouts by trying to meet the equation

$$\sum_{i=1}^n \sum_{c=1}^C [\alpha_{c_n} (c_{n_i} - \bar{c}_{n_c}) + \alpha_{d_n} (d_{n_i} - \bar{d}_{n_c})], \quad (8.1)$$

where n is the number of pipes of the DWDS, C the total number of clusters and the α 's are the associated weights to each continuous (c) and discrete (d) variables and \bar{c}_c is the respective averages by cluster, and \bar{d}_c the median for the discrete variables. The

model is validated by the corresponding stabilization of this value that we attempt to minimize.

By this new complementary viewpoint of the more classical discriminant analysis, it is possible to achieve homogeneous groups where various characteristics in relation to biofilm development can be described. In addition, this new division offers an interesting starting point for further attempts to divide a given DWDS into hydraulic sectors.

8.3 Graph Theory Measurements to Assess the Importance of the Edges

Graph theory is a useful approach for the treatment of complex networks of real systems, whose techniques facilitate their representation and analysis. The framework is based on a set of measurements that enable to capture the global properties of such networks and model them as graphs. Formally, a graph $G = (V, E)$ is a pair that consist of two sets V and E , where $V \neq \emptyset$ is the set of vertices (nodes or points) $V = \{v_1, v_2, \dots, v_n\}$ and E is a set of unordered (or ordered) pairs of vertices $E = \{(v_1, v_2), (v_2, v_3), \dots, (v_j, v_k), (v_{n-1}, v_n)\}$ named edges $E = \{e_1, e_2, \dots, e_n\}$ (links or lines). In this regard, DWDSs are complex networks, which can be abstracted and analysed as graphs; the nodes would represent junctions, reservoirs, tanks and pumps, while links would be the pipes and valves. In the context of DWDSs, we are interested in knowing the structurally important edges, which might have implications on where the impact of biofilm development is higher. Below, we introduce the concept of graph theory typically used to measure edge importance, the *edge betweenness centrality*.

8.3.1 Edge betweenness centrality

Betweenness is one of the standard measurements of node centrality, originally introduced to quantify the importance of an individual in a social network. For such a reason, the concept *betweenness centrality* focus on the centrality of a node in terms of the degree to which the node falls on the shortest path between other pairs of nodes. If a node

has a high betweenness centrality, then it lies on the path of many pairs of nodes. The communication of two non-adjacent nodes, j and k , depends on the nodes belonging to the connecting paths going through it, and defining the *node betweenness*. In this regard, the Girvan-Newman algorithm (by generalizing Freeman's proposal [219]) extends this definition to the case of edges and define the *edge betweenness centrality* as the number of the *shortest paths* that go through an edge in a graph or network [220]. If there is more than one shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity. Besides, each edge in the network can be associated with an *edge betweenness centrality* value. An edge with a high *edge betweenness* score represents a bridge-like connector between two parts of a network, and their removal may severely affect the communication between many pairs of nodes through the shortest paths between them. The *edge betweenness* of edge i is defined by

$$b(e_i) = \sum_{i \neq j} \frac{n_{ij}(e_i)}{n_{ij}} \quad (8.2)$$

where $n_{ij}(e_i)$ is the number of paths from node i to node j through edge e_i , and n_{ij} is the total number of *shortest paths* of the network.

In this regard, in a DWDS a pipe with high *edge betweenness* would be between many potential upstream contamination events and downstream receptor populations [221]. Also, pipes with high *edge betweenness* could be potential locations for chlorination points or sensors.

8.4 Case Study

The Example 3 of Epanet [222] (Figure C.1 a) has been chosen as a given DWDS where to apply this methodology. With the aim of making the network as real as possible, the material and age of the pipes were randomly assigned - within the ranges indicated in Table 8.2 - depending on the average age of the area (see Figure C.1 b).

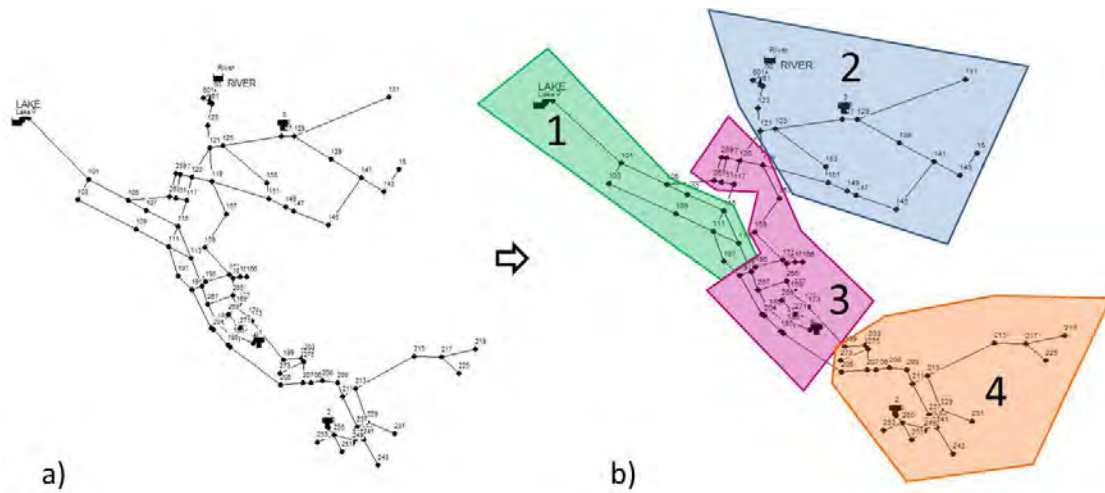


FIGURE 8.2: Areas based on pipe average age used to design the network.

TABLE 8.2: Range of ages and materials of the pipe materials.

Area	Average age (years)	Maximum age (years)	Minimum age (years)	Material 1	Material 2	Material 3
1	60	86	54	concrete	asbestos cement	iron cast
2	45	58	33	asbestos cement	iron cast	-
3	30	38	24	asbestos cement	iron cast	polyethylene
4	15	25	5	iron cast	polyethylene	-

Once the network was ready, using the obtained medoids ¹, discriminant analysis and label propagation were applied (Figure 8.3). The model has been developed in the NetLogo software [223].

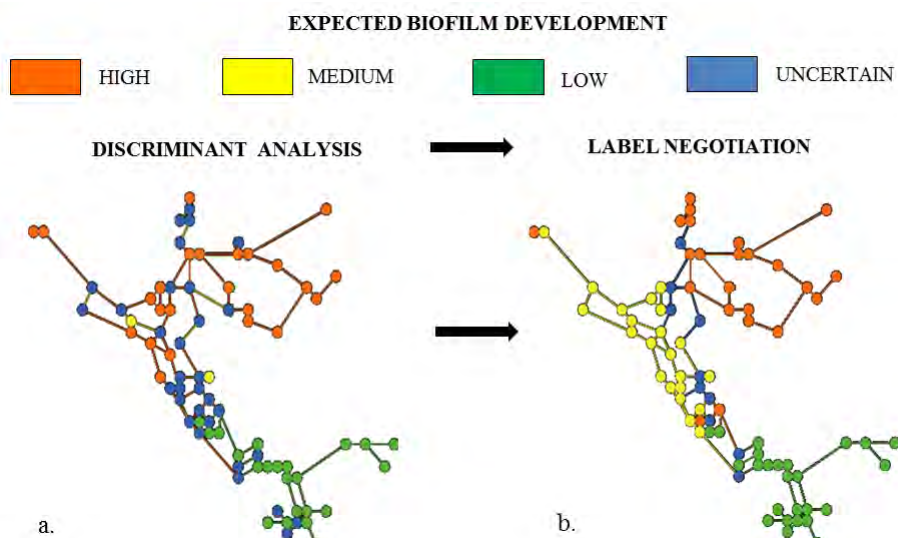


FIGURE 8.3: Results of the discriminant analysis via label propagation.

¹Due to the progress of the investigation the database used in this Chapter is an earlier version of the previously presented database

After performing the discriminant analysis (Figure 8.3 a) in the given DWDS, most of the pipes are prone to suffer high biofilm development. However after the propagation process (Figure 8.3 b) three homogeneous and clear areas associated with different degree of biofilm development appear. The area with high susceptibility to biofilm development is observed in the NorthWest zone of the network. It is an old area with no plastic pipes, that are know to support less biofilm development.

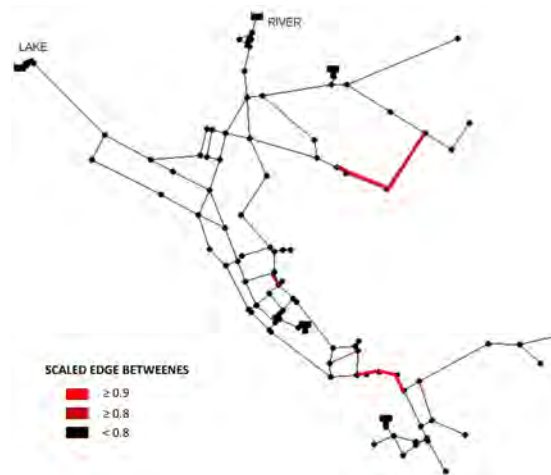


FIGURE 8.4: Results of the *edge betweenness* score.

When applying the *edge betweenness* algorithm to the network, the obtained values in each pipe were scaled to facilitate the observation of the results (Figure 8.4). It is worth to highlight that the appearance of these types of pipes in the area prone to high biofilm development raises the importance of focusing management efforts in this zone. Because of the importance of these pipes in the network operation, avoiding, as much as possible, biofilm development within them must be crucial to guarantee a service of quality in DWDSs. These highlighted pipes (Figure 8.4) are also important because they are strategic points where carrying out targeted monitoring to control the quality of the water that goes through them, developing cleaning processes to remove the biofilm adhered to its walls, as well as, locating chlorination points to reduce the development of these communities. They represent the biofilm hot spots of the network, where the management efforts must be focused.

8.5 Further application: Biofilm susceptibility as criteria for rehabilitation actions in DWDSs

We aim to detect the most susceptible locations to biofilm development within the biofilm hot spot area of the network (Figure 8.3 b) to study how just the replacement of these specific pipes could reduce the susceptibility of the whole area. We claim that this kind of approaches are the next step that have to be made in DWDS management in order to mitigate the decline of water quality in distribution systems while trying to save resources and reduce costs [28].

To find the key pipes to replace, with the aim of minimizing the area of the DWDS susceptible to high biofilm development, we identified the pipes that were found to exhibit high biofilm development in both, the discriminant analysis and label propagation. After that, according to the results of the clustering and the bibliography, we selected the metal pipes which are known to tend to support more biofilm development [224]. Among them, the older pipes were selected, obtaining the pipes susceptible to be replaced. The accumulation of corrosion and dissolved substances in older pipes can increase their roughness and a rough surface has greater potential for biofilm growth [84]. The replaced pipes would be substituted by new plastic pipes that, as found in the clustering process and in the bibliography, are the ones less susceptible to present biofilm development.

After the label propagation, an area with high susceptibility to biofilm development is observed in the North-West zone of the network. We focus on this area and look for the pipes that were found to present high biofilm development in the discriminant analysis. Then we select the metallic ones that meet this requirement. Finally, we obtain 9 pipes susceptible to be replaced (Figure 8.5).

With the aim to try to save resources, we have decided to start studying the variations in the area susceptible to high biofilm development replacing first the shortest pipe (Figure 8.5) and adding pipes, one by one, since arriving to the longest one (Figure 8.6). The results (Figure 8.7) show that as the pipes are replaced the number of pipes susceptible

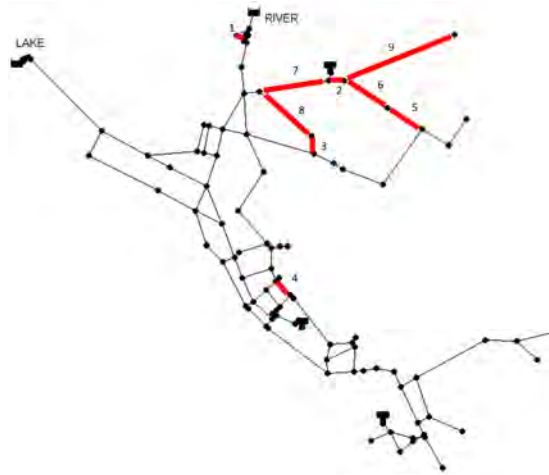


FIGURE 8.5: Pipes susceptible to be replaced.

to support high biofilm development decreases. However, it is observed that after the fourth replacement a stabilization in the number of pipes susceptible to high biofilm development occurs. In the last replacements (8^{th} and 9^{th}) a reduction in the number of pipes is observed again. This suggests that the replacement of some pipes is more influential than the replacement of others. Certainly, the spatial position in the network of pipes has an important role.

Although the replacement criteria implemented in this paper are just an approach, in the studied network the incidence of pipes susceptible to support high biofilm development has been reduced from 25% to 10% (Figure 8.7). As a result, the risk of developing high biofilm development has decreased.

8.6 Conclusions

A new methodology is developed where data mining techniques and multi-agent systems are integrated in order to assess the susceptibility to biofilm development of homogeneous groups of pipes where various characteristics in relation to biofilm development can be described. It has been shown that label negotiation via discriminant analysis and label propagation as interesting tools enable the use of knowledge gained in the development of biofilm in DWDSs in a practical and efficient manner. This methodology enables an advanced visualization of the case-study database. According to the results obtained in

this work, there are some areas within a DWDS more vulnerable to support high biofilm development, thus, biofilm is not uniform in space.

In the same way, the introduction of the *edge betweenness* score has demonstrated to be of great help to improve the efficiency of DWDS management. Thanks to it the most problematic pipes can be easily detected. These pipes represent the critical elements of the network. Thus, special attention must be focused on these elements to prevent its deterioration and mitigate, as much as possible, the negative effects derived of biofilm development in DWDSs. Beside, the effect of pipe replacement is studied in order to observe the influence on the susceptibility of DWDSs to biofilm development. An example of replacement criteria is applied and a reduction from the 25% to the 10% in the incidence of high biofilm development has been observed. However, this is just an approach and much more work must be done in this area, in order to optimize, as much as possible, the invested resources and the obtained benefits. The results obtained in this work suggest that the replacement of some pipes is more influential than the replacement of others, probably due to their spatial position in the network. The importance of this characteristic must be more deeply studied.

In summary, in this chapter the effect that rehabilitation actions in a DWDSs would have on biofilm development trends and how helpful they could be to reduce the susceptibility of these systems to the development of these microbial communities have been analyzed. Although more work has to be done in this direction, we claim that this kind of new approaches could represent a clear improvement in the future of DWDS management.

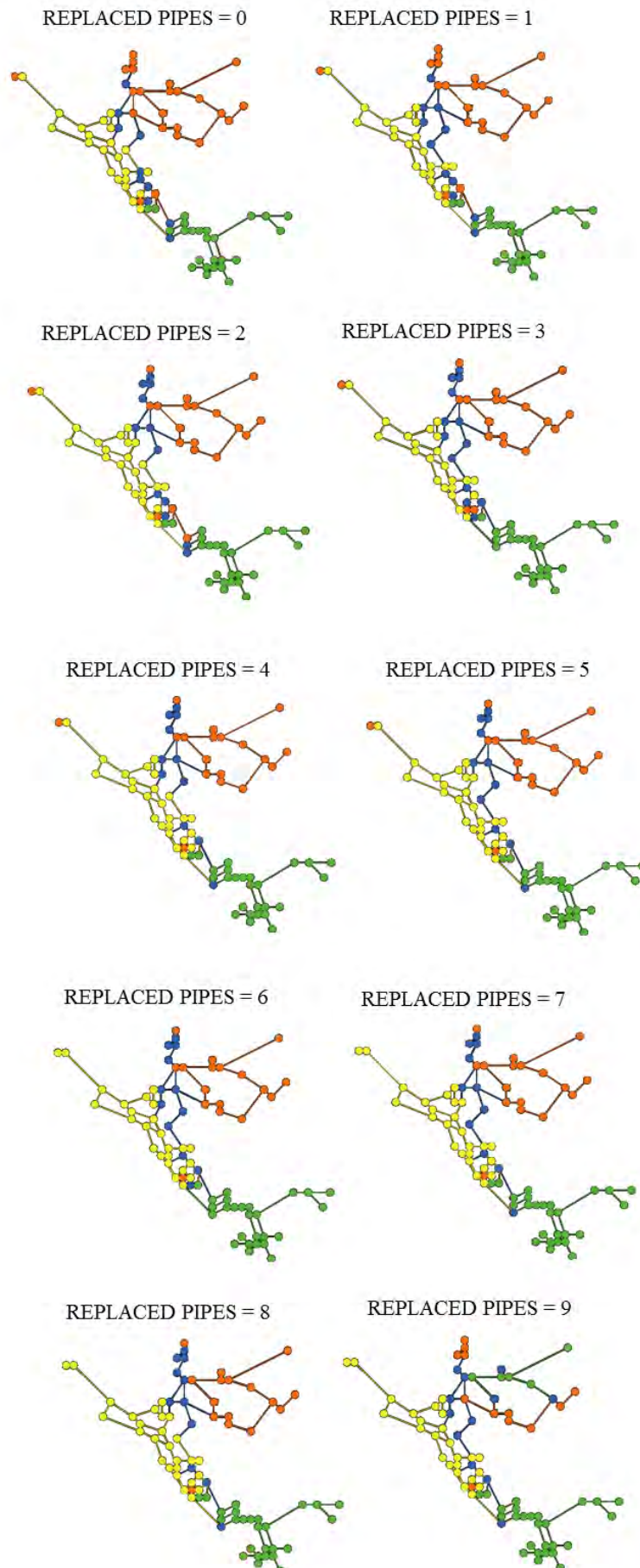


FIGURE 8.6: Biofilm susceptibility after progressive pipe replacement.

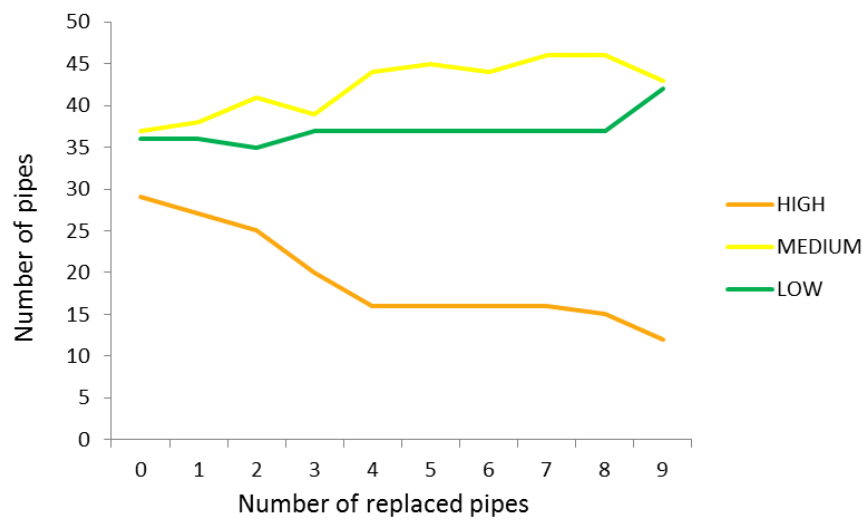


FIGURE 8.7: Evolution of biofilm susceptibility when replacing pipes.

Chapter 9

Conclusions and Future Work

An important part of engineering is about solving human-made problems. In this endeavour scientific understanding, the laws of physics, chemistry, biology, etc., and the formulations of mathematics are applied to effect as appropriately as possible. Engineering is multi-disciplinary, even transdisciplinary, and like many other disciplines continually evolves to become even more relevant and effective as a practical approach to achieving worthwhile objectives [225]. This is the context in which this thesis has been developed.

Biofilm development in DWDSs is a real problem negatively affecting the service and water quality offered by water utilities, 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, but it is now that technology and data have been available to support the new approach that we present in this thesis. Through the combination of various disciplines we have gathered knowledge and works carried out in this field and developed a multidisciplinary approach based, mainly, on an intensive preprocessing and the implementation of Machine learning (ML) techniques. We develop a practical decision-making tool to assist in DWDS management in order to maintain, as much as possible, biofilm at the lowest level, thus mitigating its negative effects on the service and on the consumers.

9.1 Merits of the new approach

This work proposes data preprocessing 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. This compilation represents a hard task for the researcher that should merge and preprocess data from different sources for posterior analysis. Data science, an interdisciplinary field to extract knowledge from data, is a hard and challenging discipline because it requires expertise in a broad range of subjects and technologies. Various formal process models have been proposed for knowledge discovery and machine learning, as reviewed by [226]. These models estimate the data preprocessing stage to take 50% of the overall process effort, while the data mining task takes less than 10-20%. The high workload required to achieve this preprocessing is reflected in the arduous work that has been developed in Chapter 5 of this thesis. However, the step forward that could represent this new approach in this field is huge.

Data preprocessing is required in all knowledge discovery tasks. Our proposal is to achieve preprocessing of all the work already developed in this field, preparing a case-study database to do inferences by posterior ML analyses. Thanks to it, we can develop a scalable and interesting set of tools to understand biofilm behaviour respect its environment and develop models that can be used as decision-making tools in DWDS management to mitigate its negative effects on the service.

The benefits of implementing ML algorithms are huge. ML is a subfield of computer science related to the artificial intelligence. It is the systematic study of algorithms and systems that improve their knowledge or performance with experience. That is, the ML models are built from example inputs to get data-driven predictions. In a data-rich world, data-driven solutions are suffering a rapid evolution, increasing their sophistication and enhancing their performance. In summary, these techniques are making data more understandable to computers and analysts. ML algorithms are able to make intelligent decisions, modify themselves and make multiple iterations of the model in order to get the highest accuracy. ML allows to perform highly sophisticated pattern recognition.

The implementation of this family of techniques in the study of biofilm development in DWDSs opens a vast field to explore with promising results. Some of the possibilities that these techniques offer have been presented in this thesis, obtaining very good results (see Chapter 7).

In this dissertation the benefits of combining these ML techniques with modelling and more visual techniques, such as multi-agent systems (MASs) [215] have also been presented. Visualization is a natural way to identify patterns because human mind excels in prompt interpretation of visual information [192] and it plays a crucial role. This combination of techniques allows a rapid and easy interpretation of the obtained results. It makes more appealing the application of these techniques and more implementable in operating utilities, since it makes not necessary the presence of a data scientist to interpret the results. Thus, the developed tools can become daily management tool in DWDS management.

9.2 Practical implications

Nowadays, regarding biofilm development in DWDSs, there is a need for a deeper understanding of how the large spectrum of conditions interacts and affects biofilm formation potential and accumulation with the final purpose of predicting the total and cultivable bacteria attached to real DWDS pipes, based on the system characteristics [227]. We believe that the methodology and the models that are presented in this work represent a step forward necessary to achieve this final aim. This could be the beginning for a new paradigm in the study of biofilm development in DWDSs and its management in the water utilities.

- The large number of variables that are affecting biofilm development can be analysed and its importance evaluated. Thus, studies could offer a global vision of the biofilm environment, where the physico-chemical water characteristics and the physical and hydraulic conditions of the systems are taking into account, thus avoiding a biased perception of the reality. The possibility of studying a large spectrum of variables makes it possible to analyse the influence that the sampling

and incubation conditions have in the final obtained bacteria count. Knowing how these variables, related to the samples obtaining and manipulation, affect the results could represent a serious incentive to standardize these procedures, or strictly follow the protocols that already exist [228].

In summary, nowadays, there is a lack of unified and consensus criteria to be followed, as it has been observed in the number of papers that have been discarded due to this issue during the preprocessing in Chapter 5. Being able to study the system as whole would enable to take into account a higher number of variables and emphasise their importance.

- Having a tool able to detect the most susceptible areas to biofilm development in DWDSs offers a huge variability of applications that can be implemented to improve the service in these systems. In this work, one of this possible applications is further studied (See Chapter 8), namely the effect that the pipe replacement criteria can have in the extension of these susceptible areas. However, there are much more applications that could be developed. Some of them are presented below.
 - This tool can be very useful in the prevention and maintenance works of the supply networks. On the one hand, knowing which are the areas more prone to biofilm development, directed flushing can be undertaken and thereby, thus saving invested time and money and increasing the process efficiency. Moreover, taking into account the fact that biofilm can increase the rates of corrosion in metal pipes, this tool can also help to improve the efficiency of damage prevention methods and reduce leaks and service failures in the network.
 - Likewise, the implementation of this tool can be hygienically relevant as biofilm is involved in the consumption of residual disinfectant in DWDSs. Knowing the tendency of each pipe or sector of pipes to biofilm development can be useful for optimizing disinfectant consumption modelling in the wall pipe. It could help to achieve a greater precision when locating the chlorination points.

- Also of note, the usefulness of this tool is relevant in the design of distribution networks. The susceptibility to biofilm development could be taken into account in this previous phase and, as far as possible, the existence of problematic areas could be avoided in future DWDSs.

In short, the implementation of a tool which can give us an idea of the expected biofilm development would help to effectively mitigate the negative effects associated with biofilm development in DWDSs, improving the quality of the service and the tap water, while reducing the costs. It could be a very helpful decision support system enhancing the efficiency and efficacy in these systems' management.

9.3 Future perspectives

This thesis proposes some approaches to follow in the future. All of these lines are related with keeping improving and validating the obtained models and tools. Specially, it is intended to obtain accessibility to test if the good results obtained at pipe level are maintained at network level. In order to get the attention and interest of water utilities stakeholders in the developed network model a web page has been developed (Figure 9.1). This web has been designed as a research outreach tool.



FIGURE 9.1: QR code of the web page.

In order to make more appealing this project and get stakeholders attention, an informative model of the biofilm developing process in pipes has been created [26]. An

agent-based modelling environment has been used with simulation purposes. The model has been developed in the NetLogo framework [223] (see Appendix D). This is one of the most popular agent-based modelling tools for environmental science and ecology. This model has been cast into a video that has been uploaded to Youtube (<https://youtu.be/cIxorP81fBo>) and embedded in the web.

Intro Biofilm in DWDSs More

Biofilm development in DWDSs

One of the main challenges of drinking water utilities is to ensure high quality supply, in particular, in chemical and microbiological terms. However, biofilms invariably develop in all drinking water distribution systems (DWDSs), despite the presence of residual disinfectant. As a result, water utilities are not able to ensure total bacteriological control.

CURRENTLY BIOFILM REPRESENTS A PARADIGM IN WATER QUALITY MANAGEMENT FOR ALL DWDSs

FluIng **MODELING BIOFILMS FORMATION AND EVOLUTION IN DRINKING WATER DISTRIBUTION SYSTEMS BY MULTI-AGENT SYSTEMS** Universitat Politècnica de València **IMAFluIng** Eiva Ramon Martínez **UNIVERSITAT POLITÈCNICA DE VALÈNCIA**

Let's work together to solve it!

FIGURE 9.2: The NetLogo model embedded in the web page.

Through this web page it is also intended to enhance the networking and get in contact with others researchers interested in this field. Collaborating with another research groups would be the perfect way to keep enlarging the present database. The more cases are represented in the database, the greater the performance of the final model. The project has been entitled “Biofilm for All” (BfA) and the web platform would be used as a repository to share biofilm data at international level (Appendix E). In the web page a detailed description of the project can be found as well as the up-to-now obtained results and publications (Appendix E). There is also a section where the contact details (Appendix E) of the FluIng research group (<https://fluIng.upv.>

es), of the Universitat Politècnica de València, that hosts this project are detailed and the academics/stakeholders interested in collaborating or that require more information about the project can get in contact with us.

9.4 Final conclusions

This work offers an innovative perspective of work in the study of biofilm in DWDSs. Data science techniques are introduced in the study of development of biofilm in DWDSs. The importance of multidisciplinary and the need for a shift to more practical and real life implementable approaches is highlighted. In this way, the benefits that a consensus in biofilm sampling and analyzing procedure would report are also demonstrated.

The developed methodology, helps to understand how the number of conditions interact and affect biofilm formation in DWDSs. Good performance values have been obtained when predicting the cultivable bacteria attached to real DWDS pipes. The extension of this model to network scale opens the possibilities of implementing a variety of cost-effective procedures that would significantly improve the quality of the service and of the distributed water in DWDSs. Keep working in this direction could mean a great step forward in biofilm management in DWDSs.

Appendix A

Compiled variables with less than
the 15% of data

TABLE A.1: Compiled variables with less than the 15% of data.

Hydraulic characteristics	
Shear stress	Expressed in Pa
Hydraulic retention time	Expressed in h-1
Physico-chemical characteristics of the water	
Turbidity	Expressed in NTU
Conductivity	Expressed in $\mu\text{S}/\text{cm}$
Oxygen	Expressed in $\text{mg O}_2/\text{l}$
Biodegradable dissolved organic carbon	Expressed in $\text{mg C}/\text{l}$
Dissolved organic carbon	Expressed in $\text{mg C}/\text{l}$
Inorganic carbon	Expressed in $\text{mg C}/\text{l}$
Biodegradable organic matter	Expressed in $\text{mg C}/\text{l}$
Total dissolved solids	Expressed in $\text{mg C}/\text{l}$
Ammonia (NH_3)	Expressed in $\text{mg N}/\text{l}$
Ammonium (NH_4^+)	Expressed in $\text{mg N}/\text{l}$
Nitrogen dioxide (NO_2^-)	Expressed in $\text{mg N}/\text{l}$
Nitrate (NO_3^-)	Expressed in $\text{mg N}/\text{l}$
Total phosphorus	Expressed in $\text{mg P}/\text{l}$
Phosphate (PO_4^{3-})	Expressed in $\text{mg P}/\text{l}$
Monoammonium phosphate ($\text{NH}_4\text{H}_2\text{PO}_4$)	Expressed in $\mu\text{g}/\text{l}$
Sulphate (SO_4^{2-})	Expressed in mg/l
Silicon dioxide (SiO_2)	Expressed in mg/l
Calcium	Expressed in mg/l
Magnesium	Expressed in mg/l
Sodium	Expressed in mg/l
Iron	Expressed in mg/l
Manganese	Expressed in mg/l
Aluminium	Expressed in mg/l
Zinc	Expressed in mg/l
Bicarbonate (HCO_3^-)	Expressed in mg/l
Calcium carbonate (CaCO_3)	Expressed in mg/l
Bacteria	
Total cell in water	Expressed in $\log \text{ cell}/\text{ml}$

Appendix B

Extract of the first 50 elements of
the synthetic database

TABLE B.1: Extract of the first 50 elements of the synthetic database.

device	material	pipe_like	c_type	c_constant	removal	insert	inc_time	inc_temp	culture	itinerary	w_source	w_temp	freel	hpc	
1	AR	TP	N	SP	Y	S	C	7	28	S	TR	S	15.85	0.00	4.19
2	P	C	Y	SP	Y	S	D	7	22	P	T	S	14.63	0.05	4.00
3	P	TP	Y	SP	Y	S	D	7	22	P	T	S	17.40	0.38	3.39
4	P	TP	Y	SP	Y	M	D	7	22	S	TR	G	7.05	0.00	5.96
5	P	TP	Y	SP	Y	S	D	7	22	P	T	S	14.90	0.45	4.26
6	AR	S	N	SP	Y	S	C	7	28	S	TR	S	15.85	0.00	3.33
7	D	C	Y	SP	N	S	D	7	22	P	T	S	14.63	0.08	6.42
8	P	TP	Y	SP	Y	M	D	7	20	S	TR	S	17.70	0.00	4.51
9	RD	TP	Y	SP	N	L	S	5	20	S	T	S	10.70	0.00	5.52
10	D	TP	Y	SP	N	S	D	7	22	P	T	S	14.50	0.51	3.15
11	P	C	Y	SP	Y	S	D	7	22	P	T	S	14.90	0.45	4.86
12	P	S	Y	SP	Y	M	S	7	28	S	T	S	15.50	0.00	3.05
13	P	TP	Y	SP	Y	S	D	7	22	P	T	S	14.70	0.05	3.48
14	D	I	Y	SP	N	L	D	7	20	S	T	G	10.70	0.01	5.04
15	D	C	Y	SP	N	M	D	7	20	S	T	G	10.70	0.03	5.00
16	P	TP	Y	SP	Y	S	D	7	22	P	T	S	8.90	0.01	3.72
17	P	C	Y	SP	Y	S	D	7	22	P	T	S	17.47	0.01	4.32
18	FC	TP	N	NC	Y	S	C	7	22	S	T	S	21.00	0.15	6.68
19	P	C	Y	SP	Y	S	D	7	22	P	T	S	5.30	0.44	3.76
20	P	C	Y	SP	Y	S	D	7	22	P	T	S	17.40	0.38	4.16
21	D	C	Y	SP	N	M	D	7	20	S	T	G	10.70	0.13	2.18
22	P	C	Y	SP	Y	S	D	7	22	P	T	S	8.90	0.01	3.87
23	P	TP	Y	SP	Y	M	D	7	22	S	TR	G	9.10	0.34	5.12
24	RD	S	Y	SP	N	L	S	5	20	S	T	S	10.70	0.00	5.70
25	AR	I	N	C	Y	S	C	7	28	S	T	S	22.00	0.40	5.16
26	P	TP	N	SP	Y	S	S	7	20	S	TR	S	18.28	0.00	4.83
27	D	I	Y	SP	N	S	D	7	22	P	T	S	14.50	0.51	4.64
28	P	TP	Y	SP	Y	S	D	7	22	P	T	S	14.50	0.11	2.19
29	P	S	Y	SP	Y	S	D	7	28	S	T	S	15.50	0.00	2.81
30	D	TP	Y	SP	N	M	D	7	20	S	T	G	10.70	0.00	3.46
31	D	S	Y	SP	N	M	D	7	20	S	T	G	10.70	0.13	5.08
32	P	TP	Y	SP	Y	S	D	7	22	P	T	S	17.47	0.01	4.28
33	P	TP	Y	SP	Y	S	D	7	22	P	T	S	5.30	0.44	3.08
34	D	C	Y	SP	N	S	D	7	22	P	T	S	8.90	0.11	4.32
35	AR	S	N	C	Y	S	C	7	28	S	T	S	22.00	0.40	3.42
36	P	TP	Y	SP	Y	M	D	7	22	S	T	S	10.70	0.08	5.58
37	P	TP	Y	SP	Y	M	D	7	22	S	TR	S	9.30	0.00	4.68
38	P	TP	Y	SP	Y	S	D	7	22	P	T	S	5.30	0.06	4.23
39	PE	S	N	NC	Y	M	S	7	28	S	T	S	21.20	0.00	5.25
40	P	C	Y	SP	Y	S	D	7	22	P	T	S	5.30	0.06	4.77
41	P	S	N	SP	Y	S	S	7	20	S	TR	S	18.28	0.00	4.87
42	P	C	Y	SP	Y	S	D	7	22	P	T	S	14.57	0.11	2.43
43	P	TP	Y	SP	Y	S	D	7	22	P	T	S	8.90	0.11	3.02
44	P	TP	Y	SP	Y	M	D	7	22	S	T	S	19.80	0.06	5.69
45	D	I	Y	SP	N	L	D	7	20	S	T	G	10.70	0.07	2.40
46	AR	S	Y	SP	Y	S	C	7	20	S	TR	S	12.78	0.00	5.57
47	P	TP	Y	SP	Y	M	D	7	22	S	T	S	14.90	0.02	6.42
48	P	TP	Y	SP	Y	M	D	7	22	S	T	G	7.40	0.08	5.61
49	PE	S	N	NC	Y	S	S	7	28	S	TR	S	18.28	0.00	7.73
50	P	TP	Y	SP	Y	M	D	7	22	S	T	G	8.80	0.16	4.93

Appendix C

Ensemble of naïve Bayesian approaches for the study of biofilm development in drinking water distribution systems

C.1 Naïve Bayesian approaches

This paper focuses on naïve bayesian methods and a number of variants in order to assess the biofilm development degree in DWDSs. A naïve Bayesian network classifier, which is sometimes called naïve Bayes classifier (NBC for short), has a very simple structure while its classification performance in practice is surprisingly high. The structure assumes that all the attributes are mutually independent given the class. This simplify the way in which the process works.

Let T be a training set of samples, each with their class labels. There are k classes, C_1, \dots, C_k . Each sample is represented by an n -dimensional vector, $\mathbf{X} = \{x_1, \dots, x_n\}$, depicting n measured values of the n attributes. Then, the classifier will predict that \mathbf{X} belongs to the class having the highest *a posteriori* probability, conditioned on \mathbf{X} (see Equation C.1).

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \text{ for } 1 \leq j \leq n, j \neq i. \quad (\text{C.1})$$

The probabilities involved in this model can be approximately calculated using Equation C.2.

$$P(C_h|\mathbf{X}) \propto P(C_h) \prod_{i=1}^n P(X_i|C_h), \quad (\text{C.2})$$

where $P(C_h)$ represents the *a priori* information with respect to the classification of the variable of interest in the class h .

In order to predict the corresponding class of \mathbf{X} , the expression $P(C_i)P(\mathbf{X}|C_i)$ is evaluated for each class C_i . The classifier predicts that the class label of \mathbf{X} is C_i if and only if it is the class that maximizes $P(C_i)P(\mathbf{X}|C_i)$. Thus, a final classifier is obtained by Equation C.3.

$$\arg \max_c P(C) \prod_{i=1}^n P(X_i = x_i|C = c). \quad (\text{C.3})$$

Despite the fact that the far-reaching independence assumptions are often inaccurate, an NBC has several properties that make it exceptionally useful in practice. In particular, the decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This, for example, helps alleviate problems stemming from the curse of dimensionality and also allows working with missing and scarce data.

C.1.1 Augmented Bayesian Classifiers

The tree augmented naïve (TAN) classifier [229] is obtained by allowing each attribute to have at most one other attribute as a parent, in addition to the class. Therefore a maximum of $n - 1$ number of edges can be added to an NBC to obtain a TAN classifier. Then, this algorithm outperforms the accuracy of the naïve Bayes algorithm by relaxing the conditional independence assumption [230].

In order for the algorithm to be computationally efficient, Keogh & Pazzani [230] proposes the following approach for each TAN classifier to be built. In the first step, the results of equation C.2 are stored in a $J \times I$ matrix, (J is the number of instances in the training set, I is the number of distinct classes) where each element is the probability that example j belongs to class C_i . When testing a new classifier that has an arc from node X_b to node X_a , we adjust the matrix by multiplying element (i, j) by

$$\frac{P(X_a = x_{a_j} | C_i, X_b = x_{b_j})}{P(X_a = x_{a_j} | C_i)}. \quad (\text{C.4})$$

This approach means that the time taken to evaluate one instance of a TAN classifier will be independent of the number of attributes. So, the speed-up achieved by this optimization is approximately of order n , the number of nodes.

C.1.2 A combined approach: bagging naïve bayes

Bootstrap aggregating, *bagging*, predictors are used to generate multiple versions of a predictor that are then used to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets [231]. Bagging then weighs classifiers generated by different bootstrap samples: S_1, \dots, S_B . From each sample S_i a classifier is induced by the same learning algorithm (NBC in this case). Classifiers obtained in this manner are then combined by majority voting respect to the B classifiers (see Figure C.1). This aggregation process helps mitigate the impact of random variation and provides stability to the classifier method [232].

The procedure, iterated for B bootstrap samples, results in an ensemble of B NBCs, each one with a possibly different set of features. Unseen subjects are then classified by making each NBC estimate output class probabilities, and by averaging the probabilities across all B NBCs. Such an approach increases the robustness of the predictions [231].

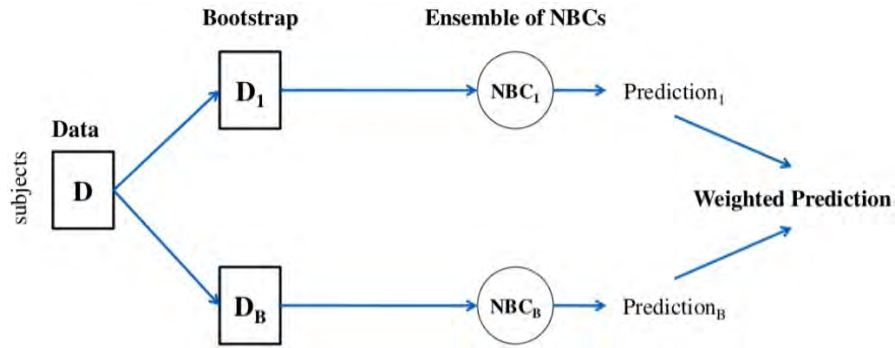


FIGURE C.1: Bagging naïve Bayes process.

C.1.3 A hybrid approach: Bagging leafs of naïve Bayesian trees

A decision tree is a decision support tool that uses a schematic tree-shaped diagram graph which model decisions and their possible consequences. Each branch of the decision tree represents a possible decision or occurrence. The tree structure shows how one choice leads to the next, and the use of branches indicates that each option is mutually exclusive. Decision trees are learned in a top-down fashion, with an algorithm known as Top-Down Induction of Decision Trees (TDIDT), recursive partitioning, or divide and conquer learning. The algorithm selects the best attribute for the root of the tree, splits the set of examples into disjoint sets, and then adds corresponding nodes and branches to the tree [233].

A naïve Bayesian tree applies different NBCs to different regions of the input space inducing a hybrid decision tree classifier: the decision tree nodes contain univariate splits as regular decision trees, but their leafs contain NBCs [234]. In this way, the main part of this approach is by classical recursive partitioning schemes as in usual decision trees (such as the above-mentioned TDIDT). However, the corresponding leaf nodes created are NBCs instead of nodes predicting a single class.

Besides the NBT approach, this paper also proposes a new strategy on leaf nodes. It consists on bootstrapping the elements at the leaf nodes, followed by a bagging process based on NBCs. This approach tries to take advantage of the tree structure of the data, which obtains, thus, a suitable starting point to apply a re-sampling method. As a consequence, it represents a first step where the process diminishes variability and prevents bias in the creation of the bootstrap process; this helps optimize the bagging

classifier. Due to the nature of the proposed ensemble learning method, the overall process still remains simple while computationally efficient.

C.1.4 Summary of the results and conclusions

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 approached this problem through the naïve Bayes algorithm showing that the intricacy of the problem under study is a big handicap to get the final aim.

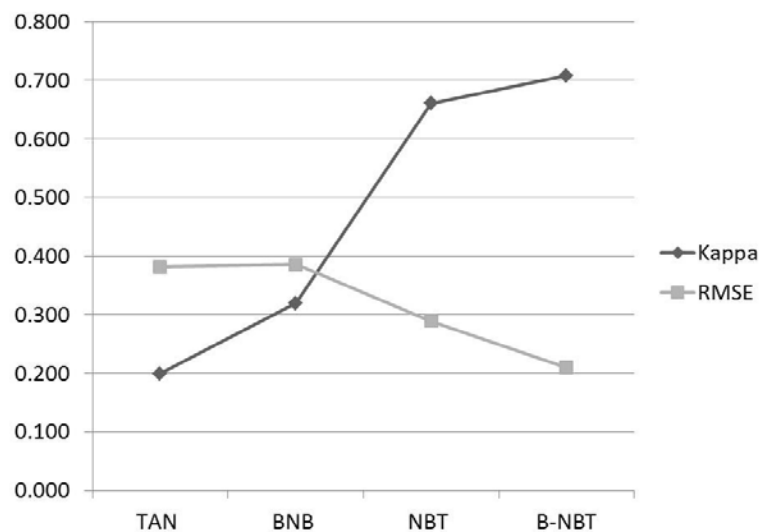


FIGURE C.2: Kappa statistic value and RMSE for TAN, BNB, NBT and B-NBT.

It has been demonstrated that ensemble techniques are more useful in this complex case, obtaining better results than the simpler methods because the iterations increased the robustness of the process. However, this has not been enough to get a good model. Hybrid ensemble techniques have been necessary to achieve good results (Figure C.2). The cumulative experience on the performance of multiple applications of different learning systems is the suitable way to achieve our aim, thus, reducing the uncertainty and improving the overall prediction accuracy of the model. Furthermore, the approach proposed in this paper, has demonstrated to be a suitable way to achieve a good model in this case. It has shown to be able to exploit the advantages of the different techniques used. Avoiding bias and decreasing the uncertainty with the classification trees,

improving the efficiency through the naïve Bayes classifier and, finally, gaining accuracy by applying bagging.

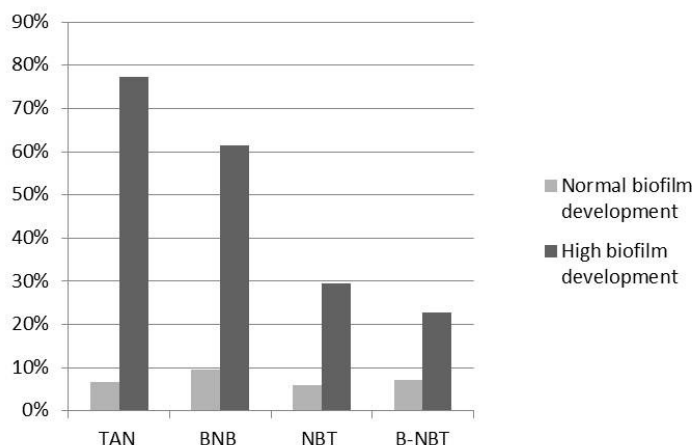


FIGURE C.3: Error percentages of the confusion matrix.

The improvement of the output is not shown only in the goodness indexes, but also in the results (Figure C.3). Although, in the cases with normal biofilm development, the error percentage of the B-NBT method is a little bit bigger than the obtained with the NBT, the error rate of the cases with high biofilm development, in which we are interested to due to their implication in numerous DWDS problems, is greatly reduced. As a consequence, we claim that the methodology that we have developed is able to deal suitably with the problem tackled in this paper, and outperforms previous approaches found in the literature.

Appendix D

Modelling the Biofilm Development Process within pipes with Multiagent systems

D.1 Modelling the Biofilm Development Process

The model has been developed in the NetLogo software [223]. One of the purposes of this study was to build a model as generic as possible, with no assumptions about the nature of the biofilm or the type of the microorganisms that compose it, that can develop the biofilm formation stages in DWDSs. Due to computational constraints, and the selected simulation scale, the high concentration of microorganisms occurring in biofilm does not allow us to model each individual bacterium. The agents were defined as clusters of colonies of bacteria due to the high bacterial densities reached in these systems. Each agent represents a core, a bacteria colony, and is capable of binding to the pipe wall, excrete glycocalix, reproduce (create new agents), die and detach from the biofilm. This last action will depend on the flow velocity and the position of the agent in the matrix model. The environment model has been described as the inside of a pipe.

In the instant that a clean pipe is filled with water, biofilm begins to form. Any surface immersed in water instantly attracts, both, organic and inorganic molecules from

the water that surrounds it, forming a preparation film. The formation of this initial film is especially important in environments that are low in nutrients, such as drinking water, where the accumulation of organic molecules on the surface creates a localized area relatively rich in nutrients. Some of the planktonic bacteria will approach the pipe wall and become entrained within it [235]. This initial attachment is based on the electrostatic attraction and physical forces, not on any chemical attachments. Some of the adsorbed cells begin to make preparations for a lengthy stay by forming structures that may permanently attach the cell to the surface [236]. Biofilm bacteria excrete extracellular polymeric substances, or sticky polymers (glycocalix), which hold biofilm together and cement it to the pipe wall. As nutrients accumulate, the pioneer cells proceed to reproduce [235]. The glycocalyx net, apart from trapping nutrient molecules, snares other types of microbial cells through physical restraint and electrostatic interaction (second colonizers) [236].

In summary, the steps to develop a mature biofilm are: surface conditioning, adhesion of pioneer bacteria, glycocalix formation and incorporation of secondary colonizers (Figure 2.1). All these steps have been incorporated in our model. True biofilm steady state is never achieved, since selection is continually occurring, and slight changes in environment conditions may favour the growth of different organisms [124]. Shear forces or residual disinfectant are some of these factors that cause this biofilm instability. Shear forces exerted by flowing water impact on the mechanical stability of biofilm causing the continuous erosion of the surface layers and population succession. Indeed hydraulic shear can limit biofilm thickness [7]. Increasing the shear force decreases the thickness of the boundary layer. Agents interact with each other to find the balance between density and spatial growth (Figure D.1).

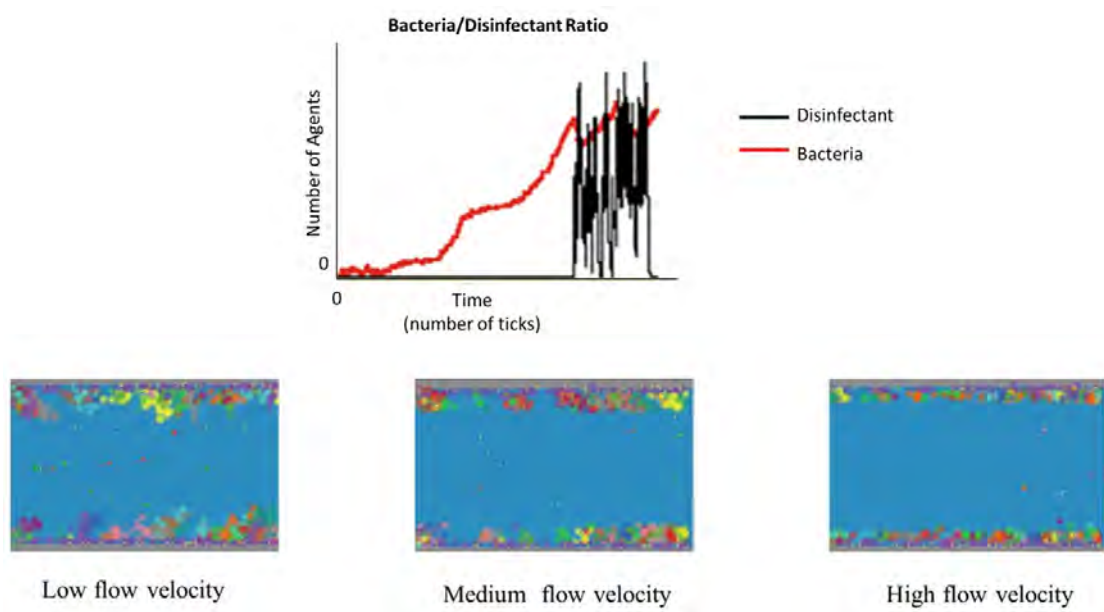


FIGURE D.1: Modelling biofilm development within a pipe.

Appendix E

Presentation of the web page sections

E.1 Presentation of the web page sections

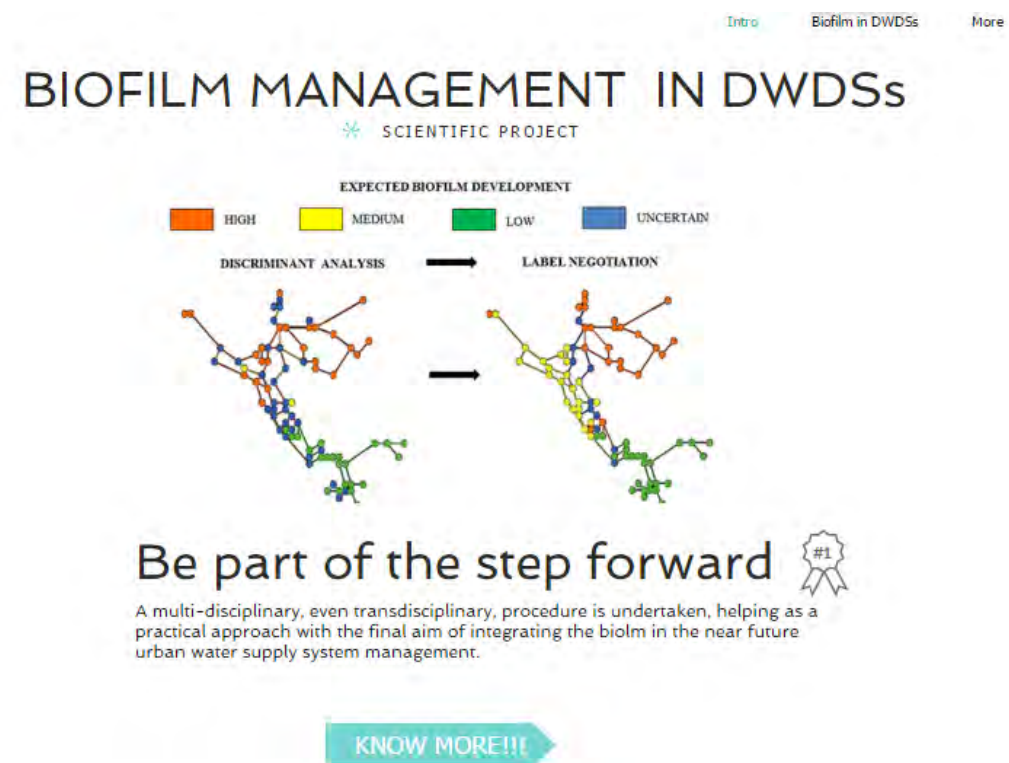


FIGURE E.1: The appearance of the web page.


Intro Biofilm in DWDSs More

Biofilm for All

After observing the good results that can be obtained when pre-processing and applying the data science techniques to the study of biofilm development in DWDSs.

We launch this platform in order to enhance the networking and get in contact with other people interested in this field.

We believe that collaboration with research groups or/and water utilities would be the perfect way to keep enlarging the biofilm in DWDSs database and **SHIFT TO THE DEVELOPMENT OF A MORE PRACTICAL AND REAL LIFE IMPLEMENTABLE APPROACHES.**



[Contact us!](#)

FIGURE E.2: The “Biofilm for All” project presentation in the web page.

Intro Biofilm in DWDSs More

CONTACT US:

Grupo Fluing - IMM

Camino de Vera, S/N
(Edificio 5C-Bajo)
46022 Valencia - Spain

T +34 682 17 93 61
F +34 628 02 88 04

	Name
	Email
	Subject
	Message


 [Send](#)

FIGURE E.3: The “Contact us” section in the web page.

Intro Biofilm in DWDSS More

Outstanding Publications

Publications in scientific Journals and Books

E. Ramos-Martinez; M. Herrera; J. Izquierdo; R. Perez-Garca. Multi-Agent Approach to Biofilm Development in Water Supply Systems. Athens: ATINER'S Conference Paper Series, WAT2015, No. 1693, 2015.

M. Herrera; E. Ramos-Martinez; J. Izquierdo; R. Perez-Garca. Graph constrained label propagation on water supply networks. *AI Communications* 28,47-53, 2015.

E. Ramos-Martinez; M. Herrera; J. Gutierrez-Perez; J. Izquierdo; R. Perez-Garca. Rehabilitation Actions in Water Supply Systems: Effects on Biofilm Susceptibility. *Procedia Engineering* 89, 225-231, 2014.

E. Ramos-Martinez; M. Herrera; J. Izquierdo; R. Perez-Garca. Pre-processing and visualization of biofilm development in drinking water distribution systems. *Water Utility Journal* 7, 3-11, 2014.

E. Ramos-Martinez; J. A. Gutierrez-Perez; M. Herrera; J. Izquierdo; R. Perez-Garca. Pipe database analysis transduction to assess the spatial vulnerability to biofilm development in drinking water distribution systems. Ch. 7 in: J.C. Cortes, L. Jodar Sanchez and R.J. Villanueva (eds.), *Mathematical Modeling in Engineering & Social Sciences*, Nova Science Publishers, Hauppauge, NY, 2014, pp. 71-80.

Works presented in International Conferences

E. Ramos-Martinez; M. Herrera; J. Izquierdo; R. Perez-Garca. Multi-Agent Approach to Biofilm Development in Water Supply Systems. 3rd Annual International Conference on Water. Athens, Greece. 13-16/07/2015

E. Ramos-Martinez; M. Herrera; J. Izquierdo; R. Perez-Garca. Rehabilitation action in water supply systems effects on biofilm development. Water Distribution System Analysis (WDSA). Bari, Italia. 14-17/07/2014

E. Ramos-Martinez; J.A. Gutierrez Perez, M. Herrera; J. Izquierdo; R. Perez-Garca. Biofilm susceptibility in a drinking water distribution system regarding 24 hours curve demand. 7th International Congress on Environmental Modelling and Software (iEMSs). San Diego, California. 15-19/06/2014

E. Ramos-Martinez; J.A. Gutierrez Perez, M. Herrera; J. Izquierdo; R. Perez-Garca. Metadata on biofilm development in drinking water distribution systems. XVI International Congress of the Catalan association for Artificial Intelligence. Vic, Spain. 23-25/10/2013

M. Herrera; E. Ramos-Martinez; J. Izquierdo; R. Perez-Garca. Graph constrained label propagation on water supply networks. XVI

FIGURE E.4: The “Already done” section of the web page.

Bibliography

- [1] P. Brennenstuhl, A. and Doherty, P. King, and T. Dunstall. *Electrochemical interpretation of the role of microorganisms in corrosion*. Houghton DR, Smith RN, Eggins HOW (eds) Biodeterioration. Elsevier Applied Science, London, England, 1988.
- [2] G. H. Koch, M. P.H. Brongers, N. G. Thompson, Y. P. Virmani, and J.H. Payer. *Corrosion costs and preventive strategies in the United States*, chapter Publication NO. FHWA-RD-01-156. 2002.
- [3] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Multi-agent approach to biofilm development in water supply systems. In *Third Annual International Forum on Water*. Gregory T. Papanikos - Athens Institute for Education and Research, 2015.
- [4] C.M. Manuel, O.C. Nunes, and L.F. Melo. Dynamics of drinking water biofilm in flow/non-flow conditions. *Water Research*, 41:551–562, 2007.
- [5] M. Batté, B. Koudjonou, P. Laurent, L. Mathieu, J. Coallier, and M. Prévost. Biolm responses to ageing and to a high phosphate load in a bench-scale drinking water system. *Water Research*, 37:1351–1361, 2003.
- [6] S. Kalmbach, W. Manz, and U. Szewzyk. Dynamics of biofilm formation in drinking water: phylogenetic affiliation and metabolic potential of single cells assessed by formazan reduction and in situ hybridization. *FEMS Microbiology Ecology*, 22: 265–279, 1997.

- [7] The Cooperative Research Centre (CRC) for Water Quality and Treatment. Understanding the impact on water quality and water treatment processes : Management implications from the research programs of the cooperative research centre for water quality and treatment. *Australia*, 2005.
- [8] P. Deines, R. Sekar, S.P. Husband, J. B. Boxall, A. M. Osborn, and C. A. Biggs. A new coupon design for simultaneous analysis of in situ microbial biofilm formation and community structure in drinking water distribution systems. *Applied Microbiology and Biotechnology*, 87:749756, 2010.
- [9] W. Furnass, I. Douterelo, R. Collins, S. Mounce, and J. Boxall. Controlled, realistic-scale, experimental study of how the quantity and erodibility of discoloration material varies with shear strength. *Procedia Engineering*, 89:135142, 2014.
- [10] I. Douterelo, R.L. Sharpe, and J.B. Boxall. Influence of hydraulic regimes on bacterial community structure and composition in an experimental drinking water distribution system. *Water Research*, 47(2):503516, 2013.
- [11] I.B. Gomes, Sim oes M., and Sim oes L.C. An overview on the reactors to study drinking water biofilms. *water research*, 62:63–87, 2014.
- [12] Apha method 9215: Standard methods for the examination of water and wastewater. Technical report, American Public Health Association and American Water Works Association and Water Environment Association, 1992.
- [13] L. Gang. *Microbiological water quality in drinking water distribution systems: Integral study of bulk water, suspended solids, loose deposits, and pipe wall biofilm*. PhD thesis, Delft Univerisity of Technology, 2013.
- [14] Biyela P. Thabisile. *Water quality Decay and Pathogen Survival in Drinking Water Distribution Systems - Partial Fullfilment*. PhD thesis, Arizona State University, 2010.
- [15] C.R. Kokare, S. Chakraborty, A. N. Khopade, and K. R. Mahadik. Biofilm: Importance and applications. *Indian Journal of Biotechnology*, 8:159–168, 2009.

- [16] R. M. Donlan and J. W. Costerton. Biofilms: Survival mechanisms of clinically relevant microorganisms. *Water Resource Planning and Management*, 15(2):167193, 2002.
- [17] M.W. Cowle, A.O. Babatunde, W.B. Rauen, B.N. Bockelmann-Evans, and A.F. Barton. Biofilm development in water distribution and drainage systems: dynamics and implications for hydraulic efficiency. *Environmental Technology Reviews*, 3(3):3147, 2014.
- [18] I. Douterelo, J. B. Boxall, P. Deines, K.E. Sekar, R. Fish, and C. A. Biggs. Methodological approaches for studying the microbial ecology of drinking water distribution systems. *Water Research*, 65:134 – 156, 2014.
- [19] J. Yu, D. Kin, and T. Lee. Microbial diversity in biofilms on water distribution pipes of different materials. *Water Science and Technology*, 61(1):163–171, 2010.
- [20] H. Wu, J. Zhang, Z. Mi, S. Xie, C. Chen, and X. Zhang. Biofilm bacterial communities in urban drinking water distribution systems transporting waters with different purification strategies. *Applied Microbiology and Biotechnology*, 99(4): 1947–1955, 2015.
- [21] Y. Zhu, H. Wang, X. Li, C. Hu, M. Yang, and J. Qu. Characterization of biofilm and corrosion of cast iron pipes in drinking water distribution system with uv/cl2 disinfection. *Water Research*, 60(1):174181, 2014.
- [22] J.L.A. Shaw, P. Monis, R. Fabrisb, L. Ho, K. Braun, M. Drikas, and A. Cooper. Assessing the impact of water treatment on bacterial biofilms in drinking water distribution systems using high-throughput dna sequencing. *Chemosphere*, 117: 503516, 2014.
- [23] J. Berry, L. Fleisher, W. Hart, C. A. Phillips, and J. Watson. Sensor placement in municipal water networks. *Water Resource Planning and Management*, 131(3): 237–243, 2005.
- [24] J. Aubrecht, P. Mikosovski, and Z. Kouba. *Metadata Driven Data Pre-processing for Data Mining*. J. Pokorny, V. Snasel (Eds.) - Technical University of Ostrava, 2003.

- [25] E. Karunakaran, J. Mukherjee, B.i Ramalingam, and C. A. Biggs. Biofilmology: a multidisciplinary review of the study of microbial biofilms. *Applied Microbiology and Biotechnology*, 90(6):1869–1881, 2011.
- [26] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Multi-agent approach to biofilm development in water supply systems. *Athens: Atiner Conference Series No: WAT2015-1693*, 2015.
- [27] M. Herrera, E. Ramos-Martínez, J. Izquierdo, and R. Pérez-García. Graph constrained label propagation on water supply networks. *AI Communications*, 28: 47–53, 2015.
- [28] E. Ramos-Martínez, J. Gutiérrez-Pérez, M. Herrera, J. Izquierdo, and R. Pérez-García. Rehabilitation actions in water supply systems: Effects on biofilm susceptibility. In *16th Conference of Water Distribution Analysis, WDSA 2014*, 2014.
- [29] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Pre-processing and visualization of biofilm development in drinking water distribution systems. *Water Utility Journal*, 7:3–11, 2014.
- [30] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. *Mathematical Modeling in Social Sciences and Engineering*, chapter 7. On Kernel spectral clustering for identifying areas of biofilm development in water distribution systems. NOVA Science Publisher, 2014.
- [31] M. Herrera, E. Ramos-Martínez, J. Gutiérrez-Pérez, J. Izquierdo, and R. Pérez-García. *Mathematical Modeling in Social Sciences and Engineering*, chapter 8. Pipe Database Analysis Transduction to Assess the Spatial Vulnerability to Biofilm Development in Drinking Water Distribution Systems. NOVA Science Publisher, 2014.
- [32] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Ensemble of naive bayesian approaches for the study of biofilm development in drinking water distribution systems. *International Journal of Computer Mathematics*, 99 (1), 2014.

- [33] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Drinking water distribution systems characteristics on biofilm development: a kernel based approach. *ATINER'S Conference Paper Series*, 0773, 2013.
- [34] J. A. Gutiérrez-Pérez, M. Herrera, R. Pérez-García, and E. Ramos-Martínez. Application of graph-spectral methods in the vulnerability assessment of water supply networks. *Mathematical Computing and Modelling*, 57(7-9):18531859, 2013.
- [35] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Evaluación de la distribución espacial del biofilm en los sistemas de abastecimiento de agua. In *XXVI Congreso Latinoamericano de Hidráulica. Santiago de Chile*, 2014.
- [36] E. Ramos-Martínez, J.A. Gutiérrez Pérez, M. Herrera, J. Izquierdo, and R. Pérez-García. Biofilm susceptibility in a drinking water distribution system regarding 24 hours curve demand. In *7th International Congress on Environmental Modelling and Software (iEMSs)*, 2014.
- [37] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Métodos kernel para el estudio del desarrollo de biofilm en los sistemas de distribución de agua potable. In *Seminario Euro Latinoamericanos de Sistemas de Ingeniería*, 2013.
- [38] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Estudio de la influencia relativa en el desarrollo de biofilm de las características físicas e hidráulicas de los sistemas de distribución de agua potable. In *XII Simposio Iberoamericano sobre planificación de sistemas de abastecimiento y drenaje*, 2013.
- [39] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Biofilm: influencia del diseño y operación de los sistemas de abastecimiento de agua. In *III Jornadas de Ingeniería del Agua*, 2013.
- [40] E. Ramos-Martínez, J.A. Gutiérrez Pérez, M. Herrera, J. Izquierdo, and R. Pérez-García. Metadata on biofilm development in drinking water distribution systems. In *XVI International Congress of the Catalan association for Artificial Intelligence*, 2013.

- [41] M. Herrera, E. Ramos-Martínez, J. Izquierdo, and R. Pérez-García. Graph constrained label propagation on water supply networks. In *XVI International Congress of the Catalan association for Artificial Intelligence*, 2013.
- [42] J. A. Gutiérrez-Pérez, E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Graph spectral method to assess biofilm development in drinking water distribution systems. In *Mathematical Modelling in Engineering & Human Behaviour*, 2013.
- [43] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Drinking water distribution systems characteristics on biofilm development a kernel based approach. In *Annual International Forum on Water*, 2013.
- [44] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Preprocessing and visualization on biofilm development in drinking water distribution systems. In *8th International Conference of European Water Resources Association*, 2013.
- [45] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Biofilms en los sistemas de distribución de agua potable. aproximación basada en sistemas multi-agente. In *XI Seminario Euro Latinoamericano de sistemas de ingeniería*, 2012.
- [46] J. A. Gutiérrez-Pérez, M. Herrera, J. Izquierdo, R. Pérez-García, and E. Ramos-Martínez. Enfoque multi-agente para la identificación de elementos vulnerables en una red de abastecimiento. In *XI Seminario Euro Latinoamericano de sistemas de ingeniería*, 2012.
- [47] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Ensemble of multiple data mining approaches to biofilm development in drinking water distribution systems. In *Mathematical Modelling in Engineering & Human Behaviour*, 2012.
- [48] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Estudio del desarrollo de biofilm en tuberías mediante redes bayesianas con variables mixtas. In *XXV Congreso Latinoamericano de Hidráulica*, 2012.

- [49] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Evaluación de las características físicas e hidráulicas de los sistemas de distribución de agua que determinan el desarrollo de biofilms. In *XI Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje*, 2012.
- [50] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Modelling biofilm formation and evolution in drinking water distribution systems using a multi-agent approach. In *2nd Meeting of Young Researchers Modelling Biological Processes*, 2012.
- [51] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Assessing variations in biofilm development in a drinking water distribution system by an object oriented bayesian network approach. In *6th International Congress on Environmental Modelling and Software (iEMSs)*, 2012.
- [52] E. Ramos-Martínez, M. Herrera, J. Izquierdo, and R. Pérez-García. Evaluación del desarrollo de biofilms en los sistemas de abastecimiento de agua mediante redes bayesianas. In *X Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje*, 2011.
- [53] J. A. Gutiérrez-Pérez, M. Herrera, R. Pérez-García, and E. Ramos-Martínez. La vulnerabilidad de los sistemas de abastecimiento de agua. In *X Seminario Iberoamericano sobre Sistemas de Abastecimiento y Drenaje*, 2011.
- [54] United Nations Environment Programme UNEP. Improving the quantity, quality and use of africa's water. *Environment for development. Atlas of Our Changing Environment*, 2014.
- [55] Un-water global annual assessment of sanitation and drinking-water. Technical report, World Health Organization and United Nations, 2010.
- [56] J.W. Costerton, K.J. Cheng, G.G. Gessey, T.I. Ladd, J.C. Nickel, M. Dasgupta, and T.J Marrie. Bacterial biofilms in nature and disease. *Annual Review Microbiology*, 41:435–464, 1987.
- [57] A. Farkas, D. Citaras, and B. Bocos. *Biofilms Impact on Drinking Water Quality, Ecological Water Quality. Water Treatment and Reuse*. Dr. Voudouris (Ed.), 2012.

- [58] M. E. Davey and G. A. OToole. Microbial biofilms: from ecology to molecular genetics. *Microbiology and molecular biology reviews*, 64(4):847867, 2000.
- [59] D. López, H. Vlamakis, and R. Kolter. Biofilms. *Cold Spring Harb Perspect Biol*, pages 2421–2430, 2010.
- [60] A. W. Decho. Microbial biofilms in intertidal systems: an overview. *Continental Shelf Research*, 20(10-11):12571273, 2000.
- [61] L.C. Simes. *Biofilms in drinking water*. Simes M, Mergulho F, editors. Biofilms in bioengineering. New York (NY): Nova Science, 2013.
- [62] S. L. Yu, X. J. Yu, W. X. Shi, D. Wang, and X. X Qiu. Bacterial species and variation in the biofilm of water distribution system in harbin city. *Journal of Water Supply Research and Technology-Aqua*, 56:445–451, 2007.
- [63] C. W. Keevil and J. T. Walker. Nomarski dic microscopy and image analysis of biofilms. *Binary Computing Microbiology*, 4:93–95, 1992.
- [64] I. W. Sutherland. Biofilm exopolysaccharides: a strong and sticky framework. *Microbiology*, 147:3–9, 2001.
- [65] M. F. Gelves. Deterioro de la calidad del agua por el posible desprendimiento de las biopelículas en las redes de distribución de agua potable. *Universidad de los Andes*, 2005.
- [66] P. H. Dreeszen. The key to understanding and controlling bacterial growth in automated drinking water systems.(2nd ed). Technical report, Edstrom industries Inc, 2003.
- [67] C. Chagnot, M. A. Zorgani, T. Astruc, and M. Desvaux. Proteinaceous determinants of surface colonization in bacteria: bacterial adhesion and biofilm formation from a protein secretion perspective. *Frontiers in Microbiology*, 4(303), 2013.
- [68] N. J. Ashbolt. Microbial contamination of drinking water and human health from community water systems. *Current Environmental Health Reports*, 2(1):31–47, 2015.

- [69] A. von Graevenitz. The role of opportunistic bacteria in human disease. *Annual Review of Microbiology*, 31:447–471, 1977.
- [70] L. Nieto and J.G. Saldarriaga. Eventos de coloración del agua potable como consecuencia del desprendimiento de biopelículas: el caso de bogotá d.c. *Universidad de los Andes*, 2009.
- [71] J. Wingender and F. Hans-Curt. Biofilms in drinking water and their role as reservoir for pathogens. *Hygiene and Environmental Health*, 214:417–423, 2011.
- [72] H. Sun, B. Shi, Y. Bai, and D. Wang. Bacterial community of biofilms developed under different water supply conditions in a distribution system. *Science of The Total Environment*, 472:99107, 2014.
- [73] R.M. Batt, H.C. Rutgers, and A. A. Sancak. Enteric bacteria: friend or foe? *The Journal of Small Animals Practice*, 37(6):261–267, 1996.
- [74] W.R. Jarvis. Opportunistic pathogenic microorganisms in biofilms. *Center for Disease Control, Washington D.C.*, 1990.
- [75] P. Payment and W. Robertson. The microbiology of piped distribution systems and public health. *Safe piped water: Managing Microbial Water Quality in Piped Distribution Systems. Ainsworth, R. (Ed.), IWA Publishing*, pages 31–47, 2004.
- [76] A.B. Goncalves, R.R.M. Paterson, and N Lima. Survey and significance of filamentous fungi from tap water. *International Journal of Hygiene and Environmental Health*, 29(3):257264, 2006.
- [77] M. Steirnet, K. Birkness, E. White, B. Fields, and F. Quinn. Mycobacterium avium bacilli grow saprozoically in coculture with acanthamoeba polyphaga and survive within cyst walls. *Applied and Environmental Microbiology*, 63(6):2256–2261, 1998.
- [78] EPA United States Environmental Protection Agency. Legionella: Drinking water health advisory. *United States Environmental Protection Agency*, EPA-822-B-01-005, 2001.

- [79] U. Szewzik, R. Szewzyk, W. Manz, and K.H. Schleifer. Microbiological safety of drinking water. *Annual Review of Microbiology*, 54:81–127, 2000.
- [80] H. F. Ridway and B.H. Olso. Chlorine resistance patterns of bacteria from two drinking water distribution systems. *Applied and Environmental Microbiology*, 41: 274–287, 1982.
- [81] F. Codony, A. M. Miranda, and J. Mas. Persistence and proliferation of some unicellular algae in drinking water systems as result of their heterotrophic metabolism. *Water SA*, 29(1), 2003.
- [82] UK. Environment Agency and Standing Committee of Analysts. The assessment of taste, odour and related aesthetic problems in drinking waters. *Methods for the Examination of Waters and Associated Materials*, 1998.
- [83] R. Imran, M. and Sadiq and Y. Kleiner. Identifying research needs related to impacts of water quality on the integrity of distribution infrastructure. Technical report, National research council Canada, 2006.
- [84] S. Chowdhury. Environmental monitoring and assessment. *Environmental Monitoring and Assessment*, 2012.
- [85] C. Evins. *Small Animals in drinking Water Distribution Systems. Safe Piped Water: Managing Microbial Water Quality in Piped Distribution Systems*. World Health Organization, IWA Publishing, London, United Kingdom, 2006.
- [86] H. E. Smith-Somerville, C. V. B. Huryn, Walker, and A. L. Winters. Survival of legionella pneumophila in the cold-water ciliate tetrahymena vorax. *Applied Environmental Microbiology*, 57:2742–2749, 1991.
- [87] A. O. Al-Jasser. Pipe service age effect on chlorine decay in drinking-water transmission and distribution systems. *CLEAN Soil, Air, Water*, 39(9):827832, 2011.
- [88] Y. Lévi L. Kiéné, W. Lu. Relative importance of the phenomena responsible for chlorine decay in drinking water distribution systems. *Water Science and Technology*, 38(6):219227, 1998.

- [89] D. de Beer, R. Srinivansa, and P.S. Stewart. Direct measurement of chlorine penetration into biofilms during disinfection. *Applied Environmental Microbiology*, 60(12):4339–4344, 1994.
- [90] R. A. Adhikari, A. Sathasivan, and K. C. Bal Krishna. Effect of biofilms grown at various chloramine residuals on chloramine decay. *Water Science & Technology: Water Supply*, 12(4):463–469, 2012.
- [91] Y. Seo. Monitoring the role of biofilm biopolymers against disinfectants in water distribution systems. *Water Resources Center Annual Technical Report FY 2009*, 2009.
- [92] Z. Xue, V. R. Sendamangalam, C. L. Gruden, and Y. Seo. Multiple roles of extracellular polymeric substances on resistance of biofilm and detached clusters. *Environmental Science & Technology*, 46(24):1321213219, 2012.
- [93] W. Lu, L. Kin, and Y. Lvi. Chlorine demand of biofilms in wter distribution systems. *Water Research*, 33(3):827–835, 1998.
- [94] I. Beech, A. Bergel, A. Mollica, H. C. Flemming, V. Scotto, and W. Sand. Simple methods for the investigation of the role of biofilms in corrosion. *Biofilms Publication*, 2000.
- [95] I. Beech and C. C. Gaylarde. Recent advances in the study of biocorrosion - an overview. *Revista de Microbiologia*, 30(3), 1999.
- [96] F.A. Lopes, P. Morin, R. Oliveira, and L.F. Melo. Impact of biofilms in simulated drinking water and urban heat supply systems. *Environmental Engineering*, 1(3), 2009.
- [97] B. J. Little and Jason S. L. Microbiologically influenced corrosion. *John Wiley & Sons, Inc., Hoboken, New Jersey*, 2007.
- [98] P. Brennenstuhl, A. and Doherty, P. King, and T. Dunstall. *Microbially Influenced Corrosion and Biodeterioration*. Edited by N. Dowling, M. Mittleman and J. Danko, Knoxville, Tennessee, 1991.

- [99] III J.E. Cromwell, H. Reynolds, J. Pearson, and M. Grant. Costs of infrastructure failure. Technical report, Denver, CO: AwwaRF, 2002.
- [100] M.A. Shockling, J.J. Allen, and A.J. Smits. Roughness effects in turbulent pipe flow. *Journal of Mathematical Fluid Mechanics*, 564:267285, 2006.
- [101] A.F. Barton, M.R. Wallis, J.E. Sargison, A. Buia, and G.J. Walker. Hydraulic roughness of biofouled pipes, biofilm character, and measured improvements from cleaning. *Journal of Hydraulic Engineering*, 134(6):852857, 2008.
- [102] A.F. Barton. *Friction, roughness and boundary layer characteristics of freshwater biofilms in hydraulic conduits*. PhD thesis, Hobart, Tasmania: School of Engineering, University of Tasmania, 2006.
- [103] A. Townsend. *The structure of turbulent shear flow*. Cambridge University Press, Cambridge, 1976.
- [104] M.P. Schultz and G.W. Swain. The effect of biofilms on turbulent boundary layers. *Journal of Fluids Engineering*, 121(1):4451, 1999.
- [105] W.G. Characklis and K.C. Marshall. *Biofilms*. Ed. John Wiley & Sons, Inc., Nueva York, 1990.
- [106] M. Simoes, L.C. Simoes, and M.J. Vieira. A review of current and emergent biofilm control strategies. *LWT - Food Science and Technology*, 43(4):573583, 2010.
- [107] M. W. LeChevallier. The case for maintaining a disinfectant residual. *Journal - American Water Works Association*, 91:86–94, 1999.
- [108] D. van der Kooij. Biological stability: a multidimensional quality aspect of treated water. *Water, Air, & Soil Pollution*, 123:2534, 2000.
- [109] F. Ramírez. Desinfección del agua con cloro y cloraminas. *Técnica Industrial*, 260, 2005.
- [110] M.W. LeChevallier. Biocides and the current status of biofouling control in water systems. *Biofouling and biocorrosion in industrial water systems*, pages 113–132, 1991.

- [111] F. Hammes, C. Berger, O. Koster, and T. Egli. Assessing biological stability of drinking water without disinfectant residuals in a fullscale water supply system. *Journal Of Water Supply Research And Technology - AQUA*, 59(1):31–40, 2010.
- [112] T. Griebe and H.-C. Flemming. Biocide-free antifouling strategy to protect ro biofouling. *Desalination*, 118:153156, 1998.
- [113] H.C. Flemming and H. Ridgway. Biofouling on membranes. a microbiological approach. *Desalination*, 70:95–119, 1998.
- [114] H.C. Flemming and H. Ridgway. Biofilm control: Conventional and alternative approaches. *Springer Series on Biofilms*, 2008.
- [115] A. Carvajal, L. F. and Gomez and S. Ochoa. Simulación de un lavado hidráulico en tuberías para el control del crecimiento de biopelícula. *Dyna rev.fac.nac.minas*, 74(152):153156, 2007.
- [116] Y. J. Hasit, A. J. DeNada, and R. S. Raucher. *Cost and Benefit Analysis of Flushing*. American Water Works Association (AWWA) Research Foundation, 2004.
- [117] C. Mains. Biofilm control in distribution systems. *National Environmental Services Center at West Virginia University*, 8(2), 2008.
- [118] E.R. Holm, B.T. Nedved, N. Phillips, K.L. Deangelis, M.G. Hadfield, and C.M. Smith. Temporal and spatial variation in the fouling of silicone coatings in pearl harbour, hawaii. *Biofouling*, 15:95–107, 2000.
- [119] J.S. Louie, I. Pinnau, I. Ciobanu, K.P. Ishida, A. Ng, and M. Reinhard. Effects of polyetherpolyamide block copolymer coating on performance and fouling of reverse osmosis membranes. *Journal of Membrane Science*, 280:762–770, 2006.
- [120] T Matsunaga, T. Nakayama, H. Wake, M. Takahashi, M. Okochi, and N. Nakamura. Prevention of marine biofouling using a conductive paint electrode. *Biotechnology and Bioengineering*, 59:374–378, 1998.

- [121] K.E. Fish, R. Collins, N.H. Green, R. L. Sharpe, I. Douterelo, A. M. Osborn, and J. B. Boxall. Characterisation of the physical composition and microbial community structure of biofilms within a model full-scale drinking water distribution system. *PLOS ONE*, 2015.
- [122] K. Pedersen. Method for studying microbial biofilms in flowing-water systems. *Applied and Environmental Microbiology*, 43(1):6–13, 1982.
- [123] N.B. Hallam, J.R. West, C.F. Forster, and J. Simms. The potential for biofilm growth in water distribution systems. *Water Resources*, 35(17):4063–4071, 2001.
- [124] R. Boe-Hansen, H.J. Albrechtsen, E. Arvin, and C. Jrgensen. Bulk water phase and biofilm growth in drinking water at low nutrient conditions. *Water Research*, 36(18):44774486, 2002.
- [125] J.E. Hobbie, R.J. Daley, and S. Jasper. Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Applied Environmental Microbiology*, 33:1225–1228, 1997.
- [126] H.C. Schaule, G. and Flemming and H.F. Ridgway. Use of 5-cyano-2,3-ditolyl tetrazolium chloride for quantifying planktonic and sessile respiring bacteria in drinking water. *Applied environmental Microbiology*, 59:3850–3857, 1993.
- [127] H.C. Schaule, G. and Flemming. Quantification of respiratory active bacteria in water and biofilms with a fluorescent redox dye. *Microbiologically Influenced Corrosion Materials*, pages 159–165, 1996.
- [128] G. Liu, J. Q. J. C. Verberk, and J. C. Van Dijk. Bacteriology of drinking water distribution systems: an integral and multidimensional review. *Applied Microbiology and Biotechnology*, 97:92659276, 2013.
- [129] L. Boulos, M. Prevost, B. Barbeau, J. Coallier, and R. Desjardins. Ulive/dead baclightTM application of a new rapid staining method for direct enumeration of viable and total bacteria in drinking water. *Journal of Microbiological Methods*, 37(1):77–86, 1999.

- [130] M. Berney, M. Vital, I. Hlshoff, H.U. Weilenmann, T. Egli, and F. Hammes. Rapid, cultivation-independent assessment of microbial viability in drinking water. *Water Research*, 42(14):40104018, 2008.
- [131] D. Santić, N. Krstulović, and M. Solić. Comparison of flow cytometric and epifluorescent counting methods for marine heterotrophic bacteria. *Acta Adriatica*, 48(2):107114, 2007.
- [132] J. Bartram, J. Cotruvo, M. Exner, C. Fricker, and A. Glasmacher. *Heterotrophic plate counts and drinking-water safety: The significance of HPCs for water quality and the human health*. IWA Publishing on behalf of the World Health Organization, 2003.
- [133] W. Robertson and T. Brooks. *The role of HPC in managing the treatment and distribution of drinking-water*. J. Bartram, J. Cotruvo, M.Exner, C. Fricker, A. Glasmacher. On behalf of WHO by IWA Publishing, 2003.
- [134] E. Theres Gensberger, E. M. Gossel, L. Antonielli, A. Sessitsch, and T. Kostic. Effect of different heterotrophic plate count methods on the estimation of the composition of the culturable microbial community. *PeerJ*, 2015.
- [135] M. J. Allen, S. C. Edberg, and D. J. Reasoner. Heterotrophic plate count bacteria what is their significance in drinking water? paper 93. Technical report, U.S. Environmental Protection Agency Papers, 2004.
- [136] D.J. Reasoner. Heterotrophic plate count methodology in the United States. *International Journal of Food Microbiology*, 92(3):307, 2004.
- [137] D.J. Reasoner, J.C. Blannon, and E.E. Geldreich. Rapid seven-hour fecal coliform test. *Applied and Environmental Microbiology*, 38(2):229, 1979.
- [138] R. M. Atlas. Handbook of media for environmental microbiology. *CRC Press. Boca Raton. Fla. USA*, 1995.
- [139] R.K. Yin. *Case study research: Design and methods*. Sage Publications, Thousand Oaks, CA, 2003.

- [140] B. E. White, S. J. Gandhi, A. Gorod, V. Ireland, and B. Sauser. On the importance and value of case studies. In *Systems Conference (SysCon), 2013 IEEE International*, pages 114–122, 2013.
- [141] D. Assimacopoulos. Water & sanitation services in greece and the sustainability challenges. Technical report, School of Chemical Engineering National Technical University of Athens, Greece, 2012.
- [142] N.J. Xanthopoulou, M.V. Papagianni, A. Papaioannou, and A. Haralambidou. Volatile organic compounds in the finished water of the water treatment in thesaloniki, greece. *Global NEST Journal*, 7(1):119–127, 2005.
- [143] Noel Munoz. *Estudio exploratorio en la caracterización microbiológica y fisicoquímica la interior de las tuberías de la red de distribución baja de la ciudad de Cali-Colombia*. PhD thesis, Instituto CINARA-Universidad del Valle, 2007.
- [144] R.T. Christensen. *Age Effects on Iron-Based Pipes in Water Distribution Systems*. PhD thesis, Utah State University, 2009.
- [145] US Environmental Protection Agency. Effects of water age on distribution system water quality. Technical report, International Water Association, 2002.
- [146] S. Mounce, I. Douterelo, R. Sharpe, and J. Boxall. A bio-hydroinformatics application of selforganizing map neural networks for assessing microbial and physicochemical water quality in distribution systems. In *10th International Conference on Hydroinformatics HIC 2012, Hamburg, Germany*, 2012.
- [147] P.S. Husband, J.B. Boxall, and A.J. Saul. Laboratory studies investigating the processes leading to discolouration in water distribution networks. *Water Research*, 42(16):4309–4318, 2008.
- [148] A. C. Smith and M. A. Hussey. Gram stain protocols. *American Society for Microbiology - ASM Conference for Undergraduate Educators 2005*, 2005.
- [149] J.B. Patel. 16s rrna gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular Diagnosis*, 6(4):313–21, 2001.

- [150] J. R. Marchesi, T. Sato, A. J. Weightman, T. A. Martin, J. C. Fry, S. J. Hiom, and W. G. Wade. Design and evaluation of useful bacterium-specific per primers that amplify genes coding for bacterial 16s rrna. *Applied and Environmental Microbiology*, 64(2):795799, 1998.
- [151] USEPA United States Environmental Protection Agency. The effectiveness of disinfectant residuals in the distribution system. Technical report, Office of Water (4601M) Office of Ground Water and Drinking Water Total Coliform Rule Issue Paper, 2006.
- [152] D. J. Reasoner and E. E. Geldreich. A new medium for the enumeration and subculture of bacteria from potable water. *Applied and Environmental Microbiology*, 49(1):1–7, 1985.
- [153] P. Stoodley, K. Sauer, D. G. Davies, and J.W. Costerton. Biofilms as complex differentiated communities. *Annual Review of Microbiology*, page 187209, 2002.
- [154] D. Janjaroen, F. Ling, G. Monroy, N. Derlon, E. Mogenroth, S. A. Boppart, W.T. Liu, and T. H. Nguyen. Roles of ionic strength and biofilm roughness on adhesion kinetics of escherichia coli onto groundwater biofilm grown on pvc surfaces. *Water Research*, 47:2531–2542, 2013.
- [155] L.D. Chambers, K.R. Stokes, F.C. Walsh, and R.J.K. Wood. Modern approaches to marine antifouling coatings. *Surface and Coating Technologies*, 201:3642–3652, 2006.
- [156] A. Chaves, M. Simoes, and N. Lima. Adhesion and biofilm formation by drinking water-isolated microorganisms: The role of surface properties and inter-kingdom interactions. In *European Culture Collections as tools in Research and Biotechnology–ECCO XXXIV*, 2015.
- [157] A. C. Martiny, T. M. Jorgensen, H.J. Albrechtsen, E. Arvin, and S. Molin. Long-term succession of structure and diversity of a biofilm formed in a model drinking water distribution system. *Applied and Environmental Microbiology*, 69(11): 68996907, 2003.

- [158] K. Gibert, J. Izquierdo, G. Holmes, I. Athanasiadis, J. Comas, and M. Sànchez-Marrè. On the role of pre and post-processing in environmental data mining. In *Proceedings of iEMSs 2008 International Congress on Environmental Modelling and Software.*, pages 1937–1958, 2008.
- [159] L. Chaves Simoes and M. Simoes. Biofilms in drinking water: problems and solutions. *RSC Advances*, 3(2520), 2013.
- [160] J.A.S. Pérez, E.M.R. Porcel, J.L.C. López, J.M.F. Sevilla, and Y. Chisti. Shear rate in stirred tank and bubble column bioreactors. *Chemical Engineering Journal*, 124(1-3):1–5, 2006.
- [161] ELGA. VEOLIA Water. Pure labwater guide. an essential overview of lab water purification applications, monitoring and standards. Technical report, Water Science and Technology, 2012.
- [162] Health Canada. *Guidance on the Use of Heterotrophic Plate Counts in Canadian Drinking Water Supplies*. Federal-Provincial-Territorial Committee on Drinking Water of the Federal-Provincial-Territorial Committee on Health and the Environment, Ottawa, Ontario, 2012. ISBN 978-1-100-21736-9.
- [163] R. H. Taylor, M. J Allen, and E. E. Geldreich. Standard plate count: A comparison of pour plate and spread plate method. *Research and Technology*, 75(1):35–37, 1983.
- [164] A. A. Van Soestbergen and C. H. Lee. Pour plates or streak plates? *Applied Microbiology*, 18(6):10921093, 1969.
- [165] CDC SWS Project. *Chlorine Residual Testing Fact Sheet*. Centers for disease control and prevention.
- [166] *Standard Methods for the Examination of Water and Wastewater*. American Public Health Association, American Water Works Association, Water Environment Federation, 1999.

- [167] M. W. LeChevallier. *Heterotrophic Plate Counts and Drinking-water Safety*, chapter Conditions favouring coliform and HPC bacterial growth in drinking water and on water contact surfaces, page 77197. IWA Publishing, London, UK., 2003.
- [168] *Guidelines for drinking-water quality 2nd ed. Health criteria and other supporting information*. World Health Organization, Geneva, 1996.
- [169] Standard Methods Committe. *Total Organic Carbon TOC (5310)*. STANDARD METHODS, 1998.
- [170] D. Page and P. Dillon. Measurement of the biodegradable fraction of dissolved organic matter relevant to water reclamation via aquifers. *Water for a Healthy Country report series*, pages 163–171, 2007.
- [171] D. van der Kooij. Assimilable organic carbon as an indicator of bacterial regrowth. *American Water Works Association*, 84(2):57–65, 1992.
- [172] Miettinen I. T., Vartiainen T., and Martikainen P. J. Determination of assimilable organic carbon in humus-rich drinking waters. *Water Research*, 33:22772282, 1999.
- [173] D. van der Kooij, A. Visser, and W.A.M. Hijnen. Determining the concentration of easily assimilable organic carbon in drinking water. *American Water Works Association*, 74(10):540–545, 1982.
- [174] G. A. Gagnon and R. M. Slawson. An efficient biofilm removal method for bacterial cells exposed to drinking water. *Journal of Microbiological Methods*, 34:203214, 1999.
- [175] D. Yiran and J. P. Chao-Ying. Principled missing data methods for researchers. *SpringerPlus*, 2(222):457–467, 2013.
- [176] Z. Tsvetanova. Study of biofilm formation on different pipe materials in a model of drinking water distribution system and its impact on microbiological water quality. *Chemicals as Intentional and Accidental Global Environmental Threats*, pages 463–468, 2006.
- [177] I. Ben-Gal. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.

- [178] K. Singh and S. Upadhyaya. Outlier detection: Applications and techniques. *International Journal of Computer Science Issues*, 1(3):307–323, 2012.
- [179] C. C. Aggarwal. *Supervised Outlier Detection*. Arfken and Weber. Springer, 2012.
- [180] Z. Xue, Y. Shang, and A. Feng. Semi-supervised outlier detection based on fuzzy rough c-means clustering. *Mathematics and Computers in Simulation*, 80(9):1911–1921, 2010.
- [181] K. Leung and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Australasian Computer Science Conference, Newcastle, NSW, Australia*, 2005.
- [182] H.P. Kriegel, M. M. Breunig, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.*, pages 93–104, 2000.
- [183] L. Torgo. *Functions and data for Data Mining with R. Version 0.4.1*, 2015.
- [184] J. Demsar and B Zupan. Orange: from experimental machine learning to interactive data mining. *White Paper (www.ailab.si/orange)*, Faculty of Computer and Information Science. , Slovenia University of Ljubljana, 2004.
- [185] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.
- [186] L. Kong, M. Xia, X. Y. Liu, M. Y. Wu, and X. Liu. Data loss and reconstruction in sensor networks. In *Proceedings IEEE INFOCOM*, 2013.
- [187] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.
- [188] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011.
- [189] Environmental protection agency advice note on disinfection by-products in drinking water. advice note no. 4. version 2. Technical report, United States Environmental Protection Agency, USEPA.

- [190] L. Wagenet, A. Heidekamp, and A. Lemley. Chlorination of drinking water. *Water Treatment NOTES. Cornell Cooperative Extension, College of Human Ecology*, 1988 (Updated October 2005).
- [191] Federal-Provincial-Territorial Committee on Drinking Water of the Federal-Provincial-Territorial Committee on Health and the Environment. Guideline technical document chlorine. *Guidelines for Canadian Drinking Water Quality*, pages 1–39, 2009.
- [192] L. Nováková and O. Stepankova. Radviz and identification of cluster in multi-dimensional data. In *13th International Conference Information Visualisation*, 2009.
- [193] G. Leban, B. Zupan, G. Vidmar, and I. Bratko. Vizrank: Data visualization guided by machine learning. *Data Mining and knowledge discovery*, 13:119–136, 2006.
- [194] William S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.
- [195] J. Fox, S. Weisberg, D.l Adler, D. Bates, G. Baud-Bovy, S. Ellison, D. Firth, M. Friendly, G. Gorjanc, S. Graves, R. Heiberger, R. Laboissiere, G. Monette, D. Murdoch, H. Nilsson, D. Ogle, B. Ripley, W. Venables, A. Zeileis, and R-Core. *car: Companion to Applied Regression*. MASS R Package, Version 2.1-0. Available from <http://cran.R-project.org>, 2015.
- [196] X. Bai, F. Wu, B. Zhou, and X. Zhi. Biofilm bacterial communities and abundance in a full-scale drinking water distribution system in shanghai. *Water and Health*, 8(3), 2010.
- [197] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.
- [198] I. Davidson and S. S. Ravi. *Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results*. Arfken and Weber. Series Lecture Notes in Computer Science. Springer, 2005.

- [199] M. Maechler, M. Rousseeuw, P. Struyf, A. Hubert, M. Hornik, K. Studer, and P. Roudier. *Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al.* R Package, Version 2.0.2. Available from <http://cran.R-project.org>, 2015.
- [200] J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–874, 1971.
- [201] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65, 1987.
- [202] G. Pison, A. Struyf, and P. J. Rousseeuw. Displaying a clustering with clusplot. *Computational Statistics & Data Analysis*, 30:381392, 1999.
- [203] C. Hennig. *fpc: Flexible Procedures for Clustering*. fpc R Package, Version 2.1-10. Available from <http://cran.R-project.org>, 2015.
- [204] J. Hernández, M.J. Ramírez, and C. Ferri. *Introducción a la minería de datos*. Ed. Pearson, 2004.
- [205] Wei-Yin Loh. Classification and regression trees. *John Wiley and Sons, Inc.*, 2011.
- [206] L. Breiman, J. Friedman, C. J. Stone, and R.A. Olshen. *Classification and regression trees*. CRC press, 1984.
- [207] C. Shalizi. *Regression Trees*. Department of Statistics. Carnegie Mellon University, 2006.
- [208] T. Therneau, B. Atkinson, and B. Ripley. *Recursive Partitioning and Regression Trees*. R Package, Version 4.1-10. Available from <http://cran.R-project.org>, 2015.
- [209] J. Benesty, J. Chen, Y. Huang, and I. Cohen. *Springer Topics in Signal Processing*, chapter Pearson Correlation Coefficient, pages 1–4. Springer, 2009.
- [210] V.F. Rodríguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P.M. Atkinson, and C. Jeganathan. Random forest classification of Mediterranean land cover using

- multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 121:93107, 2012.
- [211] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [212] K. R. Gray, P. Aljabar, R. A. Heckemann, and A. Hammers. Random forest-based similarity measures for multi-modal classification of alzheimer’s disease. *NeuroImage*, 65:167175, 2013.
- [213] B. Lantz. *Machine Learning with R*. PACKT Publishing, 2013.
- [214] Fortran original by Leo Breiman, R port by Andy Liaw Adele Cutler, and Matthew Wiener. *Breiman and Cutler’s Random Forests for Classification and Regression. Version 4.6-12*, 2015. URL <https://www.stat.berkeley.edu/~breiman/RandomForests/>.
- [215] K. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations*. Cambridge University Press; Cambridge, MA, USA., 2009.
- [216] M. Wooldridge. *An Introduction to MultiAgent Systems*. Second Edition, John Wiley & Sons, 2009.
- [217] Sycara. K. P. Multiagent systems. *AI Magazine*, 19(2), 1998.
- [218] M. Herrera, J. Izquierdo, R. Pérez-García, and I. Montalvo. Multi-agent adaptive boosting on semi-supervised water supply clusters. *Advances in Engineering Software*, 50:131136, 2012.
- [219] Freeman L. A set of measures of centrality based upon betweenness. *Sociometry*, 40(1):35–41, 1977.
- [220] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Applied Mathematics. PNAS*, 99(12):7821–7826, 2002.
- [221] J. Xu, P. S. Fischbeck, M. J. Small, J. M. Van Briesen, and E. Casman. Identifying sets of key nodes for placing sensors in dynamic water distribution networks.

- Journal of Water Resources Planning and Management ASCE*, 134(4):378–385, 2008.
- [222] L. Rossman. Wrpmd'99. In *The EPANET Programmer's Toolkit for Analysis of Water Distribution Systems.*, 1999.
- [223] U. Wilensky. *NetLogo*. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL., 1999. URL <http://ccl.northwestern.edu/netlogo/>.
- [224] P. M. Niquette, P. Servais, and R. Savoir. Impacts of pipe materials on densities of fixed bacterial biomass in a drinking water distribution system. *Water Resources*, 34(6):1952–1956, 2000.
- [225] A. Gorod, B. E. White, V. Ireland, S. J. Gandhi, and B. Sauseri. CRC Press, 2005.
- [226] L. Kurgan and P. Musilek. A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review*, 21(1):1–24, 2006.
- [227] L.C. Simoes, Azevedo N., Pacheco A., Keevil C.W., and Vieira M.J. Drinking water biofilm assessment of total and culturable bacteria under different operating conditions. *Biofouling*, 22:91–99, 2006.
- [228] M. Hamilton, J. Heersink, K. Buckingham-Meyer, and D. Goeres. *The Biofilm Laboratory*. MSU Center for Biofilm Engineering, 2003.
- [229] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
- [230] Eamonn J. Keogh and Michael J. Pazzani. Learning the structure of augmented bayesian classifiers. *Artificial Intelligence Tools*, 11(4):587–601, 2002.
- [231] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.
- [232] S. Kotsiantis. Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, 35(3):223–240, 2011.

-
- [233] Geoffrey I. Sammut, Claude; Webb, editor. *Encyclopedia of Machine Learning*. Springer, 2010.
- [234] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *2nd International Conference on Knowledge Discovery and Data Mining*, pages 202–207. AAAI Press, 1996.
- [235] P. H. Dreeszaen. *The key to understanding and controlling bacterial growth in Automated Drinking Water Distribution Systems. Second Edition*. John Wiley and Sons, 2003.
- [236] M. G. Paraje. *Antimicrobial resistance in biofilms. Science against microbial pathogens: communicating current research and technological advances*. Vilas Ed., 2011.