UNIVERSITAT POLITÈCNICA DE VALÈNCIA

máster en ingeniería hidráulica y medio ambiente mihma

Título del Trabajo Fin de Máster:

EVALUACIÓN DEL DESARROLLO DE BIOFILMS EN LOS SISTEMAS DE DISTRIBUCIÓN DE AGUA POTABLE MEDIANTE LA EXTRACCIÓN DE CONOCIMIENTO A TRAVÉS DE LOS DATOS (KNOWLEDGE DISCOVERY IN DATABASES - KDD)

Intensificación:

HIDRÁULICA URBANA

Autor:

RAMOS MARTÍNEZ, EVA

Director/es:

DR. PÉREZ GARCÍA, RAFAEL
DR. IZQUIERDO SEBASTIÁN, JOAQUÍN

Fecha: **SEPTIEMBRE, 2012**

Título del Trabajo Fin de Máster:

EVALUACIÓN DEL DESARROLLO DE BIOFILMS EN LOS SISTEMAS DE DISTRIBUCIÓN DE AGUA POTABLE MEDIANTE LA EXTRACCIÓN DE CONOCIMIENTO A TRAVÉS DE LOS DATOS (KNOWLEDGE DISCOVERY IN DATABASES - KDD)

Autor: ***RAMOS MARTÍNEZ, EVA***

Tipo	**A** ☐ **B** ☒		
Director **Codirector1** **Codirector2** **Tutor**	***RAFAEL PÉREZ GARCÍA*** ***JOAQUÍN IZQUIERDO SEBASTIÁN***	**Lugar de Realización** **Fecha de Lectura**	***VALENCIA*** ***SEPT, 2012***

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 microbiológica en su abastecimiento. 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. Son varias las investigaciones que se han llevado a cabo en este área. Sin embargo, los estudios realizados en relación a la influencia conjunta de las distintas características de los DWDSs en el desarrollo de biofilms, excepto notables excepciones, son escasos, debido a la complejidad de la comunidad y del entorno estudiado. El presente trabajo trata de cubrir esta carencia, estudiando el efecto de la interacción del conjunto de características físicas e hidráulicas de los DWDSs relevantes en el desarrollo de biofilms. Para ello utilizamos la metodología de extracción de conocimiento a través de los datos (KDD, del inglés Knowledge Discovery in Databases). Además, introducimos técnicas de ensamblaje adecuadas que nos permiten aumentar la robustez y precisión de los resultados obtenidos y así mejorar la metodología final propuesta de

ayuda a la toma de decisiones. La realización de este trabajo ha servido para confirmar la necesidad de estudiar el impacto que el conjunto de las características de los DWDSs tienen en el desarrollo de biofilms. Mostramos que el efecto que una variable tiene sobre este desarrollo depende del valor que tomen el resto de variables y así identificamos condiciones conjuntas, físicas e hidráulicas, que determinan el mayor o menor desarrollo de biofilms en el interior de las tuberías.

Resum:

Un dels principals objectius de les empreses encarregades de la gestió dels sistemes de distribució d'aigua potable (DWDS, de l'anglès Drinking Water Distribution Systems) és assegurar una alta qualitat microbiològica en el seu proveïment. No obstant açò, degut a l'existència de biofilms en tots ells, malgrat la presència de desinfectants residuals, fa que no es puga assegurar un control bacteriològic total; per açò, avui dia, els biofilms representen un paradigma en la gestió de la qualitat de l'aigua als DWDSs. Els biofilms són comunitats complexes de microorganismes recobertes d'un polímer extracel•lular que els dona una estructura i els ajuda a retenir l'aliment i a protegir-se d'agents tòxics. A més del risc sanitari que suposen pel seu paper com a refugi de patògens, existeixen molts altres problemes associats al desenvolupament de biofilms als DWDSs, com deteriorament estètic de l'aigua, biocorrosió i consum de desinfectants, entre d'altres. Son varies les investigacions que s'han dut a terme en aquest materia. No obstant, els estudis realitzats en relació amb la influència de les distintes característiques dels DWDSs en el desenvolupament de biofilms, excepte notables excepcions, són escassos, degut a la complexitat de la comunitat i de l'entorn estudiat. El present treball tracta de cobrir aquesta mancança, estudiant l'efecte de la interacció del conjunt de característiques físiques i hidràuliques dels DWDSs rellevants en el desenvolupament de biofilms. Per a aquest fi, emprem la metodología d'extracció de coneixement a través de les dades, (KDD, de l'anglès Knowledge Discovery in Databases). A més, la introducció de tècniques de acoblament adequades ens permet augmentar la robustesa i precisió dels resultats obtinguts i així millorar la metodologia final proposada d'ajuda a la presa de decisions. La realització d'aquest treball ha servit per confirmar la necessitat d'estudiar l'impacte que el conjunt de les característiques dels DWDSs tenen en el desenvolupament de biofilms. Mostrem com l'efecte que una variable té sobre aquest desenvolupament depèn del valor que prenguen la resta de variables i així identifiquem condicions conjuntes, físiques i hidràuliques, que determinen el major o menor desenvolupament de biofilms a l'interior de les canonades.

Abstract:

One of the main challenges of drinking water utilities is to ensure microbial high quality supply. 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 a total bacteriological control. Currently biofilms represent a real paradigm in water quality management for all DWDSs. Biofilms are complex communities of microorganisms bound 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. Numerous investigations have been carried out in this field. Nevertheless, the joint influence of the various DWDS characteristics in biofilm development, apart from a few exceptions, has been scarcely studied, due to the complexity of the community and the environment under study. The present work aims to help solve this problem studying the effect of the interaction among relevant hydraulic and physical characteristics of the DWDSs in biofilm development. To achieve this purpose we have chosen the framework of the KDD (Knowledge Discovery in Databases). Ensamble methods have been introduced to increase the robustness and the precision of the obtained results. The final aim is to improve the proposed methodology to assist in decision making. This work confirms the necessity of studying the impact that the joint characteristics of the DWDSs has in biofilm development. We show that the effect of one variable depends on the values of the rest of variables and, as a result, we are able to identify some joint physical and hydraulic scenarios that determine greater or lesser biofilm development in pipe walls.

Palabras clave:

Biofilm, sistemas de distribución de agua potable, KDD, minería de datos, meta-learning

Agradecimientos

En primer lugar me gustaría dar las gracias a mis directores, Rafael Pérez García y Joaquín Izquierdo Sebastián, por haber depositado su confianza en mí y poner toda su experiencia a mi disposición. También quiero agradecer a todos mis compañeros del grupo FluIng la buena acogida que he recibido, su compañerismo y los buenos ratos que pasamos juntos. Me gustaría agradecer especialmente a Manuel Herrera, por su implicación en este proyecto y por haber compartido conmigo su gran valía personal y profesional.

A nivel personal, nunca podré agradecer suficiente a mis padres, Modesto e Inma, y a mi super hermana, Alba, todo lo que han hecho y hacen por mí, gracias por siempre ayudarme a conseguir mis metas. También me gustaría agradecer muy mucho el apoyo de toda mi familia, especialmente el de mi tía Marian, mi tío Manu y mis primas Aitziber e Iratxe, que a pesar de la distancia siguen atentos mis pasos. Y, por supuesto, a mi extraordinaria abuela. Dar las gracias a mis amigos; a los de toda la vida por apoyarme estén donde estén y esté donde esté; y a los nuevos, por compartir alegrías y penas. Por último, un enorme gracias a Akis, por haberse ofrecido a acompañarme en esta aventura.

Índice

1. Introducción **1**

1.1. Objetivos del trabajo 5

1.2. Estructura del trabajo 6

1.3. Contribuciones del trabajo 7

2. Los biofilms en los sistemas de distribución de agua potable **9**

2.1. Formación de los biofilms en los DWDSs 11

2.2. Problemática asociada al desarrollo de biofilms en los DWDSs 14

2.2.1. Riesgo sanitario para el ser humano 14

2.2.2. Deterioro estético del agua 17

2.2.3. Proliferación de organismos superiores 18

2.2.4. Problemas operacionales 20

2.2.4.1. Consumo de desinfectante 20

2.2.4.2. Biocorrosión 22

2.3. Control del desarrollo de biofilms en los DWDSs 25

3. Extracción del conocimiento a través de los datos (KDD) **29**

3.1. Recopilación y selección de los datos 33

3.2. Pre-procesamiento y transformación de los datos 34

3.3. Aplicación de técnicas de minería de datos 36

4. Recopilación y selección de datos **43**

4.1. Comprensión del tema a tratar y planteamiento de objetivos 43

4.2. Generación de la base de datos 51

4.2.1. Recopilación de la información de las diferentes fuentes de datos 51

4.2.2. Selección de las variables relevantes y extracción de la información de interés 53

4.2.3. Unificación de la información y generación de la base de datos 53

5. Pre-procesamiento, transformación y visualización de los datos 57

5.1. Pre-procesamiento de los datos 59

5.1.1. Técnicas de detección: *clustering* 60

5.1.2. Técnicas de transformación: redes neuronales artificiales 61

5.1.3. Aplicación de las técnicas de detección y transformación 63

5.2. Transformación de los datos 64

5.3. Visualización de los datos 65

5.3.1. Visualización inteligente de los datos 66

5.3.1.1. Resultados visualización: RadViz 67

5.3.1.2. Resultados visualización: *Scatterplot* 70

6. Aplicación y resultados de las técnicas de minería de datos 73

6.1. Métodos de aprendizaje no supervisado 74

6.1.1. Clustering: algoritmo *Farthest-First* 74

6.1.1.1. Resultados: Clustering, algoritmo *Farthest-First* 75

6.1.2. Reglas de asociación: algoritmo *Ripple-Down* 78

6.1.2.1. Resultados: Reglas de asociación, algoritmo *Ripple-Down* 80

6.2. Métodos de aprendizaje supervisado 83

6.2.1. Reglas de decisión: algoritmo *Nearest-neighbor-like* 83

6.2.1.1. Resultados: Reglas de decisión, algoritmo *Nearest-neighbor-like* 84

6.2.2. Árboles de clasificación: algoritmo *Best-First* 86

6.2.2.1. Resultados: Árboles de clasificación, algoritmo *Best-First* 88

7. Ensamblaje **91**

7.1. Algoritmos de ensamblaje 92

7.1.1. El algoritmo *bagging* 92

7.1.2. El algoritmo *boosting* 94

7.1.3. El algoritmo *AdaBoost* 95

7.2. Caso de estudio: algoritmo EBARAMA 96

7.2.1. Resultados: algoritmo EBARAMA 98

8. Conclusiones **103**

8.1. Aplicaciones 105

8.2. Futuras líneas de investigación 107

8.3. Difusión del trabajo 108

Bibliografía **111**

Lista de Figuras

Figura 1.1. El sistema de distribución como un reactor de biofilm [MSU-CBE] **2**

Figura 1.2. Proceso del KDD **3**

Figura 2.1. Proceso de formación de un biofilm **13**

Figura 2.2. Interacciones tróficas generalizadas en los DWDSs [Evins, 2004] **18**

Figura 2.3. Biocorrosión [Videla, 1988] **23**

Figura 2.4. Tubería metálica dañada por MIC **24**

Figura 2.5. Costes asociados a la corrosión en EE.UU. [NACE] **24**

Figura 3.1. Esfuerzo requerido por cada fase del proceso KDD [Molina, 2006] **36**

Figura 3.2. Ejemplo de un árbol de clasificación sencillo [Sierra, 2006] **40**

Figura 4.1. Varios mecanismos participando en la acumulación de biofilms sobre una superficie en contacto con agua potable [le Puil, 2004] **44**

Figura 4.2. Crecimiento de microorganismos según la temperatura **47**

Figura 5.1. Proceso de pre-procesamiento y visualización **58**

Figura 5.2. Una red típica de tres capas de alimentación hacia adelante **62**

Figura 5.3. Estructura de la ANN utilizada **63**

Figura 5.4. Resultados obtenidos tras la aplicación de RadViz, VizRank y FreeViz **69**

Figura 5.5. Aplicación del algoritmo VizRank en los *scatterplots* **70**

Figura 6.1. Algoritmo *Farthest-First* trasversal para un conjunto de datos **75**

Figura 6.2. Visualización del clustering **76**

Figura 6.3. Adquisición del conocimiento en las reglas *Ripple-Down* [Wada, 1999] **79**

Figura 6.4. Una de las posibles representaciones de la estructura de las reglas *Ripple-Down* **80**

Figura 6.5. Hipotético árbol de clasificación *Best-First* **87**

Figura 7.1. Modelo de ensamblaje **91**

Figura 7.2. Gráfica de la evolución del proceso de remuestreo de EBARAMA **99**

Lista de Tablas

Tabla 2.1. Medidas a tomar en los sistemas de abastecimiento para el control de biofilms [Mains, 2008] **27**

Tabla 3.1. Forma general de la base de datos para técnicas no supervisadas **38**

Tabla 3.2. Forma general de la base de datos para técnicas supervisadas **39**

Tabla 3.3. Conjunto de reglas de decisión [Sierra, 2006] **40**

Tabla 4.1. Recopilación de datos de la bibliografía **52**

Tabla 4.2. Información disponible en la base de datos generada **55**

Tabla 5.1. El método de *clustering* para la detección de *outliers* **61**

Tabla 5.2. Variables y categorías de la base de datos **64**

Tabla 6.1. Medoides de los *clusters* **76**

Tabla 6.2. Aleaciones hierro-carbono **78**

Tabla 6.3. Valoración en función del índice kappa **81**

Tabla 6.4. Validación cruzada estratificada de las reglas de asociación **81**

Tabla 6.5. Matriz de confusión de las reglas de asociación **81**

Tabla 6.6. Resultado de las reglas de asociación **82**

Tabla 6.7. Resultado simplificado de las reglas de asociación **83**

Tabla 6.8. Validación cruzada estratificada de las reglas de decisión **84**

Tabla 6.9. Matriz de confusión de las reglas de decisión **85**

Tabla 6.10. Resultado de las reglas de decisión **86**

Tabla 6.11. Validación cruzada estratificada del árbol de clasificación **89**

Tabla 6.12. Matriz de confusión del árbol de clasificación **89**

Tabla 6.13. Resultado del árbol de clasificación **90**

Tabla 7.1. Entrenamiento del algoritmo *bagging* **93**

Tabla 7.2. Clasificación *bagging,* proceso de votación **94**

Tabla 7.3. El algoritmo *boosting* **95**

Tabla 7.4. Entrenamiento *AdaBoost* **96**

Tabla 7.5. Algoritmo EBARAMA **97**

Tabla 7.6. Tabla de la evolución del proceso de remuestreo de EBARAMA **99**

Tabla 7.7. Resultados de EBARAMA **100**

Tabla 7.8. Reglas obtenidas en EBARAMA **100**

Evaluación del desarrollo de biofilms en los sistemas de distribución de agua potable mediante la extracción de conocimiento a través de los datos (Knowledge Discovery in Databases– KDD)

Capítulo 1

Introducción

En los últimos años, diferentes factores han hecho que aumente el interés en la investigación de la calidad del agua de consumo humano con el fin de mejorar su protección y control en su distribución tras el tratamiento. Las nuevas técnicas analíticas, especialmente técnicas de conteo de bacterias, junto con el creciente grado de regulación de la calidad del agua y el hecho de que los consumidores, hoy en día, estén mucho más informados, han aumentado las expectativas sobre la calidad del agua servida. Es por ello que los gestores encargados de los servicios de agua intentan producir y hacer llegar a sus consumidores, cada vez, agua de mayor calidad.

Es un hecho que el propio diseño de los sistemas de distribución hace que el decaimiento de la calidad del agua durante el transporte sea inevitable. Sin embargo, muchas veces el diseño del sistema, por sí solo, no sirve para explicar la magnitud de dicho decaimiento. Las razones para un alto deterioro de la calidad del agua en los sistemas de distribución no están del todo claras, pero es sabido que uno de los principales agentes que influyen en este deterioro es la formación de biofilms en el interior de las tuberías. Los sistemas de distribución son los componentes mayoritarios de los servicios de agua y en su interior se dan numerosos procesos, físicos, químicos y biológicos. Se puede decir que las tuberías de los sistemas de distribución se asemejan a reactores de crecimiento de biofilm, con un complejo conjunto de componentes y reacciones que varían con el tiempo (Figura 1.1).

Los biofilms son estructuras colectivas de microorganismos que se forman en presencia de agua y se adhieren a superficies. Estas comunidades están revestidas por una capa protectora que ellas mismas segregan. De este modo, las bacterias que forman parte de los biofilms son capaces de resistir a los biocidas y a los antibióticos de un modo más eficaz que aquellas que viven como organismos libres, soportando dosis considerablemente más altas de antimicrobianos.

En los últimos años, los avances en las técnicas microscópicas han evidenciado el hecho de que los biofilms pueden convertirse en hábitats transitorios o a largo plazo de

microorganismos higiénicamente relevantes. Estos patógenos que se encuentran, incluso por debajo del límite de detección, en el agua pueden accidentalmente unirse a biofilms, los cuales pueden actuar como su reservorio ambiental, representando una potencial fuente de contaminación del agua [Wingender & Flemming, 2011]. En los sistemas de distribución de agua potable (DWDSs, del inglés Drinking Water Distribution Systems) los biofilms se fijan fuertemente a la pared interior de las tuberías y la modifican mientras captan más nutrientes y nuevas bacterias.

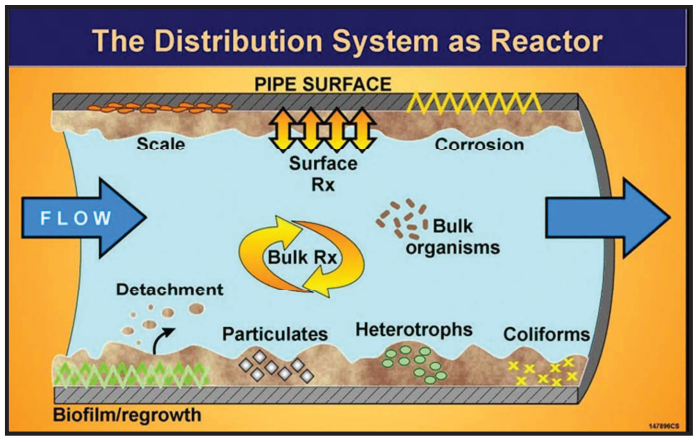

Figura 1.1. El sistema de distribución como un reactor de biofilm [MSU-CBE]

Aparte del riesgo microbiológico que suponen los biofilms en los DWDSs, su presencia en estos sistemas también lleva asociada muchos otros aspectos negativos que favorecen el decaimiento de la calidad del agua en los sistemas de distribución. De hecho, estas comunidades microbianas son responsables de muchos de los problemas que se dan en estos sistemas. Los más destacados son: deterioro estético del agua [Gelves, 2005], proliferación de organismos superiores [Chowdhury, 2011] y problemas operacionales [Lopes *et al.*, 2009], como aumento de las tasas de corrosión [Videla & Herrera, 2005] y consumo de desinfectante [Dirk de Beer, 1994]. Aunque, en la mayoría de países se mantienen cantidades reguladas de desinfectante residual en los DWDSs, estas no son suficientes para evitar la formación de biofilms en el interior de las tuberías. Es por ello que, hoy en día, los biofilms representan un paradigma en la gestión de la calidad del agua en todos los DWDSs.

La complejidad del ambiente que se crea en el interior de las tuberías hace que existan hábitats heterogéneos a lo largo del tiempo y del espacio, haciendo que los biofilms existan a

diferentes niveles dentro de los DWDSs. La supervivencia y el rebrote de microorganismos en estos sistemas, pueden estar afectados no sólo por factores biológicos sino también por la interacción de diferentes factores [Yu *et al.*, 2010]. Son numerosos los estudios que se han llevado a cabo en relación a la influencia de las distintas características de los DWDSs en el desarrollo de biofilms [Tsai, 2005; Tsvetanova, 2006; Zhou *et al.*, 2009; Silhan *et al.*, 2006], sin embargo, excepto notables excepciones [Simoes *et al.*, 2006], apenas se ha investigado su influencia conjunta (más de dos factores) debido a la complejidad de la comunidad y del entorno estudiado. El presente trabajo pretende cubrir esta carencia, estudiando el efecto de la interacción del conjunto de características físicas e hidráulicas de los DWDSs relevantes en el desarrollo de biofilms. De esta manera, se pretende lograr identificar las condiciones que determinan el mayor o menor desarrollo de biofilms en el interior de las tuberías. Con este fin, se ha recurrido a la utilización de la metodología de extracción de conocimiento a través de los datos (KDD, del inglés Knowledge Discovery in Databases) como herramienta en este estudio.

La metodología KDD es un *proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles* [Fayaad, 1996]. El KDD surge de la necesidad de manejar grandes y/o complejos conjuntos de datos; de donde además de llevar a cabo una extracción de la información eficiente, se encarga de la preparación de dichos datos, su análisis y la interpretación de los resultados obtenidos [Witten *et al.*, 2011]. El KDD es un proceso interactivo e iterativo que se divide en diferentes etapas (Figura 1.2).

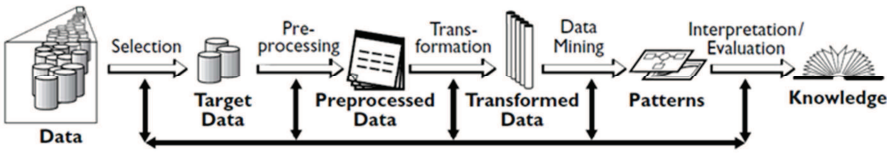

Figura 1.2. Proceso del KDD

En un primer paso se procede a la compresión del tema a tratar mediante el conocimiento de antecedentes y planteamiento de objetivos. A continuación, se recopila información de las diferentes fuentes de datos y se integran y seleccionan las variables relevantes.

Una vez obtenidos los datos se pre-procesan y transforman. En esta etapa se procede a la búsqueda de los datos atípicos, eliminación de los incorrectos, reconstrucción de los perdidos

y normalización. De esta manera se obtiene una extensa base de datos a la que aplicar técnicas de minería de datos.

Tras este paso, se aplican la o las técnicas de minería de datos que se consideren más adecuadas. En nuestro caso, utilizamos métodos de aprendizaje automático tanto supervisados como no supervisados. El aprendizaje no supervisado se distingue del aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. Así, el aprendizaje no supervisado, típicamente, trata los objetos de entrada como un conjunto de variables aleatorias. En cambio, el aprendizaje supervisado se usa para deducir una función a partir de datos de entrenamiento.

En el estudio experimental propuesto en este trabajo el aprendizaje automático no supervisado se lleva a cabo mediante *clustering* y reglas de asociación. Mientras que las técnicas de aprendizaje supervisado utilizadas para nuestro estudio son árboles de clasificación y reglas de decisión.

El proceso KDD, propiamente dicho, finaliza con la interpretación de los patrones obtenidos y la consolidación del conocimiento descubierto, mediante la incorporación del conocimiento en la interpretación del sistema o, simplemente, mediante la documentación de las partes de interés [Díaz-Arevalo, 2010].

Sin embargo, en nuestro caso se ha pretendido incorporar un paso más al proceso KDD. Se trata de la integración (de los resultados obtenidos en los diferentes métodos de aprendizaje automático aplicados) en un único modelo. Este proceso se ha llevado a cabo con la ayuda de técnicas de ensamblaje. La idea subyacente en estas técnicas es combinar "reglas simples" para formar un conjunto de manera que el rendimiento del modelo obtenido, unificado en un solo proceso final, mejore [Meir & Rätsch, 2003]. De esta manera, mediante un proceso de voto y remuestreo de los *outputs* de las técnicas de minería de datos (que actúan como *inputs* del ensamblaje propuesto), se logra mejorar la robustez y precisión de la metodología de ayuda a la toma de decisiones finalmente propuesta. Además, adicionalmente, también se consigue reducir la incertidumbre asociada al desarrollo de biofilms en los DWDSs.

1.1 Objetivos del trabajo

El principal objetivo del presente estudio es determinar qué tuberías serán propensas al desarrollo de biofilms en su interior, en función de las características físicas e hidráulicas de las mismas. A través de este conocimiento se desarrollarán las bases para la implementación de las herramientas necesarias para localizar las áreas de los DWDSs que presentan un mayor riesgo de desarrollo de biofilms. Esto puede suponer una gran mejora de la calidad del agua que circula por los DWDSs, permitiendo mitigar la problemática asociada al desarrollo de biofilms en estos sistemas.

El camino para conseguir nuestro objetivo pasa por diferentes objetivos parciales a satisfacer, entre los que se encuentran los siguientes:

- Recopilación y puesta al día de la información existente en relación al desarrollo de biofilms en los DWDSs, abarcando tanto los aspectos microbiológicos como los relacionados con la hidráulica de los sistemas. Este proceso se llevará a cabo con el fin de asimilar el conocimiento alcanzado en relación al comportamiento de los biofilms en estos sistemas, unificar criterios y determinar cuáles son las principales características, físicas e hidráulicas, de los DWDSs que afectan a su desarrollo.

- Generación de una base de datos completa y extensa mediante la recopilación de datos obtenidos de diferentes fuentes (estudios de cuantificación de biofilms en DWDSs reales o simulados en laboratorio), en una primera etapa, y la aplicación posterior de técnicas de pre-procesamiento y transformación de los datos. Se procederá a la búsqueda de datos atípicos, eliminación de datos incorrectos, reconstrucción de datos perdidos y discretización de las variables continuas para normalizar la base de datos. Finalmente, mediante técnicas de visualización se realizará un estudio exploratorio de los metadatos obtenidos para observar si existe algún tipo de inconsistencia en la base de datos y comprobar su fiabilidad.

- Aplicación de diferentes técnicas de minería de datos sobre los metadatos, que nos permita identificar, en función de las características estudiadas, patrones de comportamiento de los biofilms en los DWDSs. Se aplicarán varias técnicas de

minería de datos con el fin de poder realizar distintos tipos de análisis mediante diferentes algoritmos que nos permitan extraer la máxima información posible de la base de datos generada.

- Combinación de los resultados obtenidos mediante las diferentes técnicas de minería de datos empleadas. Este proceso se llevará a cabo a través de técnicas que combinen estos resultados. El objetivo será mejorar la robustez y precisión de los resultados obtenidos y de esta manera optimizar la metodología final propuesta de ayuda a la toma de decisiones. Además, de esta manera, se conseguirá reducir la gran incertidumbre asociada al proceso de desarrollo de biofilms en los DWDSs

1.2 Estructura del trabajo

Este trabajo se divide en 8 capítulos:

- Capítulo 1. Corresponde al presente capítulo. Expone el perfil del trabajo, sus objetivos y contribuciones.

- Capítulo 2. Profundiza en el conocimiento de los biofilms, los factores que determinan su desarrollo en los DWDSs y los problemas asociados a su presencia en estos sistemas.

- Capítulo 3. Describe la metodología KDD, su origen, evolución y aplicaciones actuales.

- Capítulo 4. Recopila las fuentes de datos e integra y selecciona las variables relevantes. Genera una base de datos definitiva mediante técnicas de pre-procesamiento y transformación de los datos.

- Capítulo 5. Realiza un análisis exploratorio de los metadatos obtenidos mediante técnicas de visualización inteligente.

- Capítulo 6. Aplica diferentes técnicas de minería de datos. Interpreta y evalúa los resultados obtenidos.

- Capítulo 7. Combina los resultados obtenidos anteriormente con ayuda de técnicas de ensamblaje.

- Capítulo 8. Resume las principales conclusiones, aplicaciones y futuras líneas de investigación de este trabajo.

1.3 Contribuciones del trabajo

El propósito del presente trabajo es lograr una mayor comprensión de las consecuencias reales que las interacciones de las diferentes características físicas e hidráulicas de los DWDSs tienen en el desarrollo de biofilms en estos sistemas.

Hasta ahora, la mayoría de los enfoques en este área se han centrado en el estudio del efecto de uno o dos factores en el desarrollo de los biofilms. Este trabajo, en conocimiento de la autora, es el primero que trata de estudiar el efecto combinado de las características físicas e hidráulicas de los DWDSs más relevantes en el desarrollo de biofilms.

Esta propuesta, que consideramos innovadora, pretende sentar las bases para el desarrollo de una herramienta capaz de identificar y predecir las condiciones que favorecen un alto desarrollo de biofilms en los DWDSs y, por lo tanto, las áreas de los DWDSs que son propensas a albergar este elevado desarrollo. De esta manera, los gestores de los servicios de agua contarían con una herramienta complementaria de ayuda a la toma de decisiones que aumentaría la eficacia en la gestión de los servicios de agua y ayudaría a mejorar la calidad del agua servida.

Son numerosos los aspectos relacionados con la gestión de los DWDSs que pueden verse favorecidos con la introducción de esta herramienta en los servicios de agua. A continuación se presentan algunos de ellos.

Esta herramienta puede ser muy útil en las tareas de prevención y mantenimiento de la red de abastecimiento. Por una parte, conociendo las áreas propensas a un alto desarrollo de biofilm pueden llevarse a cabo lavados hidráulicos dirigidos y, de esta manera, conseguir un ahorro en el tiempo y en el dinero invertidos, logrando aumentar la eficiencia del proceso. Por otra parte, si se tiene en cuenta el hecho de que los biofilms pueden aumentar las tasas de

corrosión de las tuberías metálicas, localizar qué tuberías tienden a desarrollar más biofilm puede ayudar a mejorar los métodos de prevención de averías y fugas en la red.

De la misma manera, su uso puede ser higiénicamente relevante ya que los biofilms están involucrados en el consumo de desinfectante residual de los DWDSs y conocer qué tuberías tienden a tener un mayor desarrollo de biofilm puede ser útil para optimizar la modelación del consumo de desinfectante en pared y lograr una mayor precisión en la localización de los puntos de cloración.

Finalmente, cabe destacar, la utilidad de dicha herramienta en el diseño de redes de abastecimiento de agua potable. El conocimiento de qué características físicas e hidráulicas de los DWDSs determinan un alto grado de desarrollo de biofilms puede ser de gran utilidad en el diseño de redes de abastecimiento. De esta manera, se podrá evitar, en la medida de lo posible, la existencia de dichas áreas en los futuros DWDSs.

En definitiva, la implementación de una herramienta capaz de identificar qué tuberías tienden a presentar un mayor desarrollo de biofilm servirá para mitigar los efectos negativos asociados al desarrollo de biofilms en los DWDSs. De esta manera, se ayudará a la mejora en la gestión de dichos servicios de agua y se logrará aumentar la calidad del agua suministrada.

Capítulo 2

Los biofilms en los sistemas de distribución de agua potable

Los biofilms son comunidades complejas de microorganismos que se forman en medios acuosos adheridos a una superficie y recubiertos de un polímero extracelular de tipo polisacárido, el *glicocalix*, que les da estructura y protección y les ayuda a retener el alimento y a protegerse de agentes tóxicos. Es por ello que las bacterias que forman parte de los biofilms son capaces de resistir a los biocidas y a los antibióticos de un modo más eficaz que aquellas que viven como organismos libres, y soportan dosis considerablemente mayores de productos antimicrobianos. Por este motivo, un biofilm desarrollado es muy resistente y representa un problema cuando se precisa un entorno limpio y desinfectado.

Actualmente, uno de los principales objetivos de las empresas encargadas de la gestión de los DWDSs es asegurar una alta calidad microbiológica en su abastecimiento. Sin embargo, pese a que en los sistemas de abastecimiento se llevan a cabo tratamientos de agua y procesos de desinfección que eliminan la mayor parte de las bacterias que se encuentran en el agua, el agua que se produce no es estéril y la presencia de desinfectantes residuales en los sistemas de distribución no es suficiente para evitar el desarrollo de biofilms.

Si bien los desinfectantes clorados son los únicos desinfectantes con permanencia en el sistema, se encuentran limitados debido a que pueden generar efectos perjudiciales para la salud humana. El Real Decreto 140/2003 de la legislación española, que fija los valores paramétricos a cumplir desde el punto de captación hasta el punto donde se pone el agua a disposición del consumidor, establece que cuando se utilice el cloro, o derivados, como método de desinfección del agua para consumo humano, el cloro libre residual no podrá superar la cantidad de 1.0 mg/l en la red de distribución. Si el método de desinfección empleado es la cloraminación los parámetros a controlar serán la cantidad de nitritos y de

cloro combinado residual, y en este caso, no podrá superar los 2.0 mg/l en la red de distribución.

La cantidad de desinfectante presente en el agua de los sistemas de distribución está limitada por sus implicaciones para la salud de los consumidores ya que, aparte de la toxicidad propia del cloro, los compuestos clorados reaccionan con la materia orgánica presente en el agua generando, entre otros, compuestos carcinógenos como los trihalometanos (THM). Los estudios epidemiológicos asocian determinadas exposiciones a THM y, en general, la exposición a subproductos de la desinfección (SPD) con efectos sobre la salud; como el cáncer de vejiga y determinados defectos de nacimiento en recién nacidos de madres expuestas a estos SPD. Los THM que se encuentran de manera más frecuente en el agua de consumo humano son el cloroformo, el bromodiclorometano (BDCM), el dibromoclorometano (DBCM) y el bromoformo. La Agencia internacional de investigación sobre el cáncer clasifica el cloroformo y el BDCM como posibles carcinógenos para los humanos en ciertas condiciones de exposición. Esto quiere decir que, a pesar de que existen indicios de su carcinogeneidad en animales en estudios experimentales, la evidencia es limitada en humanos [Agència de salut pública de Barcelona]. El bromoformo y el DBCM no se han llegado a clasificar como cancerígenos. Debido al riesgo asociado a estos SPD el Real Decreto 140/2003 de la legislación española y la normativa europea de criterios sanitarios de la calidad del agua de consumo humano fijan una concentración máxima permitida de THM totales (suma de cloroformo, BDCM, DBCM y bromoformo) de 100µg/l. Estos límites legislativos han sido fijados estableciendo unos márgenes de seguridad que garanticen un elevado grado de protección a la población.

Aparte de la limitación existente en la cantidad de desinfectante residual utilizado en los DWDSs debido a la implicación para la salud humana de sus subproductos, se debe tener en cuenta que, además, el desinfectante también es consumido a lo largo del sistema de distribución. El gasto de desinfectante en los DWDSs se produce en parte por el agua que circula por la red, donde es consumido por sustancias presentes en el agua y por condiciones físico-hidráulicas tales como temperatura, agitación y turbulencias, entre otras. A este consumo hay que añadir el que se produce en la interfase con las paredes de las conducciones. En el agua, el consumo depende principalmente de la temperatura y del contenido en materias orgánicas disueltas y otras inorgánicas capaces de ser oxidadas. En la interfase con las paredes, el consumo de cloro se produce por la interacción con los productos de corrosión, los

depósitos y la biomasa fijada en las paredes, los biofilms [Ramírez, 2005]. Es por ello que adicionalmente existe una disminución exponencial de cloro residual con el tiempo que hace que la presión sobre el biofilm, en los DWDSs, disminuya según nos alejamos del punto de cloración.

La resistencia propia de los biofilms, la limitación en las dosis de desinfectantes residuales y el consumo de desinfectante que se da a lo largo del propio DWDS hacen que, aunque en la mayoría de países se mantengan cantidades reguladas de desinfectante residual en los sistemas de distribución, los biofilms actualmente representen un reto en la gestión de la calidad del agua en todos los DWDSs.

2.1 Formación de los biofilms en los DWDSs

Las aguas naturales contienen miles de especies diferentes de bacterias, muchas de las cuales no han sido cultivadas y tampoco identificadas y el número de estos microorganismos varía considerablemente entre diferentes tipos de agua. Los microorganismos implicados en la formación de biofilms en los DWDSs son aquellos que han sido liberados directamente en la planta de tratamiento o que han sido introducidos en el sistema de distribución en algún punto aguas abajo de la planta de tratamiento.

Las técnicas de potabilización de agua actuales logran eliminar la mayor parte de bacterias presentes, sin embargo, no consiguen eliminar su totalidad. Algunos de los medios que tienen estos organismos para superar estas técnicas incluyen su asociación a partículas de turbiedad que logren pasar la barrera de la filtración en la planta de tratamiento o penetrar coligados a partículas de los finos del carbón activado utilizado en la filtración [Gelvés, 2005].

Actualmente se asume que, si una planta de tratamiento funciona adecuadamente, el número de bacterias patógenas liberadas al sistema de distribución será bajo. Sin embargo, los microorganismos también pueden entrar en los DWDSs por uniones, tuberías rotas o deficientes y/o por procesos como los transitorios hidráulicos debidos a fallos en el funcionamiento de los sistemas de distribución donde pueda haber intrusión de aguas contaminadas.

Los biofilms se forman como respuesta de las bacterias ante un medio adverso, como lo es el agua potable, debido a su bajo contenido en nutrientes. Dependiendo de la especie, las bacterias adoptarán diferentes estrategias para la formación de los biofilms. Unas cambiaran su cubierta para hacerla más hidrófoba y dirigirse hacia las paredes; otras irán moviéndose directamente con sus *flagelos* o *pili*, y otras caerán al fondo por gravedad [Piera, 2002]. Los biofilms se forman siempre que haya agua en contacto con una superficie sólida (por lo que los DWDSs son ambientes propicios para su desarrollo). Se fijan fuertemente, contra la repulsión inicial, a la pared interior de las tuberías y la modifican, mientras captan más nutrientes y nuevas bacterias.

El desarrollo de un biofilm maduro puede llevar varias horas o varias semanas, dependiendo del agente que origine su adhesión, el tipo de tratamiento del agua, las condiciones de operación del sistema, la velocidad del flujo u otros.

Según el *The Cooperative Research Center for Water Quality and Treatment* esta es la secuencia de eventos que lleva a la formación de un biofilm sobre una superficie (Figura 2.1).

1. Cualquier superficie inmersa en agua, instantáneamente, atrae moléculas tanto orgánicas como inorgánicas del agua que la rodea, formando una película de preparación. La formación de esta película inicial es de especial importancia en ambientes con poco contenido en nutrientes como el agua potable, donde la acumulación de moléculas orgánicas en la superficie crea un ambiente localizado relativamente rico en nutrientes.

2. Las primeras bacterias colonizadoras se adhieren a la superficie. Debido a la presencia de esta película de preparación rica en nutrientes, las bacterias que encuentran el camino a esta superficie están en ventaja con respecto a las que están en la fase acuosa baja en nutrientes. Las primeras bacterias colonizadoras se multiplican y así condicionan la superficie para favorecer la colonización por otras bacterias u organismos superiores, dando lugar a la maduración del biofilm y a la sucesión de la población.

3. Las fuerzas de corte, ejercidas por el agua circulante, producen un impacto en la estabilidad mecánica del biofilm causando una erosión continua de las capas

superficiales. De hecho, las fuerzas hidráulicas pueden limitar el grosor del biofilm. Otro fenómeno importante es la *muda*, desprendimientos ocasionales de grandes porciones de biofilm cuando este alcanza una densidad crítica. Estos desprendimientos pueden ser causados por un cambio repentino de las fuerzas de corte, por cambios en las concentraciones de desinfectante o por las propias bacterias. Como resultado, la formación del biofilm es un continuo proceso de adherencia, desarrollo y pérdida (o desprendimiento). El desprendimiento de biofilm ofrece a los microorganismos desprendidos el potencial para colonizar superficies aguas abajo y, por lo tanto, propagarse por los sistemas de distribución.

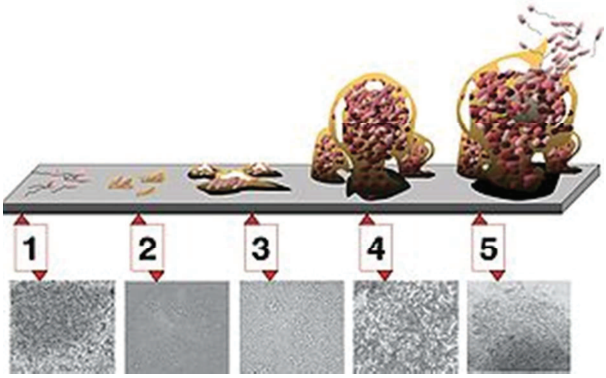

Figura 2.1. Proceso de formación de un biofilm

La formación de un biofilm es un proceso dinámico, lo que produce desconfianza acerca de la calidad del agua que llega a las casas de los usuarios y que podría producir problemas en la salud humana. Sin embargo, no es un proceso aleatorio sino que sigue una sistemática que permite su predicción [Piera *et al.*, 2002]. Un modelo que ayude a determinar tasas de formación de biofilms en tuberías sería de gran interés y ayuda para desarrollar programas de control de biofilms en los DWDSs, lo que supondría una gran mejora en la gestión de estos sistemas.

2.2 Problemática asociada al desarrollo de biofilms en los DWDS

La presencia de biofilms en los DWDSs lleva asociada numerosos aspectos negativos. De hecho, estas comunidades microbianas son responsables de muchos de los problemas que se dan en estos sistemas.

Los problemas asociados al desarrollo de biofilms en los DWDSs siguen la siguiente clasificación:

- Riesgo sanitario
- Deterioro estético del agua
- Proliferación de organismos superiores
- Problemas operacionales
    - Consumo de desinfectante
    - Biocorrosión

2.2.1 Riesgo sanitario para el ser humano

Las bacterias tienen capacidad para adaptarse y desarrollarse en los biofilms de los DWDSs, los cuales les ofrecen todos los microambientes nutritivos y electroquímicos necesarios para su evolución y protección contra los desinfectantes [Nieto *et al.*, 2009]. Es por ello que los biofilms pueden convertirse en reservorio de bacterias patógenas, ya que actúan como refugio protegiéndolas de los desinfectantes. Muchas de estas bacterias han sido identificadas en los biofilms de los DWDSs [Momba *et al.*, 2000] aunque se desconoce su tiempo de supervivencia en ellos, ya que este depende de la especie y de las condiciones del sistema.

Dentro de las bacterias patógenas se pueden diferenciar las bacterias patógenas primarias y las secundarias.

Las primarias son aquellas capaces de producir por sí solas enfermedad en el huésped (en este caso, el ser humano). Entre los patógenos de los DWDSs estas bacterias suelen ser minoritarias, aunque existen algunos casos, como por ejemplo los de la familia *Enterobacteriaceae,* a la que pertenecen bacterias como *Escherichia coli* o la *Salmonella*.

Las bacterias patógenas secundarias no pueden establecer por sí solas ninguna enfermedad ya que no resisten los mecanismos de defensa del huésped. Sólo logran colonizar cuando estos mecanismos son deprimidos por varias causas, entra las que se encuentran agentes físicos, como la temperatura o la humedad y agentes químicos, como los corticosteroides y las infecciones por un patógeno primario.

Los patógenos secundarios también son llamados patógenos oportunistas y son de especial importancia en los DWDSs porque pueden causar enfermedades en personas con el sistema inmune deprimido. Las personas mayores, niños, enfermos de cáncer recibiendo quimioterapia o radioterapia, personas infectadas con el VIH y pacientes con quemaduras o trasplantes en hospitales son especialmente susceptibles a infecciones por bacterias oportunistas [Jarvis, 1990].

En la actualidad, en los DWDSs se usan las bacterias *coliformes* como indicadores de la presencia de microorganismos patógenos. La denominación genérica *coliformes* designa a un grupo de especies bacterianas que tienen ciertas características bioquímicas en común e importancia relevante como indicadores de contaminación del agua. Su presencia representa un riesgo potencial para la salud pública y es por ello que estas bacterias son usadas como el principal indicador microbiológico de la calidad del agua. Su ausencia indica que el agua es bacteriológicamente segura, por lo que en los DWDSs las bacterias *coliformes* no deben ser detectadas en más de cierto tanto por ciento de las muestras de agua tomadas. Sin embargo, debido a que los monitoreos sólo se llevan a cabo para bacterias planctónicas presentes en el agua, se suele subestimar la cantidad de bacterias en el sistema [Parsek & Fuqua, 2003], ya que en los DWDSs el 95% de la biomasa se encuentra en las paredes de las tuberías [Flemming *et al.*, 2002].

En los países desarrollados, los organismos tradicionalmente asociados con suministros de agua contaminada, como *Vibrio chlolerae* y otros patógenos entéricos han dejado de ser un problema serio. Aunque, desafortunadamente, existen algunos patógenos que siguen causando contaminaciones periódicas de los DWDSs, entre los que se incluyen virus entéricos y protozoos.

Si bien, la mayoría de las bacterías presentes en los biofilms de los DWDSs no son patógenas [Szewzyk *et al.*, 2000], es importante tener en cuenta que, aparte del hecho de que existen bacterias patógenas con capacidad para adaptarse y desarrollarse en los biofilms de los DWDSs, también se ha demostrado que un tratamiento prolongado puede seleccionar subpoblaciones de bacterias resistentes al cloro [Ridway & Olso, 1982; Walsh, 1989]. Por ejemplo, se ha demostrado que todas las especies del genero *Mycobacterium* son resistentes a los métodos de desinfección estándares [Taylor, 2000] y persisten durante largos periodos de tiempo en los DWDSs [von Reyn *et al.*, 1994].

Entre otros aspectos a destacar, se encuentra el hecho de que algunos patógenos pueden crecer y persistir en los DWDSs utilizando los productos metabólicos producidos por miembros de comunidades de biofilms no patógenas [Steirnet *et al.*, 1998]. Este hecho tiene especial relevancia para organismos como *Legionella pneumophila*, que no puede crecer únicamente con los nutrientes presentes en el agua [Wadowsky *et al.*, 1983]. Por último, hay que señalar que también existen casos en los que los biofilms pueden aumentar la patogeneicidad de las bacterias, como es el caso de *L. pneumophila.* Numerosos estudios sugieren que la interacción de esta con otras especies en biofilms mixtos puede aumentar su patogenicidad [Morris *et al.*, 1979; Harf, 1988].

Los seres humanos sanos necesitan una dosis de exposición a patógenos relativamente alta, entre $10^6 – 10^{10}$ células, para causar infección o enfermedad en los humanos sanos, ya sea por vía oral o intranasal. En ocasiones, estas concentraciones de patógenos han sido encontradas en los biofilms de los DWDSs. Olson en 1982 detectó en la capa superficial de una tubería de argamasa *Acinetobacter* en niveles superiores a $10^9/cm^2$, cantidad más que suficiente para causar enfermedad en cualquier ser humano y/o animal [EPA, 2002].

A pesar de que las bacterias son las más comunes en los biofilms de los DWDSs, también se han identificado otros tipos de organismos, potencialmente patógenos, como los virus. La mayoría de virus encontrados en los DWDSs, y que tienen incidencia sobre la salud, son los llamados virus entéricos que son conocidos por causar enfermedades gastrointestinales. En los biofilms, los virus pueden acumularse, pero no reproducirse. Aunque es sabido que se encuentran diez veces más virus en los biofilms que en el flujo de agua en presencia de cloro, y en ausencia de este, 20 veces más que en el flujo de agua [Gelves, 2005], lo que subraya el papel protector y de reservorio de organismos de los biofilms.

También hay que destacar el hecho de que los biofilms ofrecen a las bacterias los microambientes nutritivos y electroquímicos necesarios para la evolución y protección contra los desinfectantes. Adicionalmente, los biofilms también pueden tener implicaciones para la salud humana de manera indirecta ya que la proliferación excesiva de la actividad microbiológica puede generar interferencias con los métodos utilizados para monitorear parámetros transcendentes para la salud [Chowdury, 2011], como el recuento de *coliformes.*

2.2.2 Deterioro estético del agua

Los biofilms están asociados con los problemas de decoloración del agua y la generación de malos sabores y olores.

Los biofilms representan pequeños ecosistemas dentro de los DWDSs y, si bien, las bacterías son las formadoras de los biofilms, y las más abundantes, existen otros tipos de organismos asociados a estos ecosistemas. De hecho, la principal causa para el deterioro estético del agua en los DWDSs, no son las bacterias, sino los hongos que se encuentran en los biofilms. Esto se debe a que muchos de los productos y subproductos del metabolismo de estos organismos tienen la capacidad de infundir al agua tratada sabor y olor, lo cual afecta directamente al consumidor final [Gelves, 2005].

Se ha encontrado que las algas que tienen la capacidad de crecer en los sistemas de distribución en ausencia de luz y alojadas en los biofilms también pueden deteriorar las características organolépticas del agua en los DWDSs. Estas algas pueden proliferar en la oscuridad debido a su capacidad de desarrollar metabolismos heterotróficos, utilizar el carbono como fuente de energía y desarrollarse en los biofilms [Codony *et al.*, 2003]. Otros organismos presentes en los biofilms también pueden aportar sustancias generadoras de sabores y olores, como nematodos y amebas [Env. Agency UK, 1998].

La decoloración del agua que circula por los DWDSs está causada por las sales de hierro y magnesio desprendidas como producto de los procesos de corrosión de las paredes internas de las tuberías. Es sabido, que los procesos de corrosión pueden verse favorecidos por el metabolismo de algunos de los microorganismos presentes en los DWDSs. Este fenómeno es de especial importancia en EE.UU. y en algunos países europeos donde el hierro de fundición fue el primer material utilizado en los DWDSs. El problema de la decoloración es de gran impacto para el consumidor final, generando problemas en el lavado de la ropa o eventos de agua roja. Los problemas asociados a episodios de aguas decoloradas son principalmente estéticos. Sin embargo, una consecuencia más importante puede ser la pérdida de desinfectante residual y el posterior aumento del desarrollo de biofilms en los DWDSs [Imran *et al.*, 2006].

2.2.3 Proliferación de organismos superiores

Los biofilms en los DWDSs pueden servir como base de la cadena alimentaria de hongos, protozoos, gusanos y crustáceos, entre otros. Estos organismos pueden estar presentes en los DWDSs incluso en presencia de desinfectante residual [Chowdhury, 2011].

Las cadenas tróficas en los sistemas de distribución son relativamente cortas y la mayoría de las especies de animales presentes pertenecen al mismo nivel trófico. La mayor parte son detritívoros y ramoneadores, y aunque se han encontrado algunas especies de pequeños carnívoros, carnívoros de mayor tamaño son muy escasos o inexistentes. De manera general, las interacciones tróficas existentes en los DWDSs se encuentran resumidas en la Figura 2.2.

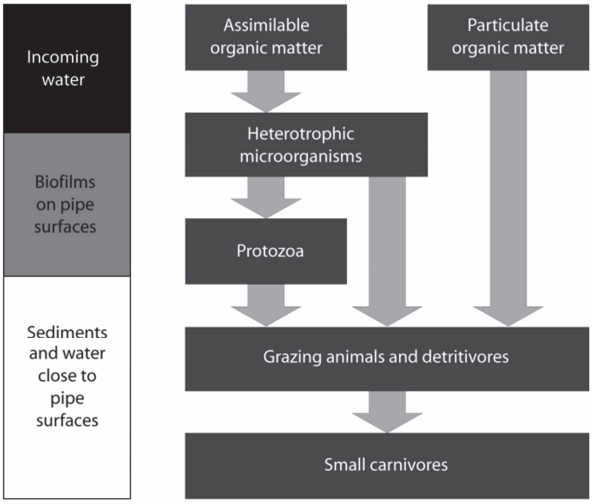

Figura 2.2. Interacciones tróficas generalizadas en los DWDSs [Evins, 2004]

En la literatura existen informes que atestiguan la presencia de animales en los DWDSs en Norte América, África, Sur y Este de Asia desde finales del siglo XIX (antes de que se extendiese el uso de la filtración y la desinfección) hasta el siglo XXI. Por ejemplo, en el Reino Unido en los años 60 y 70 se estudió la población animal de los DWDSs; se muestrearon unos 50 DWDSs, y se encontraron animales en todos ellos, aunque los gestores de los servicios de agua y sus consumidores a menudo no se percataban de su presencia [Evins, 2004].

Los animales pueden estar presentes en los DWDSs porque:

- Entran en el DWDSs con el agua, habiendo superado los procesos de tratamiento o habiendo colonizado parte de la planta de tratamiento.
- Entran por defectos en la integridad del sistema de distribución, como reservorios mal tapados.
- Forman poblaciones reproductivas en los propios sistemas de distribución.

La presencia de animales puede ser visible y desagradable para el consumidor si aparecen en el grifo, lo que puede generar quejas por parte de los consumidores. Además la presencia de animales en los DWDSs está asociada con problemas de decoloración del agua. Pueden ser tanto causa como efecto. Se ha observado que los animales crecen especialmente en puntos con bajo flujo, como puntos muertos de tuberías, donde los sedimentos se acumulan. Pero también se ha descubierto, al examinar muestras de agua decolorada, que parte de la materia articulada consiste en fragmentos de animales, como caparazones vacíos, que están teñidos con hierro.

Por otra parte, la descomposición de los animales y sus heces pueden crear problemas de olores y sabor en el agua que circula por los DWDSs. Aunque, alternativamente, los animales al alimentarse de partículas de materia orgánica limitan la capacidad de crecimiento de microorganismos como los actinomicetos, que pueden generar problemas en el olor y sabor del agua. Ambas observaciones son conjeturas, no se sabe exactamente la relevancia de estos hechos ya que la biomasa de los microorganismos en los DWDSs es mucho mayor que la de los metazoos [Evins, 2004].

En países tropicales y subtropicales se conoce que algunas especies de invertebrados acuáticos actúan como huéspedes intermediarios de parásitos. En cambio, en los países de clima templado, no hay evidencia de que ningún animal encontrado en los DWDSs sea directamente dañino para los seres humanos [Evins, 2004]. Sin embargo se sabe que algunas bacterias patógenas, como la *Legionella*, pueden crecer y sobrevivir dentro de ciertas amebas (protozoos) en los DWDSs [Smith-Somerville *et al.*, 1991] y que estas existen en estos sistemas alimentándose de los biofilms (Figura 2.2).

2.2.4 Problemas operacionales

Los problemas operacionales asociados con el desarrollo de biofilms en los DWDSs son diversos. Se ha demostrado que la formación de biofilms en las conducciones de agua potable reduce la velocidad y la capacidad de circulación, lo que conduce a consumir más energía y obtener un menor rendimiento [Lopes *et al.*, 2009]. De la misma manera, la presencia de biofilms puede aumentar la disminución de presión en las tuberías de distribución [Characklis & Marshall, 1990]. Sin embargo, los problemas operacionales, causados por el desarrollo de biofilms en los DWDSs, más destacados, debido a su relevancia sanitaria y económica, son el consumo de desinfectante y la biocorrosión. Estos son explicados con detalle en los siguientes apartados.

2.2.4.1 Consumo de desinfectante

Al agua que circula por los DWDSs, normalmente, antes de entrar en el sistema de distribución, se le aplica algún tipo de desinfectante residual con la intención de mantener los niveles de calidad adquiridos en la planta de tratamiento, durante el tiempo que permanezca en el sistema. Los desinfectantes secundarios comúnmente usados para este fin son el cloro libre y las cloraminas.

Existen cuatro criterios que deben considerarse a la hora de llevar a cabo la elección del desinfectante secundario:

1. Estabilidad del residual
2. Toxicidad del residual
3. Control del biofilm
4. Formación de subproductos

En la actualidad el cloro libre es el desinfectante más comúnmente utilizado por su bajo precio, efectividad matando bacterias y su estabilidad química en el agua [Kowalska *et al.*, 2006], aunque presenta un mal comportamiento en cuanto a la formación de subproductos por su alta reactividad.

Las cloraminas, en cambio, parecen más estables, permaneciendo durante un mayor tiempo en el sistema, y siendo más efectivas penetrando en los biofilms. Sin embargo, el uso de

cloraminas como desinfectante secundario también presenta inconvenientes. La nitrificación es uno de los principales problemas que pueden darse. Además, las cloraminas presentan menor capacidad de oxidación y desinfección que el cloro libre. No oxidan ciertas sustancias frecuentes en muchas aguas como el hierro, manganesos y sulfuros. Por último, conviene destacar que también generan subproductos de la desinfección, aunque la mayoría en menor medida que el cloro libre, hay algunos como el cloruro de cianógeno que puede formarse en mayor cantidad [Ramírez, 2005].

Si se implanta el mantenimiento de un desinfectante residual se debe minimizar este residual y los subproductos formados [Ramírez, 2005] por lo que se recomienda tener en consideración los siguientes factores que pueden estar presentes en el agua que sale de la planta de tratamiento:

- Concentración de carbono orgánico asimilable (COA)
- Formación potencial de subproductos de la desinfección
- Tiempo de retención en el sistema de distribución

A pesar de que la principal característica de los desinfectantes secundarios es su capacidad de permanencia en el agua, es sabido que estos son consumidos en los sistemas de distribución. Este consumo de desinfectante se produce por reacciones que se dan tanto en la masa de agua circulante como en la pared de las tuberías de los sistemas de distribución. Depósitos, productos de corrosión, microorganismos, impurezas orgánicas, compuestos amónicos y metálicos (como iones ferrosos y manganeso), son algunos de los constituyentes del agua que reaccionan con el cloro y lo consumen. Los factores que se ha demostrado influyen en el decaimiento de cloro asociado a la pared de las tuberías son el material y el diámetro de la tubería, la concentración inicial de cloro, los depósitos y productos de corrosión y los biofilms.

Aunque el consumo de cloro por parte de los biofilms es un proceso que aún se entiende pobremente, se sabe que la presencia de biofilms en las paredes de las tuberías tiene implicación en la tasa de decaimiento de cloro debido a que el biofilm reacciona con él y lo consume. Dick de Beer *et al.* en 1994 observó que las concentraciones de cloro medidas en los biofilms eran tan sólo de un 20%, o menos, de la concentración existente en el agua circulante. Los datos obtenidos demostraban que existe difusión de cloro hacia el interior de la matriz del biofilm, y allí es consumido. Concluyó que la limitación en la penetración del

cloro en el biofilm no se debe únicamente a una difusión transitoria, sino que está causada por la neutralización del cloro en la matriz del biofilm.

También se observó que existía variabilidad en las tasa de penetración del cloro en el biofilm bajo condiciones comparables. Este hecho sugiere la existencia de diferencias locales en los biofilms con respecto a la resistencia a la eficacia del cloro. Las zonas con una alta resistencia al cloro pueden tener una mayor capacidad de reducción del cloro que las zonas que presentan una mayor rapidez en la penetración del cloro. Esto puede deberse a la existencia de una mayor densidad celular, subpoblaciones con mayor poder reductor por célula o mayor densidad o poder reductor de las sustancias poliméricas extracelulares. Seo (2009), observó que la cubierta de polímeros extracelulares de los biofilms está implicada en la interacción con el desinfectante y que puede atribuírsele parte del consumo.

El desarrollo de biofilms en las paredes de las tuberías tiene una implicación directa en el consumo de desinfectante, por lo que ha sido propuesto como un factor a tener en cuenta en la modelación del decaimiento de cloro en los DWDSs [Lu *et al.*, 1998].

2.2.4.2 Biocorrosión

La biocorrosión (corrosión influenciada por microorganismos, MIC, del inglés Microbially Influenced Corrosion) se refiere a la influencia de los microorganismos en la cinética de los procesos de corrosión de metales. La MIC está causada por los microorganismos adheridos a la interfase, es decir, los biofilms [Beech *et al.*, 2000] (Figura 2.3). El proceso de biocorrosión está asociado con los microorganismos o con los productos de su actividad metabólica, incluyendo enzimas, exopolímeros, ácidos orgánicos e inorgánicos, así como componentes volátiles como el amoniaco y el sulfuro de hidrógeno [Beech & Gaylarde, 1999].

Los microorganismos implicados en la MIC de metales como el hierro, cobre y aluminio, y sus aleaciones son, fisiológicamente, diversos. Desde un punto de vista electroquímico, la corrosión es una reacción química donde se transfieren electrones desde un metal con valencia cero a un aceptor de electrones externo, causando la liberación de iones metálicos al medio circundante [Lopes *et al.*, 2009]. La capacidad que tienen muchas bacterias de sustituir el oxígeno por otro compuesto oxidable, como aceptor final de electrones en la respiración, les

permite estar activas en un gran rango de condiciones propicias para que se dé la corrosión metálica. La habilidad de producir un amplio espectro de productos metabólicos corrosivos en un gran rango de condiciones ambientales hace que los microorganismos sean una amenaza real para la estabilidad de los metales, incluso para los que han sido diseñados para resistir la corrosión.

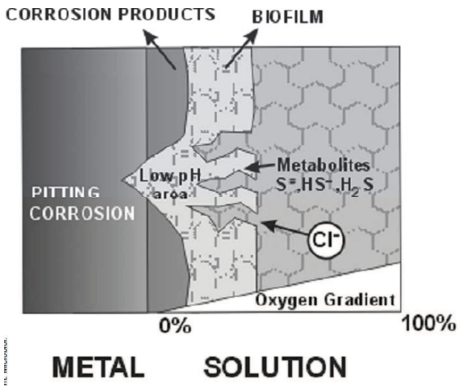

Figura 2.3. Biocorrosión [Videla, 1988]

Las condiciones básicas que conducen a la MIC están presentes en los DWDSs: las bacterias y las superficies metálicas entran en contacto con la formación de biofilms en las paredes de las tuberías [Lopes *et al.*, 2009]. El principal grupo de bacterías asociado con fallos por corrosión en estructuras metálicas, son las bacterias reductoras de azufre (SRB, del inglés Sulphate Reducing Bacteria), aunque existen muchos otros grupos capaces de llevar a cabo la MIC en los DWDSs (Figura 2.4). Todos ellos coexisten en los biofilms, a menudo formando comunidades capaces de afectar a los procesos electroquímicos a través de metabolismo cooperativo, el cual las especies individualmente no son capaces de iniciar [Beech *et al.*, 2000]. La colonización microbiana de superficies metálicas produce importantes cambios en el tipo y la concentración de iones, el pH y el potencial de oxidación-reducción, alterando el comportamiento pasivo, o activo, del sustrato metálico y sus tasas de corrosión.

El coste de la corrosión y las estrategias de prevención en Estados Unidos se calcularon en unos 276 billones de dólares al año, lo que supone el 3,1% del producto interior bruto (PIB). Otros estudios en Reino Unido, Japón, Australia y Alemania han considerado que el coste asociado a la corrosión corresponde a entre el 1 y el 5% del PIB de los respectivos países [Little & Jason, 2009]. El coste de la corrosión mediada por organismos se estima que supone el 20% del coste total asociado a la corrosión.

Figura 2.4. Tubería metálica dañada por MIC

Aunque no existen datos oficiales sobre el coste que genera la MIC en los DWDSs, se puede obtener una idea de su repercusión observando la magnitud de los costos asociados a la corrosión en los sistemas de distribución de agua (Figura 2.5).

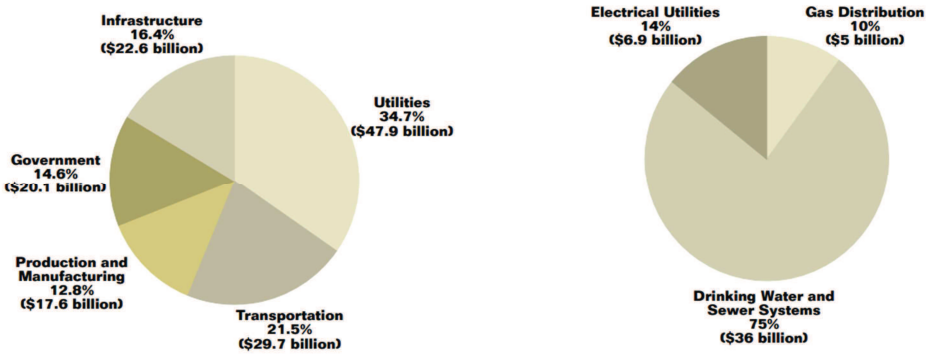

Figura 2.5. Costes asociados a la corrosión en EE.UU. [NACE]

En las empresas energéticas sí se lleva un especial seguimiento del costo que genera la MIC, debido a la magnitud de los daños que produce. En la planta de generación de energía nuclear operada por la empresa Ontario Hydro (Canadá) los tubos de refrigeración se vieron dañados por MIC y se estimó que el coste asociado al reemplazamiento de dichos tubos supuso a la corporación 300.000 $ por día y tubo [Brennenstuhl *et al.*, 1990].

Aunque no haya datos concretos del impacto económico de la MIC en los DWDSs, no es difícil hacerse a la idea de la importancia económica que debe tener en estos sistemas. Entre los costes atribuibles a este proceso se encuentran una parte de los gastos asociados a tratamientos anti-corrosión, reemplazo de tuberías y estructuras dañadas y costes relacionados con averías o fugas, entre otros.

Es importante tener en cuenta, aparte del coste económico asociado a la MIC, que el deterioro de las infraestructuras de los DWDSs es una de las principales causas de la pérdida de calidad y cantidad de agua que llega al consumidor [Imran *et al.*, 2006]. Es por ello que los fallos en los sistemas de distribución pueden causar tanto daños económicos como sociales, entre los que hay que añadir daños a propiedades adyacentes y negocios, retrasos en el tráfico y otras molestias a la población [Cromwell III *et al.*, 2002]; incluyendo la mala imagen que se lleva el consumidor cuando se encuentra con agua decolorada como consecuencia de la presencia de productos derivados de los procesos de corrosión [Lopes *et al.*, 2009].

2.3 Control del desarrollo de biofilms en los DWDSs

Un buen programa de control del desarrollo de biofilms en los DWDSs debe incorporar múltiples enfoques.

Una vez que el agua ya se encuentra en el sistema de distribución, aparte del control con sustancias químicas (desinfectantes residuales - cuya aplicación, acción y efecto sobre los biofilms en los DWDSs ya han sido explicados en los apartados anteriores) también debe llevarse a cabo un control periódico mediante lavados hidráulicos (LH) y mecanismos físicos (*pigging*).

Un programa de mantenimiento de los sistemas de distribución debe incluir LHs regulares ya que estos ayudan a redistribuir el desinfectante residual a todas las secciones del sistema y a eliminar los biofilms y sedimentos existentes [Carvajal *et al.*, 2007]. El LH está definido como la apertura de hidrantes en un área específica, durante un tiempo determinado, hasta conseguir que el agua de la salida sea de la calidad deseada. El LH no tiene un efecto duradero y su proceso debe ser repetido periódicamente.

En la actualidad, los programas de LH se establecen como medidas correctivas dando respuesta a las quejas de los usuarios después de que se han hecho instalaciones o reparaciones, y así expeler los contaminantes introducidos inadvertidamente en el sistema. Un programa de LH también puede emplearse como práctica de mantenimiento preventivo.

Algunas tuberías presentan tubérculos de corrosión que requieren el uso de técnicas mecánicas de *pigging* para eliminarlos. Estos tubérculos son cúmulos de material oxidado que consumen el desinfectante y sirven de refugio para los microorganismos, facilitando la formación de biofilms. La técnica de *pigging* consiste en utilizar los denominados *pigs*

(*pipeline inspection gauges*) para realizar diversas operaciones de mantenimiento en una tubería. Este proceso se realiza sin detener el flujo en la tubería.

Ni el LH ni el *pigging* son soluciones permanentes y es posible que no sean suficientes para controlar un biofilm bien establecido. Por ello, en algunos casos, el reemplazamiento de la tubería es la opción más sensata [Mains, 2008].

Un método preventivo del desarrollo de biofilms en los DWDSs puede ser evitar, en la medida de lo posible, la existencia de puntos muertos en los sistemas de distribución o aumentar el control sobre ellos, ya que en estos puntos el consumo de desinfectante es muy alto y la presencia de biofilms aumenta. También es recomendable llevar a cabo tratamientos anti-corrosión en los DWDSs mediante inhibidores químicos, o ajustes de pH, que pueden ser de ayuda para evitar la formación y acumulación de productos derivados de la corrosión que favorecen el desarrollo de biofilms.

Aparte de las diferentes técnicas de control y prevención, que se pueden realizar en la etapa de distribución del agua (en la que este trabajo se centra), también existen diferentes procesos que se pueden llevar en las etapas de captación y potabilización del agua, que ayudan a disminuir el posterior desarrollo de biofilms en los DWDSs. Minimizar la cantidad de nutrientes y componentes potencialmente formadores de depósitos que entran en el sistema de distribución, ayudará a prevenir la proliferación de biofilms. De hecho, la optimización del tratamiento del agua debería ser el primer paso en cualquier plan para asegurar la calidad microbiológica del agua durante la distribución [Levi, 2004]. En la Figura 2.6 se expone un resumen de las posibles medidas a tomar para el control del desarrollo de biofilms en los sistemas de abastecimiento por las agencias de encargadas del suministro de agua potable.

A pesar de las medidas anteriormente mencionadas, los biofilms hoy en día se encuentran, en mayor o menor medida, en todos los DWDSs. Conocer cómo contribuyen los diferentes factores al crecimiento de biofilms en estos sistemas y las maneras de controlar esos parámetros se presenta como la mejor prevención.

<table>
<tr><th colspan="2">POTENTIAL COMPONENTS OF A BIOFILM CONTROL PROGRAM</th></tr>
<tr><td colspan="2">Source Water Protection</td></tr>
<tr><td colspan="2">Monitoring and Maintenace of Adequate Plant Performance

• monitor individual filter effluents as well as plant effluent</td></tr>
<tr><td>Reduce Organic Carbon/Nutrients Levels

• coagulation

• precipitative softening

• activated carbon filters

• mixed carbon/sand filters

• biologically activated filters</td><td>Appropiate Disinfection Practices

• increase free chlorine residual

• use alternate disinfectant</td></tr>
<tr><td>Corrosion Control

• use chemical inhibitors

• adjust pH</td><td rowspan="2">Distribution System Maintenance

• regular flushing

• eliminating cross-conection

• modifying dead ends

• line pigging

• pipe replacement</td></tr>
<tr><td>Reservoir Maintenance

• rinse prior to use

• limit retention times

• maintain adequate residuals

• monitor sediment accumulation

• keep covered and secure</td></tr>
<tr><td colspan="2">Personnel Trainning</td></tr>
</table>

Tabla 2.1. Medidas a tomar en los sistemas de abastecimiento para el control de biofilms [Mains, 2008]

Capítulo 3

Extracción de conocimiento a través de los datos (KDD)

Con el propósito de alcanzar el objetivo principal del presente trabajo, a saber, evaluar el efecto de la interacción de las diferentes características físicas e hidráulicas de los DWDSs en el desarrollo de biofilms, se propone en este estudio la utilización del paradigma de extracción de conocimiento a través de los datos (KDD, del inglés Knowledge Discovery in Databases) como metodología que permita la obtención de modelos que expliquen el desarrollo de biofilms en los DWDSs, desde el punto de vista de su posible interacción con el resto de variables incluidas en el estudio.

El KDD es un *proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles* [Fayaad, 1996]. Los datos recogen un conjunto de hechos (una base de datos), mientras que los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto) [Molina & García, 2006]. El KDD surge de la necesidad de manejar grandes y/o complejos conjuntos de datos; de donde, además de llevar a cabo una extracción eficiente de la información, se encarga de su preparación, análisis e interpretación de los resultados obtenidos [Witten *et al.*, 2011].

El método tradicional de convertir los datos en conocimiento ha consistido y consiste todavía, en muchas situaciones, en un análisis e interpretación realizados de forma manual. Esta forma de actuar es lenta, cara y altamente subjetiva, ya que muchas decisiones importantes se realizan, no sobre la cantidad de datos disponibles, sino siguiendo la propia intuición del usuario que no dispone de las herramientas necesarias [Hernández *et al.*, 2004]. Sin embargo, el KKD explora exhaustivamente volúmenes de datos para determinar relaciones y extrae información de calidad que puede usarse para obtener conclusiones basadas en relaciones o modelos dentro de los datos mediante el uso de técnicas de minería de datos (DM, del inglés Data Mining).

No es la primera vez que en los sistemas de abastecimiento de agua se recurre a esta herramienta. El primer artículo al respecto fue escrito por Abbott *et al.* en el 2001. A partir de entonces y debido a su gran aplicabilidad, diferentes problemas de los DWDSs se han tratado de predecir y solventar mediante el uso de estas técnicas. Entre otros, se ha determinado el riesgo de roturas de tuberías [Babovic *et al.*, 2002], se ha modelado la ocurrencia de trihalometanos en los DWDSs [Millot *et al.*, 2002], se ha llevado a cabo el predimensionado de calderines antiariete [Izquierdo *et al.*, 2002] y se ha intentado predecir el tipo de daño producido en el sistema de abastecimiento [Díaz-Arévalo, 2010].

Actualmente, se almacena una gran cantidad de información, en forma de datos, en las diferentes etapas de anteproyecto, diseño, construcción y funcionamiento de un sistema de abastecimiento de agua. Esta información es potencialmente importante, sin embargo, solo se extrae una cantidad muy reducida de la información contenida en esos datos que, por tanto, apenas es aprovechada. El KDD posibilita lograr ese conocimiento, permitiendo encontrar las relaciones y patrones que ayudan al entendimiento del sistema. Es por ello que el KDD se convierte en una excelente herramienta de ayuda a la toma de decisiones en la gestión de los DWDSs.

El KKD, no es un proceso automático, sino un proceso iterativo e interactivo. Es iterativo debido a que la salida de algunas fases puede hacer volver a pasos anteriores y, a menudo, son necesarias varias iteraciones para extraer conocimiento con el que poder generar conclusiones que realmente ayuden a un posterior proceso de toma de decisiones. Es interactivo porque el usuario, o un experto en el dominio del problema, debe ayudar a la preparación y validación del conocimiento extraído.

Las propiedades deseables del conocimiento descubierto mediante el proceso KDD deben ser:

- Validez: los patrones deben también explicar datos nuevos (con un cierto grado de certidumbre), y no sólo aquellos que han sido usados para (entrenamiento y) obtención del modelo.
- Novedad: que aporte algo desconocido para el sistema y preferiblemente para el usuario.
- Utilidad: el modelo debe descubrir información que resulte útil tanto para el sistema como para (preferiblemente) el usuario.

- Comprensibilidad: la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento (al menos desde el punto de vista de su utilidad). Es por ello que se debe obtener un modelo interpretable que pueda describir y resumir la información disponible, a la vez que permita tomar conclusiones y decisiones mediante su correcta interpretación.

El KDD es un proceso complejo que se divide en diferentes etapas: selección de la información, pre-procesamiento, transformación, exploración e interpretación (Figura 1.2).

Esta metodología combina herramientas de análisis tales como la gestión de base de datos, la inteligencia artificial, la ingeniería del conocimiento, la estadística y el aprendizaje automático, entre otros [Díaz-Arevalo, 2010]. En cada etapa, según las necesidades del estudio, se pueden llevar a cabo diferentes procesos:

1) Recopilación y selección de los datos

- Comprensión del tema a tratar y planteamiento de objetivos
- Recopilación de la información de las diferentes fuentes de datos
- Selección de las variables relevantes y extracción de la información de interés
- Unificación de la información y generación de la base de datos

2) Pre-procesamiento de los datos

- Búsqueda y eliminación de datos atípicos (*outliers*)
- Reconstrucción de datos perdidos

3) Transformación

- Uso de técnicas de discretización
- Uso de técnicas de escalado

4) Aplicación de técnicas de minería de datos

- Métodos de aprendizaje no supervisado
- Métodos de aprendizaje supervisado

5) Evaluación e interpretación de los datos

- Evaluación de modelos
- Inferencia del conocimiento

Los negocios de distribución y la publicidad dirigida han sido tradicionalmente las áreas en las que más se han empleado estos métodos, ya que han permitido reducir costes o aumentar la receptividad de ofertas. Sin embargo, la integración del proceso KDD y de las técnicas de DM en diferentes actividades en el ámbito de la gestión de información tanto pública como privada se está convirtiendo en algo habitual. De hecho podemos encontrar ejemplos en todo tipo de aplicaciones: financieras, sanitarias, educativas, biológicas e industriales, entre otras [Hernández *et al.*, 2004]. En particular, en los últimos años se observa un interés creciente en la utilización de técnicas de KDD en temas medioambientales (ver, por ejemplo, Izquierdo *et al.* (2008), Gibert *et al.* (2012) y la serie de workshops DMTES, Data Mining as a Tool for Environmental Scientists, en las ediciones del iEMSs, international congress on Environmental Modelling and Software, desde el 2006) y, en particular, como se ha destacado más arriba, en el campo de la Hidráulica Urbana.

La perspectiva de la utilidad de aplicaciones futuras del KDD en Hidráulica Urbana es de largo alcance. El aumento de la complejidad y modernidad de las ciudades hace plantear una evolución de sus modelos de gestión, originando el concepto de *Smart City*. El KDD se presenta como la herramienta para llevar a cabo una gestión automática y eficiente del sistema de abastecimiento de agua en una ciudad inteligente. Gracias al mismo se podrá obtener el mayor partido a los datos e incluso realizar actividades de previsión de comportamientos (consumo de agua) y situaciones (fugas y eventos de contaminación) que ayuden a la toma de decisiones. Además, permite presentar la información agregada de diferente manera y a diferentes niveles según el objetivo de los análisis: sectorización, planes de rehabilitación, evaluación de vulnerabilidades o ubicación de sensores en la red, entre otros. En unos años, el concepto de *Smart City* se extenderá, convirtiéndose en una herramienta fundamental en el desarrollo futuro de la sociedad y sus servicios [Herrera *et al.*, 2012a].

3.1 Recopilación y selección de los datos

La primera etapa del proceso de extracción de conocimiento involucra la identificación de los objetivos del proceso y una planificación avanzada sobre qué tipo y nivel de información se piensa capturar.

Una vez que se han definido los objetivos del proceso, se deben seleccionar los datos y la información que se ha de utilizar. Los datos son hechos que describen un suceso o unas entidades y son la fuente principal de trabajo del proceso KDD. La importancia de los datos está en su capacidad de asociarse dentro de su propio contexto para convertirse en información. La información reduce nuestra incertidumbre (sobre algún aspecto de la realidad) y, por tanto, nos permite tomar mejores decisiones [Izquierdo *et al.*, 2008].

Para poder analizar los datos con garantía es necesario que exista una cierta estructura y coherencia entre los mismos. Generalmente, la información que se quiere investigar se dispone en bases de datos y otras fuentes diversas. Surge aquí la necesidad de conjugar los distintos ficheros y bases de datos de manera que se les pueda utilizar para extraer conclusiones adecuadas en el curso de su análisis y estudio, partiendo de posibles vistas unificadas de la(s) base(s) de datos de partida.

Solucionados los inconvenientes de heterogeneidad de las fuentes, surgen otros problemas relacionados a la estandarización de los datos:

- Diferentes tipos de datos representando el mismo concepto
- Diferentes claves para representar el mismo elemento
- Diferentes niveles de precisión al representar un dato

Las bases de datos se dividen en una parte destinada a entrenar los modelos de DM, otra a validar dichos modelos y una última que evalúe las predicciones a las que se ha sido capaz de llegar. A la hora de su generación se recomienda que los datos presenten las siguientes características:

- Representatividad de los datos: la base de datos de entrenamiento debe ser representativa. Debe tener en cuenta el posible rango de valores completo que presentan los datos.

- Presencia de datos límites: se deben definir adecuadamente los límites de los subconjuntos de datos que se pueden encontrar en la base de datos. Uno o más casos deben incluirse si se necesita identificar las diferencias reales entre 2 clases diferentes.
- Datos sin información limitada: En este caso el sistema de DM no tiene ninguna manera de distinguir entre dos tipos de registros. Esto ocurre normalmente cuando dos datos con el mismo valor para atributos de condición tienen diferente valor de clasificación. Esto ofrecerá problemas de redundancia de la información disponible en la base de datos.

Generar una estructura adecuada de los datos no es una tarea sencilla y es por eso que la calidad de los resultados está directamente relacionada con la correcta comprensión de los datos almacenados. La capacidad de extraer conocimiento válido y útil a partir de la información original viene determinada en gran medida por las primeras fases del KDD [Hernández *et al.*, 2004]. La mayoría de los trabajos en KDD centran su atención en el paso de DM, sin embargo, en la practica, el resto de pasos son de gran importancia para obtener una aplicación satisfactoria del KDD [Fayyad *et al*., 1996].

3.2 Pre-procesamiento y transformación de los datos

Después del proceso de recopilación y selección de los datos, el siguiente paso en el proceso KDD es preparar el conjunto de datos que se va a estudiar. Este paso es necesario ya que los buenos resultados se obtienen con datos íntegros, limpios y consistentes, sin embargo, los datos que se suelen obtener no cumplen estas características. Además, generalmente, debido a las características propias de las técnicas de DM, se debe realizar una transformación de los datos para obtener una "materia prima" que sea adecuada para el propósito concreto y las técnicas que se quieren emplear. Con ese fin se recurre a técnicas de pre-procesamiento [Hernández *et al.*, 2004].

En la mayoría de las bases de datos existe mucha información que es incorrecta respecto al dominio de la realidad que se desea cubrir y un número menor, pero a veces también relevante, de datos inconsistentes. Estos problemas se acentúan cuando realizamos la integración de distintas fuentes de información. No obstante, mientras los datos erróneos crecen de manera lineal respecto al tamaño de los datos recopilados, los datos inconsistentes se multiplican. La integración también produce una disparidad de formatos, nombres, rangos,

que podría no existir, o hacerlo en menor medida, en las fuentes originales. Esto dificulta los procesos de análisis y extracción de conocimiento.

Entre los problemas que afectan a la calidad de los datos se encuentra la presencia de valores que no se ajustan al comportamiento general de los datos, datos atípicos. Estos datos pueden representar errores en los datos o pueden ser valores correctos que son, simplemente, diferentes a los demás. Algunos algoritmos de DM ignoran estos datos, otros los descartan considerándolos ruido o excepciones, pero otros son muy sensibles y el resultado se ve claramente perjudicado por ello [Hernández *et al.*, 2004].

La presencia de datos faltantes o perdidos puede ser también un problema pernicioso que puede conducir a resultados poco precisos. No obstante, es necesario reflexionar primero sobre el significado de los valores faltantes antes de tomar ninguna decisión sobre como tratarlos ya que estos pueden deberse a causas muy diversas, como a un mal funcionamiento del dispositivo que hizo la lectura del valor, a cambios efectuados en los procedimientos usados durante la colección de los datos o al hecho de que los datos se recopilen desde fuentes diversas [Hernández *et al.*, 2004].

Estos dos problemas son dos claros ejemplos de la necesidad del pre-procesamiento de los datos, es decir, de la mejora de su calidad. Esta etapa consume una parte sustancial del tiempo del proyecto (Figura 3.1). A menudo, alrededor del 70% del tiempo de procesamiento de datos reales mediante técnicas de DM se consume en el pre-procesamiento de los datos. Las operaciones que se llevan a cabo durante el pre-procesamiento pueden reducirse en dos grandes familias de técnicas: Técnicas de Detección (TD) para detectar imperfecciones en el conjunto de datos y Técnicas de Transformación (TT) orientadas a obtener datos más manejables. Las TD incluyen detección de datos atípicos, datos faltantes, y observaciones influyentes y evaluación de la normalidad, linealidad e independencia. Por otro lado, las TT incluyen tratamiento de los datos atípicos, reconstrucción de los datos faltantes, técnicas de reducción de las escalabilidad o de proyección de los datos, obtención de nuevas técnicas de atributos, filtración y remuestreo.

La transformación de los datos engloba, en realidad, cualquier proceso que modifique la forma de los datos. Prácticamente todos los procesos de preparación de datos entrañan algún tipo de transformación. Entre las técnicas de transformación propiamente dichas se encuentran las técnicas de aumento o reducción de la dimensionalidad de los datos, técnicas de discretización y numeración y técnicas de normalización de rango.

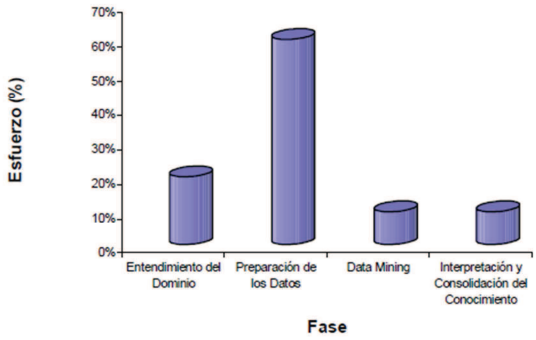

Figura 3.1. Esfuerzo requerido por cada fase del proceso KDD [Molina & García, 2006]

Conjuntamente, la preparación de datos tiene como objetivo la eliminación del mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba), y trata de presentar los datos de la manera más apropiada para la DM.

Una vez que los datos están recopilados, integrados y limpios, es necesario, además, realizar un reconocimiento, o análisis exploratorio de los datos, antes de pasar a aplicar las técnicas de DM. Este análisis exploratorio se hace con el objetivo de conocer mejor los datos de cara a la siguiente fase de análisis de los mismos. Este proceso cubre un conjunto de técnicas diversas: algunas técnicas simples del análisis exploratorio de datos, técnicas de visualización previa, de agrupamiento exploratorio y técnicas de selección entre otras [Hernández *et al.*, 2004].

3.3 Aplicación de técnicas de minería de datos

Minería de datos es un término genérico que engloba resultados de investigación, técnicas y herramientas usadas para extraer información útil y patrones de interés de grandes bases de datos [Molina & García, 2006].

La DM difiere de la estadística tradicional en cuanto a que en la estadística una hipótesis es formulada, de manera previa al conocimiento de los datos, para su posterior validación haciendo uso de los mismos; mientras que en DM se extraen patrones y modelos, automáticamente, de la información dispuesta en las bases de datos.

Las técnicas estadísticas proporcionan un análisis descriptivo, inferencial y multivariante de los datos, mientras que la DM profundiza más, buscando patrones ocultos que se escapan a técnicas tradicionales. El análisis que se realiza con la DM es exploratorio, no confirmatorio, se trata de descubrir conocimiento nuevo, no de confirmar o desmentir hipótesis. En el caso de la DM se buscan relaciones entre los diferentes *inputs* o características de la base de datos,

no es necesario sospechar la existencia de relación entre dos variables para encontrarla [Hernández *et al.*, 2004]. Con la DM es el sistema y no el usuario el que encuentra la hipótesis, además de comprobar su validez.

Las técnicas de DM son diferentes unas a otras en términos de la representación del problema, parámetros a optimizar, exactitud, complejidad, tiempo de ejecución, transparencia e interpretación.

Las técnicas de DM se pueden dividir en técnicas de aprendizaje supervisado y aprendizaje no supervisado. Estas técnicas se diferencian principalmente en que las técnicas no supervisadas no predicen un valor objetivo, sino que se centran en la estructura intrínseca, las relaciones, y la interconexión de los datos. En el aprendizaje no supervisado no existe conocimiento *a priori* y el conjunto de datos de objetos de entrada es tratado sin contrastarlo con ningún tipo de *output* esperado.

En el aprendizaje supervisado, en cambio, existe conocimiento *a priori* y se deduce una función a partir de datos de entrenamiento. En un escenario típico se tiene una medición de los resultados (por ejemplo, sobre biofilm), cuantitativa, por lo general, o categórica, que se quiere predecir en base a un conjunto de características (características fiscas e hidráulicas de los DWDSs). Se tiene un grupo de datos de entrenamiento, en los que se observa las mediciones de los resultados y las características para un conjunto de objetos (tuberías). Con estos datos se construye un modelo predictivo o clasificatorio mediante técnicas de DM que permita predecir el resultado para nuevos objetos [Hastie *et al.*, 2009].

- Técnicas de aprendizaje no supervisado

  Estas técnicas se asocian con el concepto de técnicas descriptivas. La descripción puede ayudar a la comprensión del comportamiento y es un paso previo fundamental para un posterior análisis en profundidad de los datos. Si definimos E como el conjunto de todos los posibles elementos de entrada. Las instancias posibles, dentro de E, generalmente se representan como un conjunto de valores para una serie de atributos, A_i; $i = 1, \ldots, n$ (sean nominales o numéricos), que forman una partición del espacio de entrada del modelo. Es decir, $E = A_1 \times A_2 \times \ldots \times A_n$. En el caso de las técnicas de aprendizaje no supervisado definimos un *ejemplo* o realización muestral del espacio de entrada del modelo no supervisado como un conjunto $\delta = \{<a_1, \ldots, a_n>: a_i \in A_i\}$, sin etiquetar mediante ningún tipo de información de salida (Tabla 3.1).

Nº Caso	A_1	A_2	...	A_n
1	a_{12}	a_{21}	...	a_{n2}
2	a_{11}	a_{23}	...	a_{n1}
3	a_{13}	a_{21}	...	a_{n3}
…	...	...	...	...
N	a_{11}	a_{22}	...	a_{n2}

Tabla 3.1. Forma general de la base de datos para técnicas no supervisadas

A continuación se presentan las técnicas no supervisadas utilizadas en el presente trabajo.

- Agrupamiento (*clustering*): el objetivo de esta tarea es obtener grupos o conjuntos entre los elementos de δ, de modo que los que están dentro de cada grupo están más estrechamente relacionados entre sí que los objetos asignados a grupos diferentes. Un objeto de la base de datos puede ser descrito por una serie de mediciones, o por su relación con otros objetos. Por lo tanto, los elementos dentro de un grupo son más "similares" entre ellos que a los objetos de otro grupo [Herrera, 2011a]. La noción de similitud, en términos de proximidad, entre los objetos que se agrupan (en caso contrario, la disimilitud se utiliza para explicar la diferencia) es fundamental para el *clustering*. El *clustering* trata de agrupar los objetos en base a la definición de similitud que se le suministra [Gibert & Pérez-Bonilla, 2005]. Lo importante del *clustering* respecto a la clasificación es que son precisamente los grupos y la pertenencia a los grupos lo que se quiere determinar y, *a priori*, no se sabe ni cómo son los grupos ni cuántos hay. En algunos casos se puede proporcionar el número de grupos que se desea obtener. Otras veces, este número se determina por el algoritmo de agrupamiento, según las características de los datos. La función a obtener es idéntica a la de clasificación, λ: E $\rightarrow$ S, con la diferencia que los valores de S (etiquetas) y sus miembros se “crean o inventan”, durante el proceso de aprendizaje [Hernández *et al.*, 2004].

- Reglas de asociación: esta técnica se utiliza con atributos nominales con el objetivo de ver la relevancia de los atributos, detectar redundancias o dependencias entre atributos, o seleccionar un subconjunto. Dados los ejemplos del conjunto *E*, una regla de asociación se define, generalmente, de la

siguiente forma: "si $A_i = a \wedge A_j = b \wedge \ldots \wedge A_k = h$ entonces $A_r = u \wedge A_s = v \wedge \ldots \wedge A_z = w$'', donde todos los atributos son nominales y las igualdades se definen utilizando algún valor de los posibles para cada atributo [Sierra, 2006].

- Técnicas de aprendizaje supervisado

En este caso, se define el *ejemplo* de E o realización muestral del espacio de entrada del modelo supervisado, como un conjunto etiquetado, δ, mediante algún tipo de información de salida; es decir, $\delta = \{<(a_i, s_i)> : a_i \in A_i, s_i \in S; i = 1, \ldots, n\}$, donde S es el conjunto de valores de salida.

En la Tabla 3.2 se puede ver un ejemplo de una base de datos de este tipo [Hernández *et al.*, 2004; Sierra, 2006].

N^o Caso	A_1	A_2	...	A_n	S
1	a_{12}	a_{21}	...	a_{n2}	s_2
2	a_{11}	a_{23}	...	a_{n1}	s_M
3	a_{13}	a_{21}	...	a_{n3}	s_1
...	...	...	...	...	...
N	a_{11}	a_{22}	...	a_{n2}	s_2

Tabla 3.2. Forma general de la base de datos para técnicas supervisadas

Estas técnicas supervisadas se asocian con el concepto de técnicas de clasificación. En el presente trabajo las técnicas utilizadas han sido las siguientes.

  - Reglas de decisión: se realizan mediante un proceso inductivo en el cual se lleva a cabo la generación de un conjunto de reglas de decisión con el propósito de obtener hipótesis que traten de explicar un determinado sistema. El sistema viene determinado por un conjunto de ejemplares y las reglas de decisión van a representar conceptos que describen ese sistema. Para un problema de clasificación *n*-ario, i.e. la etiqueta de clase de un ejemplar se va a encontrar en un conjunto de valores $\{s_1, s_2, \ldots, s_n\}$, el conjunto de reglas creado tendrá la forma representada en la Tabla 3.3, donde R_i es la regla *i*-ésima, $antecedente_i$, es el conjunto de tests sobre los atributos de entrada que se realiza para determinar si hay emparejamiento del ejemplar para su clasificación y cada uno de los s_j se corresponde con la etiqueta de clase *j*-ésima [ver, entre otras muchas referencias, Sierra, 2006].

R_1	$antecedente_1$	entonces	s_1
R_2	$antecedente_2$	entonces	s_1
	...		
R_k	$antecedente_k$	entonces	s_1
R_{k+1}	$antecedente_{k+1}$	entonces	s_2
	...		
R_m	$antecedente_1$	entonces	s_{n-1}
		demás	$s_{n,}$

Tabla 3.3. Conjunto de reglas de decisión [Sierra, 2006]

- Árboles de clasificación: el funcionamiento de los árboles de clasificación consiste en dividir el espacio de clasificación en zonas, de manera que a los patrones que pertenecen a cada zona se les asigna una de las posibles clases. La definición más cercana al paradigma de los árboles de clasificación sería que un clasificador es una partición del espacio de clasificación A en M subconjuntos disjuntos $X_1, X_2, \ldots, X_M$, siendo A la unión de todos ellos y para todo a perteneciente a X_m la clase predicha es S_m. En la Figura 3.2 se muestra un árbol de clasificación simple [ver, entre otros, Sierra, 2006].

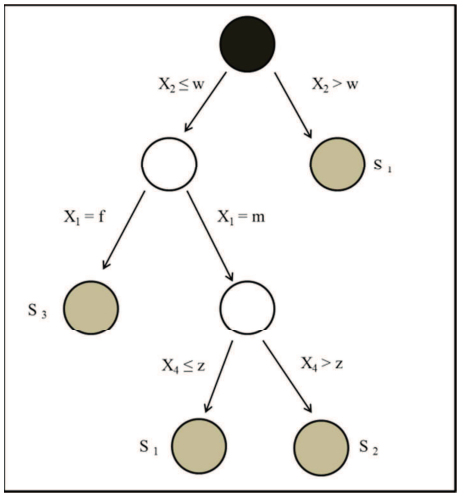

Figura 3.2. Ejemplo de un árbol de clasificación sencillo [Sierra, 2006]

En todos los casos es necesario establecer una o varias medidas de interés que consideren la validez, utilidad y simplicidad de los patrones obtenidos mediante las técnicas de DM. De esta manera se evalúa la fiabilidad de la interpretación de los resultados obtenidos y se alcanza el objetivo final del proceso KDD. Este objetivo consiste en consolidar el conocimiento obtenido mediante su incorporación en algún sistema real, toma de decisiones a partir de los

resultados alcanzados o, simplemente, registro de la información conseguida y suministro a quien esté interesado. Idealmente, los patrones descubiertos deben tener tres cualidades: ser precisos, comprensibles e interesantes.

Capítulo 4

Recopilación y selección de los datos

La primera etapa del proceso de extracción de conocimiento involucra la identificación de los objetivos del estudio y el tipo y nivel de información que se piensa emplear. Esta fase del proceso del KDD comienza con la comprensión del tema a tratar, en este caso, el desarrollo de biofilms en los DWDSs. Mediante la recopilación de información y antecedentes a través de la bibliografía se llega a comprender qué factores se encuentran implicados en el desarrollo de biofilms en estos sistemas. Una vez alcanzado ese conocimiento se plantean los objetivos y se realiza una planificación avanzada sobre qué tipo y nivel de información manejar. En el presente estudio se propone como objetivo estudiar el efecto de la interacción de las diferentes características físicas e hidráulicas de los DWDSs en el desarrollo de biofilms, por lo que se recopila la información y se seleccionan las variables relevantes para este fin. Finalmente, se unifica la información y se genera la base de datos.

4.1. Comprensión del tema a tratar y planteamiento de objetivos

Los biofilms se comportan como sistemas metaestables. Por un lado son "alimentados" por la entrada de bacterias con el flujo de agua (la velocidad de fijación/deposición de las células está correlacionada con la densidad celular en el flujo de agua) y por la multiplicación de las bacterias. Por otro lado, sin embargo, sufren el desprendimiento de bacterias del biofilm hacia el flujo de agua [le Puil, 2004].

La supervivencia y el rebrote de microorganismos en los DWDSs están afectados no solo por factores biológicos sino también por la interacción de diferentes factores (Figura 4.1) [Yu, *et al.,* 2010]. Actualmente estos factores están bien estudiados a escala de laboratorio aunque, raramente se estudian en campo (en un sistema de distribución real).

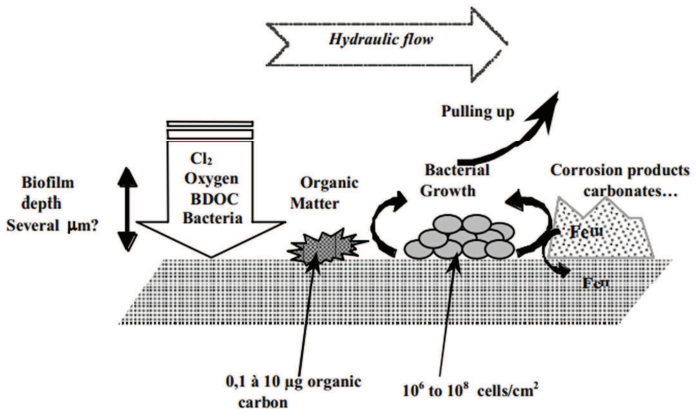

Figura 4.1. Varios mecanismos participando en la acumulación de biofilms sobre una superficie en contacto con agua potable [le Puil, 2004]

Dentro de los DWDSs encontramos que los factores más destacados que influyen en desarrollo de los biofilms son los siguientes:

- Naturaleza de los materiales utilizados en los DWDSs

  Todos los materiales son colonizados por microorganismos, pero el tipo de material juega un papel importante. De hecho, los materiales determinan la eficiencia de adsorción de las bacterias *pioneras* y puede ser una fuente de nutrientes o factores de crecimiento. Análisis de biofilms adsorbidos en cristal o polietileno han mostrado diferentes proporciones de detección de bacterias en función del material (4 y 26% respectivamente) [Kalmbach *et al.*, 2000].

  Las tuberías de los DWDSs pueden ser clasificadas en tres grandes grupos de materiales: metálicas, poliméricas o de cemento (completas o como recubrimiento interno) [Imran *et al.*, 2006]. Las tuberías de hierro o aleaciones de hierro tienden a desarrollar más biofilm que las tuberías de cemento y estás, a su vez, más que las tuberías de plástico [Thabisile, 2010]. Este hecho se explica, principalmente, por la rugosidad de los materiales, ya que las superficies rugosas protegen al biofilm del desprendimiento y le proporcionan mayor área de protección y colonización [Chowdury, 2011].

  En el caso de los materiales susceptibles a sufrir corrosión, ha sido demostrado que la presencia de productos de corrosión favorecen la actividad y la producción de biomasa en los biofilms. La especial proliferación de biofilms en estos puntos puede deberse a que el

cloro es consumido por la oxidación del hierro ferroso (Fe^{2+}) producido por la corrosión del hierro férrico (Fe^{3+}) [Thabisile, 2010]. También influye el hecho de que las zonas corroídas proporcionan un ambiente propicio para albergar el crecimiento de comunidades de microorganismos, como bacterias oxidantes del hierro, que de otra manera no se desarrollarían bien en condiciones normales en los sistemas de distribución de agua. De esta manera, aumenta la diversidad y el número de microrganismos que forman los biofilms.

- Carbono orgánico

La presencia de carbono orgánico es esencial para el desarrollo de bacterias heterotróficas. El carbono orgánico se divide principalmente en carbono orgánico disuelto biodegradable (CODB) y carbono orgánico asimilable (COA). El CODB es la fracción biodegradable del carbono orgánico y representa la actividad metabólica bacteriana [Chowdhury, 2011]. La mayoría de estudios muestran una correlación positiva entre la concentración de CODB y el crecimiento bacteriano en los DWDSs [Lu *et al.*, 1999; Ollos *et al.*, 2003; Ndiongue *et al.*, 2005]. El COA representa el potencial de crecimiento bacteriano, ya que mide solo el carbono que ha sido convertido en biomasa. El crecimiento bacteriano se estimula con el aumento de COA [Chowdhury, 2011]. En la mayoría de los casos se usa el acetato como referencia del COA [Chandy & Angles, 2001] ya que su totalidad (o casi) puede ser convertida en biomasa [Lu & Chu, 2005].

La mayor o menor presencia de carbono orgánico es la responsable de que aguas subterráneas (que tienden a tener bajos contenidos de materia orgánica) sean menos propensas al desarrollo de biofilms que aguas superficiales (con mayor carga de materia orgánica).

- Nutrientes

El crecimiento bacteriano en los DWDSs está potenciado por la presencia de determinados compuestos (nutrientes), los cuales están presentes en el agua tratada o son producidos por los materiales (tuberías) que están en contacto con el agua potable [Van der Kooij 1998]. También se encuentran en las células de las bacterias muertas y en los productos derivados de la desinfección. Incluso, algunos aceites lubricantes, usados en las bombas de agua,

pueden verter cantidades sustanciales de nutrientes al agua [White & LeChevallier, 1993]. Los dos principales nutrientes son el nitrógeno y el fósforo.

El nitrógeno en los DWDSs está presente principalmente en forma de nitratos, en condiciones aerobias, o aminas/amonio, en condiciones anaerobias. Los nitratos en el agua potable normalmente se encuentran en un rango entre 0-10 mg/l, con una media de 1,4 mg/l [Zaldivar & Wetterstrand, 1978], cantidad suficiente para el desarrollo de biofilms en los DWDSs [Donlan, 2002]. El nitrógeno**,** normalmente, no es un factor de crecimiento limitante de los biofilms en los DWDSs [Qin, 2009].

La concentración de fosfatos en los DWDSs está normalmente por debajo de 1mg/l [Lu & Chu, 2005; Lehtola *et al.*, 2006]. Sin embargo, los fosfatos se suelen añadir al agua de los DWDSs como inhibidores de los procesos de corrosión [Rompré, 1999; Butterfieldl *et al.*, 2002]. El aumento de la concentración de fósforo promueve el crecimiento de biofilms y también cambia su composición [Keinänen *et al.*, 2002]. El hierro y el acero usados en los DWDSs, normalmente, contienen entre un 0.03 y un 0.2% de fósforo en su forma reducida, que puede ser liberado al flujo de agua debido a la corrosión y convertirse en una fuente de fósforo que sustente el crecimiento bacteriano [Morton & Edwards, 2005].

- Régimen hidráulico

Simoes *et al.* (2006) observó un mayor número de bacterias en los biofilms formados en régimen turbulento (Re = 11000), que en los biofilms formados bajo régimen laminar (Re = 2000). También se ha observado que los biofilms de monocultivo de *P. fluorescens,* en régimen turbulento, tienden a ser más activos, tener más masa por cm^2, mayor densidad celular y distinta morfología que los biofilms en régimen laminar [Simoes *et al.*, 2007].

En general, el flujo turbulento favorece el desarrollo de biofilms respecto al régimen laminar debido a que, además de proporcionar mejores condiciones de absorción de nutrientes, mejora la nucleación química, favoreciendo la precipitación de compuestos que contribuyen al desarrollo de los biofilms.

- Temperatura

La temperatura juega un papel importante en el crecimiento de los biofilms. Las temperaturas altas favorecen la tasa de crecimiento de las bacterias frente a las temperaturas bajas, siempre que estas se encuentren dentro del rango de tolerancia de las bacterias estudiadas (Figura 4.2).

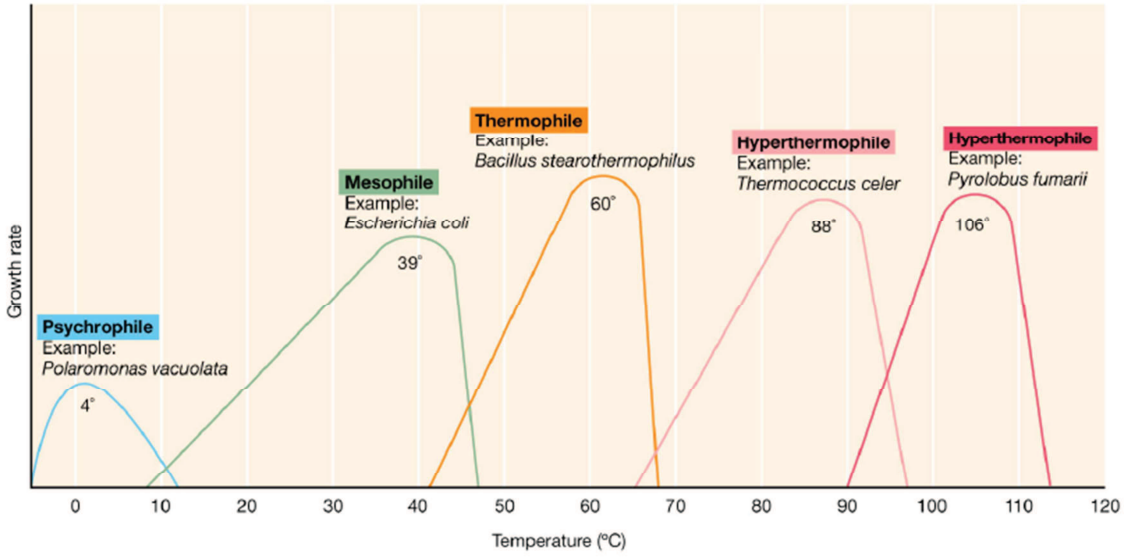

Figura 4.2. Crecimiento de microorganismos según la temperatura

Una investigación llevada a cabo en el sistema de distribución de agua de Shanghai determinó que la temperatura era uno de los factores más influyentes tanto en la abundancia como en la composición del biofilm [Bai *et al.*, 2010]. Silhan *et al.*, 2006, demostró que el número de bacterias heterotróficas en los biofilms era mayor a 35ºC que a 15ºC. También se han observado variaciones estacionales en la diversidad de especies de poblaciones de bacterias heterotróficas presentes en los DWDSs. LeChevallier *et al.* (1980), encontró que la diversidad de bacterias en los DWDSs, presenta varias tendencias apreciables en su distribución anual por su capacidad para crecer y sobrevivir en un amplio rango de temperaturas. Por ejemplo, la diversidad de especies fue mayor en el periodo estival que en los meses de invierno. También se ha encontrado que, cambios de temperatura a corto plazo (del orden del tiempo de residencia) no producen cambios significativos en los biofilms en sistemas cloraminados.

El aumento de la temperatura también contribuye a un aumento de la eficacia de los desinfectantes [Ramírez, 2005], por lo tanto, al aumento del estrés sobre los biofilms. Sin

embargo, según los resultados obtenidos en trabajos como el anteriormente comentado [Bai *et al.*, 2010], este aumento de la eficacia de los desinfectantes con la temperatura no parece limitar el desarrollo de biofilms en los DWDSs. En cambio, el aumento de las tasas de crecimiento de los microorganismos con el incremento de la temperatura sí que parece tener gran relevancia en el desarrollo de estas comunidades.

- Velocidad del flujo

  La formación de biofilms aumenta con el incremento de la velocidad del flujo, probablemente debido al aumento de la transferencia de masa de nutrientes desde el agua al biofilm [Lehtola *et al.*, 2006]. También se ha observado que las velocidades altas de flujo forman biofilms más compactos de larga duración [Tsvetanova, 2006]. Cloete *et al.* (2003) demostró que a partir de una velocidad específica, entre 3-4 m/s, el biofilm tiende a desprenderse. Mayores cantidades de bacterias han sido encontradas en el flujo de agua a mayores velocidades de flujo en sistemas con tuberías de diferentes materiales, probablemente debido al mayor desprendimiento de biofilms [Qin, 2009].

  Fluctuaciones frecuentes de la velocidad de flujo en los sistemas de distribución afectan al flujo de nutrientes y de desinfectante residual, así como al desprendimiento de los biofilms [Telghmann *et al.*, 2004]. Del mismo modo, cambios repentinos en la velocidad del flujo también pueden resultar en desprendimientos de biofilm [Lehtola *et al.*, 2006]. Se ha observado que los niveles de bacterias planctónicas en un sistema de tuberías experimental aumentan en unas 10 veces cuando el flujo es parado y reanudado.

- Depósitos

  Los productos de corrosión y las partículas, que se depositan en la red de distribución, forman sedimentos y depósitos sueltos que pueden tener un efecto significativo en el desarrollo del biofilm [Chowdury, 2011]. Estos depósitos pueden proporcionar refugio a los microorganismos frente al desinfectante y consumir cloro residual. Los microrganismos son protegidos de los desinfectantes oxidantes si las partículas contienen compuestos reductores, tales como óxidos de hierro o materia orgánica [Muñoz *et al.*, 2007]. Los sedimentos, los depósitos sueltos y los productos de corrosión también proporcionan más superficie para el crecimiento de los biofilms y reducen la fuerza de arrastre del agua sobre

estos. Los productos derivados de los procesos de corrosión también pueden ser usados como nutrientes por el biofilm, favoreciendo de este modo su crecimiento [Chowdury, 2011].

- Tiempo de retención hidráulica

  Los DWDSs suelen estar sobredimensionados debido a que se diseñan para poder dar una buena respuesta en situaciones de protección contra incendios. Por lo que, las tuberías y los tanques de almacenamiento suelen ser mayores de lo necesario para su propósito de abastecimiento de agua para consumo humano, dando lugar a velocidades menores y mayores tiempos de retención hidráulica (HRT, del inglés Hydraulic Retention Time) en condiciones de funcionamiento normal [Crozes & Cushing, 2000]. Sin embargo, cuanto mayor tiempo esté el agua retenida en el sistema de distribución, mayor será el decaimiento de desinfectante residual que sufra, produciéndose, a su vez, un aumento de la temperatura y la deposición de sedimentos [EPA, 2002]. Todos estos aspectos favorecen el desarrollo de biofilms. La reducción de la concentración de desinfectante reduce la presión sobre el biofilm, la temperatura favorece el crecimiento bacteriano y los sedimentos reducen la exposición del biofilm al desinfectante y a la fuerza de arrastre del flujo de agua. Normalmente el agua tiende a estar mayor tiempo retenida en los puntos muertos y en otras áreas de los sistemas de distribución donde el agua tiende a estancarse.

- pH

  Se ha encontrado que en reactores biológicos, a valores de pH por encima de 8.6, la actividad metabólica del biofilm se reduce [Lee & Rittmann, 2003].

  En conjunto, la eficacia de los desinfectantes disminuye al aumentar el pH, al hacerse el medio más básico. El poder bactericida del cloro libre es mayor a pH = 7 (neutro) o menor, por lo que valores superiores a este reducirán la efectividad del cloro como desinfectante y favorecerán el desarrollo de biofilm [Ramírez, 2005].

- Desinfectante residual

  En el desarrollo de biofilms en los DWDSs influye tanto la cantidad de desinfectante residual, como el tipo de desinfectante utilizado.

  Respecto a las cantidades de desinfectante residual permitidas y el consumo del mismo por diferentes factores a lo largo del sistema de distribución ya se ha hablado anteriormente en este trabajo (Capítulo 2). El consumo de desinfectante en el sistema reduce el estrés sobre los biofilms según nos alejamos del punto de cloración, viéndose favorecido su desarrollo según se avanza en el sistema.

  Si nos fijamos en el tipo de desinfectante residual los más utilizados son el cloro gas y las cloraminas. En relación al desarrollo de biofilms en los DWDSs las cloraminas son más efectivas que el cloro limitando su crecimiento. Esto es debido a que las cloraminas son menos reactivas que el cloro y, por lo tanto, tienen una mayor capacidad de penetración en los biofilms [Ramírez, 2005].

El avance, en las últimas décadas, de las técnicas microscópicas y el desarrollo de diferentes sistemas que permiten estudiar el desarrollo de biofilms han permitido investigar el amplio espectro de factores que afectan al crecimiento de biofilms en los DWDSs. Sin embargo, la mayoría de los estudios solo evalúan uno o dos factores a la vez [Tsai, 2005; Tsvetanova, 2006; Epa 2002; Zhou *et al.*, 2009; Silhan *et al.*, 2006]; excepto notables excepciones [Lehtola *et al.*, 2006; Simoes *et al.*, 2006], escasos intentos se han llevado a cabo para estudiar las interrelaciones y comparar la importancia relativa de los diferentes factores en el desarrollo de biofilms en los DWDSs. Además, la complejidad del microambiente en estudio e, incluso, el uso de diferentes metodologías y sistemas de crecimiento de biofilms llevan, en algunos casos, a resultados ambiguos o no fácilmente comparables [Simoes *et al.*, 2006].

El presente trabajo se propone como objetivo cubrir esta carencia, estudiando el efecto de la interacción de un conjunto de características de los DWDSs, relevantes en el desarrollo de biofilms.

Tras estudiar la información relativa a los diferentes factores que se conoce influyen en el desarrollo de biofilms en los DWDSs, se observa que estos factores pueden clasificarse en características físico-químicas del agua que circula por los DWDSs y características físicas e

hidráulicas de los propios sistemas de distribución. Con el fin de lograr identificar las condiciones que determinan el mayor o menor desarrollo de biofilms, en el interior de las tuberías de un DWDS, se propone centrar el trabajo en el estudio del resultado de la interacción de las diferentes características físicas e hidráulicas de los DWDSs.

Como objetivo final del presente estudio se pretende determinar qué tuberías serán más o menos propensas al desarrollo de biofilms en su interior, en función de las características físicas e hidráulicas de las mismas. A través de este conocimiento, se quieren desarrollar las bases de la implementación de las herramientas necesarias para localizar las áreas de los sistemas de distribución de agua potable que presentan un mayor riesgo de desarrollo de biofilms.

4.2. Generación de la base de datos

Una vez establecido el objetivo del trabajo se procede a recopilar la información de las diferentes fuentes de datos, seleccionar las variables relevantes y extraer la información de interés, para finalmente, unificar la información y generar la base de datos.

4.2.1. Recopilación de la información de las diferentes fuentes de datos

Se recopilan datos de mediciones de crecimiento de biofilms de diferentes fuentes bibliográficas que estudian el desarrollo de biofilms en los DWDSs (Tabla 4.1).

En este caso las fuentes de información han sido artículos científicos que estudiaban el desarrollo de estas comunidades en los DWDSs, ya sea en sistemas reales o simulados en laboratorio. Aparte de las mediciones de biofilms también se recopiló la información asociada relativa a las condiciones de estudio. El número total de casos recopilados fue de 303.

AUTOR	Nº MUESTRAS	VARIABLES MEDIDAS
Muñoz *et al.* (2007)	16	régimen hidráulico, velocidad, distancia al punto de cloración, material, edad_tubería, biofilm
Manuel *et al.* (2010)	4	nº reynolds, velocidad, HRT, temperatura, material, edad_tubería, biofilm
Lehtola *et al.* (2006)	4	nº reynolds, velocidad, temperatura, material, edad_tubería, biofilm
Chu *et al.* (2005)	1	velocidad, HRT, temperatura, material, edad_tubería, biofilm
Ndiongue *et al.* (2005)	3	HRT, temperatura, material, edad_tubería, biofilm
Chu & Lu (2004)	1	velocidad, HRT, temperatura, material, edad_tubería, biofilm
Lehtola *et al.* (2004b)	2	nº reynolds, velocidad, distancia al punto de cloración, temperatura, material, edad_tubería, biofilm
Pozos *et al.* (2004)	2	velocidad, HRT, temperatura, material, edad_tubería, biof
Tsai *et al.* (2004)	2	nº reynolds, velocidad, HRT, temperatura, material, edad_tubería, biofilm
Batté *et al.* (2003)	5	velocidad, HRT, temperatura, material, edad_tubería, biofilm
Volk & LeChevallier (1999)	8	HRT, material, edad_tubería, biofilm
van der Kooij *et al.* (1995)	4	velocidad, temperatura, material, edad_tubería, biofilm
Sylvestry-Rodriguez *et al.* (2008)	20	temperatura, material, edad_tubería, biofilm
Obst & Schwartz (2007)	3	material, biofilm
Tsvetanova (2006)	51	velocidad, material, edad_tubería, biofilm
Momba & Binda (2002)	18	temperatura, material, edad_tubería, biofilm
Zacheus *et al.* (2000)	19	velocidad, temperatura, material, edad_tubería, biofilm
Lehtola *et al.* (2004a)	6	régimen hidráulico, velocidad, distancia al punto de cloración, temperatura, material, edad_tubería, biofilm
Wingender & Flemming (2004)	18	régimen hidráulico, material, edad_tubería, biofilm
Langmark *et al.* (2007)	80	régimen hidráulico, HRT, temperatura, biofilm
Tsai (2004)	3	velocidad, biofilm
Storey & Ashbolt (2002)	8	nº reynolds, velocidad, material, edad_tubería, biofilm
Bai *et al.* (2010)	15	régimen hidráulico, distancia al punto de cloración, temperatura, material, biofilm
Simoes *et al.* (2006)	4	régimen hidráulico, material, edad_tubería, biofilm
Schwartz *et al.* (2003)	6	distancia al punto de cloración, material, biofilm

Tabla 4.1. Recopilación de datos de la bibliografía

4.2.2. Selección de las variables relevantes y extracción de la información de interés

Las principales características físicas e hidráulicas de los sistemas de distribución que se conoce influyen en el desarrollo de biofilms son las desarrolladas anteriormente (Sección 4.1), a las que pertenecen:

- Naturaleza de los materiales utilizados en los DWDSs
- Tiempo de retención hidráulica
- Depósitos
- Régimen hidráulico
- Velocidad del flujo

Respecto al biofilm, se ha optado por el recuento de heterótrofos en placa (HPC, del inglés Heterotrophic Plate Count) como método de cuantificación. Existen métodos más completos, que reflejan de manera más veraz la realidad, como el recuento total de células (TCC, del inglés Total Cell Count) o el recuento de células metabólicamente activas (MAC, del inglés Metabolically Active Cells), pero estos métodos se han empezado a utilizar recientemente y son mucho más complejos y costosos de llevar a cabo. Esto hace que el HPC sea el método más utilizado, habiendo muchos más datos reales disponibles medidos de esta manera. De este modo, se dispone de un número mucho mayor de datos, aumentando así la fiabilidad de los resultados y la consistencia de los parámetros con los que se trabaja.

4.2.3. Unificación de la información y generación de la base de datos

Para aumentar el número de datos disponibles y minimizar la pérdida de información se ha decidido usar la edad de la tubería como medida de rugosidad y de acumulación de depósitos. Esta decisión se basa en que la acumulación de productos de corrosión y sustancias disueltas en las tuberías aumenta con el tiempo y, de la misma manera, aumenta la rugosidad de las tuberías [Christensen, 2009]. También se sabe que los depósitos más viejos suelen tener más biomasa y contenido bacteriano [Chowdhury, 2011].

Con el mismo objetivo, para tratar la variable del HRT se ha decidido generar un índice sintético de edad del agua utilizando el HRT y la distancia al punto de cloración (km). En

cada caso, a los valores reales se les ha restado el mínimo valor y se les ha dividido entre la diferencia del máximo valor y el mínimo. En la realización del índice sintético se juntan dos variables y se convierten en una sola, por lo que con el fin de no sesgar el estudio se decide mantener las proporciones existentes en los datos originales. Estos presentan un ratio de 2.5 favorable para el HRT, por lo que el HRT se multiplica por un factor de 0.7 y la distancia al punto de cloración por un factor de 0.3.

$$\text{Edad del agua} = 0.7 \times \left(\frac{HRT_i - HRT_{min}}{HRT_{max} - HRT_{min}}\right) + 0.3 \times \left(\frac{Dist_i - Dist_{min}}{Dist_{max} - Dist_{min}}\right) \quad \textbf{[1]}$$

De esta manera, se obtiene un índice sintético adimensional de edad del agua en el que los valores más cercanos a uno corresponden con los que mayor edad del agua.

Finalmente, los factores físicos e hidráulicos de los DWDSs seleccionados para su estudio mediante el proceso KDD fueron el material y la edad de las tuberías, el índice sintético de edad del agua, la velocidad de flujo y el régimen hidráulico. La base de datos final, con el tipo de información disponible en cada caso se presenta en la Tabla 4.2.

AUTOR	Nº MUESTRAS	VARIABLES MEDIDAS
Muñoz *et al.* (2007)	16	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Manuel *et al.* (2010)	4	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Lehtola *et al.* (2006)	4	régimen hidráulico, velocidad, material_tubería, edad_tubería, biofilm
Chu *et al.* (2005)	1	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Ndiongue *et al.* (2005)	3	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Chu & Lu (2004)	1	velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Lehtola *et al.* (2004b)	2	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Pozos *et al.* (2004)	2	velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Tsai *et al.* (2004)	2	régimen hidráulico, velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Batté *et al.* (2003)	5	velocidad, edad_agua, material_tubería, edad_tubería, biofilm
Volk & LeChevallier (1999)	8	edad_agua, material_tubería, edad_tubería, biofilm
van der Kooij *et al.* (1995)	4	velocidad, material_tubería, edad_tubería, biofilm
Sylvestry-Rodriguez *et al.* (2008)	20	material_tubería, edad_tubería, biofilm
Obst & Schwartz (2007)	3	material_tubería, biofilm
Tsvetanova (2006)	51	velocidad, material_tubería, edad_tubería, biofilm
Momba & Binda (2002)	18	material_tubería, edad_tubería, biofilm
Zacheus *et al.* (2000)	19	velocidad, material_tubería, edad_tubería, biofilm
Lehtola *et al.* (2004a)	6	régimen hidráulico, edad_agua, material_tubería, edad_tubería, biofilm
Wingender & Flemming (2004)	18	régimen hidráulico, material_tubería, edad_tubería, biofilm
Langmark *et al.* (2007)	80	régimen hidráulico, edad_agua, biofilm
Tsai (2005)	3	velocidad, biofilm
Storey & Ashbolt (2002)	8	velocidad, material_tubería, edad_tubería, biofilm
Bai *et al.* (2010)	15	régimen hidráulico, edad_agua, material_tubería, biofilm
Simoes *et al.* (2006)	4	régimen hidráulico, material_tubería, edad_tubería, biofilm
Schwartz *et al.* (2003)	6	edad_agua, material_tubería, biofilm

Tabla 4.2. Información disponible en la base de datos generada

Capítulo 5

Pre-procesamiento, transformación y visualización de los datos

La obtención de datos en relación al desarrollo de biofilms en los DWDSs se encuentra muy limitada. Datos *in situ* son muy complicados de obtener debido a la dificultad de acceso y de toma de muestras en estos sistemas; aparte de las restricciones que los administradores de los servicios de agua, por lo general, imponen en la toma de muestras y en la difusión de los datos. Esto hace que los estudios experimentales sean, casi, la única forma de obtención de datos. Sin embargo, se encuentra una gran cantidad de inconvenientes cuando se llevan a cabo. Por un lado, son necesarios investigadores altamente cualificados tanto en microbiología, como en ingeniería hidráulica, para diseñar y entender todos los procesos que se producen en el sistema. Por este motivo, muchas veces, los resultados están sesgados, centrándose sólo en un área de conocimiento. Por otro lado, la reproducción de las condiciones de crecimiento de los biofilms en los DWDSs es una tarea muy difícil. Es por ello que la mayoría de los estudios se simplifican, estudiando sólo uno o dos aspectos relacionados con el desarrollo biofilms en los DWDSs, sin tener en cuenta la influencia de todo el entorno. Además, debido a la complejidad del ambiente estudiado se utilizan diferentes metodologías y reactores de crecimiento de biofilm. Esto hace que la recopilación de datos llevada a cabo, sobre biofilms en los DWDSs, sea ambigua y difícil de comparar, además de incompleta.

También se debe tener en cuenta que la complejidad aumenta cuando se analizan datos proporcionados por diversos estudios y fuentes de información [Aubrecht *et al.*, 2003]. En estos casos, por lo general, se deben manejar problemas como la heterogeneidad en las medidas de los datos, la multi-escalaridad y una presencia importante de datos faltantes, entre otros inconvenientes.

La primera parte de este capítulo aborda estos problemas, mediante el pre-procesamiento y transformación de los datos [Gibert *et al.* 2008]. Se propone un pre-proceso inicial de la base de datos con la que se va a trabajar. El primer apartado de este pre-proceso consiste en una fase de detección donde se buscan y eliminan los datos atípicos (*outliers*) aplicando técnicas de *clustering*. A continuación, se propone reconstruir los datos perdidos mediante la utilización de redes neuronales artificiales y, posteriormente, se realiza una transformación de los datos discretizando las variables continuas para así normalizar la base de datos.

El segundo objetivo de este capítulo es hacer una visualización inteligente de los metadatos obtenidos con el propósito de hacer un análisis exploratorio e identificar posibles patrones de interés y grupos relacionados con el desarrollo de biofilm en los DWDSs. Las técnicas de visualización utilizadas son el *scatterplot* y el RadViz (Radial Coordinate Visualization). Ambas se optimizan mediante la aplicación de algoritmos de visualización inteligente de datos.

Este proceso viene resumido en la Figura 5.1.

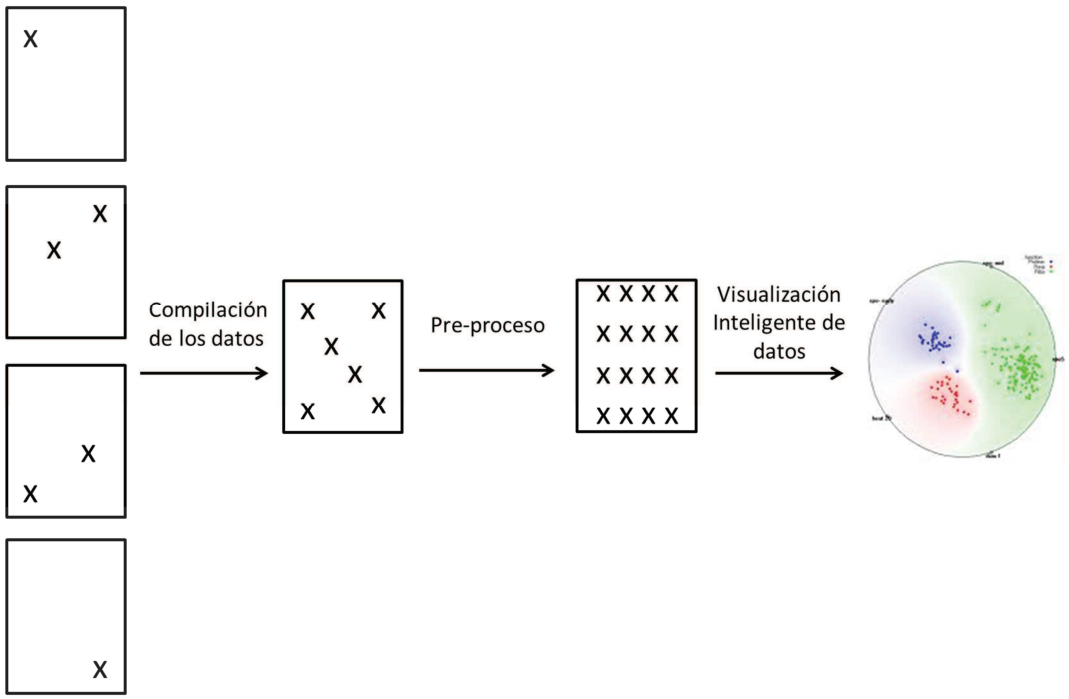

Figura 5.1. Proceso de pre-procesamiento y visualización

5.1. Pre-procesamiento de los datos

En prácticamente todos los estudios de descubrimiento del conocimiento a través de los datos se requiere un pre-procesamiento de los mismos. Normalmente, los metadatos incluyen múltiples escalas, y habitualmente presentan heterogeneidad y una importante cantidad de datos faltantes. La presencia de datos faltantes, por lo general, se debe a la existencia de datos provenientes de diferentes fuentes en la base de datos.

Por lo tanto, se requieren herramientas para [Gibert *et al.*, 2008]:

- Entender mejor el conjunto de datos.
- Detectar imperfecciones en el conjunto de datos y gestionarlas de la manera más adecuada.
- Preparar correctamente los datos para la(s) técnica(s) seleccionada(s) para su posterior análisis.

El pre-procesamiento, dependiendo de la naturaleza de los datos y de los objetivos del propio análisis, va desde las técnicas descriptivas más simples a los métodos de análisis de datos más sofisticados. La mayoría de las operaciones llevadas a cabo en la etapa de pre-procesamiento se pueden agrupar en dos familias de técnicas:

- Técnicas de detección

  Son aquellas orientadas a detectar imperfecciones en el conjunto de datos o a verificar el cumplimiento de los supuestos requeridos para un análisis particular (valores atípicos, datos faltantes y detección de observaciones influyentes).

- Técnicas de transformación

  Se trata de aquellas orientadas a realizar transformaciones en el conjunto de datos con el fin de corregir las imperfecciones detectadas anteriormente, o para lograr las condiciones técnicas para aplicar una técnica de análisis determinada (tratamiento de valores atípicos, imputación de datos faltantes, técnicas de reducción de dimensionalidad, creación de nuevas variables transformadas, filtrado, remuestreo).

Aparte de las técnicas estadísticas clásicas de limpieza de datos, las técnicas inductivas también son una alternativa cuando los métodos analíticos/tradicionales fallan, son demasiado lentos o simplemente no existen.

5.1.1. Técnicas de detección: *clustering*

El *clustering* es una técnica utilizada popularmente para agrupar puntos de datos u objetos similares en grupos o *clusters* [Jain & Dubes, 1988]. Se han desarrollado varias técnicas de detección de *outliers* basadas en el *clustering*. La mayoría de estas técnicas se basan en la suposición de que los objetos "normales" pertenecen a grupos grandes y densos, mientras que los *outliers* forman agrupaciones de pequeño tamaño [Loureiro *et al.*, 2004; Niu *et al.*, 2007]. El *clustering* es una herramienta importante para el análisis de *outliers*.

El presente trabajo, que estudia el desarrollo de biofilms en los DWDSs mediante metadatos, necesita administrar conjuntos de datos que contienen *inputs* de varios tipos. Esto hace que se requiera un algoritmo escalable y capaz de manejar atributos de diferentes tipos, por lo que los métodos clásicos no son la solución. En este caso particular, para la detección de *outliers* en metadatos se propone utilizar los algoritmos CLARA / PAM [Kaufman & Rousseeauw., 1990]. Por ejemplo, el algoritmo PAM (*Partitioning Around Medoids*) puede manejar varios tipos de atributos pero no es eficiente con grandes bases de datos. El algoritmo *k-means* [Hartigan & Wong, 1979; Likas *et al.*, 2003] pueden manejar gran cantidad de información pero sólo trata conjuntos de datos formados por variables de intervalo escaladas. Sin embargo, el algoritmo CLARA (*Clustering Large Applications*) es una combinación de un método de muestreo y el algoritmo PAM. CLARA extrae una muestra del conjunto de datos y utiliza el algoritmo PAM para seleccionar un conjunto óptimo de medoides de la muestra [Wei *et al.*, 2003]. Para paliar el sesgo de muestreo, CLARA repite la toma de muestras, y el proceso de *clustering,* múltiples veces y selecciona el mejor conjunto de medoides para definir el *clustering* final.

Un *cluster pequeño* es definido como un *cluster* que contiene menos puntos que la mitad de la media del número de puntos en los *k clusters* [Loureiro *et al.*, 2004]. Para detectar los *outliers* (si hay), se calcula la distancia entre el medoide del *cluster* candidato a representar los elementos de los valores extremos, μ_o, y cada uno de los medoides de los *clusters*, μ_i, que

representan la "normalidad". Se toma el máximo de estas distancias. Si el valor producido es mayor que un umbral calculado. Tentativamente el umbral, *T*, se calcula como el promedio de todos los valores de la distancia multiplicada por un factor constante, y en este trabajo vamos a proponer = 2, entonces el grupo se considera como *outlier*, de lo contrario, no.

algoritmo: detección de *outliers* con el algoritmo CLARA
1. Aplicar el algoritmo de *clustering* CLARA para producir un conjunto de *k clusters*
2. Comprobar si existe algún *cluster pequeño* y considerar los puntos (objetos) que pertenecen a estos *clusters* como candidatos a ser *outliers*
2.1 Calcular la distancia entre el medoide de este *cluster*, μ_o, y el resto de medoides $D_{oi} = \|\mu_o - \mu_i\| \forall_i \neq 0$
2.2 Calcula $M\ D_o = \arg\max_i \|\mu_o - \mu_i\| \forall_i \neq 0$
2.3 Si $M_o > T$ entonces el *cluster* está compuesto por *outliers* volver al paso 2]

Tabla 5.1. El método de *clustering* para la detección de *outliers*

5.1.2. Técnicas de transformación: redes neuronales artificiales

El uso de redes neuronales artificiales (ANNs, del inglés *Artificial Neural Networks*) encaja perfectamente en la filosofía de trabajo con metadatos como herramienta para la interpolación, restaurando automáticamente los datos necesarios para llevar a cabo el análisis posterior y la creación de modelos *data driven* [Herrera *et al.*, 2011b].

Una ANN es un grupo interconectado de neuronas artificiales. Cada neurona ejecuta un cálculo no lineal basado en los valores del *input* y el valor resultante alimenta a otras neuronas. Las neuronas normalmente se disponen como series de capas interconectadas. Se utiliza un algoritmo (generalmente de retro-propagación) que, basándose en los datos presentados a la red, ajusta iterativamente los pesos de conexión de la neurona de tal manera que el rendimiento de predicción de la red mejore [Bishop, 2005]. Los pasos para desarrollar un modelo ANN vienen detallados en Bishop (2005), entre otras referencias.

La red neuronal más común es la red de alimentación hacia adelante (*feed-forward network*) (Figura 5.2), la cual utiliza para su entrenamiento el algoritmo de retro-propagación [Bougadis *et al.*, 2005]. La obtención de este tipo de redes es un proceso iterativo, donde cada caso de la muestra se presenta varias veces a las neuronas de entrada de la ANN.

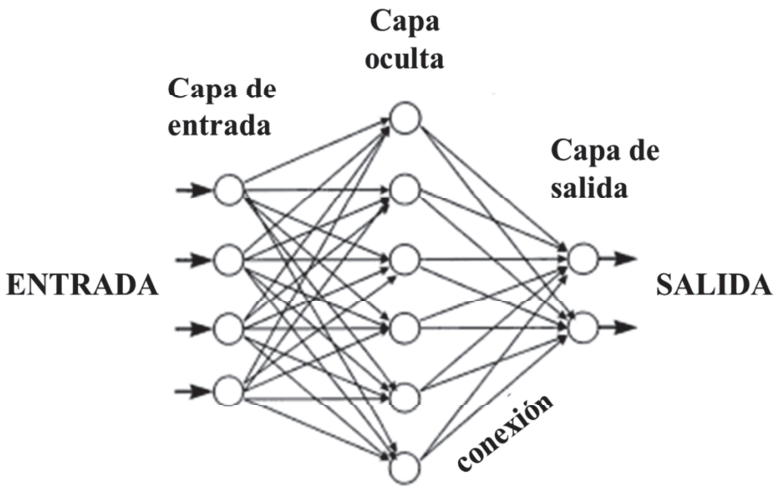

Figura 5.2. Una red típica de tres capas de alimentación hacia adelante

Normalmente, para predicciones se utiliza un modelo típico de tres capas alimentado hacia adelante [Lingireddy & Ormsbee, 1973]. Los nodos ocultos (h) con las correspondientes funciones de transferencia no lineales son usados para procesar la información recibida por los p nodos de entrada, cada uno asociado a uno de los predictores. La función de transferencia se utiliza para acotar la salida de la neurona y generalmente viene dada por la interpretación que queramos darle a dichas salidas. Algunas de las más utilizadas son la función sigmoidea (para obtener valores en el intervalo [0,1]) y la tangente hiperbólica (para obtener valores en el intervalo [-1,1]). Finalmente, el modelo puede ser expresado como [Zhang & Qi, 2005]:

$$Y_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j f\left(\sum_{i=1}^{h} \beta_{ij}\, y_{t-j} + \beta_{0j}\right) + \in_t \qquad \textbf{[2]}$$

donde p es el número de nodos de entrada, h es el número de nodos ocultos, f es una función de transferencia sigmoidal; α_j, con j = 0, 1, …, h, como el vector de los pesos desde los nodos ocultos a los nodos de salida y β_{ij}, con i = 0, 1, …, p y j = 1, …, h, como los pesos desde los nodos de entrada a los nodos ocultos. Los pesos son coeficientes capaces de adaptarse dentro de la red para determinar la intensidad de la señal de entrada registrada por cada neurona artificial. Pueden ser modificados en respuesta de los ejemplos de entrenamiento (de acuerdo a la topología específica o debido a las reglas de entrenamiento). α_0 y β_{0j} son los pesos de los arcos que tratan de evitar un posible sesgo en el modelo final. y_{t-j} se refiere al valor de salida de la red neuronal en la iteración $(t-j)$. La expresión de la formula la

tenemos para $t = 1, \ldots, T$; siendo T el número de iteraciones propuestas para llevar a cabo la ANN. $\in_t$ se refiere al término del error en la iteración t-ésima.

5.1.3. Aplicación de las técnicas de detección y transformación

Se parte de los 303 datos iniciales, obtenidos en la primera etapa del proceso KDD (Capítulo 4). A estos casos se les aplica el algoritmo de *clustering* divisivo PAM [Kaufman *et al.*, 1987] dada la diferente naturaleza de los datos (continuos y discretos). PAM detecta 8 *outliers* que son eliminados de la base de datos original, dado que todos ellos además están conformados por una gran parte de información faltante. Este es un problema que, aunque no en la misma media que los anteriores, siguen teniendo los 295 registros que ahora forman la base de datos. Por ello se propone un plan especial de recuperación de la máxima cantidad de información.

El uso de las redes neuronales como herramienta de interpolación [Herrera *et al.*, 2011b] es el punto de partida para llevar a cabo esta tarea. La idea es realizar una interpolación progresiva, de manera que se van reconstruyendo las variables una a una, apoyándose en los valores de las demás. Se propone comenzar a trabajar con el input (variable) del que se tiene más información en la base de datos, que en nuestro caso es la variable edad del agua. De esta manera, se maximiza el uso de información real (no reconstruida) en el proceso de interpolación de los datos. En todos los casos que tratamos se ha comprobado que una red neuronal de tres capas, alimentada hacia adelante, con arquitectura 5-3-1 y un parámetro de decaimiento de los pesos de 0.001 es la propuesta más adecuada (contando con un menor RMSE - raíz del error cuadrático medio - en la fase de validación que otras propuestas) (Figura 5.3).

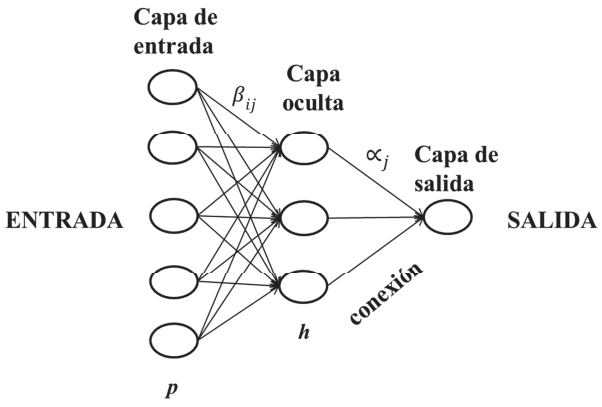

Figura 5.3. Estructura de la ANN utilizada

Descrito el proceso a seguir, se comienza con la interpolación de la edad del agua. Es necesario interpolar 72 datos, ya que se cuenta con 223 valores de edad del agua del total de 295 registros de la base de datos (141 serán datos de entrenamiento - aproximadamente el 63% de los datos - y 82 serán los datos que se reservan para la fase de validación). El siguiente paso es interpolar los datos faltantes de la variable edad de la tubería. De este *input* se tienen 192 datos y es necesario interpolar 103. El proceso de interpolación se repite con las variables velocidad y flujo; sin embargo, dado el número de datos ausentes de estas últimas sólo llegamos a tener una base de datos completa de 210 registros (de los 295 registros incompletos iniciales), dado que no podremos interpolar (predecir) un número mayor de datos del que tenemos disponible.

5.2. Transformación de los datos

Una vez completada la base de datos se procede a realizar una discretización en base a la bibliografía y el conocimiento de expertos. Es importante destacar que el pre-proceso, en las variables en las que ha sido posible, se ha llevado a cabo con datos continuos. De esta manera, por una parte se usa toda la información disponible para el pre-procesamiento, sin perder información, y por otra, se consigue reducir la incertidumbre asociada al pre-proceso.

VELOCIDAD (m/s)		EDAD TUBERIA (años)		MATERIAL		BIOFILM (HPC/cm^2)	
Baja [0-0.7]	L	Baja [0-10]	Y	Plástico	P	Bajo [0-10^3]	L
Media [0.8-1.7]	M	Media [11-30]	M	Cemento	C	Medio [10^4-10^6]	M
Alta [1.8-3.5]	H	Alta [≥ 31]	O	Metal	M	Alto [≥ 10^7]	H
FLUJO				EDAD DEL AGUA			
Laminar		L		Baja		L	
Turbulento		T		Media		M	

Tabla 5.2. Variables y categorías de la base de datos

Las variables a discretizar fueron la velocidad de flujo (m/s), la edad del agua y el biofilm (HPC/cm^2). La velocidad de flujo se discretizó basándonos en la opinión de expertos. De 0 a 0.7 m/s se consideró baja, de 0.8 a 1.7 media y por encima de 1.7 se consideró alta. Para discretizar la edad del agua se tuvo en cuenta que el índice empleado podía tomar valores

únicamente de 0 a 1, por lo que se decidió dividir equitativamente el rango, de manera que por debajo de 0.3 se consideraba baja y por encima media. Valores de 0.6 o superiores se considerarían altos, pero no se encontraron. A la hora de discretizar el biofilm se observó que los datos seguían una distribución normal, por lo que se decidió tomar los valores centrales como valores medios y las colas como extremos. De esta manera, de 0 a 10^3 se consideró un desarrollo bajo de biofilm, de 10^4 a 10^6 medio y de 10^7 en adelante alto.

Finalmente la base de datos obtenida consiste en 210 casos completos con las variables y categorías que se especifican en la Tabla 5.2.

5.3. Visualización de los datos

La visualización juega un papel crucial en la identificación de grupos y patrones de interés en el análisis exploratorio de datos [Leban *et al.*, 2006]. Debido a que la mente humana destaca en la rápida interpretación de la información visual [Nováková & Stepankova, 2009], la visualización de datos es la forma más natural de identificar grupos. Una herramienta esencial en el análisis exploratorio de datos.

Existen varias técnicas que pueden ser usadas para visualizar datos multidimensionales. En este caso, el *scatterplot* y el RadViz han sido las dos elegidas. El *scatterplot* es uno de los métodos de visualización más antiguos y más utilizados. Proyecta los dos atributos seleccionados de una forma clara y fácil de comprender. El RadViz, en cambio, es adecuado para datos multidimensionales. Se ha diseñado para mapear datos descritos por tres o más atributos en una imagen bidimensional [Nováková & Stepankova, 2009]. Aunque el *scatterplot* y el método RadViz se usan principalmente para la visualización de los atributos continuos, también se pueden aplicar a atributos discretos [Leban *et al.*, 2006]. El s*catterplot* es una herramienta muy útil para el análisis de datos. Supongamos que (x_i, y_i), para i desde 1 hasta n, son mediciones emparejadas de dos variables: x e y. Un *scatterplot* de y_i contra x_i puede decirnos mucho acerca de la cantidad de relación existente entre x e y [Cleveland & McGill, 1984]. RadViz, en cambio, se trata de una técnica de visualización multidimensional no lineal, en la que puntos de datos n-dimensionales se presentan como puntos espaciados uniformemente alrededor del perímetro de un círculo.

5.3.1. Visualización inteligente de los datos

Encontrar proyecciones interesantes puede ser una tarea difícil y requiere tiempo para el analista, ya que el número de posibles proyecciones aumenta exponencialmente con el número de atributos visualizados simultáneamente [Leban *et al.*, 2006]. Se utilizan herramientas de visualización inteligente de datos debido a su gran utilidad en la identificación de patrones en grupos de datos multidimensionales. Por esta razón, se utiliza el método VizRank. Este, busca de forma exhaustiva a través de todas las combinaciones de características dentro de los parámetros especificados y evalúa las proyecciones usando el clasificador *k-nearest neighbors* [Demšar, 2005].

El VizRank se utiliza para seleccionar la proyección más informativa de los datos clasificados que nos permita inducir visualmente una regla exitosa en la separación de las distintas clases. En el caso de la visualización RadViz, también se ha implementado el algoritmo FreeViz para encontrar una buena proyección bidimensional de los datos dados. FreeViz es una herramienta que complementa perfectamente el análisis VizRank porque relaja las restricciones de colocación de los anclajes en RadViz y hace computacionalmente eficiente la visualización.

- **VizRank**: se puede aplicar en datos clasificados para encontrar automáticamente proyecciones de datos más útiles (proyecciones donde los grupos con valores de clase diferentes están bien separadas y no se superponen). Se puede utilizar con cualquier método de visualización que asigna puntos a valores atributos en un espacio de dos dimensiones. Vizrank es capaz de clasificar, automáticamente, las proyecciones visuales de datos clasificados por su éxito en mostrar diferentes valores de clase bien separados. El interés de una proyección específica se calcula mediante la inducción del clasificador *k-nearest neighbors* y evaluando su precisión en la clasificación de un conjunto de datos que consta de las posiciones de los puntos proyectados y su información de clase. Esta calificación puede proporcionar una buena estimación de la utilidad de la proyección, ya que las proyecciones con clases muy separadas se asocian con una alta precisión de la clasificación, mientras que cuando las clases se superponen obtienen puntuaciones más bajas. Las proyecciones con una separación de clases perfecta (no hay solapamiento entre las clases) recibe el valor 100, mientras que

las proyecciones menos informativas reciben valores correspondientemente más bajos [Leban *et al.*, 2006].

- **FreeViz:** es una técnica de visualización para el análisis de datos multi-dimensionales clasificados. Las visualizaciones FreeViz pueden presentar datos sobre varias características en el mismo gráfico, pero a través de la optimización de los procedimientos elige la proyección que mejor separa los casos que corresponden a diferentes clases, mediante la flexibilización de las restricciones de colocación de los anclajes de los atributos. En RadViz, estos se colocan en el límite de un círculo [Demsar *et al.*, 2005].

5.3.1.1 Resultados visualización: RadViz

Previo al análisis, se ha realizado un tratamiento de *outliers* para mejorar el aspecto de la visualización (distancia métrica: Euclídea; puntos más cercanos: 15; *outlier* Z: 2.0). Después, la visualización RadViz permite (ya sin sesgos por presencia de *outliers*) representar nuestros datos multidimensionales en un plano (Figura 5.4). Este método es de gran ayuda para la identificación de las relaciones entre los datos.

En nuestro caso de estudio, el biofilm se trata como atributo oculto y el resto de atributos son de anclaje en el límite del círculo. Los puntos verdes representan los casos con una menor formación de biofilm, los amarillos los de desarrollo medio y los rojo son los que presentan un alto desarrollo de biofilm. Al resultado obtenido se le aplica el algoritmo VizRank. VizRank es capaz de clasificar automáticamente las proyecciones visuales de datos clasificados por su éxito en mostrar los diferentes valores de clase bien separados. En nuestro caso la mejor proyección (Figura 5.4) tiene una puntuación de 78,9, y el número de atributos ha sido reducido a 4. El atributo eliminado ha sido material de la tubería. La buena puntuación obtenida indica que las características físicas e hidráulicas de los DWDSs estudiadas, así como los procesos de pre-procesamiento y transformación de los datos, han sido los adecuados ya que permiten realizar una distinción bastante clara entre los diferentes grupos formados en relación al desarrollo del biofilm.

El hecho de que el material de la tubería haya sido eliminado puede deberse en cierta medida a que está parcialmente representado en la edad de la tubería. Debido a la evolución histórica,

se puede decir que, en general, los tubos metálicos tienden a ser más antiguos y los de plástico más nuevos. Aun así, la siguiente mejor configuración mantiene todas las variables, es decir, incluye el material de tubería y presenta un valor de 77.95, muy próximo al anterior.

Por último a la proyección que mejor puntuación ha obtenido se le ha aplicado el algoritmo FreeViz, para mejorar lo máximo posible la visualización de las clases. Tras su aplicación se nota una mejora en la observación de los *clusters*. Aunque existe solapamiento se observa un cambio gradual desde un desarrollo de biofilm bajo a medio, y desde medio a alto.

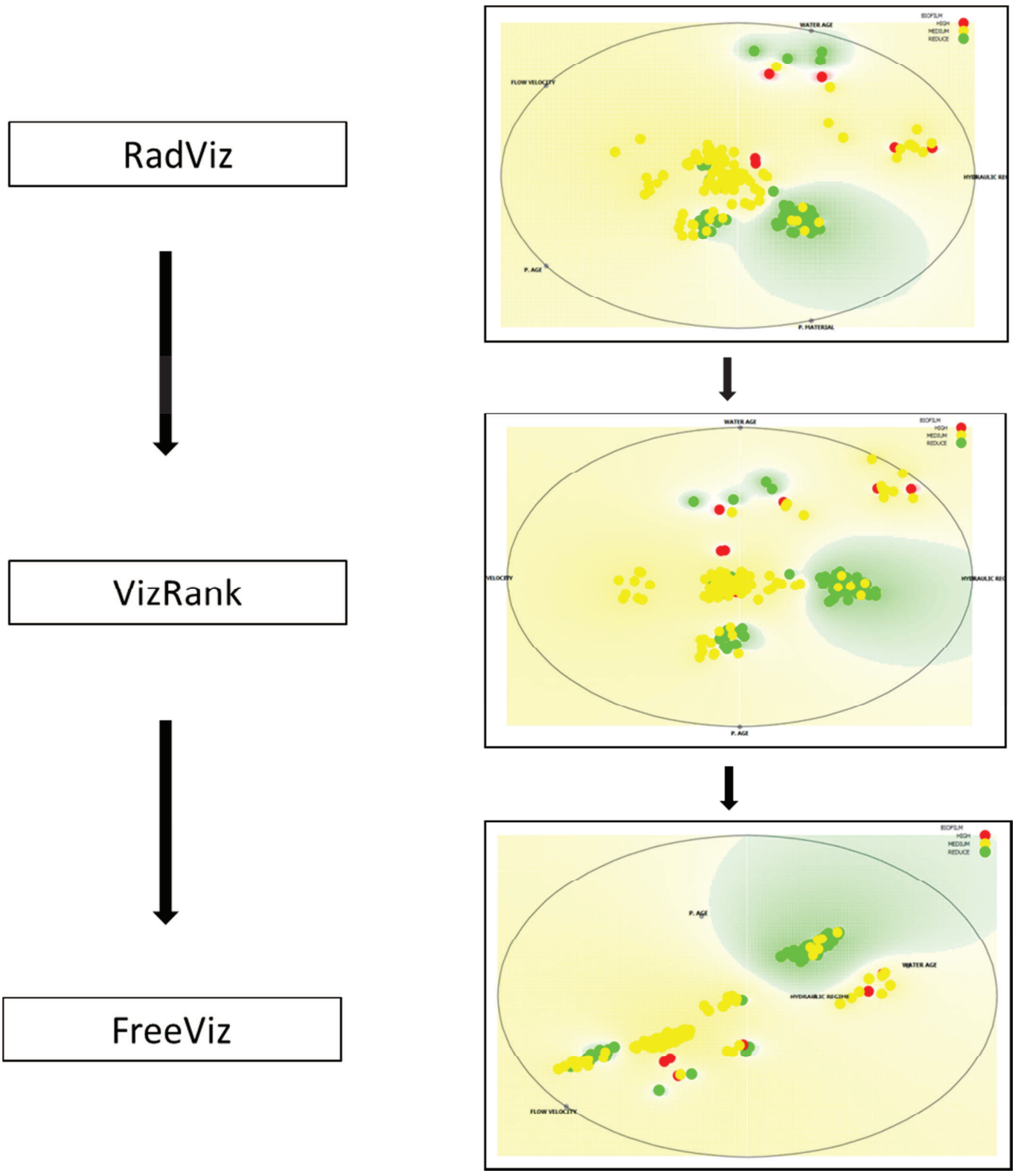

Figura 5.4. Resultados obtenidos tras la aplicación de RadViz, VizRank y FreeViz

5.3.1.2 Resultados visualización: *Scatterplot*

Este método es muy utilizado porque puede proyectar dos atributos de forma muy clara y fácil de comprender. En este caso, con el fin de mejorar la interpretación de las visualizaciones, también se utiliza el tratamiento de *outliers* (ver apartado 5.3.1.1 del presente capítulo) y el algoritmo de optimización VizRank. En la Figura 5.5 se muestran aquellas que presentaron los índices más altos.

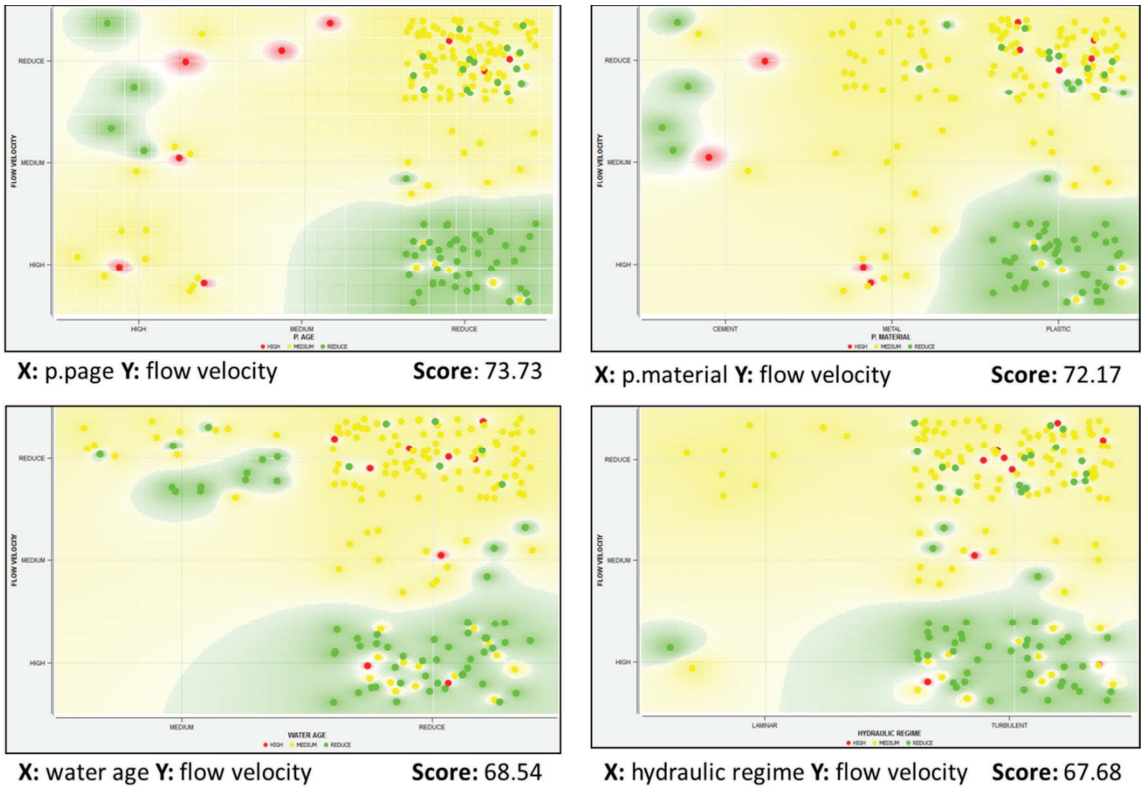

Figura 5.5. Aplicación del algoritmo VizRank en los *scatterplots*

En el primer *scatterplot* (1.1), se observa el efecto de la velocidad de flujo y la edad de la tubería sobre el biofilm. Como era de esperar, según la literatura, la probabilidad de que se dé un alto desarrollo de biofilm es mayor cuando la edad de la tubería es media o alta. Por otro lado, cuando la tubería es joven y la velocidad del flujo es alta la probabilidad de que se de un bajo desarrollo de biofilm es muy elevada. Según la bibliografía consultada, se esperaría encontrar un elevado desarrollo de biofilm con el aumento de la velocidad de flujo [Lehtola *et al.*, 2006], como parece ocurrir cuando la edad tubería es media o alta, pero no es así cuando la edad del tubo se reduce. Esto puede ser explicado por el hecho de que las tuberías aumentan su rugosidad y el contenido de depósitos con la edad, aumentando el potencial de desarrollo de biofilms en su interior [Chowdury, 2011]. La rugosidad y los depósitos ofrecen a los biofilms más superficie de colonización y de refugio, protegiéndolos del desinfectante y

de la fuerza de arrastre del agua. Por lo tanto, es posible que en este caso se observe un efecto de la interacción de estas variables y que el biofilm se vea reducido con el aumento de la velocidad en tuberías jóvenes, pero no en tuberías de mayor edad. Esto se explicaría por el hecho de que los biofilm, que crecen en las tuberías nuevas, al tener estas una superficie poco rugosa, pueden desprenderse con más facilidad cuando las fuerzas de arrastre de agua aumentan al aumentar la velocidad del flujo.

En el siguiente *scatterplot* (1.2), se estudia el efecto conjunto del material de la tubería y la velocidad sobre el desarrollo de biofilm; se encuentra algo similar que en el anterior. Como ya se ha mencionado anteriormente, un mayor desarrollo de biofilm se espera cuando se incrementa la velocidad de flujo [Lehtola *et al.*, 2006], pero esto no sucede en el caso de tuberías de plástico. Este fenómeno puede ser explicado, de nuevo, por la rugosidad de los materiales. La rugosidad del plástico es muy baja y puede que este hecho haga que la unión de los biofilms a la superficie interior del tubo sea más débil en este caso que en otros casos, donde las tuberías son de otros materiales más rugosos. De esta manera, los biofilms se desprenderían más fácilmente a altas velocidades de flujo. También se observa que las tuberías metálicas, como se encuentra en la bibliografía, tienden a tener mayor desarrollo de biofilms; casi no se presentan casos con bajo desarrollo del biofilm.

En el tercer *scatterplot* (2.1), se observa el efecto en el desarrollo de biofilm de la edad del agua y la velocidad de flujo. Aunque la probabilidad de un bajo desarrollo de biofilm sigue siendo alta cuando la velocidad de flujo es alta, también se encuentran casos que presentan alto desarrollo de biofilm. Además se observa, que no se da ningún caso con un alto desarrollo de biofilm cuando la edad del agua es media. Según la bibliografía, esto sería contrario a lo esperado, ya que cuanto más tiempo esté el agua en el sistema de distribución más se favorecerá el desarrollo de biofilm [EPA, 2002]. Sin embargo, estos resultados se pueden ver sesgados por el hecho de que en todos los casos donde la edad del agua es media la velocidad es baja. Esto puede deberse a que ambos parámetros están, en cierto grado, negativamente correlacionados; ya que las menores velocidades de flujo de agua se dan en las tuberías más ramificadas del sistema y esta ramificación aumenta con la distancia al punto inicial de distribución.

En el último *scatterplot* (2.2), donde se estudia el efecto de la rugosidad y la velocidad de flujo en el desarrollo de biofilm, al igual que anteriormente, se encuentra una alta

probabilidad de un bajo desarrollo de biofilm cuando la velocidad de flujo es alta y el régimen es turbulento. Este hecho, en cambio, no sucede cuando el flujo es laminar. Muy probablemente se explique porque a igual velocidad, el régimen turbulento induce fuerzas de corte más altas que el régimen laminar, por lo que el desprendimiento del biofilm se ve favorecido.

Aunque sólo se trata de un análisis exploratorio de los metadatos obtenidos, tras el pre-procesamiento se han encontrado algunos aspectos interesantes. La variable velocidad de flujo tiene especial importancia en el desarrollo del biofilm en los DWDSs, ya que se encuentra en todos los *scatterplots* con los índices de VizRank más altos. También se observa los diferentes efectos de la velocidad de flujo sobre el biofilm dependiendo del material, la edad y el régimen hidráulico de la tubería. La respuesta del biofilm frente a una variable varía según el valor que tomen el resto de variables. Estos resultados indican que el estudio del efecto que la interacción de las diferentes características de los DWDSs tiene en el desarrollo de biofilms es necesario.

La cantidad de información obtenida sólo con una visualización exploratoria de los datos, reafirma la validez de la elección de esta metodología para el estudio del desarrollo de biofilms en los DWDSs. Hace pensar que en los siguientes pasos del KDD donde se aplican técnicas de DM con algoritmos más potentes pueden generar resultados muy interesantes. Al mismo tiempo, refuerza el propósito del presente trabajo, estudiar el efecto de la interacción de los diferentes factores de los DWDSs sobre el desarrollo de biofilm, ya que se observan diferencias según interactúen unos con otros.

Los análisis correspondientes a la primera parte del capítulo se han realizado con el software Weka 6.0 [Witten *et al.*, 2000] y los correspondientes a la segunda parte de visualización con el software Orange Canvas 2.0b [Curk *et al.*, 2005].

Capítulo 6

Aplicación y resultados de las técnicas de minería de datos

En este capítulo se hace un desarrollo teórico de las técnicas de minería de datos utilizadas en el presente trabajo. Se realiza una clasificación general de estas técnicas de DM, en base a los métodos de aprendizaje de los algoritmos en que se basan, dividiéndolas así en técnicas de aprendizaje supervisado y no supervisado.

Dentro del aprendizaje no supervisado las técnicas utilizadas han sido el *clustering* y las reglas de asociación. Para realizar el *clustering* se ha aplicado el algoritmo *Farthest-First*. Se trata de una variante del algoritmo *k-means*, que destaca por su rapidez y sencillez. Las reglas de asociación se han llevado a cabo mediante el algoritmo *Ripple-Down*. Este algoritmo permite que la adquisición de conocimiento se realice sin la ayuda de un experto de una manera sencilla y clara.

Las técnicas de aprendizaje supervisado empleadas han sido las reglas de decisión y los árboles de clasificación. Las reglas de decisión se han obtenido empleando el algoritmo *Nearest-neighbor-like*. Este algoritmo se caracteriza por usar una métrica, la distancia entre un nuevo ejemplo y un conjunto de ejemplos en la memoria. Los árboles de clasificación se han realizado a través del algoritmo *Best-First*. Se ha elegido este algoritmo porque en cada paso se agrega el "mejor" nodo de división. Este "mejor" nodo corresponde con el que tiene una separación mayor del resto, centrándose sobre él dicho algoritmo para lograr una clasificación más clara.

También se presentan los resultados obtenidos tras la aplicación de las técnicas de DM y se lleva a cabo una evaluación e interpretación de los mismos. Los resultados obtenidos tras la aplicación de las diferentes técnicas, tanto de aprendizaje supervisado, como no supervisado,

presentan, en general, un grado de concordancia moderado, siendo mayor en los casos clasificados con desarrollo de biofilm medio y bajo.

El estudio conjunto de las diferentes características físicas e hidráulicas de los DWDSs ha permitido observar diferencias en los resultados según los casos estudiados y en la importancia relativa en el desarrollo de biofilm de las diferentes características estudiadas según los valores que estas tomen en cada caso.

6.1. Métodos de aprendizaje no supervisado

Las características y propiedades de los métodos no supervisados ya han sido explicadas de manera detallada en el Capítulo 3 del presente trabajo. Por ello, en este capítulo, solo se recuerda el hecho de que en el aprendizaje no supervisado no existe conocimiento *a priori* y el conjunto de datos de objetos de entrada es tratado sin contrastarlo con ningún tipo de output esperado.

6.1.1 *Clustering*: algoritmo *Farthest-First*

Las técnicas de *clustering* tratan de obtener grupos o conjuntos entre los elementos, de modo que los que pertenezcan al mismo grupo estén más estrechamente relacionados entre sí que los objetos asignados a grupos diferentes (cf. Capítulo 5 de este trabajo).

El *Farthest First clustering* implementa el algoritmo *Farthest-First* trasversal [Hochbaun & Simons, 1985; Dasgupta, 2002]; es rápido y simple. Es una variante del algoritmo *k-means* que sitúa cada centro del *cluster* que estudia lo más alejado posible de los centros de *clusters* ya existentes. De esta manera, se acelera el *clustering*, en la mayoría de los casos, ya que son necesarios menos procedimientos de reasignación y ajuste.

Este algoritmo comienza seleccionando aleatoriamente un caso que pasa a ser el centro del *cluster*. Se calcula la distancia entre cada una de las instancias y el centro. La distancia que se encuentre más alejada del centro más cercano es seleccionada como el nuevo centro del *cluster* en la siguiente iteración. Después, el punto más alejado de los dos primeros (la distancia de un punto x de un conjunto S es el habitual $\text{mín}\{d(x, y) : y \in S\}$), y así

sucesivamente hasta que se obtienen *k* puntos. Estos puntos son tomados como centros de los *clusters* y cada punto restante se asigna al centro más cercano (Figura 6.1.). Y así, en cada iteración, los *clusters* mejoran sus puntos peor representados [Dasgupta & Long, 2010]. Este proceso se repite hasta alcanzar el número de *clusters* buscado.

Algoritmo *Farthest-First* traversal

- Toma cualquier punto de datos y lo etiqueta como punto 1
- Para $i = 2,3,\ldots,n$
  - Encuentra el punto no etiquetado más alejado de $\{1, 2, ..., i\text{-}1\}$ y lo etiqueta como i.

//Usa $d(x,S) = \min_{y \in S} d(x,y)$ para identificar la distancia //de un punto del conjunto

  - $\Pi(i) = \text{argmin}_{j<i} d(i,j)$
  - $R_i = d(i,\pi(i))$

Figura 6.1. Algoritmo *Farthest-First* trasversal para un conjunto de datos.

6.1.1.1 Resultados: *Clustering*, algoritmo *Farthest-First*

Se han obtenido tres *clusters*; el número de *clusters* deseados ha sido especificado antes de realizar el análisis. El número elegido ha sido tres, debido a que la variable de interés, que en nuestro caso es el biofilm, está dividida en tres categorías: bajo, medio y alto. De esta manera, se quiere ver si, con la información contenida en la base de datos, es posible diferenciar estas tres categorías en función de las variables estudiadas. En la Figura 6.2 se presenta el resultado obtenido.

Se observa que el *Cluster* 0, es el grupo al que más casos pertenecen, en él se encuentran casos en las tres categorías de desarrollo de biofilm, pero especialmente casos de biofilm bajo y medio. El *Cluster* 1 está formado por el 29% de los casos estudiados y todos ellos, excepto uno, corresponden con un desarrollo de biofilm medio. El *Cluster* 2 está representado por el menor número de casos y aunque presenta observaciones en las tres categorías, la mayoría de ellas pertenecen a un desarrollo de biofilm alto.

Al observar el medoide de cada *cluster* (el objeto representativo del conglomerado, aquel objeto para el cual la disimilitud promedio con todos los objetos en el conglomerado es mínima) se observa que cada uno de los medoides corresponde con cada uno de los estados de desarrollo definidos para el biofilm (Tabla 6.1). El *Cluster* 0 está definido por presentar un desarrollo de biofilm bajo, el *Cluster* 1 por un desarrollo medio y el *Cluster* 2 por un desarrollo alto.

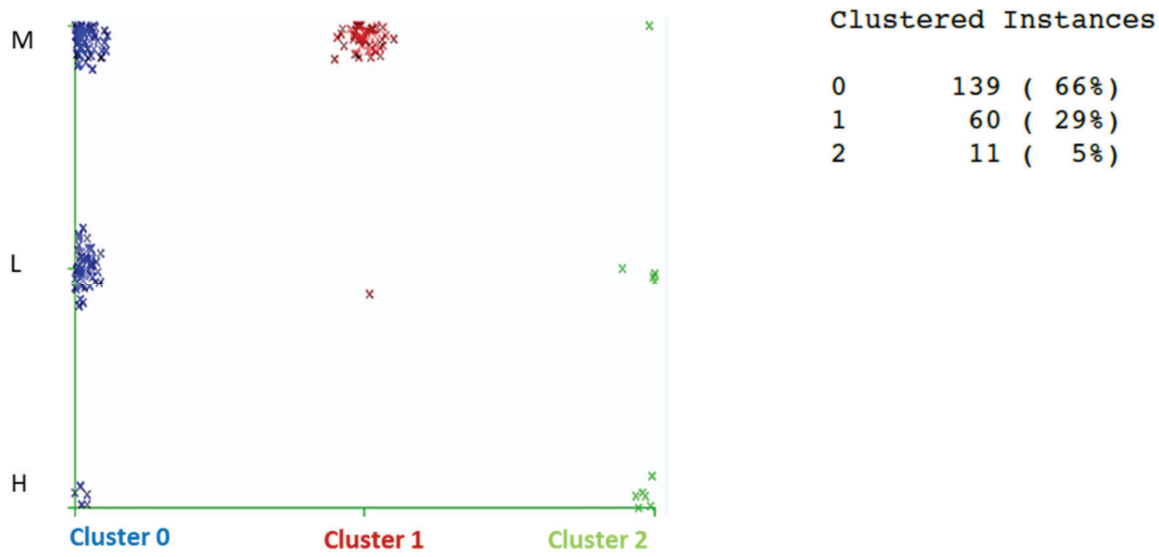

Figura 6.2. Visualización del *clustering*

MEDOIDE	FLUJO	VELOCIDAD	EDAD DEL AGUA	MATERIAL	EDAD TUBERÍA	BIOFILM
Cluster 0	T	H	L	P	Y	L
Cluster 1	T	L	M	M	O	M
Cluster 2	T	L	L	C	M	H

Tabla 6.1. Medoides de los *clusters*

Cada uno de los medoides también está definido por un tipo de material de tubería y una edad. Así, el desarrollo de biofilm bajo corresponde con tuberías plásticas y tuberías jóvenes. Este resultado concuerda a la perfección con la información obtenida de la bibliografía consultada [Thabisile, 2010], donde las tuberías plásticas y las tuberías jóvenes vienen descritas como aquellas que menor desarrollo de biofilm soportan. Esto se explica por el hecho de que cuanto mayor es la rugosidad del material, mayor es el desarrollo de biofilm, ya que las superficies rugosas protegen al biofilm del desprendimiento y le proporcionan mayor área de protección

y colonización [Chowdury, 2011]. Las tuberías plásticas y las tuberías jóvenes son las que menor rugosidad presentan, por lo tanto, también, un menor desarrollo de biofilm. En el caso del medoide del *Cluster* 0 es posible que el hecho de que presente una velocidad de flujo alta también haya influido en que este definido por un desarrollo de biofilm bajo, ya que aunque las velocidades de flujo altas aumentan la difusión de nutrientes también pueden favorecer el desprendimiento de biofilm.

El medoide definido por un desarrollo medio de biofilm está definido por tuberías metálicas y de elevada edad. El medoide que corresponde a un alto desarrollo de biofilm, se define por tuberías de cemento de mediana edad. Según la teoría explicada anteriormente [Chowdury, 2011] esta relación debería ser opuesta, ya que las tuberías metálicas y las de elevada edad son las que mayores rugosidades presentan. El medoide definido por tuberías metálicas viejas con un desarrollo de biofilm medio, presenta una edad del agua superior al medoide correspondiente a un desarrollo de biofilm alto. Cuanto mayor es la edad del agua, más tiempo se encuentra el agua en el sistema de distribución, por lo que el desinfectante residual sufre un mayor decaimiento y se produce un aumento de la temperatura y de la deposición de sedimentos [EPA, 2002], aspectos, todos ellos, que favorecen el desarrollo de biofilm, por lo que en este caso también se hubiese esperado encontrar un elevado desarrollo del mismo.

En los DWDSs, en comparación con los tipos de cemento, se utiliza una gran variedad de metales y aleaciones de metales, y no todos son igual de vulnerables al desarrollo de biofilm. Se sabe que las aleaciones de hierro presentan la mayor densidad de biofilm entre todos los materiales testados en las mismas condiciones, incluyendo hierro de fundición [Kalmbach *et al.*, 2000; Appenzeller *et al.*, 2001], acero al carbón [Tsvetanova, 2006], hierro dúctil [Ollos *et al.* 2003; Camper, 2004], fundición gris y acero embreado [Camper *et al.*, 1996; Niquette *et al.*, 2000; Tsvetanova, 2006]. Esto se debe a que las tuberías de aleaciones de hierro, al tener más cantidad de hierro (Tabla 6.2), sufren más corrosión, la cual favorece el desarrollo de biofilm [Lehtola *et al.*, 2004; Gauthier *et al.*, 1999; Camper, 2004]. Esta diversidad de tipos de materiales metálicos puede haber influido en el hecho de que no quede clara la relación de las tuberías metálicas y un alto desarrollo de biofilm frente a las tuberías de cemento.

Volviendo a los resultados del *clustering* (Tabla 6.1), el hecho de encontrar que cada uno de los medoides se corresponde con cada una de las categorías de biofilm estudiadas, reafirma el hecho de que las características seleccionadas en nuestro estudio para el estudio del desarrollo

de biofilms en los DWDSs son válidas, ya que son capaces de explicar las diferencias encontradas en el grado de desarrollo de los biofilms en los DWDSs. Este resultado también valida el proceso de obtención, procesamiento y transformación de los datos empleados en el presente estudio al observar concordancia en los mismos (cf. Capítulo 5 del presente trabajo).

Aleaciones hierro-carbono	
Hierros	del 0 al 0.10% de C.
Aceros	del 0.10 al 1.76% de C.
Fundiciones	del 2.5 al 4% de C.

Tabla 6.2. Aleaciones hierro-carbono

6.1.2 Reglas de asociación: algoritmo *Ripple-Down*

Las reglas de asociación se utilizan con el objetivo de ver la relevancia de las variables estudiadas, detectar redundancias (dependencias entre atributos) y seleccionar un subconjunto. Se emplean exclusivamente con datos discretos.

En este caso, el algoritmo seleccionado para obtener las reglas de decisión ha sido el algoritmo *Ripple-Down*. Destaca porque permite que la adquisición de conocimiento se lleve a cabo de una manera rápida y simple, posibilitando un rápido desarrollo de las bases de conocimiento por parte de los expertos.

Basándose en la larga experiencia en el desarrollo de sistemas de conocimiento, se ha observado que los expertos humanos no son buenos para el suministro de información sobre cómo llegar a las conclusiones, y no pueden justificar que sus conclusiones son correctas. Las reglas *Ripple-Down* son una metodología de adquisición del conocimiento donde el componente crucial es que el experto justifique por qué su nueva interpretación es mejor que la interpretación dada para el caso (tal vez no haya interpretación) en un ámbito dado [Compton *et al.*, 1991].

Las reglas *Ripple-Down* son reglas con excepciones jerárquicas que se utilizan en la adquisición de conocimiento, ya que proporcionan representaciones de sistemas expertos fácilmente comprensibles, incluso cuando estos son de gran tamaño [Scheffer, 1996].

Las reglas *Ripple-Down*, son listas donde cada regla está asociada a otras. Estas otras reglas se definen como excepciones de la primera. Una regla *Ripple-Down* es una regla del tipo "Si P entonces C_1, excepción: Q entonces C_2 ...", donde Q es un predicado que no ha de cumplirse para clasificar ejemplos con clase C_1, de lo contrario, estos serán etiquetados con clase C_2 [Giráldez, 2003]. El concepto implícito de localidad (una excepción sólo es aplicable si su siguiente regla general es aplicable) hace que las reglas *Ripple-Down* sean muy cómodas para las necesidades humanas: las reglas en bruto se pueden expresar primero y las excepciones de estas pueden ser modeladas como un refinamiento de las hipótesis [Scheffer, 1996].

Este sistema se construye como un árbol con una regla en cada nodo; y con dos ramas en función de si la regla se cumple, o no, por los datos que están siendo considerados. Los vínculos entre las reglas indican dónde se han añadido las reglas para corregir la interpretación dada por una regla anterior. Las reglas se almacenan como una larga lista única, con nuevas reglas añadidas al final de la lista (Figura 6.3).

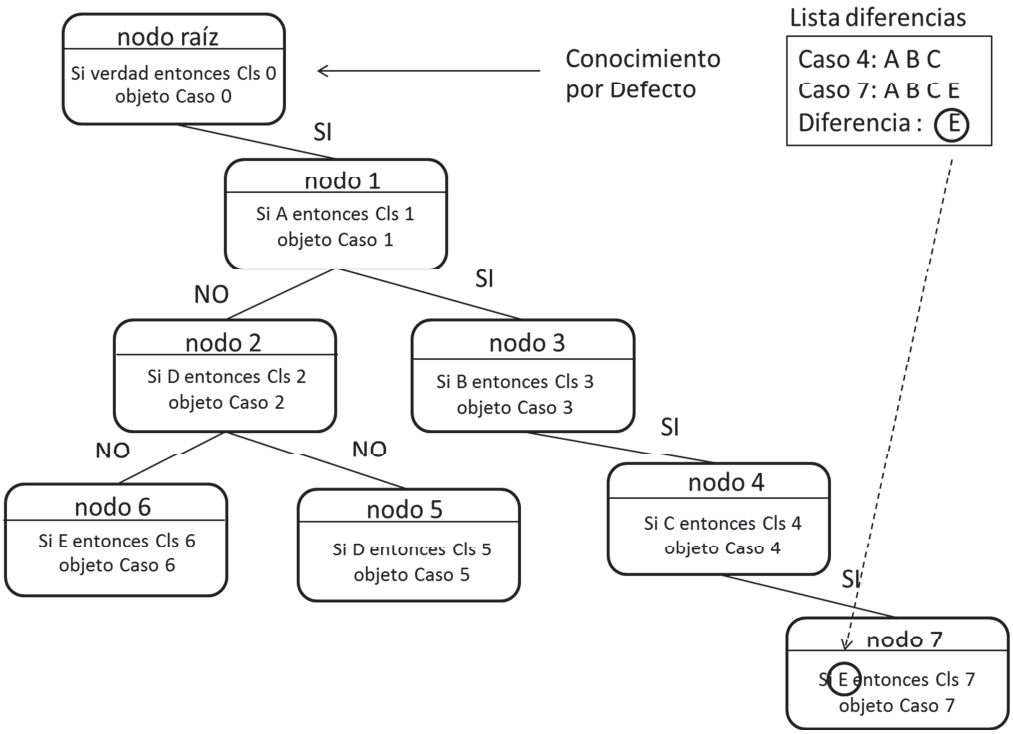

Figura 6.3. Adquisición del conocimiento en las reglas *Ripple-Down* [Wada, 1999]

También existen estructuras alternativas como la presentada en la Figura 6.4, donde cada barra horizontal representa una regla y las nuevas reglas se añaden a la parte inferior de cada pila de reglas.

El problema más evidente de las reglas *Ripple-Down* es la probabilidad de adquisición de conocimiento repetitivo. Es decir, puede darse el caso de terminar con conocimiento repetido en diferentes lugares del árbol. Pero ya que facilita tanto la adquisición de conocimiento se considera una tarea adicional pequeña. De cualquier forma, estas redundancias normalmente se solucionan con la sugerencia de que el árbol sea reorganizado, evitando, así, desequilibrios entre las diferentes ramas y la posibilidad de existencia de desórdenes entre las clasificaciones. El principal problema que presenta este tipo de representación es su difícil interpretación cuando el número de excepciones es elevado [Giráldez, 2003].

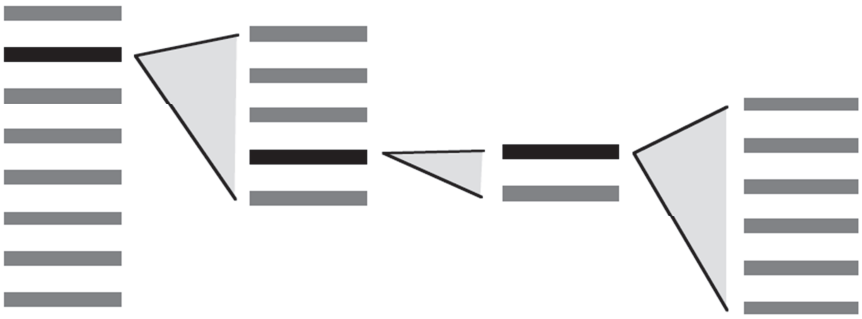

Figura 6.4. Una de las posibles representaciones de la estructura de las reglas *Ripple-Down*

Para mejorar la comprensión de las bases de conocimiento de gran tamaño se han desarrollado diferentes esquemas de representación que ofrecen modularidad mediante el uso de excepciones [Vere, 1980; Dimopoulos & Kakas, 1995]. Sin embargo, el más general para representaciones basadas en atributos y con un esquema más fuertemente estructurado son las reglas *Ripple-Down* [Scheffer, 1996].

6.1.2.1 Resultados: Reglas de asociación, algoritmo *Ripple-Down*

Según el estadístico kappa, que determina hasta qué punto la concordancia observada es superior a la que es esperable obtener por puro azar, y siguiendo los márgenes para valorar el grado de acuerdo en función del índice kappa que propusieron Landis & Koch en 1977 (Tabla

6.3), se puede decir que el grado de acuerdo alcanzado en las reglas de asociación obtenidas mediante la aplicación del algoritmo *Ripple-Down* es moderado (Tabla 6.4).

Kappa	Grado de acuerdo
<0	sin acuerdo
0 - 0.2	insignificante
0.2 - 0.4	bajo
0.4 - 0.6	moderado
0.6 - 0.8	bueno
0.8 - 1	muy bueno

Tabla 6.3. Valoración en función del índice kappa

Mediante las reglas de asociación obtenidas se han conseguido clasificar correctamente más de las de tres cuartas partes de los casos presentes en la base de datos. Concretamente, se han clasificado correctamente 161 casos de los 210 estudiados (Tabla 6.4).

Casos clasificados correctamente	161	77%
Casos clasificados incorrectamente	49	23%
Estadístico Kappa final	0,5259	

Tabla 6.4. Validación cruzada estratificada de las reglas de asociación

Respecto a la exactitud en la clasificación según las clases estudiadas, se ha observado que ha sido menor en los casos de biofilm alto, donde la mayoría han sido clasificados como de desarrollo de biofilm medio. Los caso de desarrollo medio de biofilm han sido, especialmente, bien clasificados (Tabla 6.5).

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
   3   0   8 |   a = H
   3  47  18 |   b = L
   3  17 111 |   c = M
```

Tabla 6.5. Matriz de confusión de las reglas de asociación

En las reglas de asociación resultantes, un primer resultado (por defecto) ha sido la existencia de un desarrollo de biofilm alto (Tabla 6.6). La mayoría de las excepciones de esta regla son para casos de edad de tubería baja, como sería de esperar según la bibliografía estudiada, por

el hecho de que las tuberías jóvenes presentan menor rugosidad y el desarrollo de biofilm se ve favorecido por esta. Para los caso de tuberías jóvenes el desarrollo de biofilm pasa de ser medio a ser bajo, cuando la velocidad de flujo es alta o la edad del agua es media (Tabla 6.7). Un mayor desarrollo de biofilm en el caso de una velocidad de flujo alta, podría ser explicado por un posible aumento de la fuerza de arrastre del agua que haya producido un aumento del desprendimiento de biofilm. Sin embargo, según la bibliografía consultada [EPA, 2002], no se encuentra explicación al hecho de que la mayor edad del agua reduzca el desarrollo de biofilm, ya que se esperaría el efecto contrario. Esta contradicción también se ha observado en los resultados del *clustering* y de la visualización (cf. Capítulo 5 del presente trabajo). Este hecho puede indicar que si el agua permanece circulando durante un largo periodo de tiempo en el sistema se produce un agotamiento de los nutrientes que limita el crecimiento bacteriano, lo que hace que el resto de factores asociados con largos periodos de retención hidráulica, a pesar de producirse, dejen de tener efecto sobre el desarrollo del biofilm.

```
RIpple DOwn Rule Learner(Ridor) rules
--------------------------------------

biofilm = H  (210.0/199.0)
           Except (page = Y) and (velocity = H) => biofilm = L  (28.0/0.0)
[20.0/0.0]
           Except (page = Y) and (material = M) => biofilm = M  (25.0/0.0)
[12.0/0.0]
           Except (page = Y) and (wage = M) => biofilm = L  (15.0/0.0) [9.0/0.0]
           Except (page = Y) and (velocity = M) => biofilm = M  (6.0/0.0)
[1.0/0.0]
           Except (page = Y) and (flow = L) => biofilm = M  (5.0/0.0) [2.0/0.0]
           Except (page = Y) => biofilm = M  (33.0/2.0) [20.0/1.0]
                    Except (velocity = H) => biofilm = L  (30.0/4.0) [18.0/2.0]
                    Except (wage = M) => biofilm = L  (19.0/11.0) [5.0/1.0]
           Except (page = O) => biofilm = M  (18.0/2.0) [8.0/2.0]
                    Except (material = C) => biofilm = L  (4.0/1.0) [2.0/1.0]

Total number of rules (incl. the default rule): 11
```

Tabla 6.6. Resultado de las reglas de asociación

Si las tuberías son de elevada edad sufrirán un desarrollo de biofilm medio, pero si estas tuberías son de cemento el desarrollo de biofilm que se dé en su interior será reducido. Estos resultados son contrarios a lo que cabría esperar según la teoría estudiada. Estas contradicciones pueden deberse a la interacción con otros factores que provocan la reducción del biofilm, o bien a que como indica el índice kappa aunque existen concordancias en los resultados de las reglas de asociación algunos aspectos pueden ser mejorados.

BIOFILM	ALTO
EDAD = JOVEN	MEDIO
VELOCIDAD = ALTA	BAJO
EDAD AGUA = MEDIA	BAJO
EDAD = VIEJA	MEDIO
MATERIAL = CEMENTO	BAJO

Tabla 6.7. Resultado simplificado de las reglas de asociación

6.2. Métodos de aprendizaje supervisado

Los métodos de aprendizaje supervisado se caracterizan por hacer uso de conocimiento *a priori* en el análisis de los datos. De esta manera, el conjunto de datos de objetos de entrada se trata de un conjunto etiquetado mediante algún tipo de información de salida y es tratado contrastándolo con el tipo de *output* esperado. Estos métodos han sido explicados con mayor detalle en el Capítulo 3 de este trabajo.

6.2.1 Reglas de decisión: algoritmo *Nearest-neighbor-like*

Las reglas de decisión generan un conjunto de argumentos con el propósito de obtener hipótesis que traten de explicar un determinado sistema, representando una serie de conceptos que lo describan.

El algoritmo *Nearest-neighbor-like* ha sido el algoritmo elegido para construir estas reglas de decisión. Este algoritmo usa una métrica que mide la distancia entre un nuevo ejemplo y un conjunto de ejemplos existentes en la memoria de dicho algoritmo. El nuevo ejemplo se clasifica de acuerdo a la clase de su vecino más cercano. Un sistema *Nearest-neighbor-like* puro almacena todos los ejemplos en la memoria de manera textual. Luego clasifica nuevos ejemplos al encontrar el caso más similar en la memoria y adopta su clase. Para los atributos numéricos esto se basa generalmente en la distancia euclidiana, donde se trata cada ejemplo como un punto en un espacio de características *n*-dimensional.

Los datos discretos, que es el caso que nos ocupa, son más problemáticos, ya que no encajan en el modelo de función de espacio euclidiano. Para superar esto, la similitud entre los datos discretos se determina contando las características coincidentes. Esta es una función más débil

ya que puede haber varios conceptos basados en características totalmente diferentes, todos los cuales coincidan con el ejemplo estudiado en el mismo grado [Martin, 1995].

El ruido y los atributos irrelevantes también plantean problemas. Así mismo, el sesgo adoptado por la especificidad basada en ejemplos de aprendizaje, aunque a menudo es una ventaja, puede sobre-representar pequeñas reglas a expensas de los conceptos más generales, lo que lleva a una marcada disminución en el rendimiento de clasificación de algunos ejemplos.

6.2.1.1 Resultados: Reglas de decisión, algoritmo *Nearest-neighbor-like*

El índice kappa obtenido mediante estas reglas es de 0.40, por lo que se puede decir que su grado de acuerdo es moderado. El porcentaje de casos clasificados correctamente es del 70.48% (Tabla 6.8).

Casos clasificados correctamente	148	70%
Casos clasificados incorrectamente	62	30%
Estadístico Kappa final	0,3928	

Tabla 6.8. Validación cruzada estratificada de las reglas de decisión

Al observar el nivel de concordancia en la clasificación según las clases estudiadas, se tiene que ha sido menor en los casos de biofilm alto, donde la mayoría aparecen clasificados como desarrollo de biofilm medio. En los casos de desarrollo de biofilm medio y bajo, la mayoría de los casos han sido correctamente clasificados. Especialmente, en el caso de desarrollo de biofilm medio (Tabla 6.9).

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
   2   2   7 |   a = H
   1  39  28 |   b = L
   7  17 107 |   c = M
```

Tabla 6.9. Matriz de confusión de las reglas de decisión

En los resultados obtenidos se observa que son dos las reglas que definen las condiciones para que se dé un alto desarrollo de biofilm (Tabla 6.10). De nuevo, nos centramos en estas debido a que un alto desarrollo de biofilm es la condición que más nos interesa por sus implicaciones en los DWDSs. Una de las reglas que predice un alto desarrollo de biofilm corresponde a un flujo turbulento, tubería metálica y vieja, edad del agua baja y velocidad de flujo alta. Todas estas características, excepto la edad del agua baja, aparecen definidas en la bibliografía como favorecedoras del desarrollo de biofilms en los DWDSs. En este caso una velocidad de flujo alta puede no provocar desprendimiento de biofilm, por el hecho de que las tuberías metálicas y de mayor edad presentan una alta rugosidad en su pared interna. De este modo, los biofilms se encuentran más protegidos, tienen mayor superficie para su desarrollo y pueden fijarse con más fuerza a las paredes de las tuberías evitando su desprendimiento a altas velocidades de flujo, resultados que apoyan a los ya obtenidos en el estudio exploratorio de visualización (Capítulo 5). Sin embargo, en el caso de la segunda regla que predice un desarrollo de biofilm alto, su explicación no está tan clara. Este caso se diferencia del anterior en que la velocidad de flujo puede no ser alta y la tubería puede ser metálica o plástica y de mediana edad. Este resultado, donde las tuberías metálicas y plásticas de la misma edad son equiparadas, a pesar de las diferencias existentes en el potencial de formación de biofilm de estos dos materiales, es difícilmente explicable. Sin embargo, estos resultados pueden deberse a la interacción con otros factores y hasta ahora no haber sido observados por no haberse realizado este tipo de trabajos, donde se estudia el efecto conjunto de las diferentes características físicas e hidráulicas de los DWDSs en el desarrollo de biofilms. Es por ello que, a pesar de las diferencias en el desarrollo de biofilm encontradas según el material de la tubería, puede que estas diferencias se vean suavizadas según el resto de características y condiciones de funcionamiento del sistema.

```
Rules generated :
        class L IF : flow in {T,L} ^ velocity in {H} ^ wage in {L} ^ material in {P} ^ page in {Y}
(4)
        class L IF : flow in {T} ^ velocity in {L} ^ wage in {L,M} ^ material in {P} ^ page in {Y}
(3)
        class M IF : flow in {L} ^ velocity in {L} ^ wage in {L} ^ material in {P} ^ page in {Y}
(7)
        class M IF : flow in {T} ^ velocity in {H} ^ wage in {L} ^ material in {P} ^ page in {O}
(1)
        class M IF : flow in {T,L} ^ velocity in {L,M} ^ wage in {L} ^ material in {M} ^ page in {Y}
(35)
        class L IF : flow in {T} ^ velocity in {H} ^ wage in {L} ^ material in {P} ^ page in {M}
(1)
        class L IF : flow in {T} ^ velocity in {M} ^ wage in {L} ^ material in {P} ^ page in {O,Y}
(2)
        class H IF : flow in {T} ^ velocity in {H} ^ wage in {L} ^ material in {M} ^ page in {O}
(1)
        class H IF : flow in {T} ^ velocity in {L,M} ^ wage in {L} ^ material in {M,P} ^ page in {M}
(3)
        class M IF : flow in {T} ^ velocity in {M} ^ wage in {L} ^ material in {C,M} ^ page in {O}
(3)
        class M IF : flow in {T} ^ velocity in {L} ^ wage in {L,M} ^ material in {C,M} ^ page in {O}
(4)
        class L IF : flow in {T} ^ velocity in {M} ^ wage in {L} ^ material in {C} ^ page in {M}
(2)
        class M IF : flow in {T} ^ velocity in {L} ^ wage in {L} ^ material in {C} ^ page in {M}
(1)
```

Tabla 6.10. Resultado de las reglas de decisión

6.2.2 Árboles de clasificación: algoritmo *Best-First*

En los árboles de clasificación cada nodo interno representa una elección entre varias alternativas, y cada nodo terminal, también llamado nodo hoja, se caracteriza por una clasificación. Los árboles de clasificación son predictores, potencialmente, de gran alcance y proporcionan una descripción explícita de un conjunto de datos. En la práctica, los árboles de clasificación son una de las técnicas más populares en la clasificación, ya que son rápidos y producen modelos con un rendimiento razonable.

La idea básica de cómo se construye un árbol mediante el algoritmo *Best-First* es la siguiente. En primer lugar, seleccionar un atributo para colocar en el nodo raíz y crear algunas ramas para este atributo basándose en ciertos criterios. Luego, dividir los ejemplos de entrenamiento en subconjuntos, uno para cada rama que se extiende desde el nodo raíz. En nuestro caso, se ha considerado un árbol de decisión binario y por lo tanto el número de ramas es dos. A continuación, este paso se repite para una rama elegida, utilizando sólo aquellos casos que realmente la alcanzan. En cada paso, se elige el "mejor" subconjunto de todos los que están disponibles para, así, expandir el árbol. Este proceso de construcción del árbol continúa hasta que todos los nodos son puros o se alcanza un número específico de expansiones.

Es decir, los árboles de clasificación *Best-First* agregan en el árbol, en cada paso, el "mejor" nodo para su posible posterior división. Este "mejor" nodo es el nodo cuya separación corresponde a la máxima reducción de la impureza entre todos los nodos disponibles para la separación (es decir, no etiquetados como nodos terminales). La Figura 6.5 muestra la estructura de un árbol binario *Best-First* hipotético [Shi, 2007].

El problema en el crecimiento de árboles *Best-First* de clasificación es la forma de determinar qué atributo dividir y en qué forma dividir los datos. Debido a que el objetivo más importante de los árboles de clasificación es la búsqueda de modelos precisos y pequeños, tratamos de encontrar los nodos puros (nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente), tan pronto como sea posible.

Con el fin de encontrar el "mejor" nodo de división en cada paso de los árboles de clasificación *Best-First* deben de concretarse los criterios de división. Estos criterios están diseñados para medir las impurezas de los nodos. La impureza de los nodos está basada en la distribución de las clases. Recordemos que el principal objetivo del aprendizaje de los árboles de clasificación es la obtención de modelos precisos y pequeños. Por lo tanto, cuando se divide un nodo, se debe encontrar nodos sucesores puros lo antes posible [Shi, 2007].

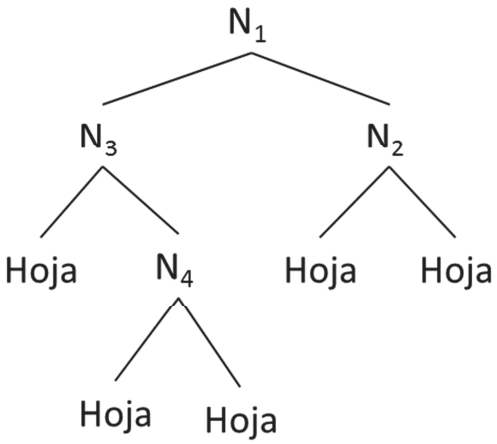

Figura 6.5. Hipotético árbol de clasificación *Best-First*

Para atributos nominales, en problemas multi-clase, como es nuestro caso, se utilizan tanto la búsqueda exhaustiva, como la búsqueda heurística. La búsqueda exhaustiva se lleva a cabo mediante el índice Gini [Breiman *et al.*, 1984]. El tiempo de cálculo de la búsqueda exhaustiva es exponencial respecto al número de valores de un atributo nominal. La búsqueda

heurística puede reducir el tiempo de búsqueda a lineal. Es, por tanto, obvio que para un atributo que tiene muchos valores la búsqueda heurística es la mejor opción porque puede reducir significativamente el tiempo de cálculo [Shi, 2007].

La única diferencia entre árboles de clasificación estándar y árboles de clasificación *Best-First* es que, el aprendizaje de árboles de clasificación estándar se expande a nodos en profundidad de primer orden, mientras que en el árbol de clasificación *Best-First* se expande al "mejor" primer nodo.

6.2.2.1 Resultados: Árboles de clasificación, algoritmo *Best-First*

En este caso el índice kappa es de 0.56 y el porcentaje de casos correctamente clasificados es del 78.10%. Los árboles de clasificación utilizan algoritmos más potentes que las reglas de decisión, suelen ser más robustos frente al ruido, y es por eso que se mejora la predicción respecto a estas. En este caso se ha obtenido un grado de acuerdo moderado, próximo a bueno (Tabla 6.11).

En la proporción de aciertos en la clasificación según las clases de desarrollo de biofilm se observa, como en los casos anteriores, que ha sido menor en los casos de biofilm alto, donde la mayoría aparecen clasificados como desarrollo de biofilm medio. Nuevamente, en los casos de desarrollo de biofilm medio y bajo la mayoría de los casos han sido correctamente clasificados (Tabla 6.12).

En el árbol obtenido (Figura 6.13) la velocidad aparece como el primer nodo de clasificación, diferenciándose los casos con alta velocidad del resto de casos. La influencia de las velocidades altas en el desarrollo del biofilm también se ha visto reflejada en los resultados de las anteriores metodologías estudiadas y en la visualización (Capítulo 5). Este hecho puede ser indicador de que esta característica es de especial importancia en el estudio de los biofilms en los DWDSs.

Dentro de los casos con alta velocidad, se diferencia entre las tuberías de mayor edad y el resto. Según estos resultados las tuberías de mayor edad presentan un desarrollo de biofilm medio y el resto un desarrollo de biofilm bajo. Este resultado refuta la idea de que a velocidades altas de funcionamiento de los DWDSs no existe desprendimiento de biofilm si la

rugosidad de la pared interior de las tuberías es suficientemente alta (como es el caso de las tuberías de mayor edad). Sin embargo este hecho parece ser independiente del material. Esto puede ser debido a que la acumulación de depósitos en la tubería con la edad [Christensen, 2009] puede ser más influyente en el aumento de la rugosidad de la pared interna de las tuberías que el propio deterioro de los materiales. Por otro lado, también se observa un mayor ratio de casos bien clasificados con desarrollo de biofilm bajo en el caso de las tuberías que no son viejas y tienen un flujo turbulento, que en las que tienen flujo laminar, lo que puede explicarse por el hecho de que a igual velocidad si el flujo es turbulento la fuerza de corte del agua es mayor que si el flujo es laminar, lo que aumenta el desprendimiento de biofilm, dándose por ello un desarrollo reducido de biofilm.

Casos clasificados correctamente	164	78%
Casos clasificados incorrectamente	46	22%
Estadístico Kappa final	0,5599	

Tabla 6.11. Validación cruzada estratificada del árbol de clasificación

```
=== Confusion Matrix ===

   a   b   c   <-- classified as
   3   2   6 |   a = H
   7  47  14 |   b = L
   0  17 114 |   c = M
```

Tabla 6.12. Matriz de confusión del árbol de clasificación

En la rama del árbol correspondiente a velocidades no altas de flujo, al igual que en el caso de la rama anteriormente explicada, el siguiente nodo de clasificación es el de edad de la tubería. En este caso se diferencia entre tuberías jóvenes y no jóvenes. Como era de esperar, según la bibliografía, los casos clasificados como altos se encuentran todos en la rama de tuberías no jóvenes. Sin embargo, ninguno de los casos de biofilm alto corresponde a tuberías metálicas, como se esperaría en un principio. Los casos clasificados con un desarrollo de biofilm alto, son los de aquellas tuberías plásticas no jóvenes con baja velocidad de flujo y los de las tuberías de cemento de edad media y baja velocidad de flujo.

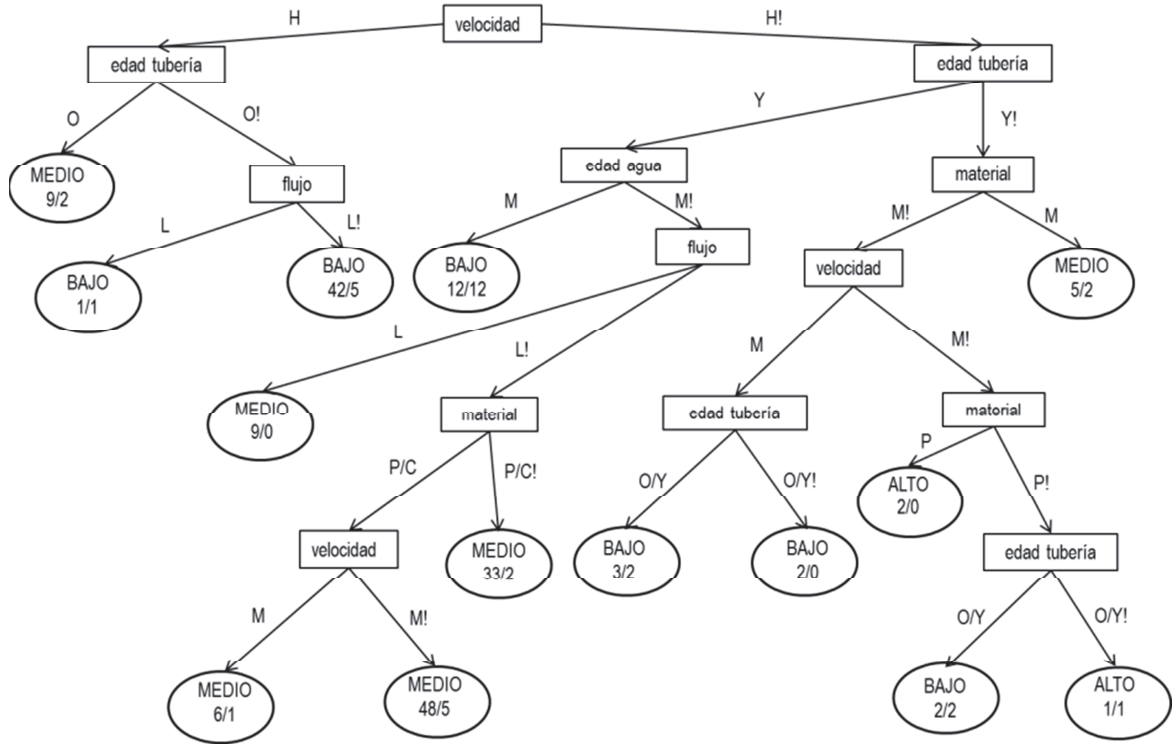

Tabla 6.13. Resultado del árbol de clasificación

Capítulo 7

Ensamblaje

Las técnicas de ensamblaje tratan de mejorar el rendimiento de los algoritmos de aprendizaje automático, combinando técnicas "débiles" con el fin de encontrar un clasificador de alta precisión, o la mejor forma para el conjunto de entrenamiento [Herrera *et al.,* 2012b].

Los modelos de ensamblaje aumentan la precisión y la robustez frente a los métodos de modelos únicos (Figura 7.1). Son de gran eficiencia porque un problema complejo puede ser descompuesto en múltiples sub-problemas que son fáciles de entender y solucionar. Se pueden encontrar aplicaciones satisfactorias de los métodos de ensamblaje en diferentes áreas de estudio, como: finanzas [Leigh *et al*., 2002], bioinformática [Tan & Gilbert, 2003], medicina [Mangiameli *et al*., 2004] y geografía [Bruzzone *et al*., 2004], entre otras.

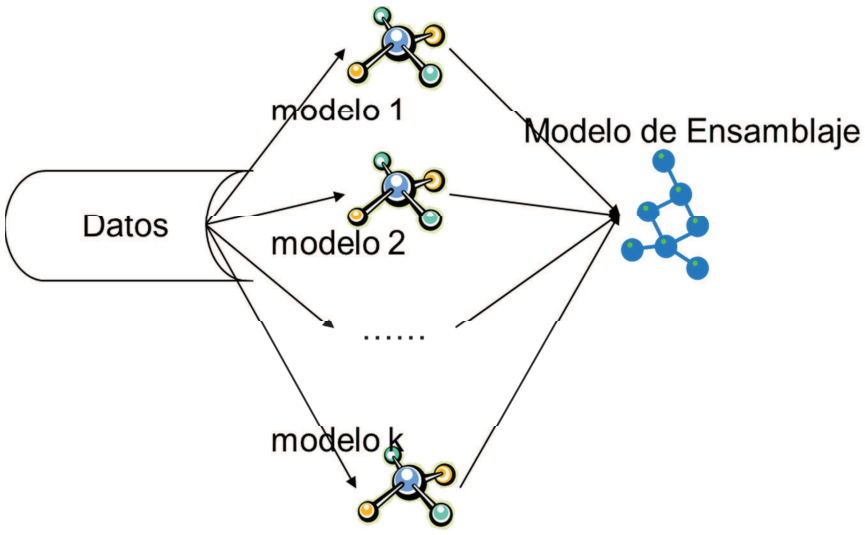

Figura 7.1. Modelo de ensamblaje

La principal idea detrás de la metodología de ensamblaje es ponderar varios clasificadores de patrones individuales, y combinarlos para obtener un clasificador que supera a cada uno de ellos [Rokach, 2010]. En los últimos años, se observó en estudios experimentales que al

combinar los resultados de múltiples clasificadores se reducía la generación de error [Domingos, 1996] del clasificador individual. En efecto, los métodos de ensamblaje pueden usar efectivamente esa diversidad para reducir la varianza sin aumentar el sesgo de las estimaciones. Incluso, en algunas situaciones, el método de ensamblaje también puede disminuir dicho sesgo [Shawe-Taylor *et al.*, 1998].

7.1. Algoritmos de ensamblaje

Existen diferentes modelos de algoritmos de ensamblaje. El caso más sencillo es el algoritmo *bagging*. Este algoritmo une los resultados de varios clasificadores de aprendizaje, en una única clasificación. El *boosting*, en cambio, es una técnica para mejorar el rendimiento de los algoritmos de aprendizaje automático que combina técnicas "débiles" con el fin de encontrar un clasificador de alta precisión, o la mejor forma para el conjunto de entrenamiento [Herrera *et al.*, 2012b]. Existen varios algoritmos diferentes de *boosting*, dependiendo de la forma exacta de ponderar los casos y los modelos [Shi, 2007]. Por último se encuentra el algoritmo *AdaBoost*, se trata de una adaptación del *boosting*, que se utiliza para obtener mejores resultados de los casos individuales con peor salida y así optimizar el rendimiento global del algoritmo *boosting*.

7.1.1. El algoritmo *bagging*

El *bagging* (*boostrap aggregating)* es un método simple para generar un ensamble de clasificadores. El ensamble de clasificadores, creado con este método, consolida los resultados de varios clasificadores de aprendizaje, en una única clasificación. Esto resulta en un clasificador cuya precisión es mayor que la precisión de cada uno de los clasificadores individuales. Específicamente, cada clasificador en el ensamble está entrenado en una muestra de casos (permitiendo repeticiones) tomada del conjunto de entrenamiento. Todos los clasificadores son entrenados usando el mismo inductor [Rokach, 2010].

Para asegurar que hay un número suficiente de casos de entrenamiento en cada muestra, es común fijar el tamaño de cada muestra al tamaño del conjunto de entrenamiento original. En la Tabla 7.1 se observa el pseudo-código para construir un ensamble de clasificadores utilizando el algoritmo *bagging* [Breiman, 1996]. El algoritmo recibe un algoritmo de inducción *I* que es usado para entrenar a todos los miembros del ensamble. El criterio de parada, en la línea 6, termina el entrenamiento cuando el tamaño del ensamble alcanza *T*. Una de las mayores ventajas del *bagging* es que puede ser implementado fácilmente en un modo

paralelo entrenando los diferentes clasificadores de ensamble en diferentes procesadores [Rokach, 2010].

Entrenamiento *bagging*
Requiere: I (un inductor base), T (número de interacciones), S (el conjunto de entrenamiento original), μ (el tamaño de muestra).
1: $t \leftarrow 1$
2: **repetir**
3: $S_t \leftarrow$ una muestra de μ casos de S (muestreo con reemplazamiento)
4: Construir clasificadores M_t usando I con S_t como conjunto de entrenamiento
5: $t \leftarrow t + 1$
6: **hasta** $t > T$

Tabla 7.1. Entrenamiento del algoritmo *bagging*

Como se usa muestreo con reemplazamiento, algunos de los casos originales S pueden aparecer más de una vez en cada una de las iteraciones del proceso, S_t, y algunos pueden no ser incluidos en ningún caso. Además, si se usa un tamaño de muestra grande causa que las muestras individuales se solapen considerablemente, apareciendo muchos casos iguales en la mayoría de las muestras. Así que aunque los conjuntos de entrenamiento en S_t puedan ser diferentes unos de otros, no son independientes desde un punto de vista estadístico. Para asegurar la diversidad entre los miembros del ensamble, se debe utilizar un inductor relativamente inestable. Resultando en un ensamble de clasificadores suficientemente diferenciados que puede ser obtenido aplicando pequeñas perturbaciones al conjunto de entrenamiento. Si se utiliza un inductor estable, el ensamble estará compuesto por un conjunto de clasificadores que producen clasificaciones casi iguales, y por tanto, será improbable que se mejore la precisión del proceso [Rokach, 2010].

Para clasificar un nuevo caso, cada clasificador devuelve la clasificación de clase para el nuevo caso desconocido. El clasificador compuesto resultante del *bagging* devuelve la clase con el mayor número de predicciones (también conocido como el más votado) [Rokach, 2010] (Tabla 7.2).

Clasificación *bagging*

Requiere: x (el caso a clasificar)

Garantiza: C (clase predecida)

1: $Contador_{1,...,} Contador_{|dom(y)|} \leftarrow 0$ {inicia el contador de votos de clase}

2: **para** $i = 1$ a T

3: $voto_i \leftarrow M_i(x)$ {toma la clase predicha para del miembro i}

4: $Contador_{votoi} \leftarrow Contador_{votoi} + 1$ {aumenta 1 el contador de la clase correspondiente}

5: **final para**

6: $C \leftarrow$ la clase con el mayor número de votos

7: Se predice C

Tabla 7.2. Clasificación *bagging,* proceso de votación

7.1.2. El algoritmo *boosting*

El *boosting* es un método general para mejorar el rendimiento de un clasificador débil. El método funciona invocando iterativamente un clasificador débil, en datos de entrenamiento que se toma de varias distribuciones. De manera parecida al *bagging*, los clasificadores son generados por remuestreo del conjunto de entrenamiento. Los clasificadores son entonces combinados en un fuerte clasificador compuesto único. Al contrario que en el *bagging*, el mecanismo de remuestreo en el *boosting* mejora el muestreo en cada iteración para facilitar la muestra más útil para cada uno de los pasos del método [Rokach, 2010].

El algoritmo *boosting* viene descrito en la Tabla 7.3. Este algoritmo genera tres clasificadores. La muestra S_t, que es usada para entrenar el primer clasificador, M_1, es seleccionada aleatoriamente del conjunto de entrenamiento inicial. El segundo clasificador, M_2, se entrena sobre una muestra, la mitad de la cual consiste en casos que están mal clasificados por M_1 y la otra mitad está compuesta por casos que están correctamente clasificados por M_2. El último clasificador M_3 entrena con los casos en los que los dos clasificadores anteriores discrepan. Para clasificar un nuevo caso, cada clasificador produce su clase predicha. El clasificador del ensamble devuelve la clase que tiene la mayoría de los votos [Rokach, 2010].

Entrenamiento *boosting*

Requiere: I (un inductor debil), S (el conjunto de entrenamiento original), k (el tamaño de muestra para el primer calsificador)

Garantiza: M_1, M_2, M_3

1: $S_1 \leftarrow$ Aleatoriamente seleccionados $k < m$ casos de S sin reemplazamiento;

2: $M_1 \leftarrow I(S_1)$

3: $S_2 \leftarrow$ Aleatoriamente seleccionados casos (sin reemplazamiento) de $S - S_1$ de manera que la mitad de ellos están correctamente clasificados por M_1

4: $M_2 \leftarrow I(S_2)$

5: $S_3 \leftarrow$ cualquier caso en $S - S_1 - S_2$ que sea clasificado diferente por M_1 y M_2

Tabla 7.3. El algoritmo *boosting*

7.1.3. El algoritmo *AdaBoost*

Adaboost (del inglés, Adaptative *Boosting),* es un popular algoritmo de ensamblaje que mejora el algoritmo *boosting* mediante un proceso iterativo. La principal idea detrás de este algoritmo es centrarse más en los patrones difíciles de clasificar. La cantidad de atención que requiere cada patrón es cuantificada por el peso que es asignado a cada patrón en el conjunto de entrenamiento. Inicialmente, se asigna el mismo peso a todos los patrones. En cada iteración los pesos de todos los casos mal clasificados aumentan mientras que los pesos de los casos clasificados correctamente decrecen. Como consecuencia, el clasificador débil es forzado a centrarse en los casos difíciles del conjunto de entrenamiento llevando a cabo iteraciones adicionales y creando más clasificadores. Además, un peso es asignado a cada clasificador individual. Este peso mide la *precisión global* del clasificador y es una función del peso total de los patrones correctamente clasificados. Por lo tanto, se dan mayores pesos a los clasificadores más precisos. Estos pesos son usados para clasificar nuevos patrones [Rokach, 2010].

El pseudo-código del algoritmo *AdaBoos*t se encuentra en la Tabla 7.4. El algoritmo asume que el conjunto de entrenamiento consiste en m casos, los cuales son etiquetados como -1 o +1. La clasificación de un nuevo caso se obtiene votando todos los clasificadores $\{M_t\}$, cada uno teniendo una precisión global de α_t. Matemáticamente puede ser escrito como:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t \times M_t(x)) \qquad \textbf{[3]}$$

Entrenamiento *AdaBoost*

Requiere: I (un inductor débil), T (el número de iteraciones), S (el conjunto de entrenamiento original)

Garantiza: M_t, α_t, $t = 1, \ldots, T$

1: $S_1 \leftarrow$ Aleatoriamente seleccionados $k < m$ casos de S sin reemplazamiento;

2: $D_1(i) \leftarrow 1/m;\ i = 1, \ldots, m$

3: **Repetir**

4: Construir clasificador M_t usando I y distribución D_t

5: $\varepsilon_t \leftarrow \sum_{i:M_t(x_i)\neq y_i} D_i\ (i)$

6: **si** $\varepsilon_t > 0.5$ **entonces**

7: T ← t - 1

8: salir del bucle

9: **acabar si**

10: $\alpha_t \leftarrow \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

11: $D_{t+1}(i) = D_t(i) \times e^{-\alpha_t y_t M_t(x_i)}$

12: Normaliza D_{t+1} para ser una distribución adecuada

13: $t \leftarrow t + 1$

14: **hasta** $t > T$

Tabla 7.4. Entrenamiento *AdaBoost*

7.2. Caso de estudio: algoritmo EBARAMA

En nuestro caso, mediante el método de ensamblaje se ha realizado un remuestreo de las técnicas de aprendizaje supervisadas aplicadas anteriormente. Este proceso se ha llevado a cabo mediante la implementación del algoritmo EBARAMA (del inglés, *Ensemble Boosting Algorithm to Resampling All Methods Applied*). El pseudo-código del algoritmo EBARAMA se encuentra en la Tabla 7.5.

1: Muestreo al azar del 20-50% del tamaño total de la base de datos [Diettrich, 2000; Mallapragada *et al.,* 2009].

En nuestro caso tomamos n = 60 individuos que representan un 35% del tamaño total de la base de datos (n = 210)

2: Fijamos el número de iteraciones, *T*, del proceso

En nuestro caso *T* = 10.

3: **Repetir**

4: Sorteo de la técnica de clasificación a elegir: árbol – regla.

El sorteo se realizará inversamente proporcional al índice de Kappa medio obtenido por cada uno de estas técnicas.

5: Guardamos predicciones.

Los errores de las predicciones corrigen el próximo muestreo

Damos más peso para que aparezcan muestreados individuos peor clasificados; en nuestro caso este peso será el mismo para todos los individuos mal clasificados y otro para los que lo estén correctamente, dado que no existe un grado de error en la predicción sino una variable dicotómica 0-1.

6: **Si** *T* < número de iteraciones fijadas **entonces**

7: actualizar pesos y volver a muestrear (volver al paso 3)

8: **acabar si**

9: *T* = número de iteraciones fijadas

10: votamos las predicciones guardadas

Tabla 7.5. Algoritmo EBARAMA

Este proceso conlleva un doble remuestreo. Primero en la técnica de *Data Mining* a utilizar en cada iteración y después en los individuos con los que se trabaja.

En el primer de los casos los pesos, a_1 y a_2, vienen dados por la expresión:

$$a_1 = (1-K_1)/(2-K_1-K_2) \quad \textbf{[4]}$$

$$a_2 = (1-K_2)/(2-K_1-K_2) \quad \textbf{[5]}$$

donde K_1 y K_2 son índices de Kappa promedio de árboles y reglas de clasificación, respectivamente.

En el proceso de remuestreo integrado en el primero, donde el objetivo de remuestreo son los individuos a los que aplicar el algoritmo, se definen los pesos, b_1 y b_2, de la siguiente manera:

$$b_1 = 1/2\mathrm{E} \quad \textbf{[6]}$$

$$b_2 = 1/2(n\text{-E}) \quad \textbf{[7]}$$

donde E es el número de errores en la predicción de la iteración anterior y n el tamaño total de la base de datos

7.2.1. Resultados: algoritmo EBARAMA

El proceso de EBARAMA se inicia dando pesos equiprobables a los dos métodos de análisis de datos que se desea conjuntar: regla y árbol de clasificación. La primera muestra de 60 individuos (35% del total) es obtenida de manera aleatoria (pesos equiprobables para todos los individuos). Después se sigue el proceso descrito anteriormente en la Tabla 7.5, finalizando el mismo con una votación de las clasificaciones de los individuos, clasificación que determina la mayoría. En todos los casos se usó la validación cruzada, dividiendo en 10 tramos la muestra analizada en cada iteración. La tabla de la evolución del proceso de remuestreo es la siguiente:

Estadístico Kappa	Método	1-k
0.3466	Regla	0.6534
0.4724	Árbol	0.5276
0.6822	Regla	0.3178
0.5642	Árbol	0.4358
0.5098	Árbol	0.4902
0.5622	Árbol	0.4378
0.6822	Regla	0.3178
0.6012	Regla	0.3988
0.6664	Árbol	0.3336
0.7053	Árbol	0.2947

Tabla 7.6. Tabla de la evolución del proceso de remuestreo de EBARAMA

En el gráfico que se presenta en la Figura 7.2 se puede observar mejor su tendencia a decrecer, sin embargo (y como es de esperar) con el número de iteraciones el proceso tiende a estabilizarse. Quizá con un mayor número de iteraciones del proceso se hubiera podido observar más claramente esta característica.

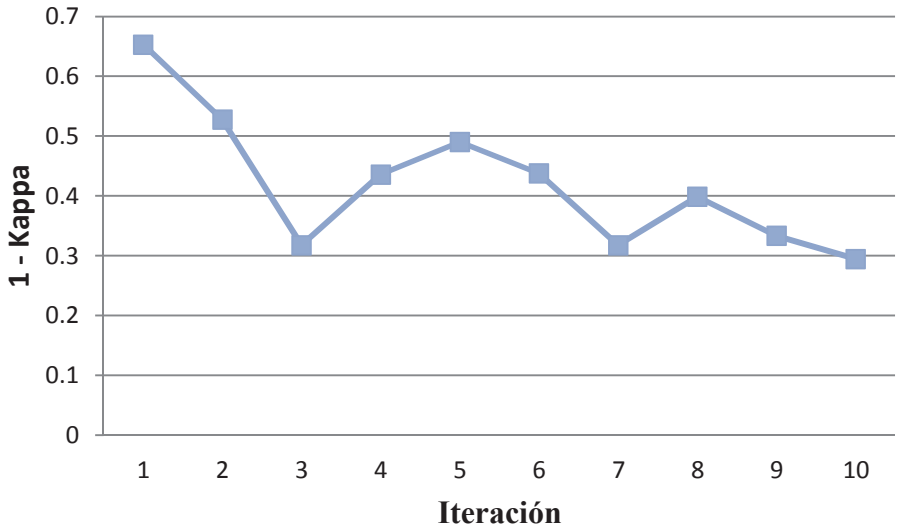

Figura 7.2. Gráfica de la evolución del proceso de remuestreo de EBARAMA

Tras realizar el paso final de votación (Tabla 7.7) se obtiene un índice Kappa de 0.73 muy por encima del máximo obtenido en los métodos de aprendizaje supervisado (0.56). De este modo se pasa de tener un grado de acuerdo moderado, a tener un grado de acuerdo bueno, mejorando notablemente el modelo. En este caso, se consigue clasificar correctamente el 85% de los casos.

Casos clasificados correctamente	178	85%
Casos clasificados incorrectamente	32	15%
Estadístico Kappa final	0,7304	

Tabla 7.7. Resultados de EBARAMA

Como resumen de los resultados obtenido se tienen algunas reglas, de las que destacan, por su mayor aparición en la base de datos, las presentadas en la Tabla 7.8.

Biofilm = Bajo
Edad de la tubería = Joven & Velocidad = Baja & Edad del agua = Media
Edad de la tubería = Joven & Velocidad = Alta
Biofilm = Medio
Edad de la tubería = Joven & Velocidad = Baja & Edad del agua = Baja
Edad de la tubería = Vieja
Biofilm = Alto
Edad de la tubería = Media & Velocidad = Alta

Tabla 7.8. Reglas obtenidas en EBARAMA

Según los resultados obtenidos, con un acierto del 83% en la fase test, se da un desarrollo de biofilm bajo en tuberías jóvenes, que bien tienen una velocidad de flujo baja y una edad del agua media, o una alta velocidad de flujo. El segundo caso refuerza los resultados encontrados anteriormente (ver Capítulo 6), donde se observaba que las altas velocidades favorecían un desarrollo de biofilm bajo en condiciones donde la rugosidad de las paredes de las tuberías era baja, como es el caso de las tuberías jóvenes. Respecto al primer caso que predice un desarrollo de biofilm bajo, se sabe, según la bibliografía, que tanto la edad de tubería baja (por la baja rugosidad de la pared de la tuberías [Chowdury, 2011]), como la velocidad de flujo baja (por la reducida difusión de los nutrientes [Lehtola *et al*., 2006]), favorecen un desarrollo de biofilm bajo, como se observa en los resultados. Sin embargo, según la bibliografía a mayor edad del agua más favorecido se vería el desarrollo de biofilm [EPA, 2002], por lo que en este caso, donde el desarrollo de biofilm es bajo, se esperaría encontrar una edad del agua baja, en vez de una edad del agua media como es el caso. Esta contradicción que se observa, se explica por el hecho de que en la base de datos estudiada todos los casos que presentan una edad del agua media, también presentan una velocidad de flujo baja. Este sesgo que se encuentra en la base de datos, puede deberse, en cierto grado, a que cuanto menor sea la velocidad a la que circula el agua, mayor será el tiempo que esta

permanezca en el sistema, por lo que ambas variables pueden encontrarse parcialmente correlacionadas. El hecho de que prevalezca el efecto de la baja velocidad sobre el de la mediana edad del agua, hace entender que, en este caso, la velocidad del flujo es más relevante para el desarrollo de biofilm que la edad del agua.

Respecto al desarrollo de biofilm medio, se ha obtenido un 87% de aciertos en la fase de test, en los casos donde, o bien, la tubería es joven y la velocidad de flujo y la edad del agua son bajas, o la edad de la tubería es alta. En este caso, los procesos que ocurren dentro de la tubería no son tan fáciles de explicar. El primer caso, sólo se diferencia del caso anterior, donde el desarrollo de biofilm es bajo, en que la edad del agua es baja en vez de media. Sin embargo, como se ha explicado anteriormente, según la bibliografía [EPA, 2002], la edad del agua baja debería favorecer un desarrollo de biofilm bajo, contrariamente a lo que parece ocurrir en este caso. Este efecto contradictorio de la edad del agua respecto al biofilm ya se observó en los análisis anteriormente realizados (Capítulo 6). El hecho de que todas las tuberías viejas se clasifiquen con un desarrollo de biofilm medio, entra dentro de lo esperado, ya que con la edad aumenta la rugosidad interna de las tuberías y la rugosidad favorece el desarrollo de biofilm [Christensen, 2009]. Sin embargo, en este caso, también se esperaría que se diese un elevado desarrollo de biofilm. Es posible que estos casos no se encuentren, debido a que la interacción de las diferentes variables confunda este efecto.

Por último, se ha observado que un 100% de aciertos en la fase de test, que donde la velocidad de flujo es baja y la edad de la tubería es media, el desarrollo de biofilm es alto. Como ya se ha explicado anteriormente, la edad de la tubería está relacionada con la rugosidad de la pared interna de las tuberías y esta, con el desarrollo de biofilm. Respecto a la velocidad de flujo se esperaría un mayor desarrollo de biofilm cuanto mayor fuese esta. Sin embargo, puede que al presentar una velocidad de flujo baja, la fuerza de arrastre del agua se vea muy reducida, favoreciendo mucho el desarrollo de biofilm. Este resultado nos muestra que la interacción entre estas variables favorece el aumento del desarrollo de biofilm, por lo que será de especial importancia tener en cuenta los casos en que se den estas condiciones.

Las clasificaciones obtenidas no predicen correctamente en las siguientes proporciones: cuando el desarrollo de biofilm es bajo el 0.05% de las veces, cuando es medio el 0.1% y cuando es alto el 50%. Aunque los datos de biofilm medio y bajo son muy buenos, los de desarrollo de biofilm alto aun deben de ser mejorados. Esta diferencia en el porcentaje de

acierto según las diferentes clasificaciones de desarrollo de biofilm probablemente se deba a que estas no se encuentran igualmente representadas en la base de datos, ya que los casos con un alto desarrollo de biofilm solo representan el 5 % de la misma.

Además de las conclusiones respecto al desarrollo de biofilm, cabe destacar el importante papel de la edad de la tubería en la clasificación final de EBARAMA, presente en todas sus conclusiones más importantes. Así como el de la velocidad de flujo, que también aparece en la mayoría de las reglas. El hecho de que el tipo de material parezca no tener un efecto significativo en el desarrollo de biofilm, a pesar de las diferentes rugosidades de los mismos, puede que se explique por el hecho de que las incrustaciones y depósitos que se acumulan con el tiempo produzcan cambios en las rugosidades internas de las tuberías más relevantes que las diferencias de rugosidad iniciales de los materiales. El tipo de régimen hidráulico, turbulento o laminar, tampoco parece ser especialmente influyente, respecto al resto de variables, en el desarrollo de biofilm en los DWDSs.

Capítulo 8

Conclusiones

Este trabajo se caracteriza por ofrecer una perspectiva de trabajo innovadora en el estudio de los biofilms en los DWDSs.

Por un lado, se introducen técnicas de análisis inteligente de datos en el estudio del desarrollo de biofilms en los DWDSs, como son el KDD y las técnicas de remuestreo y meta-learning. Aunque el proceso de descubrimiento de la información a través de los datos ya se ha utilizado anteriormente en trabajos relacionados con los DWDSs, es la primera vez que se emplea en el estudio de los biofilms en los abastecimientos de agua. El empleo de esta nueva herramienta ha permitido mejorar el nivel de conocimiento alcanzado sobre el desarrollo de biofilms en los DWDSs de una manera práctica y eficiente. Del mismo modo, con la introducción del algoritmo de ensamblaje EBARAMA se ha conseguido aumentar la precisión y robustez final de los resultados.

Por otro lado, se estudia el efecto de la interacción del conjunto de las características físicas e hidráulicas de los DWDSs relevantes en el desarrollo de los biofilms. Actualmente, debido a la complejidad de la comunidad y del ambiente estudiados, en la mayoría de los casos, el efecto de las diferentes características físicas e hidráulicas de los DWDSs sobre el desarrollo de biofilms se estudia individualmente. Sin embargo, el propósito de este trabajo es lograr una comprensión más profunda de las consecuencias que la interacción de estos factores de los DWDSs tiene en el desarrollo de biofilms. Por este motivo, se trata del primer trabajo que estudia conjuntamente un número tan elevado de variables en relación al desarrollo de biofilm en los DWDSs, ya que hasta ahora sólo se han estudiado relaciones de variables sueltas en su comportamiento, pero no en su conjunto, ni la interacción entre ellas. Esto supone un gran avance en el estudio del desarrollo de biofilms en los DWDSs ya que permite profundizar en el conocimiento de la ecología de estas comunidades y facilita una mejor comprensión de los

procesos e interacciones que se producen en los sistemas de distribución en relación al desarrollo de estas comunidades.

De esta manera se ha conseguido identificar qué tuberías serán propensas al desarrollo de biofilms en función de las características físicas e hidráulicas de las mismas. Sin embargo, durante la realización de este trabajo, aparte de su contribución principal, también se han conseguido otras aportaciones de interés en este ámbito de estudio.

- Se ha profundizado en el conocimiento de los biofilms, los factores que determinan su desarrollo en los DWDSs y los problemas asociados a su presencia en estos sistemas. Esto se ha conseguido mediante la recopilación y puesta al día de la información existente en relación al desarrollo de biofilms en los DWDSs, abarcando tanto los aspectos microbiológicos como los relacionados con la hidráulica de los sistemas. De esta manera se ha conseguido asimilar el conocimiento alcanzado en relación al comportamiento de los biofilms en los DWDSs, unificar criterios y determinar cuáles son las principales características, físicas e hidráulicas, de los DWDSs que afectan a su desarrollo.

- Se ha generado una base de datos completa y extensa mediante la recopilación de datos obtenidos de diferentes fuentes (estudios de cuantificación de biofilms en DWDSs reales o simulados en laboratorio), en una primera etapa, y la aplicación posterior de técnicas de pre-procesamiento y transformación de los datos. De esta manera, se ha conseguido obtener una base de datos completa sobre el desarrollo de biofilms en DWDSs y las características físicas e hidráulicas de los mismos. Esto, hasta ahora, no se había realizado debido a la complejidad del ambiente estudiado y a la utilización de diferentes metodologías y reactores de crecimiento de biofilm. Sin embargo, el trabajo de pre-proceso llevado a cabo, enmarcado en el adecuado proceso de KDD que se ha realizado en este trabajo lo ha hecho posible.

- Se ha resaltado la importancia y efectividad de las técnicas de visualización como herramienta para el estudio exploratorio de los metadatos obtenidos, observar si existe algún tipo de inconsistencia en la base de datos y comprobar su fiabilidad. Tras este análisis se han encontrado patrones interesantes como la especial influencia de la velocidad de flujo en el desarrollo de biofilms en los DWDSs. La visualización

también ha servido para observar que la respuesta del biofilm frente a una variable varía según el valor que tomen el resto de variables, indicando que el presente trabajo (estudio del efecto que la interacción de las diferentes características de los DWDSs tienen en el desarrollo de biofilms) es necesario.

- Se han identificado patrones de comportamiento de los biofilms en los DWDSs, en función de las características estudiadas, mediante la aplicación de diferentes técnicas de minería de datos sobre los metadatos. En este caso, la velocidad de flujo ha vuelto a aparecer como variable de especial importancia, a tener en cuenta. La velocidad de flujo alta parece ser crítica para el crecimiento o desprendimiento de biofilm en función de los valores que tomen el resto de variables. Tras estos análisis, se observa que la velocidad de flujo alta tiende a favorecer el desprendimiento de biofilm en los casos en los que la rugosidad interior de la tubería no es elevada. Estos casos corresponden, principalmente, con edades de tubería bajas o materiales plásticos.

- Se ha desarrollado un novedoso algoritmo de remuestreo estratificado, EBARAMA, capaz de remuestrear no sólo basándose en los resultados de las clasificaciones sino también en el algoritmo utilizado en cada iteración. La aplicación de este algoritmo ha conseguido mejorar los resultados obtenidos anteriormente mediante las técnicas de minería de datos. Se ha pasado de tener un grado de acuerdo moderado (Kappa = 0.56) a obtener un grado de acuerdo bueno (Kappa = 0.73); obteniendo un porcentaje de casos correctamente clasificados del 85%. Tras este análisis, la variable edad de la tubería toma especial protagonismo, al igual que la velocidad de flujo. Sin embargo, los aciertos en las predicciones decrecen en gran media en los casos de alto desarrollo de biofilm. Probablemente esto ocurra por la baja representatividad que estos tienen en la base de datos. En cambio, el porcentaje de error en las predicciones para los casos de desarrollo de biofilm medio y bajo son excelentes, del 0.1% o menos.

8.1 Aplicaciones

Este trabajo supone la base para el desarrollo de una herramienta más compleja de ayuda a la toma de decisiones capaz de predecir qué condiciones de los DWDSs favorecen un alto desarrollo de biofilm y por tanto, ser capaz de identificar qué tuberías presentarán una mayor

tendencia a sufrir un elevado desarrollo de biofilm en su interior. De esta manera, se conseguirá llevar a cabo una gestión de la calidad del agua de los DWDSs más efectiva y se mitigarán de manera más eficiente los problemas derivados del desarrollo de biofilms en estos sistemas. Existe una gran cantidad de problemas relacionados con la gestión de los DWDSs que podrían verse mejorados gracias a esta herramienta.

Su aplicación podría ser sanitariamente relevante, ya que los lavados hidráulicos de tuberías podrían realizarse de manera dirigida (intensificando los esfuerzos en aquellas tuberías donde se sabe que por sus características tienden a sufrir un alto desarrollo de biofilm) con el fin de reducir el riesgo asociado al desarrollo de biofilms. Esta práctica además de con un propósito higiénico también puede llevarse a cabo como una operación de mantenimiento, con la intención de reducir los problemas relacionados con la biocorrosión, consecuencia del desarrollo de biofilms en las tuberías metálicas. Esta práctica podría reducir el número de incidencias y roturas en el sistema, mejorando el servicio y aumentando la satisfacción del consumidor. También supondría un ahorro importante de dinero para la agencia de servicio de agua, ya que, como se ha visto en el Capítulo 2 de este trabajo, la biocorrosión tiene una alta repercusión económica.

De la misma manera, disponer de este conocimiento (qué tuberías tienden a soportar un mayor desarrollo de biofilm en su interior) también puede servir como criterio adicional a la hora de la toma de decisiones en diferentes aspectos relacionados con la gestión de los DWDSs. Por un lado, puede ayudar a decidir la localización de los puntos de re-cloración en el sistema. De este modo, se lograría mantener elevadas concentraciones de desinfectante en el agua que circula por las tuberías más problemáticas y reducir el desarrollo de biofilm en su interior. Por otro lado, ayudaría a la toma de decisiones en relación al diseño de nuevos tramos de abastecimiento o de sistemas de distribución completos, siendo el posible grado de desarrollo de biofilm un criterio más a tener en cuenta en el diseño. Por ejemplo, el conocimiento adquirido sobre las consecuencias en el desarrollo de biofilm de la interacción de las diferentes características físicas e hidráulicas de los DWDSs puede influir en la elección de un tipo u otro de material de tubería en función de las condiciones de funcionamiento del sistema o viceversa. En los casos donde el sistema de distribución ya está en funcionamiento, sería posible condicionar, en cierto grado, sin afectar al suministro, el funcionamiento hidráulico del sistema en función de las características de la tubería ya instalada, de manera que el desarrollo de biofilm se viese reducido lo máximo posible.

Según los resultados obtenidos en este estudio destacan las siguientes recomendaciones a tener en cuenta por los gestores de los servicios de agua. Principalmente, se debería evitar, que en las tuberías con edades medias (entre 11 y 30 años) se diesen velocidades del flujo de agua lentas (entre 0 y 0.7 m/s). Como regla general, se recomienda que se centren los esfuerzos de monitoreo y control en aquellas tuberías con edades medias o elevadas (superiores a 10 años) y poner especial atención en los casos donde el flujo de agua sea lento.

8.2 Futuras líneas de investigación

Se ha comenzado a diseñar un modelo de formación de biofilms en los DWDSs mediante sistemas multi-agente (MASs, Multi-Agent Systems), en base a un número de interacciones entre bacterias, el sistema hidráulico y las características físicas de las tuberías, que sea capaz de representar la variación espacial y temporal de la cantidad de biofilm adherido, simulando su comportamiento [Ramos-Martínez et al., 2012a; 2012b].

Esta modelación de los biofilms en los DWDSs permitirá profundizar en el conocimiento de estas comunidades, comprobar hipótesis sobre sus mecanismos de funcionamiento y predecir su evolución en el tiempo.

De hecho, una vez alcanzado este objetivo se quiere profundizar en el efecto que supone el desarrollo de biofilms en el consumo de desinfectante y producción de THM en los DWDSs. Ya que, como se explica en el Capítulo 2 del presente trabajo, existe una relación directa entre el biofilm y estos procesos. Por lo que conociendo con exactitud el desarrollo de biofilm dentro de las tuberías de los DWDSs, también será posible aumentar la precisión de los modelos de consumo de desinfectante y producción de THM en estos sistemas. Por lo que este modelo facilitaría aun más la toma de decisiones en la gestión de los servicios de abastecimiento y la mejora la calidad del agua servida.

En futuras etapas se contempla, aumentar en la medida de lo posible, el número de casos de la base de datos, especialmente casos que correspondan a un desarrollo de biofilm alto. También se quiere incluir en el estudio las características físico-químicas del agua para poder comparar los niveles de desarrollo de biofilms entre diferentes DWDSs.

Diferentes Agencias de Agua y Universidades europeas se han interesado en los resultados del presente trabajo, poniendo a nuestra disposición los datos de los sistemas de abastecimiento con los que ellos trabajan para aplicar esta línea de investigación con sus datos. Esto representa un prometedor e interesante reto a enfrentar en un futuro próximo.

8.3 Difusión del trabajo

Tanto los resultados de este trabajo, como diferentes estudios derivados de este, han sido presentados en diferentes congresos a nivel internacional.

- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael 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. Nov., 2012 (aceptado).
- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael Pérez-García. ESTUDIO DEL DESARROLLO DE BIOFILMS EN TUBERÍAS MEDIANTE REDES BAYESIANAS CON VARIABLES MIXTAS. XXV Congreso Latinoamericano de Hidráulica. San José, Costa Rica. Sept., 2012.
- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael Pérez-García. ENSEMBLE OF MULTIPLE DATA MINING APPROCHES TO BIOFILM DEVELOPMENT IN DRINKING WATER DISTRIBUTION SYSTEMS. Mathematical Modelling in Engineering & Human Behaviour 2012. Valencia, España. Sept, 2012.
- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael Pérez-García. MODELING BIOFILMS FORMATION AND EVOLUTION IN DRINKING WATER DISTRIBUTION SYSTEMS USING A MULTIAGENT APPROACH 2nd Meeting of Young Researchers Modelling Biological Processes. Granada, España. Jul., 2012.
- Eva Ramos-Martínez, Manuel Herrera, Rafael Pérez-García, Joaquín Izquierdo. 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. Jul., 2012.

- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael Pérez-García. ASSESSING VARIATION IN BIOFILMS 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. Jun., 2012.
- Eva Ramos-Martínez, Manuel Herrera, Joaquín Izquierdo, Rafael Pérez-García. EVALUACIÓN DEL DESARROLLO DE BIOFILMS EN SISTEMAS DE ABASTECIMIENTO DE AGUA MEDIANTE REDES BAYESIANAS. X Seminario Euro Latinoamericano de Sistemas de Ingeniería. Valencia, España. Dic., 2011.

Bibliografía

Abbott, M. B., Babovic, V. M. & Cunge, J. A. *Towards the hydraulics of the hydroinformatics era*. J. Hydr. Research, Vol. 39(4), pp. 339-349, 2001.

Agència de salut pública de Barcelona. *Los trihalometanos (THM) en el agua de consumo.* Documento informativo.

Appenzeller, B. M. R., Batté, M., Mathieu, L., Block, J.C., Lahoussine, V., Cavard, J. & Gatel. D. *Effect of adding phospahte to drinking eater on bacterial growth in slightly and highly corroded pipes*. Water. Res. Vol. 35, pp. 1100-1105.

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

Babovic, V., Drescourt, J., Keijzer M. & Hansen., P.F. *A data mining approach to modeling of wáter supply assets*. Urban water, Vol. 4, pp. 401-414, 2002.

Bai X., Wu F., Zhou B. & Zhi X.: *Biofilm bacterial communities and abundance in a full-scale drinking water distribution system in Shanghai*. Journal of water and health, Vol. 8(3), 2010.

Batté, M., Koudjonou, B., Laurent, P., Mathieu, L., Coallier, J. & Prévost, M. *Biofilm responses to ageing and to a high phosphate load in a bench-scale drinking water system.* Water Research. Vol. 37, pp. 1351–1361, 2003.

Beech, I., & Gaylarde, C. C. *Recent advances in the study of biocorrosion - an overview*. Rev. Microbiol. Vol.30(3), São Paulo Jul./Sept. 1999.

Beech, I., Bergel, A., Mollica, A., Flemming, H. C., Scotto, V. & Sand, W. *Simple methods for the investigation of the role of biofilms in corrosión*. Biofilms Publication, Sept., 2000.

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press. 2005.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees*. Monterey, CA: Wadsworth, 1984.

Breiman, L. *Bagging Predictors*. Machine Learning, Vol. 24, pp. 123-140. 1996.

Brennenstuhl, A., Doherty, P., King, P., Dunstall, T. *The effects of biofouling on the corrosión of nickel heat exchanger alloys at Ontario Hydro.* Microbially Influenced Corrosion and Biodeterioration. Edited by N. Dowling, M. Mittlemanand J. Danko, Knoxvilles, Tennessee, Oct. 7 – 12, pp 4-25 -4-31, 1991.

Bruzzonea, L., Cossua, R., Vernazzab, G. *Detection of land-cover transitions by combining multidate classifiers*. Pattern Recognition Letters. Vol.25 (13), pp. 1491–1500, 2004.

Butterfield, P. W., Bargmeyer, A. M., Camper, A.K. & Biederman, J. A. *Modified enzyme activity assay to determine biofilm biomass,* J. of Microbiol. Methods, Vol. 50, pp. 23-31, 2002.

Bougadis, J., Adamowski, K. & Diduch, R., *Short-term municipal water demand forecasting*. Hydrological Processes. Vol. 19, pp. 137-148, 2005.

Camper, A. K., Jones, W. L. & Hayes, J. T. *Effect of growth conditions and substratum composition on the persistence of coliforms in mixed-population biofilms*. Appl. Environ. Microbiol. Vol. 62(11), pp. 4010-4018, 1996.

Camper, A. K. *Involvement of humic substances in regrowth* . Int. J. Food Microbiol. Vol. 92, pp. 355-364, 2004.

Carvajal, L. F., Gomez, A. & Ochoa, S. *Simulación de un lavado hidráulico en tuberías para el control del crecimiento de biopelícula*. Dyna rev.fac.nac.minas Vol. 74(152), Medellín May/Aug. 2007.

Curk, T., Demsar, J.., Xu, Q., Leban, U.P., Petrovic, U., Bratko, I, Shaulsky, G. & Zupan, B. *Microarray data mining with visual programming*. Bioinformatics. Vol. 21(3), pp. 396-398. 2005.

Chandy, J.P. & Angles, M.L. *Determination of nutrients limiting bioiflm formation and the subsequent impact on disinfectant decay*. Water Res. Vol. 35, pp 2677 – 2682, 2001.

Characklis, W.G. & Marshall, K.C. *Biofilms*. Ed. John Wiley & Sons, Inc., Nueva York 1990.

Chowdhury, S. *Heterotrophic bacteria in drinking water distribution system: a review*. Environmental Monitoring and Assessment, DOI 10.1007/s10661-011-2407-x, 2011.

Christensen Ryan T. *Age Effects on Iron-Based Pipes in Water Distribution Systems*. All Graduate Theses and Dissertations. Paper 505. Utah State University, 2009.

Cloete T.E, Westard D. & van Vuuren S.J.: *Dynamic response of biofilm to pipe surface and fluid velocity*. Water Science and Tecnology, Vol. 47 (5), pp. 362 57-9, 2003.

Chu, C. & Lu, C. *Effects of oxalic acid on the regrowth of heterotrophic bacteria in the distributed drinking water*. Chemosphere. Vol. 57, pp. 531–539, 2004.

Chu, C., Lu, C. & Lee, C. *Effects of inorganic nutrients on the regrowth of heterotrophic bacteria in drinking water distribution systems.* Journal of Environmental Management. Vol. 74, pp. 255–263, 2005.

Cleveland, W.S. & McGill, R. *The many faces of a scatterplot*. Journal of the american statistical association. Vol 79(338), pp. 807-822, 1984.

Codony, F., Miranda, A. M. & Mas, J. *Persistence and proliferation of some unicellular algae in drinking water systems as result of their heterotrophic metabolism*. Water SA Vol. 29(1), En. 2003.

Compton, P., Edwards, G., Kang, B., Malor, R., Menzies, T., Preston, P., Srinivasan, A. & Sammut, S. *Ripple down rules: possibilities and limitations. Boose,* J.H. & Gaines, B.R., Ed. Proceedings of the Sixth AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop. pp. 6-1-6-20. Calgary, Canada, University of Calgary.1991.

Cromwell III JE, Reynolds H, Pearson J. & Grant M. *Costs of Infrastructure Failure.* Denver, CO: AwwaRF. 2002.

Crozes, G.F. & Cushing, R.S. *Evaluating biological regrowth in distribution systems*. AWWARF report 1P-6C-90796-7/00-CM. AWWARF/AWWA Denver, CO. Subject área: distributin systems. 1993.

Dasgupta, S., *Performance Guarantees for Hierarchical Clustering*. 15th Annual Conference on Computational Learning Theory, pp. 351-363, 2002.

Dasgupta, S. & Long, P. M. *Performance guarantees for hierarchical clustering*. Preprint submitted to Elsevier Science 24 Julio 2010.

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

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

Díaz Arevalo, J.L. *Utilización de técnicas avanzadas en el tratamiento y manejo de datos. Aplicación a la gestión de sistemas de abastecimiento de agua*. Tesis doctoral, Universitat Politècnica de València, 2010.

Dietterich, T. G. *Emsemble Methods in Machine Learning*. Multiple clssifiers systems, LCBS-1857. Springer., pp. 1-15. 2000.

Dimopoulos, Y. & Kakas, A. *Learning non monotonic logic programs: Learning exceptions.* Nada Lavrac and Stefan Wrobel editors. Machine LearningECML-95 (Proc, European Conference on Machine Learning). Lecture Notes in Artificial Intelligence, Vol. 912, pp. 122-127. Springer Verlag, 1995.

Domingos, P. *Linear-time rule induction*. Proc. Second International Conference of Knowledge Discovery and Data Mining. Portland OR: AAAI Press. Pp. 96-101. 1996.

Donlan, R.M. *Biofilms: microbial life on surfaces*. Emerging Infect. Dis. Vol. 8, pp. 881-890, 2002.

Environment Agency, UK. 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.

Evins, C. *Safe Piped Water: Chapter 6. Small Animals in drinking Water Distribution Systems*. 2004.

Fayyad, U. M., Piaetsky-Shapiro, G. & Smyth, P. *From Data Mining to Knowledge Discovery: An Overview. In Advances In Knowledge Discovery and Data Mining.* AAAI/MIT press, Cambridge mass, 1996.

Flemming, H. C., Percival, S. L. & Walker, J. T. *Contamination potential of biofilms in wáter distribution systems.* Wáter Supply, Vol. 2(1), pp. 271-280, 2002.

Gauthier, V., Redercher, S. & Block, J. C. *Chlorine inactivation of Sphingomonas cells attached to goethite particles in drinking water*. Appl. Environ. Microbiol., Vol 65, pp. 355-357, 1999.

Gelves, M.F. *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, Bogota, Colombia, 2005.

Gibert, K. & Pérez-Bonilla A. *Ventajas de la estructura jerárquica del clustering en la interpretación automática de clasificaciones*. In Procs. III Taller Nacional de Minería de Datos y Aprendizaje (TAMIDA). I Congreso Español de Informática (CEDI'2005). Actas Digitales ISBN: 84-609-6891-X. Granada, España. 2005.

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

Gibert, K., Sánchez-Marré, M., Izquierdo, J, Chen, S., Rodríguez-Roda, I. & Holmes G. *Environmental Data Mining Trends, Methods and Challenges*, Environmental Modelling & Software, enviado, 2012.

Giráldez, R. *Mejoras en eficacia y eficiencia en algoritmos evolutivos para aprendizaje supervisado*. Memoria del periodo de investigación. Sevilla, Septiembre de 2003.

Hartigan, J. A. & Wong, M. A., *A k-means clustering algorithm*. Applied Statistics. Vol. 28, pp.100-108. 1979.

Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Second Edition. Springer. Feb. 2009.

Herrera, M. *Improving water network management by efficient division into supply clusters*. PhD in Hydraulic Engineering and Environmental Studies Valencia, Jul. 2011a.

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

Herrera, M., Izquierdo, J., Pérez-García, R. & Ayala, D. *El sistema de abastecimiento de agua en ciudades inteligentes.* SEREA, Seminario Hispano-Brasileño sobre Planificación, Proyecto y Operación de Redes de Abastecimiento de Agua Potable. Coimbra, Portugal. Julio, 2012a.

Herrera, M., Izquierdo, J., Pérez-García, R. & Montalvo, I. *Multi-agent adaptive boosting on semi-supervised water supply clusters*. Advances in Engineering Software. Vol. 50, pp. 131–136. 2012b

Hernández, J. , Ramírez, M.J. & Ferri, C. *Introducción a la minería de datos*. Ed. Pearson, 2004.

Hochbaum, D.S., Shmoys, D.B., *A best possible heuristic for the k-center problema*. Mathematics of Operations Research. Vol. 10(2), pp. 180-184, 1985.

Imran, M., Sadiq, R. & Kleiner, Y. *Identifying research needs related to impacts of water quality on the integrity of distribution infrastructure*. National research council Canada, NRCC-48703. 2006.

Izquierdo, J., Escribano, A., Iglesias, P. L. & Díaz, J. L. *Predimensionado de calderines antiariete mediante una red neuronal*. SEREA, Seminario Hispano-Brasileño sobre Planificación, Proyecto y Operación de Redes de Abastecimiento de Agua Potable. Universidad Politécnica de Valencia (Spain). December, 2002.

Izquierdo, J., Díaz, J. L. & Pérez-García, R. *Knowledge Discovery in Environmental Data*. Integrated Water Management Series: NATO Science series: IV: Earth and Environmental Sciences; Springer, pp. 51 - 68, 2008.

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

Jarvis, W.R. *Opportunistic pathogenic microorganisms in biofilms*. Center for Disease Control, Washington D.C., 1990.

Kalmbach, S., Manz, W. & Szewzyk U. *Development of a new method to determine the metabolic potential of bacteria in drinking water bioifilms: probe active counts (PAC*). In Biofilms: investigative methods and applications, pp. 107-121, Technomic Publishing, Lancaster, 2000.

Kaufman,L. & Rousseeauw,P.J. *Clustering by means of medoids*. In Dodge,Y. (ed.), Statistical Data Analysis Based on the L1-norm and Related Methods. North Holland, Amsterdam, pp. 405–416.1987.

Kaufman, L. & Rousseeauw, P. J., *Finding groups in data: an introduction to cluster analysis*. Wiley. 1990.

Keinänen, M.M., Korhonen, L.K., Lehtola, M.J., Miettinen, I.T., Martikainen, P.J., Vartiainen, T. & Suutari, M.H. *The microbial community structure of drinking water biofilms can be affected by phosphorus availability*. Appl. Environ. Microbiol. Vol. 68, pp. 434-439. 2002.

Kowalska, B., Kowalski, D. & Musz, A. *Chlorine decay in wáter distribution system*s. Environment Protection Engineering, vol. 32 (2), 2006.

Landis, J.R. & Koch, G.G., *The measurement of observed agreement for categorical data.* Biometrics, Vol. 33, pp. 159-174, Mar. 1977.

Langmark,J., Storey, M.V., Ashbolt, N.J. & Stenstrom. T.A. *The effects of UV disinfection on distribution pipe biofilm growth and pathogen incidence within the greater Stockholm area, Sweden*. Water Research Vol. 41, pp. 3327 – 3336, 2007.

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

LeChevallier, M.W., Seidler, R.J. & Evans, T.M. *Enumeration and characterization of standard plate count bacteria in chlorinated and raw water supplies*. Appl. Environ. Microbiol., Vol. 40, pp. 922–930, 1980.

Lee, K.C., Rittmann, B.E. *Effects of pH and precipitation on autohydrogenotrophic denitrification using the hollow-fiber membrane-biofilm reactor*. Water Research Vol. 37, pp. 15551-1556, 2003

Lehtola, M.J., Juhna, T., Miettinen, I.T., Vartiainen, T. & Martikainen, P.J. *Formation of biofilms in drinking water distribution networks,a case study in two cities in Finland and Latvia* . J Ind Microbiol Biotechnol, Vol. 31, pp. 489–494. DOI 10.1007/s10295-004-0173-2, 2004a.

Lehtola, M.J., Miettinen, I.T., Keinänen, M.M., Kekki, T.K., Laine, O., Hirvonen, A., Vartiainen, T. & Martikainen, P.J. *Microbiology, chemistry and biofilm development in a pilot drinking water distribution system with copper and plastic pipes*. Water Research Vol. 38, pp. 3769–3779, 2004b.

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

Leigh, W., Purvisb, R. & Ragusaa, J. M. *Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support*. Decision Support Systems. Vol. 32(4), pp. 361–377. 2002.

Le Puil, M. *Biostability in Drinking Water Distribution Systems Study at Pilot-Scale*. A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Civil and Environmental Engineering in the College of Engineering and Computer Science at the University of Central Florida Orlando, Florida. 2004

Levi, Y. *Safe Piped Water: Managing Microbial Water Quality in Piped Distribution* Systems. *Minimizing potential for changes in microbial quality of treated water*. Edited by Richard Ainsworth. ISBN: 1 84339 039 6. Published by IWA Publishing, London, UK. 2004.

Likas, A., Vlassis, N. & Vebeek, J. L., *The global k-means clustering algorithm*. Pattern Recognition Vol. 28, pp. 451-461. 2003.

Lingireddy, S. & Ormsbee, L. E., *Neural Networks in Optimal Calibration of Water Distribution Systems*. Vol. 3. Flood and N. Kartam (ASCE). 1973.

Little, B. J. & Jason S. L. *Microbiologically Influenced Corrosion*. DOI: 10.1002/0471238961.micrlitt.a01. 2009.

Lopes, F.A., Morin, P., Oliveira, R. & Melo, L.F. *Impact of biofilms in simulated drinking water and urban heat supply systems.* International Journal of Environmental Engineering, Vol. 1(3), 2009.

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

Lu, W., Kiéné, L. & Lévi, Y. *Chlorine demand of biofilms in wáter distribution systems*. Water Research, Vol. 33 (3), pp 827 – 835. 1998.

Lu, C.S. & Chu, C.H. *Effects of acetic acido n the regrowth of heterotrophic bacteria in the drinking water distribution system*. World J. Microbiol. Biotechnol. Vol. 21, pp. 989 – 998. 2005.

Mains, C. *Biofilm control in distribution systems*. Biofilm Control in Distribution Systems, Vol. 8(2), 2008.

Mallapragada, P. K , Jin, R, Jain, A. K., Liu, Y. *SemiBoost: boosting for semi-supervised learning*. IEEE Trans Pattern Anal Mach Intell. Vol. 31(11), pp. 2000-14. 2009.

Mangiamelia, P., Westb, D. & Rampalc, R. *Model selection for medical diagnosis decisión support systems*. Decision Support Systems. Vol. 36(3), pp. 247–259. 2004.

Manuel, C.M., Nunes, O.C. & Melo, L.F. *Unsteady state flow and stagnation in distribution systems affect the biological stability of drinking water*. Biofouling Vol. 26(2), pp. 129-139, 2010.

Martin, B. *Instance-Based learning: Nearest Neighbor With Generalization*. Hamilton, New Zealand, 1995.

Meir, R. & Rätsch, G. *An Introduction to Boosting and Leveraging*. Advanced Lectures on Machine Learning, Springer, 2003. Review Paper.

Millot, J., Rodríguez, M. J. & Sérodes, J. B. *Contribution of Neural Networks for Modelling Trihalomethanes Occurrence in Drinking Water*. J. Water Res. Plan. and Mgmnt. Vol. 128(5), pp. 370-376, 2002.

Molina, J.M. & García, J. *Técnicas de Análisis de Datos: Aplicaciones prácticas utilizando Microsoft, Excel y Weka.* Universidad Carlos III de Madrid, 2006.

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

Momba , M.N.B. & Binda M.A. *Combining chlorination and chloramination processes for the inhibition of bio®lm formation in drinking surface water system models*. Journal of Applied Microbiology, Vol. 92, pp. 641-648, 2002.

Morris G. K., Patton C. M., Feeley J. C., Johnson S. E., Gorman G., Martin W. T., Skaliy P., Mallison G. F., Politi B. D. & Mackel D. C. *Isolation of the Legionnaires' disease bacterium from environmental samples*. Ann Intern Med. Vol. 90(4), pp. 664–666, 1979.

Morton, S.C. & Edwards, M. *Reduced phosphorus compounds in the environment.* Crit. Rev. Environ. Sci. Technol. Vol. 35, pp. 333-364. 2005

Muñoz, A., Craik, S., & Kresta, S. *Computational fuid dynamics for predicting performance of ultraviolet disinfection - sensitivity to particle tracking inputs*. Journal of Environmental Engineering and Science, Vol. 6(3), pp. 285-301, 2007.

NACE. *Corrosion costs and preventive strategies in the United States*.

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

Nieto, L. & Saldarriaga, J.G. *Eventos de coloración del agua potable como consecuencia del desprendimiento de biopelículas: el caso de Bogotá D.C*. Universidad de los Andes. Enero, 2009.

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

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

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

Obst, U. & Schwartz, T. *Microbial Characteristics of Water Distribution: Compiled Investigations in a German Drinking Water Distribution System*. DOI: 10.1061/(ASCE)1090-025X(200)11:2(78). Abril, 2007.

Ollos, P. J., Huck, P. M. & Slawson, R. M. *Factors affectiong biofilm accumulation in model distribution systems*. J. Am. Water Works Assoc. Vol. 95(1), pp. 87-97, 2003.

Olson, B.H. *Assesment and implications of bacterial regrowth in water distribution systems*. EPA-600/52-82-072. US Environmental Protection Agency, Cincinnati, OH. 1982.

Parsek, M. R. & Fuqua, C. *Biofilms: Emerging themes and challenges in studies of surface-associated microbial life*. J. Bacteriol. Vol. 186, pp.4427-4440. 2003.

Piera, G. *Estudio del biofilm: formación y consecuencias*. Escola de Prevenció i Seguretat Integral, 2002-2003.

Pozos, N., Scow, K., Wuertz, S. & Darby, J. *UV disinfection in a model distribution system: biofilm growth and microbial community*. Water Research, Vol. 38, pp. 3083–3091, 2004.

Qin, X. *Biofilms in drinking water distribution systems*. A thesis submitted for the deggree of Doctor of philosophy by the University of Hong Kong. Enero 2009.

Ramírez, F. *Desinfección del agua con cloro y cloraminas*. Técnica Industrial, Vol. 260. Diciembre, 2005.

Ramos-Martínez, E., Herrera, M., Izquierdo, J., Pérez-García, R. *Modeling biofilms formation and evolution in drinking water distribution systems using a mutiagent approch*. 2nd Meeting of Young Researchers Modelling Biological Processes. Granda, España. Jul., 2012a.

Ramos-Martínez, E., Herrera, M., Izquierdo, J., Pérez-García, R. *Biofilms en los sistemas de distribución de agua potable: Aproximación basada en sistemas multi-agentes.* XI Seminario Euro Latinoamericano de Sistemas de ingeniería. La Habana, Cuba. Nov., 2012b (aceptado).

Ridway, H. F. & Olso, B.H. *Chlorine resistance patterns of bacteria from two drinking water distribution systems*. Appl. Environ. MIcrobiol. Vol. 41, pp. 274-287, 1982.

Rokach, L. Pattern Classification Using Ensemble Methods. Series inMachine Perception Artificial Inteligence, Vol. 75. 2010.

Rompré A. *Evolution of bacterilogical parameters in drinking water distribution networks*. M. Sc. Thesis Civil Engineering-Environment. Ecole Polytechnique de Montréal. 1993.

Scheffer, T. *Algebraic Foundation and Improved Methods of Induction of Ripple Down Rules*. Proc. Pacific Knowledge AcquisitionWorkshop, Sydney, 1996.

Shawe-Taylor, J. Bartlett, P.L. ; Williamson, R.C. & Anthony, M. *Structural risk minimization over data-dependent hierarchies*. IEEE Transactions on Information Theory. Vol 4 (5), pp. 1926-1940. 1998.

Shi, H. *Best-first decision tree learning*. Hamilton, NZ, 2007.

Schwartz, T., Hoffmann & Obst, U. *Formation of natural biofilms during chlorine dioxide and U.V. disinfection in a public drinking water distribution system*. Journal of Applied Microbiology, Vol. 95, pp. 591–601, 2003.

Seo, Y., *Monitoring the Role of Biofilm Biopolymers against Disinfectants in Water Distribution Systems*. Report as of FY2009 for 2009OH89B, 2009.

Sierra, B. *Aprendizaje automático: conceptos básicos y avanzados*. Pearson Prentice Hall, ISBN: 10: 84-8322-318-X, 2006.

Silhan, J. , Corfitzen, C.B. & Albrechtsen H.J. *Effect of temperature and pipe material on biofilm formation and survival of Escherichia coli in used drinking water pipes: a laboratory-based study*. Water Science & Technology Vol 54(3), pp 49–56, 2006.

Simoes, M., Pereira, M.O., Sillankova, S., Azeredo, J. & Vieira, M.J. *The effect of hydrodynamic conditions on the phenotype of Pseudomonas fluorescens biofilms*. Biofouling: The Journal of Bioadhesion and Biofilm Research. Vol. 23(4), 2007.

Simoes, L.C., Azevedo, N., Pacheco, A., Keevil, C.W. & Vieira, M.J. *Drinking water biofilm assessment of total and culturable bacteria under different operating conditions*, Biofouling, Vol. 22(2), pp. 91 – 99, 2006.

Smith-Somerville, H. E., V. B. Huryn, C. Walker & A. L. Winters. *Survival of Legionella pneumophila in the cold-water ciliate Tetrahymena vorax*. Appl. Environ. Microbiol. Vol. 57, pp.2742-2749. 1991.

Steirnet, M., Birkness, K., White, E., Fields, B. & Quinn, F. *Mycobacterium avium Bacilli Grow Saprozoically in Coculture with Acanthamoeba polyphaga and Survive within Cyst Walls.* App. And Environ. Microb. Vol 63(6), pp 2256 – 2261, 1998.

Storey, M.V. & Ashbolt, N.J. *A comparison of methods and models for the analysis of water distribution pipe biofilms*. International Water Association. World water congress No2, Berlin , ALLEMAGNE (15/10/2001) , Vol. 2(4), pp. 73-80[Note(s) : VI, 238 p.,] [Document : 8 p.] (1 p.1/4) ISBN 1-84339-427-8. 2002.

Sylvestry-Rodriguez, N., Bright, K.R., slack, D.C., Uhlmann, D.R. & Gerba, C.P. *Silver as a residual disinfectant to prevent biofilm formation in water distribution systems*. Appl. Environ. Microbiol., Vol. 74(5) pp.1639. DOI: 10.1128/AEM.02237-07. 2008.

Szewzyk, U., Szewzyk, R., Manz, W. & Schleifer, K.H. *Microbiological safety of drinking water*. Ann. Rev. Microbiol. Vol. 54, pp. 81–127, 2000.

Tan, A C & Gilbert, D. *Ensemble machine learning on gene expression data for cancer classification*. Proceedings of New Zealand Bioinformatics Conference, Te Papa, Wellington, New Zealand, ,13-14 February 2003.

Taylor, R. H. *Disinfectant Susceptibility of Mycobacterium avium*. Virginia Polytechnic Institute and State University, Blacksburg, Virginia. Dic. 4, 1998

Telghmann, U., Horn, H., and Morgenroth, E. *Influence of growth history on sloughing and erosion from biofilms*. Water Res. Vol. 38, pp. 3671-3684. 2004.

Thasibisile, P., *Water quality Decay and Pathogen Survival in Drinking Water Distribution Systems*. PhD partial fulfillment. Arizona State University, 2010.

Tsai, Y.P., Pai, T.Y., Qiu, J.M. *The impacts of the AOC concentration on biofilm formation under higher shear force condition* . Journal of Biotechnology Vol. 111, pp. 155–167, 2004

Tsai, Y.P. *Impact of flow velocity on the dynamic behaviour of biofilm bacteria.* Biofouling, Vol. 21(5/6), pp. 267-277, 2005.

Tsevetanova, Z. *Study of biofilm formation in different pipe material in a model of drinking water distribution system and its impact in water quality*. Chemicals as intentional and accidental global environmental threats NATO. Security through Science Series, 463-468, DOI:10.1007/978-1-4020-5098-5_46 , 2006.

United States Environmental Protection Agency. *Control of biofilm growth in drinking water distribution systems*.EPA/625/R-92/001. Jun., 1992.

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

Van der Kooij, D. *Potential for biofilm development in drinking water distribution systems.* Journal of Applied Microbiology Vol. 85(S1), pp.39S–44S, Dic.1998.

Van der Kooij, D., Veenendaal, H.R., Baars-Lorist, C., Van der Klift, D. W. & Drost, Y.C. *Biofilm formation on surfaces of glass and teflón exposed to treated water*. Wat. Res, Vol. 29(7), pp. 1655-1662, 1995.

Vere, S. A., *Multilevel counterfactuals for generalizations of relational concepts and Productions*. Artificial Intelligence, Vol. 14, pp. 139-164, 1980.

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

Videla H. A. *Electrochemical interpretation of the role of microorganisms in corrosion*. In: Houghton DR, Smith RN, Eggins HOW (eds) Biodeterioration . Elsevier Applied Science, London, England, pp 359-371. 1988.

Volk, C.J., LeChevallier, M.W.*Impacts of the reduction of nutrients levels on bacteria water quality in distribution systems*, Appl. Environ. Microbiol., Vol. 65(11), pp. 4957. 1999.

Von Reyn, C. F. , Marlow, J. N., Arbeit, R. D., Barber, T. W. & Falkinham, J.O. *Persistent colonisation of potable water as a source of Mycobacterium avium infection in AIDS*. The Lancet, Vol. 343(8906), pp. 1137–1141, 7 Mayo 1994.

Wada, T., Horiuchi, T., Motoda, H. & Washio, T. *Characterization of Default Knowledge in Ripple Down Rules Method*. Institute of Scientic and Industrial Research, Osaka University, 1999.

Wadowsky R. M. & Yee R. B. *Satellite growth of Legionella pneumophila with an environmental isolate of Flavobacterium breve*. Appl Environ Microbiol., Vol. 46(6), pp.1447–1449. 1983.

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

White, D.R. & LeChevallier, M.W. *AOC associated with oils from lubricating well pumps.* J. Am. Water Works Assoc. Vol. 85(8), pp. 112–114.1993.

Wingender J. & Flemming, H.-C. *Contamination potential of drinking water distribution network biofilms*. Water Science and Technology Vol 49(11–12), pp. 277–286, 2004.

Wingender, J. & Flemming H.C. *Biofilms in drinking water and their role as reservoir for pathogens.* International Journal of Hygiene and Environmental Health, Vol. 214, pp. 417–423, 2011.

Witten, I H, Eibe Frank, Len Trigg, Mark Hall Geoffrey Holmes & Sally Jo Cunningham. *Weka: Practical machine learning tools and techniques with java implementations*. Department of Computer Science. University of Waikato. New Zealand. http://www.cs.waikato.ac. nz/~ml/publications/1999/99IHW-EF-LT-MH-GH-SJC%-Tools-Java.pdf. 2000.

Witten, I. H, Frank, E. & Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, ISBN 978-0-12-374856-0, 2011.

Yu, J., Kin, D. & Lee, T. *Microbial diversity in biofilms on water distribution pipes of different materials*. Water Science & Technology, Vol. 61(1), pp. 163-171, 2010.

Zaldivar, R. & Wetterstrand, W.H. *Nitrate nitrogen levels in drinking water of urban areas with high- and low-risk populations for stomach cancer: an environmental epidemiology study*. Z Krebsforsch Klin Onkol Cancer Res. Clin. Oncol. Vol. 92(3), pp. 227-234. 1978.

Zhang, G., Qi, M., *Neural network forecasting for seasonal and trend time series.* European Journal of Operational Research Vol. 160, pp. 501-514.2005.

Zacheus, O. M., Iivanainen, E. K., Nissinen, T. K., Lehtola, M. M .J. & Martikainen, P.J. *Bacterial biofilm formation on polybinyl chloride, polyethylene and stainless Steel exposed to ozonated water*. Wat. Res. Vol. 34(1), pp. 63-70, 2000.

Zhou, L.L., Zhang, Y.L. & Li, G.B, *Effect of pipe material and low level disinfectants on biofilm development in a simulated drinking water distribution system,* Journal of Zhejiang University ISSN 1673-565X (Print); ISSN 1862-1775 (Online), Vol. 10(5), pp. 725-731, 2009.