

## XII Simposio Iberoamericano sobre planificación de sistemas de abastecimiento y drenaje

### “MÉTODOS KERNEL PARA EL ESTUDIO DEL DESARROLLO DE BIOFILM EN LOS SISTEMAS DE DISTRIBUCIÓN DE AGUA POTABLE”

*E. Ramos Martínez (1), M. Herrera (2), J. Izquierdo (3), R. Pérez García (4)*

- (1) IMM-FluIng, Universitat Politècnica de València. C. de Vera s/n, 46022 Valencia, España  
+34 96 387 7007, [evarama@upv.es](mailto:evarama@upv.es)
- (2) BATir - Université libre de Bruxelles. Av. F. Roosevelt, 50 (CP 194/2), B-1050 Bruselas, Bélgica  
+32 2 650 2759, [mherrera@ulb.ac.be](mailto:mherrera@ulb.ac.be)
- (3) IMM-FluIng, Universitat Politècnica de València. C. de Vera s/n, 46022 Valencia, España  
+34 96 387 7007, [jizquier@gmmf.upv.es](mailto:jizquier@gmmf.upv.es)
- (4) IMM-FluIng, Universitat Politècnica de València. C. de Vera s/n, 46022 Valencia, España  
+34 96 387 7007, [rperez@gmmf.upv.es](mailto:rperez@gmmf.upv.es)

#### RESUMEN

El biofilm genera numerosos problemas en los sistemas de distribución de agua potable. Los estudios sobre la influencia conjunta que las características de estos sistemas tienen en el desarrollo de biofilm, son escasos debido a su complejidad. Recurriendo a técnicas de aprendizaje automático hemos generado una base de datos completa y lo suficientemente extensa como para ser capaces de estudiar el efecto que la interacción del conjunto de características físicas e hidráulicas de estos sistemas tiene sobre el desarrollo de biofilm. Para ello proponemos los métodos Kernel por su precisión y sencillez en el descubrimiento de patrones en problemas complejos.

**Palabras claves:** Biofilm, sistemas de distribución de agua potable, métodos Kernel, aprendizaje automático.

#### ABSTRACT

Biofilm is responsible of many problems in the drinking water distribution systems. Studies about the joint influence that the characteristics of these systems have on biofilm development are scarce due to its complexity. Using machine learning techniques we have generated a comprehensive and extensive enough database to be able to study the effect that the set of the physical and hydraulic characteristics of these systems have on biofilm development. We propose Kernel methods to this aim for their accuracy and simplicity discovering patterns when addressing complex problems.

**Key words:** Biofilm, drinking water distribution systems, Kernel methods, machine learning.

#### SOBRE EL AUTOR PRINCIPAL

**Eva Ramos Martínez:** Estudiante de doctorado en el Departamento de Ingeniería Hidráulica y del Medio Ambiente de la Universitat Politècnica de València (España) y miembro del grupo de investigación FluIng-IMM. Licenciada en Biología por la Universidad del País Vasco y con dos masters; Máster en Biodiversidad, Funcionamiento y Gestión de Ecosistemas (Universidad del País Vasco, España) y Máster en Ingeniería Hidráulica y del Medio Ambiente (Universitat Politècnica de València, España). Ha participado en varios congresos a nivel internacional y posee publicaciones en revistas científicas. Actualmente sus líneas de investigación se centran en la evaluación del desarrollo del biofilm en los sistemas de distribución de agua potable mediante el uso de herramientas de ayuda a la toma de decisiones y análisis inteligente de datos.

## ANTECEDENTES E INTRODUCCIÓN

En los últimos años, diferentes factores han hecho que aumenten las expectativas sobre la calidad del agua servida, aumentando así el interés en la investigación, protección y control de la calidad del agua de consumo humano. Es por ello que los gestores encargados de los servicios de agua, actualmente, están centrando sus esfuerzos en la etapa de distribución, tras el tratamiento, por ser la etapa en la que la calidad del agua puede experimentar un mayor deterioro y que adolece de un menor control.

Recientemente, se está tomando conciencia creciente del papel que el biofilm juega en el interior de las tuberías como uno de los principales agentes que influyen en el deterioro de la calidad del agua durante su distribución. El biofilm está formado por complejas comunidades de microorganismos lo que supone un riesgo sanitario por su papel como refugio de patógenos; además, también puede ser responsable del 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 sistemas de distribución de agua potable (DWDSs del inglés, Drinking Water Distribution Systems) en el desarrollo de biofilm, excepto notables excepciones, son escasos, debido a la complejidad de la comunidad y del entorno estudiado. Si bien, compilando datos de diferentes estudios sobre el desarrollo de biofilm en tuberías y recurriendo a técnicas de aprendizaje automático hemos generado una base de datos completa y lo suficientemente extensa como para ser capaces de estudiar el efecto que la interacción del conjunto de características físicas e hidráulicas de los DWDSs relevantes en el desarrollo de biofilm tiene sobre estas comunidades.

El presente trabajo se centra en estudiar las interacciones existentes entre el conjunto de estas variables y el desarrollo de biofilm en función del grado de desarrollo del mismo (bajo, medio o alto). De esta manera, se pretende profundizar en el estudio del biofilm en los DWDSs, logrando una mayor comprensión de las causas reales que hacen que el biofilm exista a diferentes niveles dentro de estos sistemas. Para alcanzar este objetivo se han propuesto los métodos Kernel por la precisión y sencillez con la que abordan problemas complejos. Estos métodos proporcionan un marco poderoso y unificado para el descubrimiento de patrones, dando lugar a algoritmos que pueden actuar sobre tipos

generales de datos y buscar tipos generales de relaciones. También proporcionan una forma natural de combinar e integrar los diferentes tipos de datos. La combinación del apropiado diseño Kernel y algoritmos Kernel relevantes ha dado origen a una poderosa y coherente clase de métodos, cuyas propiedades computacionales y estadísticas son ampliamente utilizadas.

Sintetizando, este proyecto profundiza en el estudio del biofilm y logra una mayor comprensión de su interacción con el medio en los DWDSs, sentando las bases para el desarrollo de una herramienta capaz de identificar y predecir las condiciones que favorecen un alto desarrollo de biofilm en estos sistemas.

## BASE CIENTÍFICO – TEÓRICA

En el presente trabajo se ha optado por el uso de las máquinas de soporte vectorial (SVMs, del inglés, Support Vector Machines) para estudiar la influencia que tienen en el desarrollo de biofilm el conjunto de las características hidráulicas y físicas de los DWDSs que individualmente se sabe son relevantes sobre estas comunidades de microorganismos. La aplicación de las SVMs para la clasificación del biofilm utilizará, de manera conjunta, las variables físicas y las hidráulicas, tenidas en cuenta en la base de datos (Ramos-Martínez et al., 2013). El estudio se completará, comparando la bondad de los resultados obtenidos con los obtenidos mediante otras técnicas de aprendizaje automático que han demostrado una alta precisión en estudios de clasificación (Witten et al., 2011). Las SVMs se basan en los métodos Kernel, que son algoritmos especializados en el análisis de patrones (Shawe-Taylor y Cristianini, 2006; Schölkopf y Smola, 2002) que proporcionan una forma eficiente de detectar relaciones no lineales. Su funcionamiento representa una ventaja especial en los casos no-lineales (de resolución compleja). Éstos se proyectan desde su espacio inicial a un espacio de alta dimensionalidad donde se pueden analizar mediante funciones lineales. Esta característica se complementa en la práctica con el llamado "kernel trick" (Aizerman et al., 1964), gracias al cual no es necesaria una representación explícita de los datos en ese espacio de alta dimensión sino que bastará con conocer la función que mapea los datos de un espacio a otro para poder hacer los análisis. Estas funciones son conocidas por funciones kernel y han de cumplir muy escasos requisitos, como ser semi-definidas positivas (el caso más sencillo es el producto interior de un vector). Resumiendo, gracias a las SVMs podremos clasificar mediante hiperplanos (lineales) grupos

distribuidos de manera no-lineal, obteniendo, además, una solución exacta y reproducible del problema (a diferencia de otros métodos que también tratan estas clasificaciones no-lineales, tales como las redes neuronales: que tienen una solución heurística y matemáticamente no reproducible). Las SVM clásicas trabajan con una clasificación binaria. En este artículo se hace uso de clasificaciones multi-clase que nos permitan clasificar el biofilm en las 3 categorías prefijadas por la base de datos: alto, medio y bajo. Esto nos permite determinar qué tuberías serán propensas a desarrollar una mayor cantidad de biofilm en su interior.

## METODOLOGÍA

En este trabajo se estudia el efecto conjunto que las características de los DWDSs tienen sobre el desarrollo de biofilm. Para ello se dispone de una base de datos obtenida mediante la compilación de datos de diferentes estudios de desarrollo de biofilm en tuberías, y la aplicación de técnicas de aprendizaje automático para poder lidiar con la heterogeneidad en las medición de los datos, la multiescalaridad, la falta de datos y las diferentes codificaciones utilizadas (Ramos-Martínez et al., 2013). De esta manera hemos generado una base de datos con 210 casos completos, lo suficientemente extensa como para permitir estudiar el efecto que la interacción del conjunto de características físicas e hidráulicas de los DWDSs relevantes en el desarrollo de biofilm tiene sobre estas comunidades. Las variables que conforman la base de datos fueron encontradas relevantes para el desarrollo de biofilm cuando fueron estudiadas individualmente por diferentes investigadores. Estas variables son:

- (i) Velocidad de flujo. Con la velocidad de flujo aumenta la transferencia de masa de nutrientes favoreciendo el desarrollo del biofilm (Lehtola et al., 2006). Sin embargo, velocidades específicas de entre 3-4 m/s pueden favorecer su desprendimiento (Cloete et al., 2003).
- (ii) Régimen hidráulico. Puede ser laminar o turbulento. Algunos biofilms en régimen turbulento tienden a ser más activos, tener mayor densidad celular, y distinta morfología, que los biofilms en flujo laminar (Simoes et al., 2007).
- (iii) Material de la tubería. Puede ser de metal, plástico, o cemento. En general, las tuberías de metal tienden a desarrollar más biofilm que las de cemento, y éstas más que las de plástico (Niquette et al., 2000). Esto se debe a que las tuberías

con una superficie más rugosa tienen un mayor potencial para el crecimiento de biofilm (Chowdhury, 2011). Las superficies rugosas proporcionan una mayor superficie de crecimiento para el biofilm y lo protegen de las fuerzas de corte del agua.

- (iv) Edad de las tuberías. La acumulación de sustancias disueltas y de corrosión en las tuberías de mayor edad puede aumentar su rugosidad (Christensen, 2009), lo que favorece el desarrollo de biofilm. Además, los depósitos más viejos tienden a tener mayor biomasa y contenido de bacterias (Chowdhury, 2011). Dividimos los tubos en jóvenes, de edad media y viejos (Tabla 1).
- (v) Edad del agua. Cuanto más tiempo está el agua en el sistema, mayor será el consumo de desinfectante residual, la deposición de sedimentos, y el aumento de la temperatura (EPA, 2002). Todos ellos son factores que favorecen el desarrollo del biofilm. En nuestro caso, hemos creado un índice sintético llamado "edad del agua". Para ello utilizamos el tiempo de retención hidráulica (h) (HRT, del inglés, Hydraulic Retention Time) y la distancia hasta el punto de cloración (km) ya que ambos aumentan con la edad del agua en el sistema. Con el fin de normalizarlos, escalamos cada variable, HRT y la distancia hasta el punto de cloración. El valor mínimo se resta al valor actual y se divide por la diferencia entre los valores máximo y mínimo. Al combinar dos variables en una sola, a fin de no sesgar el estudio se utiliza la proporción inversa existente en los datos originales. HRT se multiplica por un factor de 0,3, mientras que la distancia hasta el punto de cloración se pondera con un factor de 0,7: ya que disponemos de 2,5 veces más datos de HRT que de distancia al punto de cloración, por lo que HRT se multiplica por un factor de casi 2,5 veces más pequeño que el factor que multiplica la distancia al punto de desinfección. En consecuencia, las dos variables tienen una influencia comparable en la generación del índice. Por último, se re-escalamos, una vez más, los valores obtenidos. Por lo tanto, la edad del agua es un índice entre 0 y 1, donde los

valores cercanos a uno corresponden con las mayores edades de agua.

- (vi) Biofilm. Elegimos el recuento de heterótrofos en placa (HPC/cm<sup>2</sup>) como el método de cuantificación de biofilm. Aunque hay otros métodos, este es el más utilizado, y para el que más datos hay disponibles. Basándonos en la distribución de datos observada y en el criterio de expertos, se divide en desarrollo de biofilm en bajo, medio y alto (Tabla 1).

Las variables y categorías estudiadas en esta base de datos se describen en la Tabla 1. Es en esta base de datos donde se ha intentado clasificar, de la manera más eficiente posible, las diferentes categorías de desarrollo de biofilm en función de las características estudiadas. Para ello nos hemos centrado en las SVM basadas en los métodos Kernel beneficiándonos de las ventajas que ofrecen y comparando la bondad de sus resultados con el de otras técnicas de clasificación.

Con el fin de encontrar los mejores resultados se han utilizado diferentes funciones Kernel (Tabla 2). En todos los casos, una tercera parte de la base de datos se ha utilizado para test y las otras dos partes para entrenamiento y validación. Los parámetros  $C$  y  $\gamma$  se buscan por *Grid Search*, estableciendo una malla de posibles combinaciones donde se busca su óptimo, entre los límites [1, 100] y [0,1], respectivamente. La malla se organiza así para todos los posibles Kernels. La bondad de los resultados se ha estimado a través de los índices Diagonal y Kappa. Diagonal calcula el porcentaje de datos que se encuentran en la diagonal principal de cada matriz de confusión para cada prueba y Kappa es una corrección del índice Diagonal, que determina hasta qué punto la concordancia observada es superior a la que es esperable obtener por puro azar para cada solución (Landis y Koch, 1977).

Las otras técnicas de clasificación utilizadas han sido las reglas y los árboles de clasificación. En el caso de las reglas se han usado el algoritmo JRip (Cohen, 1995; Rajput et al., 2011) y NNge (Brent, 1995; Sylvain, 2002) y en el de los árboles los algoritmos J48 (implementación del algoritmo C4.5) (Quinlan, 1993; Rajput et al., 2011) y NBTree (Webb et al., 2005).

## PRESENTACIÓN DE RESULTADOS

Al aplicar las diferentes metodologías de clasificación expuestas anteriormente sobre la base

de datos discretizada, se observa que en el caso de las SVM el mejor resultado se obtiene con la función RBF (Tabla 3), mientras que el mejor resultado al aplicar otras técnicas de clasificación se observa tras aplicar el árbol de clasificación con el algoritmo J48. Aunque los resultados en ambos casos son muy parecidos, el valor de Kappa en el caso de la SVM con RBF es mayor que en el del árbol de clasificación con J48, siendo este índice más completo que el diagonal al tener en cuenta también los falsos negativos no solo los aciertos. Por lo que se puede decir, que observando todos los análisis, el mejor resultado se obtiene al aplicar la SVM con la función RBF.

**Tabla 1. Variables y categorías de la base de datos.**

BIOFILM (HPC/cm <sup>2</sup> )	VELOCIDAD DE FLUJO (m/s)
Bajo [0-10 <sup>3</sup> ] Medio [10 <sup>4</sup> -10 <sup>6</sup> ] Alto [ 10 <sup>7</sup> ]	Baja [0-0.7] Media [0.8-1.7] Alta [1.8-3.5]
REGIMEN HIDRÁULICO	EDAD DEL AGUA
Laminar Turbulento -	Baja [0-0.3] Media [0.4-0.6] Alta [0.7-1]
EDAD TUBERIA (años)	MATERIAL TUBERIA
Joven [0-10] Mediana [11-30] Vieja [ 30]	Metal Plástico Cemento

**Tabla 2. Variables y categorías de la base de datos.**

Kernel	Expresión
RBF	$k(x, y) = \exp\left(-\frac{\ x - y\ ^2}{2\sigma^2}\right)$
Lineal	$k(x, y) = x^T y + c$
Polinómica	$k(x, y) = (ax^T y + c)^d$
Sigmoidea	$k(x, y) = \tanh(ax^T y + c)$

**Tabla 3. Resultados obtenidos al aplicar las diferentes funciones de las SVM en la base de datos discreta**

MÉ-TODOS KERNEL	RBF	Lineal	Polinómica	Sigmoidea
Diag.	0.757	0.757	0.614	0.685
Kappa	0.533	0.512	0.283	0.359

Según este resultado y siguiendo los márgenes para valorar el grado de acuerdo en función del índice Kappa que propusieron Landis & Koch en 1977 (Tabla 5) se concluye que se ha obtenido un grado de acuerdo moderado. Así, con el objetivo de mejorar los resultados obtenidos, seguimos

trabajando con las SVM utilizando, en este caso, la máxima información disponible. Así decidimos aplicar las diferentes SVM a la base de datos que teníamos en un principio, antes de su discretización, es decir, a la base de datos con datos mixtos (Tabla 1). El procedimiento a seguir fue el mismo que el caso de la base de datos discreta. En este caso, el mejor resultado se obtiene en el caso de la función lineal. Se observa una mejora en los resultados, pasando de tener un grado de acuerdo moderado a tener un grado de acuerdo bueno (Tabla 6). Una vez obtenidos estos resultados con el fin de observar si es posible mejorar aún más la clasificación y aprovechando las posibilidades que las SVM y los métodos Kernel ofrecen aplicamos Multi-kernel a la base de datos mixta. Se trata de uno de los últimos retos en métodos Kernel, trabajar con Multi-kernel en el caso de tratar con múltiples tipos de datos (Gonen y Alpaydin, 2011). El Multi-kernel en las SVM está idealmente adaptado para el problema de la integración de datos, ya que permite convertir distintos tipos de datos en un formato común utilizable llevando a cabo una combinación ponderada de tantos diferentes Kernel como tipos de datos hay en la base de datos. En nuestro caso al haber dos tipos de datos (discretos y continuos) utilizamos dos tipos de funciones Kernel diferentes, RBF, para los datos continuos, y lineal para los datos discretos. En este caso, los resultados vuelven a mejorar (Tabla 7). Aunque se sigue teniendo un grado de acuerdo bueno, ahora se aproxima más un grado de acuerdo muy bueno.

**Tabla 4. Resultados obtenidos al aplicar las diferentes técnicas de clasificación en la base de datos discreta**

MÉTODOS DE CLASIFICACIÓN	Reglas de clasificación		Árboles de clasificación	
	JRip	NNge	J48	NBTree
Diag.	0.7605	0.7464	0.7764	0.7605
Kappa	0.5008	0.4758	0.5239	0.5008

## ANÁLISIS DE RESULTADOS

Este estudio ofrece una visión general de un trabajo innovador que utiliza los métodos Kernel como una herramienta interesante en este área, permitiendo el uso de los conocimientos adquiridos sobre el desarrollo de biofilm en los DWDSs de una manera práctica y eficiente. Además, posibilitando tener en cuenta el efecto de la interacción entre las características hidráulicas y físicas de los DWDSs, relevantes en el desarrollo del biofilm.

**Tabla 5. Valoración en función del índice Kappa**

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

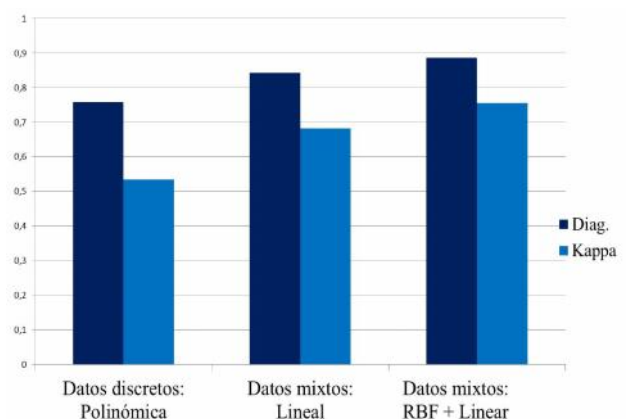
**Tabla 6. Resultados obtenidos al aplicar las diferentes funciones de las SVM en la base de datos discreta y mixta**

MÉTODOS KERNEL		RBF	Lineal	Polinóm.	Sigmoi.
Datos mixtos	Diag.	0.800	0.842	0.585	0.771
	Kap.	0.565	0.681	0.204	0.496
Datos Discretos	Diag.	0.757	0.757	0.614	0.685
	Kap.	0.533	0.512	0.283	0.359

**Tabla 7. Resultados obtenidos al aplicar multi-kernel**

MULTI-KERNEL		RBF + Lineal
Datos mixtos	Diag.	0.885
	Kappa	0.754

Los métodos Kernel han resultado ser los más efectivos a la hora de clasificar el grado de desarrollo de biofilm en los DWDSs en función de las características hidráulicas y físicas de los mismos al compararlos con otros métodos de clasificación comúnmente utilizados. Además, gracias a ellos se ha conseguido mejorar la clasificación, pasando de un grado de acuerdo moderado a un grado de acuerdo bueno, próximo a muy bueno (Figura 1).



**Figura 1. Resumen de los resultados al aplicar los métodos Kernel**

## CONCLUSIONES, RECOMENDACIONES, Y TRABAJO FUTURO

La complejidad de la comunidad y el medio ambiente estudiados es la razón por la que existe una escasez de trabajos que estudien la influencia conjunta que las características de los DWDSs tienen sobre el desarrollo del biofilm. En este trabajo hemos elegido los métodos Kernel para abordar este problema por su capacidad de recoger la información de una manera eficiente y adecuada, además, de por el hecho de que su adaptación es simple, en contraste con otros métodos de aprendizaje automático. Multi-kernel ha demostrado ser el mejor enfoque para este objetivo. La combinación de los métodos lineales y RBF permite utilizar todo el conocimiento disponible, sin perder la información al discretizar los datos.

El conocimiento obtenido mediante este estudio persigue el desarrollo de una herramienta más compleja de ayuda a la toma de decisiones capaz de predecir qué condiciones de los DWDSs favorecen el desarrollo de biofilm y qué medidas tomar para evitar, en lo posible, la existencia de estas localizaciones de mayor riesgo. De esta manera, se mitigarán de manera más eficiente los problemas derivados del desarrollo de biofilm en estos sistemas, por lo que se conseguirá llevar a cabo una gestión de la calidad del agua y del servicio de los DWDSs más eficiente y efectiva, minimizando así la repercusión sobre el consumidor y aumentando su satisfacción.

## BIBLIOGRAFÍA

- Aizerman M.A., Braverman E.M., Rozonoer L.I., (1964) Theoretical foundations of the potential function method in pattern recognition learning, Automation Remote Control, 25, pp. 821-827.
- Brent Martin (1995). Instance-Based learning: Nearest Neighbor With Generalization. Hamilton, New Zealand.
- Chowdhury, S. (2011). Heterotrophic bacteria in drinking water distribution system: a review, Environmental Monitoring and Assessment, pp. 2407–2415.
- Christensen, R.T. (2009). Age Effects on Iron-Based Pipes in Water Distribution Systems, Utah State University.
- Cloete, T.E. and Westard, D. and van Vuuren, S.J. (2003). Dynamic response of biofilm to pipe surface and fluid velocity, Water Science and Technology 45, pp. 57–59.
- Cohen W. W. (1995) Fast Effective Rule Induction. In: 12th International Conference on Machine Learning, pp. 115-123.
- Gonen, M., Alpaydin, E. (2011). Multiple kernel learning algorithms, Journal of Machine Learning Research 12, pp. 2211-2268.
- Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.
- Landis J. and Koch, G. (1977). The measurement of observed agreement for categorical data, Biometrics 33 pp.159–174.
- Lehtola, M.J. and Laxandera, M. and Miettinen, I.T. and Hirvonec, A. and Vartiainen, T. and Martikainen, P.J. (2006). The effects of changing water flow velocity on the formation of biofilms and water quality in pilot distribution system, Water Research 40, pp. 2151–2160.
- Niquette, P. M. and Servais, P. and Savoie, R. (2000). The role of hydrodynamic stress on the phenotypic characteristics of single and binary biofilms of *Pseudomonas fluorescens*, Water Resources 64.
- Rajput, A., Aharwal, R.P., Dubey, M., Saxena, S.P. and Raghuvansi M. (2011) J48 and JRIP rules for E-governance data. In: International Journal of Computer Science and Security, Volume (5) : Issue (2).
- Ramos-Martínez, E., Herrera, M, Izquierdo, J., Pérez-García, R. (2013). Pre-processing meta-data on biofilm development in drinking water distribution systems, Hydroinformatics, under review.
- Schölkopf, B., Smola, A. J. (2002). Learning with kernels. MIT Press.
- Shawe-Taylor, J., Cristianini, N. (2006). Kernel Methods for Pattern Analysis. Cambridge University Press.
- Simoës, M. and Pereira, M.O. and Vieira, M.J. (2007). The role of hydrodynamic stress on the phenotypic characteristics of single and

- binary biofilms of *Pseudomonas fluorescens*, Water Science and Technology 55, pp. 437–445
- Sylvain Roy (2002). Nearest Neighbor With Generalization. Christchurch, New Zealand.
- United States Environmental Protection Agency (2002) Effects of water age on distribution system water quality. Paper Issue.
- Webb, Geoffrey I., Janice R. Boughton, and Zhihai Wang. (2005): Not so naive bayes: Aggregating one-dependence estimators. Machine Learning 58.1 pp. 5-24.
- Witten, I. H, Frank, E. & Hall, M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, ISBN 978-0-12-374856-0.