Simulación de un controlador basado en el algoritmo Q-learning en un sistema de lodos activados

C.A. Peña Guzmán
Grupo de Investigación Agua y Sociedad, Departamento de Ingeniería Ambiental, Universidad Santo Tomas, Bogotá, Colombia, carlos.pena@usantotomas.edu.co

J.A. Lara-Borrero
Grupo de Investigación Ciencia e Ingeniería del Agua y el Ambiente, Departamento de Ingeniería Civil, Pontificia Universidad Javeriana, Bogotá, Colombia, laraj@javeriana.edu.co

ABSTRACT

This document presents a computational tool for control in a activated sludge reactor through simulations, based on the concept of reinforcement learning, where an agent runs an unknown environment and take action to achieve a specific objective. The control is execute to seraching flow for air injected into a reactor aerobic and the recycle sludge flow rate, to obtain a substrate of less than 100 mg/L and dissolved oxygen concentrations between 1 and 2 mg/L.

Keywords: Aeration control, Activated sludge, Reinforcement learning, Optimisation

RESUMEN

Este documento presenta una herramienta computacional para el control en un tanque de aireación de los lodos activados mediante simulaciones, la herrameinta se basa en el concepto de aprendizaje por refuerzo, donde un agente recorre un ambiente desconocido y realizando acciones para alcanzar un objetivo especifico. El control se ejecuta buscando un caudal de aire a inyectar a un reactor aeróbico y un caudal de recirculación de lodos, para obtener un efluente con una concentración inferior a 100 mg/L de DQO y concentraciones de oxígeno disuelto entre 1 y 2 mg/L.

Palabras claves: Control de aireación, lodos activados, aprendizaje por refuerzo, caudal de recirculación de lodos, optimización.

1. INTRODUCCIÓN

El alto costo (por construcción, mantenimiento y operación) en una gran parte de los procesos de tratamientos de aguas residuales, preocupa a la mayoría de gobernantes y poblaciones del mundo (incluyendo países desarrollados). Esto ha llevado a la ingeniería a investigar, crear métodos, sistemas, funciones etc. que permitan tener bajos costos y altas eficiencias (Tsagarakis et al., 2003).

Estos costos por operación y mantenimiento pueden dividirse en cuatro categorías: personal, energía, químicos y mantenimiento, donde los costos por personal y consumo energético son los más altos. La cantidad de personal en la mayoría de las Plantas de Tratamiento de Aguas Residuales (PTAR), es función del tamaño, tipo de PTAR y grado de tecnificación de esta. Sin embargo el consumo energético es el mayor aportante en el total de costos de operación en una PTAR aproximadamente una tercera parte (Tsagarakis et al., 2003), de esta parte la energía consumida por el proceso de aireación en una planta de lodos activados, se encuentra aproximadamente entre el 50 y 65% del total (Ferrer et al., 1998).

Los sistemas de lodos activados utilizan el oxígeno para realizar el proceso de oxidación de la materia orgánica, este oxígeno debe ser introducido de forma mecánica, lo que convierte a la aireación en un proceso con un fuerte consumo energético, por lo tanto controlar la concentración de oxígeno disuelto (OD) que ingresa al reactor aeróbico es esencial para este tipo de tratamientos (Rieger et al., 2006). Una concentración baja de OD podría generar un pobre crecimiento del lodo y una baja remoción de los contaminantes, a su vez, una alta concentración de OD podría presentar una pobre eficiencia de sedimentación del lodo al igual que un bajo rendimiento en la remoción (Fernández et al., 2011) , adicionalmente el exceso de OD requiere un alto caudal de aire (Lindberg, 1997).

De acuerdo con lo anterior, muchos investigadores, entidades gubernamentales y Universidades, se dieron a la tarea de buscar soluciones para el alto consumo energético utilizando controladores sobre las diferentes partes del proceso, por ejemplo se han utilizado controladores feedback y feedforward (Proporcional Integral Derivativo PID), Modelos Predictivos de Control (MPC) y adicionalmente se han implementado herramientas de la inteligencia artificial como: la lógica difusa y el ANFIS.

Numerosos trabajos hechos se han enfocado en el control de las concentraciones de oxígeno disuelto en el tanque de aireación, ya que la tasa de crecimiento de microorganismos y la concentración de sustrato del efluente son altamente dependientes de los niveles del oxígeno disuelto (Akyurek et al., 2009). Esta concentración es típicamente controlada por un valor constante deseado (set point) mediante el ajuste automático del caudal de aire (Olsson and Newell, 1999), este set point puede variar en su magnitud entre 1.7 a 2.5 mg/l, siendo 2 mg/l el valor más empleado (Kalker et al., 1999).

Por otra parte, otros procesos que son sujetos de control dentro de los sistemas de lodos activados, son el caudal de recirculación de lodos y el caudal de purga de estos, ya que al ser controlados y manipulados se puede manejar los niveles de distribución de lodos, la altura de lodos en el sedimentador y los tiempos de retención, los cuales afectan directamente la calidad final del efluente (Ma et al., 2006).

1.1 Aprendizaje por refuerzo (AR)

La idea de aprender por la interacción con nuestro ambiente, es probablemente la primera forma que pensamos de aprendizaje natural, cuando un infante juega, mueve sus brazos o mira a su alrededor, este no necesita de un profesor explicito ya que tiene una conexión directa mediante un sensor que lo conecta con el ambiente, este ejercicio de esta conexión produce una gran cantidad de información, a través de causa y efecto por consecuencia de las acciones y el cumplimiento de sus objetivos (Sutton and Barto, 1998).

En los algoritmos del aprendizaje por refuerzo, el alumno o agente observa e interactúa en un estado del sistema para producir una salida, este recibe un refuerzo (recompensa o penalidad), la cual se puede entender como una señal de evaluación que indica la utilidad de la salida, este paso lleva a la selección de la mejor acción y así afrontar un nuevo estado y repetir el proceso, esto puede entenderse como un aprendizaje de ensayo y error (ver Figura 1) (Kretchmar, 2000).

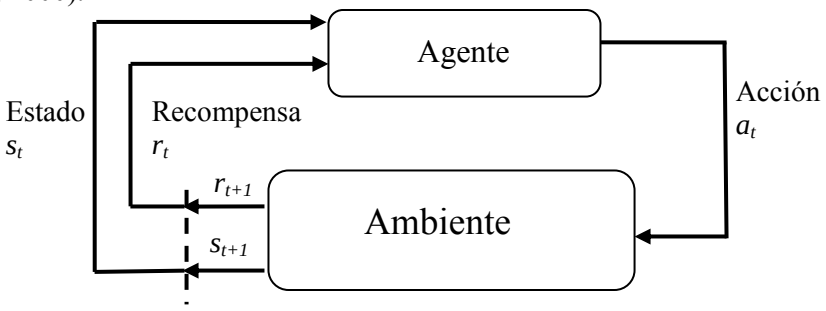

Figura 1: Esquema del aprendizaje por refuerzo

1.2 Q-learning

El Q-learning es uno de los algoritmos de aprendizaje por refuerzo más importantes y fue presentado por Watkins en 1989 (Guo et al., 2004), este algoritmo es una combinación de la programación dinámica, más concretamente el algoritmo de Iteración de Valores y la aproximación estocástica (Azar et al., 2011). Se considera un método que involucra el concepto de "modelo libre de aprendizaje por refuerzo", esto quiere decir que se proporciona al agente con la capacidad de aprender, para actuar en un dominio Markoviano por la experiencia dada por la consecuencia de acciones (Watkins, 1989, Watkins and Dayan, 1992).

De acuerdo a lo anterior, si en cada paso de tiempo discreto t = 1, 2,…., el controlador observa el estado s_t del proceso de Markov, selecciona una acción a_t, recibe una recompensa resultante r_t y observa el resultado del siguiente estado s_{t+1}. La distribución de probabilidad para r_t y s_{t+1} dependen solo de r_t y a_t tiene un valor finito esperado. El objetivo es encontrar una regla de búsqueda que maximice en cada paso de tiempo la suma de las recompensas futuras esperadas (Sutton et al., 1992).

$$E\{\sum_{j=0}^{\infty}\gamma^{j}r_{t+j}\} \qquad (1)$$

Donde γ es un factor discontinuo que se encuentra $0 \leq \gamma \geq 1$.

De acuerdo a Sutton et al. (1992), una idea básica del Q-learning es estimar una función de valores, donde $Q(s,a)$ es la suma esperada de las futuras recompensas por el desempeño de una acción a en el estado s y un rendimiento óptimo a partir de entonces. Esta función satisface la siguiente relación recursiva:

$$Q(s,a) = E\{\gamma_t + max_b Q(s_{t+1},b) | s_t = s, a_t = a\} \qquad (2)$$

Adicionalmente, el Q-learning es insensible a la exploración, esto quiere decir que los valores Q convergerán a los valores óptimos, independientemente de cómo el agente se comporte mientras inicia la colecta de datos, esto significa que aunque la exploración-explotación debe ser abordado en el Q-learning, los detalles de la estrategia de exploración no afectarán la convergencia de aprendizaje del algoritmo, por esta razón el Q-learning es el algoritmo de modelo-libre más popular y al parecer el más efectivo (Kaelbling et al., 1996).

Por lo tanto, este artículo propone una alternativa para el control de la concentración de oxígeno disuelto y el caudal de recirculación de lodos en un sistema de lodos activados, aplicando un controlador a través de un *modelo libre de aprendizaje*, basado en un algoritmo del concepto de *aprendizaje por refuerzo* (*AR*) mediante procesos de simulación.

2. MÉTODOS

Para la elaboración y simulación del controlador sobre una planta de lodos activados, se han planteado en este artículo la elaboración de dos etapas básicas, la predicción y el control. La primera hace referencia a la predicción de variables estimadas sobre los procesos de la planta de lodos activados, en la segunda etapa, el controlador se nutre de estos valores estimados, para realiza las acciones necesarias para alcanzar un valor objetivo deseado.

2.1 Simulación de reactor aerobio de un sistema de lodos activados.

Antes de la elaboración del agente, es primordial la construcción y delimitación del ambiente en el cual se va a desenvolver el mismo, para este caso, el ambiente está representado por el sistema de lodos activados, de acuerdo a esto, el objetivo de la simulación del sistema de lodos cumple un papel fundamental en la comprensión del

ambiente para el agente, ya que este debe ejecutar sus acciones y analizar qué repercusiones (recompensas o penalidades) tuvo estas sobre el ambiente y sobre los objetivos planteados.

Las siguientes ecuaciones muestran el balance en el sistema de lodos para obtener el comportamiento de la Demanda Química de Oxígeno (DQO) o sustrato (*S*), el de la biomasa (*SSV* o *X*), así como la concentración de oxígeno disuelto OD (O_2) en el reactor y también los *SSV* en el sedimentador (X_t) (Martinez, 2005), por lo tanto se tiene:

- ***En el reactor***

$$Sustrato = \frac{d_S}{d_t} = \frac{Q_f}{V} S_f - \frac{Q_o}{V} S - \frac{\mu X}{Y} \quad (3)$$

$$Biomasa = \frac{d_X}{d_t} = \frac{Q_r}{V} X_r - \frac{Q_o}{V} X - \mu X - k_d X \quad (4)$$

$$Oxígeno\ disuelto = \frac{dC_{O_2}}{dt} = \frac{Q_f}{V} C_{O_{2f}} - \frac{Q_o}{V} C_{O_2} - \frac{\mu X}{Y_{O_2}} - b * X + Kla_w * (C_{sat} - C_{O_2}) \quad (5)$$

- ***En el sedimentador***

$$Biomasa = \frac{d_{X_r}}{d_t} = \frac{Q_u}{V_s} X_r - \frac{Q_o}{V} X \quad (6)$$

Para llevar a cabo la simulación del modelo matemático, se tomó la (DQO) como el sustrato de entrada al modelo, esto por la importancia que tiene este parámetro en el diseño y en la modelación de procesos de lodos activados (Henze, 2008).

Las mediciones de DQO se llevaron a cabo dentro de un convenio realizado entre la Pontificia Universidad Javeriana y la Empresa de Acueducto y Alcantarillado de Bogotá-ESP en el año 2011, estas se realizaron a través de un Spectro::lyser, el cual es un espectrómetro sumergible, capaz de medir en línea los espectros de absorción (UV-Visible) directamente en medio líquidos (in situ) y con alta calidad. Las mediciones fueron hechas en continuo cada minuto durante 25 días, sin embargo no todos los días reportaron datos, ya que se presentaron algunas discontinuidades en el suministro de aguas residuales a través del sistema de bombeos, por lo tanto solo se seleccionaron 13 días, los cuales se puede observar en la figura 2.

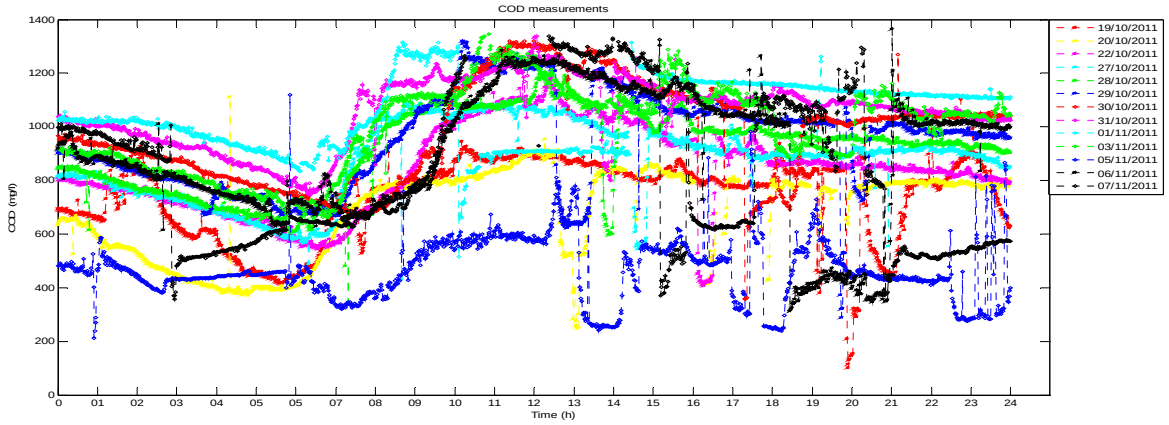

Figura 2: Mediciones de DQO durante 13 diferentes días

Es claro que las concentraciones de DQO varían según su día y hora, sin embargo se identifica una tendencia en el comportamiento de la misma, por consiguiente para comprobar el funcionamiento del controlador, se consideró realizar una simulación con un día patrón de DQO, para lo cual se llevó a cabo un cálculo de valores atípicos mediante un diagrama de cajas sobre las mediciones hechas cada minuto con respecto a los 13 días seleccionados. Finalmente, al llevar a cabo la totalidad de evaluaciones hechas sobre las series de datos y eliminando los valores atípicos, el percentil 50 (la media) se tomó como el día patrón.

2.2 Desarrollo del controlador

Luego de tener un ambiente definido en el cual va a interactuar el agente, es necesario configurar los demás elementos y subelementos que existen dentro del aprendizaje por refuerzo. La idea principal del controlador, es mantener el sustrato de salida del reactor aerobio inferior a 100 mg/L, siendo cualquiera de estas concentraciones el estado objetivo, por consiguiente, para la obtención de este, el agente toma dos acciones, las cuales consisten en modificar el caudal de aire inyectado al reactor aeróbico y/o el caudal de recirculación de lodos.

Como estas acciones podrían tomar valores no representativos para la realidad del sistema, se condicionó al agente a una cantidad de posibles acciones (rangos), ya que él podría tomar como decisión no inyectar aire, situación que tornaría al reactor en anaeróbico o inyectar más de lo necesario. Por lo tanto, para el caudal de recirculación de lodos se tomaron rangos de recirculación que van desde el 0% hasta 120%, en cuanto al caudal de aire, este valor afecta directamente a las concentraciones de oxígeno disuelto en el reactor aerobio, de acuerdo a esto, se planteó que el caudal de aire no supere una concentración de 2 mg/l o sea inferior a 1 mg/l de OD.

Después de delimitar los caudales, el agente recorre la totalidad de las acciones conjuntas, las cuales ingresan al ambiente, donde recibe una recompensa de diferentes magnitudes. Estas magnitudes son función de que tan alejado se encuentra el agente de los objetivos deseados (sustrato y concentración de oxígeno).

2.3 Rutina del Controlador

A diferencia de la gran mayoría de controladores, en donde la acción de control se lleva luego de identificar o medir una alteración sobre un valor deseado dentro de un proceso, este controlador propuesto mide su objetivo luego de llevarse a cabo las acciones de control (como se mencionó anteriormente), las cuales son tomadas en función del tiempo de retención hidráulico en el tanque de aireación y el tiempo total de mediciones hechas sobre el sustrato, como se observa a continuación:

$$Control = \frac{\text{Tiempo total de mediciones}}{\text{Tiempo de retención hidráulico}} \quad (7)$$

Al definir cuantos controles se van a llevar a cabo, el agente crea dos límites para cada acción de control, el primero es el $sustrato_i$ y el segundo es el $sustrato_n$ los cuales representan el primer y el último de los sustratos que ingresan al reactor durante cada tiempo de retención hidráulico (para este control se tomó un tiempo de retención de 3 horas). Con los límites identificados el agente busca la mejor acción, la cual luego de ser identificada es replicarla sobre cada sustrato medido entre los límites establecidos, así verificará si esta acción es pertinente para todos los sustratos existentes sobre el ciclo de control evaluado o si por el contrario debe modificarla nuevamente, hasta encontrar uno que sea aplicable sobre la totalidad de sustratos de ese ciclo.

Para la ejecución de las acciones sobre el caudal de recirculación, estas se llevaron sobre las ecuaciones 3, 4 y 5, en cuanto a las acciones sobre el caudal de aire, se realizó una modificación a la estructura principal de la ecuación del balance de oxígeno, específicamente sobre el coeficiente de transferencia de oxígeno (kla_w), ya que este valor se relaciona con el cambio en la intensidad de la aireación y el caudal de aire (Qa) suministrado (Makinia and Wells, 2007, Chai and Lie, 2008, Makinia, 2010).

Por tal motivo se usa la ecuación de Goto and Andrews de 1985 $K_{La} = m_1 Q_A - b_1$, la cual se remplazará en la ecuación 5, quedando de la siguiente manera:

$$\frac{dC_{O_2}}{dt} = \frac{Q_f}{V} C_{O_{2f}} - \frac{Q_R}{C} C_{O_2} - \frac{\mu X}{Y_{O_2}} - b * X + * m_1 Q_A - b_1 (C_{sat} - C_{O_2}) \quad (8)$$

Donde $m_1 = 0.0081$ y $b_1 = 2.85$ valores obtenidos de un ensayo de Makinia and Wells en 2007.

3. RESULTADOS Y DISCUSIÓN

Para la comprobación del funcionamiento del agente, se inició empleando los valores obtenidos del día típico calculado, donde se generó un primer escenario en el cual se encontró que la acción que más se llevó a cabo fue la variación del caudal de recirculación de lodos, esto con el fin de permitir un comportamiento de la biomasa óptimo en los tanques, por otra parte el control sobre las concentraciones de oxígeno al inicio de la simulación no dio presentó cumplimiento sin embargo el sustrato siempre se encontró dentro del rango deseado, como se observa en la fig. 3.

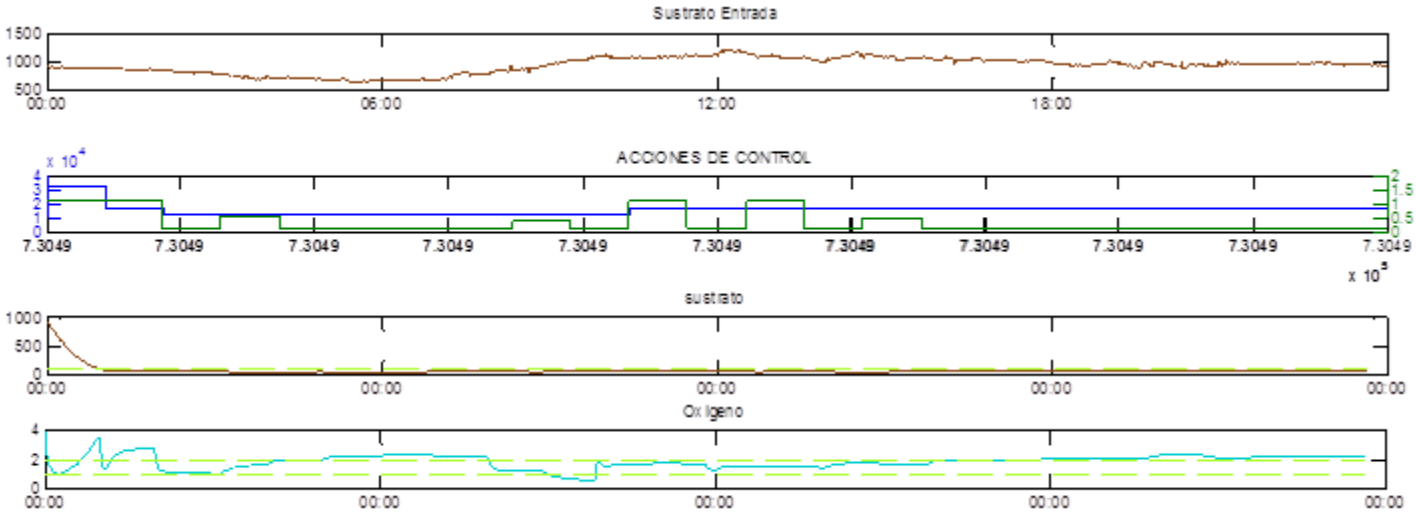

Figura 3: Resultado y comportamientos de la simulación del controlador sobre el día típico calculado

La figura 4 muestra los resultados a la función de recompensa para esta simulación, en la cual se puede ver claramente como las acciones conjuntas con menores penalizaciones presentan pocos rangos, sin embargo el agente estas las repite para dar cumplimiento a los objetivos planteados.

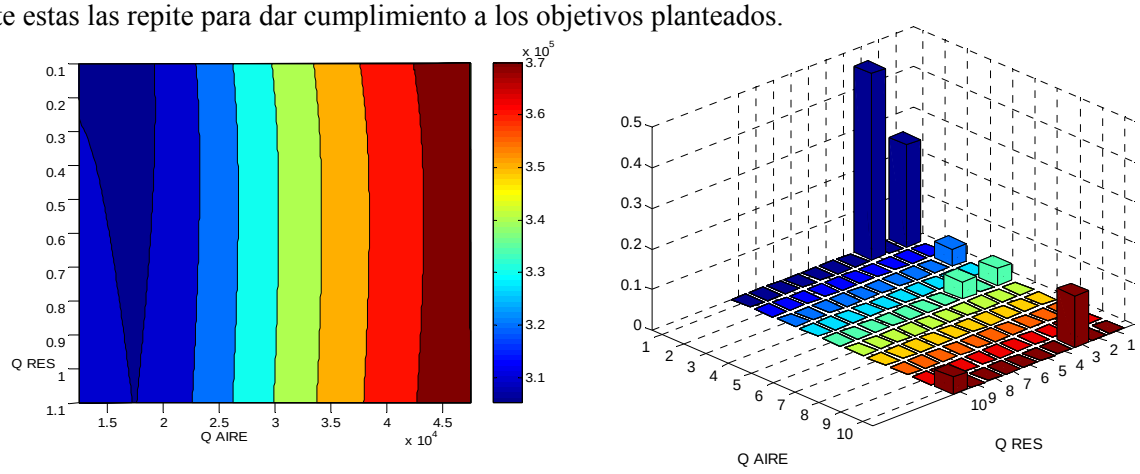

Figura 4: Penalidades y distribución de probabilidad conjunta de acciones

De acuerdo al comportamiento del agente dentro del día típico, se decidió llevar acabo un segundo escenario, donde se tomaran todas las mediciones de DQO medidas durante los 13 días, haciendo la suposición que estos días eran continuos y llevando a cabo reconstrucción de datos faltantes.

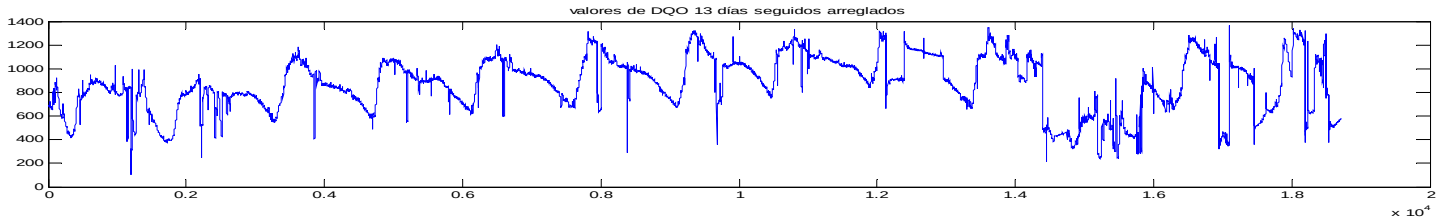

Figura 5: Concentraciones de DQO durante 13 días con intervalos de medición de un minuto

De acuerdo a la simulación, el agente realizó aumentó las acciones de control sobre el caudal de oxígeno a medida que se disminuía de manera drástica el sustrato, comportamiento muy parecido con el caudal de recirculación de lodos, sin embargo se evidencia fuertemente la sensibilidad de este parámetro a las altas perturbaciones de DQO, ya que en pocas ocasiones se mantiene constante el grado de recirculación de lodos, estas mismas perturbaciones presentan grandes dificultades para el agente en el control de las concentraciones de oxígeno, sin embargo es de resaltar que aun así el agente intenta rápidamente mantener las concentraciones de oxígeno dentro del objetivo deseado como se observa en la figura 6.

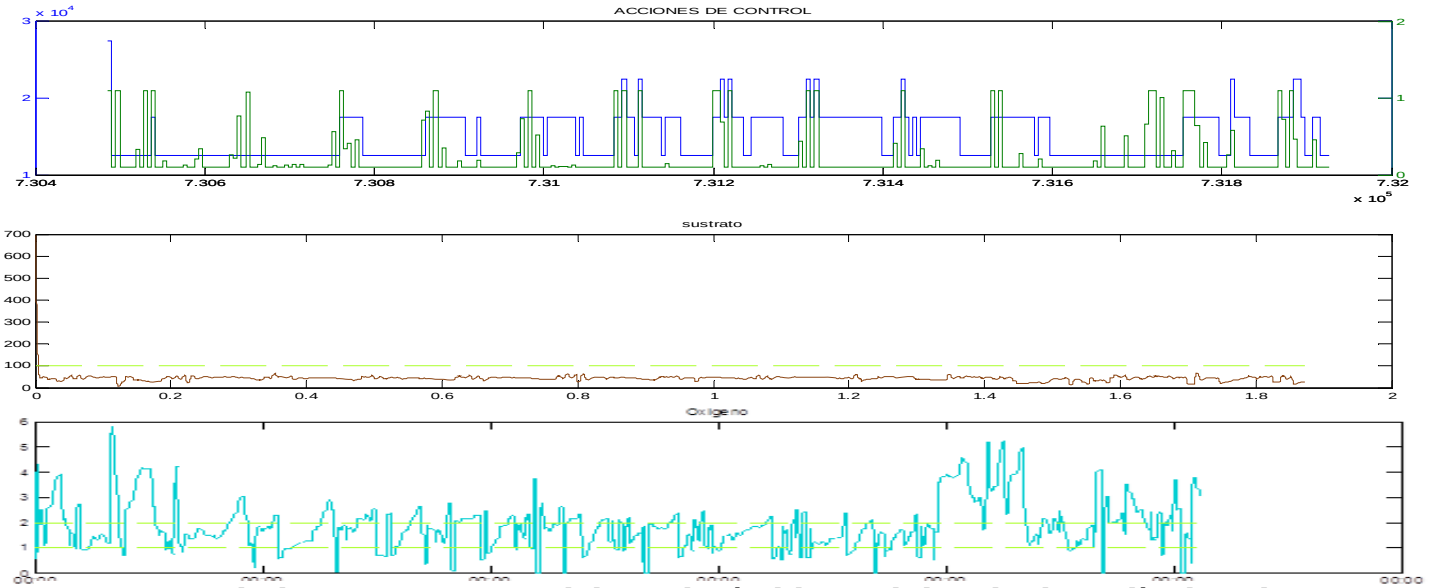

Figura 6: Resultado y comportamientos de la simulación del controlador sobre los 13 días de mediciones

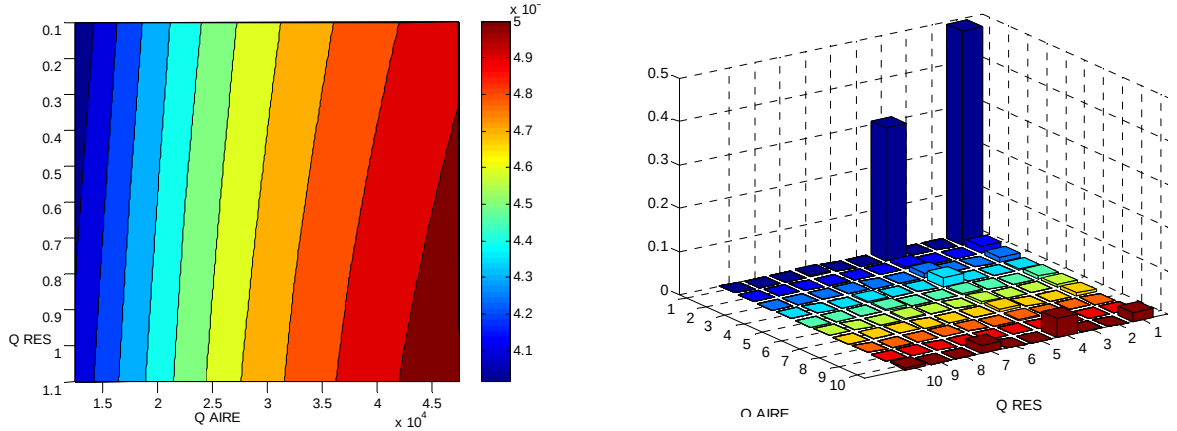

Figura 7: Penalidades y distribución de probabilidad conjunta de acciones

Para verificar el funcionamiento del agente bajo diferentes condiciones y su sensibilidad, se planteó un escenario donde se variaba el volumen en los tanques de aireación y de sedimentación al 50%.

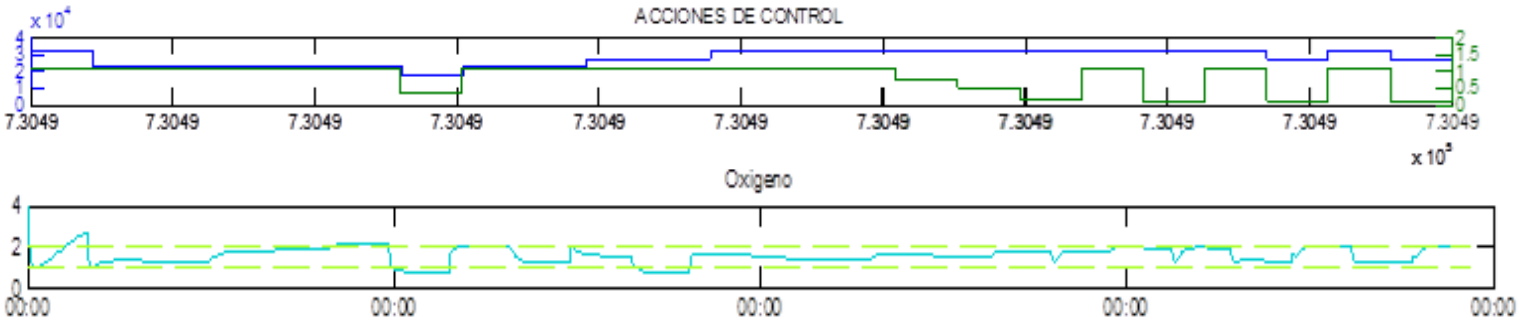

Figura 8 Resultado y comportamientos de la simulación del controlador sobre el día típico calculado

Como resultado se obtuvo, que el agente llevo a cabo un mayor control sobre caudal de aire como se observa en la figura 8, adicionalmente las concentraciones de oxígeno disuelto dentro de la franja objetivo fueron más altas, encontrándose que solo el 19.4% de las concentraciones estuvieron por fuera del rango.

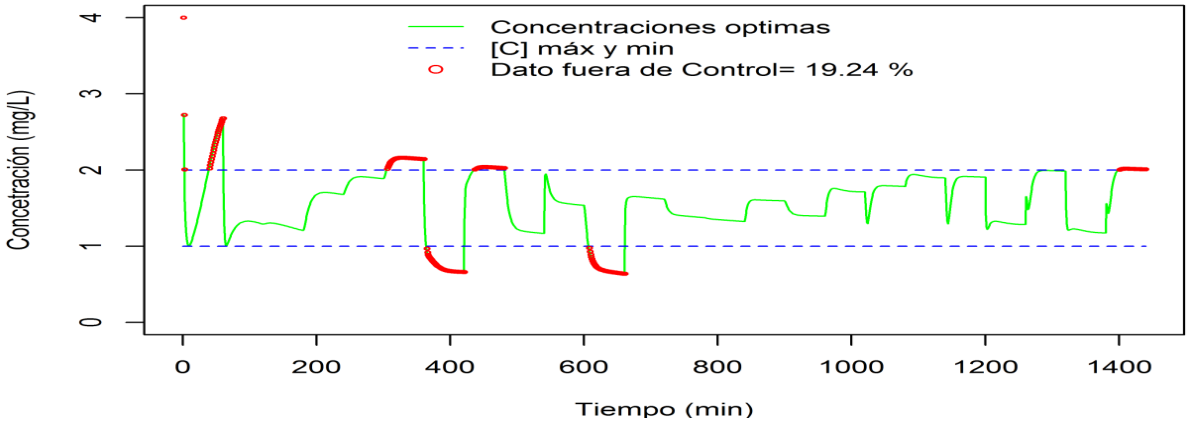

Figura 9: Concentraciones de OD fuera del rango de control

Por último se realizó el mismo escenario con la totalidad de días, donde se encontró un mayor cumplimiento en el control sobre el OD, pero a su vez ejecutó más acciones de control, tanto como en los caudales de oxígeno como en los caudales de recirculación de lodos (ver figura 10), situación que lleva a identificar como bajo mayor esfuerzo por cumplir el objetivo, el agente ejecuta mayor acciones de control.

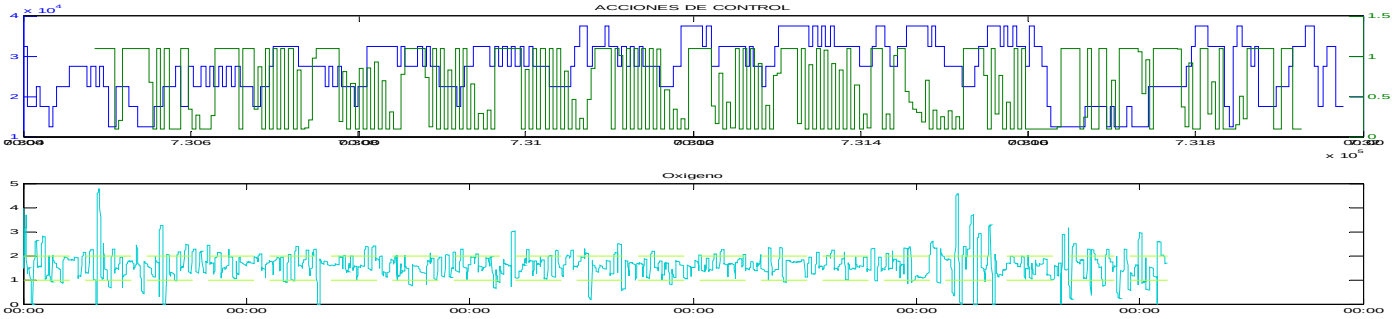

Figura 10: Resultado y comportamientos de la simulación del controlador sobre los 13 días

Al igual que en el anterior caso, el comportamiento del oxígeno disuelto fue mucho mejor, reflejado por la acción continua del controlador de caudal de aire, donde solo el 27. 84% de los valores estuvieron fuera del rango (3744 valores de 18720) como se puede ver a continuación:

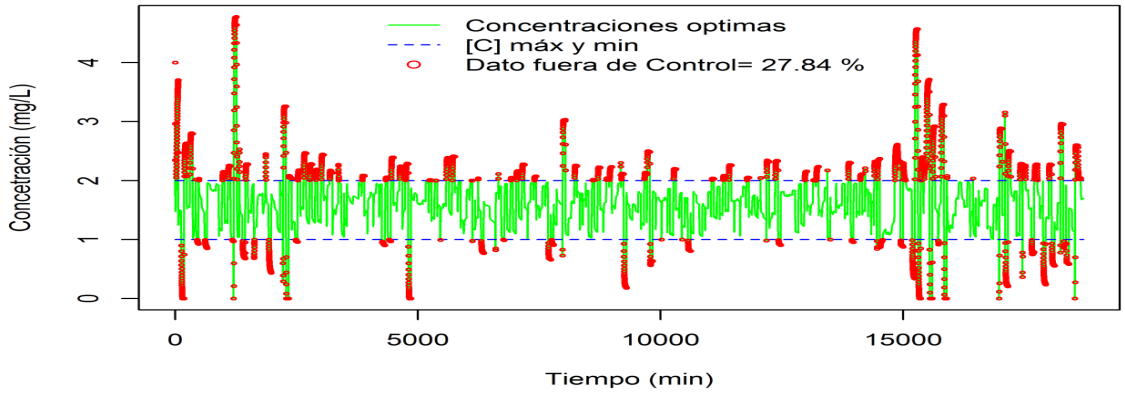

Figura 13: Concentraciones de OD fuera del rango de control

4. CONCLUSIONES

De acuerdo a lo observado en el comportamiento del agente y de los parámetros del sistema de lodos activados, en los diferentes escenarios generados para la prueba del controlador, se puede establecer que el caudal de recirculación de lodos es más sensibles bajo diferentes condiciones (concentraciones y dimensiones), presentando un alto impacto sobre el crecimiento de la biomasa y el consumo del sustrato.

Por otra parte se encontró que el agente presenta problemas de control, al sentir en su ambiente cambios bruscos (perturbaciones) de concentraciones de DQO, situación que conllevó al no cumplimiento en los límites de concentración de OD, donde más del 50% de los datos estuvieron por fuera del objetivo, cabe resaltar que a pesar de esta situación el agente siempre buscó la mejor opción sobre el caudal de recirculación de lodos.

Se encontró que el agente no presenta una gran sensibilidad a la variación de volúmenes en los tanques y que por el contrario su desempeño para el escenario con todos los datos fue superior al valor normal de diseño.

BIBLIOGRAFÍA

Akyurek E, Yuceer M, Atasoy I, Berber R (2009) Comparison of Control Strategies for Dissolved Oxygen Control in Activated Sludge Wastewater Treatment Process. Computer Aided Chemical Engineering 26:1197-1201.

Azar MG, Munos R, Ghavamzadeh M, Kappen HJ (2011) Speedy Q-Learning.

Chai Q, Lie B (2008) Predictive control of an intermittently aerated activated sludge process. pp 2209-2214: IEEE.

Fernández F, Castro M, Rodrigo M, Cañizares P (2011) Reduction of aeration costs by tuning a multi-set point on/off controller: A case study. Control Engineering Practice 19:1231–1237.

Ferrer J, Rodrigo, Seco, Penya-roja (1998) Energy saving in the aeration process by fuzzy logic control. Water Science & Technology 38:209-217.

Guo M, Liu Y, Malec J (2004) A new Q-learning algorithm based on the metropolis criterion. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34:2140-2143.

Henze M (2008) Biological wastewater treatment: principles, modelling and design: Intl Water Assn.

Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: A survey. Journal of Artificial Intelligence Research 237-285.
Kalker T, Van Goor C, Roeleveld P, Ruland M, Babuška R (1999) Fuzzy control of aeration in an activated sludge wastewater treatment plant: design, simulation and evaluation. Water Science and Technology 39:71-78.
Kretchmar RM (2000) A synthesis of reinforcement learning and robust control theory. In: Department of Computer Science, vol. Doctor of Philosophy Fort Collins: Colorado State University.
Lindberg C-F (1997) Control and estimation strategies applied to the activated sludge process. In: Materials Science Systems and Control Group, vol. Doctor of Philosophy in Automatic Control, p 214: Uppsala University.
Ma Y, Peng Y, Wang S (2006) New automatic control strategies for sludge recycling and wastage for the optimum operation of predenitrification processes. Journal of Chemical Technology and Biotechnology 81:41-47.
Makinia J (2010) Mathematical Modelling and Computer Simulation of Activated Sludge Systems: Intl Water Assn.
Makinia J, Wells SA (2007) Improvements in modelling dissolved oxygen in activated sludge systems. Portland State University 751:1-9.
Martinez SA (2005) Tratamiento de aguas residuales con MATLAB: Reverte.
Olsson G, Newell B (1999) Wastewater treatment systems: modelling, diagnosis and control: Intl Water Assn.
Rieger L, Alex J, Gujer W, Siegrist H (2006) Modelling of aeration systems at wastewater treatment plants. Water Science & Technology 53:439-447.
Sutton RS, Barto AG (1998) Reinforcement learning: An introduction: Cambridge Univ Press.
Sutton RS, Barto AG, Williams RJ (1992) Reinforcement learning is direct adaptive optimal control. Control Systems, IEEE 12:19-22.
Tsagarakis K, Mara D, Angelakis A (2003) Application of cost criteria for selection of municipal wastewater treatment systems. Water, Air, & Soil Pollution 142:187-210.
Watkins CJCH (1989) Learning from delayed rewards. vol. Ph.D Cambridge: King's College, Cambridge.
Watkins CJCH, Dayan P (1992) Q-learning. Machine learning 8:279-292.

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.