CONTROL DE OXIGENO DISUELTO EN UN TANQUE DE AIREACIÓN DE UNA PLANTA PILOTO DE LODOS ACTIVADOS

Ing. CARLOS ANDRÉS PEÑA GUZMÁN

BOGOTÁ D.C.
PONTIFICA UNIVERSIDAD JAVERIANA
MAESTRÍA EN HIDROSISTEMAS
FACULTAD DE INGENIERÍA
2012

CONTROL DE OXIGENO DISUELTO EN UN TANQUE DE AIREACIÓN DE UNA PLANTA PILOTO DE LODOS ACTIVADOS

Ing. CARLOS ANDRÉS PEÑA GUZMÁN

Trabajo de Grado presentado como requisito
parcial para optar el título de Magister en Hidrosistemas
Director: JAIME ANDRÉS LARA BORRERO
I.C., M.Sc., PhD.

BOGOTÁ D.C.
PONTIFICA UNIVERSIDAD JAVERIANA
MAESTRÍA EN HIDROSISTEMAS
FACULTAD DE INGENIERÍA
2012

Nota de Aceptación

_____

_____

_____

Director

_____
JAIME ANDRES LARA BORRERO

Jurado

_____
JUAN MANUEL GUTIERREZ

Jurado

_____
CARLOS EDUARDO COTRINO BADILLO

Tabla de contenido

Tabla de figuras

Tabla de tablas

Tabla de fotos

1. JUSTIFICACIÓN

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 sociedades 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, Mara et al. 2003).

Los 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) está en función del tamaño, tipo de PTAR y al grado de optimización de esta. Sin embargo el consumo energético es el mayor aportante en el total de costos de operación en una PTAR (Tsagarakis, Mara et al. 2003), lo que lo convierte en un ítem primordial a inspeccionar.

El consumo energético, es aproximadamente una tercera parte del costo total de operación de una PTAR (Tsagarakis, Mara et al. 2003; Fika, Chachuat et al. 2005), 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 consumo total (Ferrer, Rodrigo et al. 1998; Duchène, Cotteux et al. 2001; Ingildsen, Jeppsson et al. 2002; Fika, Chachuat et al. 2005; Rieger, Alex et al. 2006; Vrecko, Hvala et al. 2006).

Los sistemas de lodos activados utilizan el oxígeno para realizar el proceso oxidación de la materia orgánica, lo que convierte a la aireación en un proceso con un fuerte consumo energético, ya que este debe ser inyectado por máquinas, por lo tanto controlar la concentración de oxígeno disuelto (OD) que ingresa al reactor aeróbico, es esencial para este tipo de tratamientos (Samuelsson and Carlsson 2002; Chachuata, Rocheb et al. 2005; Rieger, Alex et al. 2006). Debido a que una muy baja concentración de OD podría generar un pobre crecimiento del lodo y una baja remoción en 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, M.C.Castro et al. 2011), adicionalmente el exceso de OD requiere de una alta tasa de caudal de aire (Lindberg 1997).

Por lo tanto los sistemas de automatización o control, juegan un rol muy importante en la reducción de costos de operación en plantas de tratamiento de aguas residuales e industriales (Bongards 1999).

2. OBJETIVOS

2.1. Objetivo general

Controlar el caudal de oxígeno que ingresa en un tanque de aireación de una planta piloto de lodos activados mediante una herramienta de aprendizaje por refuerzo (reinforcement learning).

2.2. Objetivos específicos

- Desarrollar un controlador de caudal de oxígeno a través una herramienta computacional de aprendizaje por refuerzo.
- Establecer el ahorro energético sobre una planta piloto a través de la implementación del controlador de caudal de oxígeno.

3. MARCO TEÓRICO

Bogotá es una ciudad que ha venido creciendo de manera abrumadora, no solo en espacio si no también en población, pasando aproximadamente de 2.9 millones de habitantes en 1973 a 6.8 millones en 2005, de acuerdo a este crecimiento para el año 2020 se espera que en la ciudad se cuente alrededor de 8.4 millones de habitantes (Rodríguez, Díaz-Granados et al. 2008), lo que conlleva no solo un mayor consumo de recursos naturales sino también una mayor generación de impactos ambientales.

Uno de estos impactos, es la producción de aguas residuales sin tratar que son vertidas a los diferentes cuerpos hídricos del Distrito, ya sea por la falta de alcantarillado o de sistemas de tratamiento para estas. Cabe mencionar que el sistema de alcantarillado de Bogotá está compuesto por tres Subcuencas: Salitre, Fucha y Tunjuelo, donde cada una de estas drena a los tres ríos del Distrito Capital que poseen los mismos nombres (Rodríguez, Díaz-Granados et al. 2008; Giraldo J.M., Leirens S. et al. 2010) y a su vez estos desembocan en el río Bogotá.

De acuerdo con la Empresa de Acueducto y Alcantarillado de Bogotá (EAAB-ESP) en 2005, la cobertura del servicio de alcantarillado para Bogotá era del 96.9%, del cual cerca del 19% son tratadas por la Planta de Tratamiento de Aguas Residuales el Salitre (PTAR Salitre). La cual fue construida entre 1997 y 2000, para tratar las aguas residuales de la cuenca del Salitre con un caudal de 4 m^3 s^{-1}, mediante procesos físicos y químicos (Rodríguez, Díaz-Granados et al. 2008), sin embargo las demás aguas son vertidas a ríos, quebradas o canales (EAAB-ESP 2005).

A pesar de esta y otras acciones por parte de diferentes entidades del Distrito para mejorar la calidad hídrica, es claro que no es suficiente para tener un ambiente más sano, situación que llevó a la población capitalina a buscar de manera legal la recuperación del río Bogotá, por esto el Tribunal Administrativo de Cundinamarca mediante el fallo 01-479 del 25 de agosto de 2004, obligó a las autoridades gubernamentales; la ampliación de la PTAR Salitre a 8 m^3 s^{-1} y la construcción de una nueva PTAR que trate las aguas de las Subcuencas de Fucha y Tunjuelo (Rodríguez, Díaz-Granados et al. 2008).

Como resultado de esto, diferentes entidades iniciaron la búsqueda del mejor sistema de tratamiento de aguas residuales para las dos cuencas faltantes de tratamiento y la ampliación de la PTAR Salitre, la EAAB-ESP realizó diferentes estudios (Convenio No. 29-2004 Water Research Centre, Convenio No. 9-07-24100-846-2005 Universidad Industrial de Santander – UIS y el contrato 1-02-26100-806-2006 HMV Ingenieros), donde se plantearon diferentes trenes de tratamiento: Tratamiento Primario Químicamente Asistido (TPQA) y varios tipos de tratamientos biológicos. Adicionalmente la CAR firmó el 20 de octubre de 2009 el contrato de consultoría No.000680 con el consorcio Hazen and Sawyer P.C. – Nippon Koei Co. Ltd. para realizar los estudios de alternativas para el tratamiento de las aguas residuales de las cuencas de los ríos Salitre, Torca y Jaboque, por

un valor de $3.150 millones y durante un plazo de 7 meses contados a partir del 4 de noviembre de 2009 (Consejo Nacional de Política Económica y Social 2009).

De acuerdo a lo anterior, se consideró por parte de los evaluadores, que la mejor opción de tratamiento es realizar un sistema biológico para la planta nueva y para la ampliación de la PTAR Salitre, mediante el proceso de lodos activados.
Éste método de tratamiento de aguas residuales, fue desarrollado en Inglaterra en 1914 por Ardern y Lockett y su nombre proviene de la producción de una masa activada de microorganismos capaz de estabilizar un residuo por vía aerobia (Metcalf and Eddy 1991) y cabe mencionar que este tipo de sistema de tratamiento biológico, es el más usado en el mundo. (Morla 2004; Holenda, Domokos et al. 2007; Dias, I. Moita et al. 2008; Zhang, Yuan et al. 2008; O'Brien, Mack et al. 2010).

3.1. Sistema de tratamiento de lodos activados

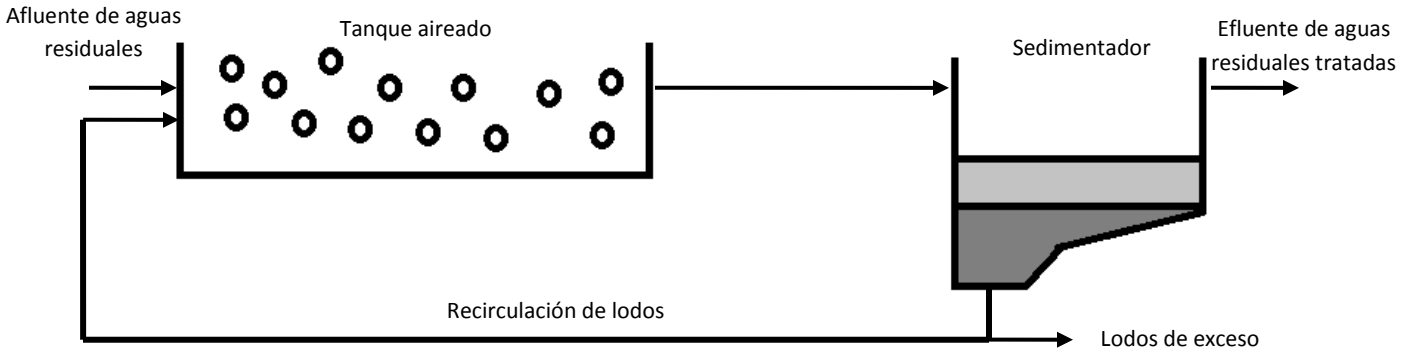
Los sistemas de lodos activados, son un tipo de tratamiento biológico de aguas residuales domésticas o industriales, en el cual microorganismos oxidan y mineralizan la materia orgánica (Lindberg 1997; Seyssiecq, Ferrasse et al. 2003; Mulas 2006).

Como ya se mencionó, el proceso fue desarrollado en Manchester, Inglaterra, en 1914 por Ardern y Lockett, para lo cual en esta década, las investigaciones sobre el tratamiento de aguas residuales se enfocaban en la aireación. Ardern y Lockett pudieron observar que después de un periodo de aireación, al detener este proceso, los flocs se sedimentaban decantado aquello que estaba flotando, el trabajo continuaba ingresando más agua residual y repitiendo este ciclo varias veces. Después de la acumulación de una cierta cantidad de biomasa, obtuvieron un efluente totalmente nitrificado en un periodo de 6 horas, para lo cual el lodo sedimentado fue llamado “Lodo activado” (Kayser 1999).

Una planta de lodos activados se caracteriza por cuatro elementos (Kayser 1999):

- Un tanque de aireación con un equipo apropiado para esta tarea, en el cual la biomasa se mezcla con las aguas residuales y una distribución de oxígeno en el tanque.
- Un clarificador final, en el cual la biomasa es removida del agua tratada por sedimentación u otros medios.
- Retorno y almacenamiento continúo de lodos y bombeo dentro del tanque de aireación.
- Retirada de exceso de lodos para mantener la concentración apropiada del líquido de mezcla.

Figura 1: Diagrama de una planta de lodos activados.



Fuente: (Mulas 2006)

3.1.1. Descripción del proceso

El residuo orgánico, se introduce en un reactor donde se mantiene un cultivo bacteriano aerobio en suspensión, este contenido es conocido como "líquido de mezcla", el cultivo bacteriano lleva a cabo la conversión en concordancia general con la estequiometría de las siguientes ecuaciones.

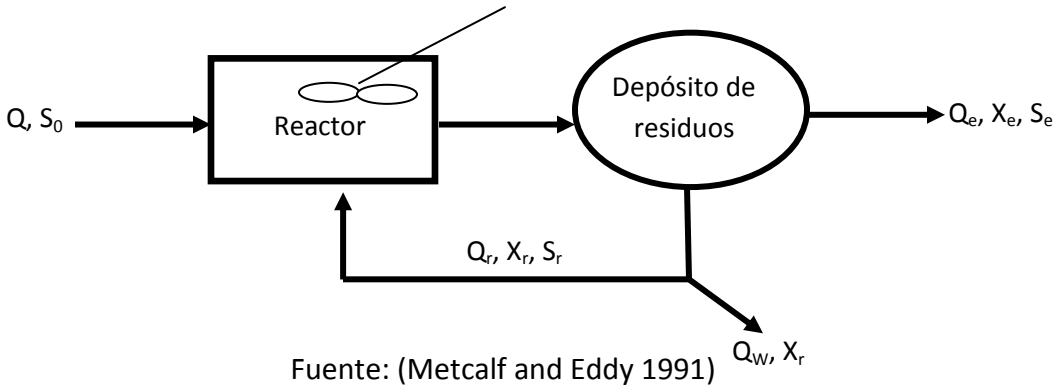
Oxidación y síntesis:

$$\boldsymbol{COHNS + O_2 + nutrientes \xrightarrow{bacterias} CO_2 + NH_3 + C_5H_7NO_2\ otros\ productos\ finales} \quad (1)$$

Respiración endógena:

$$\boldsymbol{C_5H_7NO_2 + 5\ O_2 \xrightarrow{bacterias} 5\ CO_2 + 2\ H_2O + NH_3 + energía} \quad (2)$$

Figura 2: Esquema de un reactor de mezcla completa con recirculación celular y purga.



Fuente: (Metcalf and Eddy 1991)

Donde:

Q_r = Caudal de entrada (afluente).
S_0 = Sustrato de entrada.
Q_e = Caudal de salida (efluente).
X_e = Biomasa del efluente.
S_e = Sustrato del efluente.
$Q_{r'}$ = Caudal de recirculación.
S_r = Sustrato recirculado.
$X_{r'}$= Biomasa recirculada.
Q_w = Caudal de desperdicio (purga).
X_r = Biomasa de purga.

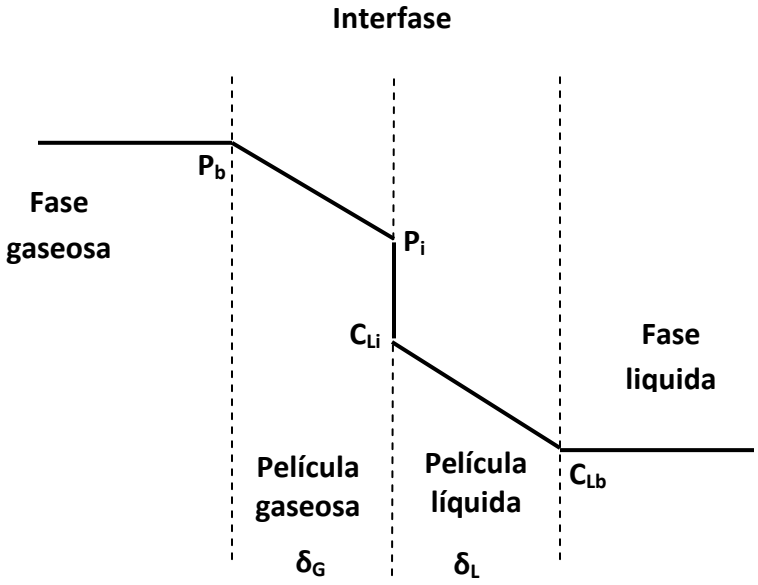
El ambiente aerobio en el reactor, se consigue mediante el uso de difusores o aireadores mecánicos que también sirven para mantener el líquido completamente mezclado (Metcalf and Eddy 1991), adicionalmente los microorganismos son mantenidos en suspensión por las burbujas de aire dentro del tanque o por el uso de agitadores. Como ya se menciono anteriormente, el oxígeno cumple varias funciones, una es para el uso de los microorganismos para oxidar la materia orgánica, la siguiente es para mantener la población microbiológica viva en el tanque(Lindberg 1997).

3.1.2. Aireación

Al ser un proceso aerobio, la trasferencia de gas al líquido es el éxito de este proceso, para esto existen diferentes teorías que explican el mecanismo de transferencia de gases, como la teoría de la doble capa por Lewis y Withman de 1924, el modelo de penetración de Higbie de 1935 y la teoría de renovación superficial por Dankwets de 1951. Sin embargo la teoría más usada para entender el mecanismo de la trasferencia de gases es la de doble capa (Metcalf and Eddy 1991).

La teoría de la doble capa es un modelo físico donde se establece, que la tasa de trasferencia se puede expresar en términos de un coeficiente global de transferencia y las resistencias en ambos lados de la interface (gas-liquido). En estas fases se presenta una resistencia al paso de las moléculas de gas de una fase a la otra. En el caso de gases poco solubles, la capa ofrece una mayor resistencia al paso de las moléculas de gas, de la fase gaseosa a la fase líquida, mientras que para el caso de gases más solubles es la capa gaseosa la que presenta una mayor resistencia (Metcalf and Eddy 1991; Stenstrom, Leu et al. 2006).

Figura 3: Esquema de la teoría de la doble capa.



Fuente: (Chapra 1997)

Donde:

δL = Espesor de película líquida.
δG = Espesor de la película de gas.
C_{Li} = Concentración de oxígeno en la película líquida de la interfase.
C_{Lb} = Concentración de oxígeno en el volumen liquido.
P_i = Presión parcial de oxigeno en la película de gas de la interfase.
P_b = Presión parcial de oxigeno en el volumen gaseoso.

En este modelos se realizan algunas suposiciones:

1. Perfil de concentración lineal a través de la película inactiva.
2. Condiciones de estado estable.
3. Equilibrio instantáneo en la interface.
4. El transporte por difusión no es una limitante.
5. Las soluciones están diluidas, por lo tanto se aplica la ley de Henry.
6. Existe flujo laminar en las películas de gas y agua.

Según la teoría del modelo, el gas se moviliza por difusión molecular a través de la película líquida y se distribuye por difusión turbulenta a través del líquido. Por tanto la velocidad de transferencia del gas depende de la resistencia de cualquier de las dos películas (gaseosa o liquida). La velocidad de difusión a través de la película de gas es proporcional a la concentración de soluto en la masa de aire y en la película delgada de gas (Kiely and Veza 1999). La difusión en la película de agua es controlada por la diferencia en

concentraciones entre C_{Li} y C_{Lb}. En la película de gas, existen menos moléculas de gas en comparación con la densidad elevada de moléculas en la película de agua, por lo tanto la resistencia a la difusión en la película del líquido es mayor que la de la película del gas (Chapra 1997; Kiely and Veza 1999).

La transferencia de masa de oxígeno, que representa la cantidad de soluto absorbido por unidad de tiempo por la difusión a través de las dos películas se muestra a continuación (Lewis and Whitman 1924).

Se tiene un N_0 que representa la trasferencia de oxígeno en condiciones de estado estable:

$$N_{oG} = N_{oL} \quad (3)$$ (no hay acumulación de gas en la película líquida)

Invocado la primera ley de Fick, se obtiene:

$$N_{oG} = K_G a\left(\frac{P_b M_W}{RT} - \frac{P_i M_w}{RT}\right) \quad (4)$$

$$N_{oL} = K_L a\left(C_{Li} - C_{Lb}\right) \quad (5)$$

Donde:

k_L = Coeficiente de transferencia de masa en la película liquida.
k_G = Coeficiente de transferencia de masa en la película gaseosa.
a = área de la interfase.
M_W = Peso molecular.

Igualando las ecuaciones 4 y 5 resulta:

$$K_G a\left(\frac{P_b M_W}{RT} - \frac{P_i M_w}{RT}\right) = K_L a\left(C_{Li} - C_{Lb}\right) \quad (6)$$

Para eliminar la presión parcial, utilizamos la ley de Henry:

$$P_b = HC_\infty^* \quad (7)$$
$$P_i = HC_{Li} \quad (8)$$

Donde:

H = Coeficiente de Henry para O_2 en agua.
C_∞^* = Concentración de oxigeno en el agua en equilibrio con el volumen de gas parcial.

El objetivo es resolver para la concentración de la interfase C_{Li}, ya que esta cantidad es esencialmente imposible de determinar. Sustituyendo las ecuaciones 7 y 8 dentro de la ecuación 6 y aplicado que $H\left(\frac{M_W}{RT}\right)$se tiene:

$$k_G(H_C C_\infty^* - H_C C_{Li}) = k_L(C_{Li} - C_{Lb}) \quad (9)$$

Resolviendo para C_{Li}

$$C_{Li}(k_L + k_G H_C) = k_G H_C C_\infty^* + k_L C_{Lb} \quad (10)$$

$$C_{Li} = \frac{k_g H_C C_\infty^* + k_L C_{Lb}}{k_L + k_G H_c} \quad (11)$$

Sustituyendo la ecuación 11 en la 5 se obtiene:

$$N_0 = k_L\left(\frac{k_G H_C C_\infty^* + k_L C_{Lb}}{k_L + k_G H_C} - C_{Lb}\right) \quad (12)$$

$$N_0 = k_L\left(\frac{k_G H_C C_\infty^* + k_L C_{Lb} - k_L C_{Lb} - k_G H_C C_{Lb}}{k_L + k_G H_C}\right) \quad (13)$$

$$N_0 = k_L\left(\frac{k_G H_C C_\infty^* - k_G H_C C_{Lb}}{k_L + k_G H_C}\right) \quad (14)$$

$$N_0 = \frac{k_L k_G H_C}{k_L + k_G H_C}(C_\infty^* - C_{Lb}) \quad (15)$$

$$N_0 = \frac{k_L}{\frac{k_L}{k_G H_C} + 1}(C_\infty^* - C_{Lb}) \quad (16)$$

Ahora tenemos

K_L = Coeficiente global de transferencia de masa

Se tiene

$$N_0 = K_L(C_\infty^* - C_{Lb}) \quad (17)$$

Aplicando la misma relación a la película de gas, se puede obtener:

$$N_0 = K_G(P_b - P_\infty^*) \quad (18)$$

Donde:

$P_\infty^* =$ = Presión parcial de oxigeno en el gas, en equilibrio el volumen de concentración líquida.

Por lo tanto, al ser este tratamiento un proceso aerobio, la capacidad de transferir oxígeno y la eficiencia de la aireación caracteriza el desempeño y los costos de las instalaciones de aireación en plantas de lodos activados (He, Petiraksakul et al. 2003) y convierte este proceso en la esencia del sistema, lo que conlleva a la necesidad de controlar la aireación.

3.2. Control de procesos

El objetivo primordial de los diferentes procesos que se realizan dentro de una planta de tratamiento de aguas residuales, es obtener un efluente de buena calidad con unas características determinadas, de forma que cumplan con los requerimientos o exigencias ambientales. Estas condiciones del efluente solo podrán ser posibles si se tiene un control exhaustivo sobre de las condiciones de operación de los sistemas.

El control de este tipo de sistemas, permite una operación del proceso más fiable y sencilla, al encargarse de obtener unas condiciones de operación estables, y corregir toda desviación que se pudiera producir en ellas, respecto a los valores de ajuste (Altuna Guevara 2009).

De acuerdo a Dunn (2005), el *control de procesos* es el control automático de una variable de salida, mediante la detección de una amplitud del parámetro de salida del proceso y comparándola con la salida deseada o ajustando el nivel con la alimentación de una señal de error. Por otra parte, para Smith (1985) el *control automático de proceso* es una actividad instrumental que controlan las diferentes variables, sin necesidad que intervengan operadores. Definiciones que son aplicables a esta investigación.

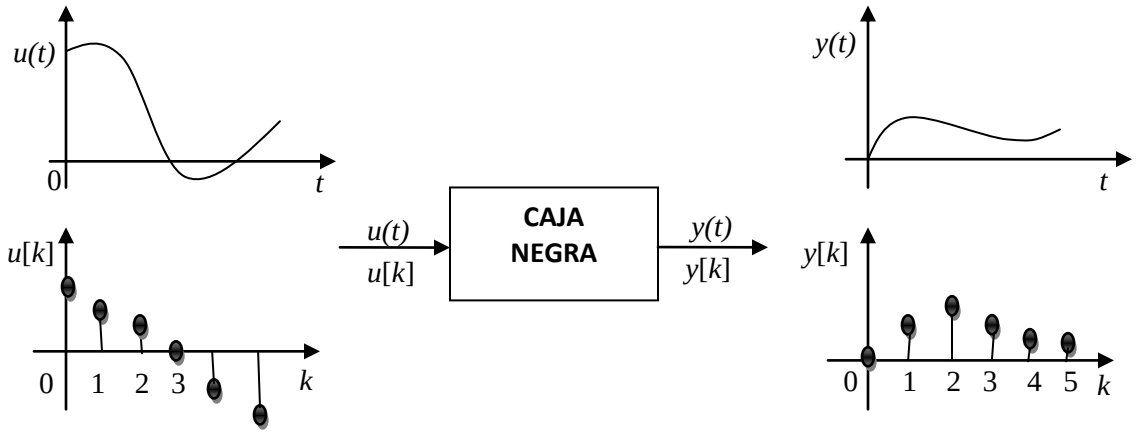
Las primeras maquinas y equipos controlados fueron primordialmente de manera natural, sin embargo a medida que avanzaba la historia y las necesidades humanas, los equipos, maquinarias y procesos cambiaban para convertirse más complejos y más aplicables para diferentes campos, por lo tanto creció la necesidad de controlar y optimizar estas acciones y equipos. De acuerdo a esto, para Raven (1978) existen dos razones primordiales para el control:

1. El control automático facilita al hombre muchas actividades monótonas que pueden dedicar constantemente a sus habilidades u otros esfuerzos.
2. El control moderno complejo puede desempeñar funciones que están fuera del alcance de las habilidades físicas humanas.

3.2.1. Sistemas de control

Todos los sistemas poseen una terminal de entrada y una terminal de salida como se muestra en la siguiente figura. Si se asume, que una excitación o entrada es aplicada sobre la terminal de entrada, una única respuesta o señal de salida puede ser medida en el terminal de salida. Esta única relación entre la excitación y la respuesta, entrada y salida, o causa y efecto es esencial en definir un sistema. Un sistema con una sola terminal de entrada y solo una terminal de salida son llamados *sistemas de una sola variable*. Un sistema de más de una variable de entrada o salida es llamado *sistema multivariable* (Chen 1998).

Figura 4. Sistema.



Fuente: (Chen 1998)

Un sistema es llamado *sistema de tiempo continuo,* si acepta señales de tiempo continuo como entrada y genera señales de tiempo continuo como salidas. La entrada se denota en cursiva $u(t)$ para una sola entrada o en negrilla $\boldsymbol{u(t)}$ para múltiples entradas. Si el sistema tiene p terminales de entrada, entonces $\boldsymbol{u(t)}$ es un vector p x 1 o $\boldsymbol{u} = [\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots\ldots \boldsymbol{u_p}]'$, donde la prima denota la traspuesta. Similarmente, la salida se denotara por $y(t)$ o por $\boldsymbol{y(t)}$. El tiempo t es asumido entre un rango de $-\infty$ hasta ∞ (Chen 1998).

Un sistema es llamado *sistema de tiempo discreto,* si este acepta señales de tiempo discreto como entradas y genera señales de tiempo discreto como salidas. Todas las señales de tiempo discreto en un sistema se asume que tiene el mismo periodo de muestreo T. La entrada y salida se denotan por $u[k] := u(kT)$ y $y[k] := y(kT)$, donde k es el

instante de tiempo discreto y es un número entero comprendido entre de -∞ hasta ∞. Estos se expresan en negrilla si tienen múltiples entradas y múltiples salidas (Chen 1998).

Ya conociendo que es un sistema, un sistema de control es una interconexión de componentes formando una configuración de sistema, que proporciona una respuesta deseada, la base para el análisis de un sistema está fundamentada por la teoría de sistemas lineales, lo que supone una relación causa efecto para los componentes. Entonces un componente o proceso a ser controlado puede ser representado por un bloque (Ver figura 5), donde la relación entrada salida representa la relación causa efecto del proceso (Dorf, Bishop et al. 2005).

Figura 5. Componentes básicos de un sistema de control.

Fuente:(Kuo and Golnaraghi 2003)

Los sistemas de control, básicamente poseen los siguientes componentes:

- Objetivo del control.
- Componentes del sistema de control.
- Resultados y salidas.

La relación básica de estos tres componentes se puede observar en la figura 5, la cual muestra una gran similitud con la figura 4. En términos generales, para comprender la figura 5, el **objetivo** puede ser identificado con las **entradas** o **señales** de **actuación** $\boldsymbol{u}$ y los resultados son también llamados **salidas** o variables controladas $\boldsymbol{y}$. En general, el objetivo del sistema de control, es vigilar las salidas de alguna manera prescrita, por medio de las entradas a través de los elementos del sistema de control (Kuo and Golnaraghi 2003).

Todo sistema de control posee cuatro componentes básicos, los cuales son (Smith and Corripio 1985):

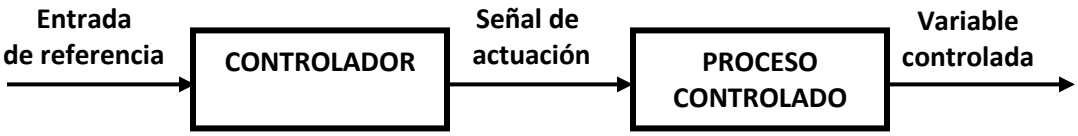
1. Sensor, que también se conoce como elemento primario.
2. Transmisor, el cual se conoce como elemento secundario.
3. Controlador, que el "cerebro" del sistema de control.
4. Elemento final de control, frecuentemente se trata de una válvula de control aunque no siempre. Otros elementos finales de control comúnmente utilizados son las bombas de velocidad variable, transportadores y motores eléctricos.

La importancia de estos componentes se fundamenta, en que realizan las tres operaciones básicas que deben estar presentes en todo sistema de control; estos son (Smith and Corripio 1985):

1. Medición ($\boldsymbol{M}$): la medición de la variable que se controla se hace generalmente mediante la combinación de sensores y transmisores.
2. Decisión ($\boldsymbol{D}$): con base en la medición, el controlador decide que hacer para mantener la variable en el valor que se desea.
3. Acción ($\boldsymbol{A}$): como resultado de la decisión del controlador se debe efectuar una acción en el sistema, generalmente esta es realizada por el elemento final de control.

Existen dos clases de sistemas de control, el primero se denomina *sistema de control de lazo abierto*, el cual utiliza un controlador o actuador para obtener la respuesta deseada, los elementos de este tipo de sistema pueden usualmente estar divididos en dos portes: el *controlador* y el *proceso controlado*, como se muestra en la siguiente figura (Kuo and Golnaraghi 2003).

Figura 6. Diagrama de bloque de un sistema de control de laso cerrado.

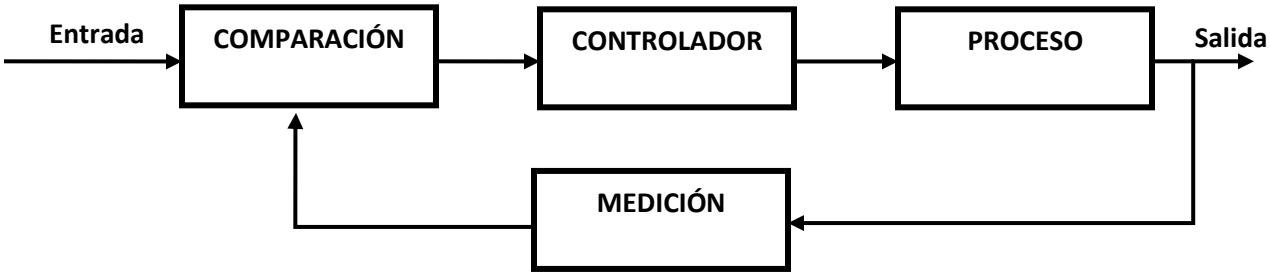


Fuente: (Kuo and Golnaraghi 2003)

Como se puede observar, la entrada es aplicada a un controlador cuya salida actúa como una señal sobre el proceso controlándolo, para obtener una variable de salida deseada (Kuo and Golnaraghi 2003).

El siguiente sistema se denomina *sistema de control de lazo cerrado* (*Sistema de Control con Retroalimentación*), este tipo de sistema cuenta con un enlace de retroalimentación, con el objetivo de ser más preciso; la señal controlada puede ser reutilizada y comparada con una referencia de entrada y así la señal de actuación es proporcional a la diferencia de la señal de entrada y la de salida, la cual debe ser enviada a través del sistema para corregir el error (Kuo and Golnaraghi 2003; Dunn William 2005).

Figura 7. Esquema de un sistema de control de lazo abierto (con retroalimentación).



Fuente: (McGraw-Hill 2005)

Existen dos tipos de retroalimentación posibles en el sistema de control cerrado: positivo y negativo, la retroalimentación positiva es una operación que aumenta el desequilibrio, por ejemplo, si un controlador de temperatura con retroalimentación positiva se utiliza para calentar una habitación, aumentaría el calor cuando la temperatura está por encima del punto de consigna y apagarlo cuando este por debajo, mostrando una propiedad que no es reguladora si no de extremos. Por el contrario el sistema de retroalimentación negativo trabaja por restablecer el equilibrio, continuando con el ejemplo, si la temperatura es demaciado alta el calor se reduce (Shinskey 1990).

Estos dos tipos de sistema pueden clasificarse dependiendo su finalidad, por ejemplo según el método de diseño y análisis, los sistemas de control se clasifican como lineales o no lineales y variables con el tiempo o no. De acuerdo con los tipos de señal encontrados en el sistema, a menudo se hace referencia a sistemas de datos continuos o discretos, modulados y no modulados (Kuo and Golnaraghi 2003).

3.2.2. Control de sistemas lineales y no lineales

El punto de partida en el análisis de un sistema de control es una representación por un modelo matemático, generalmente como un operador entre entradas del sistema, o como un conjunto de ecuaciones de diferencia y/o diferenciales. La mayoría de los modelos matemáticos usados tradicionalmente por teóricos y prácticos del control son lineales. De hecho, los modelos lineales son mucho más manejables que los no lineales y pueden representar en forma precisa el comportamiento de sistemas reales en muchos casos (Seron and Braslavsky 2001).

Prácticamente todos los sistemas físicos son no lineales de por sí, aunque muchas veces es posible describir su funcionamiento de modo aproximado mediante un modelo lineal, la caracterización matemática del comportamiento de los sistemas lineales es posible hacerla o bien en el dominio temporal o bien en el dominio trasformado. En el dominio temporal se trabaja con ecuaciones diferenciales lineales, compactables en la formulación

en ecuaciones de estado. En el dominio transformado lo habitual es la caracterización mediante funciones de trasferencia. En general en un sistema lineal es posible simultanear estas dos posibilidades de caracterización (Montoro López 1996).

Los sistemas de control con retroalimentación lineales, son modelos idealizados, para simplificar el análisis y diseño. Cuando las magnitudes de las señales en un sistema de control son limitadas a intervalos en los cuales los componentes del sistema exhiben características lineales (por ejemplo, el principio de superposición), el sistema es esencialmente lineal. Pero cuando las magnitudes de las señales se extienden más allá del alcance de la operación lineal, dependiendo de la severidad de la no linealidad, el sistema ya no se debe considerar lineal (Kuo and Golnaraghi 2003).

3.2.3. Sistemas de tiempo variable y no variable

Cuando el parámetro de un sistema de control es estacionario con respecto al tiempo durante la operación del mismo, este es llamado *sistema de tiempo no variable*. En la práctica la mayoría de los sistemas físicos contiene elementos que derivan o varían con el tiempo.

Sistemas de control de datos continuos

Un sistema de datos continuos, es uno en el cual la señal en varias partes del sistema es toda función del tiempo, variable continúa *t*. La señal en el sistema de datos continuos puede estar además clasificada como ac o dc. A diferencia de la definición general de señales ac y dc usada en la ingeniería electrónica, cuando se hace referencia al sistema de control ac, usualmente se refiere a las señales que en los sistemas son *moduladas* por alguna forma de esquema de modulación. Un sistema de control dc, por otra parte, simplemente implica que la señal es *no modulada* (Kuo and Golnaraghi 2003).

En la práctica, no todos los sistemas son estrictamente de tipo ac o dc, un sistema puede incorporar una mezcla de componentes de estos dos, usando moduladores y demoduladores para que coincidan las señales de varios puntos en el sistema.

Sistemas de control de datos discretos

Los sistemas de datos discretos difieren de los sistemas de datos continuos, en que las señales en uno o más puntos del sistema son en forma de un tren de pulso o un código digital. Usualmente estos sistemas son subdivididos dentro de una **muestra de datos** y **sistemas de control digital**. Los sistemas de muestra de datos, hacen referencia a la clase más general de sistemas de datos discretos, en el cual las señales son en forma de pulso

de datos. Un sistema de control digital hace referencia al uso de computadores digitales o controladores en el sistema, de modo que las señales son codificadas digitalmente, tal como en el código binario (Kuo and Golnaraghi 2003).

La clasificación de estos sistemas puede especificarse de acuerdo a la forma de un conjunto de ecuaciones y mezclarse entre ellos, por ejemplo, si un sistema es *no lineal*, *variable en el tiempo*, la ecuación de estado es:

$$\dot{x}(t) = a(x(t), u(t), t) \quad (19)$$

Un sistema *no líneal*, *sin variación en el tiempo* es representada por el conjunto de ecuaciones de la siguiente forma:

$$\dot{x}(t) = a(x(t), u(t)) \quad (20)$$

Si es un sistema *lineal*, *varía en el tiempo*, el conjunto de ecuaciones es:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad (21)$$

Donde $\boldsymbol{A}(t)$ y $\boldsymbol{B}(t)$ son matrices de n x n y n x m y con elementos que varían en el tiempo (Isidori 1995; Kretchmar 2000; Kirk 2004).

El conjunto de ecuaciones para sistemas *lineales*, *varían en el tiempo* tiene la siguiente forma:

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (22)$$

Donde $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices constantes (Isidori 1995; Kirk 2004).

Si las salidas son *no lineales*, *varían en el tiempo* las funciones de estado y control, se puede escribir la ecuación de salida como:

$$y(t) = c(x(t), u(t), t) \quad (23)$$

Si la salida está relacionada con los estados y el control por una relación *lineal*, *variables en el tiempo*, entonces

$$y(t) = Cx(t) + Du(t) \quad (24)$$

Donde $\mathbf{C}$ y $\mathbf{D}$ son matrices contantes de q x n y q x m.

3.3. Control de oxígeno en tanques de aireación

Para controlar el consumo energético que se da por el proceso de aireación, varios investigadores han planteado el uso de diferentes herramientas computacionales, una de las más usadas son los controladores, los cuales son dispositivos que modifican las condiciones de un sistema dinámico, ya sea por los datos de salida o entrada al sistema, estos se dividen en dos tipos: controladores de retroalimentación (feedback) y sistemas directos (feedforward), los primeros realizan un control de bucle cerrado, donde existe una retroalimentación continua de una señal de error del proceso, buscando una reducción gradual del error, en el segundo, el controlador del sistema emplea la lectura de una o más variables de entrada, para actuar sobre la variable manipulada que produce la salida deseada (Smith and Corripio 1985), cabe mencionar que estos dos tipos pueden combinarse, para compensar las aproximaciones del modelo, dado que la parte del controlador anticipativo atenúa las perturbaciones relativamente rápido comparado con los de retroalimentación (Vrecko, Hvala et al. 2003; Stare, Vrečko et al. 2007).

Los resultados encontrados en controladores de retroalimentación y directos han mostrado diferentes rangos en la reducción de costos energéticos, Vreckro et al. (2006) redujeron un 45% el caudal de flujo de aire aplicándole un controlador combinado (retroalimentación-directo), Ingildsen et al. (2002) encontraron que el consumo energético se puede reducir en un rango del 5 al 15%, por otra parte Zhang et al. (2008) mostraron que la energía usada en el proceso de aireación fue reducida a 4.7%, 7.9% y 3.7%, respectivamente bajo tres condiciones climáticas diferentes. O'Brien (2010) obtuvo una reducción del 20% en el consumo energético.

La Inteligencia artificial aplicada para el control de oxígeno

La Inteligencia Artificial (IA) se puede definir como la ciencia e ingeniería que hace maquinas inteligentes, especialmente programas inteligentes de computadoras, esto relaciona de forma similar con usar computadoras para comprender la inteligencia humana (McCarthy 2004), y se puede mencionar que (Herless and Castro 2004):

- **Como ingeniería**, el objetivo de la IA es resolver problemas reales, actuando como un conjunto de ideas acerca de cómo representar y utilizar el conocimiento, y de cómo desarrollar sistemas informáticos.
- **Como ciencia**, el objetivo de la IA es buscar la explicación de diversas clases de inteligencia, a través de la representación del conocimiento y de la aplicación que se da a éste en los sistemas informáticos desarrollados.

Muchos investigadores, dentro de la búsqueda de soluciones de problemas para diferentes campos de la ingeniería, medicina etc., aprovechan la informática; por las capacidades en la manipulación simbólica y la interferencia para resolver problemas de

razonamientos complejo y difícil en el nivel de rendimiento de los expertos humanos (Feigenbaum 1980).

Uno de estos problemas es el control en diferentes procesos, estos controles son usualmente considerados exclusivamente competencia de la teoría de control, sin embargo algunos problemas actuales sugieren combinar métodos de la teoría de control y la IA, ya que existe un incremento de problemas de control complejo, derivados de procesos que poseen comportamientos no lineales, estocásticos o no estacionarios, los cuales se convierten en menos faciles de reproducir o controlar (Gullapalli 1992).

La unión de la inteligencia artificial y la teoría del control, se genera gracias a que la inteligencia artificial, se fundó en parte para escapar de las limitaciones matemáticas de la teoría de control en los años 50. Las herramientas de inferencia lógica y computación, permitieron a los investigadores afrontar problemas relacionados con el lenguaje, visión y planificación, que estaban completamente fuera del punto de mira de la teoría del control (Russell and Norvig 2004). De acuerdo a las diferentes revisiones bibliográficas, se ha encontrado que las herramientas más utilizadas de la inteligencia artificial para el control de aireación, en un sistema de lodos activados son: la lógica difusa, los algoritmos genéticos, redes neuronales y ANFIS, cabe mencionar que actualmente se están utilizado otros métodos como el aprendizaje por refuerzo para el control de diferentes procesos.

Una de las herramientas computacionales más usada para este fin es la aplicación de lógica difusa, la cual, es un modelo presentado por Zadeh en 1965 (Sivanandam, Sumathi et al. 2007) y está clasificada como una técnica de inteligencia artificial. Es una herramienta emparentada con la técnica de los conjuntos difusos, una teoría que relaciona los objetos con límites no definidos, en los cuales la pertenencia a un conjunto es abordada desde la perspectiva de diferentes grados de certeza (Gutiérrez, Riss. et al. 2004). Este método se ha empleado con el propósito de mejorar la calidad del efluente, controlar la aireación y disminuir costos de operación (Mingzhi, Jinquan et al. 2009), debido a que las reglas son introducidas por expertos y operadores. El resultado de la lógica difusa al ser comparada con controladores convencionales ha presentando grandes beneficios energéticos sobre las PTAR (Kalker, VanGoor et al. 1999; Meyer and Popel 2003; Baroni, Bertanza et al. 2006).

Con esta técnica, Meyer y Popel (2003) encontraron en sus investigaciones una reducción del caudal de aire en el tanque del 23%, Ferrer (1998) obtuvo un ahorro del 40% de energía en comparación con un sistema convencional encendido/apagado de control de aireación, aplicándole el controlador difuso al reactor aeróbico, Baroni et al. (2006) pudieron determinar en una planta de un municipio (escala real) una reducción de 4% de energía.

Otro sistema muy aplicado son las redes neuronales artificiales, el cual es un modelo matemático inspirado en el funcionamiento biológico del sistema nervioso (red biológica neuronal). Las redes neuronales son una herramienta no lineal de modelación de datos,

usada para simulaciones complejas, relaciones entre datos de entrada y datos de salidas o para encontrar patrones de datos. En la mayoría de los casos las redes neuronales son un sistema adaptivo que cambia su estructura, basada en información externa o interna que fluye a través de la red durante la fase de aprendizaje, en otras palabras, el conocimiento es adquirido por la red a través del proceso de aprendizaje y las inter-conexiones entre los elementos que la red almacena para el conocimiento (Arbib 2003; Rustum 2009). Este tipo de controlador ha mostrado en las simulaciones que puede encontrar el apropiado caudal de entrada de aire, obteniendo una reducción en el costo de operación (Han and Qiao 2011).

La unión de la lógica difusa y las redes neuronales artificiales, llevó a los investigadores al uso de Sistemas de Inferencias Artificiales Neuro-Difusos ANFIS (por sus siglas en inglés Artificial Neuro-Fuzzy Inference Systems), estos sistemas además de tener las características de cada una de estos instrumentos, se beneficia de la transparencia propia de los sistemas difusos que mejora la comprensión de la red neuronal y además, la capacidad de aprender y adaptarse de las redes neuronales que aporta un mecanismo de sintonización automática y adaptabilidad al sistema difuso (Jang 1993; Jang and Sun 1995; Lin, Lee et al. 1996).

Mingzh et al. (2009) encontraron que aplicando un controlador basado en redes neuro-difusas, se obtuvo una reducción del 33% en el costo de operación, con respecto a la marcha normal de la planta.

La investigación en IA ha tenido grandes progresos en diferentes direcciones, uno de estos métodos es el aprendizaje por refuerzo (*reinforcement learning*), el cual se remonta a los primeros días de la IA, donde Arthur Samuel (1959) desarrollo un programa de ajedrez y del cual se han realizado avances para su aplicación en diferentes campos (Dietterich 1997), adicionalmente se ha popularizado por sus algoritmos simples y fundamentos matemáticos (Watkins 1989; Bertsekas and Tsitsiklis 1996; Sutton 1999).

Esta clase de aprendizaje puede definirse como el proceso de generación de experiencias que permite resolver un problema de decisión secuencial o control óptimo. La clave de este, se encuentra en la interacción continua entre un controlador que ejercita acciones de control que influyen en el estado del proceso y recibe una señal de recompensa o penalización, dependiendo del estado resultante de cada acción tomada (Martínez and de Prada 2003).

3.4. Tipos de aprendizaje

Para Hykin (1994) la clasificación de los algoritmos en la inteligencia artificial depende del intercambio de información que existe con el medio ambiente, de acuerdo con esto pueden desprenderse tres categorías de algoritmos de aprendizaje, basados en el tipo de información recibida por el medio ambiente: aprendizaje supervisado, aprendizaje por refuerzo y aprendizaje no supervisado (Kretchmar 2000).

El aprendizaje supervisado asume que el medio ambiente ofrece un profesor, por lo tanto el alumno supervisado recolecta la información del medio ambiente y produce una salida. El profesor brinda una retroalimentación en forma de una "salida correcta", la cual realiza cambios en los parámetros internos del algoritmo, para producir la siguiente decisión correcta cada vez que se observe el mismo estado (Kretchmar 2000).

El aprendizaje no supervisado observa los estados en el sistema y produce salidas, sin embargo un agente no supervisado no recibe retroalimentación del medio ambiente, en su lugar se ajustan vectores de parámetros estadísticos, para capturar la tendencia en la frecuencia y distribución de los estados (Kohonen 1990; Kretchmar 2000).

Entre los extremos del aprendizaje supervisado y no supervisado, se encuentra el aprendizaje por refuerzo, en el cual observando el estado del sistema produce una salida, el alumno recibe un refuerzo, una señal de evaluación desde el medio ambiente, que indica la utilidad de la salida. A través de ensayo y error, el alumno es capaz de descubrir los mejores resultados para maximizar la señal de evaluación (Kretchmar 2000).

Por lo tanto en el aprendizaje por refuerzo, el conocimiento es obtenido por el agente, ya que en el aprendizaje supervisado el conocimiento se debe dar por un fiscalizador externo(Sutton and Barto 1998), sin embargo no en todos los casos el fiscalizador conoce las respuestas correctas (Harmon and Harmon 1996). En cambio el aprendizaje por refuerzo es función del ensayo y error del agente con un medio ambiente dinámico (Kaelbling, Littman et al. 1996), donde no la primera acción indica la mejor de todas. Por otra parte otra gran diferencia con el aprendizaje supervisado es que en sistemas en línea el desempeño es alto, debido a que la evaluación de estos sistemas es a menudo concurrente con el aprendizaje (Kaelbling, Littman et al. 1996).

Figura 8 Esquema de aprendizaje supervisado.

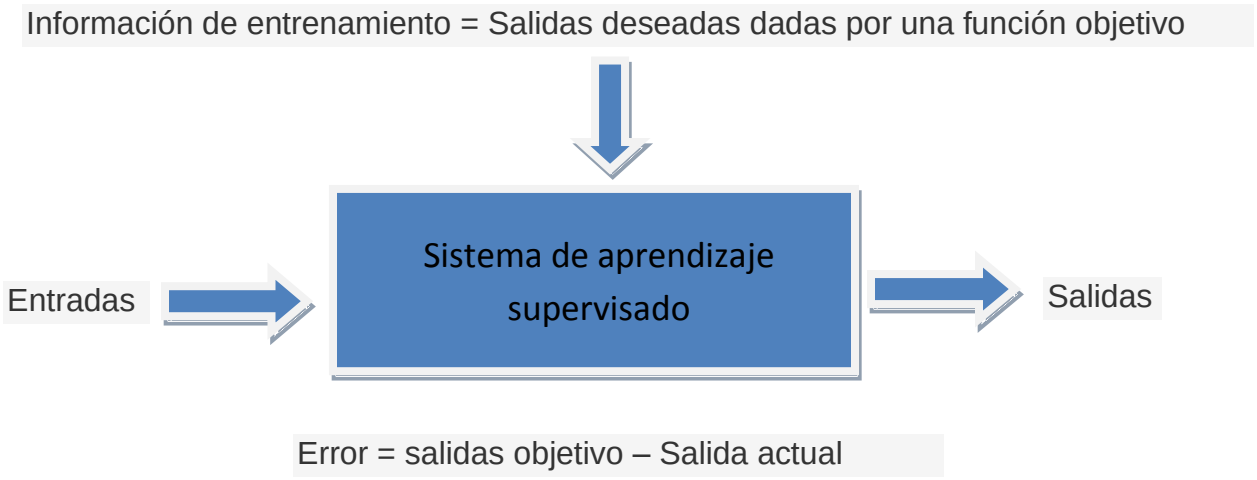
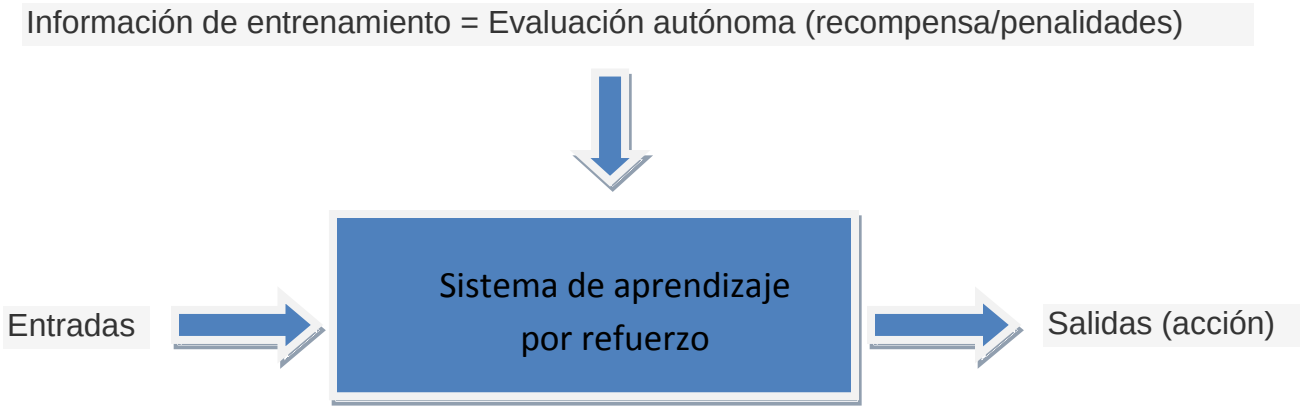


Figura 9 Esquema de aprendizaje por refuerzo.



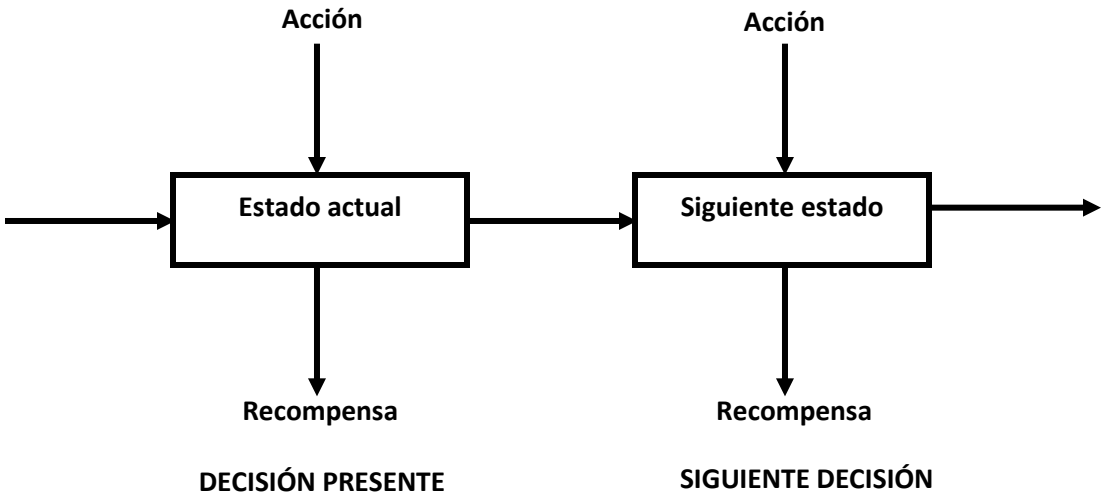
Fuente: (Dolk 2010)

De acuerdo con lo anterior, el aprendizaje por refuerzo se adapta muy bien al entorno de control, ya que tiene la capacidad de optimizar en el tiempo, especialmente si se desea minimizar el error cuadrático medio sobre el tiempo. En la mayoría de los problemas de control, cada acción de control tiene un efecto no solo inmediato si no también el error del siguiente paso (Sutton and Barto 1998).

3.5. Aprendizaje por Refuerzo (*AR*)

El aprendizaje por refuerzo tiene una gran relación con los modelos de decisiones secuenciales. Estos modelos pueden representarse gráficamente por la figura 10, que describe como en un punto del tiempo, un agente observa un estado de un sistema, de acuerdo a lo existente dentro de este, elige una acción y esta acción produce dos resultados: una recompensa y un nuevo estado, consecuencia de la distribución de probabilidad determinada por la acción elegida (Puterman 1994).

Figura 10 Representación de un modelo de decisiones secuenciales.



Fuente: (Puterman 1994)

Por lo tanto, los elementos que interviene en este modelo son los siguientes:

- Un conjunto de decisiones.
- Un conjunto de estados del sistema.
- Un conjunto de posibles acciones.
- Un conjunto de estados y acciones dependientes de una recompensa inmediata.
- Un conjunto de estados y acciones dependientes de una probabilidad de transición.

Adicionalmente se cuenta con una política y unas reglas de decisión, el cual es un modelo muy asociado al aprendizaje por refuerzo y base teórica de los procesos de decisión de Markov.

3.5.1. Generalidades

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 este tiene una conexión directa mediante un sensor que conecta con el ambiente, este ejercicio de esta conexión produce una gran cantidad de información, a través de causa y afecto por consecuencia de las acciones y el cumplimiento de sus objetivos (Sutton and Barto 1998).

Este concepto de aprendizaje mediante el ensayo y error para lograr un objetivo, es el fundamente teórico del AR, el cual puede entenderse de manera sencilla; como un agente el cual desconoce su medio ambiente y está conectado a este, mediante la percepción del mismo, el agente interactúa con el ambiente y aprende de este y la acción tomada por su interacción, nuevamente interactúa y toma una decisión de este nuevo estado, almacenando esta información y la recompensa dada por su acción, de tal manera que a medida que va aprendiendo mediante sus nuevas interacciones y lo almacenado, elige la mejor acción, la que representa el mayor valor en la recompensa dada al agente (Sathiya Keerthi and Ravindran 1994; Kaelbling, Littman et al. 1996; Sutton and Barto 1998; Sutton 1999; Ahn 2009; Gosavi 2009; Whiteson 2010).

Es importante mencionar que este tipo de problemas con estas características son mejor descritos, con la estructura de Procesos de Decisión de Markov, el cual se explicará más adelante (Szepesvári 2010).

Como se puede ver, los actores principales son el agente y el ambiente, sin embargo se pueden identificar cuatro subelementos dentro del modelo: una *política*, *una función de recompensa* y una *función de valor* (Sutton and Barto 1998).

1. **Elementos**

   - *Agente,* es un sistema computacional que habita en un entorno complejo y dinámico, con la capacidad de percibir y actuar autónomamente sobre dicho entorno, y de esta manera es capaz de cumplir un conjunto de objetivos o llevar a cabo ciertas tareas para las cuales fue diseñado (Maes 1995). Lo más importante es mencionar que su comportamiento esta dado por su propia experiencia (Ribeiro 2002).

   - *Ambiente,* es el entorno existente que no es el agente y que es de interés para llevar a cabo la tarea que se le ha asignado al agente. Este ambiente obliga al agente a realizar un conjunto de acciones sobre él (Choi, Yim et al. 2009).

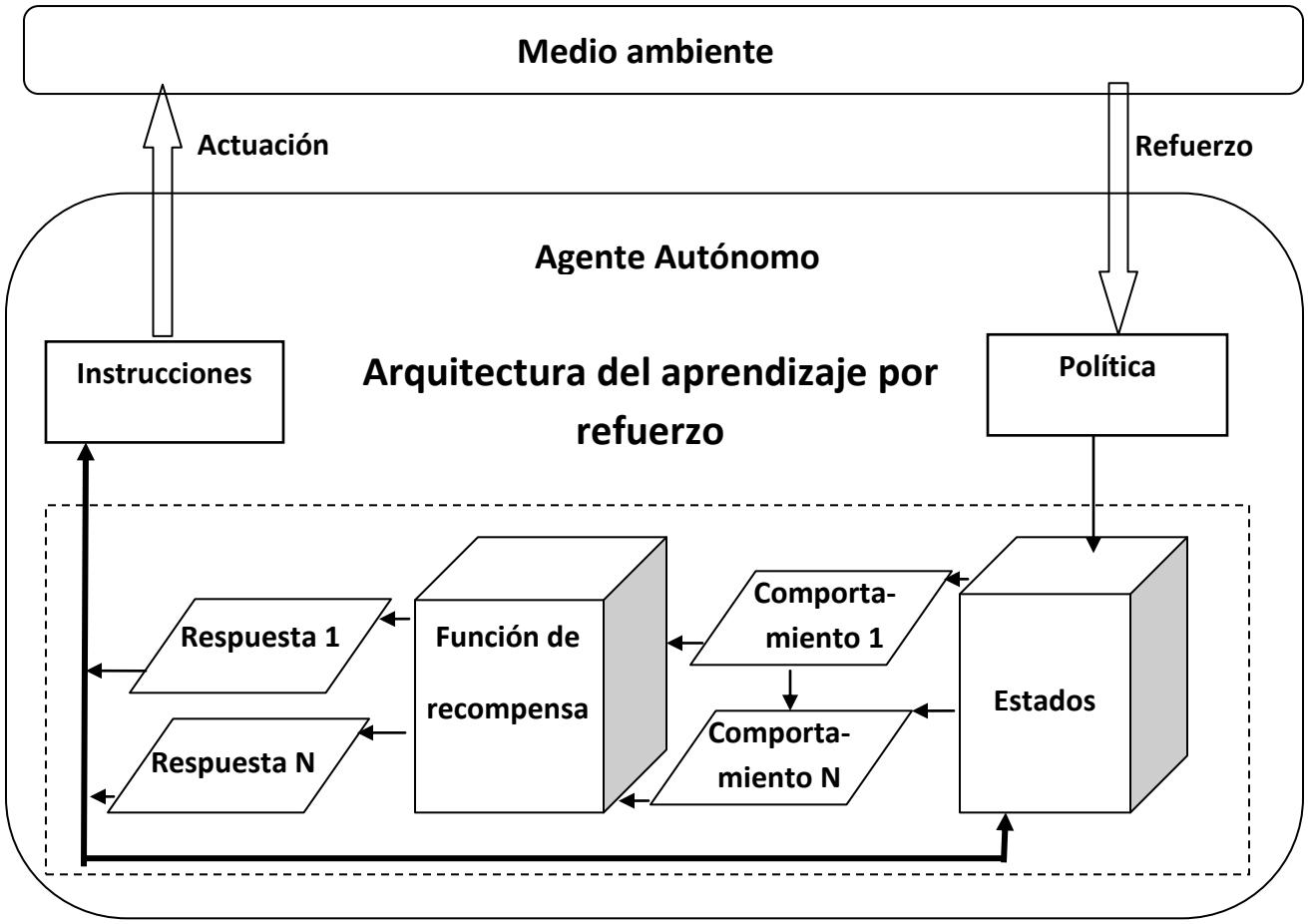
2. Subelementos

- La *política,* define la forma de aprendizaje y comportamiento del agente en un momento dado, en términos generales es un mapa de los estados percibidos en el medio para las acciones tomadas en esos estados (Sutton and Barto 1998).

- *Función de recompensa,* es un valor escalar que indica lo deseable que es una situación para un agente. La recompensa puede tomar valores tanto positivos como negativos. Fisiológicamente podría compararse un valor de recompensa negativo con el dolor y un valor positivo con el placer. Cada vez que el agente ejecuta una acción, recibe un valor de recompensa. Estas recompensas no tienen por qué estar asociadas directamente con la última acción ejecutada, sino que pueden ser consecuencia de acciones anteriores llevadas a cabo por el agente.

- La *función valor,* indica que es bueno a largo plazo, en términos generales el valor de un estado es la cantidad total de recompensas que un agente puede esperar acumular en el futuro, ya que las recompensas determinan la conveniencia inmediata.

Figura 11 Arquitectura del aprendizaje por refuerzo.

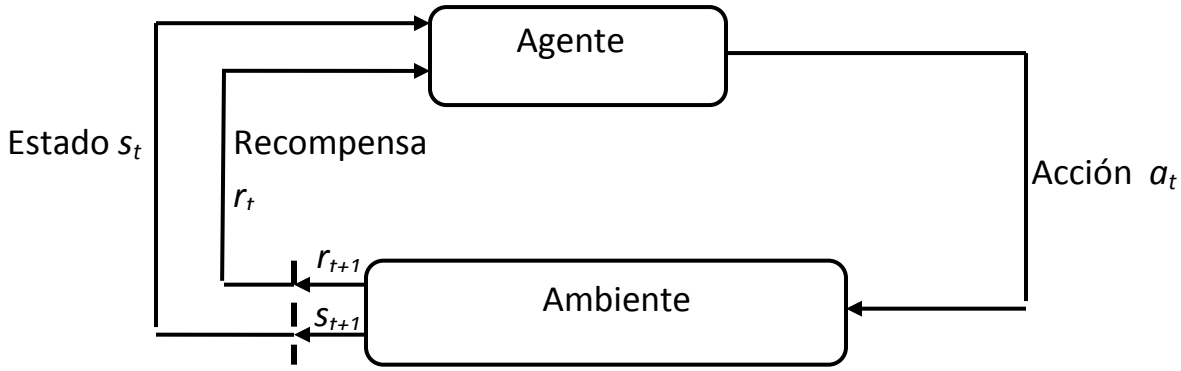


Fuente: (Choi, Yim et al. 2009)

Para entender claramente este proceso podemos describirlo de la siguiente manera (Sutton and Barto 1998; Glorennec 2000):

- En el intervalo de tiempo discreto t, el agente interactúa con el medio ambiente y recibe una representación de estado ambiental $s(t) \epsilon S$, donde S es un conjunto de todos los posibles estados.
- Éste escoge una posible acción $a(t) \epsilon A$, donde A es un conjunto de todos las posibles acciones. en el estado $s(t)$,
- Aplica la acción, que provoca;
    - la recepción del refuerzo o recompensa, $r(t) \epsilon R$;
    - pasa a un nuevo estado, $s(t+1)$,

Figura 12 Esquema de interacción del agente con el ambiente



Fuente: (Sutton and Barto 1998)

En cada intervalo de tiempo, el agente implementa un mapeo de los estados para seleccionar cada posible acción, este mapeo se denota como $\pi(t)$ (*política*).

El método de aprendizaje por refuerzo especifica, como el agente cambia su política como resultado de su experiencia, el objetivo del agente en términos generales es maximizar la cantidad total de recompensa que se recibe a largo plazo (Sutton 1992; Littman 1994; Kaelbling, Littman et al. 1996; Hu and Wellman 1998; Sutton and Barto 1998; Sutton 1999).

Por lo tanto la medición de recompensa a largo plazo es denominada *retorno* o *función de retorno* y se denota $\mathrm{R(t)}$, el cual puede ser definido como; la suma sobre un número finito o infinito en términos de un caso (Sutton and Barto 1998), para hallar el retorno, existen tres posibles expresiones para calcular el *retorno*:

Modelo de horizonte finito

En este caso, el horizonte corresponde a un número finito de pasos, en el cual existe un estado final y la secuencia de acciones entre el estado inicial y el final se llama un periodo (Sutton and Barto 1998; Glorennec 2000). En otras palabras significa que hay un tiempo prefijado *N*, después del cual no importa nada (Russell and Norvig 2004).

$$R(t) = r_t + r_t + r_{t+1} + r_{t+2} + r_{t+3} \dots\dots\dots . . r_T \quad (25)$$

Donde *T* es el tiempo final. Este modelo puede ser usado de dos formas, en la primera, el agente tendrá una política no estacionaria, que es uno de los cambios en el tiempo, en su primer paso se llevará a lo que es denominado *paso-h acción óptima*. Esta se define como la mejor acción disponible dada en el *paso-h,* ya que restantes pasos tendrán ganancia en el refuerzo. Sobre el siguiente paso se tendrá al paso (*h* -1) y así hasta que se finalice. En la segunda opción, el agente siempre actúa de acuerdo a la misma política. Este modelo no

siempre es apropiado, ya que en algunos casos no se puede saber la longitud exacta de la vida del agente (Kaelbling, Littman et al. 1996).

Retorno con descuento (modelo de horizonte infinito)

Se emplea este cuando la secuencia de acciones es infinita, existe en factor denominado *factor de descuento* y se denomina γ, que tiene valores de $0 \leq \gamma \leq 1$. Este factor asegura que la suma de las recompensas es finita y también adhiere más peso a las recompensas recibidas a corto plazo en comparación con las recibidas a largo plazo (Rummery 1995; Sutton 1999).

$$R(t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (26)$$

El valor de γ permite modular el periodo en el que el agente toma en cuenta los refuerzos, si γ es = 0 el agente es "oportunista" por considerar la recompensa actual, mientras que un factor cercano a 1 el agente se esforzará más.

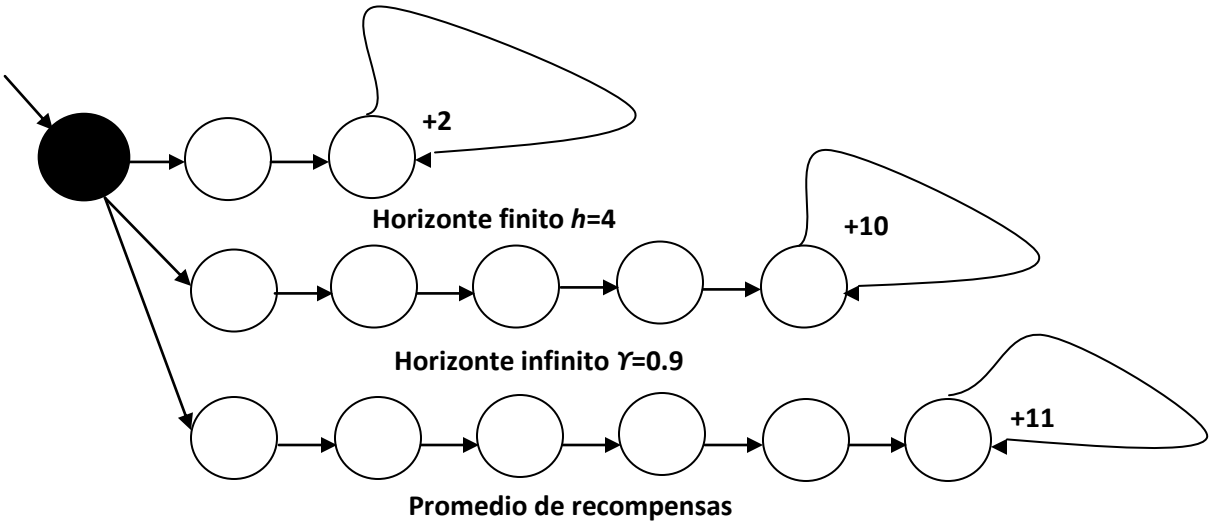
El modelo de recompensas promedio

En este, el agente supone que la acción que optimiza a lo largo del tiempo es el promedio de las recompensas (Kaelbling, Littman et al. 1996).

$$R(t) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n} r_{t+k} \quad (27)$$

Kaelbling *et al*. (1989), realizaron una comparación de los tres modelos del retorno mediante un ejemplo, el cual consistia en realizar una acción por parte de un agente sobre cada estado. Gráficamente el ejemplo se puede ver en la figura 13; los circulos representan los estados en el ambiente y las flechas son la transición entre estos, sobre cada estado se ejecuta una unica acción exeptuando el circulo negro el cual es el estado inicial, todas las recompensas son cero exceptuando donde se encuentra marcado.

Figura 13 Comparación de modelos.



Fuente: (Kaelbling, Littman et al. 1996)

Bajo el modelo de horizonte finito con h = 5, las tres acciones tiene valores de +6.0, +0.0 y +0.0, entonces la primera acción es la seleccionada, en el modelo de horizonte infinito con un γ = 0.9, las tres opciones son +16.2, +59.0 y 58.5, entonces la acción seleccionada puede ser la segunda y en el modelo de recompensa promedio, la tercera acción puede ser la elegida ya que el promedio de las recompensas es de +11.0. Se modificó el ejemplo cambiando h hasta 1000 y γ hasta 0.2, los resultados dados fueron; la segunda acción es la óptima para el modelo de horizonte finito y la primera para el modelo de horizonte infinito, sin embargo el modelo de recompensas promedio siempre prefirió el mejor de los promedios a largo plazo.

El modelo de horizonte finito es apropiado cuando se conoce el tiempo de vida del agente, por lo tanto con un horizonte finito, la acción óptima para un estado dado puede cambiar a lo largo del tiempo. Decimos que la política óptima para un horizonte finito es **no estacionaria**. Por otra parte, sin una alimentación establecida de tiempo, no hay razón de comportarse de manera diferente en los mismos estados, en diferentes instantes de tiempo. Por lo tanto la acción óptima sólo depende del estado actual y la política óptima es **estacionaria** (Russell and Norvig 2004).

El modelo más usado es el de horizonte infinito, ya que sus políticas son más sencillas que las de horizontes finitos (Russell and Norvig 2004), cabe mencionar que el modelo promedio de las recompensas es relativamente nuevo y complejo (Kaelbling, Littman et al. 1996; Glorennec 2000; Russell and Norvig 2004).

3.6.Estructura del aprendizaje por refuerzo

Procesos de decisión de Markov (PDM)

Los *Procesos de decisión de Markov* o *cadena controlada de Markov*, puede entenderse como un conjunto de posibles acciones, recompensas y probabilidades de transición que dependen solo del presente estado y su acción y no de los estados ocupados y acciones elegidas en el pasado (esto también se conoce como la propiedad de Markov). Estas acciones son ejecutadas por un agente cuyo objetivo es escoger una secuencia de acciones, las cuales causen en un sistema el desempeño óptimo con respecto a algún criterio optimo de desempeño (Puterman 1994). Por esta razón, los PDM se han estudiado en diferentes ámbitos que incluyen la Inteligencia Artificial, la investigación operativa, la economía y la teoría de control (Russell and Norvig 2004).

Este tipo de problemas están compuestos por (Watkins 1989; Puterman 1994; Feinberg and Shwartz 2002):

- Una decisión: La cual se hace en cualquier momento del tiempo, el conjunto de estas decisiones se denota con la letra $\mathcal{T}$, este conjunto puede ser clasificado de dos formas; la primera como un conjunto discreto o continuo, y como un conjunto finito o infinito. Cuando este es discreto, se toman las decisiones en las épocas de decisiones, por otra parte cuando este es continuo, la decisión puede ser hecha mediante:

    - Todas las épocas de decisiones.
    - Puntos aleatorios del tiempo cuando ciertos eventos ocurren, tales como llegadas a una cola del sistema.
    - En un tiempo oportuno elegido por el agente.

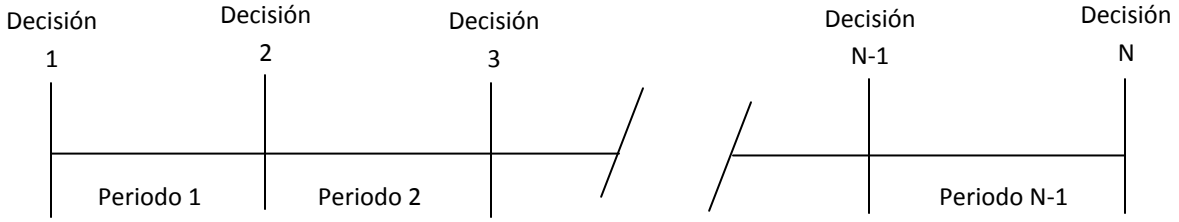
  En problemas de tiempo discreto, el tiempo es dividido en *periodos* o *etapas*, como se puede ver en la figura 14, donde la decisión de cada periodo es tomada al inicio de cada periodo, así si $\mathcal{N}$, es infinito, el problema de decisión es denominado *horizonte infinito* si no se denomina *horizonte finito*.
- Un conjunto de estados $\mathcal{S}$.
- Una función $\mathcal{A}$ que da la posibilidad de acción para cada estado, esta podrá verse un conjunto $\mathcal{A}(x)$ acciones posibles en el estado $s \in \mathcal{S}$;
- Una función de recompensa $\mathcal{R}$ denota la recompensa paso a paso usando una acción a en un estado s; entonces la recompensa podrá verse $r_t(s,a)$, cuando esta es positiva puede considerarse como una ganancia pero si esta es negativa será una pérdida. La ganancia puede verse como:

    - Una suma global recibida en un tiempo fijo o aleatorio de la próxima decisión,
    - acumulada continuamente durante todo el periodo actual,

- una cantidad aleatoria que depende del estado del sistema en la decisión posterior, o
- una combinación de las anteriores.

- La función de probabilidad de transición se denota por $p_t(j|s,a)$: es la probabilidad de ir a un estado s' dado que el agente se encuentra en el estado s y ejecuta la acción a.
- Normas de decisión: prescribe un procedimiento para la selección de una acción en cada estado,
- política específica la norma de decisión para ser usada en todas las decisiones, el objetivo de esta es maximizar el valor esperado de la suma de recompensas a largo plazo. Adicionalmente se define una política óptima, como aquella que maximiza el valor esperado de recompensa total, partiendo de un estado i para cierto número de transiciones n.

Figura 14 Decisión y periodos.



Fuente: (Puterman 1994)

Considerando como en un medio ambiente general podrá responder en el tiempo $t + 1$ a las acciones tomadas a un tiempo t, en manera más general, el causal de esta respuesta puede depender de todo lo que ha sucedido antes, en este caso, la dinámica solo se puede definir mediante la especificación de la distribución de probabilidad (Barto, Sutton et al. 1989; Sutton and Barto 1998) para todos los s', r, y todos los posibles valores de eventos pasados: s_t, a_t, r_t, …. . , r_1, s_0, a_0:

$$\Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, \ldots\ldots, r_1, s_0, a_0\} \quad (28)$$

Si la señal de estado tiene la propiedad de Markov, entonces la respuesta del medio ambiente en $t + 1$ solamente dependen del estado y la acción representada en t, en cuyo caso la dinámica del medio ambiente puede ser definido solo para todos los s', r, s_t, y a_t.

$$\Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t\} \quad (29)$$

En otras palabras una señal de estado tiene la propiedad de Markov y es un estado de Markov, si e solo si la ecuación (28) es igual a la (29) para todos los s', r, y las historias de

s_t , a_t , r_t , ……. , r_1 , s_0 , a_0 , en este caso el medio ambiente y la tarea como en su conjunto también se dice que tiene la propiedad de Markov.

Si un medio tiene la propiedad de Markov, entonces su paso dinámico (ecuación 29), permite predecir el siguiente estado y la próxima recompensa esperada, dado el estado actual y la acción. Se puede demostrar que iterando la anterior ecuación se logra predecir el futuro a todos los estados y las recompensas esperadas, conociendo solo la situación actual, así como si se obtuviera la historia completa hasta el momento actual. También se deduce que los estados de Markov proporcionan la mejor base posible para las acciones de la elección, es decir, la mejor política para la elección de las acciones en función de un estado de Markov es tan buena como la mejor política para la elección de las acciones en función de la historia completa.

Como se menciono anteriormente el procedimiento del aprendizaje por refuerzo satisface la propiedad de Markov, por lo tanto si los estados y las acciones son finitos, entonces se llama un proceso finito de decisión de Markov, los cuales son particularmente importantes en la teoría del aprendizaje por refuerzo.

Un modelo PDM finito está definido por su estado, el conjunto de acciones y por la dinámica de un solo paso en el medio ambiente, dado un estado y la acción, s y a, la probabilidad de cada posible siguiente estado s', es

$$P_{ss'}^{a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \quad (30)$$

Estas cantidades son llamadas *transición de probabilidades*, del mismo modo, dado un estado actual y la acción, s y a junto con cualquier otro estado s' el valor esperado de la próxima recompensa es

$$R_{ss'}^{a} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \quad (31)$$

Estas cantidades, $P_{ss'}^{a}$ y $R_{ss'}^{a}$, especifican completamente los aspectos más importantes de la dinámica del modelo PDM finito.

Función de valor

Casi todos los algoritmos de aprendizaje por refuerzo se basan en la estimación de las funciones de valor, los cuales estiman que tan bueno es para el agente estar en un determinado estado o lo bueno que es para llevar a cabo una acción determinada en el mismo estado, en consecuencia las funciones de valor se definen con respecto a determinadas políticas (Harmon and Harmon 1996; Sutton and Barto 1998; W. Josemans 2009).

La política π, es el mapeo que se hace en cada estado, $s \in S$ y acción $a \in A(s)$. Informalmente el valor de un estado s bajo una política π, se denota como $V^{\pi}(s)$, es el

retorno esperado cuando iniciamos en un estado s y seguimos una política π, entonces para un PDM, se puede definir $V^{\pi}(s)$ como (Sutton and Barto 1998; Glorennec 2000; Szepesvári 2010):

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\} \quad (32)$$

Donde $E_{\pi}\{\}$ denota el valor esperado, dado que el agente sigue la política π y t es cualquier paso en el tiempo, nótese que el valor del estado terminal en su caso siempre es cero, la ecuación anterior es conocida como *función de valor de estado para una política* π (Barto and Mahadevan 2003; Szepesvári 2010).

Similarmente se define que el valor de tomar un acción a en un estado s bajo una política π se denota por $Q^{\pi}(s,a)$, como el retorno esperado a partir de s tomándola acción y posteriormente siguiendo la política π:

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\} \quad (33)$$

La ecuación anterior es conocida como *función de valor de acción para una política* π (Barto and Mahadevan 2003; Szepesvári 2010).

Las funciones de valor V^{π} y Q^{π} , pueden ser estimados de la experiencia, por ejemplo, si el agente sigue una política π y mantiene un promedio para cada estado encontrado, del retorno real que se ha seguido a ese estado, entonces el promedio convergerá para el valor del estado $V^{\pi}(s)$, como el número de veces que el estado encuentra aproximaciones infinitas, si los promedios se mantiene por separado para cada acción realizada en un estado, entonces estos promedios convergerán a los valores de acción $Q^{\pi}(s,a)$ (Sutton and Barto 1998).

Para cada PDM existe una *función de valor optima* V^* de tal manera que $V^*(s) = \max_{\pi} V^{\pi}(s)$, una *función de acción optima*, Q^* de tal manera que $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$, las cuales se pueden expresar como las ecuaciones optimización de Bellman (Sutton and Barto 1998; Whiteson 2007; Whiteson 2010):

$$V^*(s) = max_a \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^*(s')] \quad (34)$$

y

$$Q^*(s,a) = \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V^*(s')] \quad (35)$$

o

$$Q^*(s,a) = \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma max_a Q^*(s',a')] \quad (36)$$

Las ecuaciones 34 y 36 son las dos formas de las ecuaciones de optimalidad de Bellman para Q^* y V^*, que formulan el principio de optimalidad, descrito a continuación:

Principio de Optimalidad: *Una política óptima tiene la propiedad de que cualesquiera que sean el estado inicial y la primera decisión tomada; el resto de decisiones debe construir una política óptima respecto al estado resultante de la primera decisión.*

Sabiendo ya con anterioridad que la política es la suma esperada de las recompensas obtenidas, donde la esperanza se toma sobre todas las posibles secuencias de estados que puedan darse, cuando se ejecuta la política. Una política óptima π^* vendrá dada por (Russell and Norvig 2004):

$$\pi^* = argmax\ E[\sum_{t=0}^{x} \gamma^t R(s_t)\,|\pi] \quad (37)$$

3.7. Métodos de solución del aprendizaje por refuerzo

De acuerdo a Sutton y Barto (1998), existen tres tipos de métodos fundamentales para resolver los problemas de aprendizaje por refuerzo: **Programación Dinámica,** métodos de **Monte Carlo** y el **Aprendizaje de Diferencias Temporales**; cada una de estas clases de métodos tiene ventajas y debilidades. La programación dinámica está muy bien desarrollada matemáticamente, sin embargo requiere un completo y preciso modelo del medio ambiente, por otra parte los modelos de Monte Carlo no requieren un modelo y son conceptualmente simples, pero no son adecuados para incrementos computacionales paso a paso. Finalmente, los métodos de diferencias temporales no requieren un modelo y son totalmente incrementales, pero son muy complejos para analizar. Adicionalmente, estos métodos también difieren en diferentes formas con respecto a su eficiencia y velocidad de convergencia (Sutton and Barto 1998). Es importante mencionar que estos métodos pueden combinarse para obtener mejores las características de cada uno de ellos.

3.7.1. Programación dinámica (PD)

El término de programación dinámica hace referencia a una colección de algoritmos que pueden ser usados para calcular una política óptima, dada en un modelo perfecto de un ambiente como un PDM. Los algoritmos de PD clásicos son de utilidad limitada en el aprendizaje por refuerzo, ya que se asume un modelo perfecto y por sus grandes gastos computacionales, pero teóricamente son muy importantes (Sutton and Barto 1998; Busoniu, Babuska et al. 2010).

Al igual que en el resto de técnicas de aprendizaje por refuerzo, la PD se basa en el mantenimiento de funciones de valor de cada estado (V (s) y/o Q(s,a)) para estructurar el conocimiento. Obteniendo valores óptimos de ambas funciones se obtiene una representación óptima de la política a seguir. (Fernández 2002).

Los métodos de PD toman tiempo para encontrar una política óptima y su aplicabilidad es limitada, por la *maldición de la dimensionalidad*, la cual hace referencia a diversos fenómenos que surgen al analizar y organizar los espacios de grandes dimensiones (a menudo con cientos o miles de dimensiones) que no se producen en bajas dimensiones. Como el número de estados crece exponencialmente con el número de variables de estado, el tiempo de cálculo se afecta en problemas con espacios de estados grandes. Ya que los métodos de la PD requieren a lo largo de todo el conjunto de los estados de PDM, se debe barrer el espacio de estado varias veces. Esto no es práctico para espacios de estados grandes. Afortunadamente, el aprendizaje se lleva a cabo en cada paso, no se tiene que esperar hasta la recompensa que se recibe para actualizar las estimaciones del valor de los estados o pares de estado-acción. Los métodos de PD utilizan una técnica llamada *bootstrappinp* (de arranque), lo que significa que las estimaciones de valor se actualizarán sobre la base de las estimaciones de otros valores (Taylor 2004).

Si se conoce el modelo del ambiente, ósea las transiciones de probabilidad ($P_{ss'}^{a}$) y los valores esperados de recompensas ($R_{ss'}^{a}$), las ecuaciones de optimalidad de Bellman representan un sistema de |S| ecuaciones y |S| incógnitas. La idea principal de la PD, es el uso de las funciones de valor para organizar y estructurar la búsqueda de una buena política (Morales 2011).

Si se considera primero calcular la función de V^{π} dada una política arbitraria π.

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} \quad (38)$$

$$V^{\pi}(s) = E_{\pi}\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s\} \quad (39)$$

$$V^{\pi}(s) = E_{\pi}\{r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s\} \quad (40)$$

$$V^{\pi}(s) = \sum_a \pi(s,a) \sum_{s'} P_{ss'}^{a} [R_{ss'}^{a} + \gamma V^{\pi}(s')] \quad (41)$$

Donde π (s,a) es la probabilidad de tomar la acción a en el estado s bajo la política π.

Se pueden hacer aproximaciones sucesivas, evaluando $V_{k+1}(S)$ en términos de $V_k(S)$.

$$V_{k+1}(s) = \sum_a \pi(s,a) \sum_{s'} P_{ss'}^{a} [R_{ss'}^{a} + \gamma V_k(s')] \quad (42)$$

Se puede entonces definir un algoritmo de evaluación ***iterativo de políticas***, el cual se puede ver en la siguiente figura (Morales 2011).

Figura 15. Algoritmo iterativo de evaluación de política.

Entrada π, la política que va a ser evaluada
Inicializa $V(s) = 0$ para toda $s \in S$
Repite
  $\Delta \leftarrow 0$
  Para cada $s \in S$:
    $v \leftarrow V(s)$
    $V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V(s')]$
    $\Delta \leftarrow max(\Delta, |v - V(s)|)$
Hasta que $\Delta < \theta$ (un número positivo pequeño)
Salida $V \approx V^\pi$

Fuente: (Sutton and Barto 1998)

Una de las razones para calcular la función de valor de una política, es tratar de encontrar mejores políticas. Dada una función de valor para una política dada, se puede probar una acción $a \neq \pi(s)$ y ver si su $V(s)$ es mejor o peor que el $V^\pi(s)$.

En lugar de hacer un cambio en un estado y ver el resultado, se pueden considerar cambios en todos los estados, considerando todas las acciones de cada estado. Entonces se puede calcular una nueva política $\pi'(s) = argmax_a Q^\pi(s,a)$ y continuar hasta que no se mejore. Esto sugiere, partir de una política (π_0) y calcular la función de valor (V^π), con la cual se encuentra una mejor política (π_1) y así sucesivamente hasta converger a π^* y V^*. A este procedimiento se llama ***iteración de políticas*** (Morales 2011).

Figura 16. Algoritmo iteración de política.

1. Inicialización:
   $V(s) \in \Re$ y $\pi(s) \in A(s)$ arbitrariamente $s \in S$

2. Evaluación de la política
   Repite
   $\Delta \leftarrow 0$
   Para cada $s \in S$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V(s')]$
   $\Delta \leftarrow max(\Delta, |v - V(s)|)$
   Hasta que $\Delta < \theta$ (un número positivo pequeño)

3. Mejora de política:
   política-estable ← verdadero
   Para cada $s \in S$
   $b \leftarrow \pi(s)$
   $\pi(s) \leftarrow argmax_a \sum_{s'} [P^a_{ss'} + \gamma V(s')]$
   Si $b \neq \pi(s)$, entonces *política-estable* ← falso
   Si *política-estable*, entonces se detiene; de otro modo ir a 2

Fuente: Sutton and Barto, 1998

Uno de los problemas de iteración de políticas, es que cada iteración involucra una evaluación de políticas que requiere recorrer todos los estados varias veces.

Sin embargo, el paso de evaluación de política se puede truncar de varias formas, sin perder la garantía de convergencia. Una de ellas es pararla después de recorrer una sola vez todos los estados. A esta forma se le llama ***iteración de valor***. En particular se puede escribir combinando la mejor en la política y la evaluación de la política truncada como sigue:

$$V_{k+1}(s) = max_a \sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V_k(s')] \quad (43)$$

Se puede ver cómo expresar la ecuación de Bellman en una regla de actualización. Es muy parecido a la regla de evaluación de políticas, solo que se evalúa el máximo sobre todas las acciones.

Figura 17. Algoritmo de iteración de valor.

Inicializa V arbitrariamente, por ejemplo $V(s) = 0$ para toda $s \in S$
Repite
$\Delta \leftarrow 0$
Para cada $s \in S$:
$v \leftarrow V(s)$
$V(s) \leftarrow max_a \sum_{s'} P^a_{ss'} [R^a_{ss'} + \gamma V(s')]$
$\Delta \leftarrow max\ (\Delta, |v - V(s)|)$
Hasta que $\Delta < \theta$ (un número positivo pequeño)
Salida a una política determinística, π, de tal manera que
$\pi(s) = argmax_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V(s')\right]$

Fuente: (Sutton and Barto 1998)

3.7.2. Métodos de Monte Carlo (MC)

Los métodos de monte Carlo, se aseguran el conocimiento final de una secuencia o intento de resolución del problema, para realizar la asignación correcta del crédito o ganancia a cada una de las acciones que se han ido realizando. Dado que en principio el modelo del problema no es conocido, es necesario aprender la función de *valor-acción* en lugar de *valor-estado*, dado que se debe estimar explícitamente el valor de cada acción con el fin de sugerir una política. (Fernández 2002).

Estos métodos difieren de la PD, en que este tipo de métodos no requieren de un modelo completo o el conocimiento del ambiente, estos solo requieren experiencia de secuencia de estados, acciones y recompensas en línea o simulados con la interacción con el medio ambiente (Barto and Duff 1994; Sutton and Barto 1998; Taylor 2004).

El aprendizaje de experiencias en línea es notable porque no requiere a priori conocer la dinámica del medio ambiente y sin embargo alcanza un comportamiento óptimo. Por otra parte el aprendizaje por experiencias simuladas, es también muy poderoso, aunque un modelo es requerido, el modelo solo necesita generar una transición de muestras, no la distribución de probabilidad completa de todas las posibles transiciones que son requeridas por los métodos de PD. Asombrosamente en muchos casos es fácil generar experiencia de la muestra, de acuerdo a la distribución de probabilidad deseada, pero inviable para obtener la distribución de forma explícita (Sutton and Barto 1998).

Como ya se mencionó, en los métodos de MC para estimar V^π y Q^π podemos tomar la estadística ya que son algoritmos de muestras, en el cual se basan en el promedio de todos los retornos de la muestra (Sutton and Barto 1998; Taylor 2004; Morales 2011). Para asegurar que los retornos bien definidos están disponibles, se define los métodos de MC solo para las tareas episódicas, es decir, se asume que la experiencia está dividida

dentro de los episodios y que en todos los episodios eventualmente terminan sin importar que acción sea seleccionada. Es solo en la realización de un episodio que las estimaciones de valor y las políticas han cambiado. Por lo tanto esto métodos tiene un incremento episodio por episodio, pero no paso por paso (Sutton and Barto 1998).

Del método de MC se distingues dos algoritmos específicos para determinar $V^{\pi}(s)$, el primer método es el de ***Cada-Visita MC***, estima el valor de un estado como el promedio de los retornos que han seguido todas las visitas de estado. El siguiente es ***Primera-Visita MC***: Estima el valor de un estado como el promedio de los retornos que han seguido la primera visita al estado, donde una primera visita es la primera vez que un ensayo que el estado es visitado (Singh and Sutton 1996).

Figura 18. Algoritmo de primera-visita MC.

Inicializa
$\pi \leftarrow$ política a ser evaluada
$V \leftarrow$ una función de *valor-estado* arbitraria
Retornos(s) $\leftarrow$ un lista vacía, para todo los $s \in S$

Repite para siempre:
(a) Generar un episodio usando π
(b) Para cada estado s que aparece en el episodios:
$R \leftarrow$ Retorno obtenido luego de la primera visita a s
Agregar R *a los retornos* (s)
$V(s) \leftarrow$ promedio (*Retornos* (s))

Fuente: (Sutton and Barto 1998)

Para estimar pares *estado-acción* (Q^{π}) se corre el peligro de no ver todos los pares, por lo que se busca mantener la exploración. Lo que normalmente se hace es considerar solo políticas estocásticas que tienen una probabilidad diferente de cero de seleccionar todas las acciones (Morales 2011).

Con MC se puede alternara entre evaluación y mejoras, en base a cada episodios. La idea es que después de cada episodio las recompensas observadas se usan para evaluar la política y la política se mejora para todos los estados visitados en el episodio, el algoritmo se puede ver a continuación (Morales 2011).

Figura 19 Algoritmo de Monte Carlos ES con exploración inicial.

Inicializa para todo los $s \in S$, $a \in A(s)$:
$Q(s,a) \leftarrow$ arbitrario
$\pi(s) \leftarrow$ arbitraria
Retornos(s) $\leftarrow$ un lista vacía

Repite para siempre:
(a) Generar un episodio usando π con una exploración
(b) Para cada estado s, a aparecen en el episodio:
$R \leftarrow$ Retorno obtenido luego de la primera aparición de s, a
Agregar R *a las recompensas* (s,a)
$Q(s,a) \leftarrow$ promedio (*Recompensas* (s))
(c) Para cada estado s en el episodio:
$\pi(s) = argmax_a Q(s,a)$

Fuente: (Sutton and Barto 1998)

Existen dos formas para asegurar que todas las acciones pueden ser seleccionadas indefinidamente (Morales 2011):

- Los algoritmos *on-policy*: Estiman el valor de la política mientras la usan para el control. Se trata de mejorar la política que se usa para tomar decisiones.

- Los algoritmos *off-policy*: Usan la política y el control en forma separada. La estimación de la política puede ser por ejemplo *greedy* y la política de comportamiento puede ser ∈-*greedy*. O sea que la política de comportamiento está separada de la política que se quiere mejorar.

Ejemplos de políticas de selección de acciones son (Morales 2011):

- ∈-*greedy*: en donde la mayor parte del tiempo se selecciona la acción del mayor valor estimado, pero con probabilidad ∈ se selecciona una acción aleatoriamente.

- *Softmax*, en donde la probabilidad de selección de cada acción depende de su valor estimado. La más común sigue una distribución de Boltzman o Gibbs y selecciona una acción con la siguiente probabilidad:

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(a)/\tau}} \quad (44)$$

3.7.3. Diferencias temporales (Temporal Difference TD)

La habilidad de los métodos de MC de trabajar sin un modelo del ambiente y aprender directamente de las experiencias es llamativa, desafortunadamente, estos no tienen la técnica de *bootstrappinp* como en los métodos de la PD, por lo tanto siempre se debe

esperar al resultado final, antes que la experiencia pueda ser grabada y la experiencia pueda ocurrir. Lo ideal es que se deseen métodos que puedan aprender directamente de la experiencia, como los de MC, pero también que tengan el *bootstrappinp* como la PD, este tipo de algoritmos son conocidos como diferencias temporales (Sutton 1988; Sutton and Barto 1998; Taylor 2004; W. Josemans 2009).

Los algoritmos de TD logran un equilibrio entre la experiencia en bruto y la no necesidad de la naturaleza un modelo de los métodos de MC y un modelo basado en la PD. Su nombre hace referencia a la apropiada distribución de las recompensas por el agente en pasos sucesivos o incrementos en tiempo discreto. La idea es que puede tomar varios pasos y varias acciones antes de la acción tomada por los resultados del agente, ya sean recompensas o castigos (Taylor 2004).

La idea central de los métodos de TD es, la de actualizar la función de valor a media que se obtienen refuerzos de parte del entorno (Uribe 2011). Los algoritmos de TD usan el error o diferencia entre predicciones sucesivas, aprendiendo de los cambios entre estas predicciones (Sutton 1988; Morales 2011). Una de sus ventajas es que el algoritmo no requiere una gran capacidad computacional y converger de forma más rápida a mejores resultados (Morales 2011).

Para explicar el concepto detrás de las TD, consideremos un problema donde una secuencia de predicción, P_t, P_{t+1},…., se está haciendo del valor esperado una variable aleatoria r_T en un tiempo futuro T. En este tiempo, la predicción P_t para todos los $t < T$ podría mejorarse mediante cambios de (Rummery 1995),

$$\Delta P_t = \alpha(r_t - P_T) \quad (45)$$

donde α es un parámetro de tasa aprendizaje. La anterior ecuación puede ser expandida en términos de errores de diferencias temporales, entre predicciones sucesivas.

$$\Delta P_t = \alpha[(P_{t+1} - P_t) + (P_{t+2} - P_{t+1}) + \cdots + (P_{T-1} - P_{T-2}) + (r_T - P_{T-1})]$$
$$\Delta P_t \ = \alpha \sum_{k=t}^{T-1}(P_{k+1} - P_k) \quad (46)$$

Donde $P_T = r_T$. Esto significa que el paso de tiempo t, cada predicción P_k para $k \leq t$ podría actualizarse mediante el error-TD, $(P_{t+1} - P_t)$. Esta idea forma la base de los algoritmos de diferencias temporales, que permiten el error-TD para ser usado en cada intervalo de tiempo, para actualizar todas las predicciones previas y eliminar la necesidad de esperar hasta el tiempo T, después de actualizar cada predicción por la aplicación de la ecuación 45 (Rummery 1995).

El método más simple es el TD (0), actualiza la función de valor de la siguiente manera (Uribe 2011):

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (47)$$

Donde s_{t+1} es el estado al que se llega luego de ejecutar la acción que indica la política estando en el estado s_t y r_{t+1} la recompensa correspondiente. Para obtener la función de evaluación V^{π} para una política π, se puede usar el siguiente algoritmo (Uribe 2011):

Figura 20. Algoritmo TD(0).

Inicializa $V(s)$ arbitrariamente, la política π a evaluar
Repite (para cada episodio):
    Inicializa s
    Repite (para cada paso del episodio):
        $a \leftarrow$ acción dada por π para s
        Toma la acción: observa la recompensa, r, y el siguiente estado s'
        $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$
        $s \leftarrow s'$
    hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

La actualización de los valores tomando en cuenta la acción sería:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_{t+1}, a_{t+1})] \quad (48)$$

Esta actualización es hecha después de cada transición de un estado terminal s_t. Si s_{t+1} es terminal, entonces $Q(s_{t+1}, a_{t+1})$ es definido como cero. Esta regla usa cada elemento de eventos quíntuples, $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$, que hace una transición de una acción estado par al siguiente. Este quíntuple da lugar al nombre del algoritmo *Sarsa* (Sutton and Barto 1998).

Figura 21. Algoritmo Sarsa.

Inicializa $Q(s,a)$ arbitrariamente
Repite (para cada episodio):
    Inicializa s
    Selecciona a de s usando la política dada por Q (e.g. $\in -greedy$)
    Repite (para cada episodio):
        Toma una acción a, observar r, s'
        Toma la acción: observa la recompensa, r, y el siguiente estado s'
        Elige a' de s' usando la política derivada de Q (e.g. $\in -greedy$)
        $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s', a') - Q(s,a)]$
        $s \leftarrow s'$; $a \leftarrow a'$
    hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

Uno de los desarrollos más importates en el TD fue el desarrollo de un algoritmo "fuera-de-política" (*off-policy*) conocido como *Q-learning*. La idea principal es realizar la actualización de la función de valores estado-acción, de la siguiente forma (Morales 2011):

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (49)$$

Y puede ser reescrita como:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha[r_{t+1} + \gamma max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (50)$$

Figura 22. Algoritmo Q-learning.

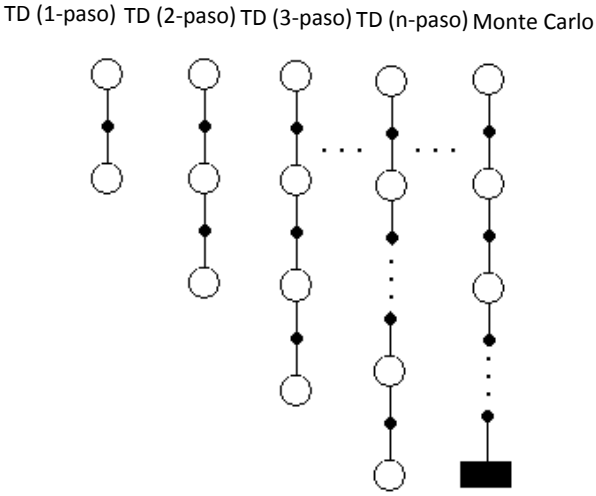
Inicializa $Q(s,a)$ arbitrariamente
Repite (para cada episodio):
  Inicializa s
  Repite (para cada episodio):
    Elige a de s usando la política derivada de Q (e.g. $\in -greedy$)
    Toma una acción a, observar r, s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_a, Q(s',a') - Q(s,a)]$
    $s \leftarrow s'$; $a \leftarrow a'$
  hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

3.7.4. Trazas de elegibilidad

Entre los métodos de TD, en los que la actualización de las funciones de valor se realiza en cada ejecución de la acción y los métodos de MC en los que esta actualización no se realiza hasta que no se termina una secuencia, existe un amplio rango de posibilidades tal y como se muestra en la siguiente figura (Sutton and Barto 1998; Fernández 2002).

Figura 23. Espectro de posibilidades desde los TD tradicionales hasta Monte Carlo.



Fuente: (Sutton and Barto 1998)

En la figura se muestra como los métodos de TD tradicionales son métodos de un único paso, en el sentido de que cada vez que se realiza una acción y se lleva a un estado nuevo, se realiza una actualización de las funciones de valor. En el lado opuesto, se encuentran los métodos de MC, donde las actualizaciones se realizan al final de un intento completo de solucionar el problema. En medio se encuentran los métodos TD de *n* pasos, en los que las actualizaciones se realizan en función de los resultados de ejecutar *n* acciones. Es decir, el refuerzo con el que se actualiza la función de valor estado en un instante de *t*, siguiendo un método TD de *n* pasos es tal y como se define a continuación (Peng and Williams 1996; Fernández 2002; Morales 2011).

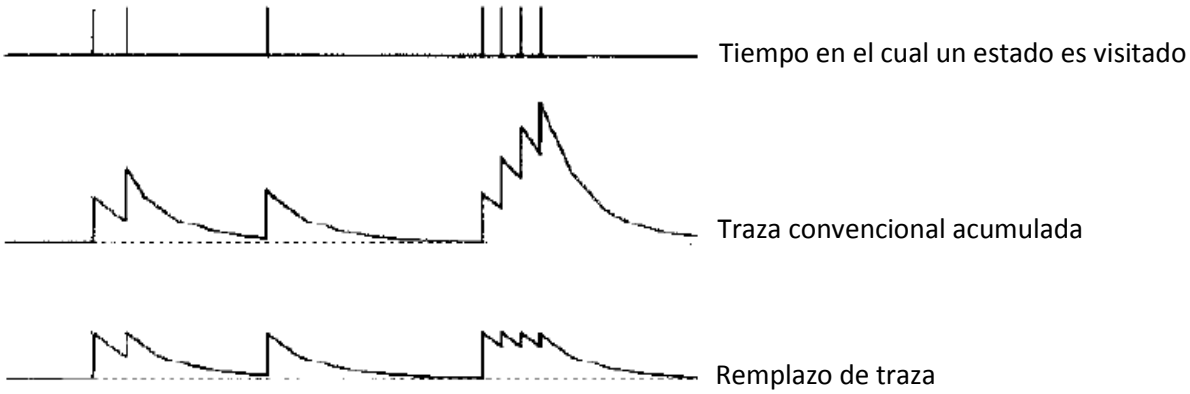
$$R_t^{(n)} = r_{t+1}\gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^n r_{t+n} + \gamma^n V_t(s_{t+n}) \quad (51)$$

Cuando n es igual a 1, la función de actualización se convierte en la función típica de los métodos TD, mientras si n es igual a la longitud de la secuencia completa, se convierte en la función de actualización de los métodos de MC.

Los métodos TD(λ) utilizan estas ideas, pero en vez de sumar directamente los refuerzos obtenidos en el futuro, los promedian con el factor λ^{n-1}, donde $0 \leq \lambda \leq 1$. El factor de normalización (1-λ) hace que el sumatorio completo sea 1 (Sutton and Barto 1998; Fernández 2002; Martínez and de Prada 2003).

Por lo tanto la forma más siencilla de implementar esta aproximación se basa en las trazas de elegibilidad. Las trazas de elegibilidad fueron introducidas por Klopf en 1972, la idea detrás de las trazas de elegibilidad es muy simple, en cada tiempo un estado es visitado al inicio en un proceso de memoria de término corto, una traza, que luego decae gradualmente sobre el tiempo. Esta traza marca el estado como elegible para el aprendizaje. Si un inesperado evento ya sea bueno o malo, ocurre mientras la traza no es cero, entonces el estado es asignado de acuerdo al crédito. En una traza acumulativa convencional, la traza se acumula cada vez que se entró en el estado. En una traza remplazada, cada tiempo el estado es visitado, la traza es reajustada hasta 1 sin tener en cuenta de la presencia de una traza anterior. La nueva traza remplaza la vieja, como se puede ver en la siguiente figura (Singh and Sutton 1996).

Figura 24. Acumulación y remplazo de trazas de elegibilidad.



Fuente: (Singh and Sutton 1996)

En la práctica, más que esperar *n* pasos para actualizar (*forward view*), se realiza al revés (*backward view*). Se guarda información sobre los estados por los que pasó y se actualizan hacia atrás las recompensas. Se puede probar que ambos enfoques son equivalentes (Morales 2011).

Para implementar la idea anterior, se asocia a cada estado o par estado acción una variable extra, representando su traza de elegibilidad que se denota por $e_t(s)$ o $e_t(s,a)$. Este valor va decayendo con la longitud de la traza creada en cada episodio. Para TD(λ):

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & \text{si } s \neq s_t \\ \gamma\lambda e_{t-1}(s) + 1 & \text{si } s \neq s_t \end{cases} \quad (52)$$

Para Sarsa se tiene lo siguiente:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s,a) & \text{si } s \neq s_t \\ \gamma\lambda e_{t-1}(s,a) + 1 & \text{si } s \neq s_t \end{cases} \quad (53)$$

Figura 25. TD (λ).

Inicializa $V(s)$ arbitrariamente y $e(s) = 0$ para todos los $s \in S$
Repite (para cada episodio):
  Inicializa s
  Repite (para cada episodio):
    $A \leftarrow$ acción dada por π para s
    Toma una acción a, observando la recompensa r y el siguiente estado s'
    $\delta \leftarrow r + \gamma V(s') - V(s)$
    $e(s) \leftarrow e(s) + 1$
    Para todos los s:
      $V(s) \leftarrow V(s) + \alpha\delta e(s)$
      $e(s) \leftarrow \gamma\lambda e(s)$
    $s \leftarrow s'$
  hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

Sarsa (λ) usa la experiencia de aprender las estimaciones óptimas Q-valores de las funciones pares que son mapeadas de s, a, al retorno óptimo sobre la acción tomada a en el estado s. La transición al paso de tiempo t, < s_t, a_t, r_t, s_{t+1}>, es usado para cargar los Q-valores estimados de todos los pares de acciones-estados en proporción de su elegibilidad (Loch and Singh 1998). El algoritmo para Sarsa (λ) se describe a continuación:

Figura 26. Sarsa (λ).

Inicializa $Q(s,a)$ arbitrariamente y $e(s,a) = 0$ para todos los s,a
Repite (para cada episodio):
  Inicializa s,a
  Repite (para cada episodio):
    Toma una acción a, observar r, s'
    Elige a' de s' usando la política derivada de Q (e.g. $\in -greedy$)
    $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    $e(s, a) \leftarrow e(s, a) + 1$
    Para todos los s,a:
      $Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$
    $s \leftarrow s'$; $a \leftarrow a'$
  hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

El algoritmo Q(λ) combina el retorno de TD(λ) (ecuación 51) para un ʎ general con la forma incremental del Q-learning (paso a paso) (Peng and Williams 1996; Glorennec 2000). Para este algoritmo tenemos que Las trazas de elegibilidad están definidas sobre el espacio por el producto de S x A. Las definiciones son por lo tanto poco modificadas, por ejemplo, para la traza de elegibilidad acumulada, una tiene (Glorennec 2000):

$$e_t(s,a) = \begin{cases} 1 + \gamma\lambda e_{t-1}(s,a) & \text{si } s = s_t \, y \, a = a_t \\ \gamma\lambda e_{t-1}(s,a) & \text{de otra manera} \end{cases} \quad (54)$$

y para la sustitución de la elegibilidad:

$$e_t(s,a) = \begin{cases} 1 & \text{si } s = s_t \, y \, a = a_t \\ \gamma\lambda e_{t-1}(s,a) \; de \; otra \; manera \end{cases} \quad (55)$$

La convergencia de Q(λ) hacia Q^* solo se asegura si λ>0. El algoritmo se muestra a continuación:

Figura 27. Q (λ).

Inicializa $Q(s,a)$ arbitrariamente y $e(s,a) = 0$ para todos los s,a
Repite (para cada episodio):
  Inicializa s,a
  Repite (para cada episodio):
    Toma una acción a, observar r, s'
    Elige a' de s' usando la política derivada de Q (e.g. $\in -greedy$)
    $a^* \leftarrow$ argmax$_b$ $Q(s',b)$ (si los a' lazos para max, entonces $a^* \leftarrow a'$)
    $\delta \leftarrow r + \gamma Q(s', a^*) - Q(s,a)$
    $e(s,a) \leftarrow e(s,a) + 1$
    Para todos los s,a:
      $Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$
      Si $a' = a^*$, entonces $e(s,a) \leftarrow \gamma\lambda e(s,a)$
        Si no $e(s,a) \leftarrow 0$
    $s \leftarrow s'$; $a \leftarrow a'$
  hasta que s sea terminal

Fuente: (Sutton and Barto 1998)

4. MATERIALES Y MÉTODOS

La presente tesis tiene como principal objetivo el control de un sistema de aireación en una planta de lodos activados, cabe mencionar, que este control se llevará a cabo solo mediante simulaciones y no se implementará sobre una planta de tratamiento de lodos activados, para esto se propuso la siguiente metodología compuesta por una serie de pasos, para el cumplimiento del objetivo.

Primero que todo se planteó un tipo de investigación documental, en el cual se pretende identificar los diferentes modelos matemáticos existentes del aprendizaje por refuerzo, vistos en el capitulo anterior. Para la selección se utilizó una matriz, donde en las columnas se introdujeron los diferentes modelos y en las filas se tuvieron criterios para la elección, como: construcción, continuidad temporal, velocidad de convergencia, etc., esto con el fin de encontrar las ventajas y desventajas de cada uno de los modelos, de acuerdo a la necesidad del problema.

Por consiguiente a esto, se debían obtener valores reales de la dinámica de la demanda química de oxígeno (DQO), el cual es un parámetro que mide la cantidad de sustancias susceptibles de ser oxidadas por medios químicos que hay disueltas o en suspensión en una muestra líquida (Sawyer, McCarty et al. 2001). La DQO dentro del modelo se convierte en una variable independiente, ya que es el sustrato (alimento) de las bacterias, el cual más adelante veremos cómo matemáticamente se conecta con la necesidad y concentración de oxígeno en el reactor aeróbico.

Luego a esto, se construyó un modelo computacional de un sistema de lodos activados mediante el software Matlab, teniendo como referencia las ecuaciones del sistema de lodos activados descritas por Sergio Alejandro Martínez presentes en el libro "*Tratamiento de aguas residuales con Matlab*", esto con el fin de simular el comportamiento de la planta piloto bajo las diferentes concentraciones de DQO encontradas anteriormente.

Adicionalmente, este modelo permitió simular el ambiente en el cual el agente interactuaría, ya que dentro de este se ejecutan las diferentes acciones necesarias, para encontrar el caudal de oxígeno necesario que permita llegar a una concentración final 100 mg/l de DQO (la cual es la política ideal y la que da una mayor recompensa). Es importante mencionar que todos los códigos elaborados y utilizados en esta tesis, se encuentran en los anexos.

A continuación se explicará con mayor detenimiento la metodología y se presentaran los resultados obtenidos.

4.1. Selección de algoritmo de control

La investigación documental se fundamentó en varios factores, el primero consistió en identificar la mayor cantidad de algoritmos para la resolución de problemas que utilicen el aprendizaje por refuerzo, ya fuesen los descritos en el titulo 3.7., algoritmos más recientes o modificaciones a los existentes, como se muestra a continuación:

Tabla 1. Algoritmos del método de la programación dinámica

PROGRAMACIÓN DINÁMICA		
Iteración de evaluación de políticas	Iteración de valores	Iteración de políticas

Fuente: (El autor 2012)

Tabla 2 Algoritmos del método de Monte Carlo.

MONTE CARLO		
Cada visita Monte Carlo	Primera visita Monte Carlos	Algoritmo de control Monte Carlo asumiendo una exploración inicial
Con política Control de Monte Carlo	Sin política Control de Monte Carlo	

Fuente: (El autor 2012)

Tabla 3. Algoritmos del método de diferencias temporales

DIFERENCIAS TEMPORALES		
TD(0)	SARSA	Q-learning
R-learning	GTD (Gradient-Descent Methods for Temporal-Difference)	GTD2 (Gradient-Descent Methods for Temporal-Difference Version 2)
LSTD (Least-squares temporal difference)	RLS-TD (Recursive Least-squares temporal difference)	Dyna
RLSTD(λ)	LSTD(λ)	TD(λ)
SARSA(λ)	Q(λ)	Dyna-Q
λ-LSPE		

Fuente: (El autor 2012)

El segundo aspecto, hacía referencia en encontrar los criterios de selección del algoritmo; la búsqueda se fundamentó en las necesidades del problema propuesto en esta tesis, por lo tanto se plantearon los siguientes criterios:

- Aplicación en control: Se deseaba conocer la aplicabilidad en problemas de control, si esta era alta o baja, en qué campos de la ingeniería se había utilizado y frecuencia de aplicación.
- Resultados de la aplicación: Se quería conocer y analizar los resultados obtenidos en los diferentes campos (control, robótica, reconocimiento de patrones, calibración de modelos, minería de datos, etc.).

- Construcción: Se miraba que complejidad tenia la construcción del modelo, cuantas variables necesitaba, si necesitaba exploración de todas las acciones, si seguía acciones estocásticas o necesitaba determinarlas en un orden.
- Continuidad temporal: Si el modelo se podía aplicar a tiempo discreto o tiempo continuo.
- Velocidad de convergencia: Se planteo este ítem para conocer qué tiempo podría tomar ejecutar una acción, recibir la recompensa.
- Velocidad de aprendizaje: Se observaba que tiempo que se demoraba en aprender del ambiente, si este aprendizaje era más minucioso o pobre.

Por último se construyó una tabla donde se compararon los modelos por método de resolución (PD, MC y TD), esto con el fin de obtener los algoritmos de manera independiente, lo anterior se llevo a cabo mediante una evaluación numérica mediante tres valores 0, 0.5 y 1, el valor se otorgó de acuerdo a la investigación documental (las tablas de comparación se anexan a la presente tesis).

Figura 28. Imagen de la tabla de comparación de algoritmos por el método de Monte Carlo

ALGORITMOS / CRITERIOS DE SELECCIÓN		*Cada-visita Monte-Carlo*	*Primera-visita Monte-Carlo*	*Algoritmo de control Monte Carlo asumiendo una exploración inicial*	*Con-Política Control de Monte Carlo*	*Sin-Política Control de Monte Carlo*
Aplicación en control	Alta		1		1	1
	Baja					
Resultados de la aplicación	Alta	1	1		1	0
	Baja					
Construcción	Simple	1	1	1	0.5	0.5
	Complejo					
Continuidad temporal	Discreto	1	1	1	1	1
	Continuo					
Velocidad de convergencia	Alta	0.5	1	0.5	1	1
	Baja					
Velocidad de aprendizaje	Alta					
	Baja					
Total		3.5	5	2.5	4.5	3.5

Fuente: (El autor 2012)

Como se puede observar en la figura 28, existen algoritmos en los cuales no se pudieron establecer valores, ya que no se encontraron referencias sobre estos ítems, sin embargo, la sumatoria tomaba esos espacios como ceros. Para cada método de resolución los algoritmos seleccionados fueron: Programación dinámica Iteración de Políticas, Primera Visita Monte Carlo y Diferencias Temporales Q-learning.

Luego de haber seleccionado a estos tres algoritmos, a continuación se realizará una comparación de ventajas y desventajas encontradas entre los algoritmos, para seleccionar un algoritmo que finalmente será el controlador.

Tabla 4. Comparación de algoritmos seleccionados mediante ventajas y desventajas.

Iteración de Políticas		Primera Visita Monte Carlo		Q-learning	
Ventajas	Desventajas	Ventajas	Desventajas	Ventajas	Desventajas
El algoritmo puede ser implementado para aprendizaje en línea (Zhang, Xu et al. 2008).	En el método de iteración de políticas se utilizan simulaciones y métodos heurísticos para determinar políticas, fallan si esas políticas están muy lejos de la política óptima (Bertsekas and Ioffe 1996).	Los algoritmos de aprendizaje por refuerzo (Primera Visita Monte Carlo) pueden ser herramientas útiles en el diseño de controladores (Bernstein, Zilberstein et al. 2001).	No es muy eficiente en grandes rangos de búsqueda (Dolk 2010) .	Con respecto a la programación dinámica se tienen dos ventajas, la primera se tiene un enfoque en línea, tiene la capacidad de aprendizaje en ambientes dinámicos, el segundo es que se pueden emplear técnicas de funciones de aproximación (Xhafa, Abraham et al. 2010) .	En casos donde los estados son muy grandes, el algoritmo puede tener problemas ya que el cálculo de los valores Q para todos los estados puede ser en muchos casos innecesario (Bertsekas and Tsitsiklis 1996; Bhatnagar and Babu 2008)
El algoritmo de la iteración de políticas converge más rápido en comparación con Q-learning (Shah and Gopal 2011).	El algoritmo requiere de un modelo de la realidad de un ambiente de un Proceso de Markov. Si bien estos algoritmos dependen fuertemente del modelo y se sabe que, a veces, determinar este modelo no es posible (Maximiliano and Laura 2004).	No necesita aprender de todo acerca de todo en los estados, por lo tanto las acciones pueden ser tomadas en menor tiempo (Sutton and Barto 1998).	El valor del estado está determinado por el promedio de las recompensas de ese estado (Aihe 2008).	Es un algoritmo muy simple y también converge con una política óptima (Wicaksono 2011).	
				En la aplicación del algoritmo en una planta piloto de pH de laboratorio, presentó un buen rendimiento en una amplia gama de valores de pH, y para diferentes procesos. (Syafiie, Tadeo et al. 2007; Syafiie, Tadeo et	

Iteración de Políticas		Primera Visita Monte Carlo		Q-learning	
Ventajas	Desventajas	Ventajas	Desventajas	Ventajas	Desventajas
				al. 2007; Syafiie, Vilas et al. 2008; Syafiie, Tadeo et al. 2011).	
				Excelentes desempeños comparados con otros algoritmos del método de las diferencias temporales (Mahadevan 1994).	
				El Q-learning es un modelo completamente libre, esto es, que no requiere el conocimiento de las probabilidades de transición de las cadenas de Markov, solo la capacidad de simular estas transiciones o de ejecutarlas sobre un sistema físico (Bhatnagar and Babu 2008; Penn-University 2011). Situación que si requiere los métodos de la PD.	

Fuente: (El autor 2012)

De acuerdo a la comparación hecha en la tabla anterior, el algoritmo que presentó mayor ventaja frente a los diferentes algoritmos fue el Q-learning, ya que sus principales fortalezas son: su velocidad de convergencia, ya que este parámetro influye de manera vital en la solución del problema de control, debido a que la acción que el agente debe tomar se debe realizar en un delta de tiempo corto. Segundo, la capacidad de exploración aleatoria de las posibles acciones que se pueden tomar durante el estado x, lo que representa un buen comportamiento del agente. Por último, a pesar de ser un algoritmo sencillo en su aplicación y construcción, presenta muy buenos resultados en sus diferentes aplicaciones.

4.1.1. Q-learning

Q-learning, uno de los algoritmos de aprendizaje por refuerzo más importantes y fue presentado por Watkins en 1989 (Guo, Liu et al. 2004), la letra "Q" no representa nada especial, solo hace referencia a la notación usada en la propuesta original de este método (Bertsekas 1995).

Este algoritmo es una combinación de la programación dinámica, más concretamente el algoritmo de Iteración de Valores y la aproximación estocástica (Azar, Munos 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 óptimamente en un dominio Markoviano por la experiencia dada por la consecuencia de acciones, sin requerir de construir mapas del dominio (Watkins 1989; Watkins and Dayan 1992).

El Q-learning también es denominado un método *off-policy*, lo que significa que la función de valor converge casi seguramente a la función de valor óptima del estado-acción, independientemente de la política que están siendo seguidas o los Q valores iniciales (Monostori and Csáji 2006).

Para entender claramente se puede ver de la siguiente manera, el agente observa el presenta estado s_t y ejecuta una acción a_t según la evaluación del retorno que se hace en esa etapa. Se actualizará la evaluación del valor de la acción mientras tome en cuenta, a) el refuerzo inmediato r_t y b) el valor estimado del nuevo estado $V_t(s_{t+1})$, que es definido por (Peng and Williams 1996; Glorennec 2000):

$$V_t(x_{t+1}) = max_{b \in A_{t+1}} Q(s_{t+1}, b) \quad (56)$$

La actualización corresponde a la ecuación:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \beta\{r_t + \gamma V_t(s_{t+1}) - Q(s_t, a_t)\} \quad (57)$$

Donde β es la tasa de aprendizaje tanto que $\beta \to 0$ como $t \to \infty$, esta ecuación puede ser escrita:

$$Q(s_t, a_t) \leftarrow (1-\beta)Q(s_t, a_t) + \beta\{r_t + \gamma V_t(s_{t+1})\} \quad (58)$$

El algoritmo presenta una serie de características:

- El método Q-learning no especifica que acciones debe tomar el agente en cada estado ya que actualiza sus estimaciones, de hecho el agente puede tomar cualquier acción que le plazca. Esto significa que el Q-learning permite experimentación arbitraria, mientras que al mismo tiempo preserva la mejor estimación actual de los valores de estado (Peng and Williams 1996).

- No se utiliza una política explicita, las acciones se van tomando de acuerdo a lo que censa del ambiente y se determina el estado en el que se encuentra el agente, por consiguiente se puede tomar una acción de acuerdo a dicho estado (Maximiliano and Laura 2004).

- Una vez el algoritmo encuentra una acción que proporciona una mayor cantidad de refuerzos, esa es la que se usa siempre.

- La función de refuerzo es conocida, si bien el algoritmo no necesita datos del ambiente trabaja sobre una función de refuerzos dada.

4.1.2. Descripción del agente diseñado

Inicialmente el agente evalúa el la señal de estado inicial, el cual es el valor de cada concentración de sustrato que llega al sistema de lodos activados, esto se llevo a cabo para poder evaluar cada concentración de sustrato y poder tomar la acción de control necesaria.

Posteriormente, como las acciones de control no se pueden tomar en deltas de tiempo muy pequeños (ya estas se ven reflejadas en un futuro) y si se llegasen a tomar así, las acciones de control no se verían reflejadas, por lo tanto se configuró al controlador para que de acuerdo a sus acciones, estas se encuentren dentro el tiempo de solución de las ecuaciones del ambiente (sistema de lodos activados).

Lo más importante para el proyecto, son las acciones que el agente puede tomar ya que estas en sí son el control, por lo tanto se creó un universo de diferentes acciones capaces de ser aplicadas dentro del sistema, estas acciones originalmente son dos, la primera es el caudal de aire inyectado al reactor aeróbico, la segunda es el caudal de recirculación de lodos (lo cual veremos más adelante).

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 el podría tomar como decisión no inyectar aire, situación que tornaría al reactor en anaeróbico, lo cual no es deseado, o inyectar más de los posible debido a que el caudal de aire es función del equipo de suministro empleado y del número de estos, lo que reduce la serie de caudales. Por otra parte la capacidad de recirculación de lodos generalmente se encuentra en porcentajes que van desde 50 hasta 120% (Water Environment Federation 2009).

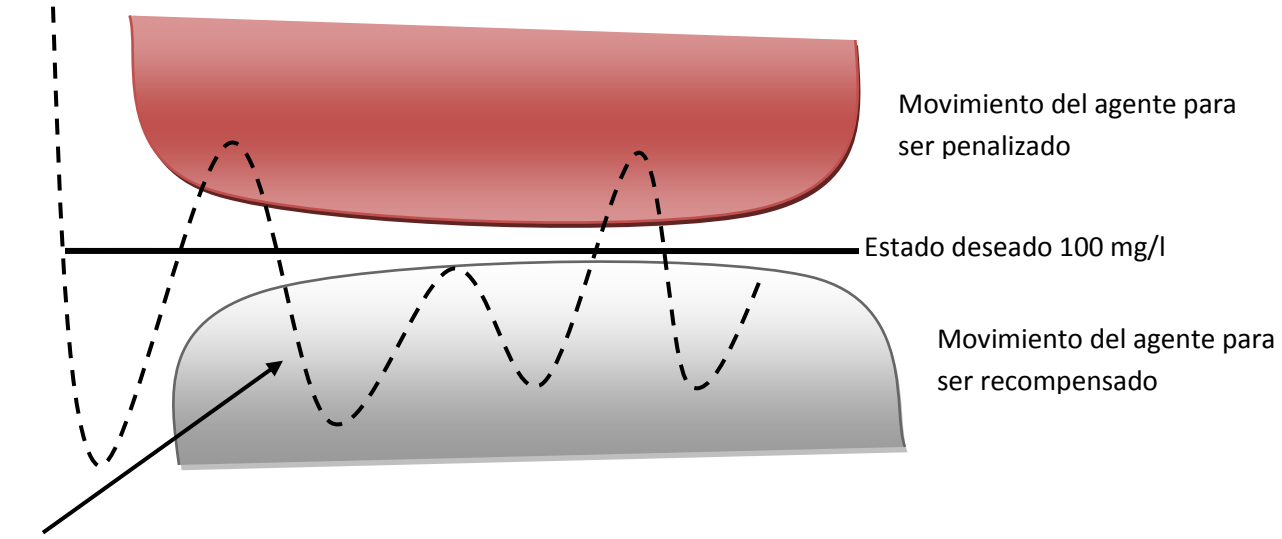
Después de delimitar los caudales, el agente recorre las acciones conjuntas con mayor probabilidad de éxito, las cuales ingresan al ambiente, de donde recibe una recompensa de diferentes magnitudes. Estas magnitudes son función de que tan alejado se encuentra el agente de los objetivos deseados.

El cálculo de la recompensa se realiza empleando una distancia euclidiana, la cual es la distancia ordinaria entre dos puntos de un espacio euclídeo, se deduce a partir del teorema de Pitágoras. Por ejemplo, la distancia entre dos puntos P_1 y P_2 de coordenadas (x_1 , y_1) y (x_1 , y_1) respectivamente es:

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (59)$$

En donde el objetivo se resta a la respuesta del sistema y así encontrar el mayor valor de recompensa de sus acciones, por ejemplo si el objetivo es mantenerme en un estado donde el sustrato de salida menor de 100 mg/l, el agente se movía de la siguiente manera:

Figura 29 Definición de estado y control objetivo.



Fuente (Autor 2012)

4.2. Estimación de la Demanda Química de Oxígeno (DQO)

La demanda de oxígeno es un parámetro muy importante para determinar la cantidad de materia orgánica contaminante en el agua. La prueba tiene su mayor aplicación en la medición de descargas hechas por redes de alcantarillados, viviendas, industrias, plantas de tratamiento, para la evaluación de la eficiencia de procesos de tratamiento y con el fin de valorar la calidad en fuentes hídricas (lagos, ríos). La prueba de la demanda de oxígeno no determina la concentración específica de una sustancia, si no que mide el efecto de una combinación de sustancias y condiciones (Boyles 1997).

Existen tres métodos ampliamente utilizados para la medición de la demanda de oxígeno, dos mediciones son directas: la Demanda Bioquímica de Oxígeno (DBO), la Demanda Química de Oxígeno (DQO) y un tercer método es el Carbono Orgánico Total (COT), el cual mide la demanda de oxígeno indirectamente (Boyles 1997).

La prueba de la DQO usa un oxidante químico muy fuerte en una solución ácida y el calor para oxidar el carbono orgánico en CO_2 y H_2O. De acuerdo al *Standard Methods for the Examination of Water and Wastewater,* la DQO es la demanda química de oxígeno, es la medida del equivalente en oxígeno del contenido de materia orgánica de una muestra que es susceptible de oxidación por un oxidante químico fuerte.

La mayoría de tipos de materia orgánica se oxidan mediante una mezcla en ebullición de ácido crómico y sulfúrico. Se somete a reflujo una mezcla en una solución ácida fuerte, con un exceso conocido de dicromato de potasio ($K_2Cr_2O_7$). Después de la digestión, el $K_2Cr_2O_7$ restante no reducido se titula con sulfato ferroso amónico para determinar la cantidad de $K_2Cr_2O_7$ consumida; la materia orgánica oxidable se calcula en términos del equivalente de oxígeno. Cuando se usan volúmenes de muestras diferentes de 50 ml, se mantienen constantes las relaciones de peso de los reactivos, los volúmenes y las concentraciones. El tiempo de reflujo estándar de 2 horas se puede reducir si se ha demostrado que un período menor conduce a los mismos resultados. Algunas muestras con la DQO muy baja o con el contenido de sólidos altamente heterogéneos pueden necesitar analizarse por duplicado para producir la mayoría de datos confiables. Los resultados además son mejorados mediante la reacción con una cantidad máxima de dicromato, tal que permanece algo de dicromato residual (INCONTEC 2002).

Las mediciones de la DQO son necesarias para los balances de masas en las plantas de tratamiento de aguas residuales, el contenido de la DQO puede ser subdividido en fracciones útiles, para su consideración en relación al diseño de procesos de tratamiento de aguas residuales (Henze 2008). Por lo tanto la DQO se convierte en un parámetro de gran importancia para la modelación de sistemas de lodos activados, debido a que este parámetro incluye materia orgánica biodegradable y no biodegradable, por ejemplo la concentración de DQO no biodegradable particulada afecta fuertemente la acumulación de lodos en el reactor y la producción de lodos diaria, por otra parte la concentración de

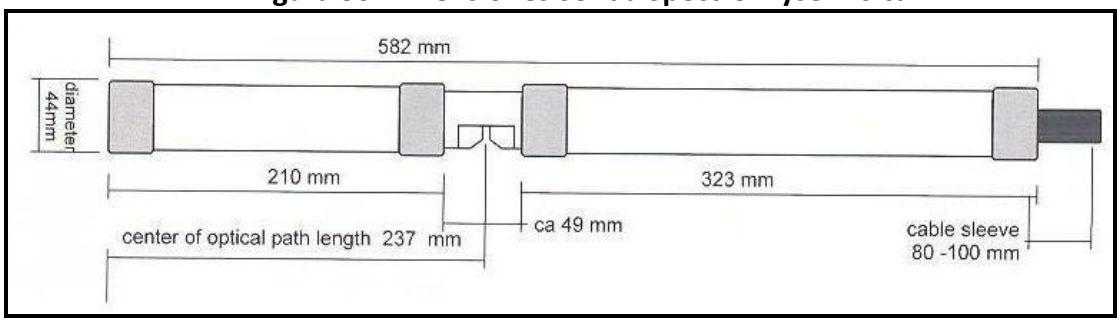
DQO no biodegradable soluble, fija la concentración de DQO filtrada en el efluente del sistema (Henze 2008).

Por tal motivo, esta tesis tomó datos de DQO extraídos del proyecto de un convenio realizado entre la Pontificia Universidad Javeriana y la Empresa de Acueducto y Alcantarillado de Bogotá en el año 2011, en el cual por medio de un Spectro::lyser, ya que una de las principales ventajas de este tipo de captores es que un número importante de parámetros (SST, DQO y nitratos) pueden ser monitoreados en continuo utilizando un sólo instrumento de medición (Gruber, Bertrand-Krajewski et al. 2006).

4.2.1. El Spectro::lyser y sus valores de DQO

La sonda Spectro::Lyser - s:can, 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, de igual forma puede ser utilizada fuera del medio líquido gracias a los accesorios y dispositivos que son acoplables a ella. Las dimensiones de la sonda se observan en la siguiente figura.

Figura 30 Dimensiones Sonda Spectro::Lyser - s:can.



Fuente: (S::CAN 2007)

El Spectro::lyser, es capaz de proporcionar informaciones del orden de una medición por minuto, que puedan traducirse en términos de concentraciones equivalentes (para nuestro caso DQO). El espectrómetro es un captor sumergible de 60 cm de longitud y 44 mm de diámetro que mide la atenuación de la luz entre 200 nm y 750 nm, muestra y/o comunica los resultados en tiempo real, el equipo es de 256 píxeles con una lámpara de flash de Xenón como fuente de luz, adicionalmente cuenta con un sistema autolimpiador usando aire presurizado otorga (Langergraber, Fleischmann et al. 2003; Langergraber, Fleischmann et al. 2004; Langergraber, Gupta et al. 2004; Hochedlinger, Hofbauer et al. 2006).

Su uso es principalmente para detectar las concentraciones de sustancias contaminantes e indicadores de calidad de agua, su aplicación va desde agua pura (potabilización) hasta aguas con vertimientos industriales (aguas residuales) (S::CAN 2007). Sin embargo, por otra parte este tipo de captores han sido probados en varias condiciones de

funcionamiento, incluyendo ríos (Staubmann, Fleischmann et al. 2001), plantas de tratamiento (Fleischmann, Langergraber et al. 2001; Winkler, Saracevic et al. 2008), aliviaderos (Gruber, Winkler et al. 2004) y sistemas de alcantarillado (Torres and Bertrand-Krajewski 2008).

Figura 31 Vista de la instalación del equipo de medida.

Fuente: (Lorenz, Fleischmann et al. 2002; Gruber, Bertrand-Krajewski et al. 2006)

La sonda trabaja de acuerdo con el principio de la espectrometría UV-VIS, emite un rayo de luz que se mueve a través del medio en estudio y mide su intensidad, es decir, mide si éste se debilita debido a las sustancias y/o partículas contenidas con un detector de medición de longitudes de onda. Cada molécula de una sustancia disuelta absorbe la radiación en ciertas longitudes de onda, por lo cual se puede determinar la concentración de las sustancias según la absorción de la muestras, a más alta concentración de agentes contaminantes más es debilitado el haz de luz (S::CAN 2007).

Su diseño consiste en tres componentes principales: (1) el emisor, (2) la celda de medición y (3) la unidad de detección (en la figura 30). El elemento central del emisor es una fuente de luz de flash de xenón, la cual se complementa con un sistema óptimo que guía el haz de luz y un sistema de control electrónico para la operación de la lámpara (S::CAN 2007)

Figura 32 Partes de la Sonda.

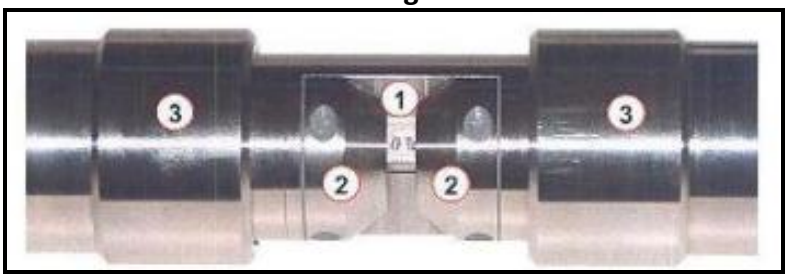
1 - Emisor
2 - Sección de Medición
3 - Detector
4 - Collarín
5 - Cable
6 – Conector

Fuente: (S::CAN 2007)

En la sección de medición la luz pasa a través de una ventana de medida que se llena con la sustancia en análisis que interactúa con él, un segundo haz de luz dentro de la sonda (haz de compensación) es guiado a la sección de comparación interna, lo cual permite

identificar alteraciones durante el proceso de medición que son compensadas de forma automática (S::CAN 2007).

Figura 33 Sección de Medición.



1 – Medida óptica
2 – Limpieza de inyectores
3 – Puntos de Medición (emisor, receptor)

Fuente: (S::CAN 2007)

La sección de medición cuenta también con accesorios para la limpieza automática de la ventana de medición, limpieza que se efectúa con aire o agua a presión en diferentes intervalos de tiempo, según programación en el software y dependiendo de la calidad del agua en estudio. Esto con el fin de optimizar la medida y que los sólidos sedimentados cerca de las ventanas no generen alteraciones de la medida, evitando el paso de luz.

La unidad de detección consta de dos componentes principales: (1) el detector y (2) la sección de operación electrónica. Un sistema óptico se centra en la medición y compensación en un puerto a la entrada del detector, donde la luz es recibida por éste y divide el haz de luz en longitudes de onda y guía los 256 fotodiodos, que transforman la señal, por lo que no es necesario el uso de componentes sensibles al movimiento. La parte electrónica de la sonda se encarga de controlar el proceso de medición y de las etapas de procesamiento para la edición y comprobación de la señal de medición y el cálculo de los parámetros (S::CAN 2007).

Como ya se mencionó anteriormente, la Pontificia Universidad Javeriana en marco del CONVENIO DE INVESTIGACIÓN Y DESARROLLO Nº 9-07-25100-0763-2010: "Estudio de la tratabilidad del agua residual afluente a las futuras plantas de tratamiento Salitre y Canoas (tratamiento secundario) de la ciudad de Bogotá en plantas piloto con el sistema de lodos activados", realizó una toma de muestras iniciando el día 18 de octubre hasta el 11 de noviembre de 2011 (para un total de 25 días), utilizando un Spectro::Lyser - s:can con paso de luz de 35 mm, donde para el proyecto de investigación se realizó un inserto para disminuir el paso de luz a 5 mm, el lugar definido para este proyecto fue la estación de Bombeo de Gibraltar de Bogotá.

Esta estación de bombeo se ubica entre el río Bogotá y el canal Cundinamarca en el barrio Osorio XII de la localidad de Kennedy, en las coordenadas Este 988,904,00 y Norte 1,006,402,00 (Camacho, Díaz-Granados et al. 2001), de la ciudad Bogotá. Esta estación eleva las aguas negras provenientes del interceptor Cundinamarca perteneciente a la cuenca hidrológica del río Fucha, junto con las aguas lluvias que provienen del canal

embalse Cundinamarca, con 4 bombas eléctricas que evacuan 1500 L/s cada una (EAAB-ESP 2007).

Figura 34 Ubicación de la estación de bobeo Gibraltar.

Fuente: (Google Earth 2012)

La toma de muestra se realizó en la cámara de recolección de las aguas residuales de la estación, donde llegan luego de ser bombeadas por dos tornillos de Arquímedes.

Foto 1 Punto de toma de muestra.

Fuente: (Zamora David 2011)

Foto 2 Punto de almacenamiento de equipos de almacenamiento de datos.

Fuente: (Zamora David 2011)

Foto 3 Estructura de contención de sondas.

Fuente: (Zamora David 2011)

Foto 4 Sondas tomando muestra en línea.

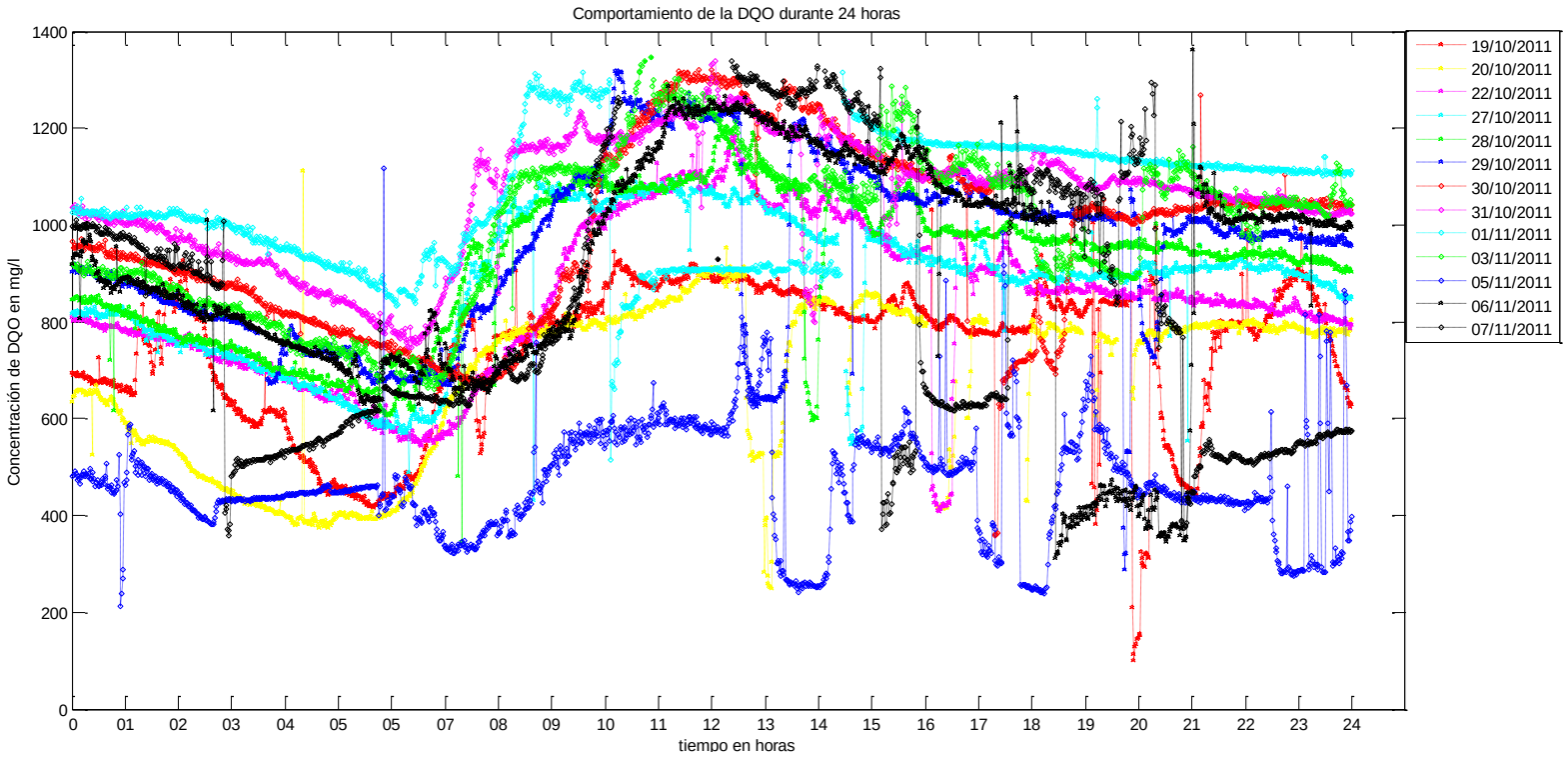
Fuente: (Zamora David 2011)

Los resultados empleados en esta tesis, fueron suministrados por el Ingeniero David Zamora estudiante e investigador de la maestría de Hidrosistemas, es importante mencionar que los valores reportados y entregados, fueron encontrados mediante la calibración con la que el equipo viene de fábrica (calibración global).

El proyecto tomó muestras cada minuto durante las 24 horas obteniendo un total de 1440 datos de DQO equivalente en mg/l, sin embargo durante los 25 días en los que sonda se encontraba funcionando, se presentaron problemas técnicos asociados a la demanda de aguas residuales, ya que en la bomba que suministra las aguas residuales a la estación, en diferentes días no se encontraba funcionando los tornillos de Arquímedes, y que estos se detenían por mantenimiento de las rejillas de cribado.

De acuerdo a lo anterior y con el fin de obtener un comportamiento típico de las concentraciones de DQO medidas para un día, se propuso tomar solo los días que contaban con la totalidad de datos encontrados o días en que falta de valores fuese muy poca (esto se debía a que la sonda inyecta aire para limpiar el lector), por lo tanto y de acuerdo a los criterios tomados solo se pudo seleccionar 13 días de la totalidad, los cuales se pueden observar a continuación.

Figura 35 Comportamiento de la DQO durante 24 horas de los 13 días seleccionados.

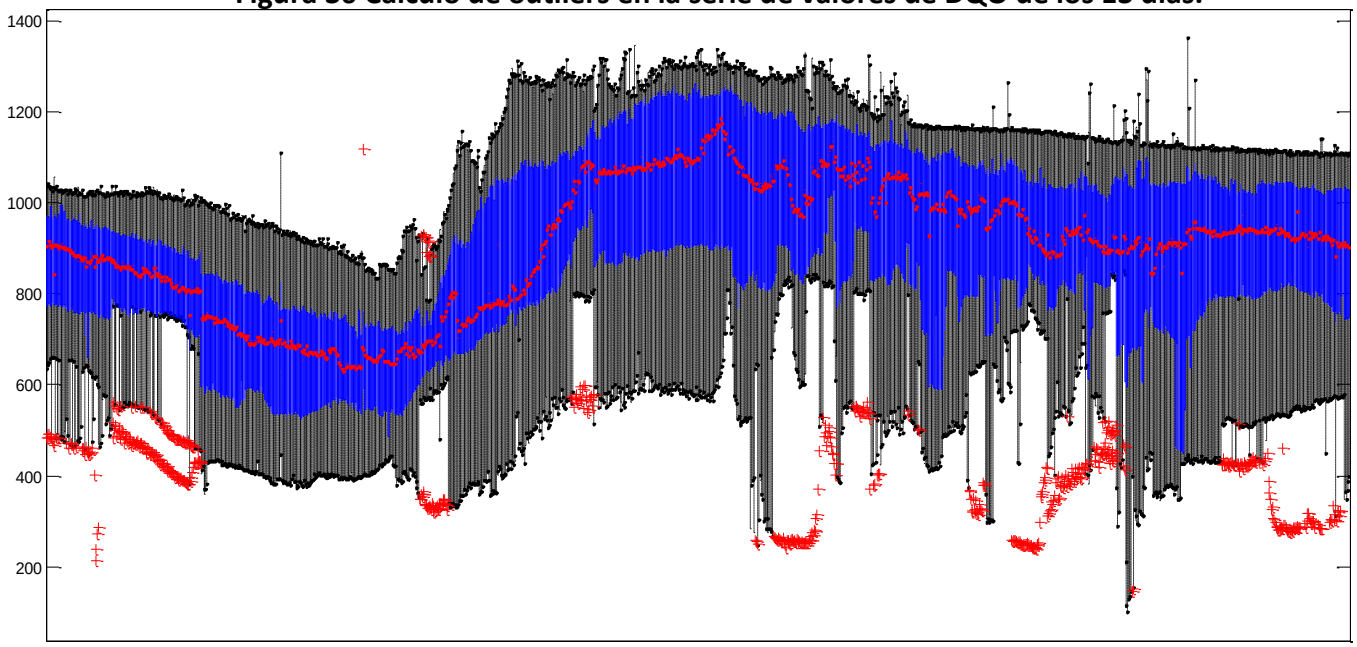


Fuente: (El autor 2012)

Como se puede observar en la anterior gráfica, la mayoría de los días presentan una misma tendencia, la cual muestra como después de las 12 de la noche la concentración de DQO disminuye, para luego desde las 06:00 a.m. aumentar de gran manera, llegando a los picos más altos entre las 11:00 a.m. hasta las 02:00 p.m., donde finalmente vuelve a disminuir. Por otra parte la gráfica presenta días y valores atípicos a diferentes horas, con respecto a lo monitoreado con la sonda.

Por lo tanto se planteó encontrar estos valores anormales o *outliers* usando el diagrama de cajas o *boxplot*, el cual utiliza percentiles, estos son tres valores que dividen un conjunto ordenado en cuatro partes iguales, los percentiles se denotan usualmente por Q1, Q2 y Q3, el primer percentil representa un cuarto de todos los datos (25%), el segundo representa la media de la serie de datos, y el tercero representa las tres cuartas partes de los datos (75%) (Glass and Stanley 1986), los brazos o “bigotes” representan los valores máximos y mínimos dentro del rango de los percentiles y los valores fuera de estos rangos, son aquellos que se consideran atípicos, por consiguiente se obtuvo:

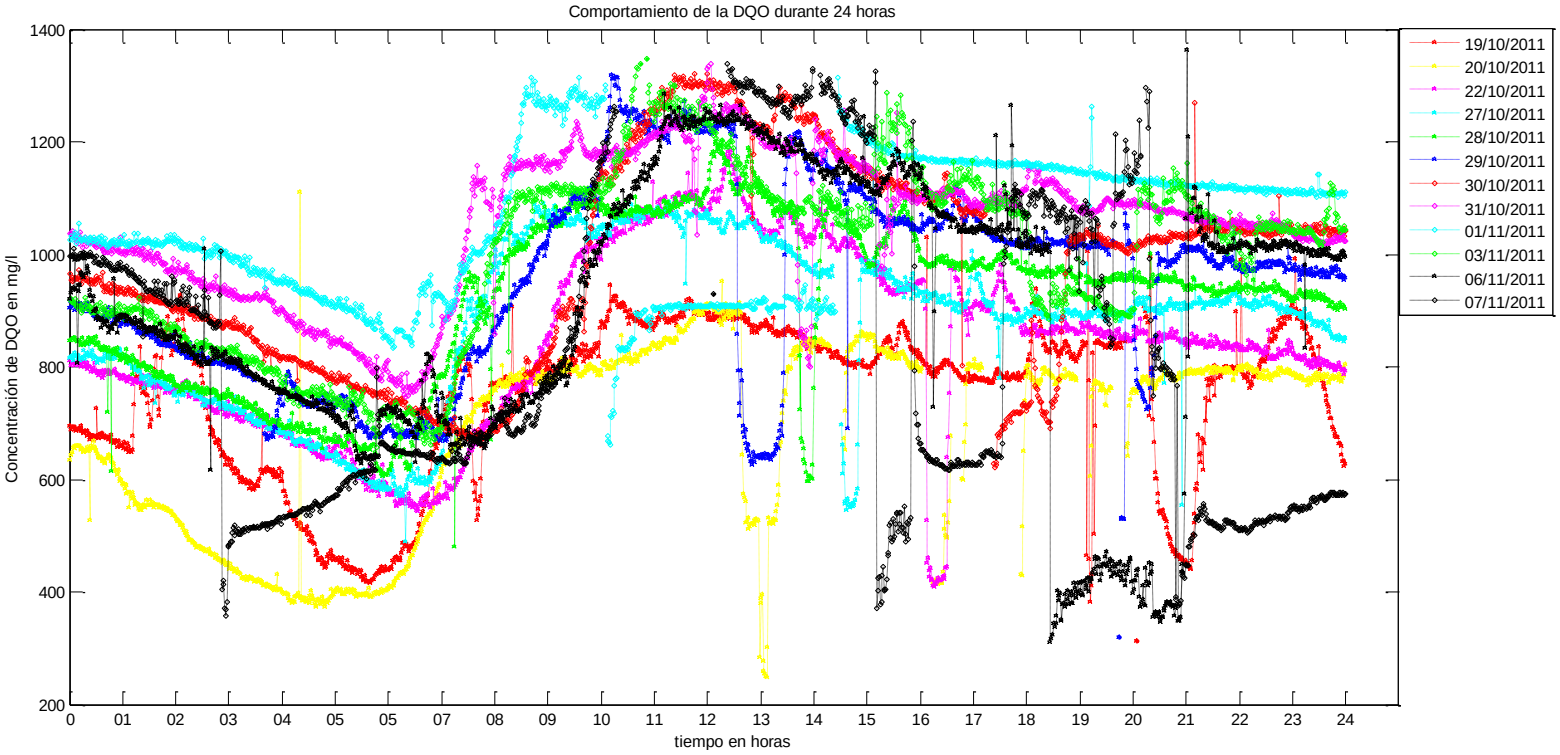
Figura 36 Cálculo de outliers en la serie de valores de DQO de los 13 días.



Fuente: (El autor 2012)

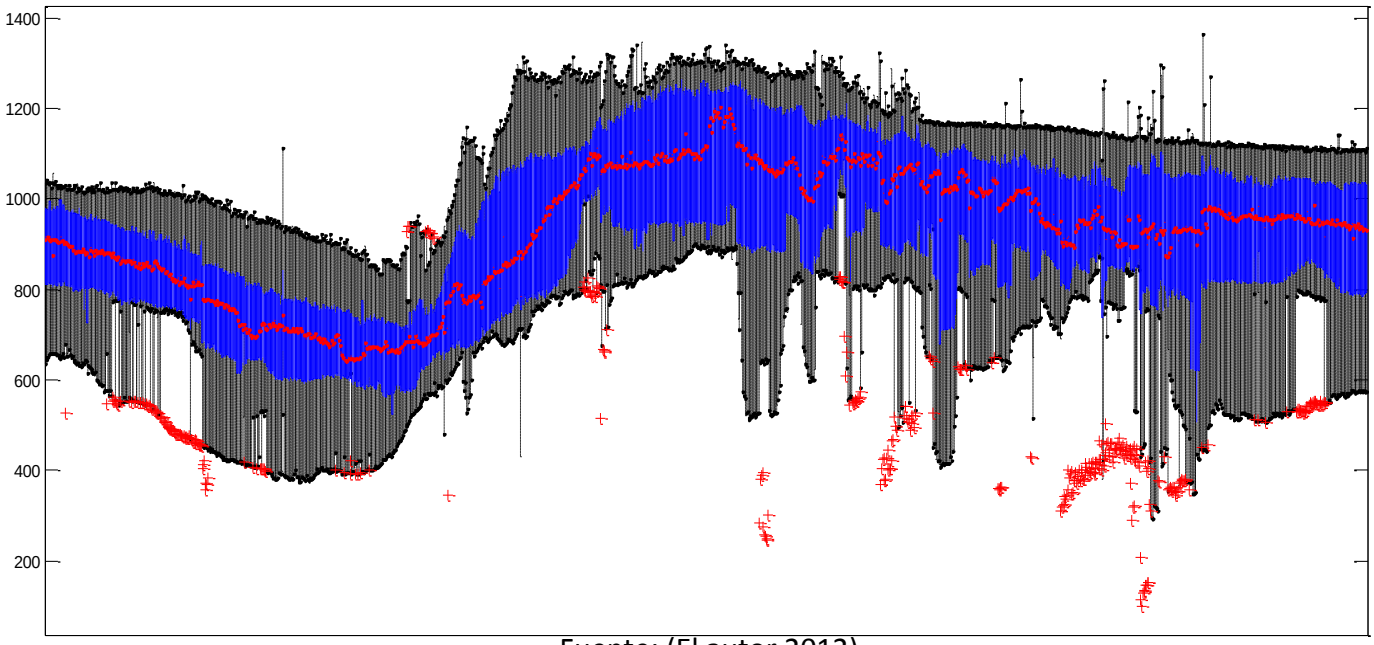
De acuerdo a este análisis, se encontró que el día 5 de septiembre de 2011 presentaba una gran diferencia en las concentraciones con respecto a los demás días, por lo tanto se decidió eliminar este día y realizar nuevamente el análisis con los demás días.

Figura 37 Comportamiento de la DQO durante 24 horas de los 12 días seleccionados



Fuente: (El autor 2012)

Figura 38 Cálculo de outliers en la serie de valores de DQO de los 12 días.

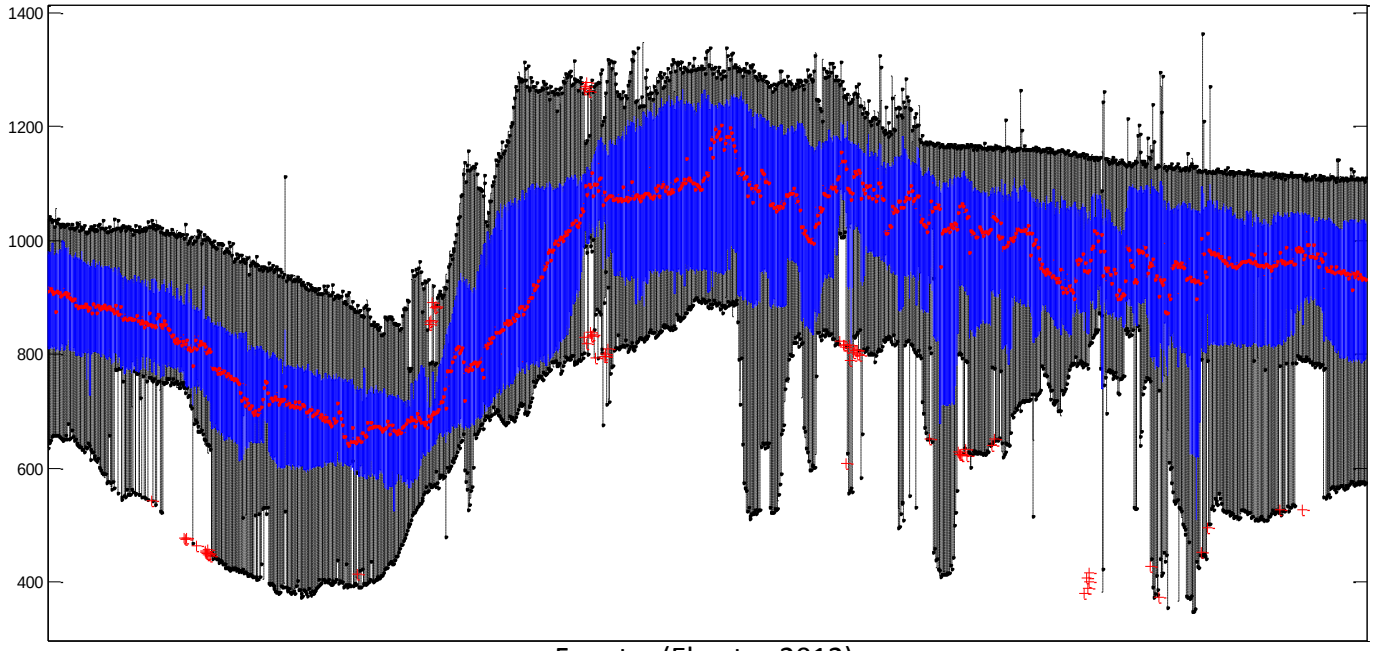


Fuente: (El autor 2012)

De acuerdo lo presentado en la gráfica 37 se presentan varios valores *outliers,* sin embargo se define que estos no se pueden asociar solamente a un día especifico, ya que se encontró que los *outliers* son de diferentes series, por consiguiente se decidió, que para no perder una o varias series de concentraciones, se cambiarían estos valores atípicos por NaN en la serie, así se garantizaría que aquellas concentraciones que se encuentren dentro de los bigotes permitirian la búsqueda del día típico.

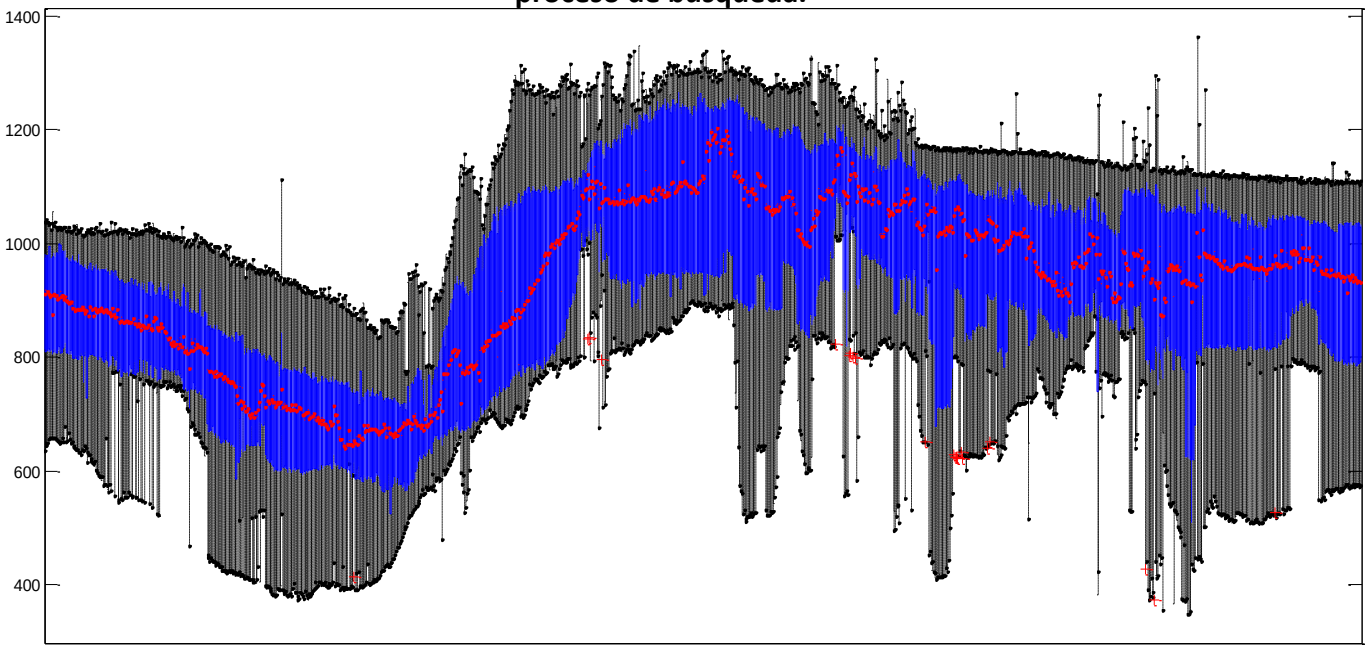
De esta manera, se decide repetir este proceso varias veces, hasta no encontrar más *outliers* y así al hallar el percentil 50 de las concentraciones de DQO dentro de los 12 días, para que estos valores de DQO simulen el día típico deseado para nuestro modelo.

Figura 39 Cálculo de outliers en la serie de valores de DQO de los 12 días en el segundo proceso de búsqueda.



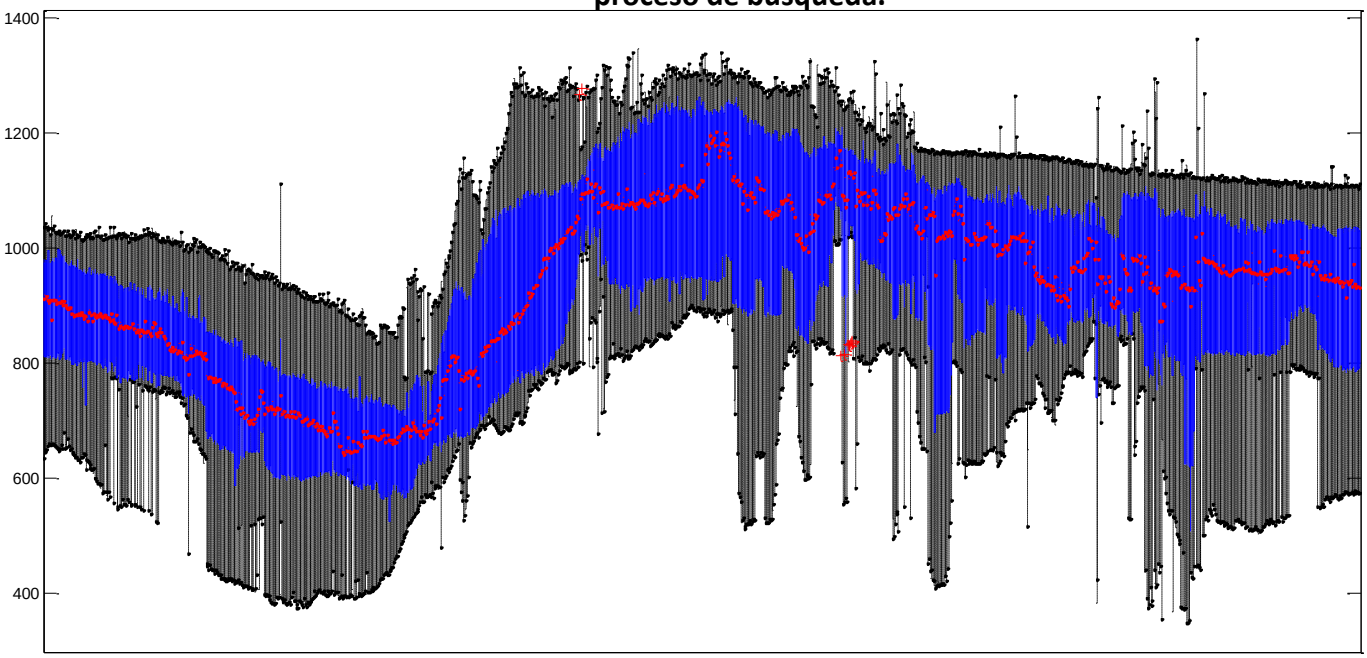
Fuente: (El autor 2012)

Figura 40 Cálculo de outliers en la serie de valores de DQO de los 12 días en el tercer proceso de búsqueda.



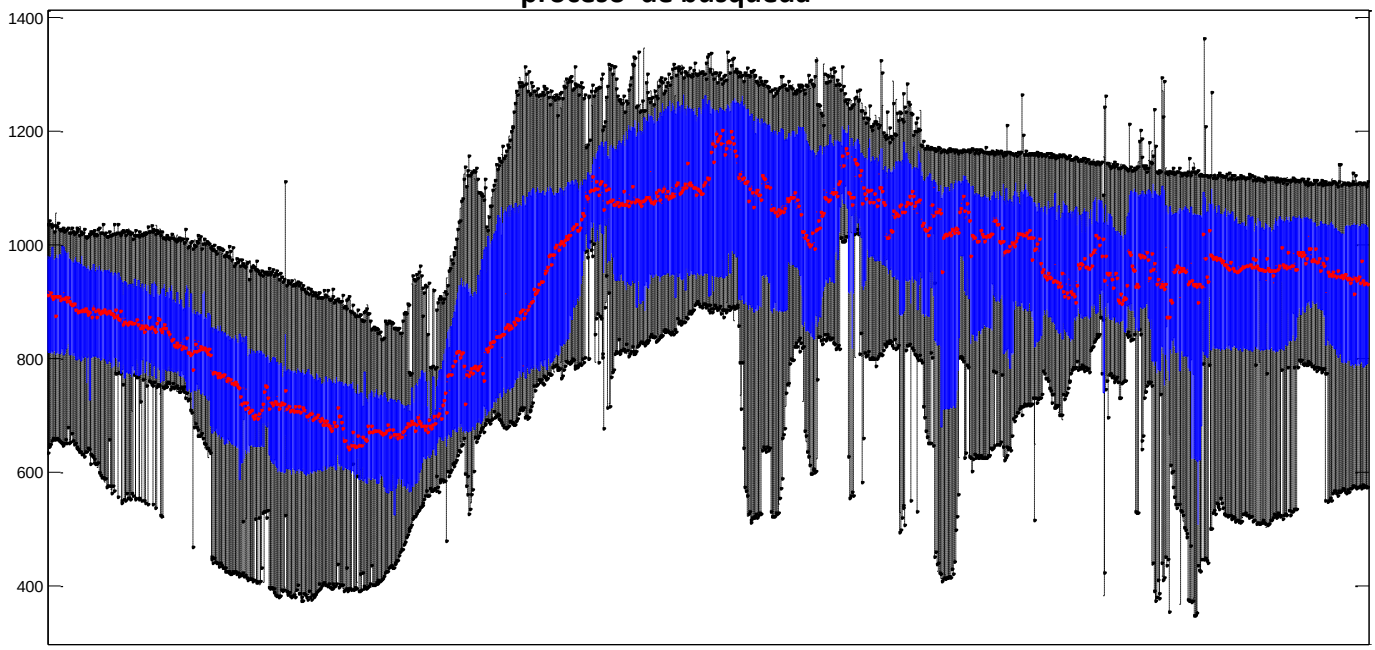
Fuente: (El autor 2012)

Figura 41 Cálculo de outliers en la serie de valores de DQO de los 12 días en el cuarto proceso de búsqueda.



Fuente: (El autor 2012)

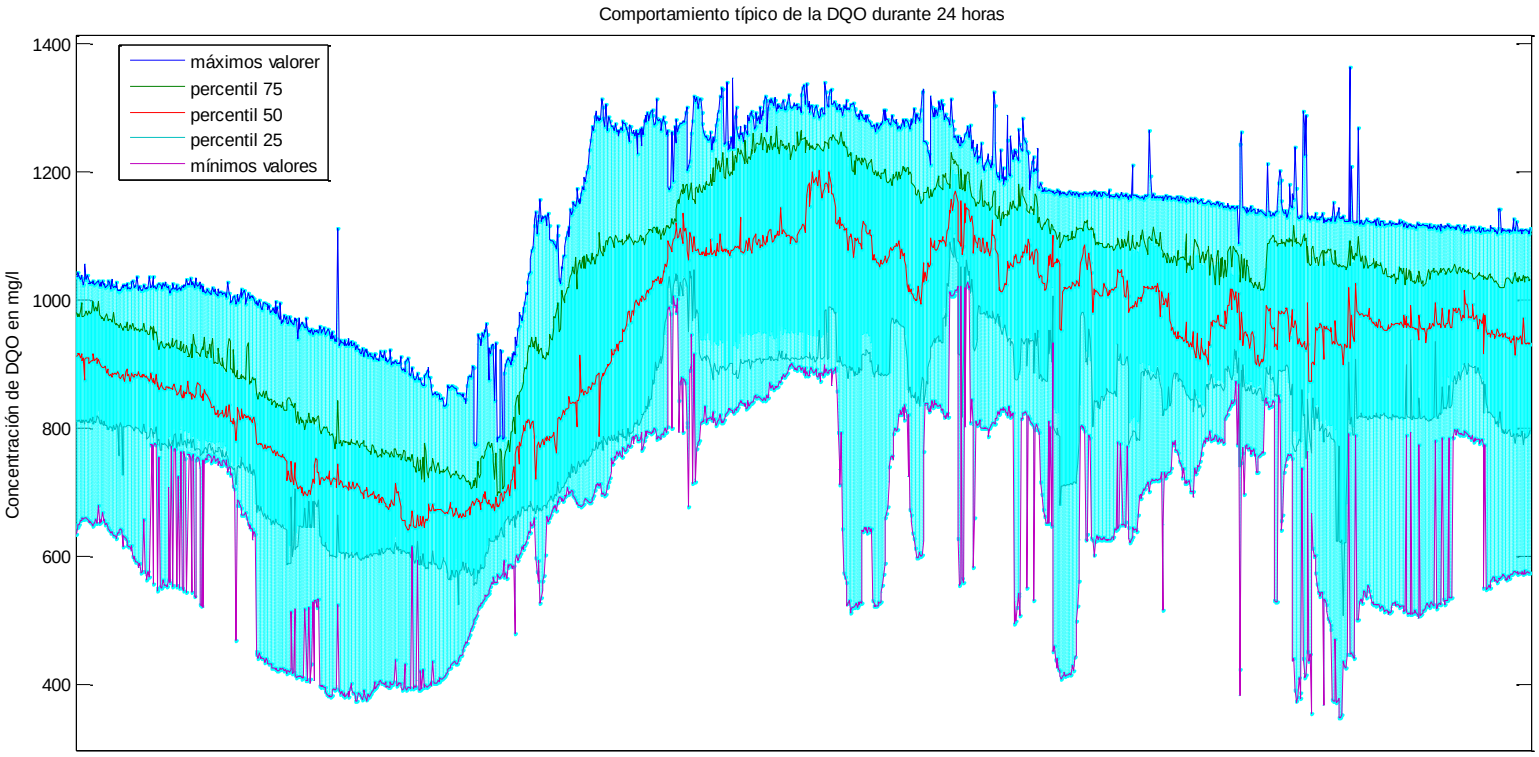
Figura 42 Cálculo de outliers en la serie de valores de DQO de los 12 días en el quinto proceso de búsqueda



Fuente: (El autor 2012)

Al quinto método de evaluación de *outliers,* ya fueron detectados más por parte del programa, por lo tanto se decidió que el percentil 50 representaría el día promedio, el cual para verlo con mayor calidad se graficó de la siguiente manera.

Figura 43 Comportamiento típico de la DQO durante 24 horas.



Fuente: (El autor 2012)

4.3. Modelación de un sistema de lodos activados

El fundamento principal del control es la retroalimentación, ya que los procesos están sujetos todo el tiempo a las perturbaciones, por ejemplo el presente estado de cualquier proceso es medido por algún sensor y este es la base de una decisión objetivo (Olsson, Nielsen et al. 2005; Olsson 2007), en otras palabras: *el control es acerca de cómo operar una planta o un proceso hacia un objetivo definido, a pesar de la perturbaciones* (Olsson and Newell 1999).

Por lo tanto, este proyecto ha planteado que para la ejecución del control sobre la planta, este se realice siguiendo dos etapas básicas, la predicción y el control. La primera hace referencia a la predicción de medidas o variables estimadas en un delta de tiempo futuro, en la segunda etapa, el controlador al obtener esta medición o variable calculada, realiza las acciones necesarias para alcanzar un valor objetivo deseado y volver a ejecutar nuevamente las acciones.

De acuerdo a esto, la modelación 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 en deltas de tiempo futuros.

Por lo tanto a continuación se contará un poco sobre el modelo de lodos activados y se describirá el proceso hecho para este proyecto de tesis.

En 1983, la Asociación Internacional en Calidad del Agua IAWQ (por sus siglas en ingles International Association on Water) formó un grupo de trabajo, el cual tenía como función promover el desarrollo y facilitar la aplicación de modelos para el diseño y operación de sistemas biológicos de tratamiento de aguas residuales. El primer objetivo fue revisar los modelos existentes y la segunda meta era llegar a un conceso en torno al más simple modelo matemático, que tuviera la capacidad en su forma más realista de predecir el comportamiento de los sistemas, donde se lleve a cabo oxidación de carbono, nitrificación y desnitrificación (Jeppsson 1996; Makinia 2010), adicionalmente el modelo está encaminado a ceder una buena descripción de la producción de fangos. La demanda química de oxígeno, se adoptó como medida de la concentración de materia orgánica, en el modelo, la amplia variedad de compuestos orgánicos de carbono y compuestos nitrogenados se subdividen en un número limitado de fracciones basadas en consideraciones de biodegradabilidad y solubilidad (Gernaey, van Loosdrecht et al. 2004).

El resultado de este trabajo fue el modelo conocido como IAWQ Activated Sludge Model No. 1 (ASM1) (Henze, Gujer et al. 2000), el cual puede ser considerado como el modelo referencia, este modelo provocó la aceptación general de la modelación de PTAR, en primer lugar en la comunidad investigadora y más tarde también en la industria. Muchos conceptos básicos del ASM1 fueron adaptados del modelo de lodos activados definido por Dold *et al.* en 1980, a la fecha el ASM1 es referencia para muchos proyectos científicos y ha sido implementado (en algunos casos con modificaciones) en la mayoría de software comerciales para la modelación y la simulación de PTAR's (Gernaey, van Loosdrecht et al. 2004) como Single Sludge Simulation Program (SSSP) desarrollado por Bistrup and Grady en 1988, ASIM de Gujer y Henze de 1991 y GPS-X de Patry y Takacs de 1990 (Rustum 2009).

El ASM1 contiene 18 parámetros, de los cuales 5 son estequiométricos y 13 coeficientes cinéticos (Rustum 2009), el modelo fue primariamente desarrollado para plantas de tratamiento de lodos activados, en la tabla 5 se muestra el listado del modelo ASM1, en la tabla 6 se define las variables de estado y en las tablas 7 y 8 el listado de coeficientes cinéticos y estequiométricos.

Tabla 5 Modelo numero 1 de lodos activados (ASM1).

Componente → i	1	2	3	4	5	6	7	8	9	10	11	12	13	**Tasas de reacción, ρj (M/(L3 T))**
j ↓ Proceso	S1	S2	X1	XS	XB,H	XB,A	XP	SO	SON	SNH	SND	XND	SALK	
1 Crecimiento aeróbico de heterótrofos		$-\frac{1}{Y_H}$			1			$-\frac{(1-Y_H)}{Y_H}$	$-\frac{(1-Y_H)}{(2.86Y_H)}$	$-i_{XB}$			$-\frac{i_{XB}}{14}$	$\mu_{H,max}\left[\frac{S_0}{K_{O,H}+S_0}\right]\left[\frac{S_s}{K_s+S_s}\right]X_{B,H}$
2 Crecimiento anóxico de heterótrofos		$-\frac{1}{Y_H}$			1			$\frac{(4.57-Y_A)}{Y_A}$	$\frac{1}{Y_A}$	$-i_{XB}$			$\left[\frac{(1-Y_H)}{(14.x2.86Y_H}\right]-\frac{i_{XB}}{14}$	$\mu_{H,max}\left[\frac{S_s}{K_s+S_s}\right]\left[\frac{K_{O,H}}{K_{O,H}+S_0}\right]\left[\frac{S_{NO}}{K_{NO}+S_{NO}}\right]\eta_g X_{B.H}$
3 Crecimiento anóxico de autótrofas						1				$-i_{XB}-\frac{1}{Y_A}$			$-\frac{i_{XB}}{14}-\frac{1}{7Y_A}$	$\mu_{A,max}\left[\frac{S_0}{K_{0,A}+S_0}\right]\left[\frac{S_{NH}}{K_{NH}+S_{NH}}\right]X_{B,A}$
4 Decaimiento de de heterótrofos				$1-f_p$	-1		f_p					$i_{XB}-f_p i_{XP}$		$b_H X_{B,H}$
5 Decaimiento de autótrofas				$1-f_p$		-1	f_p					$i_{XB}-f_p i_{XP}$		$b_A X_{B,A}$
6 Amonificación de nitrógeno orgánico soluble		1		-1						1	-1		$\frac{1}{14}$	$k_a S_{ND} X_{B,H}$
7 Hidrólisis de los orgánicos enredados														$k_h \frac{\frac{X_S}{X_{B,H}}}{K_X+\left(\frac{X_S}{X_{B,H}}\right)}\left[\left(\frac{S_0}{K_{O,H}+S_0}\right)+\eta_h\left(\frac{K_{O,H}}{K_{O,H}+S_0}\right)\left(\frac{S_{ND}}{K_{NO}+S_{NO}}\right)\right]X_{B,H}$
8 Hidrólisis del nitrógeno orgánico enredado											1	-1	-1	$\rho_7\left(\frac{X_{ND}}{X_S}\right)$
Tasa de conservación Total (observada)	$r_{i\,neta}=\sum V_{ji}\rho_i$													

Fuente: (Mulas 2006; Rustum 2009)

Tabla 6 Definición de variables de estado del modelo ASM1.

Numero del componente	Símbolo del componente	Definición
1	S_I	Materia orgánica inerte soluble $M(DQO)L^{-3}$
2	S_s	Materia biodegradable enredada $M(DQO)L^{-3}$
3	X_I	Materia particulada orgánica inerte $M(DQO)L^{-3}$
4	X_s	Sustrato biodegradable lento $M(DQO)L^{-3}$
5	X_{BH}	Biomasa heterotrófica activa $M(DQO)L^{-3}$
6	X_{BA}	Biomasa autotrófica activa $M(DQO)L^{-3}$
7	X_p	Productos del decaimiento de la biomasa $M(DQO)L^{-3}$
8	S_0	Oxígeno disuelto $M(-DQO)L^{-3}$
9	S_{NO}	Nitrato y nitritos $M(N)L^{-3}$
10	S_{NH}	Nitrógeno amoniacal $M(N)L^{-3}$
11	S_{ND}	Nitrógeno orgánico biodegradable soluble $M(N)L^{-3}$
12	X_{ND}	Nitrógeno orgánico biodegradable particulado $M(N)L^{-3}$
13	S_{ALK}	Alcalinidad - Molar

Fuente: (Rustum 2009)

Tabla 7 Coeficientes cinéticos del modelo ASM1.

Evento cinético	Símbolos	Unidad
Tasa de crecimiento especifica. Máximo Heterotrófico	$\widehat{\mu_H}$	Día^{-1}
Tasa de decaimiento heterotrófico	b_H	Día^{-1}
Coeficiente medio de saturación para heterótrofos	K_s	g DQO m^{-3}
Coeficiente medio de saturación de oxígeno para heterótrofos	K_{NO}	g NO_3N m^{-3}
Tasa de crecimiento especifica. Máximo autotrófico	$\widehat{\mu_A}$	Día $^{-1}$
Tasa de decaimiento autótrofa	b_A	Día $^{-1}$
Coeficiente medio de saturación de oxígeno para autótrofos	$K_{O,A}$	g O_2 m^{-3}
Coeficiente medio de saturación de amonio para autótrofos	K_{NH}	g NH_3N m^{-3}
Factor de corrección para el crecimiento anóxico de heterótrofos	μ_g	
Tasa de amonificación	k_a	m^3 (g DQO día)$^{-1}$
Tasa máxima especifica de hidrólisis	k_h	g de DQO biodegradable lenta (g celda DQO día)$^{-1}$
Coeficiente medio de saturación para hidrólisis de sustrato biodegradable lento	K_g	g de DQO biodegradable lenta (g celda DQO día)$^{-1}$
Factor de corrección para hidrólisis anóxica	μ_h	

Fuente: (Rustum 2009)

Tabla 8 Coeficientes estequiométricos del modelo ASM1.

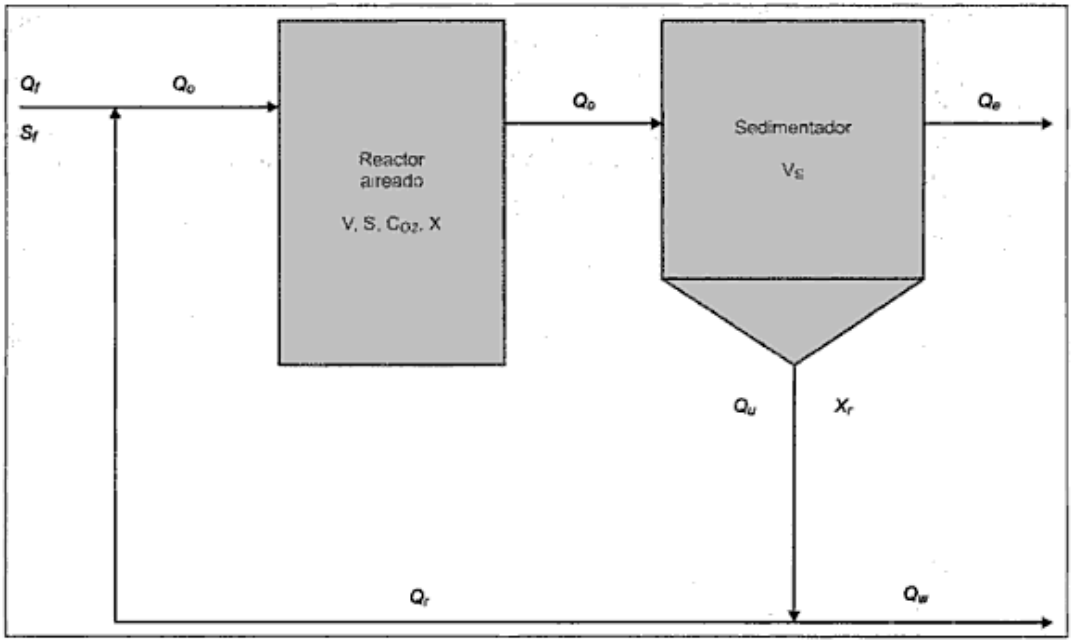
Coeficiente estequiométricos	Símbolo	Unidad
Rendimiento heterotrófico	Y_H	g de celda formada (g DQO oxidado)$^{-1}$

Coeficiente estequiométricos	Símbolo	Unidad
Rendimiento autotróficos	Y_A	g de celda formada (g DQO oxidado)$^{-1}$
Fracción de rendimiento de biomasa decaída por producto	f_p	
Masa de N/Mas de DQO en biomasa	i_{XB}	g N (g DQO)$^{-1}$ en biomasa
Masa de N/Mas de DQO en decaimiento de producto	i_{XP}	g N (g DQO)$^{-1}$ en masa endógena

Fuente:(Rustum 2009)

Del anterior modelo, Sergio Alejandro Martínez Delgadillo y Miriam Guadalupe Rodríguez Rosales en su libro "Tratamiento de aguas residuales con Matlab", capítulo sexto "Modelación del proceso de lodos activados", toman la dinámica del mismo, para presentar un set de cuatro ecuaciones basadas en el siguiente diagrama:

Figura 44 Diagrama del sistema de lodos activados



Fuente: (Martinez 2005)

Las siguientes ecuaciones muestran el balance en el sistema para obtener el comportamiento de la 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***
    - DQO

$$\frac{ds}{dt} = \frac{Q_f}{V} S_f - \frac{Q_0}{V} S - \frac{\mu X}{Y} \quad (60)$$

    - Biomasa (SSV)

$$\frac{dX}{dt} = \frac{Q_r}{V} X_r - \frac{Q_0}{V} X + \mu X - k_d X \quad (61)$$

    - Oxígeno disuelto (O_2)

$$\frac{dC_{O_2}}{dt} = \frac{Q_f}{V} C_{o_{2f}} - \frac{Q_0}{C} C_{O_2} - \frac{\mu X}{Y_{O2}} - b * X + Kla_w * (C_{sr} - C_{O_2}) \quad (62)$$

- ***En el sedimentador***
    - Biomasa (SSV)

$$\frac{dX_r}{dt} = \frac{Q_u}{Vs} X_r - \frac{Q_0}{Vs} X \quad (63)$$

Donde:

μ = Velocidad específica de crecimiento máximo.
b = Kg de (0_2) para la respiración endógena
Y = Coeficiente de rendimiento [mg(SSV)producido/mg(DQO)consumido]
Y_o = Coeficiente de rendimiento de oxígeno
k_d = Coeficiente de muerte bacteriano
S_f = Concentración de sustrato en el efluente
C_{o2f} = Concentración de oxígeno en el afluente
C_{sr} = Concentración de saturación de oxígeno
K_{law} = Coeficiente de transferencia de oxígeno
Q_f = Caudal de agua del afluente
Q_r = Caudal de recirculación
Q_w = Caudal de purga
V = Volumen del reactor
V_s = Volumen del sedimentador

Adicionalmente dentro del libro, se presenta un código en Matlab para la modelación de este sistema y su demostración mediante un ejemplo práctico en las páginas 161 y 162, por lo tanto se tomó este código para la simulación del consumo del sustrato, comportamiento de la biomasa y del oxígeno en el reactor, tomando los parámetros y factores del ejemplo de la pagina 160 y así mirar la respuesta con el día típico calculado.

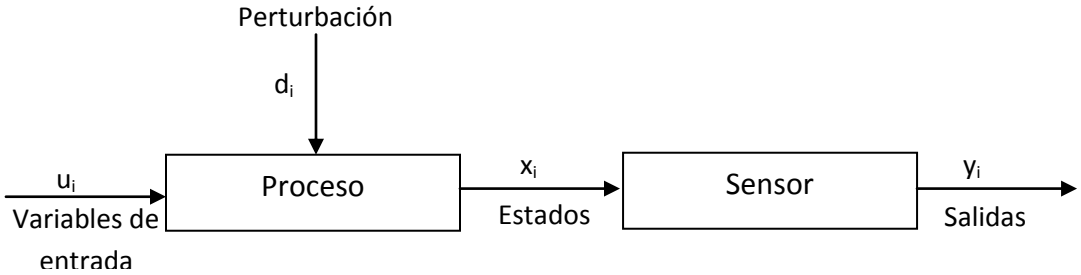
Para entender las variables a utilizar en este proyecto, las identificaremos a continuación:

- **Variables de entrada (u_i)**: Estas variables son aquellas que alimentarán el modelo de lodos activados, estas son; el sustrato de entrada, caudal de agua residual, caudal de purga de lodos, caudal de recirculación de lodos y el caudal de aire que se inyecta al reactor.

- **Variables de estado (x_i)**: Son el conjunto de variables independientes, las cuales únicamente determinan el estado del proceso, como: biomasa y oxígeno disuelto.

- **Variables de salida (y_i)**: Son variables que se pueden observar y las cuales están relacionadas con las variables de estado, como el sustrato de salida y el crecimiento bacteriano.

Figura 45 Variables del proceso.



Fuente:(Olsson and Newell 1999)

De acuerdo a lo anterior, a continuación se muestra la clasificación de variables:

Tabla 9 Clasificación de variables.

Clase de variable	Nombre	Observación
Perturbaciones	Sustrato de entrada (DQO)	Cambios fuertes de concentraciones en cortos tiempos.
Variables de entrada	Sustrato de entrada (DQO)	Entrada no constante.
	Caudal de aguas residuales	Entrada constante 10000 m^3/d
	Caudal de recirculación de lodos.	Entrada variable y controlable.
	Caudal de purga de lodos.	Entrada constante 300 m^3/d

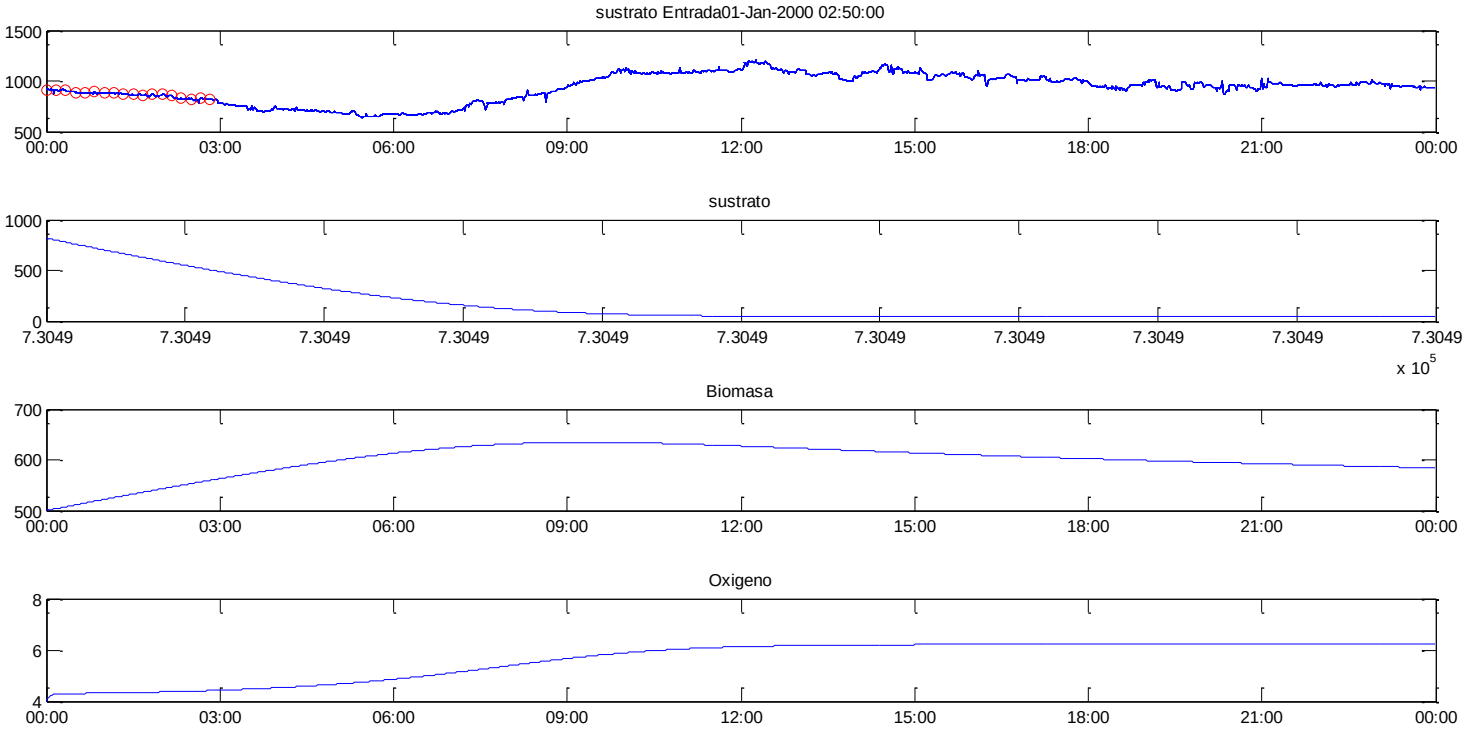
Clase de variable	Nombre	Observación
	Caudal de aire.	Entrada variable y controlable.
Estados	Sustrato calculado	Ecuación 59.
	Oxígeno disuelto	Ecuación 60.
	Biomasa en reactor	Ecuación 61.
	Biomasa en sedimentador	Ecuación 62.
Salidas	Concentraciones de oxígeno disuelto.	Variable a no superar.
	Concentraciones de sustrato a la salida.	Variable a no superar.
Parámetros conocidos	Concentración de oxígeno en el afluente.	0.3 mg/l.
	Concentración de oxígeno de saturación.	7.02 mg/l.
	Volumen del reactor.	5000 m^3.
	Volumen del sedimentador.	250 m^3.
	Coeficiente de muerte (kd).	0.0601 1/d
	Coeficiente de rendimiento de oxígeno (Yo).	0.915.
	Constante de afinidad (ks)	137.3 mg/l
	Coeficiente de rendimiento (Y)	0.33 [mg(SSV)producido/mg(DQO)consumido]
	Kg de (02) para la respiración endógena (b).	0.259 1/d.
	Tasa especifica de crecimiento máxima (μ).	1.97 1/d

Fuente:(Autor 2012)

Para esto se realizaron unas modificaciones en el programa, ya que originalmente el programa resuelve las ecuaciones diferenciales mediante el comando "*ODE45*", el cual está basado en una fórmula explícita de Runge-Kutta (4,5), lo que significa que la solución numérica *ODE45* combina un método de cuarto orden y un método de orden quinto, donde se ajusta automáticamente en el tamaño de paso de integración (Houcque 2005), para este modelo se empleó un método ODE4, con el fin que el paso de integración sea fijo y garantiza la resolución de las ecuaciones en un mismo delta de tiempo.

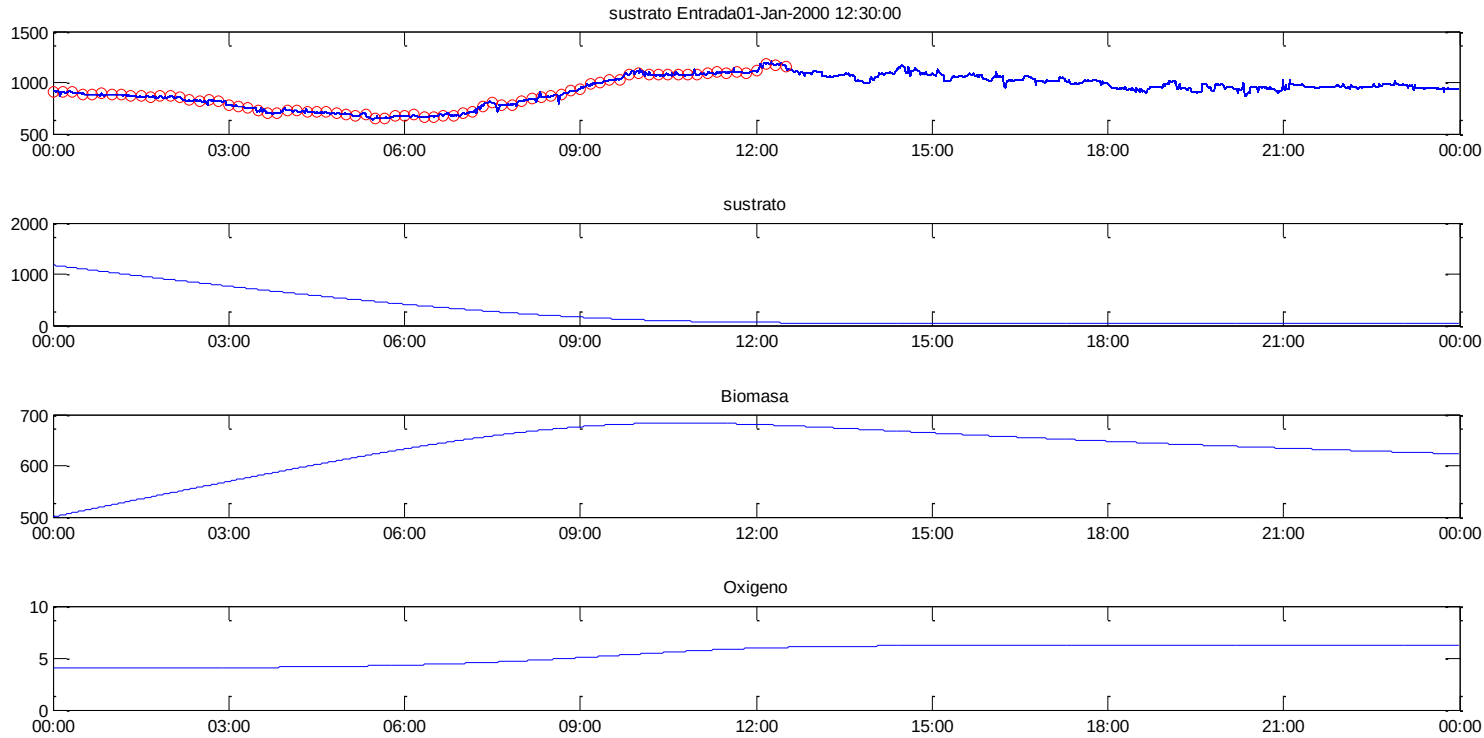
Por consiguiente, se realizó la simulación para los valores de concentración de DQO calculados durante las 24 horas, obteniéndose los siguientes comportamientos:

Figura 46 Simulación del comportamiento del sustrato, biomasa y OD para el valor de DQO de las 02:50 a.m.



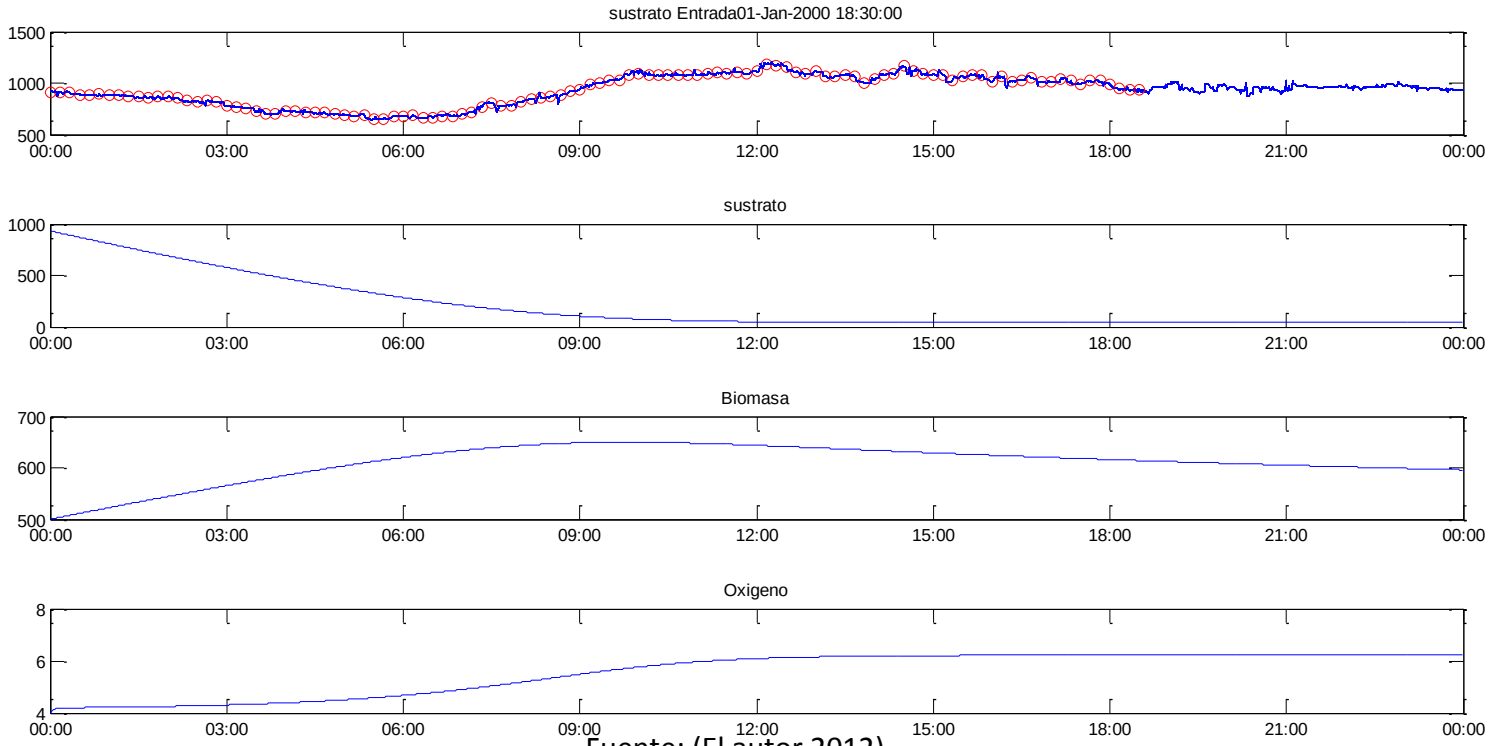
Fuente: (El autor 2012)

Figura 47 Simulación del comportamiento del sustrato, biomasa y OD para el valor de DQO de las 12:30 a.m.



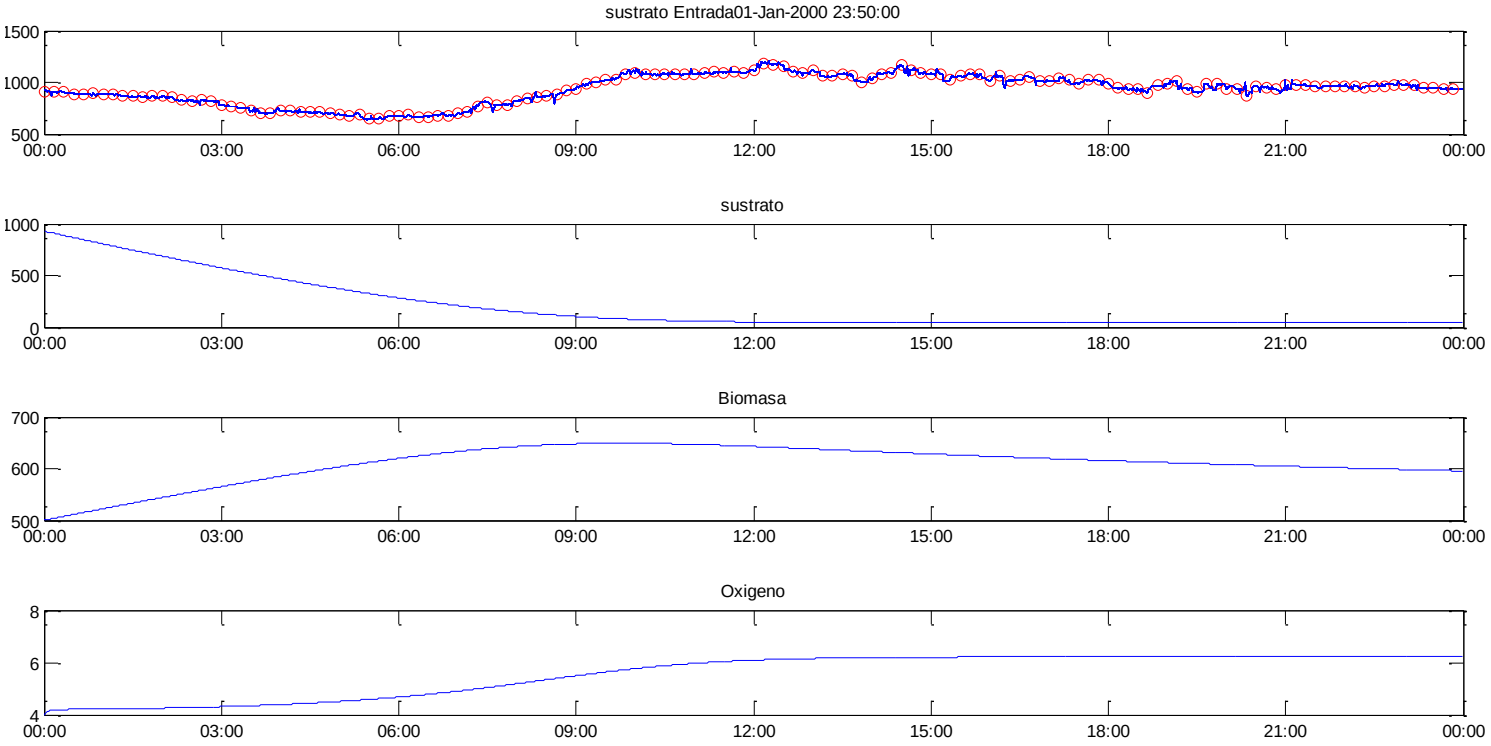
Fuente: (El autor 2012)

Figura 48 Simulación del comportamiento del sustrato, biomasa y OD para el valor de DQO de las 18:30 a.m.



Fuente: (El autor 2012)

Figura 49 Simulación del comportamiento del sustrato, biomasa y OD para el valor de DQO de las 24:00 a.m.



Fuente: (El autor 2012)

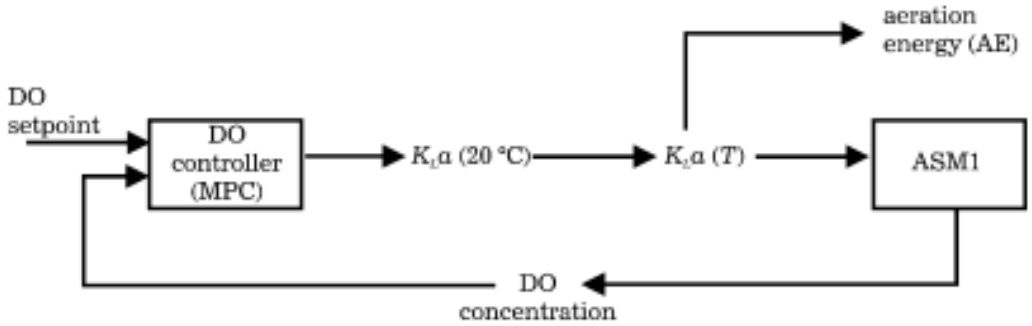
La segunda modificación, se realizó a la estructura principal de la ecuación de oxígeno en el coeficiente de transferencia de oxígeno (K_{lwa}), debido a que el valor de K_{lwa} a menudo se relaciona con el cambio en la intensidad de la aireación y el caudal de aire Q_A suministrado (Makinia and Wells 2007; Chai and Lie 2008; Makinia 2010), ya que estos dos parámetros son proporcionales el uno con el otro (Holmberg, Olsson et al. 1989; Galluzzo, Ducato et al. 2001). El cambio tiene como objetivo, buscar un caudal de aire específico para una concentración de sustrato de entrada, teniendo como referencia una concentración de sustrato deseada a la salida, ya que generalmente la concentración de oxígeno en un tanque aireado, es típicamente controlado 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, por ejemplo el rango en que más se trabaja es entre 1.7 a 2.5 mg/l, siendo 2 mg/l el valor más empleado (Kalker, VanGoor et al. 1999; Brdys, Chotkowski et al. 2002; Fiter, Güell et al. 2005; Piotrowski, Brdys et al. 2008; Thornton, Sunner et al. 2010; Zhang and Guo 2010; Zubowicz, Brdys et al. 2010).

Es importante mencionar que kla no solo depende del caudal de aire, sino también de otros factores, por ejemplo del tipo de difusor, de la composición del agua residual, de la temperatura, del diseño del tanque de aireación, de la profundidad del tanque, ubicación de los difusores, etc. (Carlsson and Lindberg 1998).

Holenda *et al.*, en 2008 utilizó el parámetro K_{lwa} para el control de oxígeno en un tanque de aireación, en donde la concentración de oxígeno disuelto se mide por un sensor en el reactor; el valor de concentración es procesada por el método de control para calcular K_{la}; el K_{la} se corrige de acuerdo con la temperatura, si es necesario; finalmente K_{La} se aplica para cambiar el nivel de concentración de oxígeno en el reactor biológico. Adicionalmente, utilizando el valor de K_{La} se puede estimar el costo energético por la aireación y el volumen de aire soplado por los difusores (Holenda, Domokos et al. 2008). El procedimiento se puede ver a continuación.

Figura 50 Vista esquemática del proceso de control de lodos activados presentado por Holenda, Domokos *et al*. 2008.



Fuente: (Holenda, Domokos et al. 2008)

De acuerdo a la revisión bibliográfica hecha, se encontraron varias formulas matemáticas para el cálculo del coeficiente de transferencia de oxígeno (Lindberg 1997; Olsson and Newell 1999; Makinia and Wells 2000; Makinia and Wells 2007; Makinia 2010):

- Chen *et al.* en 1980 presentaron la siguiente ecuación $K_{La} = m_1 Q_A^{b_1}$ (64).
- Holmberg en 1986 presentó la siguiente ecuación $K_{La} = m_1 Q_A$ (65).
- Goto and Andrews en 1985 presentaron la siguiente ecuación $K_{La} = m_1 Q_A - b_1$ (66).
- Reinius and Hultgren en 1988 presentaron la siguiente ecuación $K_{La} = m_1 Q_A + b_1$ (67).
- Holmberg *et al.* en 1989 la siguiente ecuación $K_{La} = m_1 \sqrt{Q_Q}$ (68).
- Lukasse *et al.* en 1996 plantearon la siguiente ecuación $K_{La} = m_1 Q_A + m_2 \sqrt{Q_A} + m_3$ (69).
- Olsson and Newell en 1999 plantearon la siguiente ecuación $K_{La} = m_1 Q_A + m_3$ (70).
- Olsson and Newell en 1999 plantearon la siguiente ecuación $K_{La} = m_1 \left(1 - e^{-\frac{Q_A m_3}{m_2}}\right)$ (71).

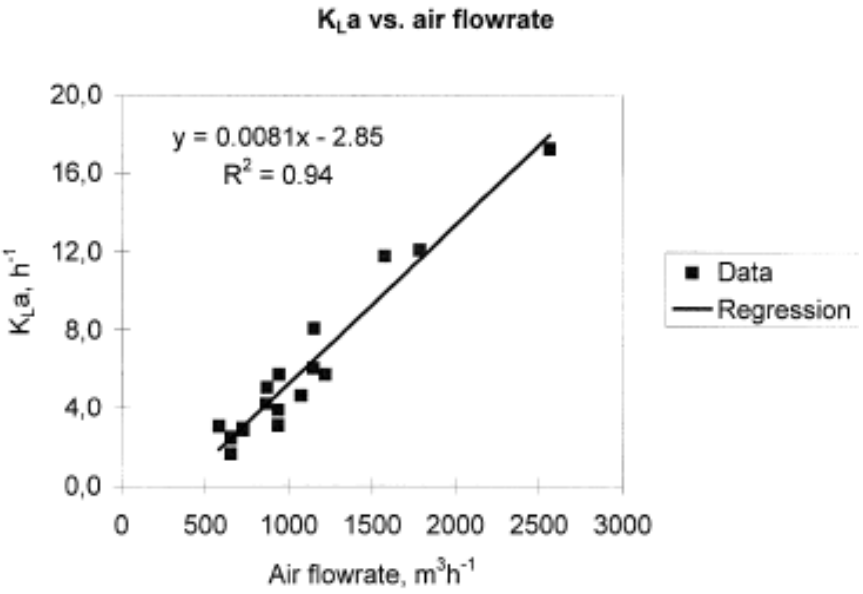
Donde m y b son parámetro empíricos.

El profesor Marsili-Libelli en 1990 utilizó le ecuación de Holmberg de 1986, encontrando que la ecuación para el ejercicio hecho por el, tiene mayor error con caudales de aire grandes (entre 5 x 10^3 m^3h^{-1} y 3 x 10^3 m^3h^{-1}), donde la DBO que entraba a la planta era mayor, esto se debe, a que menor cantidad de aire tiende a crear menos variaciones en las concentraciones de OD (Marsili-Libelli 1990), sin embrago presentó muy buenos resultados durante el día de la simulación.

La ecuación de Olsson y Newell de 1999, describe la transferencia de oxígeno de forma general no lineal y depende del caudal de aireación de accionamiento del sistema y las condiciones de los lodos (Brdys, Chotkowski et al. 2002).

Resultados presentados por Makinia en diferentes artículos, muestran que la regresión entre los valores de K_{la} y el caudal de oxígeno utilizando la ecuación de Goto y Andrews, presentan un r^2 cercano a uno, como se puede ver a continuación.

Figura 51 Regresión de valores de Kla y el caudal de aire reportado por Makina en 2000.



Fuente: (Makinia and Wells 2000)

Por tal motivo se tomará la ecuación 66, la cual se remplazará en la ecuación 62, quedando de la siguiente manera:

$$\frac{dC_{O_2}}{dt} = \frac{Q_f}{V} C_{o_{2f}} - \frac{Q_0}{C} C_{O_2} - \frac{\mu X}{Y_{O2}} - b * X + * m_1 Q_A - b_1 (C_{sr} - C_{O_2}) \quad (72)$$

Donde m_1 = 0.0081 y b_1 = 2.85.

Por consiguiente a esto y de acuerdo a lo observado en las ecuaciones 59 a 62, la ecuación 70 nos sirve para el cálculo de caudales de aire a inyectar en el reactor y así controlar esta tasa, sin embargo esta no afecta directamente en las ecuaciones de sustrato y biomasa, por lo tanto la búsqueda de un valor objetivo de sustrato a la salida del reactor por el agente presentaría dificultades, ya que puede no encontrar su objetivo y quedarse realizando acciones. Como consecuencia de esto, se buscó otro parámetro que pudiera ser controlado y se encontró que la tasa de recirculación de lodos es un método utilizado (Busby and Andrews 1975).

Existen dos estrategias para el control del caudal de recirculación de los lodos, la primera consiste en mantener constante el flujo, la segunda es tener un flujo de lodos variable proporcional al flujo del afluente del sistema. Estos dos métodos de control tradicional presentan desperdició en los consumos energéticos de la plantas de tratamiento, por lo

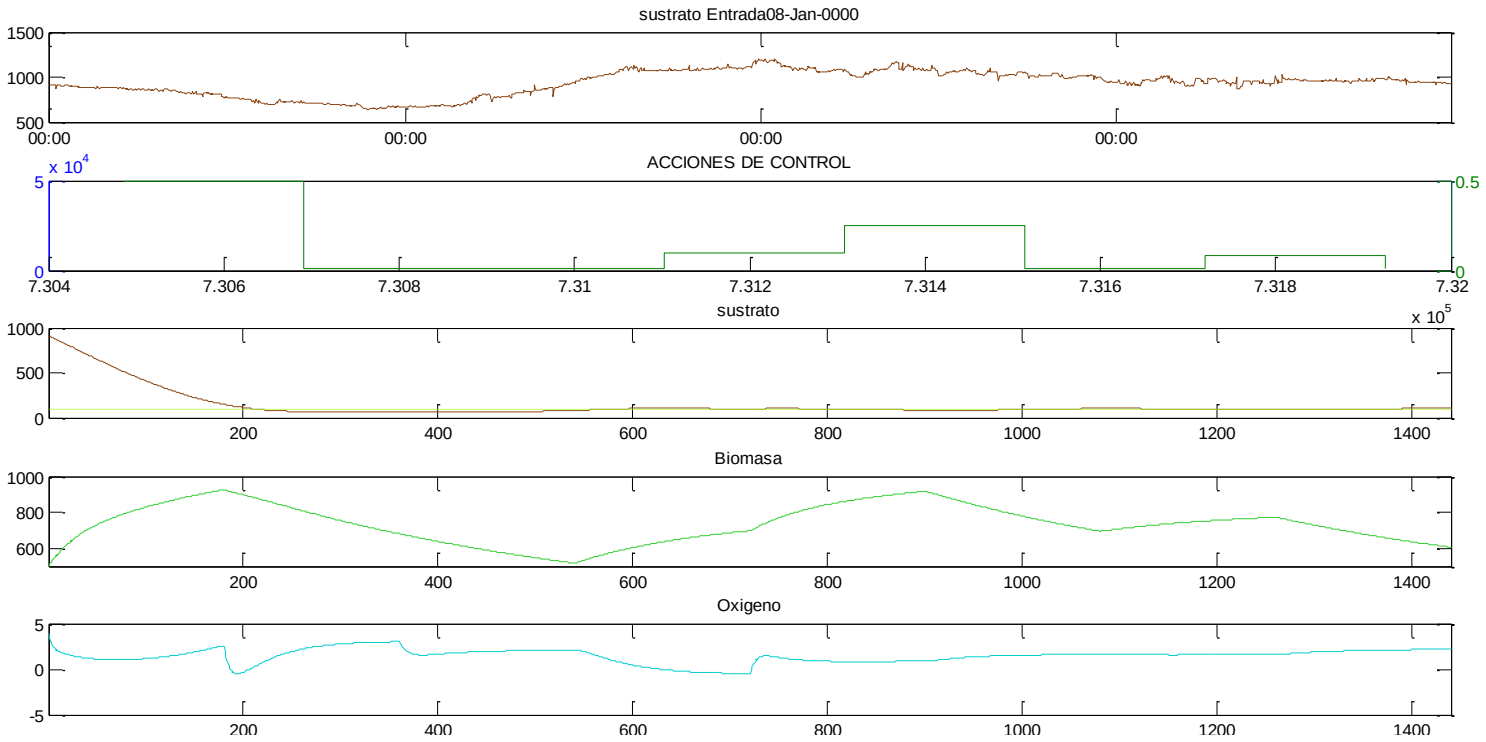
tanto si se pudiera predecir la tasa de recirculación de lodos, se podría controlar estos lodos mejor, lo que daría lugar a ahorro de energía (Long, Fei et al. 2011).

De acuerdo a esto, el agente realizará el control sobre el caudal de aire inyectado y el caudal de recirculación de lodos, al aplicar las acciones necesarias para obtener la mayor recompensa, esto para encontrar un valor de DQO objetivo (política) a la salida del tanque sedimentador, la cual es aproximadamente 100 mg/l. Es importante resaltar que el tiempo de control se llevará a cabo cada hora, debido al tiempo de retención que se llevará dentro del tanque de aireación, el cual afecta el caudal de recirculación, sin embargo cabe resaltar que dentro del algoritmo de toma de decisiones, se puede variar el tiempo de control según lo deseado y el tiempo de retención en el tanque de aireación.

5. RESULTADOS Y DISCUSIÓN

Para la comprobación del control del agente, se inició probando su aprendizaje utilizando los valores obtenidos del día típico y las ecuaciones del modelo de lodos activados, el cual se puede entender de la siguiente manera: el ambiente se representa por el sistema de lodos activados, el cual es modelado por las ecuaciones ya descritas, el agente realiza diferentes acciones de control sobre el medio ambiente, con el objetivo de que sus acciones den un valor de sustrato esperado a la salida, el cual puede generar una ganancia o penalización por esta acción. Para el primer proceso de control se obtuvo:

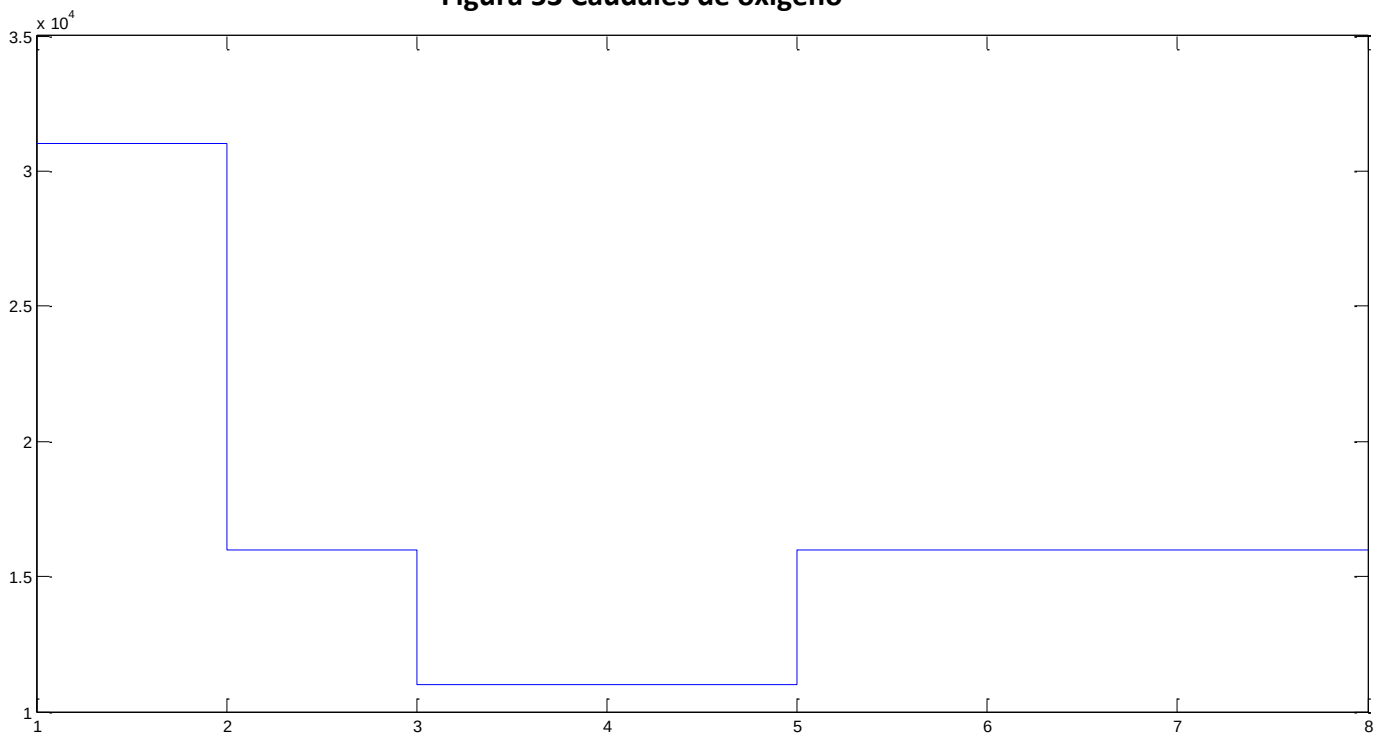
Figura 52 Control realizado sobre el día típico calculado.



Fuente: (El autor 2012)

La anterior gráfica cuenta con cinco (5) subgráficos, donde el primer cuadro presenta el sustrato (DQO) que ingresa al sistema durante las 24 horas, donde en el eje X representa el tiempo y el eje Y los valores de concentración de DQO en mg/l. En el segundo cuadro se presentan las dos acciones de control en el eje X donde los valores de magnitud entre 0 y 5 x 10^4 m^3/d de caudal de aire, en el eje X que tiene magnitudes de 0 a 1.5 es el porcentaje de recirculación de lodos. En el tercer cuadro se cuenta con se presenta el sustrato de salida después de aplicar las acciones de control, el cuarto cuadro se cuenta con el comportamiento de la biomasa en el reactor y por último se obtiene el comportamiento del oxígeno disuelto en mg/l luego de tomar las acciones de control durante el día.

Figura 53 Caudales de oxígeno



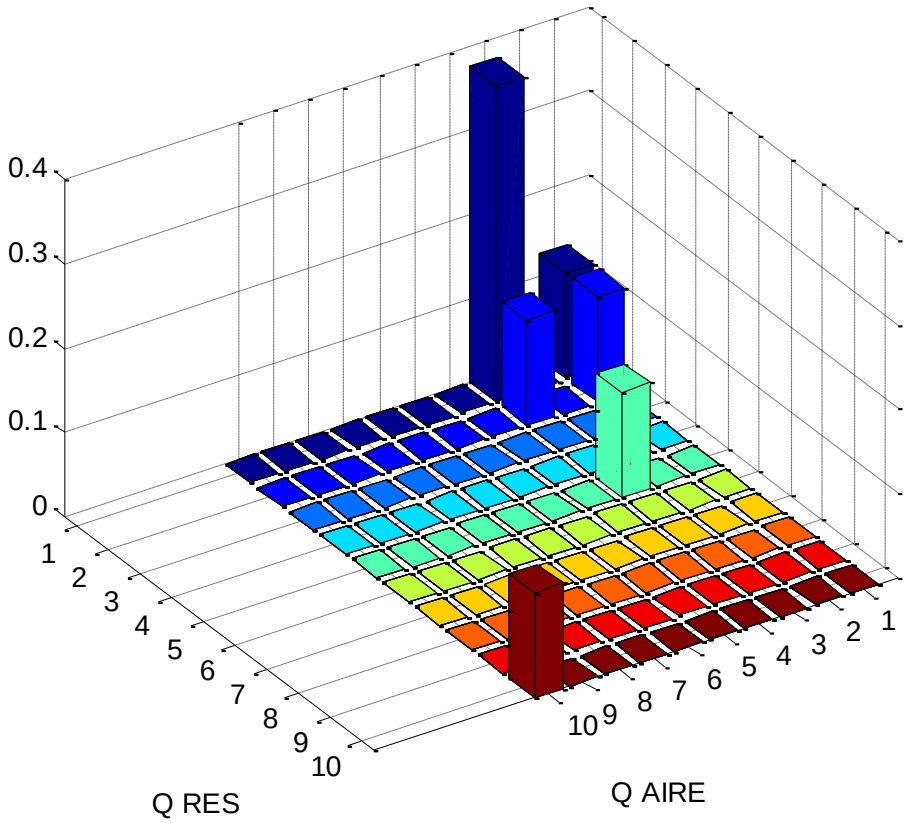
Fuente: (El autor 2012)

Como se puede ver en la figura 50, el agente realizó el control sobre el caudal de recirculación abriendo o cerrando la válvula cada tres horas, teniendo una tasa máxima de recirculación del 50% de lodos, sin embargo durante momentos del día el agente para mantener el objetivo deseado realiza una recirculación muy baja, acercándose al cierre total. En cuanto al caudal de oxígeno el agente disminuye los caudales drásticamente.

Este primer intento presentó un resultado interesante, y es que al aplicar el controlador el caudal de aire disminuyó considerablemente y permitió mantenerse dentro del rango del valor de DQO final deseado, situación se ve igualmente en el caudal de lodos, ya que en momentos del día el caudal es 0 reduciendo energía por bombeo. Sin embrago, a pesar de esto la concentración de oxígeno en el tanque normalmente debe estar entre 1 mg/l y 2 mg/l, situación que no se representa para este primer ensayo, ya que se encuentran valores cercanos a 0 y a 4, adicionalmente el comportamiento de la biomasa es muy

errático, ya que sufre caídas fuertes y grandes crecimientos debido al control hecho al caudal de recirculación.

Figura 54 Distribución de probabilidad de acciones de control.

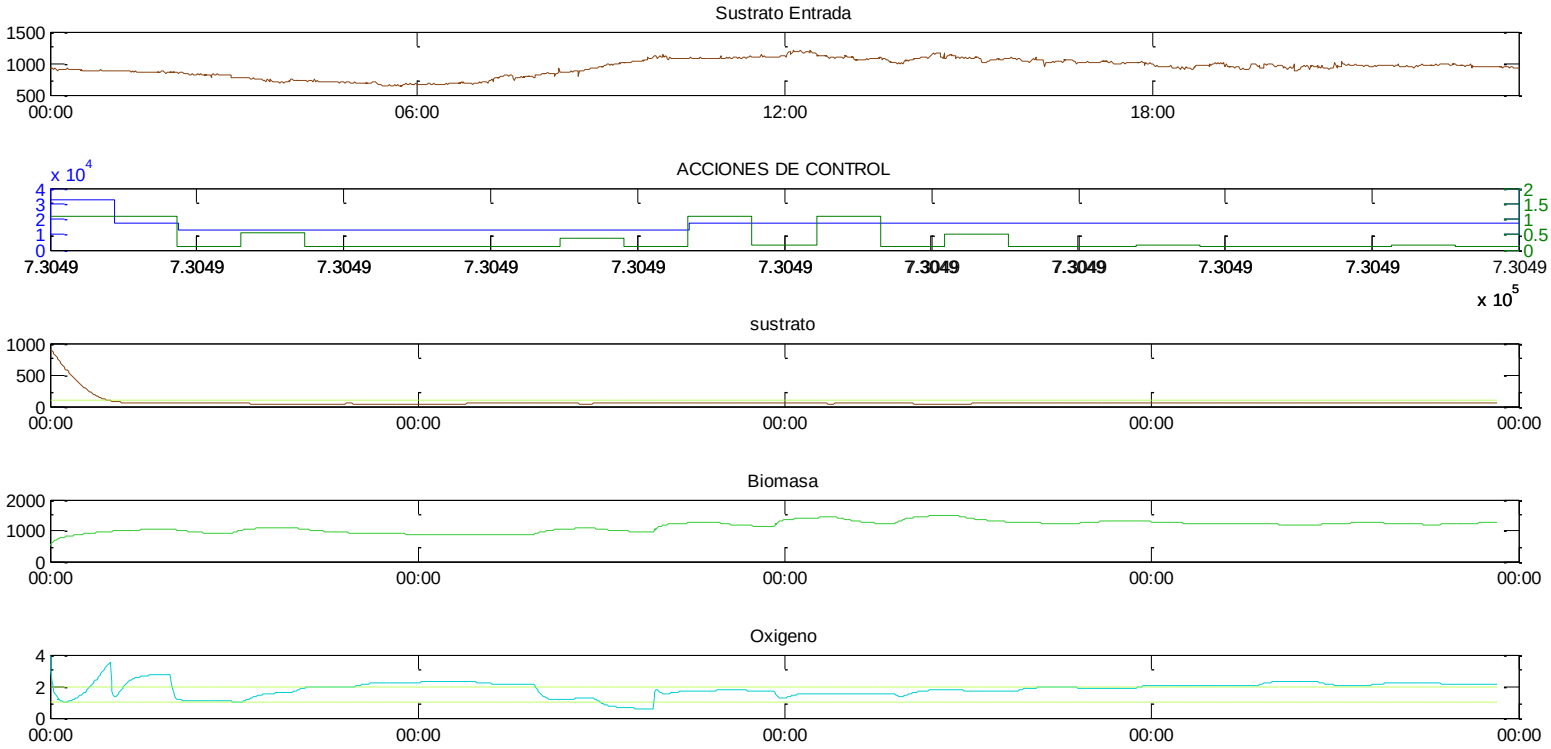


Fuente: (El autor 2012)

Luego de calcular la distribución de probabilidad conjunta del caudal de recirculación y el caudal de aire, se puede observar que el agente toma gran parte de sus acciones en un solo sector, debido a lo armónico en que se encuentra el sustrato de entrada.
Para el siguiente control se introdujo un nuevo valor de recompensa, el cual consiste en penalizar al agente si este se encuentra fuera del rango normal de concentración de oxigeno disuelto (1 a 2 mg/l).

De acuerdo a la nueva restricción hecha al agente se obtuvo la siguiente figura:

Figura 55 Control realizado con modificación de concentración de OD sobre el día típico calculado

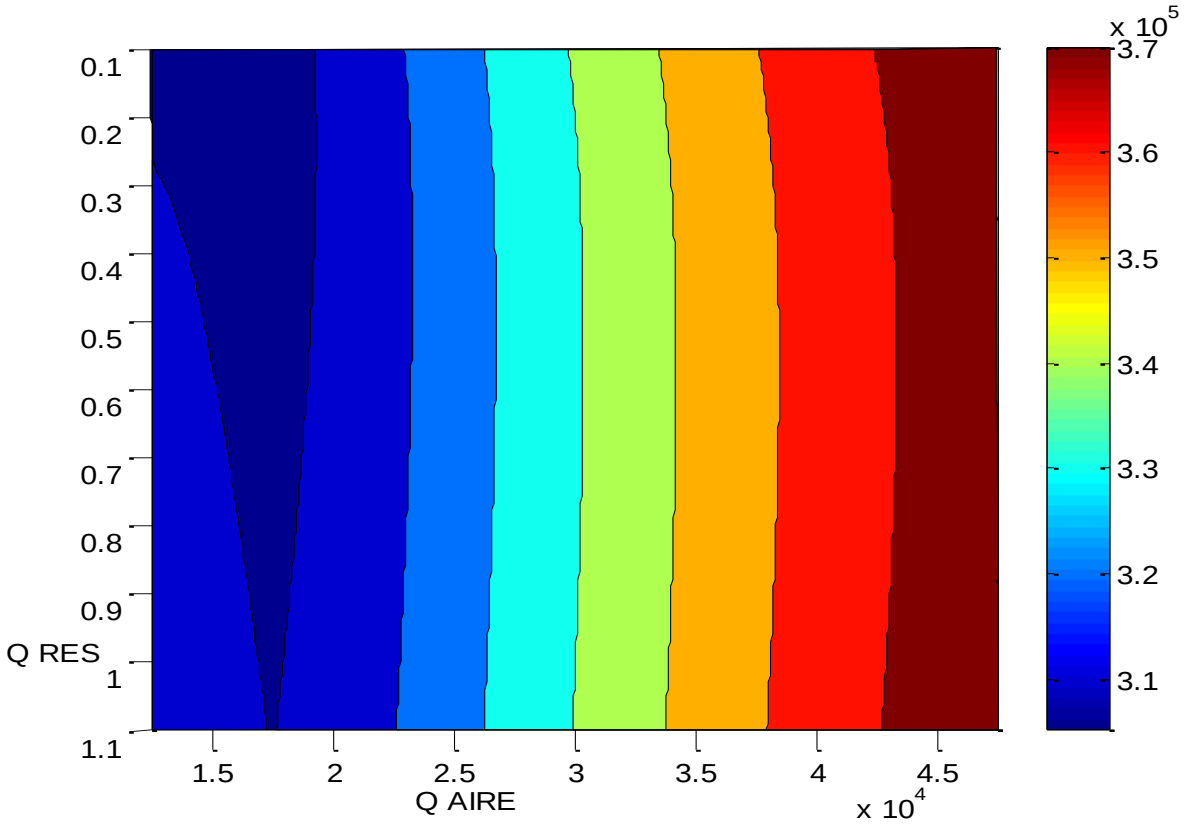


Fuente: (El autor 2012)

Comparando la figura 50 con la 52 se logra observar grandes cambios y similitudes en el comportamiento y en la dinámica del sistema, la primera de ella se da en las acciones de control, como se puede ver el caudal de recirculación de lodos tiende a variar constantemente, llegando a alcanzar tasas de recirculación del 100%, permitiendo esto que el comportamiento de la biomasa no decaiga de manera abrupta como se veía en la figura 50, donde el comportamiento de esta era función del caudal de recirculación. De acuerdo a esto se muestra que este parámetro hace que diferentes factores sean muy sensibles a cualquier cambio de caudal.

Por otra parte, a pesar que el comportamiento del oxígeno es muy parecido entre estos dos ensayos y los caudales de aire encontrados presentan la misma tendencia y los mismo valores, se puede evidenciar que en la figura 51, el agente corrige con mayor rapidez los valores que se encuentran fuera del rango, situación que es llevada a cabo por las recompensas y penalidades introducidas, adicionalmente se muestra una relación directa entre el caudal de aire y el comportamiento del oxígeno, ya que el agente cada vez que percibe que la concentración de oxigeno aumenta el reduce el caudal de aire y si por el contrario el siente una concentración baja aumenta el caudal.

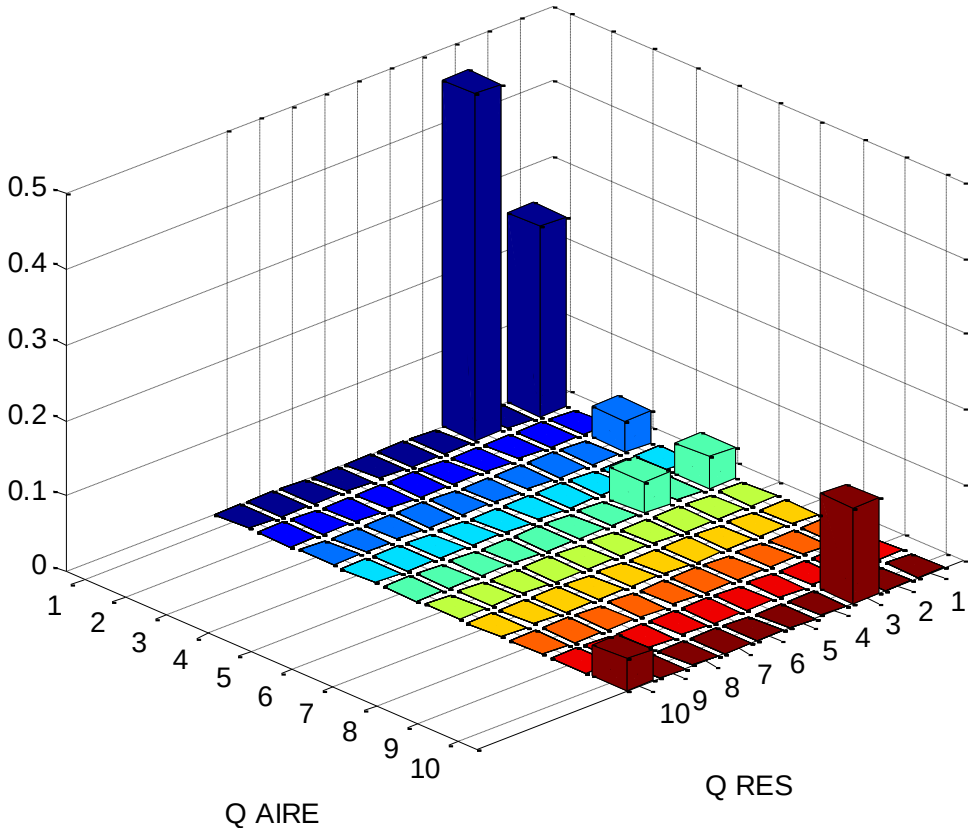
Figura 56 Penalizaciones por acciones de control realizadas.



Fuente: (El autor 2012)

La grafica nos muestra como el agente va recorriendo la totalidad de acciones dentro del sistema de control, sin embargo a pesar de la gran sensibilidad del caudal de recirculación, la superficie tiende a ser muy uniforme, lo que muestra que existe un alto porcentaje de acciones que generan el menor costo (penalizaciones).

Figura 57 Distribución de probabilidad de acciones de control con modificación de OD.

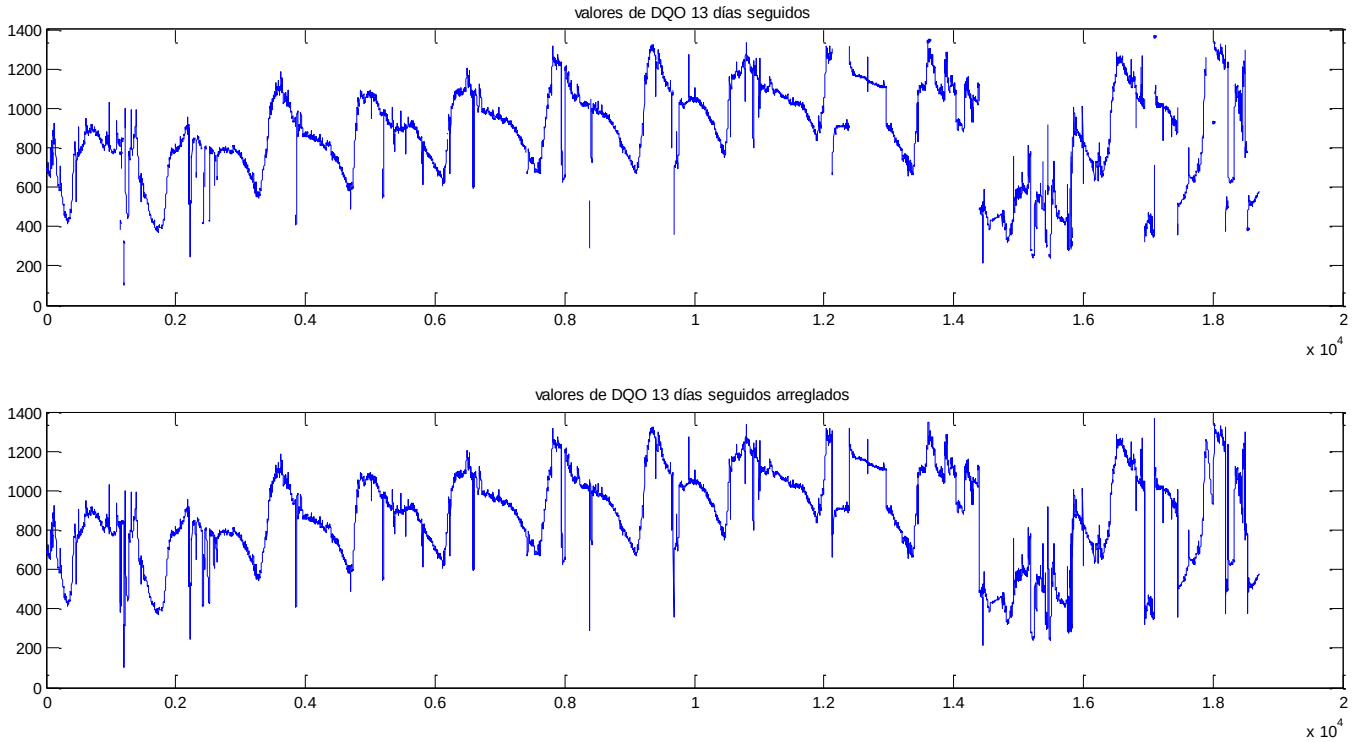


Fuente: (El autor 2012)

Como se pudo observar en la anterior figura, la distribución de probabilidad conjunta creció para las acciones de control más efectivas, quitándole acciones aquellas que se sobrepasaban los valores de OD.

De acuerdo al comportamiento del agente dentro del día típico, se decidió evaluar el funcionamiento del mismo durante los 13 días originales que se tenían en la figura 34, para lo cual primero supuso que estos días eran continuos. De acuerdo al análisis de los datos realizado en el capitulo 4.2.1., se tenía claro que los datos presentaban grandes diferencias entre ellos, por lo tanto se realizó una depuración de estos, para lo cual se calculó la diferencia entre el dato i y el dato $i+1$, encontrándose que la diferencia promedio se encontraba cercana a 300, por consiguiente se generó un programa que ubicara NaN en valores mayores a este promedio, para luego interpolar los NaN y así completar la serie de datos, la cual se muestra a continuación.

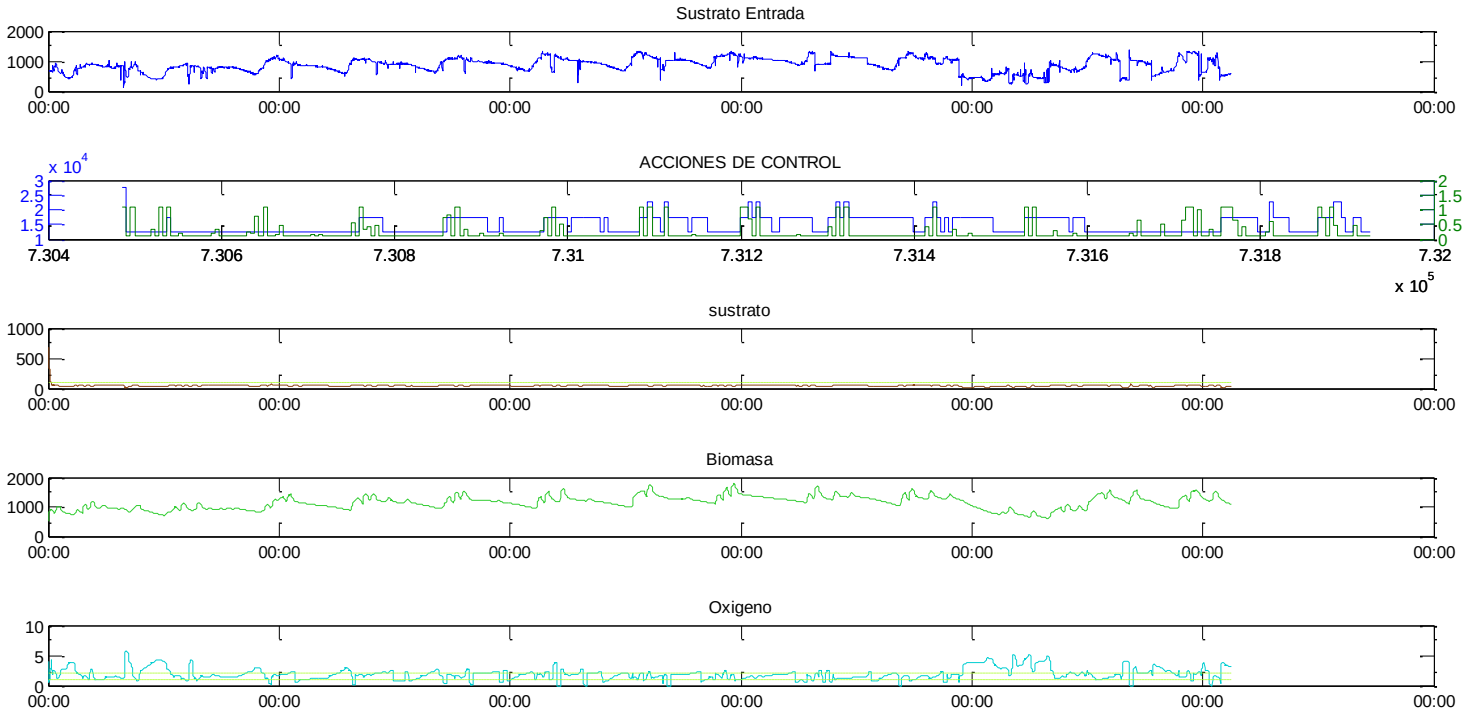
Figura 58 Valores de DQO durante 13 días para comprobación de agente.



Fuente: (El autor 2012)

Ya encontrada la serie de datos, se procedió a la implementación de estos dentro de la modelación de la planta de tratamiento para realizar el control en continuo como se esperaría que funcionase el agente, de lo cual se obtuvo:

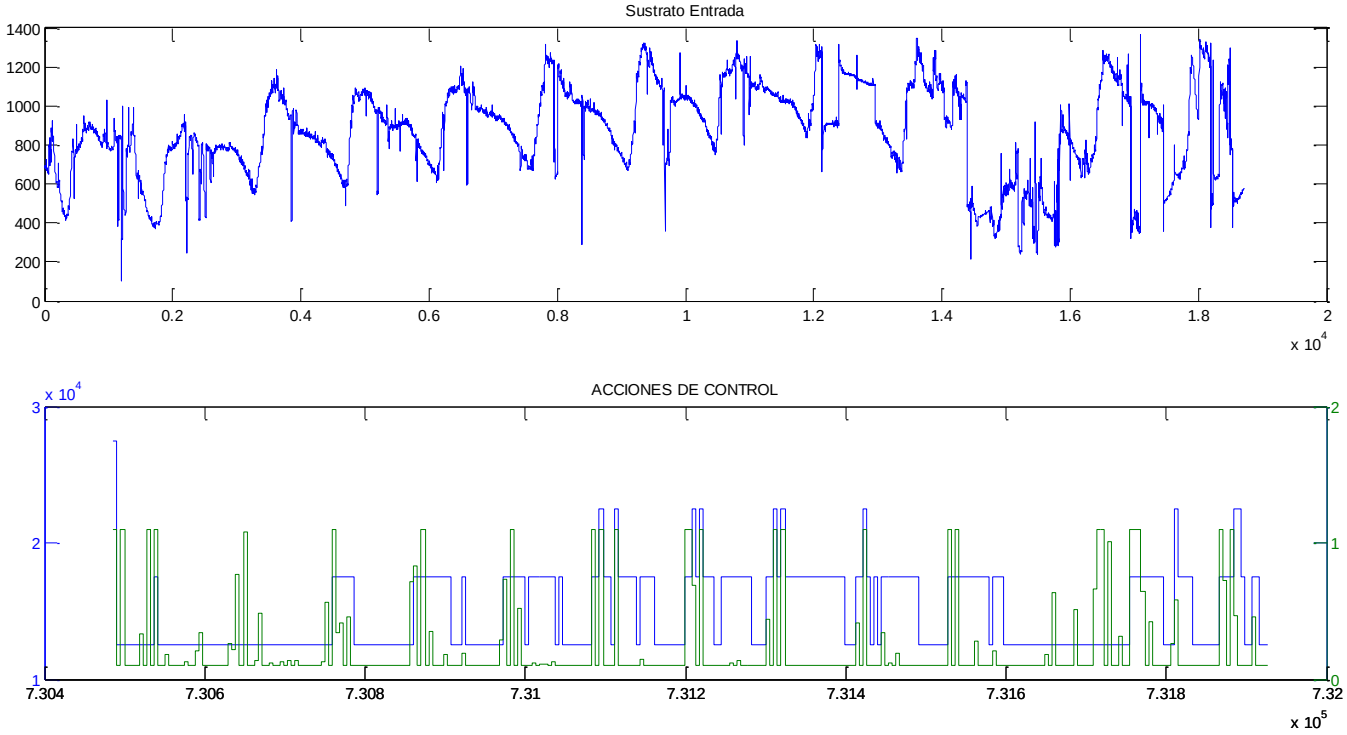
Figura 59 Control realizado sobre los 13 días encontrados.



Fuente: (El autor 2012)

Como se puede observar la el agente realiza las acciones de control sobre la planta, sin embargo esta cumple la política del sustrato pero no cumple en todo momento la condición de concentración de oxígeno, debido a la baja resolución del grafico, se decidió realizar cada uno por separado y verlo más detenidamente.

Figura 60 Acciones de control.



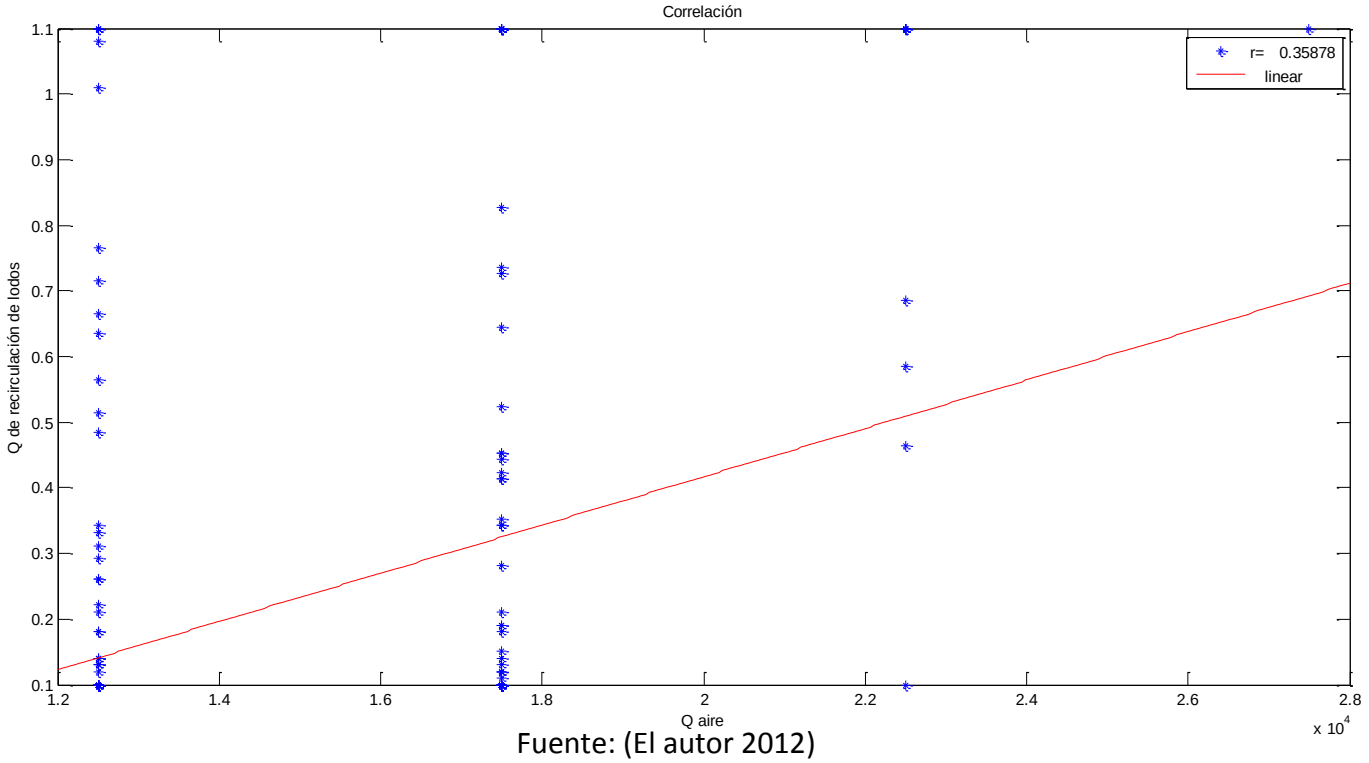
Fuente: (El autor 2012)

De acuerdo a lo observado se encuentra una relación entre el sustrato de entrada y el caudal de oxígeno, el cual aumenta a medida que se incrementa fuertemente el sustrato, cuando el sustrato baja el caudal de oxígeno disminuye pero se mantiene en un nivel mínimo de caudal para cumplir las políticas.

En cuanto al caudal de recirculación de lodos el comportamiento es similar al del caudal de aire, sin embargo este caudal bajo concentraciones muy bajas no mantiene constante el grado de recirculación, si no que lo modifica así sea en deltas de apertura muy pequeños.

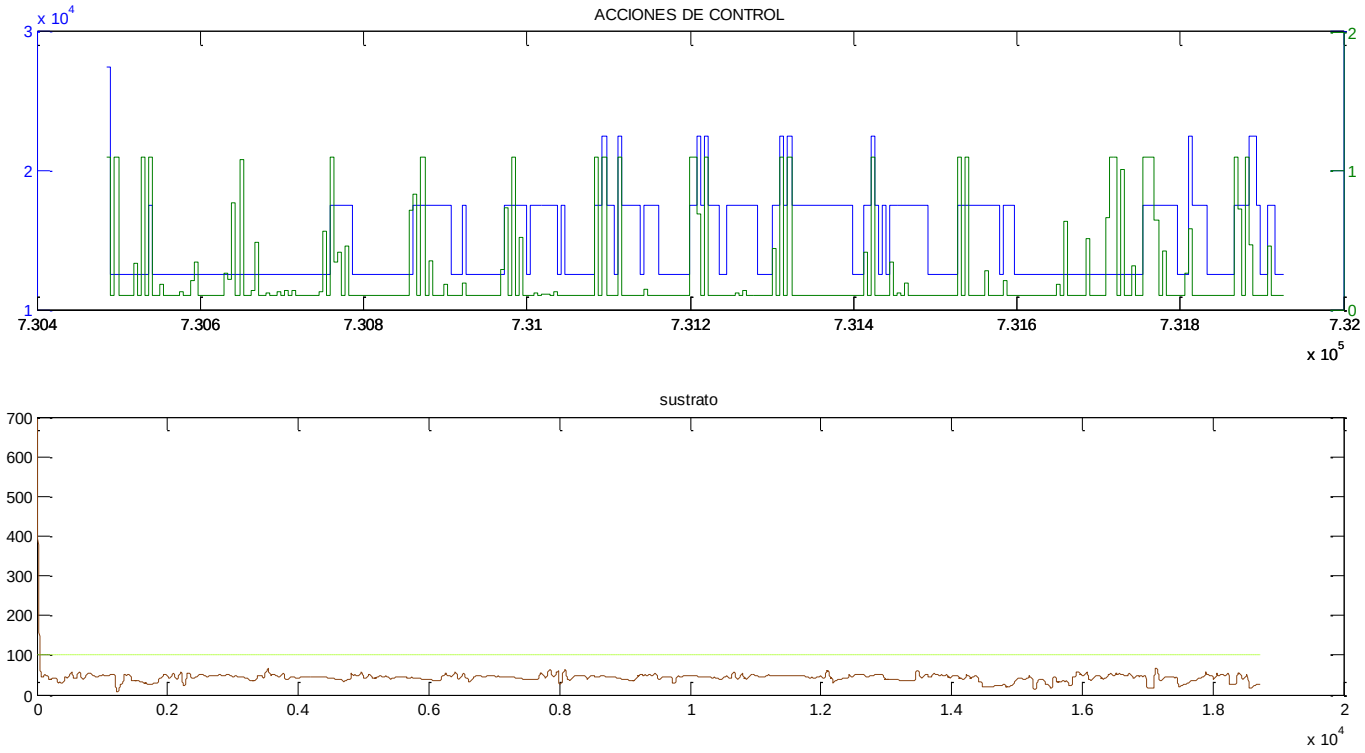
Por otra parte no se evidencia una relación entre las acciones tomadas, ya que el valor de correlación entre estos arrojo un valor de 0.358, lo que muestra que las acciones ejecutadas no afecta directamente una sobre la otra.

Figura 61 Grafica de correlación entre el caudal de aire y el caudal de recirculación.



Fuente: (El autor 2012)

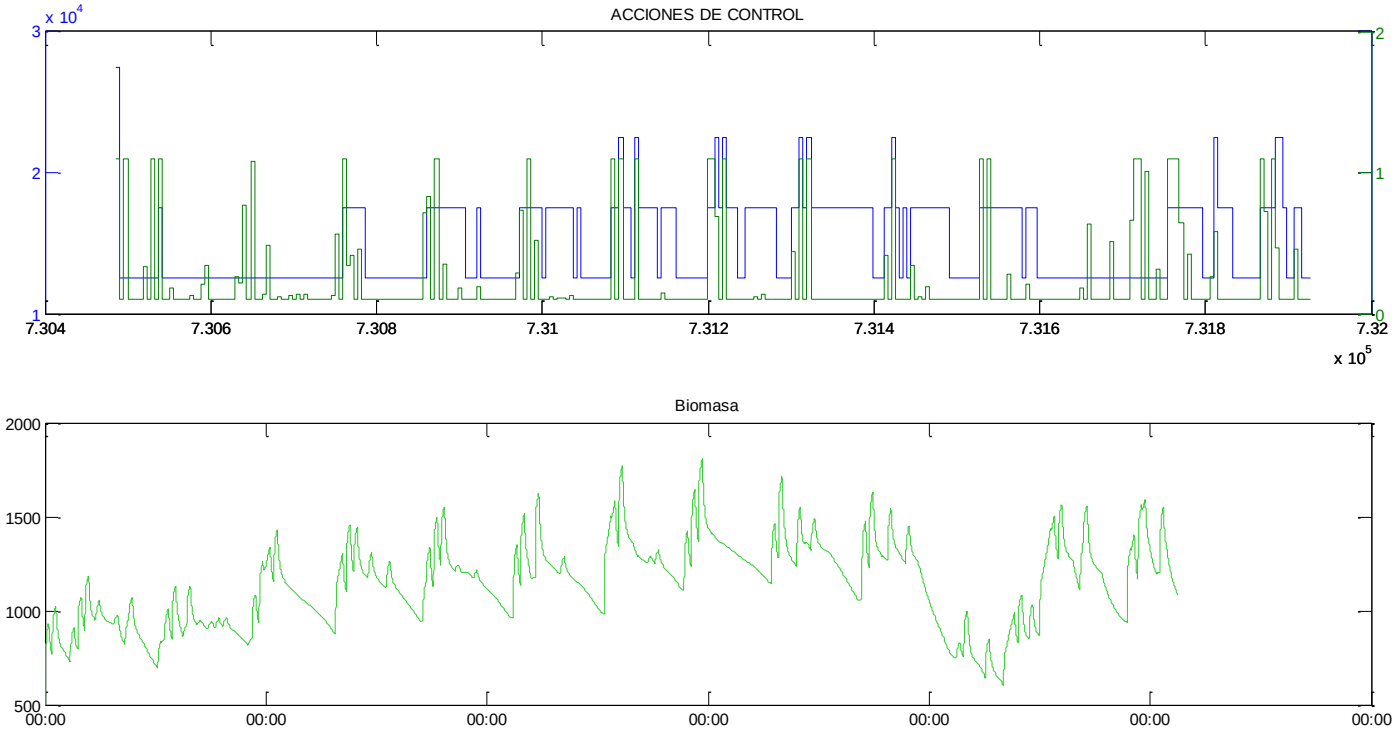
Figura 62 Acciones de control y su cumplimiento de sustrato a la salida.



Fuente: (El autor 2012)

El sustrato a la salida cumple en todo momento con la política deseada por nosotros, lo que indica que el objetivo del sustrato, se logra fácilmente con las acciones de control realizadas por el agente. Sin embargo es importante destacar que el volumen del reactor es de 5000 m^3 y del sedimentador es de 250 m^3 son altos, por lo tanto se verificará si disminuyendo los volúmenes a la mitad esta objetivo seguirá cumpliendo.

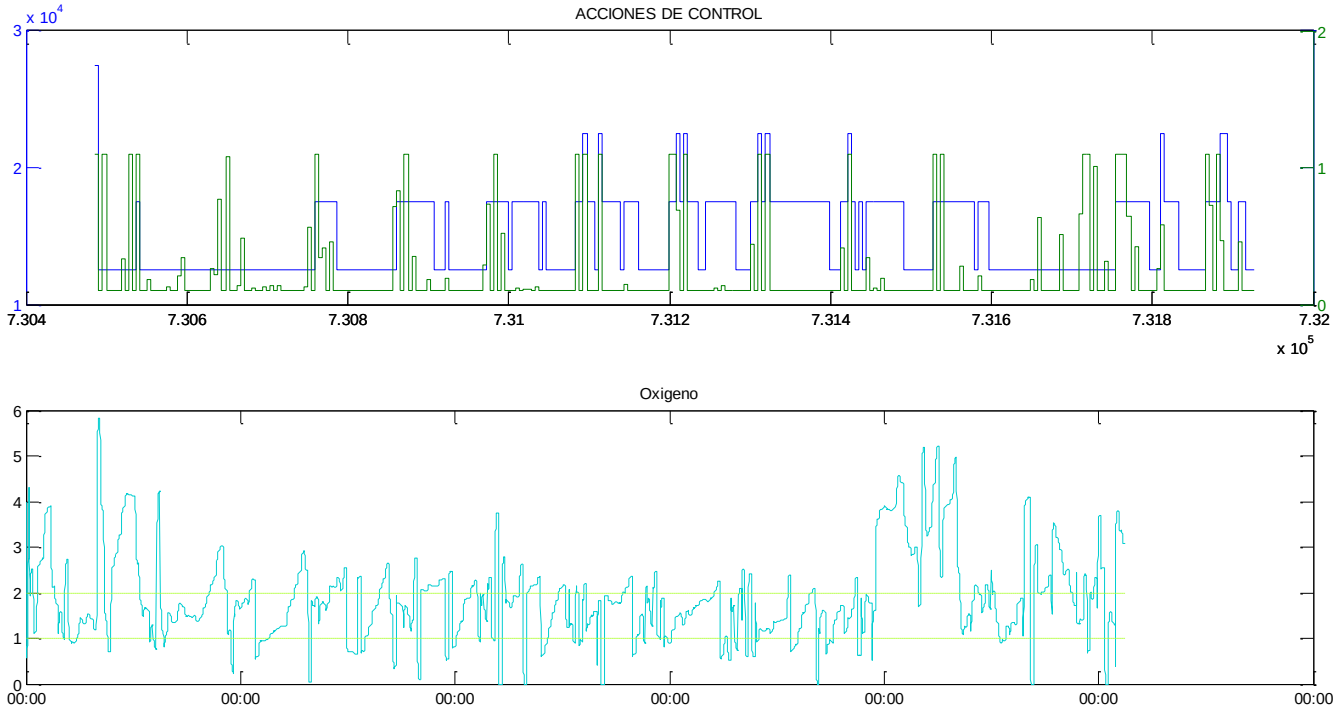
Figura 63 Acciones de control y el comportamiento de la biomasa.



Fuente: (El autor 2012)

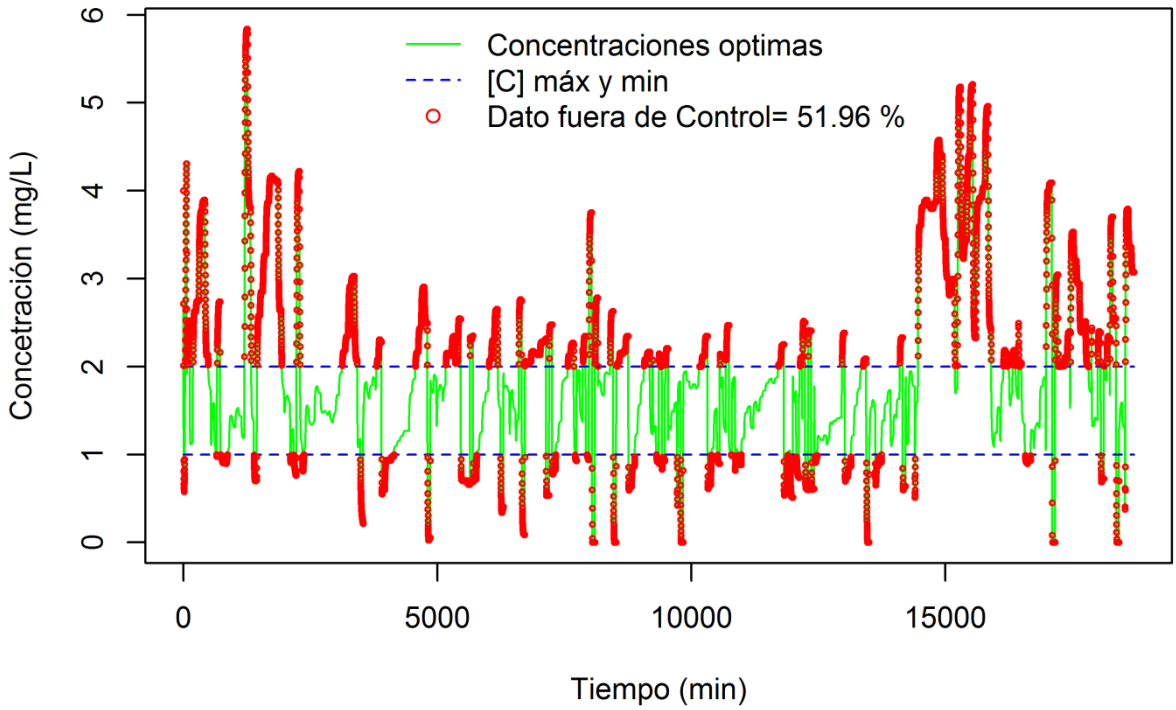
Como se mencionó anteriormente el comportamiento de la biomasa es directamente proporcional al caudal de recirculación de lodos.

Figura 64 Acciones de control y su cumplimiento de concentración de oxígeno disuelto.



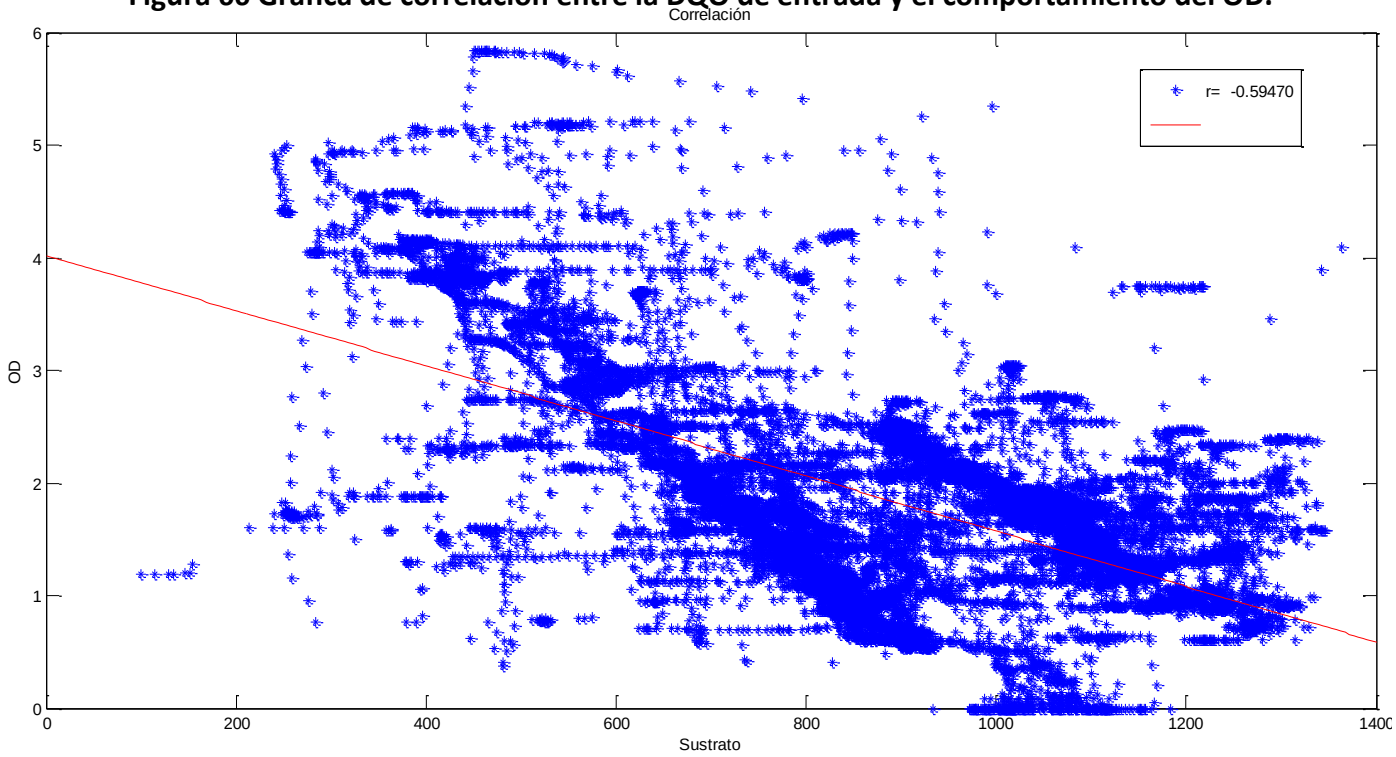
Fuente: (El autor 2012)

Figura 65Concentraciones de OD fuera del rango de control.



Fuente: (El autor 2012)

Figura 66 Grafica de correlación entre la DQO de entrada y el comportamiento del OD.

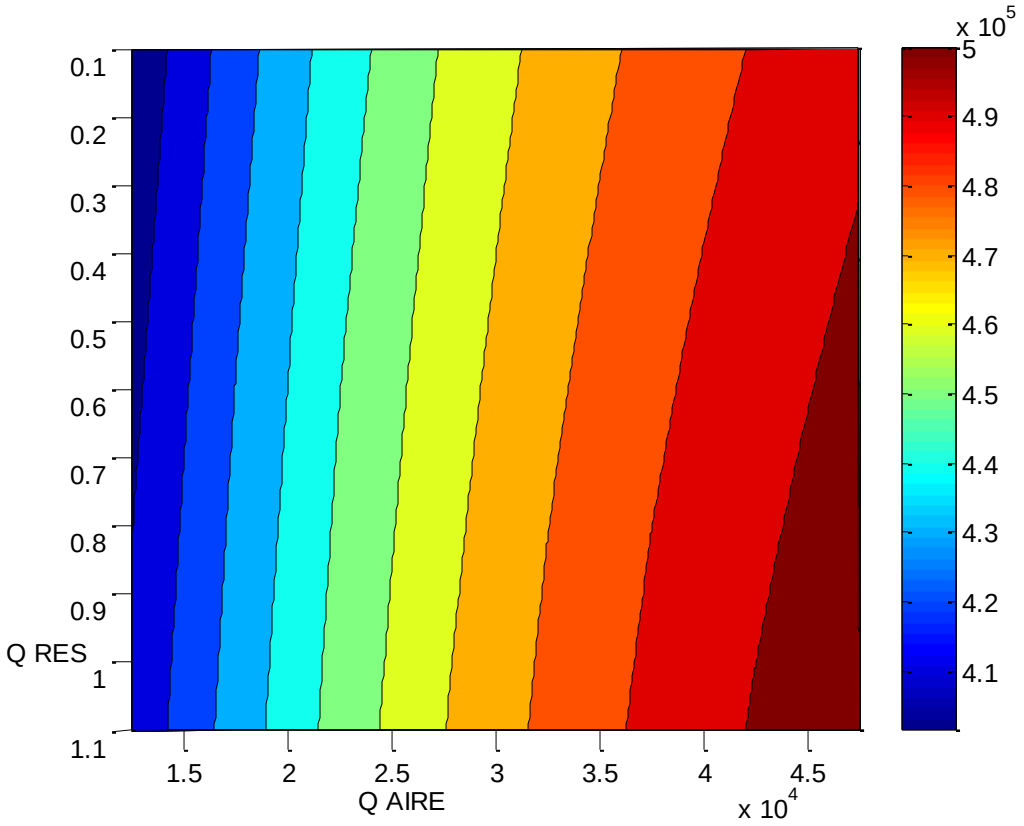


Fuente: (El autor 2012)

Se puede determinar que el comportamiento del oxígeno presenta tendencia a ser inversamente proporcional al comportamiento del sustrato de entrada, el cual al tener variaciones en sus magnitudes tan altas, no permite con facilidad mantener la concentración de oxigeno dentro del rango deseado, situación que se observa al inicio y al final del sustrato, sin embargo dentro del rango de datos de DQO que mantienen una uniformidad, es de resaltar que aun así el agente intenta rápidamente mantenerse dentro del objetivo deseado.

Por otra parte se muestra, que aunque el agente no cumple con la restricción de OD, toma una acción que intente cumplir con el objetivo del sustrato.

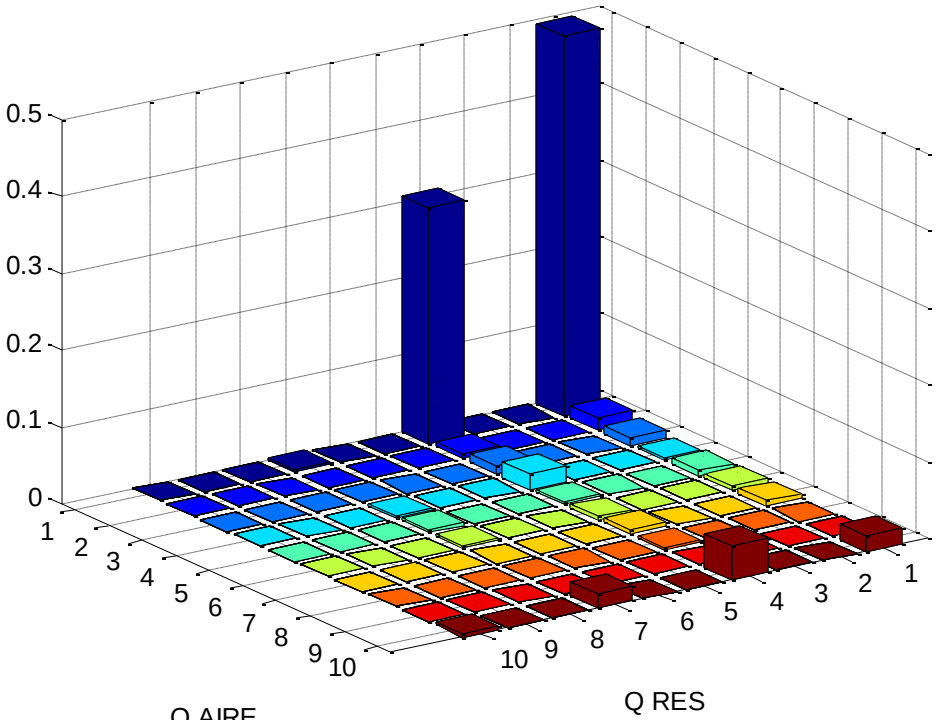
Figura 67 Penalizaciones por acciones de control realizadas durante los 13 días.



Fuente: (El autor 2012)

La situación anteriormente descrita se logra ver con claridad en la anterior figura, ya que las acciones que generan una baja penalización tienen un rango muy bajo, como lo confirma la distribución de probabilidad.

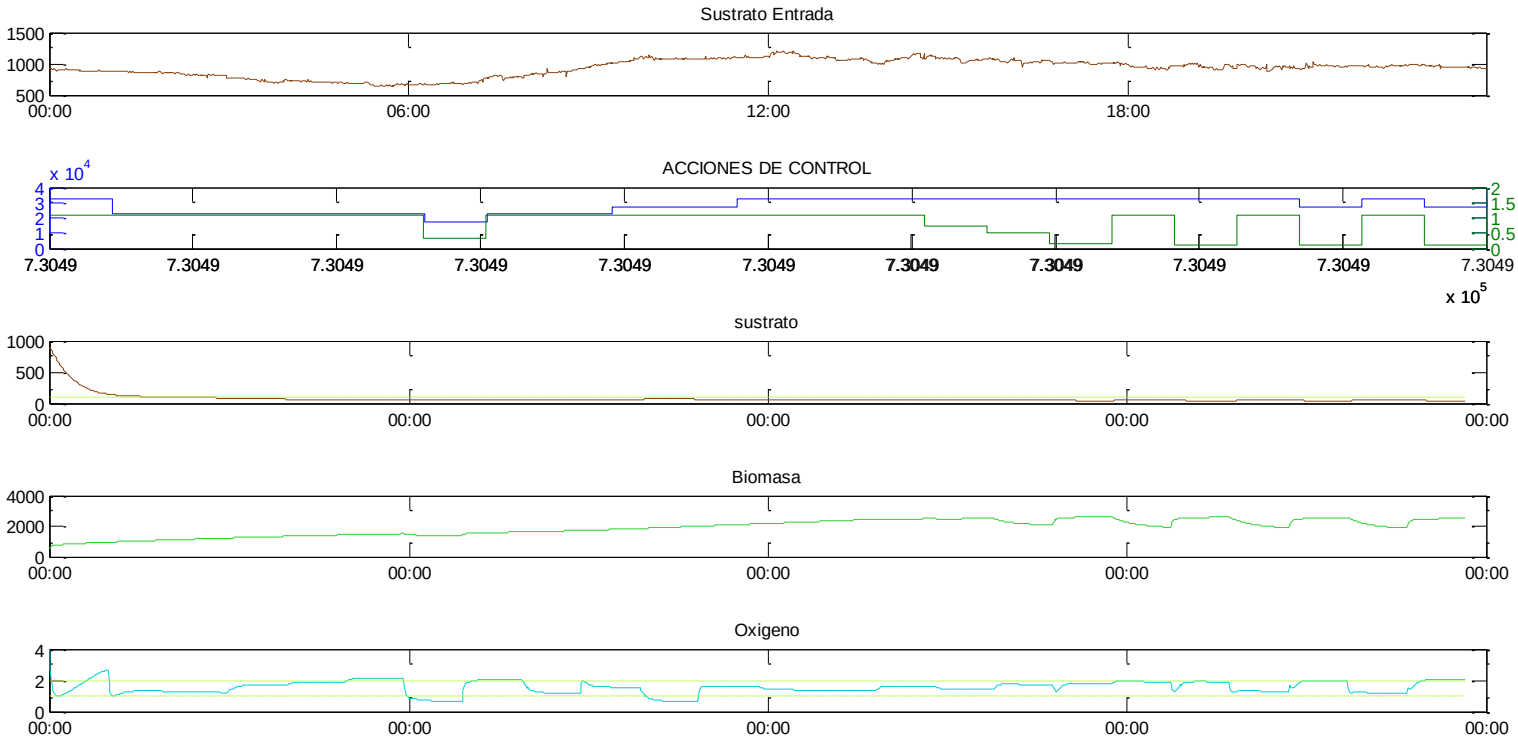
Figura 68 Distribución de probabilidad de acciones de control de los 13 días.



Fuente: (El autor 2012)

Para la verificación del funcionamiento del agente bajo diferentes escenarios, se planteó verificar que tanto afectaba las funciones del agente al presentarse un cambio de volumen al 50% en los tanques.

Figura 69 Control realizado sobre el día típico encontrado con el volumen al 50%.

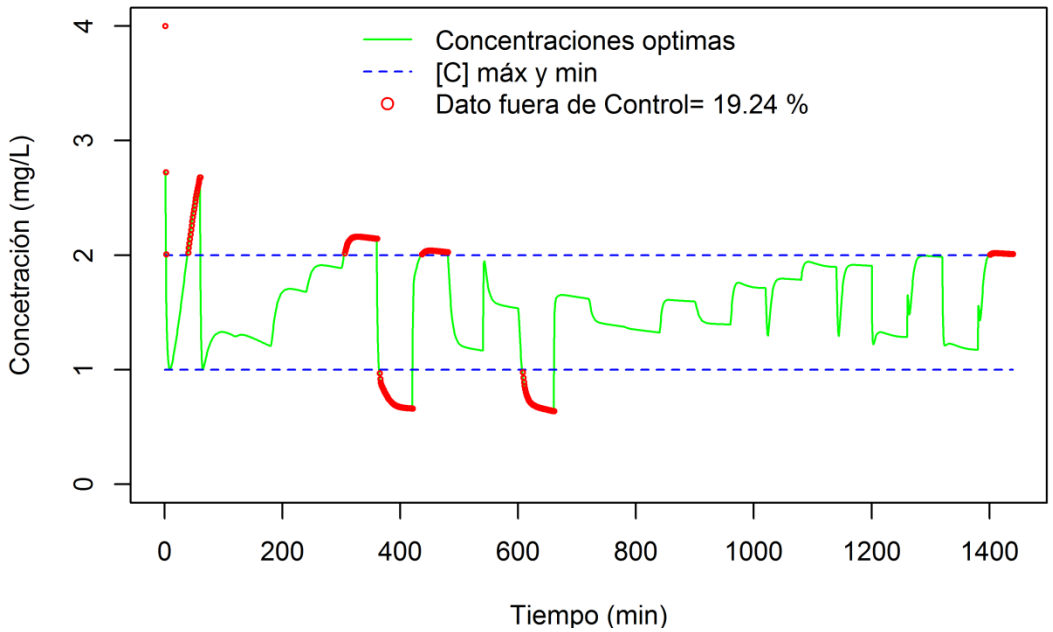


Fuente: (El autor 2012)

Como se puede observar en el día típico, al reducir los volúmenes de los tanques, se obtuvo que el caudal de aire variara mucho más en intervalos cortos de tiempo, con respecto a los valores reportados en el escenario donde los tanques se encuentran en el tamaño original, esto se debe gracias al alto crecimiento de la biomasa. Por otra parte el caudal de recirculación cuenta con una tasa del 100% en periodos largos de tiempo, lo que relaciona al comportamiento de la biomasa.

En cuanto al oxigeno disuelto su comportamiento dentro de la franja deseada es mayor, encontrándose que el 19.4% de las concentraciones estuvieron por fuera del rango, atribuido a los cambios constantes de caudal de oxígeno que hace que la condición objetivo se cumpla también rápidamente.

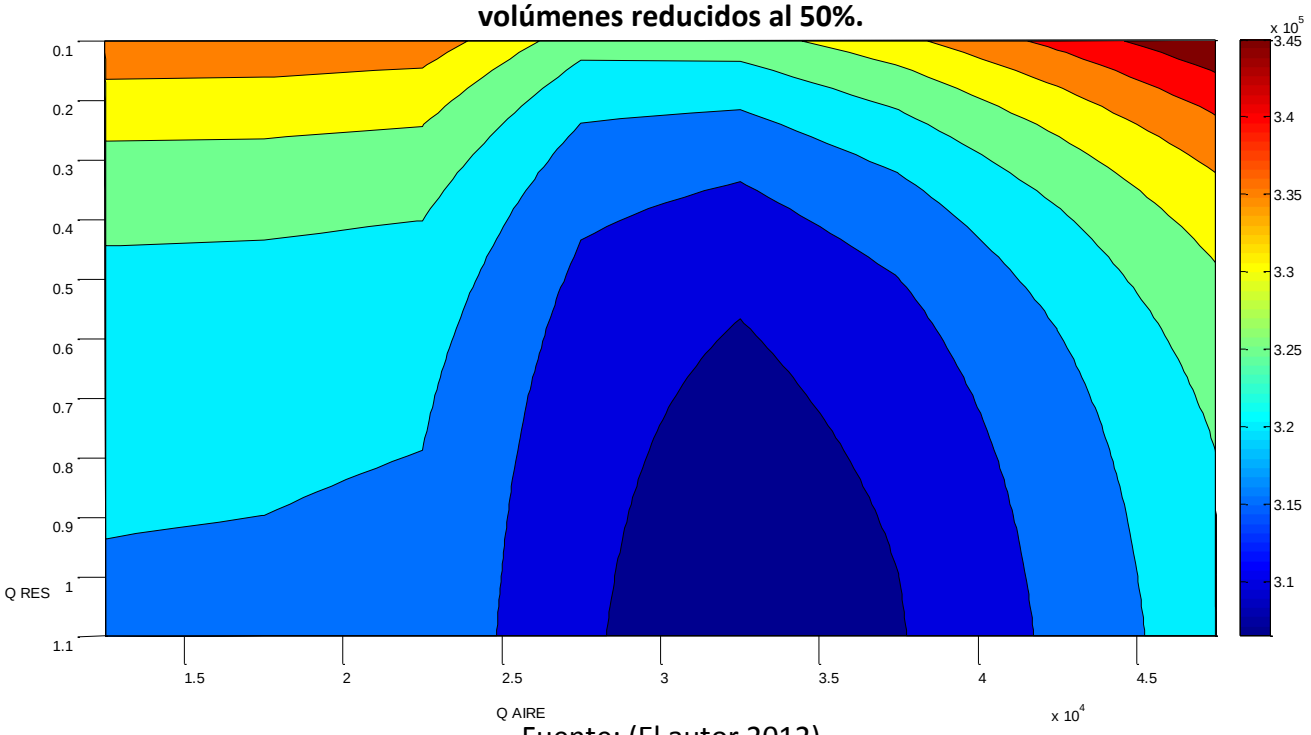
Figura 70 Concentraciones de OD fuera del rango de control.



Fuente: (El autor 2012)

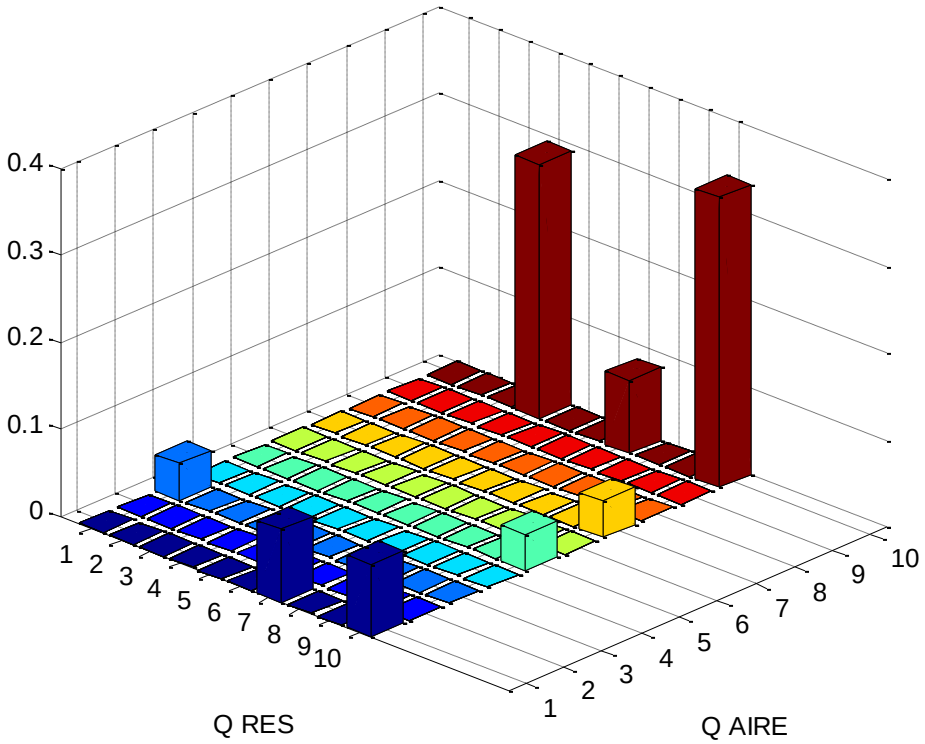
Uno de los mayores cambios se presenta en las penalizaciones, ya que estas tienen al ampliar tanto el caudal de aire estas se hacen menor, lo que nos muestra que para este ejemplo el caudal de aire tiene un mayor peso en las acciones de control.

Figura 71 Penalizaciones por acciones de control realizadas durante el día típico con volúmenes reducidos al 50%.



Fuente: (El autor 2012)

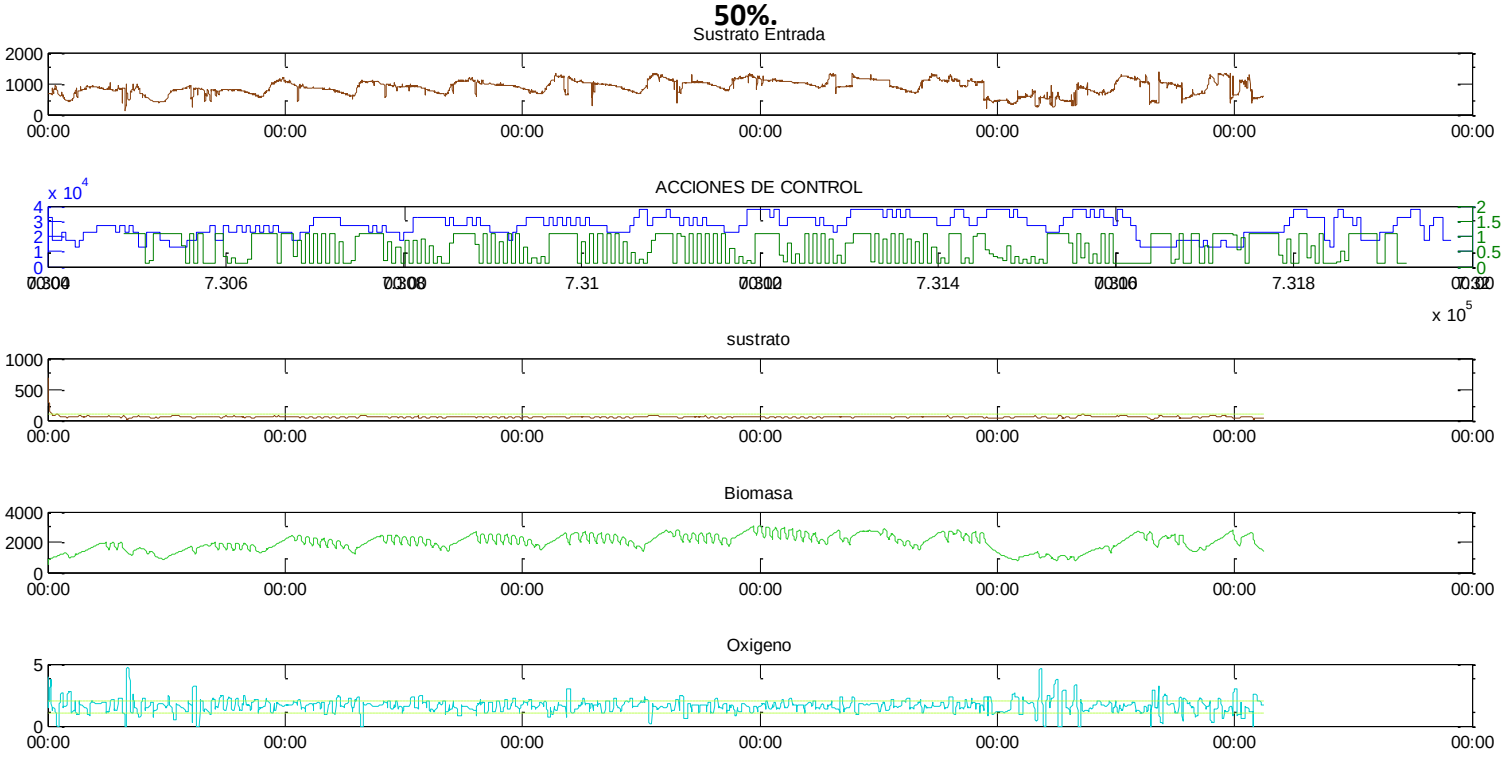
Figura 72 Distribución de probabilidad de acciones de control del día típico con volúmenes reducidos al 50%.



Fuente: (El autor 2012)

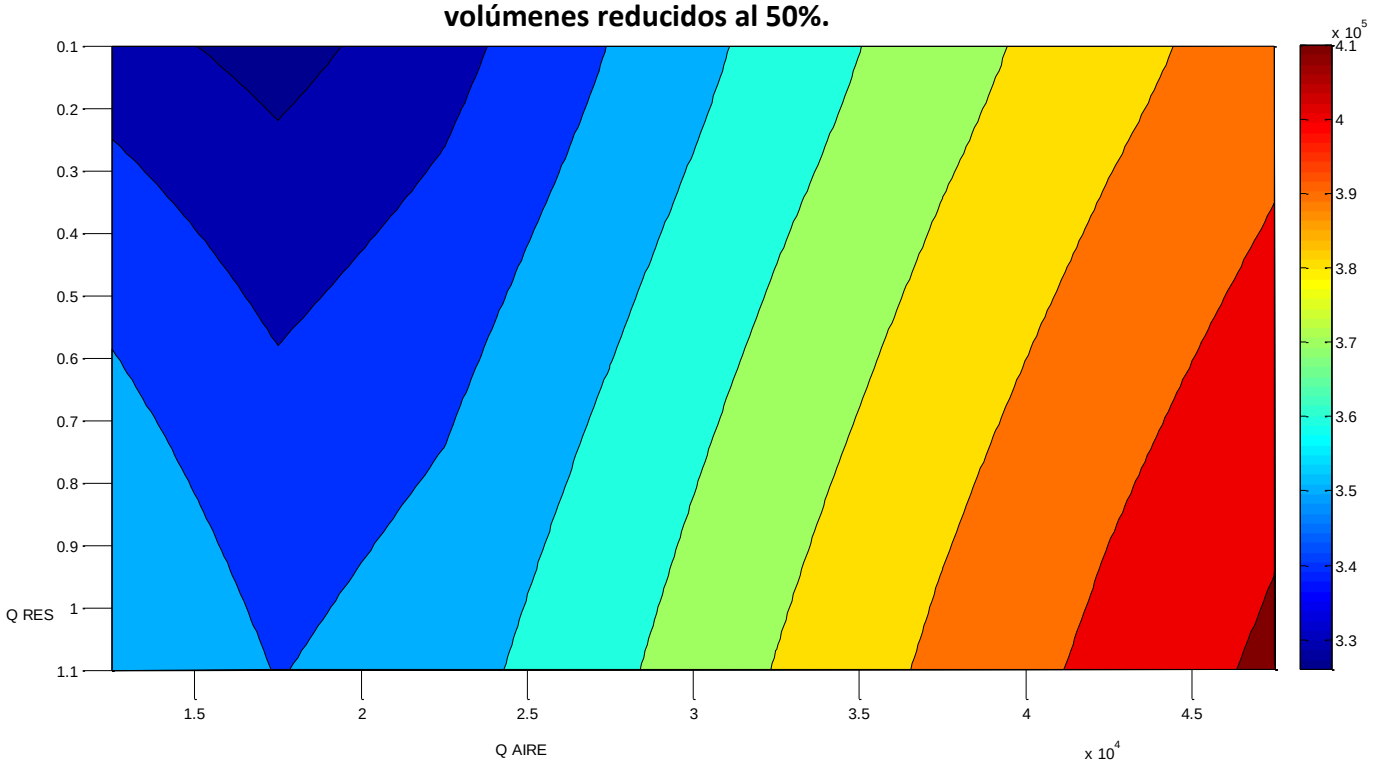
Por último se realizó el mismo escenario con la totalidad de días, del cual se encuentra un mayor cumplimiento en la franja de concentraciones de OD y muchas acciones de control en los caudales de oxígeno como en los de recirculación de lodos, situación que lleva a identificar como bajo diferentes condiciones los dos controles tienen mayor sensibilidad, presentado una mayor influencia en el control las acciones llevadas sobre caudal de recirculación de lodos.

Figura 73 Control realizado sobre los 13 días con volúmenes en los tanques reducidos al 50%.



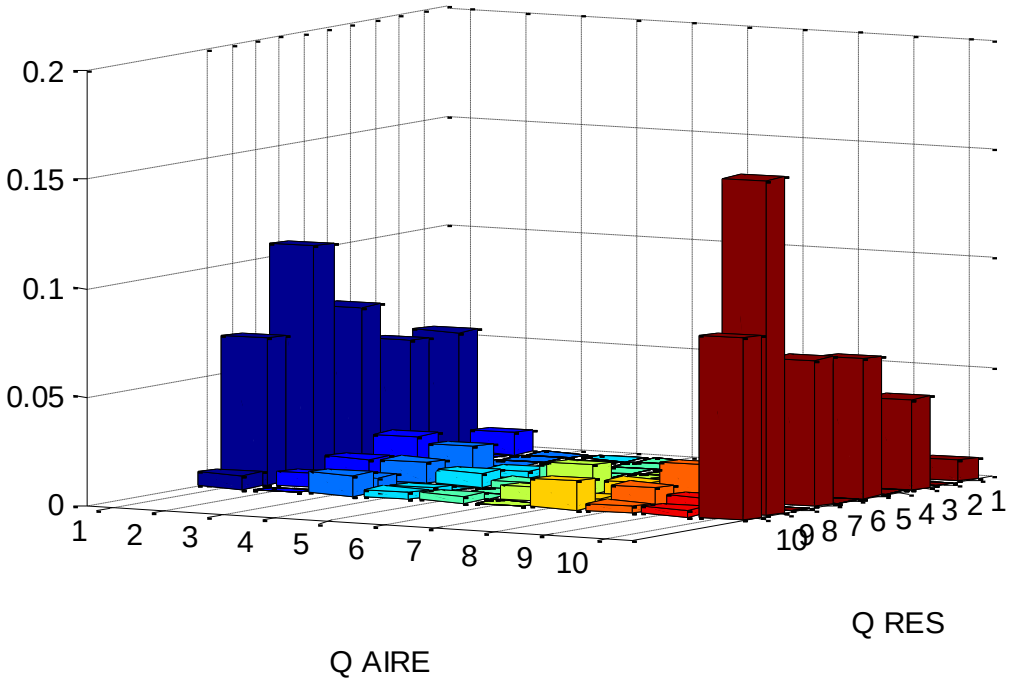
Fuente: (El autor 2012)

Figura 74 Penalizaciones por acciones de control realizadas durante 13 días con volúmenes reducidos al 50%.



Fuente: (El autor 2012)

Figura 75 Distribución de probabilidad de acciones de control de los 13 días con volúmenes en los tanques reducidos al 50%.

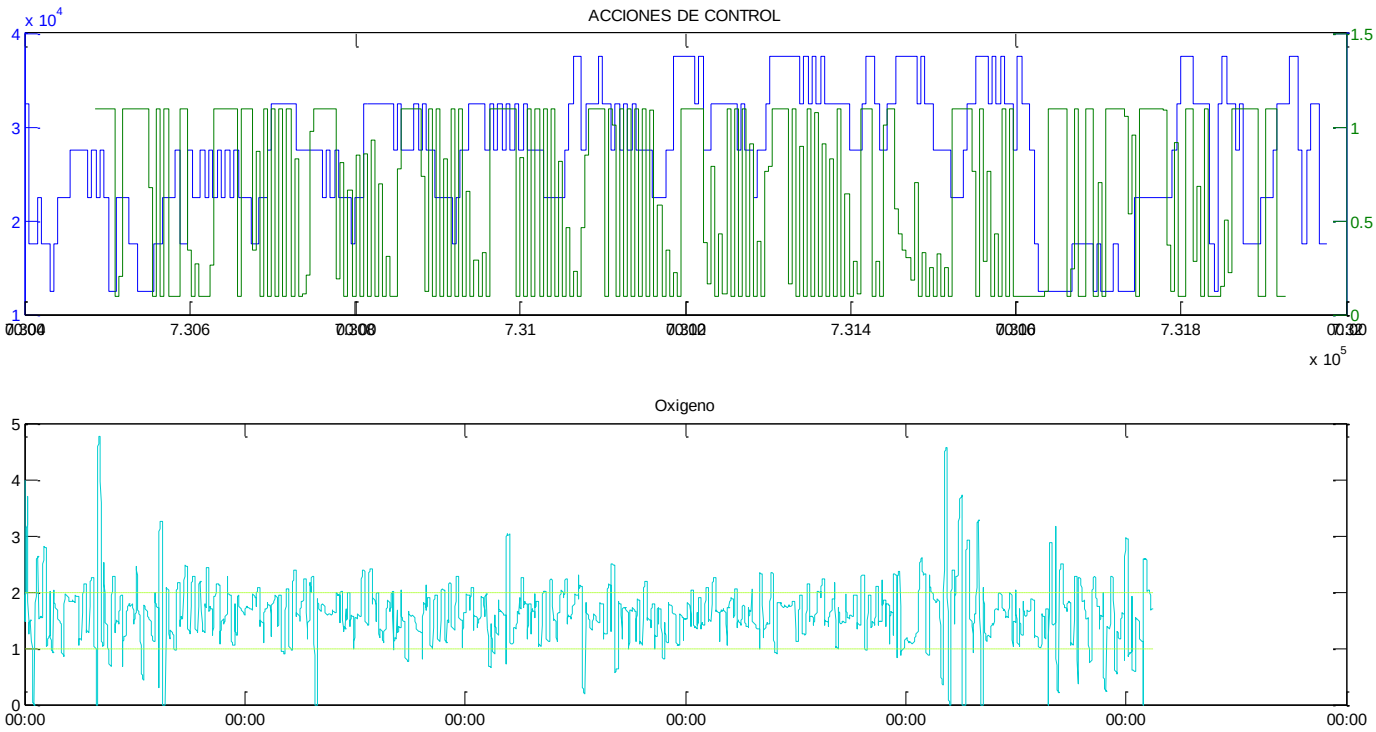


Fuente: (El autor 2012)

Las anteriores dos figuras muestran un comportamiento bastante interesante, ya que a pesar que los valores de tomar acciones que penalicen al agente son mayores que los que no, este realizó repetidamente las acciones que si se encuentran dentro del menor costo, situación que llevo al excelente comportamiento del agente en este escenario.

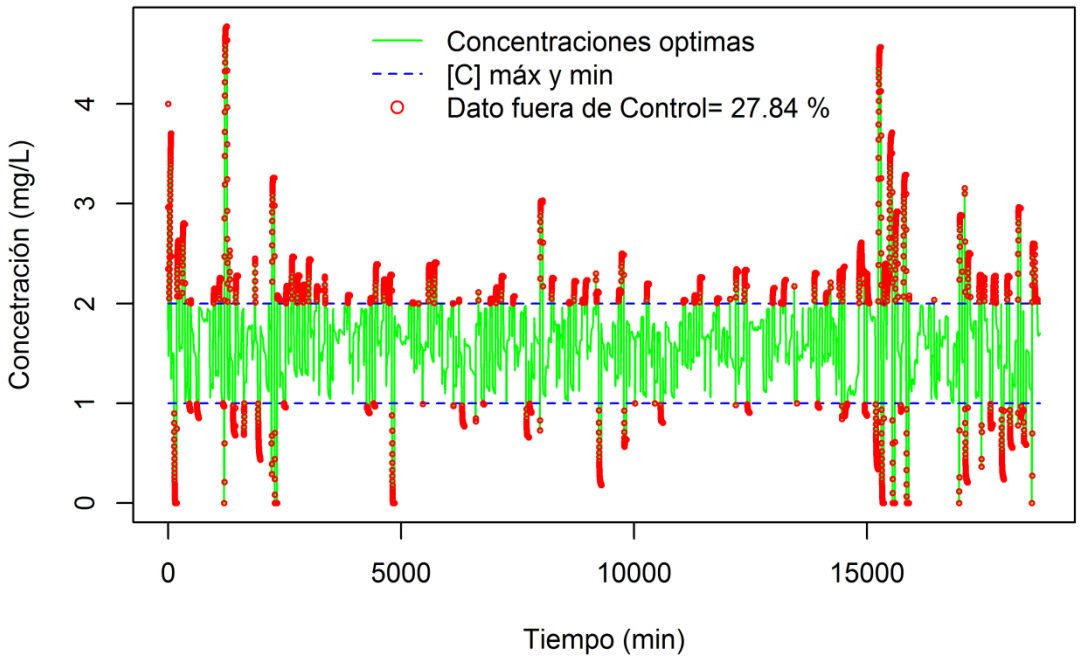
Al igual que en el anterior caso, el comportamiento del oxigeno disuelto fue muy bueno, reflejado por la acción continua del controlador de caudal de aire, donde solo el 27. 84% (3744 valores de 18720) como se puede ver a continuación:

Figura 76 Acciones de control y su cumplimiento de concentración de oxígeno disuelto.



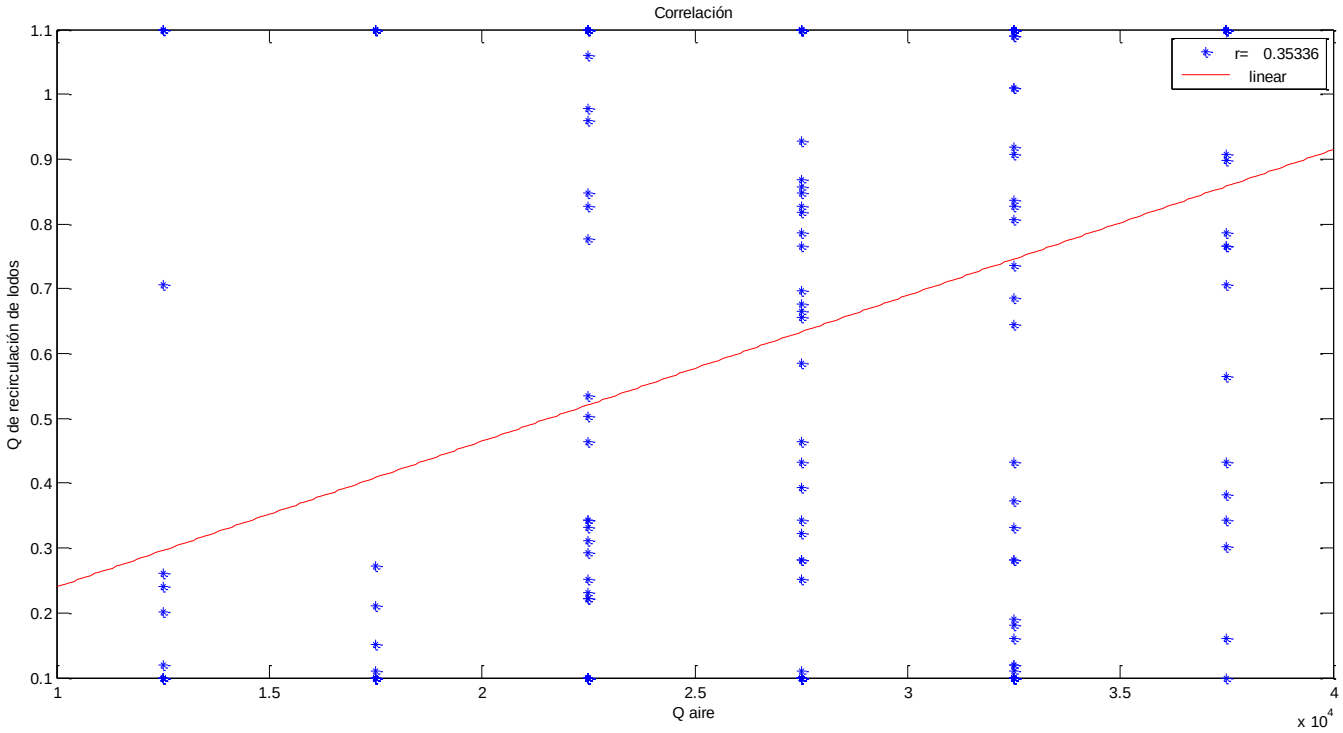
Fuente: (El autor 2012)

Figura 77 Concentraciones de OD fuera del rango de control.



Fuente: (El autor 2012)

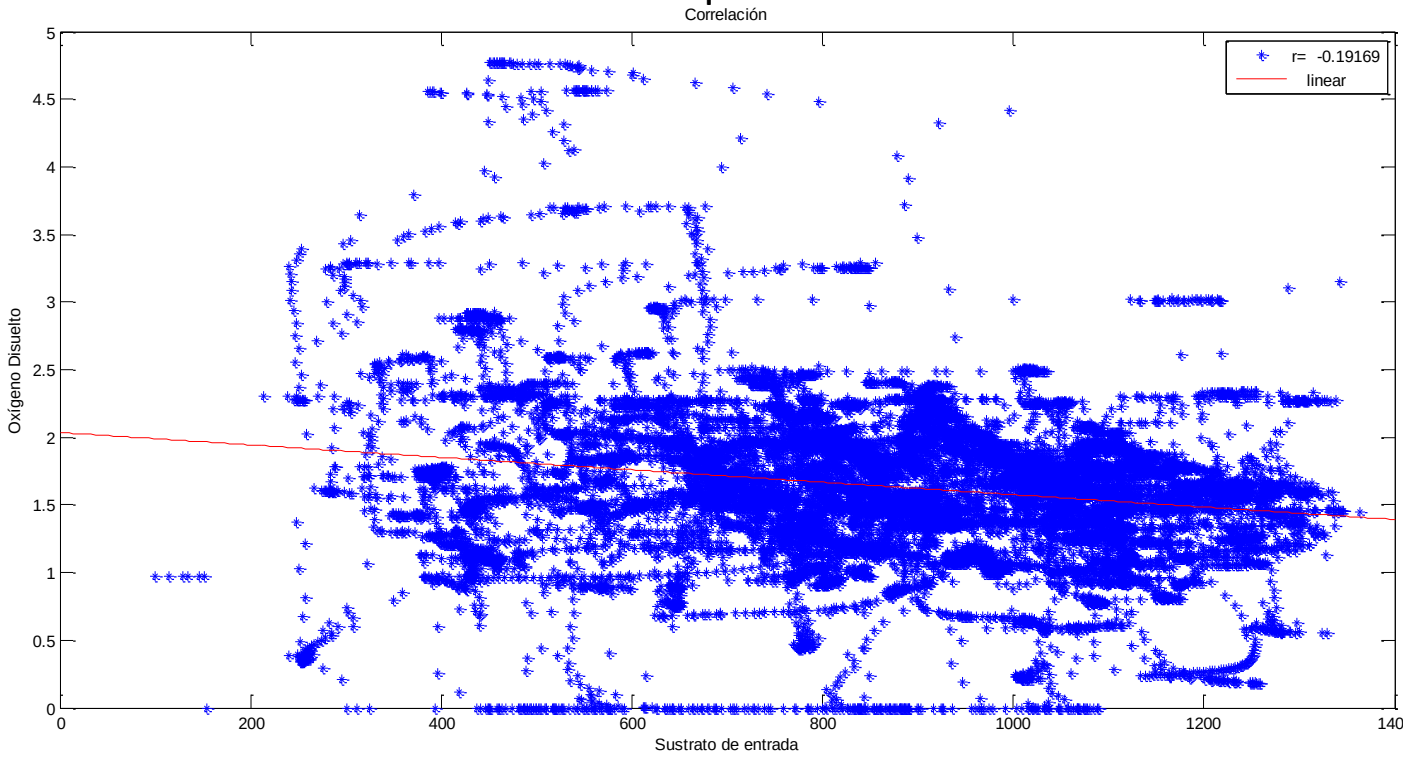
Figura 78 Grafica de correlación entre el caudal de aire y el caudal de recirculación al reducir los tanques al 50%.



Fuente: (El autor 2012)

Como se puede ver, la correlación entre las acciones de control al reducir los tanques, tiene la misma tendencia que en el escenario con el volumen original de los tanques, situación que nos muestra que a pesar de la mayor cantidad de acciones en el caudal de oxígeno hechas para este escenario (con reducción de volúmenes en los tanques), esta acción no afecta directamente a la otra acción de control mostrando clara independencia entre una y otra. Por otra parte puede verse en las dos gráficas de correlación de acciones de control puede haber dos acciones para un mismo sustrato de entrada, mostrando un comportamiento de histéresis, esto se da, ya que las recompensas pueden variar según las acciones de control previas.

Figura 79 Grafica de correlación entre la DQO de entrada y el comportamiento del OD al reducir los tanques al 50%.



Fuente: (El autor 2012)

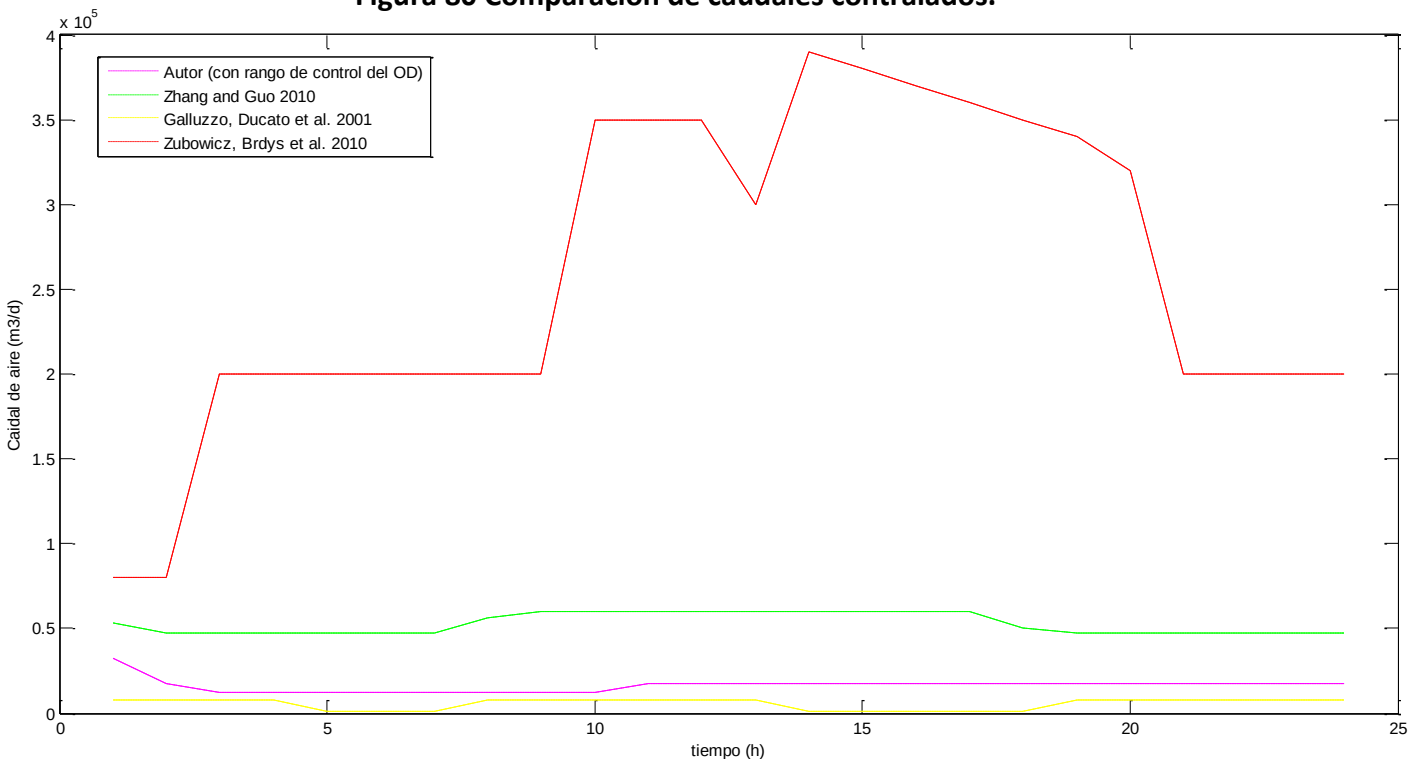
Para verificar la eficiencia del agente, se ha planteado hacer una comparación de diferentes resultados presentados en la literatura, es importante mencionar que directamente no se podría comparar caudal con caudal, ya que las dimensiones de las plantas presentan grandes diferencias y las concentraciones de DQO no alcanzan los máximos valores presentados dentro de esta tesis.

Tabla 10 Valores de Caudal de aire.

Autor	Caudal de Aire	Planta
Fiter, Güell et al. 2005	6125 (m^3/d)	Taradell WWTP California volumen reactor 1320 m^3
Zhang and Guo 2010	53000 – 60000 (10^2 m^3/d)	No se presenta
Galluzzo, Ducato et al. 2001	1200 – 7920 (m^3/d)	No se presenta
Kalker, Van Goor et al. 1999	9358 (m^3/d)	No se presenta
Piotrowski, Brdys et al. 2008	12000 – 18000 (m^3/d)	WWTP in Kartuzy, Polonia
Zubowicz, Brdys et al. 2010	80000 – 350000 (10^4 m^3/d)	Volumen 4470 m^3
Autor 2012 (sin rango de control del OD)	2300 – 8000 (m^3/d)	Volumen del tanque 7000 m^3
Autor 2012 (con rango de control del OD)	32500 – 17500 (m^3/d)	Volumen del tanque 7000 m^3

Fuente: (El autor 2012)

Figura 80 Comparación de caudales contralados.

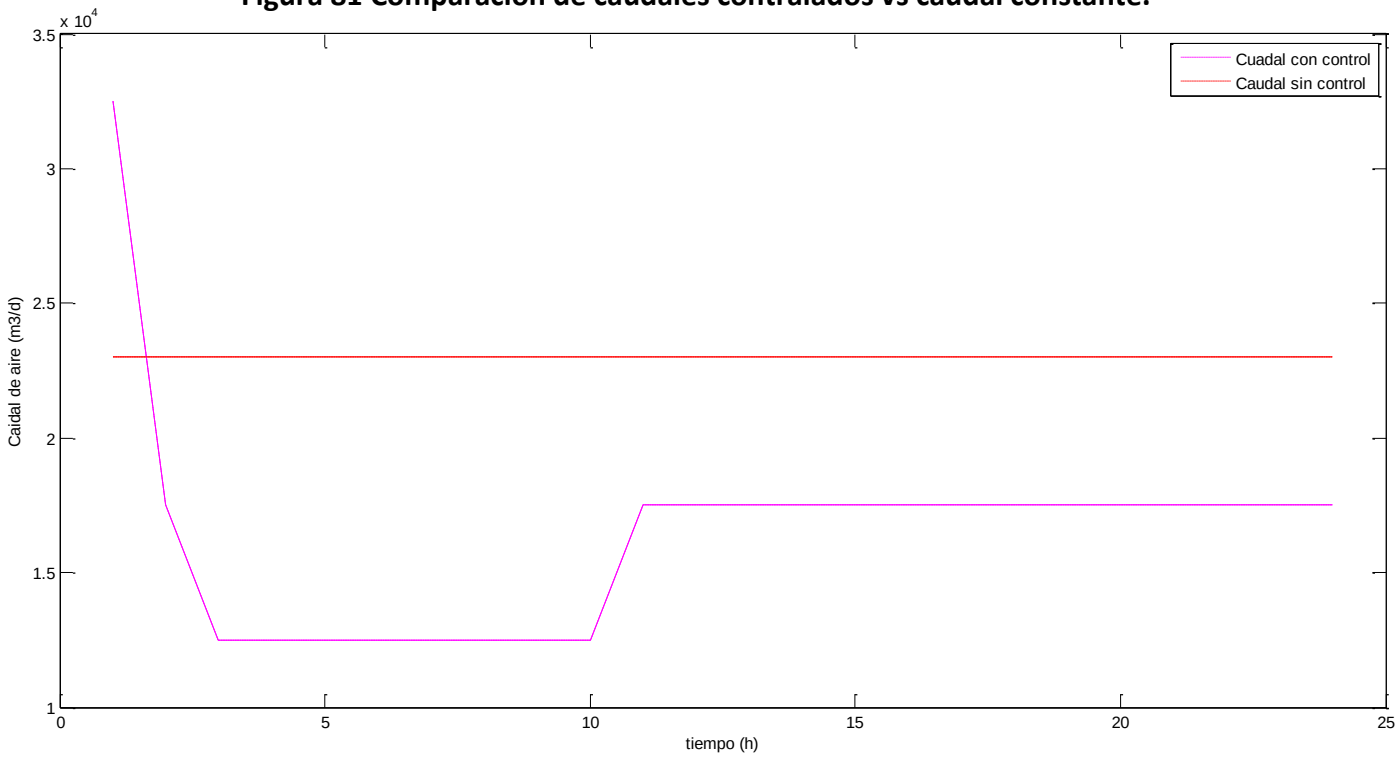


Fuente: (El autor 2012)

Al comparar el comportamiento valores de las referencias encontrados con los valores de caudal calculados por el agente, se observa que los valores del caudal de aire encontrados por el agente son bajos a pesar del tamaño del reactor y estos están dentro de los rangos encontrados en la literatura.

Por último se quiere encontrar la reducción de caudal de aire con la implementación del agente al sistema, por lo tanto se modeló la planta ingresándole el caudal del día típico, con un caudal de aire constante (como se puedo observar en las figuras 44 a 47), donde el caudal de aire mínimo para que el reactor no sea anaeróbico es de 23000 m^3/d.

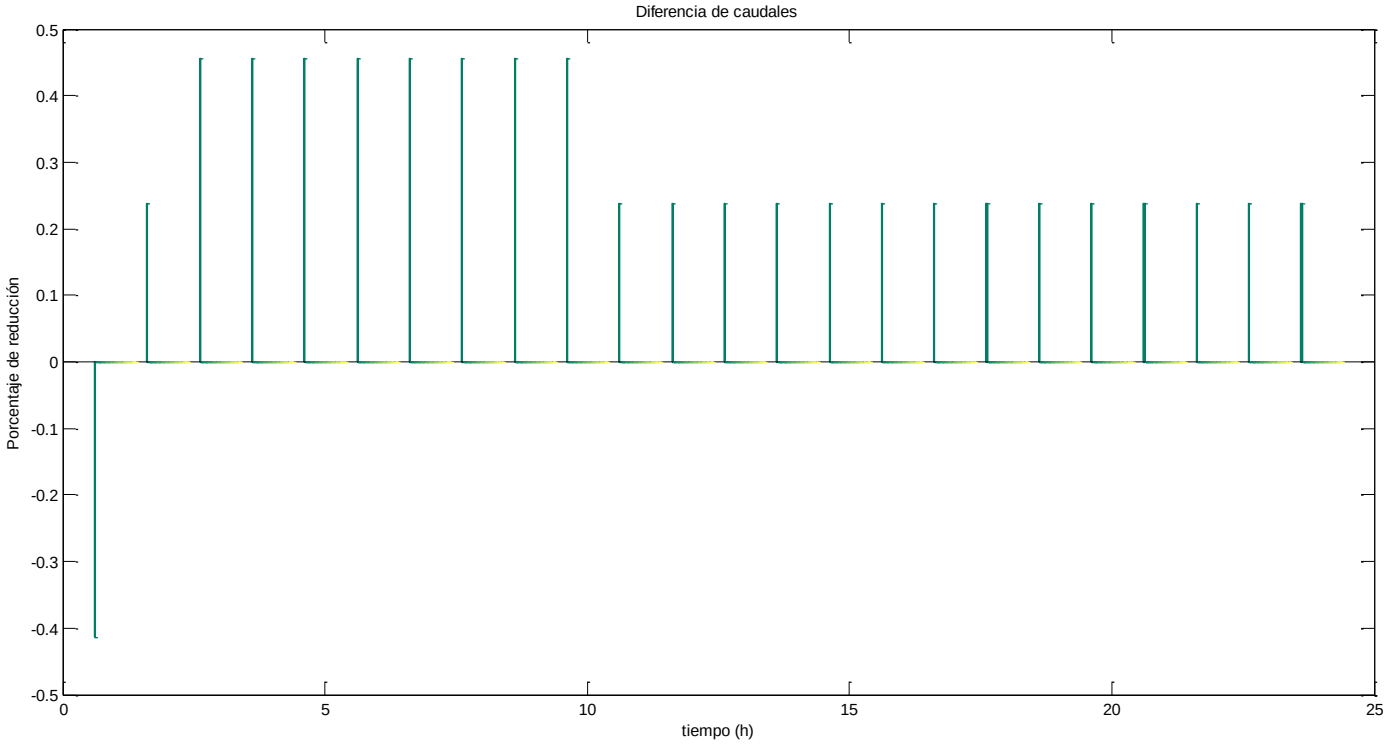
Figura 81 Comparación de caudales contralados vs caudal constante.



Fuente: (El autor 2012)

Se puede observar con claridad que la reducción de caudal de aire es alto, sin embrago durante las primeras horas de control, el agente no logra disminuir el caudal, esto gracias a que la condición inicial de oxígeno disuelto es de 4 mg/l, lo que conlleva a que el agente inicie con un alto caudal para, luego disminuirlo drásticamente para la búsqueda del objetivo (1 – 2 mg/l).

Figura 82 Porcentaje de reducción de caudal de aire.



Fuente: (El autor 2012)

De acuerdo a lo encontrado, se logró determinar que el agente redujo caudal de aire entre un rango del 20% hasta el 45%, donde de las 24 horas a las cuales se sometió el control, en 23 de estas se encontró reducción, es importante mencionar que este ahorro se encuentra dentro de los porcentajes de ahorro hallados en la literatura, por ejemplo Thornton, Sunner *et al.*, en 2010 lograron alcanzar hasta el 20%, Thunberg, Sundin *et al*. en 2009 alcanzaron el 18%, Ayesa, De la Sota *et al*. durante el 2006 obtuvieron una reducción del 15%, Stare, Vrečko *et al.* en 2007 llegaron hasta el 23% y se encontró hasta porcentajes de reducción del 45% por Vreco en 2006.

6. CONCLUSIONES

De acuerdo a lo observado dentro de las diferentes simulaciones realizadas, se pueden presentar altas variaciones en los valores de las acciones de control ejecutadas, especialmente en el caudal de recirculación de lodos, lo cual en condiciones reales puede llevar al desgasto de los equipos utilizados, por tal motivo es necesario programar al agente, para que no ejecute acciones de control dentro de una banda optima de desempeño en el sistema de lodos activados, así se disminuirían las acciones y por ende el desgaste de los equipos.

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 los caudales son sensibles bajo diferentes condiciones (concentraciones y dimensiones), sin embargo se evidencia que el control de la tasa de recirculación de lodos, presenta mayor variación (% de recirculación) durante los diferentes tiempos de control y presentando alto impacto sobre el crecimiento de la biomasa y el consumo del sustrato, situación que el oxígeno solo presento el disminuir los volúmenes de los tanques.

El comportamiento del agente mostró, que a pesar que las recompensas más altas sean muy pocas con respecto a las penalizaciones (ejemplo figuras 70 y 73), este siempre seleccionará estas acciones que lo recompensen, ya que como el Q-learning es un modelo libre donde se busca las probabilidades más altas de acciones que generen excelentes recompensas, hace que este siempre se mueva por estos sectores.

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, incumpliendo con esta restricción más del 50% de los datos, cabe resaltar que a pesar de esta situación el agente siempre buscó la mejor opción sobre el caudal de recirculación de lodos. Por otra parte es claro que con el día típico el agente cumple perfectamente realiza el control de manera adecuada alcanzado muy pocas penalizaciones.

A pesar de que el agente realiza una variedad de acciones de control cuando es variado el volumen de los tanques, no se ve afectado su desempeño, por el contrario al reducir estos se mejoró su cometido presentado los mejores resultados para los 13 días, a pesar que las recompensas eran inferiores a las penalidades.

Como recomendación para el buen funcionamiento del agente, se considera que se debe identificar con anterioridad, las concentraciones del sustrato de entrada, para en el tratamiento primario disminuir los cambios bruscos en concentraciones, de manera que las perturbaciones sean atenuadas o incluso evitarse.

El controlador alcanzó porcentaje de reducción hasta de un 45%, teniendo como referencia un caudal constante durante la operación de la planta, lo que nos muestra un ahorro energético alto y por ende económico sobre la planta. Por otra parte de las 24 horas en las que se realizó control, 23 de estas obtuvieron reducción de caudal de aire.

Por otra parte se muestra que las acciones de control no tienen correlación entre ellas, por lo tanto la variación de alguna de las dos, no afectará a la otra acción de control pero sí el comportamiento del sistema, esto se da gracias a que se muestra un comportamiento de histéresis. Por otra parte se muestra una correlación inversa entre el comportamiento del oxígeno y el sustrato de entrada, sin embargo si se eliminaran las concentraciones que se encuentran por fuera de la banda de oxígeno máximo y mínimo la correlación tendería a verse más clara.

Es importante mencionar, que se observó que el agente intenta siempre mantener las concentraciones de OD entre el rango de 1 y 2 mg/l, sin embargo en algunos casos se presentan valores fuera de este, debido a los cambios bruscos en las concentraciones de DQO, ya que estos conllevan a la variación en el caudal de oxígeno lo que muestra que estos dos factores son directamente proporcionales entre ellos.

7. BIBLIOGRAFÍA

Ahn, H. S. (2009). Reinforcement Learning and Iterative Learning Control: Similarity and Difference. International Conference on Mechatronics and Information Technology**:** 422-424.

Aihe, D. (2008). A reinforcement learning technique for enhancing human behavior models in a context-based architecture. School of Electrical Engineering and Computer Science. University of Central Florida. **Doctor of Philosophy in Computer Engineering**.

Altuna Guevara, A. F. (2009). Instrumentación de un tanque continuamente agitado (CSTR) presurizado con intercambio de calor. Quito, Quito: USFQ, 2009. **Ingeniero Elelctrico-Eliectrónico**.

Arbib, M. A. (2003). The handbook of brain theory and neural networks, The MIT Press.

Azar, M. G., R. Munos, et al. (2011). "Speedy Q-Learning."

Baroni, P., G. Bertanza, et al. (2006). "Process improvement and energy saving in a full scale wastewater treatment plant: Air supply regulation by a fuzzy logic system." Environmental technology **27**(7): 733-746.

Barto, A. and M. Duff (1994). "Monte Carlo matrix inversion and reinforcement learning." Advances in Neural Information Processing Systems: 687-687.

Barto, A. G. and S. Mahadevan (2003). "Recent advances in hierarchical reinforcement learning." Discrete Event Dynamic Systems **13**(4): 341-379.

Barto, A. G., R. S. Sutton, et al. (1989). Learning and sequential decision making, University of Massachusetts.

Bernstein, D., S. Zilberstein, et al. (2001). Planetary rover control as a markov decision process. Sixth International Symposium on Artificial Intelligence, Robotics and Automation in Space, Montreal, Canada.

Bertsekas, D. P. (1995). Dynamic programming and optimal control, Athena Scientific Belmont, MA.

Bertsekas, D. P. and S. Ioffe (1996). "Temporal differences-based policy iteration and applications in neuro-dynamic programming." Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT, Cambridge, MA.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). "Neuro-Dynamic Programming (Optimization and Neural Computation Series, 3)."

Bhatnagar, S. and K. M. Babu (2008). "New algorithms of the Q-learning type." automatica **44**(4): 1111-1119.

Bongards, M. (1999). "Controlling biological wastewater treatment plants using fuzzy control and neural networks." Computational Intelligence: 142-150.

Boyles, W. (1997). The Science of Chemical Oxygen Demand. Technical Information Series. T. I. Series.

Brdys, M., W. Chotkowski, et al. (2002). Two-level dissolved oxygen control for activated sludge processes. 15th Triennial World Congress, Barcelona, Spain.

Busby, J. B. and J. F. Andrews (1975). "Dynamic modeling and control strategies for the activated sludge process." Journal (Water Pollution Control Federation): 1055-1080.

Busoniu, L., R. Babuska, et al. (2010). Reinforcement learning and dynamic programming using function approximators, CRC Pr I Llc.

Camacho, L. A., M. A. Díaz-Granados, et al. (2001). Contribución al desarrollo de un modelo de calidad del agua apropiado para evaluar alternativas de saneamiento del río Bogotá. Bogotá, Colombia Universidad de Los Andes.

Carlsson, B. and C. F. Lindberg (1998). "Some control strategies for the activated sludge process." Systems and Control Group, Uppsala University.

Consejo Nacional de Política Económica y Social (2009). Garantía de la nación a la corporación autónoma regional de Cundinamarca – CAR - para contratar una operación de crédito público externo con la banca multilateral hasta por la suma de US $250 millones o su equivalente en otras monedas destinado a financiar parcialmente el proyecto adecuación hidráulica y recuperación ambiental del río bogotá., Documento COMPES 3631.

Chachuata, B., N. Rocheb, et al. (2005). "Optimal aeration control of industrial alternating activated sludge plants." Biochemical Engineering Journal **23**: 277–289.

Chai, Q. and B. Lie (2008). Predictive control of an intermittently aerated activated sludge process, IEEE.

Chapra, S. (1997). Surface water-quality modeling, McGraw-Hill.

Chen, C. T. (1998). Linear system theory and design, Oxford University Press, Inc.

Choi, T. H. A., E. A. Yim, et al. (2009). Environmental reinforcemnet learning: A Real-time Learning Architecture for Primitive Behavior Refinement. Gainesville, University of Florida.

Dias, A. M. A., I. Moita, et al. (2008). "Activated sludge process monitoring through in situ near-infrared spectral analysis." Water Science & Technology **57**(10): 1643-1650.

Dietterich, T. G. (1997). "Machine-learning research." AI magazine **18**(4): 97.

Dolk, V. (2010). "Survey Reinforcement Learning."

Dorf, R. C., R. H. Bishop, et al. (2005). Sistemas de control moderno, Pearson Educación.

Duchène, P., E. Cotteux, et al. (2001). "Applying fine bubble aeration to small aeration tanks." Water Science & Technology **44**(2-3): 203-210.

Dunn William, C. (2005). Fundamentals of Industrial instrumentation and Process Control, McGraw-Hill Education.

EAAB-ESP (2007). Plan de Saneamiento y Manejo de Vertimientos-PSMV, Empresa de Acueducto y Alcantarillado de Bogotá, ESP, Gerencia Ambiental.

EAAB-ESP, E. d. A. y. A. d. B. (2005). Plan Maestro de Acueducto y Alcantarillado de Bogotá. Bogotá.

Feigenbaum, E. A. (1980). Knowledge Engineering: The Applied Side of Artificial Intelligence, DTIC Document.

Feinberg, E. A. and A. Shwartz (2002). Handbook of Markov decision processes: methods and applications, Springer Netherlands.

Fernández, F. (2002). Aprendizaje por refuerzo en estados de medios contínuos. Departamento de informática. Madrid, Universidad Carlos III de Madrid **Ph. D.**

Fernández, F. J., M.C.Castro, et al. (2011). "Reductionofaerationcostsbytuningamulti-setpointon/offcontroller: A casestudy." Control Engineering Practice: 1-7.

Ferrer, J., Rodrigo, et al. (1998). "Energy saving in the aeration process by fuzzy logic control." Water Science & Technology **38**(3): 209-217.

Fika, M., B. Chachuat, et al. (2005). "Optimal operation of alternating activated sludge processes." Control Engineering Practice **13**(7): 853–861.

Fiter, M., D. Güell, et al. (2005). "Energy saving in a wastewater treatment process: an application of fuzzy logic control." Environmental technology **26**(11): 1263-1270.

Fleischmann, N., G. Langergraber, et al. (2001). "On-line and in-situ measurement of turbidity and COD in wastewater using UV/VIS spectrometry."

Galluzzo, M., R. Ducato, et al. (2001). "Expert control of DO in the aerobic reactor of an activated sludge process." Computers & Chemical Engineering **25**(4-6): 619-625.

Gernaey, K. V., M. van Loosdrecht, et al. (2004). "Activated sludge wastewater treatment plant modelling and simulation: state of the art." Environmental Modelling & Software **19**(9): 763-783.

Giraldo J.M., Leirens S., et al. (2010). Nonlinear optimization for improving the operation of sewer systems: the Bogota Case Study. International Congress on Environmental Modelling and Software. Ottawa, Canada.

Glass, G. and J. Stanley (1986). Métodos estadísticos aplicados a las ciencias sociales, Prentice-Hall Internacional.

Glorennec, P. Y. (2000). Reinforcement learning: An overview. European Symposium on Intelligent Techniques, Rennes, Citeseer.

Gosavi, A. (2009). "Reinforcement learning: a tutorial survey and recent advances." INFORMS Journal on Computing **21**(2): 178-192.

Gruber, G., J. L. Bertrand-Krajewski, et al. (2006). "Practical aspects, experiences and strategies by using UV/VIS sensors for long-term sewer monitoring." Water Practice and Technology (paper doi10. 2166/wpt. 2006.020) **1**(1): 8.

Gruber, G., S. Winkler, et al. (2004). "Quantification of pollution loads from CSOs into surface water bodies by means of online techniques." Water science and technology: a journal of the International Association on Water Pollution Research **50**(11): 73.

Gullapalli, V. (1992). Reinforcement learning and its application to control. Amherst, University Massachusetts. **Ph.D**.

Guo, M., Y. Liu, et al. (2004). "A new Q-learning algorithm based on the metropolis criterion." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **34**(5): 2140-2143.

Gutiérrez, J. D., Riss., et al. (2004). "Lógica difusa como herramienta para la bioindicación de la calidad del agua con macroinvertebrados acuáticos en la sabana de bogotá - colombia." Caldasia **26**(1): 161-172.

Han, H. G. and J. F. Qiao (2011). "Adaptive dissolved oxygen control based on dynamic structure neural network." Applied Soft Computing.

Harmon, M. E. and S. S. Harmon (1996). "Reinforcement learning: a tutorial." Wright State University.

He, Z., A. Petiraksakul, et al. (2003). "Oxygen-Transfer Measurement in Clean Water." KMITNB **13**(1): 14-19.

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

Henze, M., W. Gujer, et al. (2000). Activated sludge models ASM1, ASM2, ASM2d and ASM3, IWA Publishing.

Herless, H. and N. Castro (2004). Tutorial de inteligencia artificial y sistemas expertos.

Hochedlinger, M., P. Hofbauer, et al. (2006). Online UV-VIS Measurements–The Basis for Future Pollution Based Sewer Real Time Control in Linz. 2nd International IWA Conference on Sewer Operation and Maintenance, Vienna - Austria.

Holenda, B., E. Domokos, et al. (2007). "Aeration optimization of a wastewater treatment plant using genetic algorithm." Optimal Control Applications and Methods **28**(3): 191-208.

Holenda, B., E. Domokos, et al. (2008). "Dissolved oxygen control of the activated sludge wastewater treatment process using model predictive control." Computers & Chemical Engineering **32**(6): 1270-1278.

Holmberg, U., G. Olsson, et al. (1989). "Simultaneous DO control and respiration estimation." Water Science & Technology **21**(10-11): 1185-1195.

Houcque, D. (2005). "Applications of MATLAB: Ordinary Differential Equations (ODE)." Robert R. McCormick School of Engineering and Applied Science-Northwestern University, Evanston.

Hu, J. and M. P. Wellman (1998). Multiagent reinforcement learning: Theoretical framework and an algorithm, Citeseer.

INCONTEC (2002). NTC 3629. Calidad de agua. Demanda Química de Oxigeno -DQO, Instituto Colombiano de Normas Técnicas y Certificación, ICONTEC.

Ingildsen, P., U. Jeppsson, et al. (2002). "Dissolved oxygen controller based on on-line measurements of ammonium combining feed-forward and feedback." Water Science and Technology: 453-460.

Isidori, A. (1995). Nonlinear control systems, Springer Verlag.

Jang, J. S. R. (1993). "ANFIS: Adaptive-network-based fuzzy inference system." Systems, Man and Cybernetics, IEEE Transactions on **23**(3): 665-685.

Jang, J. S. R. and C. T. Sun (1995). "Neuro-fuzzy modeling and control." Proceedings of the IEEE **83**(3): 378-406.

Jeppsson, U. (1996). Modelling aspects of wastewater treatment processes. Department of Industrial Electrical Engineering and Automation. Sweden, Lund University. **Ph.D.**

Kaelbling, L. P., M. L. Littman, et al. (1996). "Reinforcement learning: A survey." Arxiv preprint cs/9605103.

Kalker, T. J., VanGoor, et al. (1999). "Fuzzy control of aeration in an activated sludge wastewater treatment plant: desing, simulation ans evaluation." Water Science & Technology **39**(4): 71-78.

Kayser, R. (1999). "Activated Sludge Process." Biotechnology Set: 253-283.

Kiely, G. and J. M. Veza (1999). Ingeniería ambiental, McGraw-Hill Interamericana de España.

Kirk, D. E. (2004). Optimal control theory: an introduction, Dover Pubns.

Kohonen, T. (1990). "The self-organizing map." Proceedings of the IEEE **78**(9): 1464-1480.

Kretchmar, R. M. (2000). A synthesis of reinforcement learning and robust control theory. Department of Computer Science. Fort Collins, Colorado State University. **Doctor of Philosophy**.

Kuo, B. C. and M. F. Golnaraghi (2003). Automatic control systems, Wiley Hoboken, NJ.

Langergraber, G., N. Fleischmann, et al. (2003). "A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater." Water Science & Technology **47**(2): 63-71.

Langergraber, G., N. Fleischmann, et al. (2004). "Monitoring of a paper mill wastewater treatment plant using UV/VIS spectroscopy." Water Science & Technology **49**(1): 9-14.

Langergraber, G., J. Gupta, et al. (2004). "On-line monitoring for control of a pilot-scale sequencing batch reactor using a submersible UV/VIS spectrometer." Water Science & Technology **50**(10): 73-80.

Lewis, W. and W. Whitman (1924). "Principles of Gas Absorption." Industrial & Engineering Chemistry **16**(12): 1215-1220.

Lin, C. T., C. S. G. Lee, et al. (1996). Neural fuzzy systems: a neuro-fuzzy synergism to intelligent systems, Prentice hall PTR.

Lindberg, C.-F. (1997). Control and estimation strategies applied to the activated sludge process. Materials Science Systems and Control Group, Uppsala University. **Doctor of Philosophy in Automatic Control:** 214.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning.

Loch, J. and S. Singh (1998). Using eligibility traces to find the best memoryless policy in partially observable Markov decision processes, Morgan Kaufmann Publishers Inc.

Long, L., L. Fei, et al. (2011). Predicting wastewater sludge recycle performance based on fuzzy neural network, IEEE.

Lorenz, U., N. Fleischmann, et al. (2002). Adaptation of a new online probe for qualitative measurement to combined sewer systems. 9th International Conference on Urban Drainage, Portland - Oregon.

Maes, P. (1995). "Artificial life meets entertainment: lifelike autonomous agents." Communications of the ACM **38**(11): 108-114.

Mahadevan, S. (1994). To discount or not to discount in reinforcement learning: A case study comparing R learning and Q learning, Citeseer.

Makinia, J. (2010). Mathematical Modelling and Computer Simulation of Activated Sludge Systems, Intl Water Assn.

Makinia, J. and S. A. Wells (2000). "A general model of the activated sludge reactor with dispersive flow--I. model development and parameter estimation." Water Research **34**(16): 3987-3996.

Makinia, J. and S. A. Wells (2007). "Improvements in modelling dissolved oxygen in activated sludge systems." Portland State University **751**: 1-9.

Marsili-Libelli, S. (1990). "Adaptive estimation of bioactivities in the activated sludge process." Control Theory and Applications **137**(6): 349-356.

Martínez, E. and C. de Prada (2003). "Control inteligente de procesos usando aprendizaje por interacción." XXIV Jornadas de Automática, León, Septiembre.

Martinez, S. A. (2005). Tratamiento de aguas residuales con MATLAB, Reverte.

Maximiliano, P. and R. M. Laura. (2004). "Aprendizaje por Refuerzo: Algoritmo Q – Learning análisis de diversas técnicas de exploración." from http://www-2.dc.uba.ar/materias/robotica/TPFinal2004/RachiPolimeni.pdf.

McCarthy, J. (2004). "What is artificial intelligence." URL: http://www-formal. stanford. edu/jmc/whatisai. html.

McGraw-Hill (2005). McGraw-Hill concise encyclopedia of engineering, McGraw-Hill.

Metcalf and I. Eddy (1991). Wastewater Engineering: Treatment, Disposal, Reuse, McGraw-Hill New York, NY, USA.

Meyer, U. and H. J. Popel (2003). "Fuzzy-control for improved nitrogen removal and energy saving in WWT-plants with pre-denitrification." Water Science & Technology **47**(11): 69-76.

Mingzhi, H., W. Jinquan, et al. (2009). "Control rules of aeration in a submerged biofilm wastewater treatment process using fuzzy neural networks." Expert Systems with Applications **36**(7): 10428-10437.

Monostori, L. and B. C. Csáji (2006). "Stochastic dynamic production control by neurodynamic programming." CIRP Annals-Manufacturing Technology **55**(1): 473-478.

Montoro López, G. (1996). Contribución al estudio y desarrollo de técnicas de control aplicadas a la linealización de sistemas. Departament de Teoria del Senyal i Comunicacions. Barcelona, Universitat Politècnica de Catalunya. **Tesis Doctoral**.

Morales, E. (2011). "http://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/refuerzo.pdf."

Morla, H. (2004). Modeling of activated sludge process by using artificial neural networks, Middle East Technical University. **Master Of Science in Environmental Engineering:** 108.

Mulas, M. (2006). Modelling and Control of Activated Sludge Processes. Cagliari, Universitá degli Studi di Cagliari. **Dottorato di Ricerca in Ingegneria Industriale:** 141.

O'Brien, M., J. Mack, et al. (2010). "Model predictive control of an activated sludge process: A case study." Control Engineering Practice.

Olsson, G. (2007). "Automation Development in Water and Wastewater Systems." Environmental Engineering Research **12**(5): 197-200.

Olsson, G. and B. Newell (1999). Wastewater treatment systems: modelling, diagnosis and control, Intl Water Assn.

Olsson, G., M. K. Nielsen, et al. (2005). Instrumentation, control and automation in wastewater systems, IWA Publishing.

Peng, J. and R. J. Williams (1996). "Incremental multi-step Q-learning." Machine learning **22**(1): 283-290.

Penn-University. (2011). "Q-Learninig." from http://www.seas.upenn.edu/~jeromel/teaching/DP_fall09/notes/lec12_Qlearning.pdf.

Piotrowski, R., M. Brdys, et al. (2008). "Hierarchical dissolved oxygen control for activated sludge processes." Control Engineering Practice **16**(1): 114-131.

Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming, John Wiley & Sons, Inc.

Ribeiro, C. (2002). "Reinforcement learning agents." Artificial intelligence review **17**(3): 223-250.

Rieger, L., J. Alex, et al. (2006). "Modelling of aeration systems at wastewater treatment plants." Water Science & Technology **53**(4-5): 439-447.

Rodríguez, J., M. Díaz-Granados, et al. (2008). Bogotá's urban drainage system: context, research activities and perspectives. Proceedings of the 10th National Hydrology Symposium, British Hydrological Society, Exeter, United Kingdom.

Rummery, G. A. (1995). Problem solving with reinforcement learning. Engineering Department. Cambridge, University of Cambridge. **PhD**.

Russell, S. J. and P. Norvig (2004). Artificial intelligence: a modern approach, Prentice hall.

Rustum, R. (2009). Modelling activated sludge wastewater treatment plants using artificial intelligence techniques (fuzzy logic and neural networks). School of the Built Environment. Edinburgh, Reino Unido, Heriot-Watt University. **Doctor of Philosophy**.

S::CAN (2007). Manual S::CAN spectrometer probe y ANA::PRO. Vienna: Liquid Monitoring Networks.

Samuelsson, P. and B. Carlsson (2002). "Control of the aeration volume in an activated sludge process for nitrogen removal." Water Science & Technology **45**(4-5): 45–52.

Sathiya Keerthi, S. and B. Ravindran (1994). "A tutorial survey of reinforcement learning." Sadhana **19**(6): 851-889.

Seron, M. M. and J. Braslavsky (2001). Sistemas no lineales. Notas de clase. Universidad Nacional de Quilmes.

Seyssiecq, I., J. H. Ferrasse, et al. (2003). "State-of-the-art: rheological characterisation of wastewater treatment sludge." Biochemical Engineering Journal **16**(1): 41-56.

Shah, H. and M. Gopal (2011). "Reinforcement learning framework for adaptive control of nonlinear chemical processes." Asia-Pacific Journal of Chemical Engineering **6**(1): 138-146.

Shinskey, F. G. (1990). Process control systems: application, design and tuning, McGraw-Hill, Inc.

Singh, S. P. and R. S. Sutton (1996). "Reinforcement learning with replacing eligibility traces." Recent Advances in Reinforcement Learning: 123-158.

Sivanandam, N., S. Sumathi, et al. (2007). Introduction to Fuzzy Logic using MATLAB, Springer.

Smith, C. A. and A. B. Corripio (1985). Principles and practice of automatic process control, Wiley New York.

Stare, A., D. Vrečko, et al. (2007). "Comparison of control strategies for nitrogen removal in an activated sludge process in terms of operating costs: A simulation study." Water Research **41**(9): 2004-2014.

Staubmann, K., N. Fleischmann, et al. (2001). UV/VIS spectroscopy for the monitoring of testfilters. Proceedings of the IWA 2nd World Water Congress, Berlin.

Stenstrom, M. K., S. Y. B. Leu, et al. (2006). "Theory to Practice: Oxygen Transfer and the New ASCE Standard." Proceedings of the Water Environment Federation **2006**(7): 4838-4852.
Sutton, R. (1999). Open theoretical questions in reinforcement learning, Springer.
Sutton, R. (1999). "Reinforcement learning: Past, present and future." Simulated Evolution and Learning: 195-197.
Sutton, R. S. (1988). "Learning to predict by the methods of temporal differences." Machine learning **3**(1): 9-44.
Sutton, R. S. (1992). "Introduction: The challenge of reinforcement learning." Machine learning **8**(3): 225-227.
Sutton, R. S. and A. G. Barto (1998). Reinforcement learning: An introduction, Cambridge Univ Press.
Syafiie, S., F. Tadeo, et al. (2007). Model-Free Learning Control of Chemical Processes.
Syafiie, S., F. Tadeo, et al. (2007). "Model-free learning control of neutralization processes using reinforcement learning." Engineering Applications of Artificial Intelligence **20**(6): 767-782.
Syafiie, S., F. Tadeo, et al. (2011). "Learning control for batch thermal sterilization of canned foods." ISA transactions **50**(1): 82-90.
Syafiie, S., C. Vilas, et al. (2008). Intelligent control based on reinforcement learning for batch thermal sterilization of canned foods.
Szepesvári, C. (2010). "Algorithms for reinforcement learning." Synthesis Lectures on Artificial Intelligence and Machine Learning **4**(1): 1-103.
Taylor, G. (2004). Reinforcement Learning for Parameter Control of Image-Based Applications. Waterloo - Ontario, University of Waterloo. **Master**.
Thornton, A., N. Sunner, et al. (2010). "Real time control for reduced aeration and chemical consumption: a full scale study." Water science and technology: a journal of the International Association on Water Pollution Research **61**(9): 2169.
Torres, A. and J. Bertrand-Krajewski (2008). "Partial Least Squares local calibration of a UV-visible spectrometer used for in situ measurements of COD and TSS concentrations in urban drainage systems." Water science and technology: a journal of the International Association on Water Pollution Research **57**(4): 581.
Tsagarakis, K., D. Mara, et al. (2003). "Application of cost criteria for selection of municipal wastewater treatment systems." Water, Air, & Soil Pollution **142**(1): 187-210.
Uribe, S. (2011). Aprendizaje automático de maniobras autónomas de combate áereo aplicadas a videojuegos. Facultad de ciencias exactas y naturales. Buenos Aires, Universidad de Buenos Aires. **Licenciatura en Ciencias de la Computación**.
Vrecko, D., N. Hvala, et al. (2003). "Feedforward-feedback control of an activated sludge process: a simulation study." Water science and technology: a journal of the International Association on Water Pollution Research **47**(12): 19.
Vrecko, D., N. Hvala, et al. (2006). "Improvement of ammonia removal in activated sludge process with feedforward-feedback aeration controllers." Water Science and Technology **53**(4-5): 125-132.
W. Josemans (2009). Generalization in Reinforcement Learning. Leeuwarden, UNIVERSITY OF AMSTERDAM.
Water Environment Federation (2009). Design Of Municipal Wastewater Treatment Plants, McGraw-Hill Professional.
Watkins, C. J. C. H. (1989). Learning from delayed rewards. Cambridge, King's College, Cambridge. **Ph.D**.
Watkins, C. J. C. H. and P. Dayan (1992). "Q-learning." Machine learning **8**(3): 279-292.

Whiteson, B. A. (2007). Adaptive representations for reinforcement learning. Graduate School. Austin, University of Texas at Austin. **pH.D**.

Whiteson, S. (2010). Adaptive representations for reinforcement learning, Springer-Verlag New York Inc.

Wicaksono, H. (2011). Q learning behavior on autonomous navigation of physical robot, IEEE.

Winkler, S., E. Saracevic, et al. (2008). "Benefits, limitations and uncertainty of in situ spectrometry." Water science and technology: a journal of the International Association on Water Pollution Research **57**(10): 1651.

Xhafa, F., A. Abraham, et al. (2010). Computational intelligence for technology enhanced learning, Springer Verlag.

Zhang, K. J., Y. K. Xu, et al. (2008). "Policy iteration based feedback control." automatica **44**(4): 1055-1061.

Zhang, P., M. Yuan, et al. (2008). "Improvement of nitrogen removal and reduction of operating costs in an activated sludge process with feedforward-cascade control strategy." Biochemical Engineering Journal **41**(1): 53-58.

Zhang, Z. and W. Guo (2010). Adaptive controller based on flexi-structure neural network for dissolved oxygen control, IEEE.

Zubowicz, T., M. A. Brdys, et al. (2010). "Intelligent PI controller and its application to dissolved oxygen tracking problem." Journal of Automation, Mobile Robotics & Intelligent Systems **4**(3).

8. ANEXOS

8.1. Anexo 1: Matriz de comparación de métodos de aprendizaje por refuerzo.

8.2. Anexo 2: Código en Matlab de la planta lodos activados.

```
%%%%%%%%%%%%%%%%%%%% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%Comportamiento de lodos activados modelo dinámico%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%Ecuaciones en el reactor%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%SUSTRATO
%%% dS/dt = (Qf/V)*Sf - (Qo/V)S - ((miu*X)/Y)
%%%BIOMASA
%%% dX/dt = (Qr/V)*Xr - (Q0/V)X + miu*X - kd*X
%%%OXÍGENO DISUELTO
%%% dC02/dt = (Qf/V)*Co2f - (Qo/V)Co2 - ((mui*X)/Yo2)-b*X+kla*(Csr-Co2)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%Ecuaciones en el sedimentador%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%BIOMASA
%%% dXr/dt = (Qu/Vs)*Xr - (Qo/Vs)*X

function dy=lodosactivados(t,y,Sf,control)
%ENTRADA
Qf=10000; %(m3/d)----- Caudal de agua del afluente

% ++++++++++++++++++++++++++++++++
%Parámetros de control
% ++++++++++++++++++++++++++++++++
klaw=0.0081*control(1)-2.85;%(1/d)----- Coeficiente de transferencia de oxígeno
% klaw=control(1);%(1/d)----- Coeficiente de transferencia de oxígeno
Qr=Qf*control(2); % (m3/d)----- Caudal de recirculación

%Parámetros cinéticos
Miu=1.97; %(1/d)----- Velocidad especifica de crecimeinto máx.
b=0.259; %(d/1) ----- Kg de (02) para la respiración endógena
Y=0.33; % ----- Coeficiente de rendimiento [mg(SSV)producido/mg(DQO)consumido]
ks=137.3; %(mg/L)----- Constante de afinidad
Yo=0.915; %Yo=a/Y ----- Coeficiente de rendimiento de oxígeno
kd=0.0601; %(1/d) ----- Coeficiente de muerte

%Condiciones de operación
```

```
% Sf=668; %(mg/L)----- Concetración de efluente
Co2f=0.3; %(mg/L)----- Concentración de oxígeno en el afluente
Csr=7.02; %(mg/L)----- Concentración de saturación de oxígeno
% klaw=420; %(1/d)----- Coeficiente de transferencia de oxígeno

% Qr=500; % (m3/d)----- Caudal de recirculación
Qw=300; %(m3/d)----- Caudal de purga
V=5000; %(m3) ----- Volúmen del reactor
Vs=250; %(m3) ----- Volúmen del sedimentador

dy=zeros(4,1); %vector columna

dy(1)=(Qf/V)*Sf-((Qf+Qr)/V)*y(1)-((Miu/Y)*y(1)*y(2))/(ks+y(1));%Ecuación del sustrato
dy(2)=(Qr/V)*y(3)-((Qf+Qr)/V)*y(2)+((Miu*y(1)*y(2))/(ks+y(1)))-kd*y(2);%Ecuación de biomasa
dy(3)=((Qf+Qr)/Vs)*y(2)-((Qr+Qw)/Vs)*y(3);%Ecuacion nivel del tanque
dy(4)=klaw*(Csr-y(4))+(Qf/V)*Co2f-((Qf+Qr)/V)*y(4)...
   -(((1/Yo)*Miu*y(1)*y(2))/(ks+y(1)))-b*y(1);%Ecuación de balance de oxígeno
```

8.5. Anexo 3: Código en Matlab del agente.

```
function [respuesta estadoFuturo]=evaluarAgente(estado,horizonteSol,objetivo,tipoPlanta)
% function respuesta=evaluarAgente(estado,horizonteSol,objetivo)
% estado=[Y entrada]

%La condicion inicial
Y0=estado(1:end-1);
entrada=estado(end);

%*****************************
% el timepo de solucion h
%*****************************
ini=datenum('2000-01-01 00:00:00');
fin=addtodate(ini, horizonteSol, 'hour');
tsol=linspace(ini,fin,horizonteSol*60+1);%el timepo de solucion 6min*horas +1
tsol=tsol(1:end-1);%Le quita el sobrante

%*****************************
% se estima la entrada
%*****************************
Sus_entEstimado=ones(length(tsol),1)*entrada;

%*****************************************************
% SE EVALUAN LAS DIFERENTES ACCIONES DE CONTROL
%*****************************************************
% Qaire_base=25000;
Qaire_base=25000;
Qaire=Qaire_base*.5:5000:2*Qaire_base;
Qresircula=linspace(.1,1.1,100);
```

```
% Qresircula=100:100:10000*.5;
costo=zeros(length(Qaire),length(Qresircula));
respuestas=zeros(length(Qaire),length(Qresircula),2);
for q=1:length(Qaire)
   for r=1:length(Qresircula)
      %Selecion la accion de control
      control=[Qaire(q) Qresircula(r)];
      %Realiza la acción de control
            switch tipoPlanta
               case 1
                 planta=@(t,Y,Sus_ent)lodosactivados(t,Y,Sus_ent,control);
               case 2
                 planta=@(t,Y,Sus_ent)lodosactivadosLibro(t,Y,Sus_ent,control);
            end

      %Soluciona el sistema
      [Ysol]=ode4(planta,tsol,Y0,Sus_entEstimado,tipoPlanta);

      %Se estima la diferencia entre lo obtenido y lo observado
      % costo(q)=(objetivo-Ysol(end,4)).^2;%El oxigeno

      %Las variables a controlar
      %Oxigeno Sustrato
      switch tipoPlanta
        case 1
          respuestaSistema=[Ysol(:,4) Ysol(:,1)];
        case 2
          respuestaSistema=[Ysol(:,2) Ysol(:,1)];
      end

%       respuestaSistema=[Ysol(:,4) Ysol(:,1)];
%       respuestaSistema=[Ysol(:,2) Ysol(:,1)];
      respuestas(q,r,:)=respuestaSistema(end,:);
      %Error del oxigeno + El Error del sustrato
      costo(q,r)=50*sum((objetivo(1)-respuestaSistema(:,1)).^2)+...
         50*sum((2-respuestaSistema(:,1)).^2)+...
         sum((objetivo(2)-respuestaSistema(:,2)).^2)+...
         sum((respuestaSistema(:,2)).^2);
%       costo(q,r)=var(Ysol(:,4))*10+ sum((objetivo-respuesta).^2);%El oxigeno
   end
end

% %*****************************
% %GRAFICA LA SUPERFICIE COSTO
% %*****************************
[X,Y] = meshgrid(Qresircula,Qaire);
surf(X,Y,costo)
contourf(X,Y,costo)
```

```
xlabel('Q RES')
ylabel('Q AIRE')
zlabel ('Costo')
view(54,22)
%*****************************************************
% SELECIONA LA MEJOR OPCION
%*****************************************************
% Menor costo
costoOK=costo==min(min(costo));
%saca la pos de los indices bacanos
[ii jj]=find(costoOK);
respuesta=[Qaire(ii) Qresircula(jj)];
respuesta=respuesta(1:2);
estadoFuturo=respuestas(ii,jj,:);
estadoFuturo=estadoFuturo(1:2);
```

8.5. Anexo 3: Código en Matlab que ejecuta completo los códigos.

```
% ////////////////////////////////////////////
% AGENTE CONTROLADOR DE PLANTA DE TRATAMIENTO
% ////////////////////////////////////////////
clear all;
%*********************
%El tipo de planta
%*********************
tipoPlanta=1;% 1:paper 2:Libro
%*********************
%el objetivo
%*********************
objetivo=[1 100];%Oxigeno Sustrato
% objetivo=[3 mean(media)*0.5];%Oxigeno Sustrato

%*********************************
% el timepo de control
%*********************************
horizonteSol=1; %horas

%*********************
%Carga la entrada
%*********************
% %Todos los datos
% load('datostotal.mat');
% [ndatos ndias]=size(diastotal);
% entrada=reshape(diastotal,ndatos*ndias,1);
%la media de entrada
load('media.mat');
[ndatos ndias]=size(media');
```

```
entrada=reshape(media,ndatos*ndias,1);
%*************************************************************
% el timepo de solución: longitud de la serie
%*************************************************************
ini=datenum('2000-01-01 00:00:00');
fin=datenum(['2000-01-0',num2str(ndias+1),' 00:00:00']);
t=linspace(ini,fin,ndias*24*60+1);%el timepo con una medicion de mas
t=t(1:end-1);%Le quita el sobrante

%Arregla los datos
%calcula el diferencial
 df=[0;diff(entrada)];
 %los datos atipicos los quita

%  pos=abs(df)>nanstd(df)*1.5;
pos=abs(df)>300;
 entrada(pos)=NaN;

%  ++++++++++++++++++++
%   INTERPOLA LOS NAN
%  ++++++++++++++++++++
 Sus_ent=naninterp(entrada);
%  Sus_ent(isnan(Sus_ent))=[];
%  t=t(1:length(Sus_ent));

 plot(t,Sus_ent)

%Inicializa los vectores de estado
Y=[];

%*********************************
%La condicion inicial
%*********************************
switch tipoPlanta
   case 1
      Y0=[Sus_ent(1) 500 6500 4];
   case 2
      Y0=[Sus_ent(1) 10 500];
end

%los ciclos de control
N=fix(length(Sus_ent)/(horizonteSol*60));

%*********************************
%Las acciones de control realizadas
%*********************************
tiemposControl=linspace(ini,fin,N);%el timepo con una medicion de mas
% tiemposControl=tiemposControl(1:end-1);%Le quita el sobrante
```

```
%Acciones para cada ciclo de control
accionesDeControl=zeros(N,2);
%Estados cada ciclo de control
estadosPlanta=zeros(N,1);
respuestaPlanta=zeros(N,2);
for kk=1:N
%********************
%La accion de control
%********************
% Qaire=510;
estado=[Y0 Sus_ent((kk-1)*(horizonteSol*60)+1)];%La condicion actual y la media actual
[control estadoFuturo]=evaluarAgente(estado,horizonteSol,objetivo,tipoPlanta);
accionesDeControl(kk,:)=control;
estadosPlanta(kk)=Sus_ent((kk-1)*(horizonteSol*60)+1);
respuestaPlanta(kk,:)=estadoFuturo;
%*****************************
% el timepo de solucion h
%*****************************
ini=datenum('2000-01-01 00:00:00');
fin=addtodate(ini, 6, 'hour');
tsol=linspace(ini,fin,horizonteSol*60+1);%el timepo de solucion 6min*horas +1
tsol=tsol(1:end-1);%Le quita el sobrante

%*****************************
% se estima la entrada
%*****************************
Sus_entEstimado=zeros(size(tsol));
Sus_entEstimado(:)=Sus_ent((kk-1)*(horizonteSol*60)+1)';

% LA entrada de verdad
% Sus_entEstimado(:)=media((kk-1)*(horizonteSol*60)+1:(kk)*(horizonteSol*60))';

%Realiza la acción de control
switch tipoPlanta
   case 1
      planta=@(t,Y,Sus_ent)lodosactivados(t,Y,Sus_ent,control);
   case 2
      planta=@(t,Y,Sus_ent)lodosactivadosLibro(t,Y,Sus_ent,control);
end
% planta=@(t,Y,Sus_ent)lodosactivados(t,Y,Sus_ent,control);

%Soluciona el sistema
[Ysol]=ode4(planta,tsol,Y0,Sus_entEstimado,tipoPlanta);

%Actualiza el vector de estado del sistema
   Y0=Ysol(end,:);
```

```
%Actualiza el vector de estados
Y=[Y; Ysol];
% plot(t(1:length(Y))',Y)

% %*****************************
% %GRAFICA LA funcion de prob conjunta
% %*****************************
 pxy = probxy(accionesDeControl);
bar3(pxy)
xlabel('Q AIRE')
ylabel('Q RES')
getframe();
end

%%
figure
% % Grafica el estado actual del sistema
% ------------
%La entrada
% ------------
subplot(5,1,1)
plot(t,Sus_ent,'color',[139 69 19]/255)
title('Sustrato Entrada');
datetick('x','HH:MM')
xlim([t(1) t(end)])
% ------------
% el control
% ------------
subplot(5,1,2)
[AX,H1,H2] =
plotyy(tiemposControl,accionesDeControl(:,1),tiemposControl,accionesDeControl(:,2),@stairs);
title('ACCIONES DE CONTROL');
% datetick('x','HH:MM')
xlim([t(1) t(end)])
% ------------
%el estado
% ------------
subplot(5,1,3)
plot(Y(:,1),'color',[139 69 19]/255)
xlim([1 length(Y)])
%el objetivo
hold on;
plot([1 length(Y)],[objetivo(2) objetivo(2)],'--','color',[173 255 47]/255)
xlim([1 length(Y)])
title('sustrato');
datetick('x','HH:MM')

subplot(5,1,4)
```

```
switch tipoPlanta
  case 1
    plot(Y(:,2),'color',[50 205 50]/255)
  case 2
     plot(Y(:,3),'color',[50 205 50]/255)
end

xlim([1 length(Y)])
title('Biomasa');
datetick('x','HH:MM')

subplot(5,1,5)
switch tipoPlanta
  case 1
    plot(Y(:,4),'color',[0 206 209]/255)
  case 2
    plot(Y(:,2),'color',[0 206 209]/255)
end

hold on;
plot([1 length(Y)],[objetivo(1) objetivo(1)],'--','color',[173 255 47]/255)
plot([1 length(Y)],[2 2],'--','color',[173 255 47]/255)
xlim([1 length(Y)])
title('Oxigeno');
datetick('x','HH:MM')

%tiempo | sustrato de entrada | oxigeno qresirculacion | sustrato salida a un horizonteSol

[tiemposControl' estadosPlanta accionesDeControl respuestaPlanta]

figure;
 pxy = probxy(accionesDeControl);
bar3(pxy)
xlabel('Q AIRE')
ylabel('Q RES')
```

8.5. Anexo 4: Código en Matlab ODE4.

```
function Y = ode4(fun,tspan,y0,X,restric)
%ODE4  Solve differential equations with a non-adaptive method of order 4.
%   Y = ODE4(ODEFUN,TSPAN,Y0) with TSPAN = [T1, T2, T3, ... TN] integrates
%   the system of differential equations y' = f(t,y) by stepping from T0 to
%   T1 to TN. Function ODEFUN(T,Y) must return f(t,y) in a column vector.
%   The vector Y0 is the initial conditions at T0. Each row in the solution
%   array Y corresponds to a time specified in TSPAN.
%
```

```
%   Y = ODE4(ODEFUN,TSPAN,Y0,P1,P2...) passes the additional parameters
%   P1,P2... to the derivative function as ODEFUN(T,Y,P1,P2...).
%
%   This is a non-adaptive solver. The step sequence is determined by TSPAN
%   but the derivative function ODEFUN is evaluated multiple times per step.
%   The solver implements the classical Runge-Kutta method of order 4.
%
%   Example
%       tspan = 0:0.1:20;
%       y = ode4(@vdp1,tspan,[2 0]);
%       plot(tspan,y(:,1));
%    solves the system y' = vdp1(t,y) with a constant step size of 0.1,
%    and plots the first component of the solution.
%

if ~isnumeric(tspan)
  error('TSPAN should be a vector of integration steps.');
end

if ~isnumeric(y0)
  error('Y0 should be a vector of initial conditions.');
end

h = diff(tspan);
if any(sign(h(1))*h <= 0)
  error('Entries of TSPAN are not in order.')
end

y0 = y0(:);   % Make a column vector.


neq = length(y0);
N = length(tspan);
Y = zeros(neq,N);
F = zeros(neq,4);

Y(:,1) = y0;
for i = 2:N
  ti = tspan(i-1);
  hi = h(i-1);
  yi = Y(:,i-1);
  F(:,1) = fun(ti,yi,X(i));
  F(:,2) = fun(ti+0.5*hi,yi+0.5*hi*F(:,1),X(i));
  F(:,3) = fun(ti+0.5*hi,yi+0.5*hi*F(:,2),X(i));
  F(:,4) = fun(tspan(i),yi+hi*F(:,3),X(i));
  Y(:,i) = yi + (hi/6)*(F(:,1) + 2*F(:,2) + 2*F(:,3) + F(:,4));
  if nargin >4
     switch restric
```

```
        case 1
           %Para q el sustrato no sea <0
           if Y(1,i)<=0
           Y(1,i)=0.01;
           end
           %Para q no sea ox <0
           if Y(4,i)<0
           Y(4,i)=0;
           end
        case 2
           %Para q no sea ox <0
           if Y(2,i)<=0
           Y(2,i)=0;
           end
           %Para q el sustrato no sea <0
           if Y(1,i)<=0
           Y(1,i)=0.01;
           end
      end

  end
end
Y = Y.';
```