HERRAMIENTA BASADA EN MINERÍA DE DATOS PARA AUTOMATIZACIÓN DEL DISEÑO DE SISTEMAS INTELIGENTES EN EDAR

Pascual-Pañach, Josep – Consorci Besòs Tordera, Universitat Politècnica de Catalunya
Cugueró-Escofet, Miquel Àngel – Consorci Besòs Tordera
Sànchez-Marrè, Miquel – Universitat Politècnica de Catalunya
Aguiló-Martos, Pere – Consorci Besòs Tordera

SUMARIO

Uno de los principales problemas para diseñar e implementar un sistema de supervisión y control para un proceso radica en la necesidad de establecer una solución ad-hoc para cada instalación. La interoperabilidad de los diferentes métodos utilizados para este fin es uno de los desafíos actuales relacionados con el desarrollo de Sistemas Inteligentes de Soporte a la Toma de Decisiones (IDSS), con el objetivo de garantizar la interacción y reutilización de los diferentes métodos basados en modelos, en conocimiento experto o en minería de datos.

En este trabajo se propone el uso de entornos y flujos de trabajo visuales para permitir la automatización del diseño e implementación de Sistemas Inteligentes de Control de Procesos (IPCS). Estos entornos permitirán al usuario especificar las características de un proceso concreto, así como los modelos requeridos —basados en datos y en conocimiento experto—, utilizando un entorno de desarrollo visual, con la finalidad de implementar la estrategia de control más adecuada a cada instalación particular. La herramienta propuesta se basa en una arquitectura de tres capas: la primera se corresponde a un proceso offline de generación de modelos e.g. *data-driven* a partir de datos históricos del sistema, con la finalidad de supervisarlo y controlarlo. La segunda se corresponde a un diagrama de flujo del sistema, incluyendo los distintos subprocesos que lo configuran y las señales correspondientes. Finalmente, la tercera capa es el núcleo de la aplicación, en la que se utilizan los modelos obtenidos por parte de los diferentes métodos de razonamiento inteligente, usados para supervisar el sistema, así como para generar las consignas de los actuadores.

Así, a partir de la arquitectura propuesta se podrá generar automáticamente el diseño final para el control y supervisión del proceso. La naturaleza visual de la solución propuesta permite utilizar el propio flujo de control como interfaz gráfica de usuario, pudiéndose añadir distintos parámetros configurables por el usuario, así como indicadores clave de rendimiento (en inglés, KPI), útiles para dar soporte a las decisiones relacionadas con el sistema.

El método presentado es genérico, pudiéndose implementar en aplicaciones de distinta tipología a la presentada en este trabajo, siendo la evolución natural el escalado a sistemas reales más complejos, aprovechando las ventajas que proporciona la generalidad de la solución propuesta para adaptar el método a otras instalaciones/aplicaciones.

Finalmente se muestran los resultados obtenidos con un prototipo probado en una EDAR en el ámbito del Consorci Besos Tordera (CBT), para el control de una de las variables del proceso biológico.

PALABRAS CLAVE

Estación Depuradora de Aguas Residuales, Sistema Inteligente de Control de Procesos, Minería de datos, Interoperabilidad, Flujos de trabajo visuales

INTRODUCCIÓN

Tradicionalmente, el campo de los Sistemas de Soporte a la Decisión Ambiental (en inglés, EDSS) ha intentado utilizar modelos que representen el mundo real con el fin de reproducir su comportamiento y evolución (Cugueró-Escofet et al. 2014). Históricamente, los antiguos EDSS utilizaban sólo modelos, a pesar de que solían estar disponibles grandes cantidades de datos recopilados del sistema, por lo que se empezaron a emplear nuevos modelos empíricos. Los modelos empíricos se basan en la observación directa, medidas y extensos registros de datos. Los primeros modelos empíricos utilizados fueron métodos matemáticos y estadísticos, como por ejemplo la Regresión Lineal Múltiple (MLR). Luego, el éxito de diversas técnicas de aprendizaje automático inductivo dentro del área de la Inteligencia Artificial (IA) condujo a su aplicación en los EDSS. Algunos ejemplos de estos métodos son, por ejemplo, los modelos de Reglas de Asociación (AR), modelos de Reglas de Clasificación (CR), modelos de Árbol de Decisión (DT) o redes bayesianas (BN). Desde los años 80, tanto los modelos empíricos matemáticos / estadísticos como los modelos empíricos de aprendizaje automático (*Machine Learning*) posteriores se denominaron métodos de minería de datos (*Data Mining*), ya que resultan de un proceso de minería que utiliza estos datos. Con el uso de modelos de minería de datos dentro del marco de la IA, los EDSS han evolucionado hacia Sistemas Inteligentes de Soporte a la Decisión Ambiental (IEDSS) (Sánchez-Marrè et al., 2006). Los IEDSS se pueden construir utilizando un solo modelo de IA o integrar varios modelos IA para ser más potentes, junto con otra información complementaria como puede ser información geográfica, modelos matemáticos o estadísticos, ontologías ambientales / sanitarias o alguna información económica. Los IEDSS integran conocimientos almacenados por expertos a través de años de experiencia en una determinada operación y gestión del proceso medioambiental, así como también conocimiento extraído a través del análisis inteligente de las grandes bases de datos disponibles procedentes de la explotación histórica del sistema. Así, la producción de modelos de conocimiento o datos, el razonamiento y la interoperación entre los modelos producidos son pasos clave para construir IEDSS fiables. En este contexto, los modelos de IA proporcionan una base sólida para la construcción de aplicaciones fiables, y la interoperabilidad entre los modelos IA y numéricos es, a día de hoy, uno de los principales retos abiertos en este campo. Además, el desarrollo del IEDSS se realiza de manera ad-hoc para cada tipo de sistema.

ESTADO DEL ARTE

Por un lado, la interoperabilidad se define como “la capacidad de dos o más sistemas o componentes para intercambiar información y utilizar la información que se ha intercambiado” (IEEE, 1990). Adicionalmente, la interoperabilidad semántica se consigue cuando los componentes comparten una comprensión común del modelo de información que hay detrás de los datos intercambiados (Manguinhas, 2010; Ouksel y Sheth, 1999). La integración y la interoperabilidad semántica han sido el foco de algunos trabajos de investigación en el campo del modelado de sistemas ambientales, como por ejemplo estos trabajos pioneros en integración semántica de modelos ambientales para su aplicación a sistemas de información global y toma de decisiones, especialmente relacionados con componentes y modelos de SIG (Sistema de Información Geográfica) (Mackay, 1999; Wesseling et al., 1996). En (Rizzoli et al., 1998), se presenta un trabajo relacionado con la integración de modelos y datos y su reutilización en EDSS, y en (Argent, 2004) se presenta una descripción general sobre la integración de modelos. En (Sottara et al., 2012) se presenta un trabajo interesante en esta área donde la plataforma de Bussiness Rules Management System (BRMS), Drools se utiliza como modelo de datos y entorno de ejecución, y en (Sánchez-Marrè, 2014) se propone un marco general para el desarrollo de IEDSS interoperables. En el campo de los sistemas de información se hacen varios trabajos sobre integración semántica de componentes de negocio (Elasri y Sekkaki, 2013, Kzaz et al., 2010). En cambio, otros trabajos se centran en la interoperabilidad semántica a través de arquitecturas orientadas al servicio, como (Vetere y Lenzerini, 2005). Con respecto a este tema, una de las maneras más eficaces de intercambiar información entre varios componentes de software y compartir la semántica correspondiente es a través del lenguaje XML (eXtensible Markup Language). XML es un meta-idioma destinado presenta almacenar datos de una forma legible (Erl, 2004). XML añade una capa de información inteligente a los datos que se intercambian, proporcionando la meta-información, que está codificada e incrustada como etiquetas auto-descriptivas en el documento. XML se implementa como un conjunto de elementos que se pueden personalizar para representar datos en contextos únicos. Un conjunto de elementos XML relacionados se puede clasificar como un vocabulario. Los vocabularios se pueden definir formalmente utilizando un lenguaje de definición de esquemas como *Document Type Definition* (DTD) o *XML Schema Definition Language* (XSD). Relacionando esto con el campo de la minería de datos, existe el *Data Mining Group* (DMG, 2014) un consorcio independiente que desarrolla estándares para este campo, como el Lenguaje de marcado de modelos predictivos (PMML). PMML es un estándar para los modelos estadísticos y de mine-

ría de datos con el apoyo de más de 20 proveedores y organizaciones. PMML utiliza XML para representar modelos de minería de datos, cuya estructura se describe mediante un esquema XML. Un documento PMML es, al fin y al cabo, un documento XML con un elemento raíz del tipo PMML. Puede contener más de un modelo, siendo los modelos más comunes de minería de datos compatibles con este estándar, como por ejemplo modelos asociativos, de regresión, árboles de decisión, reglas o *clustering*.

Por otra parte, la creación y ejecución de flujos de trabajo visuales puede ser una herramienta muy útil para especificar los flujos de datos en bruto, los modelos producidos, los ejecutores de estos modelos y los procesos auxiliares requeridos. Los flujos de trabajo son notaciones gráficas que se introdujeron para modelar y describir procesos empresariales (ter Hofstede et al., 2010). Permiten que el diseño y la especificación de un flujo de trabajo incluyan varios elementos, como pueden ser datos, intercambio de modelos PMML, productores de modelos, ejecutores de modelos, combinadores de soluciones, especificación de problemas, etc. Ésta filosofía para el control de procesos, por ejemplo, a nivel de organizaciones completas, se describe en la literatura (zur Muehl, 2004). En el enfoque presentado aquí se valorará el uso de diversas herramientas, como por ejemplo jBPM —java for Business Process Management (jBPM, 2017) —, que es un motor de procesos de negocio de código libre que soporta el estándar Business Process Model and Notation (BPMN 2.0) y proporciona un entorno de edición gráfico (BPMN, 2011) o YAWL (*Yet Another Workflow Language*), un lenguaje para el modelado de flujos de trabajo que también proporciona un entorno de trabajo de código abierto (ter Hofstede et al., 2010).

La idea principal de este trabajo es describir el proceso de resolución del IEDSS, a través de sus diferentes tareas y capas, utilizando flujos de trabajo, que se pueden ejecutar directamente o, alternativamente, para generar el correspondiente código de software para el IEDSS. Aunque hay algunas propuestas de arquitectura en la bibliografía para combinar algunos de estos modelos, según el conocimiento de los autores no existe un marco común para implementar IEDSS interoperables, lo que proporcionaría una manera sencilla de integrar y utilizar o reutilizar diferentes modelos de IA o estadísticos / numéricos en la misma herramienta. Hasta ahora, la mayor parte de la interoperabilidad de los modelos se logra mediante la interacción ad-hoc, y por tanto se puede mejorar considerablemente. Así, el objetivo de este trabajo de investigación es proporcionar un enfoque útil y sistemático para interoperar diferentes modelos en diferentes etapas del diseño del IEDSS y automatizar la construcción del IEDSS mediante soluciones basadas en flujos de trabajo. Los casos de estudio se centran en sistemas de saneamiento en el ámbito del Consorcio Besos Tordera (CBT), aunque el marco propuesto se considera general, hasta el punto de poder utilizarlo con sistemas de diferente naturaleza, fuera de este ámbito.

MOTIVACIÓN DEL PROYECTO

La motivación de este proyecto viene dada por la necesidad de diseñar sistemas de control y supervisión ad-hoc para cada sistema de saneamiento. Así, el diseño de la herramienta de control y supervisión para cada EDAR depende tanto de los procesos que la integran —e.g. eliminación de nutrientes, eliminación de fósforo—, como del diseño de la planta —e.g. capacidad de tratamiento, tipo de planta, sensores y actuadores disponibles. Estas características particulares de cada sistema de saneamiento implican un gran coste en tiempo y recursos, desde la fase de diseño hasta el mantenimiento del Sistema Inteligente de Control de Procesos (IPCS), pasando por la implementación y la puesta en marcha.

Así, la herramienta propuesta en este trabajo reduce el tiempo de implementación del sistema de control y supervisión de la EDAR. Las funciones de la herramienta presentada son: a) el almacenaje de datos del proceso, b) la generación de consignas de control —i.e. cómo actuar en cada situación— y c) el soporte a la decisión, mostrando algunos indicadores clave (en inglés, KPI) para el proceso. Manteniendo estas funcionalidades, y teniendo en cuenta el auge de conceptos como *Big Data* o Industria 4.0, se propone el desarrollo de una nueva herramienta de monitorización y control inteligente basada en datos, en el uso de flujos de trabajo y en lenguajes de programación visuales.

En el caso del sector del agua, las tecnologías relacionadas con la Industria 4.0 no están tan maduras como en otros sectores, e.g. automoción, y su implementación supone un reto tanto actual como a futuro. Este enfoque es especialmente relevante en el contexto actual, en el que el volumen de datos monitorizados del proceso —muchas veces en tiempo real— es cada vez mayor, debido a la reducción de coste y la mejora tecnológica de los sensores, así como el aumento de la capacidades computacionales y la aparición de técnicas adecuadas para digerir estos volúmenes crecientes de datos. Estas nuevas tecnologías actúan como facilitadores para mejorar la calidad y la veloci-

dad en la toma de decisiones, mediante e.g. la creación de IEDSS basados en estas tecnologías, apuntando hacia una gestión cognitiva del agua. La estrategia pasa por la integración de sistemas de monitorización y control de los procesos, con una tendencia creciente a la sensorización de los mismos y a una comunicación fluida, transparente y estándar de estos datos entre estos procesos. Esto implica diseño y desarrollo de objetos inteligentes i.e. objetos virtuales con IA integrada; el uso de comunicación inteligente i.e. hiperconectividad, interoperabilidad y homogeneización de datos de múltiples fuentes y formatos, disponibles en toda la cadena de valor para ser transformados en información accionable en tiempo real; el uso de datos inteligentes i.e. la transformación de los datos y la información en conocimiento, proveyendo apoyo a la decisión; y la integración en plataformas inteligentes i.e. diseño y desarrollo de tecnología basada en arquitecturas híbridas que integren datos, objetos inteligentes, servicios, herramientas analíticas en sistemas de inteligencia distribuida, como es el caso de la herramienta planteada en este trabajo.

Conceptos Básicos

Tal y como se describe en esta sección anterior, el objetivo de este trabajo es el desarrollo de una herramienta basada en datos. Esto significa que el núcleo del sistema de control serán los datos históricos del proceso a controlar y supervisar. Así, la herramienta planteada toma las decisiones en base a la información y el conocimiento que se puede extraer de estos datos históricos o del conocimiento de un experto, como puede ser un jefe de planta en el caso de una EDAR. La extracción de conocimiento de los datos históricos se realiza mediante técnicas de minería de datos, para utilizar posteriormente este conocimiento con métodos de razonamiento que permitan tomar las decisiones adecuadas. La *Figura 1* muestra de una forma simplificada los distintos pasos a seguir desde el histórico de datos hasta el conocimiento extraído a partir de éste. Partiendo de una base de datos con el histórico de operación de la planta, los pasos previos necesarios para la obtención del conocimiento intrínseco a estos datos son la preparación —recopilación y selección de datos, que pueden estar en bases de datos distribuidas— y el tratamiento de estos datos —e.g. remuestreo, limpieza de datos irrelevantes. Una vez obtenida una base de datos preprocesada, se aplican técnicas de minería de datos con el fin de extraer conocimiento de éstos. En la *Figura 2* se muestra la arquitectura del sistema completo para un sistema de saneamiento general. Para el desarrollo de esta herramienta se propone el uso de flujos de trabajo y lenguajes de programación visual e.g. Matlab/Simulink. Matlab es una herramienta de software matemático que combina un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio. Integra múltiples *toolboxes* dedicadas a una gran variedad de funcionalidades, así como el entorno de programación visual Simulink, que permite la programación de flujos de trabajo visuales. El uso de este tipo de herramientas permite una programación más rápida e intuitiva, facilita la reutilización y comprensión del código y favorece la modularidad y flexibilidad de éste. Finalmente, el concepto de interoperabilidad es también muy importante. La interoperabilidad es la capacidad de compartir información entre los distintos sistemas y métodos que forman la herramienta. En la *Figura 3* se observan las diferentes relaciones entre sistemas y métodos de la herramienta que deben ser interoperables. El sistema de control estará basado en datos, por lo que es importante garantizar la comunicación y compatibilidad de toda la información que se introduzca en el sistema —obtenida de históricos de datos o expertos del proceso— con los modelos utilizados —e.g. sensores virtuales a partir de modelos físicos del sistema para obtener información no disponible—. Adicionalmente, se debe garantizar también la comunicación entre la herramienta de control y el sistema de saneamiento, así como con las bases de datos históricos.

Figura 1. Flujo de datos – minería de datos.

Figura 2. Arquitectura del sistema.

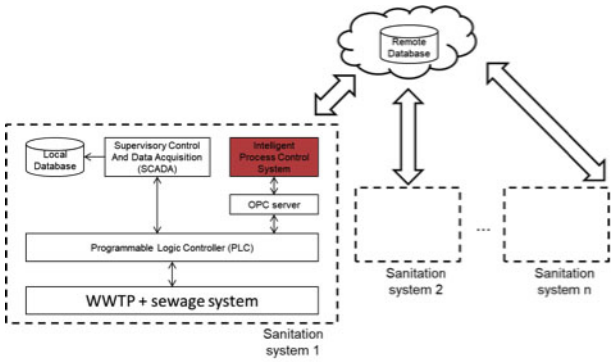


Figura 3. Interoperabilidad

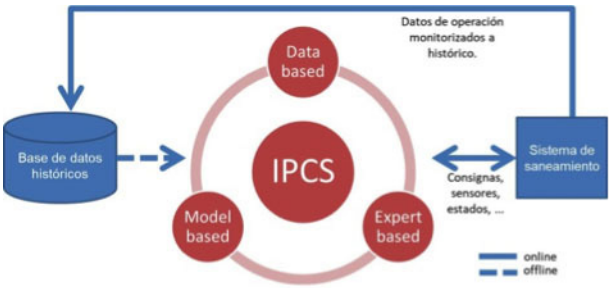


Figura 4. Estructura de la herramienta de control y supervisión propuesta

Entornos de Desarrollo

Con el fin de desarrollar el entorno presentado en este trabajo se propone el uso de flujos de trabajo visuales. El uso de lenguajes de programación visuales tiene algunas ventajas respeto los lenguajes tradicionales, como C o Java (Johnston et al., 2004), como pueden ser la programación más rápida e intuitiva, mejor comprensión del código, o facilitar la modularidad y la reutilización de las herramientas generadas.

La mayoría de lenguajes de programación disponen de librerías dedicadas a la minería de datos, como por ejemplo *scikit-learn* de Python o JDMP (*Java Data Mining Package*) de Java. A pesar de que estos lenguajes no están diseñados para la programación gráfica, pueden utilizarse para crear aplicaciones como un conjunto de cajas negras o procesos interconectados, tomando como base el concepto de *Flow Based Programming* (FBP) descrito en Morrison (2010). Por otro lado, existen algunos entornos que sí están orientados a la programación gráfica e.g. Matlab/Simulink o Labview. Estos entornos también incluyen librerías de minería de datos y otras herramientas especializadas, como pueden ser la conexión con bases de datos o el estándar de comunicación OPC, así como también compatibilidad con otros lenguajes como C y Java.

Para proporcionar la estandarización y reutilización de los modelos obtenidos en diferentes instalaciones, se propone el uso del estándar *Predictive Model Markup Language* (PMML), utilizado para representar modelos basados en datos. Para la construcción de estos modelos en la Capa 1, existen distintas plataformas basadas en programación visual dedicadas a la minería de datos, e.g. Rapid Miner.

Para la fase inicial de este proyecto se ha decidido utilizar Matlab/Simulink para desarrollar tanto los flujos de trabajo de la Capa 3 como la interfaz de usuario y configuración de la Capa 2, ya que cumple con todas las especificaciones requeridas.

Arquitectura del Sistema

El objetivo principal del IPCS presentado aquí es generar consignas para los controladores locales y el soporte a la decisión. La estructura del sistema completo se muestra en la *Figura 2*, mientras que la arquitectura del sistema de control basada en tres capas, descrita en el sumario de este trabajo, se muestra en la *Figura 4*. La primera capa o capa de minería de datos permite generar modelos a partir de datos del proceso. Se trata de una tarea *offline* en la cual se toman datos históricos de cada sistema de saneamiento, con el fin de generar modelos basados en datos válidos para el control y la supervisión de procesos. Dada la importancia de los datos para el sistema, es necesario un proceso de validación de los mismos, tratando posibles errores como la falta de valores o valores espurios, utilizando métodos como los descritos en (Cugueró-Escofet et al., 2016). Después, con una base de datos preprocesada y en un formato estándar adecuado, se pueden utilizar técnicas de minería de datos para encontrar, por ejemplo, relaciones entre variables o patrones de comportamiento que puedan luego usarse en la tercera capa o capa de control de proceso. Estos modelos basados en datos encontrados en la Capa 1 se pueden utilizar para la diagnosis y control del proceso, generando alarmas y consignas para los actuadores a partir del conocimiento extraído. En la Capa 2 o capa de diseño del proceso se definen los procesos a controlar y supervisar y qué señales están disponibles. Es decir, la Capa 2 permite diseñar cómo será la herramienta de control para cada caso particular, haciendo uso de todas las herramientas disponibles en la Capa 3 y los modelos obtenidos en la Capa 1. La naturaleza gráfica del diseño propuesto permite que el flujo de trabajo implementado en esta capa sea utilizado como interfaz de usuario de la herramienta, añadiendo los parámetros configurables que sean necesarios y mostrando los KPI para la

supervisión del proceso. Finalmente, la Capa 3 o capa de control del proceso es el núcleo de la aplicación. El proceso definido en la Capa 2 es supervisado y controlado utilizando los modelos generados en la Capa 1, mediante los flujos de trabajo construidos en esta capa.

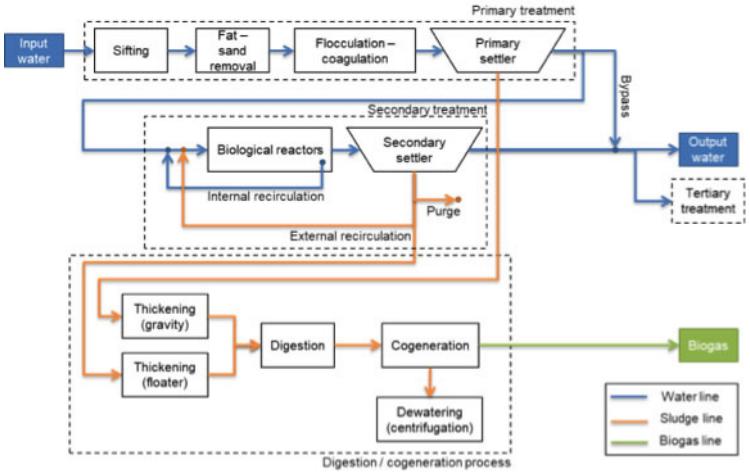
CASO DE ESTUDIO

Descripción

El caso de estudio presentado en este trabajo se encuentra en las instalaciones en el ámbito del CBT. El CBT es una administración local compuesta por 64 municipios y que da servicio a una población de 470.000 habitantes. CBT es responsable de las instalaciones de saneamiento desde la fase de proyecto y construcción hasta la operación y mantenimiento —incluyendo 315 km de colectores y 23 EDAR—, con el objetivo de preservar y mejorar la buena calidad de los ríos en su área de acción —Ripoll, Besos, Riera de Caldes, Tenes, Congost, Mogent y Tordera—. Todas las EDAR en el ámbito de CBT están basadas en la depuración biológica por fangos activos. Su capacidad de tratamiento va desde los 1000 m^3/día hasta los 40000 m^3/día o, expresado en términos de población equivalente, desde pocos centenares hasta 30000 HE. Sin embargo, el esquema general de todas las plantas es similar (Figura 5). El agua residual tanto de áreas urbanas como industriales es conducida a través de un sistema de alcantarillado a una de las depuradoras de la zona. Todas ellas incluyen línea de agua y línea de fangos, y en algunos casos línea de biogás. A pesar de ello, existen particularidades en cada instalación que implican un diseño ad-hoc del sistema de control, como puede ser el número o tipo de sensores y actuadores disponibles, o las características del influente. Un proceso crítico a controlar es el reactor biológico, donde además se da el mayor consumo eléctrico de toda la EDAR debido al proceso de aeración, aunque este consumo se puede llegar a reducir hasta en un 20% optimizando su control. El objetivo de este proceso es la eliminación de la materia orgánica (carbono orgánico) y los nutrientes (nitrógeno y fósforo) del agua residual. Para que este proceso biológico de eliminación sea posible, es necesaria la existencia de oxígeno. El oxígeno se introduce desde el ambiente al reactor biológico mediante el uso de soplantes. El proceso requiere periodos con oxígeno y periodos sin oxígeno, por lo que el control adecuado de las soplantes es clave. Los caudales de recirculación interna y externa también están involucrados en este proceso biológico: mediante la recirculación externa se mantiene un ratio adecuado de fango y microorganismos en el reactor biológico, mientras que con la recirculación interna se controla la concentración de nitratos, generados en el proceso de nitrificación y desnitrificación, por el cual se elimina el nitrógeno y el fósforo. En una primera etapa, el nitrógeno es oxidado a nitratos en presencia de oxígeno (nitrificación). Este proceso lo realizan algunas bacterias autótrofas que consumen oxígeno disuelto en el agua. Después, en la segunda etapa, el nitrato es reducido a nitrógeno en estado gaseoso (desnitrificación) por algunas bacterias heterótrofas. Esta etapa tiene lugar en una situación anóxica, ya que estas bacterias utilizan los nitratos en lugar del oxígeno para consumir el carbono de la materia orgánica.

Adicionalmente, existen otros procesos que pueden ser controlados en una EDAR, como la eliminación química de fósforo o el bypass en situaciones de sobrecarga (por ejemplo, lluvias).

Figura 5. Esquema general de una EDAR.



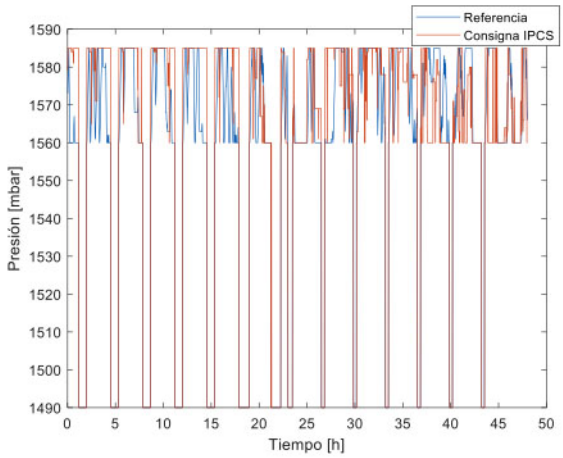
Resultados

La propuesta presentada en este trabajo se ha empezado a desarrollar en fase prototipo en una de las EDAR de CBT. Uno de los procesos que se debe controlar y supervisar en el caso considerado es la aeración del reactor biológico. Esta planta dispone de dos líneas de tratamiento biológico, es decir, dos reactores, con un sistema de aeración formado por una soplante y varias válvulas que regulan la cantidad de oxigeno que entra en cada zona del reactor. Una de las múltiples consignas con las que se gobierna este proceso es la presión del circuito de aire. Mediante esta consigna de presión, se airea (o no) el reactor biológico en función de su valor. Cuando las condiciones de ambos reactores son de desnitrificación, el sistema de control envía una consigna de presión de valor reducido para parar la soplante. Con el fin de desarrollar una metodología genérica, válida para cualquier sistema del que se dispongan históricos de datos, el primer prototipo que se plantea tiene como objetivo calcular esta consigna de presión, en función del estado de la planta.

Para generar los resultados se utilizan los datos históricos de operación de la EDAR de estudio por un periodo de un año, con el fin de disponer de un conjunto de casos representativo que considere e.g. diferencias de operación debidas a comportamientos estacionales en el estado del sistema. La cantidad de variables disponibles, considerando consignas, señales de sensores, y alarmas, entre otras, asciende a más de 300. A pesar de que la metodología propuesta permite trabajar con todos estos datos y decidir de forma automática cuales son más relevantes para cada modelo, en este primer prototipo se realiza una selección basada en el criterio de los expertos de alrededor de 50 variables relacionadas directamente con el proceso biológico, incluyendo medidas de amonio, nitrato u oxígeno, entre otras. En la Figura 6 se muestra el flujo de trabajo implementado en Simulink que se utiliza en este prototipo para generar un modelo basado en datos, así como la consigna de presión mediante los métodos de razonamiento adecuados. En la Figura 7 se muestran los resultados para una prueba del prototipo realizada durante 48 horas.

Figura 6. Flujo de trabajo utilizado implementado en Simulink.

Figura 7. *Comparación entre la consigna de presión generada por el prototipo y la consigna de presión actual*



CONCLUSIÓN

En este trabajo se presenta un marco interoperable para la automatización de la construcción de herramientas de control y supervisión inteligentes. Las bases de esta propuesta son el uso de flujos de trabajo visuales y la interoperabilidad entre las distintas herramientas de minería de datos y los modelos obtenidos. La estructura de la herra-

mienta propuesta, definida en 3 capas, permite separar la obtención de modelos basados en datos de los métodos y técnicas de razonamiento que hacen uso de estos modelos para el control y supervisión de la planta, de forma que el diseño particular para cada sistema sea una tarea sencilla y dependa únicamente de los datos disponibles.

Finalmente, destacar que la metodología desarrollada tiene el potencial de ser fácilmente escalable, tanto para generar otras consignas de la misma EDAR como de otras, si se dan las condiciones y se dispone de la información adecuada. Los flujos de trabajo desarrollados permiten generar modelos para cualquier sistema, así como utilizar estos modelos para controlar y/o supervisar los procesos correspondientes, sin necesidad de más modificaciones que el modelo de datos utilizado, es decir, sin modificar el código de la herramienta.

RECONOCIMIENTOS

Los autores agradecen el soporte en este trabajo del Programa de Doctorado Industrial (2017-DI-006) y de los Grupos/Centros de Investigación Consolidados (2017 SGR 574) por la *Agència de Gestió d'Ajuts Universitaris i de Recerca* (AGAUR) de la *Generalitat de Catalunya*.

REFERENCIAS

Argent, R. M. (2004). *An Overview of Model Integration for Environmental Applications-components, frameworks and semantics.* Environmental Modelling & Software 19, 219-234, 2004.

BMPN (2011). *The Business Process Model and Notation BMPN v2* (http://www.bpmn.org/, http://www.omg.org/spec/BPMN/2.0/) was released in January 2011.

Cugueró-Escofet, Miquel À. et al. (2016). *A Methodology and a Software Tool for Sensor Data Validation/Reconstruction: Application to the Catalonia Regional Water Network. Control Engineering Practice* 49: 159–72. http://linkinghub.elsevier.com/retrieve/pii/S0967066115300459 (September 29, 2016).

Cugueró-Escofet, Miquel À., Joseba Quevedo, Vicenç Puig, and Diego García (2014). *Inconsistent Sensor Data Detection/Correction: Application to Environmental Systems.*" In Proceeding of: IEEE World Congress on Computational Intelligence (IEEE WCCI), Beijing, China, 84-90.

DMG (2014). *The Data Mining Group (http://www.dmg.org) leads the development of the Predictive Model Markup Language (PMML).* Current version PMML 4.2. February 2014.

Elasri, H., Sekkaki, A. (2013). *Semantic Integration process of Business Components to Support Information System Designers.* Int. Journal of Web & Semantic Technologies 4(1), 51-65.

Erl, T. (2004). *Service-Oriented Architecture. A Field Guide to Integrating XML and Web Services*. Prentice-Hall, 2004.

Institute of Electrical and Electronics Engineers (1990). *IEEE Standard Computer Dictionary: a Compilation of IEEE Standard Computer Glossaries.* New York.

jBPM (2017). *The jBoss Business Process Management engine jBPM v7.0* (http://www.jboss.org/jbpm) was released in July 2017.

Johnston, M. W., Hanna, J. R. P., Millar, R. J. (2004). *Advances in dataflow programming languages. ACM Computer. Surveys* 36. 1-34. DOI: 10.1145/1013208.1013209.

Kzaz, L., Elasri, H., Sekkaki, A. (2010). *A Model for Semantic Integration of Business Components.* Int. Journal of Computer Science & Information Technology 2(1):1-12.

Manguinhas, H. (2010). *Achieving Semantic Interoperability using Model descriptions.* Bulletin of IEEE Technical Committee on Digital Libraries, Vol. 6, No. 2, Fall 2010.

Mackay, D. S. (1999). *Semantic Integration of Environmental Models for Application to Global Information Systems and Decision-Making.* ACM SIGMOD Record 28(1):13-19, March 1999.

Morrison, J. P. (2010). *Flow-Based Programming: A new approach to application development*. CreateSpace, 2010.

Ouksel, A. M., Sheth, A. (1999). *Semantic Interoperability in Global Information Systems: a Brief Introduction to the Research Area and the Special Section.* ACM SIGMOD Record 28(1):5-12, March 1999.

Rizzoli, A. E., Davis, J. R., Abel, D. J. (1998). *Model and Data Integration and re-use in Environmental Decision Support Systems.* Decision Support Systems 24:127-144, 1998.

Sànchez-Marrè, M. (2014). *Interoperable Intelligent Environmental Decision Support Systems: a Framework Proposal.* 7th International Congress on Environmental Modelling and Software (iEMSs'2014). iEMSs' 2014 Proceedings, Vol. 1, pp. 501-508.

Sànchez-Marrè, M., Gibert, K., Sojda R., Steyer, J. P., Struss, P., Rodríguez-Roda, I. (2006). *Uncertainty Management, Spatial and Temporal Reasoning and Validation of Intelligent Environmental Decision Support Systems.* 3rd International Congress on Environmental Modelling and Software (iEMSs'2006). iEMSs' 2006 Proceedings, pp. 1352-1377.

Sottara, D., Bragaglia, S., Mello, P., Pulcini, D., Luccarini, L., Giunchi, D. (2012). *Ontologies, Rules, Workflow and Predictive Models: Knowledge Assets for an EDSS. 6th International Congress on Environmental Modelling and Software (iEMSs'2012). iEMSs' 2012 Proceedings*, pp. 204-211.

ter Hofstede, A. H. M, van der Aalst, W. M. P., Adams, M., Russell, N. (2010). *Modern Business Process Automation*. Springer.

Vetere, G., Lenzerini, M. (2005). *Models for Semantic Interoperability in Service-oriented architectures.* IBM Systems Journal 44(4), 887-903, 2005.

Wesseling, C. G., Karssenberg, D., Burrough, P. A., Van Deursen, W. P. A. (1996). *Integrating dynamic environmental models in GIS: the development of a dynamic modelling language.* Transactions in GIS, 1(1), 40-48.

zur Muehlen, M. (2004). *Workflow-based Process Controlling*. Berlin: Logos-Verlag.

CONTACTO

Josep Pascual Pañach
Consorci Besòs Tordera
Avinguda Sant Julià, 241 (08103 Granollers, Barcelona)
600 88 00 38
jpascual@besos-tordera.cat