Interoperating Data-driven and Model-driven Techniques for the Automated Development of Intelligent Environmental Decision Support Systems

Josep Pascual-Pañach[a,b,c], **Miquel Àngel Cugueró-Escofet**[a,b,d], **Miquel Sànchez-Marrè**[c]
[a] *CCB Serveis Mediambientals, SAU*
[b] *Consorci Besòs Tordera*
Av. Sant Julià, 241, 08403 Granollers, Catalonia
Email: jpascual@besos-tordera.cat

[c] *Knowledge Engineering and Machine Learning Group (KEMLG)*
Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC)
Universitat Politècnica de Catalunya (UPC)
Campus Nord, building OMEGA, C. Jordi Girona 1-3, 08034 Barcelona, Catalonia
Email: miquel@cs.upc.edu

[d] *Advanced Control Systems Research Group, Universitat Politècnica de Catalunya (UPC-BarcelonaTech), Terrassa Campus, Gaia Research Bldg. Rambla Sant Nebridi, 22, 08222 Terrassa, Barcelona, Catalonia*
Email: miquel.angel.cuguero@upc.edu

Abstract: This paper proposes an Intelligent Decision Support (IDS) methodology based on the integration of data-driven and model-driven techniques for control, supervision and decision support on environmental systems. Design stage of control and decision support tools for environmental systems tend to be somehow ad-hoc regarding to the nature of the processes involved. Hence, an automated approach is proposed here for the sake of scalability to different types and configurations of environmental systems, and the methodology has been designed in a general fashion to allow scalability to further types of systems. The interoperation of a data-driven technique –Case-Based Reasoning (CBR)– and a model-driven technique –Rule-Based Reasoning (RBR)– is considered in this work. The proposed hybrid scheme provides complementarity and supervised redundancy in the set-point generation for the process controllers and actuators, increasing the reliability of the Intelligent Process Control System (IPCS), which is the core component of the IDS methodology. A Decision module selects which reasoning approach to use –i.e. CBR or RBR– depending on a metric quantifying the confidence in the CBR solution. Furthermore, the IDS methodology is flexible and dynamic enough to be able to cope with the dynamic evolution of environmental systems, learning from its relevant experienced situations. The approach presented has been implemented in a real facility within the ambit of a local water administration in the area of Barcelona.

Keywords: Case-Based Reasoning; Rule-Based Reasoning; Intelligent Environmental Decision Support System; Intelligent Process Control; Data Mining; Wastewater Treatment Plant.

1 INTRODUCTION

1.1 Background

Decision support, supervision, control and optimisation systems can be challenging because of the variability and the potential inherent complexity of the processes involved. These processes may include a wide variety of elements of different types interacting with each other —e.g. mechanical, electronic, human, biological, or chemical—, sometimes with unknown dynamics, that may raise a real challenge for the control, supervision and decision support tasks involved. In the case of environmental systems —e.g. Wastewater Treatment Plants (WWTPs) here—, the risk of dysfunction is even more severe, since the consequences can be dangerous for the environment and for human beings. To avoid this, the quality of certain monitored parameters of the environmental system —e.g. the effluent of the WWTP here— must comply with different applicable environmental regulations, e.g. European, regional, local. Specifically regarding WWTPs, these facilities are in operation 24 hours a day, 365 days a year, and hence, their energy consumption is remarkably high. Using adequate monitoring and intelligent control techniques these energy costs can be minimized, while ensuring a more reliable environmental management of the process, i.e. improving the diagnosis of possible problems providing appropriate solutions. In the field of environmental process monitoring and control systems, Artificial Intelligence (AI) techniques have been shown effective and used for some time to improve the reliability of the supervision of these processes and overcome certain shortcomings of the classic control systems. The set of AI techniques used ranges from Knowledge-Based Systems (Flanagan, 1980; Berthuex et al., 1987; Maeda, 1989; Gall and Patry, 1989; Tzafestas and Ligeza, 1989; Serra et al., 1994; Ahmed et al., 2002; Aulinas et al., 2011; Castillo et al., 2016; Corominas et al., 2018), Fuzzy Control Systems (Czoagala & Rawlik, 1989; Wang et al., 1997; Ruano et al., 2010; Santín et al., 2018; Bernardelli et al., 2020), control using Artificial Neural Networks (Capodaglio et al., 1991; Kosko, 1992; Côte et al., 1995; Syu and Chen, 1998; Hamed et al., 2004; Ráduly et al., 2007), Case-Based Reasoning (Sànchez-Marrè et al., 1997, 2002, 2005) or Genetic Algorithms (Karr, 1991; Béraud et al., 2007) to Intelligent Decision Support Systems (IDSS) (Sànchez-Marrè et al., 2004; Torregrossa et al., 2017; Nadiri et al., 2018; Han et al., 2020). AI methods can provide important improvements to the supervision and control of these processes, such as qualitative information management, expert knowledge modelling, uncertainty modelling and reasoning and learning abilities.

Regarding the water sector, the smart water concept is incipiently emerging and strongly depends on properly address several key challenges, e.g. the formulation of an integrated water information system with standardised ontologies in order to achieve real interoperability (Gourbesville, 2016), as suggested e.g. in the Smart Water Management Initiative introduced in (Choi et al. 2016). As pointed out in (Gourbesville, 2016; Sànchez-Marrè, 2014), several technical challenges are still to be tackled in order to achieve real integration and functional interoperability, needing further efforts in order to reach maturity, particularly in the definition of standards for managing workflows among various applications and models in order to produce real time information useful for decision makers. The lack of interoperability standards in Information and Communication Technology (ICT) systems for water management is also pointed out in (Laxmi and Laxmi-Deepthi, 2017; Robles et al., 2014), jeopardizing proper monitoring, control and overall efficiency of water management and preventing their evolution and improvements e.g. the adoption of Internet of Things (IoT) paradigm. The need for standards in the management of water infrastructures is also pointed out in (Di Biccari and Heigener, 2018) as an essential step for a fully integrated management and for reaching efficient levels of interoperability and communication. (Poch et al., 2017) points out that the construction of a successful Environmental Decision Support System (EDSS) should focus significant efforts on the use and transfer of the tool to the market. In (Mannina et al., 2019), a review of the state of the art in the Decision Support Systems (DSS) for WWTPs is presented. Here, the development of user-friendly applications and the challenge to reach the water market are also emphasized.

On the one hand, experts in this area have acquired over time specific knowledge related to how to successfully solve complex tasks. In order to take advantage of this knowledge, its codification in a specific type of knowledge representation formalism would be very beneficial.

The most used formalism to represent these knowledge patterns is the use of inference rules. These rules constitute the Knowledge Base of these systems, which represent this knowledge provided by the experts. As the knowledge is coded by rules, the reasoning mechanism is named Rule-Based Reasoning (RBR) (Jackson, 1999; Buchanan and Duda 1983). An RBR system contains, at least, the following three components (Figure 1): a) the knowledge base; b) a fact base or data base and; c) the inference engine. The knowledge base is composed of a set of rules that codifies the expert knowledge of a specific domain. The data base contains all the relevant information that is necessary for the rules' evaluation, i.e. historical information or parameters' values. The inference engine is the reasoning system that uses the rules in the knowledge base, the information in the data base and the data from the acquisition system to provide a decision.

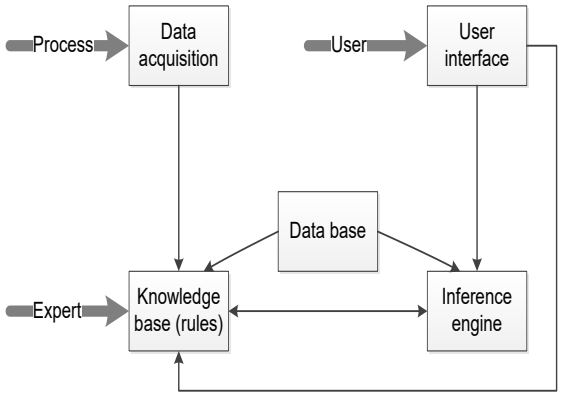

Figure 1. Rule-Based Reasoning system structure

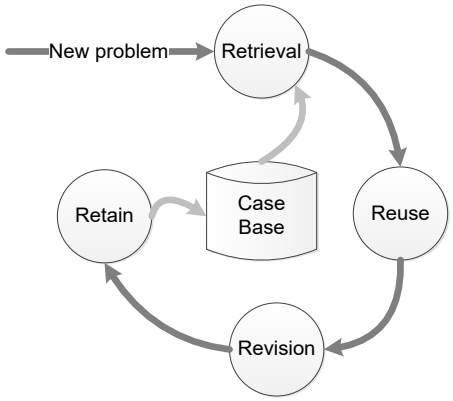

Figure 2. Case-Based Reasoning system structure

On the other hand, the Case-Based Reasoning (CBR) approach (Figure 2) tries to solve new problems in a domain reusing the previous solution provided to a similar problem in the same domain (analogical reasoning). Thus, the solved problems constitute the "knowledge" about the domain. The more experienced is the system, the better is the performance it achieves, since new relevant experiences (i.e. cases or solved problems) are stored in the Case Base (CB) or Case Library. This way, the system is continuously learning to solve new problems (Riesbeck and Schank, 1989; Kolodner, 1993; Richter and Weber, 2013). This behavior is based on the theory of dynamic memory of Roger Schank (Schank, 1982), which states that the human memory is dynamic and change with its experiences along its life. The CBR system structure showed in Figure 2 considers the four stages of the case-based reasoning method: retrieval, reuse, revision and retain (Aamodt and Plaza, 1994). The *retrieval phase* is the process by which similar problems (i.e. cases) to the new problem are searched in the CB. Then, in the *reuse phase*, the solution of the retrieved case is adapted and used to solve the new problem. The *revision phase* is to determine whether the solution found in the reuse stage has been successful or not. Finally, the *retain phase* is the stage where useful information from the new problem-solving episode is learnt into the existing CB.

As a result of research in this field, a commercial IDSS was built in a former stage of this work and used in more than 100 WWTPs around the world. This IDSS was initially based on the methodology described in (Poch et al., 2004). In (Poch et al., 2017) the gap between the research in this field and the water market is pointed out. The IDSS was drastically simplified for its commercial implementation, using only a rule-based component, and did not aim at scalability, dynamic learning and gradual competence increase, interoperation of methods and usability issues, like in the approach presented here. Furthermore, the Intelligent Decision Support (IDS) approach presented here has been designed in a general fashion for the sake of scalability to different types of environmental systems —which pose similar challenges as

the particular environmental system case study considered here, i.e. WWTPs—, but also to further types of systems beyond the environmental framework.

1.2 Overview

A common and important problem when designing a new IDSS for environmental processes is the *ad-hoc* nature of this design for each particular system, depending on its particular specifications, e.g. processes involved (e.g. nitrification, phosphorus removal, major air pollution contaminants, geographic features of a territory), particular configuration of each system or available data. This entails the investment of a large amount of time in the analysis of the different requirements of the new system to design the IDS tool. In line with the improvement of these problems, the use of RBR and CBR methods –and particularly, the interoperation of both–, is proposed here to obtain an intelligent decision support system based on a control and supervision solution that can be easily scaled to different environmental systems —without loss of generality, WWTPs here. The work presented here wants to delve into one of the layers of the framework presented in a previous study (Pascual-Pañach et al., 2018). This generic framework is based on a three-layer architecture for IDSSs deployment. One of these layers is the process control layer, where the integration of RBR and CBR approaches within the Intelligent Control Process System (ICPS) is proposed to tackle the set-points generation problem to manage the environmental system. The mainstay of this proposal is to guarantee the interoperability between the different layers and methods used, in addition to guaranteeing the scalability of the approach, as well as the high reliability and dynamic flexibility for learning from past experience environmental situations through the CBR component. This last feature makes the IDS methodology able to escape from static solution architectures which are not able to adapt to dynamic changes in the evolution of environmental systems.

The objectives of this work are, first: to propose a reliable IDS methodology and to deploy an IDSS, and its main IPCS component based on the interoperation of RBR and CBR methods in a real WWTP, as an example of a real environmental system where the proposed methodology can be applied; second: to integrate this IDS methodology in a user-friendly graphical user interface (GUI) to help in the participatory role of practitioners in the daily operation and decision-making of the process, and finally: to demonstrate how this approach can be easily scaled to different installations.

The structure of this paper is as follows: in Section 2, the methodology is presented; first, a general flowchart of the application operation is shown; then, RBR and CBR modules implementation is detailed. In Section 3, the experimental evaluation is detailed. In the first part of Section 3, the case study is described. Then, in the second part of Section 3, the methodology is validated with real data obtained from the operation of the application in a real system. The results are discussed in Section 4. Finally, some conclusions and future work are presented in Section 5.

2 METHODOLOGY

This section describes the IDS methodology used to combine both data-driven and model-driven techniques, and its integration in the core component, which is the IPCS, deployed in a real environmental system —without loss of generality, a WWTP here. In Section 2.1, the general architecture of the whole IDSS, with its major component, *i.e.*, the IPCS, presented in a previous work (Pascual-Pañach et al., 2018), is described. Next subsections describe in detail the interoperation of both reasoning methods. In Section 2.2 the control workflow is presented. Section 2.3 describes the reasoning system. In Section 2.4 how both reasoning methods –CBR and RBR– are interoperating is detailed.

2.1 IDSS and IPCS design

The main aim of the IDSS and the core IPCS component presented here is to generate the set-points for the local controllers to preserve the environmental quality of each environmental process installation (Figure 3), here a sanitation system. Figure 3 shows the integration of the IDSS and the IPCS in the current architecture of the system. This tool reads online sensor measurements from the plant and generates set-points for a lower control level Programmable Logic Controller (PLC). The standard control system available in most WWTPs consists of the combination of a Supervisory Control And Data Acquisition (SCADA) system and a PLC. The SCADA is a software system used to control, monitor and acquire data from the WWTP, while the PLC is a modular industrial computer that provides multiple inputs and outputs and contains the control loops programming. Traditionally, SCADA/PLC systems integrate classical control approaches, e.g. Proportional Integral Derivative (PID) controllers. The IPCS proposed here is based on AI techniques, with the aim of providing a scalable solution to different installations. Here both systems –i.e. SCADA/PLC and designed IPCS– are working together. The IPCS does not control all the processes in the plant –i.e. is focused on the secondary treatment– while the SCADA system is used as a backup solution in case of failure of the IPCS. This has been the design of choice since WWTPs are complex and critical installations that must be controlled and supervised 24 hours a day 7 days a week.

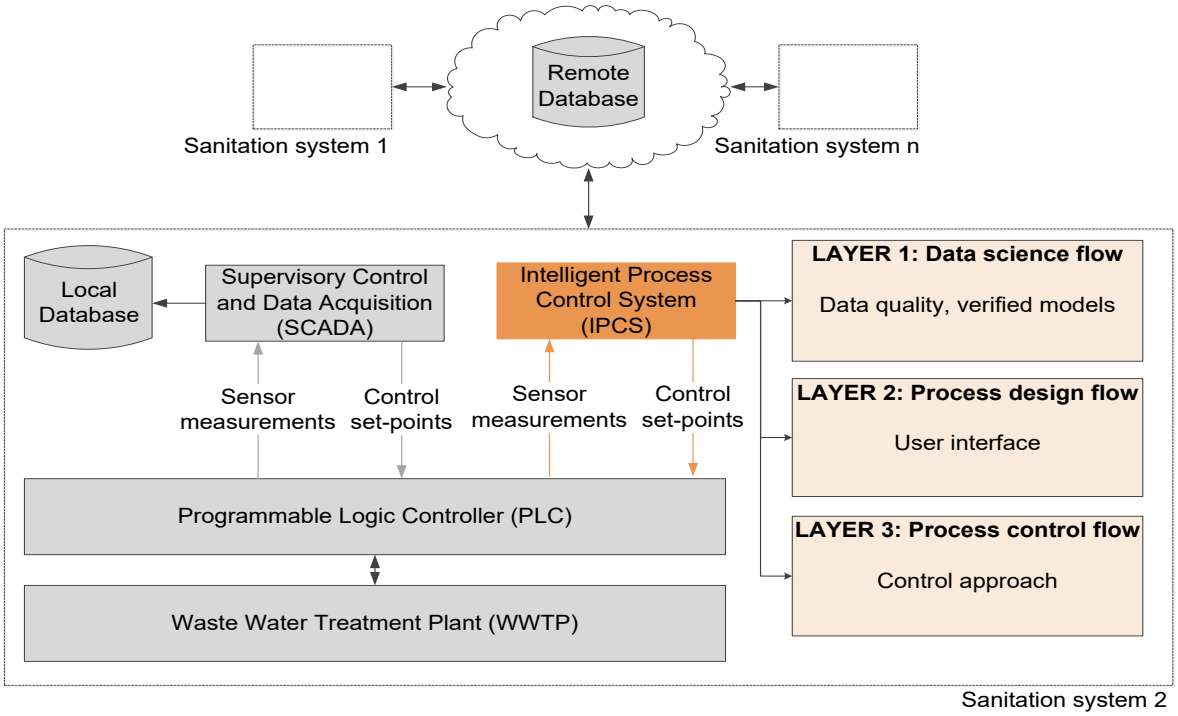

Figure 3. Architecture of the proposed Intelligent Decision Support Methodology

The IDSS, along with its IPCS core component proposal is based on a three-layer architecture (Figure 4). The data science flow layer (i.e. Layer 1) is used to generate models obtained from process data. It is an off-line procedure that takes historical available data from each system with the aim of generating valid data-driven models to supervise and control the process. The input of Layer 1 is a standardized and properly formatted raw database containing all available data for each system, namely: sensor measurements, equipment states and alarms, plant set-points and further data derived from them. First, different data validation and reconstruction methods, such as the ones in (Cugueró-Escofet et al., 2016; Gibert et al., 2010, 2018; De Mulder et al.,2018), can be applied to obtain a new filtered and valid database. Then, some of

the proposed data mining methods can be applied to the complete dataset to find relations among variables, behavioural patterns or similar methods to obtain valid models to be used in Layer 3. These models can be e.g. rule models induced from decision trees or case bases. Both types of models interoperate in Layer 3 to supervise the system by discriminating abnormal situations from normal operation, and also to control the process by generating actuator set-points based on knowledge obtained from data gathered from the process. Rule models can also include human expert knowledge of the system (model-driven method), so a user-friendly interface to integrate such human-based knowledge is considered for this layer. In the Process Design flow layer (i.e. Layer 2), the layout of the plant is designed, including all processes to be supervised and controlled, and the corresponding signals. Finally, the Process Control workflow layer (i.e. Layer 3) is the application core: the plant defined in Layer 2 is supervised and controlled using models generated in Layer 1, with the workflow designed in this Layer 3. Therefore, the online Layer 3 is directly connected with both previous offline layers.

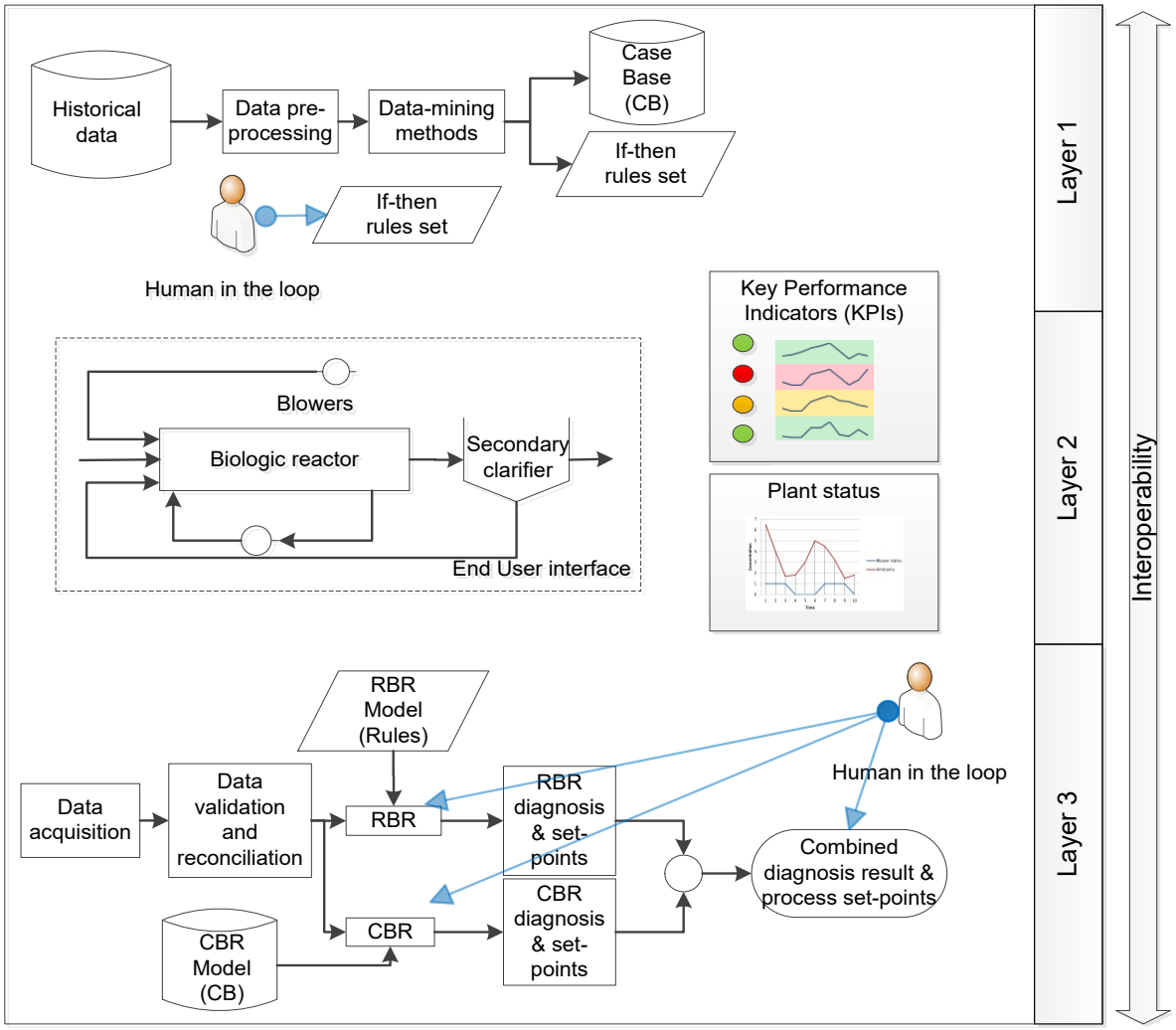

Figure 4 IDSS architecture

Hence, the proposed approach is to interoperate CBR and RBR methods, obtaining redundancy in diagnosis and/or set-points generation. CBR and RBR modules inputs are correctly fed with online data gathered from the process and the corresponding models generated offline in Layer 1. Using both methods and different models, diagnosis results and set-points can be compared in order to provide a more reliable diagnosis and set-point generation. Hence, this diagnosis and set-point generation redundancy helps on relying on

the outcome of the tool presented. Also, the human expert knowledge –provided e.g. by the plant manager– is considered in order to validate the tool outcome and also to feed the database with human-based knowledge. In the next sections, the importance of the automation of the whole process, i.e. from data acquisition to diagnosis, is emphasized. However, some situations may require the user (i.e. human in the loop) to validate the tool outputs.

In order to implement the proposed methodology and develop this tool, the use of a visual workflow is proposed. To this end, the use of graphical programming environments provides some advantages in relation to traditional languages like C or Java (Johnston et al., 2004), e.g. reusability and understandability of the code, modularity and flexibility, intrinsic parallelism, easy debugging or faster prototyping and development.
To choose a valid developing environment is necessary to define the desired specifications. Methods and algorithms needed are all related to data mining. In Gibert et al. (2010), an overview of different data mining techniques and choosing criteria is presented. Most programming languages have available libraries for data science, for example Scikit-learn for Python, or JDMP (Java Data Mining Package) for Java. Although they are not designed for graphical programming, we can find the Flow-Based Programming (FBP) paradigm described in Morrison (2010), which allows the programmer to create applications as a set of black boxes –or interconnected processes–, or some programming environments like NoFlo for JavaScript, based on the FBP concept. On the other hand, there are some programming environments and languages that make the development process easier because are oriented to graphical programming, e.g. Matlab/Simulink (Champman, Stephen J., 2020) or LabVIEW (Johnson and Jennings, 2006), or further open source equivalent options like Scilab (Nagar, S., 2017) or MyOpenLab (Ruiz Gutierrez, J. M, 2017), respectively. These environments also have available data science libraries, as well as other specialized useful tools, e.g. database connection and reading, or data acquisition, among others, so they are a convenient choice for implementation. In addition, they can be complemented with libraries from other programming languages –like C or Java– by creating new user defined blocks or tools, or using developed ones, e.g. Drools (Salantino, M., et al., 2016), a rule inference engine developed in Java.
At the current stage, the software used for prototyping the methodology and the tool presented here is Matlab-Simulink. This software provides all the necessary tools, a fourth-generation programming language (4GL) and a graphical programming environment that facilitates the standardization, allowing the tool to be easily reused in different installations.

2.2 Decision support workflow proposal

This work proposes an IDS methodology based on the interoperation of RBR and CBR modules to tackle the supervision of environmental systems, and here particularly WWTPs processes, whilst avoiding the use of ad-hoc solutions for each particular system. In this section RBR and CBR modules are described, as well as how they are integrated in a unique tool combining both methods in order to obtain a more reliable solution. In addition, the operational conditions in real applications are seldom ideal, with e.g. missing information due to bad quality measurements or non-existing monitored data. Authors want to emphasize that the methodology and the tool presented here are being developed to satisfy the different environmental supervision and decision support needs. Although all these systems, in our case WWTPs plants, are based in the same processes, the operational conditions in each installation are quite different, ranging from high levels of automation to manual control approaches. The RBR module allows the implementation of this tool in systems where historical data is of poor-quality, or even non-existent, as well as to reuse generic rules outcomes in different systems, i.e. the CB obtained for a specific system can be reused in systems with similar configuration. The CBR component provides the IDS methodology with learning capabilities. Due to the dynamic nature of CBR, the IDSS can increase its

competence skills along time, because it can learn from relevant environmental situations experienced, solved and learned facing the environmental system supervision, day after day.

The integration of both modules is shown in Figure 5. The grey part of the diagram corresponds to the classic CBR cycle, whilst in orange our proposal integrating the RBR module and the Decision module to the CBR is represented. The Decision module –after rules evaluation and case-based reasoning retrieval phase–, is used to decide which solutions are selected to be applied to the process. Then, at the revision stage, the plant Key Performance Indicators (KPIs) are evaluated for the model used –RBR or CBR– and checked within the allowed limits. Finally, in the retain phase, relevant information can be added to the case base.

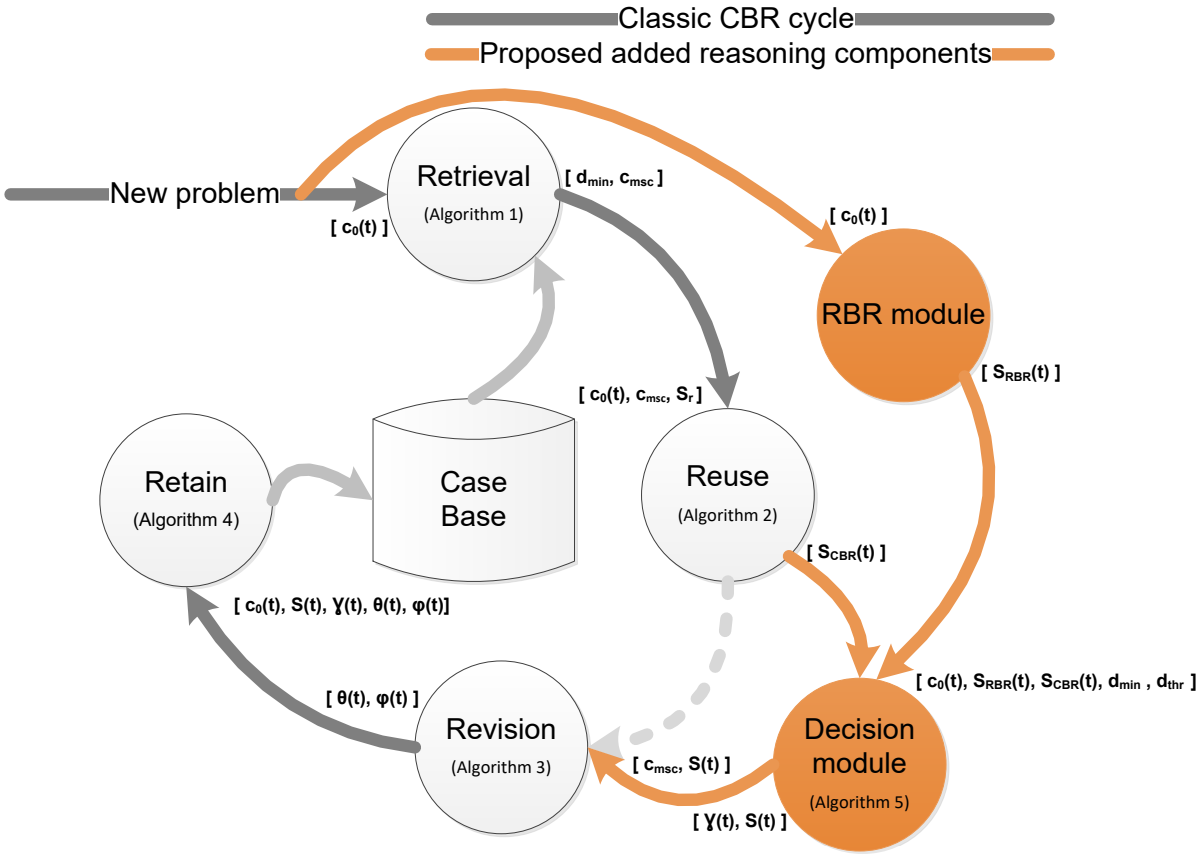

Figure 5 Interoperation of RBR, CBR and Decision modules scheme

The retrieval, reuse, revision and retain phases, as well as the Decision module and the interoperation phase, are described in the next subsections.

2.3 Reasoning system

The reasoning system presented in Figure 5 is composed of the classical CBR cycle and a proposed integration of the RBR module. The RBR module is designed following the scheme presented in Figure 1. The knowledge base consists of a set of rules to generate the set-points to control the environmental process for which they were designed. Each rule is expressed as in (1) as follows:

$$\boldsymbol{If} < condition > \boldsymbol{then} < action > \quad (1)$$

The *condition* statement in (1) depends on any measured/calculated variable/parameter related with the process that can be modified by the user's tool. It can be a simple statement

or a combination of different conditions, by means of logical {AND, OR} operations, as detailed in Table 1 and Table 2.

Table 1 AND operator

<condition1>	<condition2>	<condition1> AND <condition2>
false	false	false
false	true	false
true	false	false
true	true	true

Table 2 OR operator

<condition1>	<condition2>	<condition1> OR <condition2>
false	false	false
false	true	true
true	false	true
true	true	true

The *action* statement in (1) is related to each set-point of the process control. An action can involve setting a set-point to a specific value, or to increase/decrease the current set-point.

The set of rules of the form in (1) is designed together with the concrete system manager and experts on the process, *i.e.*, in a participatory task. All the parameters involved in the rules can be modified online and in real time by the user, even the rules themselves. According to the RBR scheme in Figure 1, the application integrates the inference engine, but not the set of rules or the database. The knowledge base and the data base blocks are out of the tool, which simplifies the implementation in different systems. In Section 3.1, the application to the pilot WWTP is described.

On the other hand, the CBR module consists of a CB obtained from historical operational data of the environmental process. First, the retrieval process is introduced in Algorithm 1, where $c_0(t)$ is the current case, c_{msc} is the retrieved case (the most similar case) and d_{min} is the distance between $c_0(t)$ and c_{msc}.

Algorithm 1 - Retrieval process

function *retrievalFcn*($c_0(t)$)
  for $i = 1$ *to* I
    $d(i) = computeDistances(c_0(t), c_i)$
  end for
  $(d_{min}, c_{msc}) = \text{identMostSimilar}(d)$
  return d_{min}, c_{msc}
end function

In the *retrieval* process, the current environmental situation is compared with stored cases in the CB. Each time new data is read by this process, they are formatted as a new case c_0, following the format of cases in the CB as in (2),

$$c_i = (f_{i1}, f_{i2}, \dots f_{iN}, s_{i1}, s_{i2}, \dots, s_{iM}); i = 1 \dots I \quad (2)$$

where f are the descriptive features, N is the number of features, s are solutions, M is the number of solutions and I is the number of cases in the CB in (3),

$$CB = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_I \end{bmatrix} \quad (3)$$

The retrieval process includes the normalization between 0 and 1, considering the range for each variable.
To solve a new problem using past experiences, it is necessary to find similar situations that have been already solved. Thus, in order to find the most similar case in the CB, the Euclidean distance in (4) is used. This is a convenient measure of similarity here since all the variables are numeric, as pointed out in (Núñez et al., 2004):

$$d(c_a, c_b) = \sqrt{\sum_{n=1}^{N} w_k (f_{an} - f_{bn})^2} \quad (4)$$

where N is the number of features, c_a and c_b are two cases a and b, respectively, and w_n is the weight of the feature n. By default, w_n is $\frac{1}{N}$ for all n, thereby, all features have the same importance.
At the current stage, the most similar case is picked, although other alternatives could be considered, e.g. the *k* most similar cases (with k being a positive integer).

The second stage –reuse process in Figure 5– is introduced in Algorithm 2. In the *reuse* process, the solution obtained in the *retrieval* process can be adapted to the new problem requirements. Since a new case may not be exactly the same as a retrieved case, the appropriate solution may not be the same either. Hence, a method must be used to adapt the retrieved case $\boldsymbol{c_{msc}}$. At the current stage, the method used to adapt the solution is the null adaptation, i.e. no action is performed to the retrieved case and this actual retrieved solution is used. However, in real systems it is necessary to consider situations where actuators used to reach these set-points may be unavailable, or its operation range limited, e.g. a valve which cannot be opened or closed, a certain blowing power is not available from the installed blower and the addition of further elements is needed. Thus, solutions –set-points in our case– obtained from CBR retrieval phase can be adapted to these abnormal situations.

Algorithm 2 - Reuse process

function $reuseFcn(c_0(t), c_{msc}, S_r)$
  for $m = 1 : M$
    if $S_{CBR}(t,m) \in S_r(k)$ *then*
      $S_{CBR}(t,m) = S_{msc}(t,m)$
    else
      $S_{CBR}(t,m) \leftarrow value\ within\ the\ limits\ of\ [s_m^{min}, s_m^{max}]$
    end if
  end for
  return $S_{CBR}(t)$
end function

S_r is the matrix in (5) with the range of valid values for each solution, where (s_M^{min}, s_M^{max}) are the minimum and the maximum values for the solution M,

$$S_r = \left((s_1^{min}, s_1^{max}), (s_2^{min}, s_2^{max}), \ldots (s_M^{min}, s_M^{max})\right) \quad (5)$$

$S_{CBR}(t,m)$ is the adapted solution obtained in the reuse process at time step t for the solution variable m, while $S_{msc}(t,m)$ is the solution of the most similar case, i.e. before the reuse process.

The *revision* process is detailed in Algorithm 3 and executed after the *decision* process and the *reuse* process.

Algorithm 3 - Revision process

```
function revisionFcn(c_msc, S_CBR(t))
    if t = 0 then
        [nuses, okuse, nokuse] = InitializeUtilityMeasures
    endif
    kpi(t) = CalculateKpiValues
    θ(t) = ObtainRevision(kpi_1..Q(t), δ_1..Q)
    if θ(t) is 1 then
        IncreaseOkuse(msc)
    else
        expert is notified and requiered for validation → φ(t) = 1
    end if
    If φ(t) = 1 // Revision θ(t) is 0
        When θ(t) is available then //Reviewed by the user
            If θ(t) is 0 then
                increase nokuses(msc)
            elseif θ(t) is 1 then
                increase okuses(msc)
            end if
            φ(t) = 0
        end when
    return θ(t), φ(t)
end function
```

This process is based on a set of KPIs that can be defined depending on the environmental application and on the environmental issues to be preserved. In the case of sanitation systems KPIs are related to the water quality and the treatment cost, mainly due to electrical consumption and reagents consumption. In Section 3.2 the KPIs used in the revision process for the case study described in Section 3.1 are detailed. In Algorithm 3 KPIs are evaluated in the function *CalculateKpiValues*. Also, different performance measures can be used for assessment and CB maintenance purposes, namely: the total number of usages per case (*nuses*), the number of incorrect usages per case (*nokuse)* and the number of correct uses per case (*okuse*).
The evaluation of the revision θ at each time step t is done in function *ObtainRevision* as described in equation (6),

$$\theta(t) = \begin{cases} 1, & if \bigwedge_{k=1}^{Q} kpi_q(t) \leq \delta_q \\ 0, & otherwise \end{cases} \quad (6)$$

where Q is the number of KPIs, $kpi_q(t)$ is the value of the q^{th} KPI at time step *t* and δ_q is the threshold corresponding to the q^{th} KPI. The δ_q threshold is fixed depending on the nature of the KPI, e.g. the stablished limits in the corresponding waste water treatment directive when it is related to effluent quality.

The effects of a certain actuation depend on the process dynamics and on different boundary conditions like for instance, the waste water charge, e.g. same actuations with different influent characterization have different effects on the process. At this stage, the KPIs designed for each environmental process are considered to detect when the system is not working as it is expected, and hence the revision process is not passed. In the latter case, i.e. $\theta(t) = 0$, the expert can verify whether the last actuations –e.g. the last few hours– are correct or not and update the revision result $\theta(t)$. If the non-fulfillment of the KPIs is caused by an *exceptional situation,* e.g. contaminant discharge over allowed limits, the expert can modify some parameters of the process to be adapted to that environmental situation, or just be aware that the environmental situation is happening, e.g. if there is already the best actuation applied to the system. On the other hand, if an *incorrect actuation* is detected, rules can be reviewed (when the solution is provided by the RBR module) or the CB analysed (when the solution is provided by the CBR module).

In the *retain process* (Algorithm 4), relevant environmental situations that are not represented in the CB can be learned and aggregated to the CB to be used in the future increasing the competence of the IDSS along time.

Algorithm 4 – Retain process

function retainFcn($c_0(t), S(t), \gamma(t), \theta(t), \varphi(t)$)
  if $\gamma(t)$ *is 1 then* //RBR is considered, $c_0(t)$ is candidate
    if $\theta(t)$ *is 1 then*
      $CB(l+1) = c_0(t)$; $l = l+1$;
    elseif $\theta(t)$ *is not 1 and* $\varphi(t) = 1$ *then*
      wait for revision
    else
      $c_0(t)$ *is not retained*
    end if
  elseif $\gamma(t)$ *is not 1 then*
    $c_0(t)$ *is not a candidate*
  end if
  return CB
end function

The identification of a candidate is based on the Decision module (Algorithm 5): when the Decision module outcome –RBR vs. CBR– is that the CBR module cannot solve the environmental problem –hence, RBR is considered–, the decision flag $\gamma(t)$ is set to 1 and the new case is a candidate to be added in the CB. When the revision result $\theta(t)$ is pending (flag $\varphi(t) = 1$) because expert's validation is required, the retain process have to be postponed. Additional details on the Decision module are given in Section 2.4.

It is worth noting that a case already validated and stored in the CB can provide incorrect solutions to a given situation, i.e. the solution provided for that case is not the one needed to tackle the actual situation occurring. Thus, it is necessary to consider a methodology to *remove* or *update* existing cases in the CB. Hence, the tool provides the possibility to remove or modify a case solution if it is proved wrong at the revision stage.
At this stage, special attention should be given to the cases *retention* to avoid an information overload. The information contained in the CB should represent all the possible situations involving the process, while storing the minimum number of cases (i.e. rows in the CB). A large CB involves more resources in terms of physical memory to allocate it and computation time in the retrieval stage. Therefore, the maintenance of the CB is another important point to be considered. *Performance measures* in Algorithm 3 can be useful for this purpose, e.g. unused cases during a certain period can be deleted.

.

2.4 Interoperation of RBR and CBR reasoning modules

The RBR and CBR modules are interoperating through the Decision module. The CBR approach provides more specific knowledge and learning capacity in comparison to the RBR approach. For this reason, the CBR module is more reliable when the computed dissimilarity measure for a new case is below a threshold. In spite of this decision can be automated, the critical nature of this application makes it essential to involve the participation of the user in different decisions along the management of the environmental process, like the *revision* and *retain* stages of the CBR cycle or the *validation* of the Decision module. At each time step the solutions (set-points to control the process in our case) are generated. The dissimilarity measure (d_{min}) of the retrieved case is used to determine the solution reliability. Here, the dissimilarity (d_{min}) of the current case to the retrieved one is compared to d_{thr} calculated in (7) and (9) to make a decision on the solution to be used. The distance threshold is statistically obtained: assuming a case base CB, a set of P experiences or cases that can be solved with CB and taking the minimum distance of each case to the most similar one, the distance threshold d_{thr} is obtained in (9) as follows. First, the average μ_d of all minimum distances is calculated in (7)

$$\mu_d = \frac{\sum_{p=1}^{P} d(c_p, c_{msc})}{P} \tag{7}$$

where $d(c_p, c_{msc})$ is the distance of the case c_p to its most similar case in the CB, i.e. the minimum distance, and P is the total number of solved cases. Then, the standard deviation of all minimum distances is calculated in (8):

$$\sigma_d = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(d_p - \mu_d)^2} \tag{8}$$

where d_p is $d(c_p, c_{msc})$. Finally, the distance threshold is calculated in (9) and used as shown in Algorithm 5 to determine which solution has to be used ($S_{RBR}(t)$ or $S_{CBR}(t)$),

$$d_{thr} = \mu_d + 3 \cdot \sigma_d \tag{9}$$

Algorithm 5 – Decision module

function $decision(c_i(t), S_{RBR}(t), S_{CBR}(t), d_{min}, d_{thr})$
  if $d_{min} \leq d_{thr}$ *then*
    $CBR\ is\ reliable \rightarrow S(t) = S_{CBR}(t)$
    $\gamma(t) = 0$
  else
    $S(t) = S_{RBR}(t)$
    $\gamma(t) = 1$
  end if
  return $\gamma(t), S(t)$
end function

Within a distance smaller than d_{thr}, it is assumed that the solution given by the CBR module can be used because the current situation is enough similar to a situation occurred in the past in the environmental system and already stored in the CB. On the other hand, a distance value over d_{thr} means that the current environmental situation is not similar enough to any stored case in the CB. d_{thr} can also be changed by the user. This value can be increased or

decreased depending on the confidence on the CBR module. Using the extreme values, it is possible to cancel one of both modules. With $d_{thr} = 0$ RBR module is used, while with $d_{thr} = 1$ CBR module is employed.

3 EXPERIMENTAL WORK

3.1 Case study

The current study, as a particular instance of an environmental system where the proposed IDS methodology can be applied, is developed in the framework of a real sanitation system regional network in Catalonia. Such system is managed by *Consorci Besòs Tordera* (CBT), a local water administration composed of 69 municipalities in four different regions of Catalonia with a population of about 470000 inhabitants. CBT is responsible for the sanitation facilities from the very beginning in project and building stages to the final facilities operation and maintenance –including 300 km of sewers and 27 WWTPs–, with the main objective of preserving and improving the good health of the rivers in its area. All WWTPs within the CBT ambit are based on the activated sludge process. Plants capacity ranges from 1000 m^3/day to 40000 m^3/day, including water and sludge lines, and in some cases, a biogas line. Despite the similar layout among CBT WWTPs, there are some particularities that imply a custom-made control system, e.g. number and type of actuators and sensors or influent characteristics. The data-driven strategy component presented in this work may provide a convenient approach to solve these particular characteristics.

One of the most important processes to be controlled and supervised in a WWTP, in order to preserve the good quality of water with a reliable treatment system, is the aeration of the biological reactor in the activated sludge treatment, since is the most critical water quality preservation and resource consuming process in these facilities, accounting for about the 50 % of the overall treatment process energy use of the WWTP (L. Feng et al., 2012; R. Oulebsir et al., 2019). The aim of this process is to supply oxygen to remove organic matter and nutrients, mainly nitrogen, from the sewage water. Nitrogen removal requires aerobic conditions for the nitrification stage –where autotrophic bacteria provide biological oxidation of ammonia to nitrate–, and anoxic conditions for the denitrification stage –where heterotrophic bacteria provide biological reduction of nitrates to produce free contaminant gaseous nitrogen–. The oxygen required for the nitrification stage is provided by means of aeration blowers.
The pilot plant considered in this work is focused on the control and supervision of the biological process of *Santa Maria de Palautordera* WWTP, in the area of the *Tordera* River. Its design capacity is 3225 m^3/day. The pollution load comes mainly from urban wastewater and it is of about 18000 population equivalent. The water line is composed of a primary treatment, two biological reactors and two secondary clarifiers. The sludge line includes thickening and dewatering processes. Currently, the plant is operated with the primary treatment, one biological reactor and two secondary clarifiers, and it is treating an input flow of about 85 m^3/hour with the characterization described in Table 3, obtained from operational data of period 2019.

Table 3. Influent characterization for the pilot WWTP

Parameter	Units	Concentration
Suspended Solids (SS)	mg/l	136
Chemical Oxygen Demand (COD)	mg/l	528
Biological Oxygen Demand (BOD)	mg/l	294
Nitrogen (N)	mg/l	58
Ammonia (NH_4)	mg/l	42
Nitrate (NO_3)	mg/l	0.7
Phosphorus (P)	mg/l	6.3

Conductivity	µS/cm	829
pH	-	8

The aeration system consists of three blowers: the main blower and two additional backups. With the current configuration, the main blower can be combined with one backup blower to reach the desired oxygen concentration in the biological reactor. Backup blowers' operation is combined in order to balance operating hours of each element. Each blower has an integrated frequency converter, so the air flow introduced in the biological reactor can be controlled using this element. The air is introduced in the reactor in two different opposite points. Each air inlet is equipped with a solenoid-controlled valve, which is currently used as an on-off valve. At the current stage, these specifications are considered. Thus, the IDSS through the IPCS proposed commands the *nitrification* (i.e. turning on the blowers and opening both valves) and *denitrification* stages (i.e. turning off the blowers and closing the valves), as well as the oxygen and pressure set-points. These set-points are used by a Proportional Integral Derivative (PID) controller for blowers' speed regulation. Figure 6 shows the basin in the area of study. Figure 7 and Table 4 show all the available measured variables in the pilot plant.

Figure 6 Basin of study

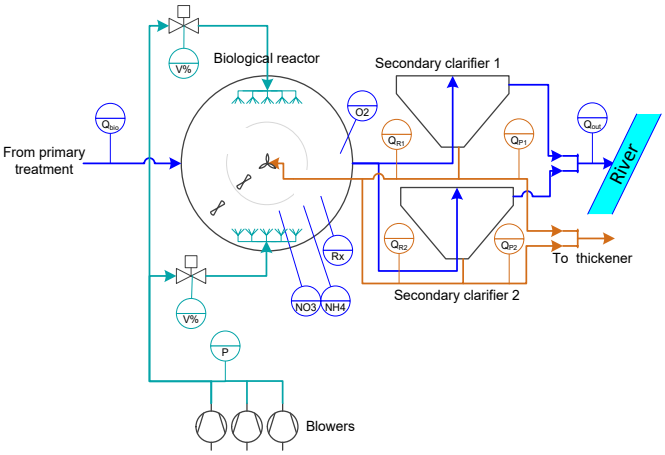

Figure 7 Pilot WWTP layout

Table 4. Available sensors for the pilot WWTP

Sensor	**Units**	**Sensor Id.**
Plant input flow	m^3/h	Q_{bio}
Plant output flow	m^3/h	Q_{out}
Sludge recirculation flow	m^3/h	Q_{R1}, Q_{R2}
Purge flow	m^3/h	Q_{P1}, Q_{P2}
Ammonia concentration	mg/l	NH4
Nitrate concentration	mg/l	NO3
Dissolved oxygen	mg/l	O2
Redox	mV	Rx
Air pressure sensor	mbar	P
Air valves position	%	V%

The former control system operating in the pilot WWTP (i.e. before the integration of the IDSS presented here) was based on the redox measurement to regulate nitrification and denitrification phases, combined with open-loop fixed timers for nitrification and denitrification, set by the operator experience. The latter approach does not make use of all the available monitored information contained in the historical data and may be inefficient in terms of resources (especially energy) use, whereas the interoperation of RBR and CBR techniques

allow considering experts' knowledge –embedded in the RBR rules–, together with further knowledge that can be also retrieved from historical operational records.

3.2 Experimental evaluation

The IDS approach presented in Section 2 is tested in the WWTP described in Section 3.1. The previous control system was based on a combined open-loop and closed-loop scheme using redox measurements. The plant sensorization was improved with ammonia and nitrate sensors in the biological reactor for a better control of the nitrification and denitrification phases. Thus, the IDSS deployed in this WWTP can include data from these new sensors. Historical available data from the process is not valid because it does not contain ammonia and nitrate measurements. The steps taken to test the methodology presented in Section 2 are:

1- Initialization of the reasoning scheme with the RBR module. In situations where the historical data is inexistent or not valid, the reasoning system can be initialized with the RBR module and an empty CB. A set of expert rules is designed with the plant manager to determine the control of the nitrification and denitrification process.
2- The CBR module is initialized with an empty CB and cases are learned following the Algorithm 4. The distance threshold (d_{thr}) defined in (7) should be initialized to a value between 0 and 1. The higher the threshold, the greater can be the difference between the cases added to the case base, i.e. the more heterogeneous they can be. In the particular case explained here, the use of the RBR module is forced during some weeks to validate the control of the process using the new available sensors – i.e. the threshold d_{thr} is set to 0 and all cases are learned.
3- Finally, after the validation of the RBR performance by the plant manager, the tool is reinitialized with both modules, RBR and CBR. The distance threshold is recalculated as explained in Section 2.4. At this stage, the IDSS is working with the complete reasoning scheme in Figure 5. The distance threshold (d_{thr}) is used to decide between RBR and CBR solutions. The amount of cases learned is analysed and a valid CB is obtained after a clustering process of the learned cases, using a selection of some representative cases of each cluster.

The knowledge base of the RBR module consists on a set of rules designed considering the expertise of the environmental system manager –plant manager in this case– in the process. In this particular case two rules with several statements are considered, following the (1). The first rule is activated when any of the conditions for the nitrification stage are satisfied. Therefore, if any of the conditions of the second rule are satisfied, the process is in the denitrification stage. The variables considered are the ammonia concentration (NH_4 [mg/l]), the nitrate concentration (NO_3 [mg/l]) and the 24h moving average (MA) of the ammonia (NH_4^{24h} [mg/l]). The RBR module is executed every minute to generate the corresponding setpoint. In a lower control layer pressure and oxygen set-points determine the operation velocity of the blowers. The plant has been operated for a long time with constant pressure and oxygen set-points. During the test phase of the RBR module it has been observed that the margin of optimisation of these set-points is quite limited because of the process requirements, which force blowers to operate at maximum power most of the time. In the future work section some improvements to this situation are proposed to be considered in further steps.
At each time step, data from the process in Table 4, as well as the set-points generated, are saved in a postgreSQL database (Hans-Jürgen Schönig, 2018). The postgreSQL database is also used to store the configuration of the IDSS, e.g. rules, the CB and the process parameters.

After two weeks of operation with the RBR module, the plant operation is validated by experts and the data stored used to generate a valid case base for the CBR module. It is known that the biological process behaviour depends on different boundary conditions, e.g. temperature. Thus, to obtain a CB including all the possible operation situations, one year of data is needed in order to consider at least a complete season of operation. Considering this dataset, data is validated, e.g. values within the limits, missing values are omitted. Then, a clustering method is used to find patterns and identifying different situations. The aim of this process is to create classes by grouping similar cases. These classes can be labelled by an expert and used to select some representative cases to reduce the size of the case base in order to avoid redundant information. Figure 8 shows a graphical representation of the clustering result obtained with real data from the biological process. In this latter case, four different classes are identified, namely: the nitrification stage; the denitrification stage; the end of nitrification stage —i.e. when it is considered completed—and; the end of denitrification stage.

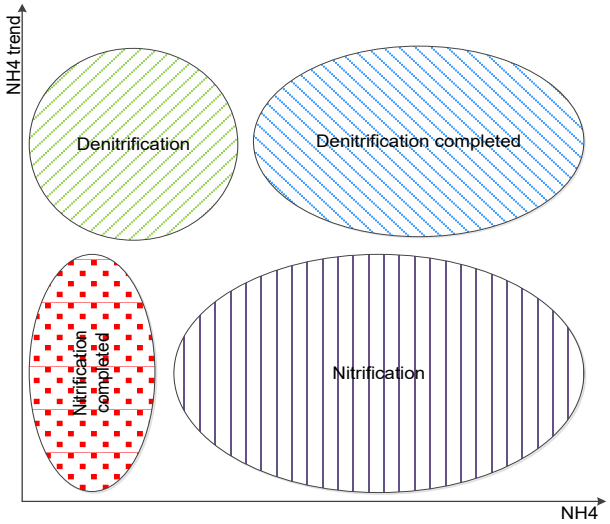

Figure 8 Clustering result obtained with real WWTP data

To evaluate the performance of the CB created from these data, a selection of cases from each class is proposed. From about twenty thousand cases obtained in two weeks operation, ten cases from each class are selected randomly. So, the initial CB consists of 40 cases.

The KPIs defined and used to supervise the performance of the biological process are related to the outflow quality and the WWTP efficiency. They are standard indicators that can be used in any WWTP to supervise the performance of the nitrification-denitrification process. At the current stage these indicators, designed together with the plant manager, are:

- *24h moving average (MA) of ammonia concentration (24MA-AC)*: The 24h MA of ammonia concentration in the effluent is stablished by applicable regulations to a maximum value of 4 mg/l.
- *Blower electrical consumption (BEC)*: The daily average consumption is calculated with historical data and used as a threshold to be compared with the current daily average consumption.
- *Total nitrogen concentration (TNC)*: Total nitrogen in the influent and in the effluent of the WWTP is not an online measure, but an offline analytic measure obtained three times per week. The total nitrogen concentration in the effluent is stablished by applicable regulations to a maximum value of 10 mg/l. In terms of nitrogen removal, an efficiency around 80% is considered a good performance.

To simplify the evaluation of the CBR module's performance, the RBR module is deactivated during a few weeks, as well as the retain stage. The results generated during this period are

used to calculate the distance threshold (d_{thr}) in (7). Then, the RBR module is activated and the Decision module selects which solution to use – RBR or CBR – as detailed in the Algorithm 5, as well as the retain stage (Algorithm 4).

The tool has been running in the plant from January 2020 to July 2020. In Section 4, obtained results are presented and discussed. First, results are evaluated from the point of view of the KPIs presented in this section. Then, the competence of the proposed reasoning system is quantified.

4 RESULTS AND DISCUSSION

For the implementation and validation of the IDSS methodology described in this work a GUI has been built. This tool includes all the necessary processes for the integration with the environmental system –a WWTP in this case–, namely: communication via OLE for Process Control (OPC) standard with the plant Programmable Logic Controller (PLC); connection with a local database and; the required configuration options for the user. In Figure 9, a screenshot of the main window of the developed IDSS and the IPCS is presented. In this window, all the process' data available in the database is displayed, including the KPI values. Figure 10 shows the configuration window for the reasoning cycle —i.e. interoperation of the RBR, CBR and Decision modules. From this window, rules can be modified in real time, as well as the variables included in the CB – i.e. variables can be included or excluded from the descriptive part of a case.

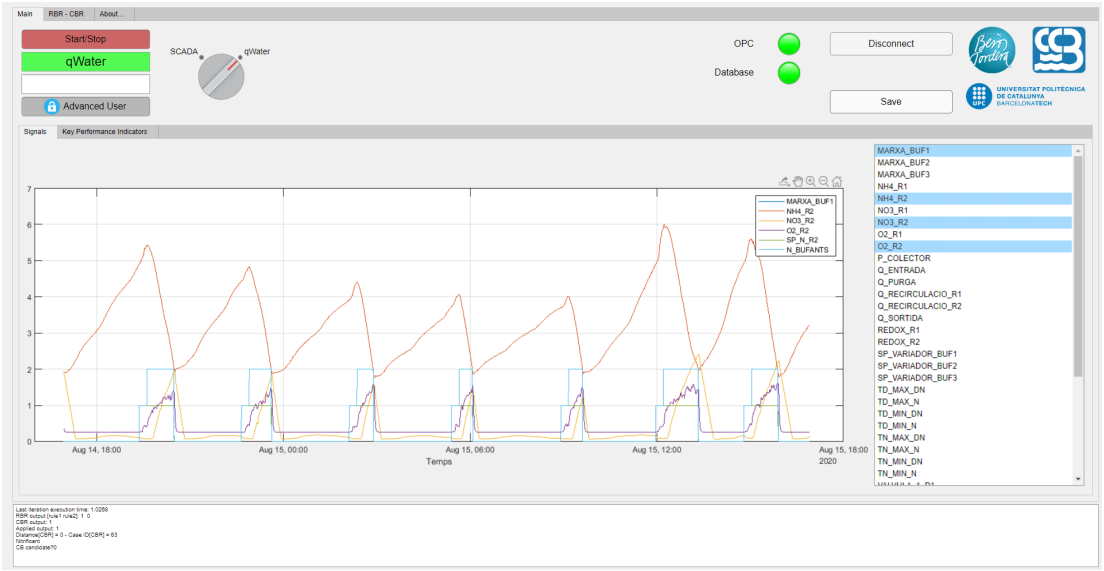

Figure 9 GUI for the IPCS – main screen

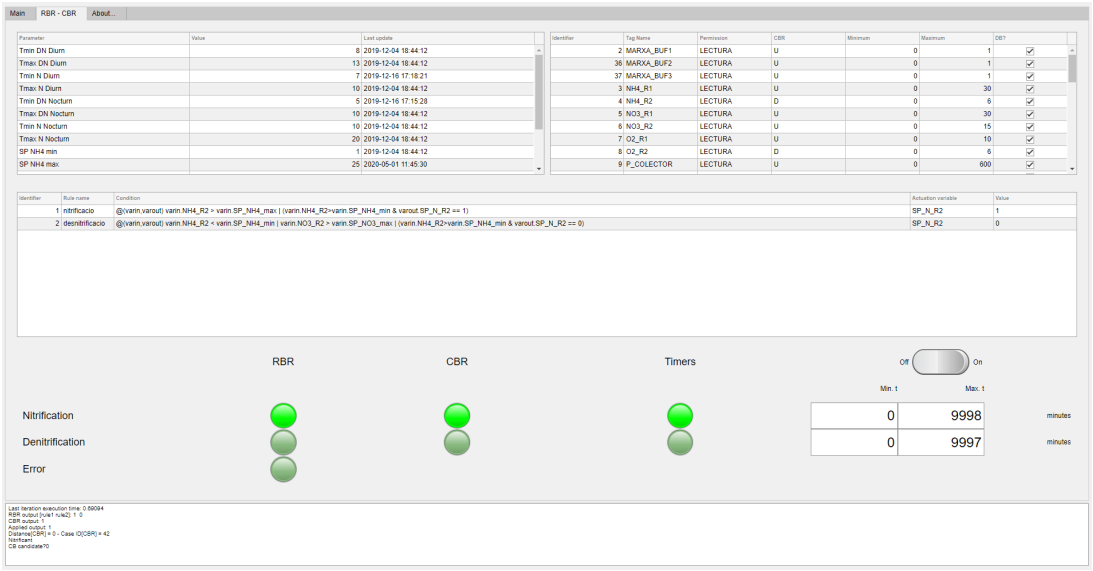

Figure 10 GUI for the IPCS – RBR and CBR configuration

In Figure 11, the distances obtained with real data for one-month period are represented. The distance threshold is calculated as explained in Section 2.4, using the first 15 days. The obtained value is $d_{thr} = 0.192$. Thus, all new cases with $d_{thr} > 0.192$ are candidates to be added to the case base, if the KPI values are within the specified limits. The second part of the month is used to validate the distance threshold and the retain phase. It can be seen that two new cases are added to the case base and how the distance values are changing. The distance threshold d_{thr} is recalculated each time a new case is added to the case base.

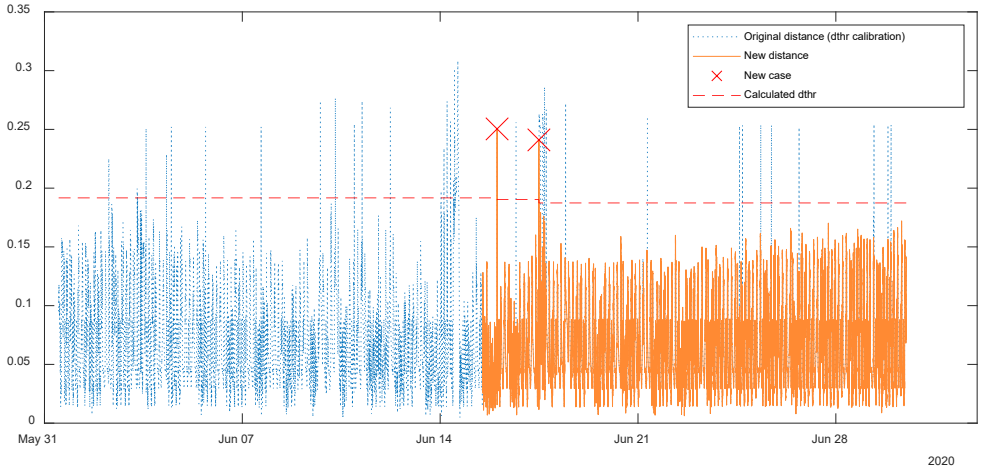

Figure 11 Distance threshold calculation example and case retain

In Figure 12, Figure 13 and Figure 14 the evolution of the proposed KPIs is shown. In Figure 12, the 24h mean average calculated by the NH4 sensor (24MA_AC) is compared with the analytical results obtained in the laboratory from samples not in the biological reactor but in the effluent, where the limit of four mg/l has to be achieved. The effluent is analysed about three times per week. It can be observed that almost all analysed samples are below the limit, i.e. the operation is correct. From February to mid-May it can be noted that the NH4 sensor

measurements are well above the limit due to a calibration problem in the sensor. Data in this period are not considered to obtain the CB.

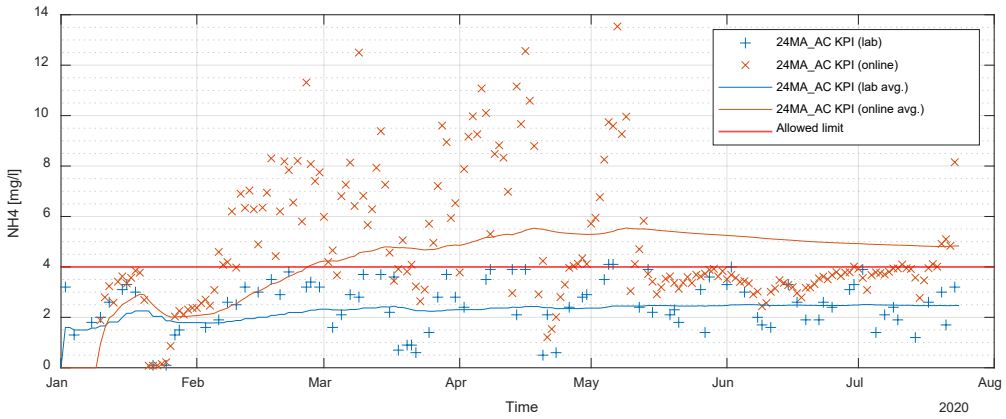

Figure 12 NH4 24h average (24MA_AC) KPI

In Figure 13, the electrical consumption (BEC) of the studied period is compared with the mean electrical consumption of the same period in 2019. In addition, the ratio of this electrical consumption related to the volume of treated water is shown. The electrical consumption is similar to the one before the deployment of the IPCS tool. The ratio between treated water and consumption is 0.3 kwh/m^3, below the 0.4 kwh/m^3 during the same period in 2019. Redox is an indirect measure of the nitrification and denitrification process progression whilst the ammonia and nitrate are direct measures. Thus, the control is expected to be more precise using ammonia and nitrate sensors, and therefore the electrical consumption could be reduced if some constraints imposed by the operator are relaxed.

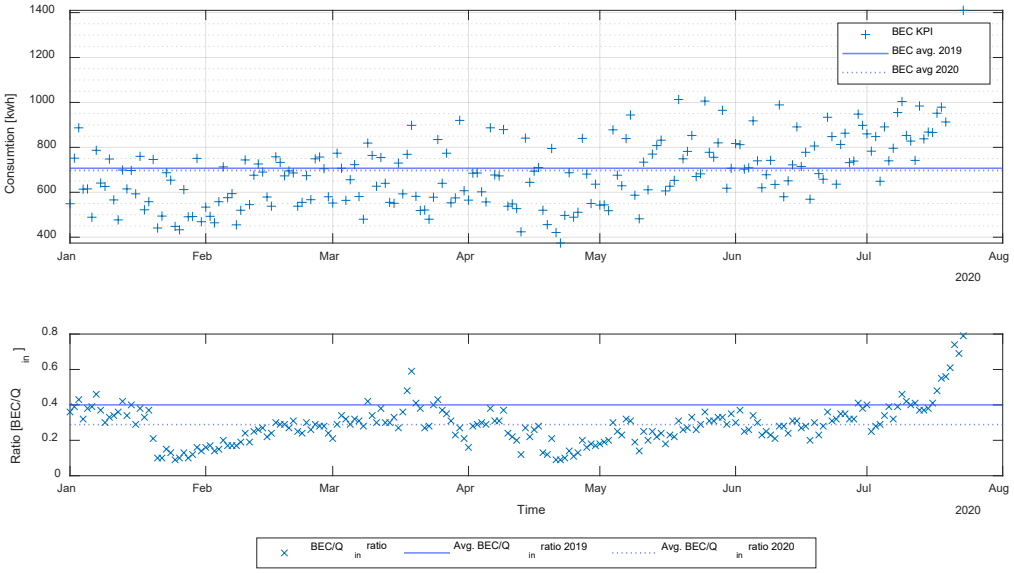

Figure 13 Electrical Consumption (BEC) KPI

Finally, Figure 14 shows the total nitrogen (TNC) in the effluent compared to the allowed limit. Some values over the limit can be observed in the last days of January 2020. These results are due to an intense rainy period that caused a high increase of the nitrate concentration in the WWTP influent. The European Union directive 91/271/CEE on urban waste water treatment stablishes the maximum nitrogen concentration in the effluent or the minimum nitrogen removal efficiency depending on the WWTP influent load expressed in population equivalent (PE) units. In terms of nitrogen removal efficiency, it can be pointed out that the

mean removal efficiency is about 80%. The European directive stablishes a minimum value between 70 and 80%. So, in terms of nitrogen removal results are also within the limits.

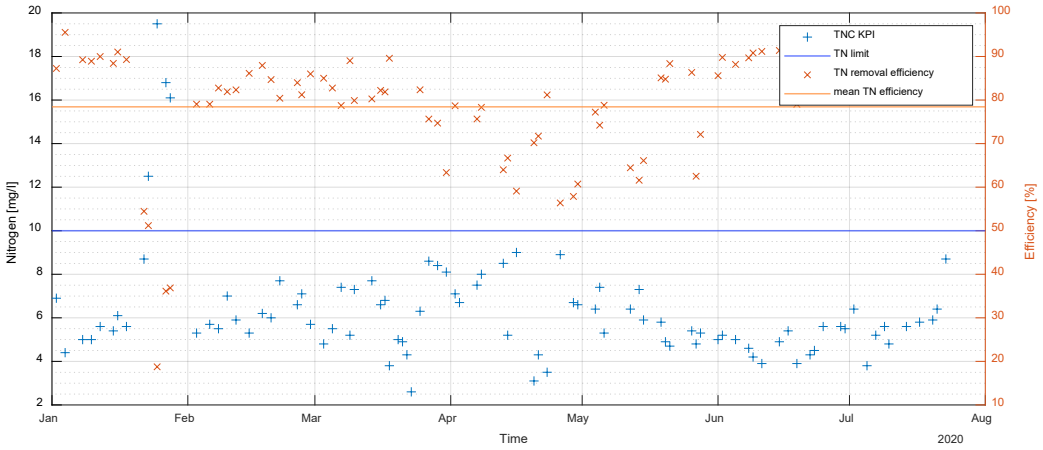

Figure 14 Total Nitrogen (TNC) KPI and Nitrogen removal

The KPIs evaluates the performance of the tool from the point of view of the process efficiency. In Table 5 the competence of the system is quantified. With this purpose different indexes are defined:

- *Solved cases* (SC): Percentage of solved cases. This index indicates the period when the tool is operating. The tool may not be operating during maintenance tasks, regarding the application itself or the WWTP.
- *CBR index* (CBR_i): Percentage of cases solved by CBR module.
- *RBR index* (RBR_i): Percentage of cases solved by RBR module.
- *Expert index* (E_i): Percentage of cases that are not solved by the reasoning cycle, i.e. the given solution is not the one used due to maintenance tasks in the WWTP or open-loop fixed timers set-points.
- *Retain index* (R_i): Number of retained cases. This index considers the new learned cases to the original case base of 40 cases described in Section 3.2.
- *Correctly solved cases (*CSC*):* Percentage of correctly solved cases by the whole reasoning system – i.e. CBR and RBR – considering the expert assessment.

4.1 Discussion

The KPIs designed to evaluate the performance of the tool provide useful information to the operators on the efficiency of the plant operation and the effluent quality, particularly in terms of nitrogen removal. The 24MA_AC KPI in Figure 12 shows that the amonia concentration in the effluent (lab value) is predominantly below the allowed limit (only two samples out of 88 obtained from laboratory analitycal tests are over this limit). The online measure obtained from the ammonia sensor in the biological reactor (NH4) is used to control the process and gives a good aproximation of the concentration of this parameter in the effluent. It is assumed that a value within the limits in the biological reactor results in a value within the limits in the effluent. In terms of total nitrogen (TNC KPI, Figure 14), most analytical results obtained in the laboratory are below the maximum allowed concentration. Only four samples out of 91 obtained from laboratory analitycal tests are out of bounds. These values over the allowed limit can be explained by the exceptional weather conditions in that period. From 19th to 25th of January 2020 a storm named *Gloria* was moving across Spain. The abundance of rainfall

produced floods and the increase of influent flow in the plant. Rain water has a high concentration of dissolved oxygen, which supposes an increase of the nitrate's concentration and consequently, an increase of the total nitrogen. The electrical consumption needed to achieve a good performance in nitrogen removal is compared to the consumption in the same period a year before the implementation of the tool presented in this work (Figure 13). The average values are quite similar, around 700 kWh/day, but it is necessary to consider some facts in the analysis of this result. The effluent quality is generally better than required by the current legislation, i.e. if the ammonia limit is 4 mg/l a lower concentration in the effluent is not required, largely due to some restrictions imposed by the operator. Hence, the operation in terms of electrical consumption may be improved by the relaxation of these restrictions, i.e. reducing temporized nitrification and denitrification cycles. Finally, in the period from May 2020 to August 2020 (Figure 13), it can be observed an increase in the consumption as a consequence of an increase of the contamination in the influent, producing higher requirements for the aeration. One blower cannot provide enough oxygen to reach the set-point; therefore, one backup blower is activated. Up until then the plant was operated using only one blower. Taking into account these results and remarks, it should be noted that the nitrogen removal target is achieved. From the point of view of the electrical consumption it is difficult to compare the results obtained with previous historical data because of several changes in the process, e.g. the control is based on different sensors or the increase of oxygen demand. But despite all of that, electrical consumption is similar to the historical one and can be reduced addressing two points, namely: a) the adjustment of the nitrogen removal to the allowed limit; b) the reduction or removal of the restrictions that avoid the use of the solutions proposed by the reasoning cycle.

Table 5 Competence of the reasoning system

Period	**SC [%]**	**CBR_i [%]**	**RBR_i [%]**	**E_i [%]**	**R_i**	**CSC [%]**
January°	71.0	0	100	nd	nd	nd
February°	99.5	0	100	nd	nd	nd
March°	88.5	nd	nd	16.0	nd	nd
April	94.7	100	0	15.6	1	85.04
May	84.6	99.98	0.02	9.2	9	92.00
June	99.3	100	0	4.4	0	96.56
July	71.4	99.92	0.08	9.1	24	91.08
August	97.2	99.94	0.06	7.4	25	94.01
Total#	89.4	99.97	0.03	9.1	59	91.74

nd: not determined
° results during this period are approximate values
period from April to August is considered

The competence of the reasoning system is summarized in the Table 5. During the first three months most indexes cannot be determined. From January 2020 to the end of February 2020 only the RBR module is working. In March 2020 the CBR module is activated, but CBR_i and RBR_i indexes are not determined due to the detection and solution of several bugs during these first weeks. From April 2020 to August 2020 the application operation is considered stable and only minor bugs are detected and solved. Considering the values shown in Table 5, it can be observed a wide scope of the case base, i.e. most of the situations occuring are included in the CB, with a 99.97% of cases solved using the case-based reasoning module. It can be also noted that the E_i index is reduced from May 2020 due to increased user confidence on the application. This fact has allowed reducing the restrictions imposed by the open-loop fixed timers, which ideally should be removed completely in the future since they correspond to an open-loop operation of the facility. It has also been shown how the RBR module may be used to solve the first cases when historical records are scarce or unexistent to inicialize the CBR module with a CB. The CBR module has been partially activated after the validation of the RBR module with the retrieval, reuse and revision phases, in the period ranging from March to April. In April, after validation of the good performance of the initial case base, the retain phase is activated, together with the Decision module. Finally, the CSC index

determines the percentage of correct solved cases by the whole reasoning system (i.e. RBR and CBR modules), taking into account the expert feedback, i.e. whether the reasoning system response is correct considering the expert criteria. For the period from April 2020 to August 2020 the percentage of correct solved cases is 91.74%, which is assumed to be a good performance from the practitioner point of view. Additionaly, in light of the water quality in the effluent and the high percentage of correct solved cases, the results may be considered enviromentally satisfactory.

5 CONCLUSIONS

This work proposes a generic and hybrid IDS methodology using the interoperation of RBR and CBR techniques for the automated development of intelligent environmental decision support systems. The aim of this proposal is to solve a common problem of this type of decision support and control systems, which is the *ad-hoc* design for different installations, as well as the lack of adaptability to dynamic changes on the environmental systems. To this end, the approach presented here has been designed in a general fashion and integrated in an environmental software tool for the sake of scalability to different types of environmental systems —without loss of generality, WWTPs here—, but also to further types of systems beyond the environmental framework. The RBR module presented here is configured with a collection of rules designed with the participation of environmental experts on the process under study. Some of these rules are generic enough to be used in any environmental system, whereas others can be used in a particular system. On the other hand, the CBR module presented here is based on historical data obtained from a particular environmental process. Therefore, using a valid case base allows the methodology to be deployed in any particular environmental system instance. The suitability of this approach has been satisfactorily tested with the experimental evaluation in a real WWTP facility.
The performance of the proposed tool is analysed using a set of designed KPIs. For the particular system under study, it has been shown how the performance achieved is within the desired values in terms of nitrogen removal and electrical consumption, compared with the same parameters calculated during the same period before the IDSS deployment. Hence, the solution presented here proposes a novel way of supervising and helping in the decision making of the processes in the system, which is robust against real-world situations —e.g. low quality measurements gathered from the system—, and provides good performance in a real case scenario. In the future work section, some guidelines to improve the performance of the methodology proposed and the tool deployed are detailed.

5.1 Future work

Further work includes the deployment of the tool proposed here in other WWTPs, with the aim of further demonstrating the scalability of the method. The case study presented in this work is complex enough to validate the methodology proposed and to deal with the challenges posed by the implementation of a novel intelligent environmental decision support software tool in a real system. The next steps will consider further systems with additional environmental preserving strategies and optimisation possibilities. The results obtained with the system studied here can be improved considering, for example, different prices of electricity depending on the day time or adjusting the quality of the effluent to the maximum values allowed.
On the other hand, the performance of the tool deployed is heavily reliant on the data quality, as shown in Section 4. Consequently, the integration of data validation and reconciliation methods should be in the scope of future steps (Cugueró et. al., 2016). In a similar way, the dynamic nature of this domain suggest temporary dependencies between cases. To deal with this problem the use of a temporal case-based reasoning approach could be useful (Sànchez-Marrè et al., 2005).

The efficiency of the biological processes like the one addressed here depends on different boundary conditions —e.g. temperature—, so its behaviour in different seasons – e.g. winter vs summer – is different. For this reason, it is necessary to consider longer datasets, hence testing this methodology for longer periods in order to enrich the CB with all possible situations. At the same time, the maintenance of the CB will grow in importance, so techniques to deal with the information increase will be used, e.g. discriminant trees or *k-d* trees to index the CB through a hierarchical structure improving the retrieval time.
Finally, in further steps the GUI will be improved in order to facilitate the integration of the environmental experts knowledge in different situations, e.g. validation of solutions proposed by the tool when KPIs are out of bounds. This point is important to take advantage of the experience of the user and to provide practitioners with more confidence on the data-based approaches, which are quite different than the traditional ones used in system operation.

ACKNOWLEDGEMENTS

The authors acknowledge the partial support of this work by the Industrial Doctorate Programme (2017-DI-006) and the Research Consolidated Groups/Centres Grant (2017 SGR 574) from the Catalan Agency of University and Research Grants Management (AGAUR), from Catalan Government. In addition, they want to acknowledge the staff from Santa Maria de Palautordera WWTP for their helpful and constant support during all the steps of the experimental work.

REFERENCES

Aamodt, A. and Plaza, E. Case-based reasoning: fundamental is-sues, methodological variations and system approaches. AI Communications 7(1):39-59, 1994.

Ahmed, S.A., Shadia R.T., Hala, A.T., 2002. Development and Verification of a Decision Support System for the Selection of Optimum Water Reuse Schemes. Desalination 152 (1-3), 339–352.

Aulinas, M., Nieves, J.C., Cortés, U., Poch, M. Supporting decision making in urban wastewater systems using a knowledge-based approach, Environmental Modelling & Software, Volume 26, Issue 5, 2011, Pages 562-572, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2010.11.009.

Béraud, B., Steyer, J. P., Lemoine, C., Latrille, E., Manic, G., & Printemps-Vacquier, C. (2007). Towards a global multi objective optimization of wastewater treatment plant based on modeling and genetic algorithms. Water Science and Technology, 56(9), 109-116.

Bernardelli, A., Marsili-Libelli, S., Manzini, A., Stancari, S., Tardini, G., Montanari, D., Venier, S. (2020). Real-time model predictive control of a wastewater treatment plant based on machine learning. Water Science and Technology, 81(11), 2391-2400. doi:10.2166/wst.2020.298

Berthuex, P.M., Lai, M., Darjatmoko, D., 1987. A Statistics-based information and expert system for plant control and improvement. In Proceeding of 5th National Conf. on Microcomputers in Civil Engineering, (W.E. Carrol, editor), Orlando, Florida, 146-150.

Buchanan B. G. and Duda R. O. Principles of Rule-Based Expert Systems. In Yovits, M.C. (ed.), Advances in Computers, Vol. 22, pp: 163-216. New York: Academic Press (1983).

Campbell, Stephen L., Chancelier, Jean-Philippe, Nikoukhah, Ramine (2010). Modeling and Simulation in cilab/Scicos with ScicosLab 4.4

Capodaglio, A.G., Jones, H.V., Novotny V., Feng, X., 1991. Sludge bulking analysis and forecasting: application of system identification and artificial neural computing technologies. Water Research 25(10), 1217-1224.

Castillo, A., Cheali, P., Gómez, V., Comas, J., Poch, M., Sin, G., 2016. An integrated knowledge-based and optimization tool for the sustainable selection of wastewater treatment process concepts. Environmental Modelling & Software. 84, 177–192. https://doi.org/10.1016/j.envsoft.2016.06.019

Chapman, Stephen J. (2020). Matlab Programming for Enginners, 6th Edition. Cengage learnng.

Choi, G. W., Chong, K. Y., Kim, S. J. and Ryu, T. S. SWMI: new paradigm of water resources management for SDGs. Smart Water, vol. 1, no. 1, pp. 1–12, 2016.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M., 2018. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling & Software 106, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023

Côte, M., Grandjean, B. P. A., Lessard, P., Thibault J., 1995. Dynamic Modelling of the Activated Sludge Process: Improving Prediction Using Neural Networks. Water Res., 29 (4), 995-1004.

Cugueró-Escofet, M. À., García, D., Quevedo, J., Puig, V., Espin, S., & Roquet, J. (2016). A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network. Control Engineering Practice, 49, 159–172.

Czoagala, E., Rawlik, T., 1989. Modelling of a Fuzzy Controller with application to the Control of Biological Processes. Fuzzy Sets and Systems 31, 13-22.

De Mulder, C., Flameling, T., Weijers, S., Amerlinck, Y., Nopens, I., 2018. An open software package for data reconciliation and gap filling in preparation of Water and Resource Recovery Facility Modeling. Environmental Modelling & Software 107, 186–198. https://doi.org/10.1016/j.envsoft.2018.05.015

Di Biccari, C. and Heigener, D. Semantic modeling of wastewater treatment plants towards international data format standards. In 30 Forum Bauinformatik (Weimar, Germany), 2018, no. September, pp. 183–190.

Feng, L., Ouedraogo, A., Manghee, S. and Danilenko, A. A primer energy efficiency for municipal water and wastewater utilities. Washington DC, USA, 2012.

Flanagan, M.J., 1980. On the Application of Approximate Resaoning to the Control of Activated Sludge Process. In Proceedings of Joint Automatic Control Conference, ASME, San Francisco, CA.

Gall, R., Patry G., 1989. Knowledge-based system for the diagnosis of an activated sludge plant. In Dynamic Modelling and Expert Systems in Wastewater Engineering. (G. Patry and D. Chapman editors), Chelsea, MI. Lewis Publishers, 1989.

Gibert, K., Sànchez-Marrè, M., Codina, V. (2010). Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation. *5th International Congress on Environmental Modelling and Software (iEMSs 2010)*. iEMSs' 2010 Proceedings, Vol. 3, pp. 1940-1947.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S.H., Rodríguez-Roda, I., Holmes, G., 2018. Which method to use? An assessment of data mining methods in Environmental Data Science. Environmental Modelling & Software 110, 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021

Gourbesville, P. Key Challenges for Smart Water. Procedia Eng., vol. 154, pp. 11–18, 2016.

Hamed, M.M., Khalafallah, M.G., Hassanien, E.A., 2004. Prediction of Wastewater Treatment Plant Performance Using Artificial Neural Networks. Environmental Modelling & Software 19, 919-928.

Han, H.-G., Zhang, H.-J., Liu, Z., Qiao, J.-F., 2020. Data-driven decision-making for wastewater treatment process. Control Engineering Practice, 96, art. no. 104305. https://doi.org/https://doi.org/10.1016/j.conengprac.2020.104305

Hans-Jürgen Schönig (2018). Mastering PostgreSQL 11 - Second Edition. Packt. https://www.postgresql.org/

Jackson, P. (1999). Introduction to Expert Systems. 3rd edition. Boston, MA: Addison-Wesley.

Johnson, G.W. and Jennings, R., 2006. LabVIEW graphical programming. McGraw-Hill, New York

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.

Karr, C.L., 1991. Genetic Algorithms for Fuzzy Controlers. AI Expert 6(2), 26-33.

Kolodner, J.L., Case-Based Reasoning. Morgan Kaufmann, 1993.

Kosko, B., 1992. Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Prentice Hall, USA.

Laxmi, P. and G. Laxmi-Deepthi, G. Smart Water Management Process architecture with IoT Based Reference. *Int. J. Comput. Sci. Mob. Comput*., vol. 6, no. 6, pp. 271–276, 2017.

Maeda, K., 1989. A knowledge-based system for the wastewater treatment plant. Future Generation Computer Systems 5, 29-32.

Mannina, G., Rebouças, T., Cosenza, A., Sànchez-Marrè, M. and Gibert, K. (2019). Decision support systems (DSS) for wastewater treatment plants – A review of the state of the art. *Bioresource Technology*. Vol. 290. 121814. 10.1016/j.biortech.2019.121814.

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

Nadiri, A.A., Shokri, S., Tsai, F.T.C., Moghaddam, A. A., 2018. Prediction of Effluent Quality Parameters of a Wastewater Treatment Plant Using a Supervised Committee Fuzzy Logic Model. Journal of Cleaner Production 180, 539–549.

Nagar, S. (2017). Introduction to Scilab for Engineers and Scientists. https://www.scilab.org/

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Martínez, M., Rodríguez-Roda, I. and Poch, M. (2004). A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. Environmental Modelling & Software 19, 809–819. https://doi.org/10.1016/j.envsoft.2003.03.003.

Oulebsir, R., Lefkir, A., Safri, A. and Bermad, A. Optimization of the energy consumption in activated sludge process using deep learning selective modeling, Biomass and Bioenergy, vol. 132, no. October 2019, p. 105420, 2020.

Pascual-Pañach, J., Cugueró-Escofet, M.A., Aquiló-Martos, P. and Sànchez-Marrè, M. (2018). An Interoperable Workflow-Based Framework for the Automation of Building Intelligent Process Control Systems. 9th International Congress on Environmental Modelling and Software, Fort Collins, Colorado, USA. https://scholarsarchive.byu.edu/iemssconference/2018/Stream-B/20/

Poch, M., Comas, J., Cortés, U., Sànchez-Marrè, M., Rodríguez-Roda, I. Crossing the death valley to transfer environmental decision support systems to the water market. "Global challenges", 10 Abril 2017, vol. 1, núm. 3, p. 1700009-1-1700009-10.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M., Cortés, U. (2004). Designing and building real environmental decision support systems. Environmental Modelling & Software 19: 857-873. 10.1016/j.envsoft.2003.03.007.

Ráduly, B., Gernaey, K. V, Capodaglio, A.G., Mikkelsen, P.S., Henze, M., 2007. Artificial neural networks for rapid WWTP performance evaluation: Methodology and case study. Environmental Modelling & Software 22, 1208–1216. https://doi.org/10.1016/j.envsoft.2006.07.003

Richter, M. and Weber, R. (2013). Case-based reasoning: a textbook. Springer

Riesbeck, C.K. and Schank, R.C.. Inside Case-Based Reasoning. Lawrence Erlbaum Associates Publishers, 1989.

Robles, T., Alcarria, R., Martín, D., Navarro, M., Calero, R., Iglesias, S. López, M. 2015. An iot based reference architecture for smart water management processes. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 6. 4-23.

Ruano, M. V, Ribes, J., Sin, G., Seco, A., Ferrer, J., 2010. A systematic approach for fine-tuning of fuzzy controllers applied to WWTPs. Environmental Modelling & Software 25, 670–676. https://doi.org/10.1016/j.envsoft.2009.05.008

Ruiz Gutierrez, J. M., 2017. https://myopenlab.org/documentos/

Salantino, M., De Maio, M., Aliverti, E. (2016). Mastering JBoss Drools 6. Packt.

Sànchez-Marrè, M., 2014. Interoperable Intelligent Environmental Decision Support Systems: a Framework Proposal. 7th International Congress on Environmental Modelling & Software (iEMSs 2014). iEMSs 2014 Proceedings, 1, 501-508. Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.), 201

Sànchez-Marrè, Miquel, Cortés, Ulises, Martínez, Montse, Comas, Joaquim, Rodríguez-Roda, Ignasi. (2005). An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). 3620. 465-476. 10.1007/11536406_36.

Sànchez-Marrè, M., Martinez, M., Rodríguez-Roda I., Alemany, J., Cortés, C., 2004. Using CBR to improve intelligent supervision and management of wastewater treatment plants: the atl_EDAR system. 7th European Conference on Case-Based Reasoning (ECCBR'2004), Proc. of Industrial day, 7th European Conference on Case-Based Reasoning (Eds. Francisco Martin and Mehmet Göker), 79-91

Sànchez-Marrè, M., Cortés, U., Rodríguez-Roda, I., Poch, M., Lafuente, F., 2002. Learning and Adaptation in Wastewater Treatment Plants Through Case–Based Reasoning. Comput. Civ. Infrastruct. Eng. 12, 251–266. https://doi.org/10.1111/0885-9507.00061

Sànchez-Marrè, M., Cortés, U., R-Roda, I., Poch, M., Lafuente, J., 1997. Learning and Adaptation in WWTP through Case-Based Reasoning. Special issue on Machine Learning. Microcomputers in Civil Engineering/Computer-Aided Civil and Infrastructure Engineering 12(4), 251-266.

Santín, I., Barbu, M., Pedret, C., Vilanova, R., 2018. Fuzzy logic for plant-wide control of biological wastewater treatment process including greenhouse gas emissions. ISA Trans. 77. https://doi.org/10.1016/j.isatra.2018.04.006

Schank, R.. Dynamic memory: a theory of learning in computers and people. Cambridge University Press, 1982.

Serra, P., Sànchez-Marrè, M., Lafuente, J., Cortés, U. and Poch, M., 1994. DEPUR: a knowledge based tool for wastewater treatment plants. Engineering Applications of Artificial Intelligence 7(1), 23-30.

Steels, L. Components of expertise. AI Magazine 11(2):28-49, 1990.

Syu, M.-J. and Chen. B.-C., 1998. Back-propagation Neural Network Adaptive Control of a Continuous Wastewater Treatment Process. Industrial & Engineering Chemistry Research, 37(9), 3625-36230.

Torregrossa, D., Hernández-Sancho, F., Hansen, J., Cornelissen, A., Popov, T., Schutz, G., 2017. Energy Saving in Wastewater Treatment Plants: A Plant-Generic Cooperative Decision Support System. Journal of Cleaner Production 167, 601–609.

Tzafestas, S. and Ligeza. A., 1989. A Framework for Knowledge Based Control. Intelligent and Robotic Systems 1(4), 407-426.

Wang, X.Z., Chen, B.H., Yang, S.H., McGreavy, C., Lu, M.L. , 1997. Fuzzy Rule Generation from Data for Process Operational Decision Support. Computers Chem. Engng. 21, S661-S666