ENVIRONMENTAL ENGINEERING SCIENCE
Volume 23, Number 2, 2005
© Mary Ann Liebert, Inc.

Constructed Wetlands: Prediction of Performance with Case-Based Reasoning (Part B)

Byoung-Hwa Lee, Miklas Scholz,[*] Antje Horn, and Alison M. Furber

Institute for Infrastructure and Environment
School of Engineering and Electronics
The University of Edinburgh
Edinburgh EH9 3JL, Scotland
United Kingdom

ABSTRACT

The aim of this research was to assess the treatment efficiencies for gully pot liquor of experimental vertical-flow constructed wetland filters containing *Phragmites australis* (Cav.) Trin. ex Steud. (common reed) and filter media of different adsorption capacities. Six out of 12 filters received inflow water spiked with metals. For 2 years, hydrated nickel and copper nitrate were added to sieved gully pot liquor to simulate contaminated primary treated storm runoff. The findings were analyzed and discussed in a previous paper (Part A). Case-based reasoning (CBR) methods were applied to predict 5 days at 20°C N-Allylthiourea biochemical oxygen demand (BOD) and suspended solids (SS), and to demonstrate an alternative method of analyzing water quality performance indicators. The CBR method was successful in predicting if outflow concentrations were either above or below the thresholds set for water-quality variables. Relatively small case bases of approximately 60 entries are sufficient to yield relatively high predictions of compliance of at least 90% for BOD. Biochemical oxygen demand and SS are expensive to estimate, and can be cost-effectively controlled by applying CBR with the input variables turbidity and conductivity.

Key words: storm runoff; gullies; constructed wetlands; case-based reasoning; sampling scheme optimization; effluent standards

INTRODUCTION

Case-based reasoning (CBR) is a method of problem solving, which has arisen out of the field of artificial intelligence, and aims to recreate the robust problem solving technique often used by humans within the constraints of a computer program (Aamodt and Plaza, 1994; Arditi and Tokdemir, 1999). When a human encounters a problem he or she tends to remember similar situations that they have come across in the past, and the methodology in which solutions were found. By recalling these events, it becomes possible to reuse the previous solution(s) to solve the current problem, perhaps adjusting the methodology and outcome slightly to meet the

[*] Corresponding author: Institute for Infrastructure and Environment, School of Engineering and Electronics, The University of Edinburgh, Faraday Building, The King's Buildings, Mayfield Road, Edinburgh EH9 3JL, Scotland, United Kingdom. *Phone*: +44 131 6 506780; *Fax*: +44 131 6 506554; *E-mail*: M.Scholz@ed.ac.uk

specific requirements of the new task (Aamodt and Plaza, 1994).

Case-based reasoning works very similarly to the human logic of data handling. A data base of past experiences that may be useful to solve a particular type of query is kept. The difficulty in CBR is the design of a system that is capable of recalling past experiences, which would provide useful information when a new problem is introduced to the system. In CBR terminology, the event in which a solution to a former problem was found is referred to as a "case," and is stored in the system's "case base." For the purpose of CBR, each case should be recorded within the case base systematically and the useful information must be stored consistently through the entire case base, the chosen structure used being referred to as the "case representation" (Ardity and Tokdemir, 1999).

When a new problem is introduced to the CBR system, it should be represented in the same format as the stored cases, and then the process of deciding which of the past cases may be of use in finding a solution to this problem can begin. The main assumption underlying a CBR methodology is that similar problems will have similar solutions. It follows that the most useful cases in the case base will be those that are most similar to the problem case.

The concept of similarity is fundamental in CBR theory, making inexact matching possible, which is required when previously unseen problems arise. A mechanism is implemented within the system that is capable of recalling past cases that are most closely matched to the problem presented in terms of the variable(s) used to describe the cases. Therefore, the variables used should be carefully chosen such that the solutions recalled will also be relevant to the problem case. Once the most similar cases have been selected, the predicted solution is found using an adaptation or learning process (Aamodt and Plaza, 1994).

CBR applied to biochemical data

Concerning general data sets, CBR systems are often seen as simple, convenient, and effective methods of artificial intelligence for multicomponent analysis (Arditi and Tokdemir, 1999). The methodology is based on assuming regularity, typicality, and consistency. Moreover, it can be characterized by the four "re" steps: retrieve, reuse, revise, and retain. The output or target variable is determined from input variables that are associated with weightings. Various methods of their determination exist: uniform (i.e., no weightings), correlated, and calibrated weightings, as well as exact and fuzzy matched meta weightings (Watson, 1997).

Case-based reasoning has been successfully applied to the development and implementation of a knowledge-based hybrid supervisory system to support the operation of a real wastewater treatment plant (Rodriguez-Roda *et al*., 2002). The CBR system can be structured into three separated levels: data gathering, diagnosis, and decision support. The different tasks of the system can be performed in a seven-step cycle: data gathering and update, diagnosis, supervision, prediction, communication, actuation, and evaluation (Rodriguez-Roda *et al*., 2002).

With respect to biochemical data sets, hybrid CBR systems for monitoring water quality based on chemical variables and algae populations have been applied previously (Policastro *et al*., 2004). A CBR system was also successfully developed to supervise complex biochemical processes such as the activated sludge process (Roda *et al*., 2001). The suitability of CBR has also been shown in aqueous solutions containing mixtures of ions of different nature and concentration. For example, CBR has been successfully applied to the rapid recognition and evaluation of mineral water samples (Colilla *et al*., 2002).

Constructed treatment wetlands are often seen as complex "black box" systems, and have therefore not been used previously for a detailed CBR analysis. The processes within an experimental constructed treatment wetland are difficult to model due to the complexity of the relationships between most water quality variables (Gernaey *et al*., 2003; Nunez *et al*., 2004). However, it is necessary to monitor, control, and predict the treatment processes to meet environmental and sustainability policies, and regulatory requirements such as secondary wastewater treatment standards (Lee *et al*., 2005). CBR methodologies (Aamodt and Plaza, 1994) could be used to make water quality predictions and to optimize the operation of treatment wetlands. Consideration should also be given to CBR as a learning tool.

The measurement of biochemical oxygen demand (BOD) and suspended solids (SS) concentrations is widely applied for wastewater before and after treatment, as they give a general indication of the water quality status. BOD is a measurement of the oxygen consumed in 5 days by organisms within the water sample stored within an incubator at 20°C. N-Allylthiourea is usually added to inhibit nitrification [American Public Health Association (APHA), 1995]. A regulator such as the Scottish Environmental Protection Agency imposes thresholds for water quality variables. The corresponding secondary wastewater treatment thresholds for BOD and SS are 20 and 30 mg/L, respectively (Scholz *et al*., 2002; Lee *et al*., 2005).

Methods of measuring or reliably predicting BOD and SS are useful for the day-to-day operation of constructed treatment wetlands. Unfortunately, taking BOD measurements is both expensive (measurements are labor intensive and capital costs of modern on-line equipment are relatively high; approximately £15,000) and only of historical value (results are not available until 5 days after the sample has been taken). Furthermore, the procedures to estimate BOD and SS concentrations are time-consuming and labor-intensive. Therefore, some method of prediction, if it could be made reliable enough, would be advantageous (Lee *et al*., 2005).

CBR could be a methodology well suited to the analysis of biochemical data. Verdenius and Broeze (1999) discussed the difficulties that arise in modeling environmental systems, highlighting particularly the complexity of the relationships between the different variables, and how these relationships change over time. The creation of any model, using traditional techniques, requires expertise of the environmental system, mathematical technique, and software packages used. Once compiled, the model will then require continual updating as the system evolves over time. In many circumstances, mathematical models have proved insufficient, and environmental experts have had to be called in to find solutions based on empirical observations. Experts do recall their experience of similar past problems to come up with a new solution (Nunez *et al*., 2004). As CBR also uses this technique to find solutions, it is thought that CBR may be beneficially applied to environmental problems. Moreover, CBR has been shown to function well with highly complex "black box" system problems, and as a CBR system can be designed to "learn" solutions to the new problems it comes across, the system is dynamic and will update itself without much intervention from the system designer (Sanchez-Marre *et al*., 1999).

The application of CBR to predict variables as part of complex biochemical data sets should be considered. The predictions of BOD and SS serve as examples of one possible application of CBR to biochemical data. Successful predictions could help to optimize the operation and maintenance of constructed treatment wetlands.

Project purpose

The major purpose of this part of the study is to improve water quality monitoring and interpretation guidelines of vertical-flow constructed treatment wetlands with case-based reasoning, and to use a case study (Lee *et al*., 2005) as an educational tool. The objectives are to assess

1. the potential of CBR for analyzing biochemical data, interpretation of wetland data, and predicting BOD and SS;
2. the most appropriate method of selecting input variables,
3. the optimum size of the case base;
4. the goodness of prediction with a CBR analysis; and
5. the potential of CBR as a teaching tool to enhance understanding of "black box" systems.

METHODOLOGY AND SOFTWARE APPLIED TO UNDERTAKE CBR

The experimental data set applied for this study has been described in detail by Lee *et al*. (2005). The CBR system used to predict the BOD and SS concentrations of treated gully pot liquor samples was created using simple mathematical functions in Microsoft Excel. Past cases were sorted in the case base represented by up to six input variables: turbidity (NTU), conductivity (μS), redox potential (mV), outflow water temperature (°C), dissolved oxygen, DO (mg/L) and pH ($-$). Total dissolved solids were not selected because of very high correlations (usually >0.9 for most filters and seasons) with conductivity (Lee *et al*., 2005). Calibrated weights were assigned to each input variable. Biochemical oxygen demand (mg/L) or SS (mg/L) were the corresponding output variables. The input variables were selected due to their potential predictive relationships (based on correlation and regression analysis) with the BOD and SS (Scholz, 2003), and the fact that they are both more cost-effective and easier to measure in comparison to BOD and SS.

If the CBR system is presented with a new problem case (measurements at a particular day), the similarity of each past case with the problem case will be calculated. The most similar cases will subsequently be selected, and used to calculate the predicted output of the new problem case. The similarity of each past case with the problem case is calculated one case at a time and by comparing one input variable at a time. The local similarity (the similarity of a past case and the problem case with respect to only one variable) is found via a mathematical function of the difference between the two cases for one variable. In Equation (1), a variable *i* for the past case and problem case is normalized over the range of the past cases by dividing each case by the mean of the past cases. The differences between each past case and the problem case are then calculated with respect to each variable. The function *f* in Equation (1) converts the local difference to the local similarity.

$$\text{local\_sim}_i = f(|(V_{ip}/MV_i) - (V_{ic}/MV_i)|) \qquad (1)$$

where: local_sim$_i$ is the local similarity of variable *i* for past case c and problem case *p*; V_{ip} is the value of variable *i* for the problem case; MV_i is the mean of variable *i* found in the case base; V_{ic} is the value of variable *i* for the past case; $|(V_{ip}/MV_i) - (V_{ic}/MV_i)|$ is the local difference; and *f* is the function, which maps the local difference onto the local similarity.

The function used to map the local difference onto the local similarity is defined in Equation (2) that applies fuzzy theory such that a difference of zero scores a similarity of one and a difference of more than two standard

Table 1. Correlation coefficients from a correlation analysis comprising input (column headings) and target (row headings) variables used for a subsequent case-based reasoning analysis.

Variable	*Turbidity (NTU)*	*Conductivity (μS)*	*Redox potential (mV)*	*Temperature (°C)*	*Dissolved oxygen (mg/L)*	*pH (−)*
BOD[a] (mg/L)	0.535	0.244	−0.374	−0.121	−0.074	−0.242
SS[b] (mg/L)	0.531	0.833	−0.338	−0.323	−0.135	0.035

[a] Five-days at 20°C N-Allylthiourea biochemical oxygen demand; [b] suspended solids; Note: 5% significance level: 0.078; 1% significance level: 0.102.

deviations scores a similarity of zero. The global similarity (similarity of the past case to the problem case considering all variables) of a past case can be found from the local similarity of each variable. Each local similarity is first multiplied by a weighting factor that corresponds to the importance of that variable in predicting the output. These should be found by calibrating the system using an independent data set. As the calibration data set is introduced to the system the weighting of each factor should be adjusted one at a time until the best possible output is achieved. Equation (3) defines how the local similarities of each variable are combined to calculate the global similarity of the past case and problem case.

$$f(x) = \exp(-0.5(x/\mathrm{SDV}_i)^2) \tag{2}$$

Table 2. Constructed treatment wetlands: Case-based reasoning (CBR) applied to predict the 5-day at 20°C N-Allylthiourea biochemical oxygen demand (BOD) and the suspended solids (SS) concentrations with the input variables turbidity (NTU), conductivity (μS), redox potential (mV), and the outflow water temperature (°C).

Filter no.	*No. of cases*	*Mean concentration$_{measured}$*	*Mean concentration$_{predicted}$*	*Error (%)*	*Correct prediction of compliance (%)*[a]	a[b]	b[b]	r^{2}[b]
			CBR for BOD (mg/L)					
3	56	7.55	10.86	30.5	89.29	−0.035	2.038	0.258
4	55	12.69	9.92	27.9	87.27	−0.005	0.991	0.614
5	59	11.41	13.35	14.4	91.53	−0.013	1.703	0.731
6	60	16.20	9.85	64.5	81.67	−0.008	0.911	0.230
7	58	6.70	9.88	32.2	89.66	−0.040	2.386	0.546
8	60	11.05	11.05	0.0	95.00	−0.013	1.509	0.751
9	58	9.83	13.44	26.9	87.93	−0.017	1.826	0.513
10	62	18.75	11.11	68.8	87.10	−0.003	0.655	0.629
11	60	10.08	12.23	17.6	81.67	0.000	1.077	0.529
12	112	11.31	13.27	14.8	83.04	0.000	0.999	0.303
3–12	640	11.60	11.65	0.4	87.03	−0.007	1.125	0.396
			CBR for SS (mg/L)					
3	60	106.67	111.72	4.5	85.00	−0.001	1.231	0.863
4	42	58.20	94.43	38.4	78.57	0.000	0.554	0.349
5	58	146.86	121.50	20.9	84.48	−0.001	1.500	0.791
6	62	114.76	114.19	0.0	85.48	−0.001	1.297	0.792
7	57	127.61	111.87	14.1	84.21	0.000	1.124	0.688
8	62	91.76	86.18	6.5	87.10	0.000	0.970	0.960
9	60	91.72	97.86	6.3	88.33	0.000	1.000	0.614
10	64	81.05	94.04	13.8	89.06	0.000	1.138	0.924
11	62	82.43	81.72	0.0	88.71	−0.001	1.246	0.851
12	110	134.86	108.84	23.9	77.27	0.000	0.884	0.808
3–12	637	111.50	100.52	10.9	84.46	−0.001	1.140	0.732

[a] The likelihoods of correct predictions, if the effluent concentrations are either below or above the thresholds for secondary wastewater treatment. The BOD and SS concentrations for compliance are 20 and 30 mg/L, respectively; [b] $\mathrm{concentration}_{\mathrm{predicted}} = a \times \mathrm{concentration}_{\mathrm{measured}}{}^2 + b \times \mathrm{concentration}_{\mathrm{measured}} + c$, where $c = 0$ and r^2 = coefficient of determination.

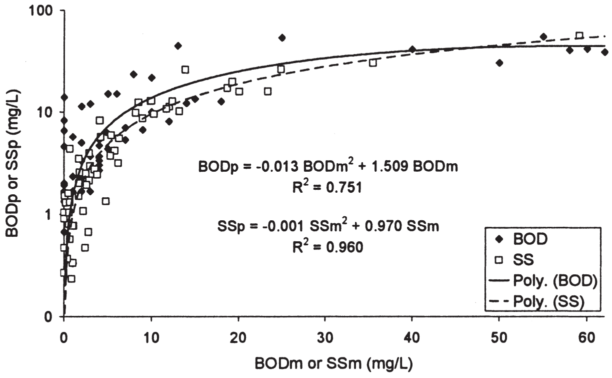

Figure 1. Regression analysis between the measured 5-day at 20°C N-Allylthiourea biochemical oxygen demand $BOD_{measured}$ (BODm) and $BOD_{predicted}$ (BODp), and between the measured suspended solids $SS_{measured}$ (SSm) and $SS_{predicted}$ (SSp) for Filter 8. The case base contained the input variables turbidity, conductivity, redox potential, and the outflow water temperature. The following SS entries are beyond the displayed range: (3.8, 0.023), (70.8, 55.234), and (82.1, 73.32).

where x is the local difference; f is the function, which converts the local difference into the local similarity; and SDV_i is the standard deviation of the differences of variable i found in the case base of past cases.

$$\text{Glob\_sim} = \Sigma[(\text{local\_sim}_i{*}w_\text{i})/\Sigma w_i \times 100] \qquad (3)$$

where: i is 1, 2, . . . , n; n is the number of variables used to represent a case; w_i is the weighting associated with variable i; and Local_sim_i is the local similarity of the past case and problem case for variable i.

When the global similarity of each past case with the problem case is found, the past cases can be ranked in order of their corresponding sum to decide which of the past cases would be deemed similar enough to be selected for adaptation. The three to five past cases with the highest similarity rankings were chosen in this study. Tests undertaken on different sets of data show that between two and six cases are usually sufficient to achieve the best performance. The CBR usually requires a relatively large data set for optimization exercises.

Equations (4) and (5) show how a prediction is made for the target variables of the problem case by combining the numerical value of the target variable for the three to five selected cases.

$$\text{Proportion } P_j = \text{Glob\_sim}_j/\text{Glob\_sim}_\text{T} \qquad (4)$$

where: j is 1, 2, 3, 4, or 5; P_j is the proportion of the prediction that is obtained from the past case j; and Glob_sim_T is the sum of the global similarities of the three to five selected cases.

$$\text{Prediction } P = \Sigma(P_j \times \text{TV}_j) \qquad (5)$$

where P_j is the proportion of the prediction that is obtained from the past case j; and TV_j is target variable of past case j.

CBR RESULTS AND DISCUSSION

Correlation analysis

T1

Table 1 summarizes the findings from a correlation analysis comprising input (turbidity, conductivity, redox potential, outflow water temperature, dissolved oxygen, and pH) and target (BOD and SS) variables. These findings are used for a subsequent CBR analysis. Correlations were strong between BOD and turbidity, SS and turbidity, and SS and conductivity (at the 1% significance level). Therefore, turbidity and conductivity are likely to be the most important input variables.

Comparison of different filters

T2
F1

Table 2 shows the application of a CBR system for the prediction of the outflow BOD and SS. Figure 1 visualizes the regression analysis between measured and predicted BOD and measured and predicted SS for Filter 8 (typical UK reed bed; Lee *et al*. 2005). The associated case base contained the following input variables: turbidity, conductivity, redox potential, and outflow water temperature. The application of second-order polynomial trendlines results in very good fits for both target variables.

The likelihoods of correct predictions if the effluent concentrations are either below or above the thresholds for secondary wastewater treatment are also shown in

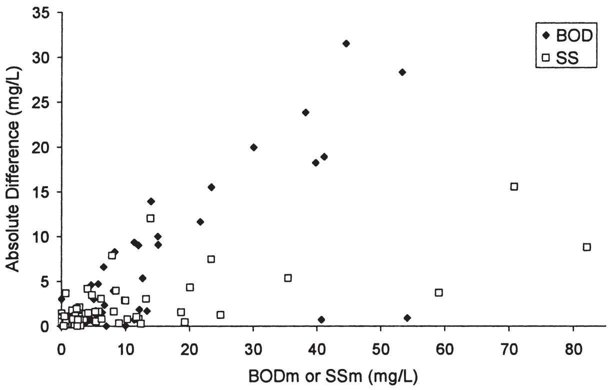

Figure 2. Distribution of absolute differences between measured and predicted concentrations for 5-day at 20°C N-Allylthiourea biochemical oxygen demand (BODm) and measured and predicted suspended solids (SSm) for Filter 8. The case base contained the input variables turbidity, conductivity, redox potential, and the outflow water temperature. The BOD entry (24.88, 75.12) is beyond the displayed range.

Table 3. Comparison between extended storage (Filters 1 and 2) and Constructed treatment wetlands (Filters 3 to 11) for a similar retention time: case-based reasoning (CBR) applied to predict the 5-day at 20°C N-Allylthiourea biochemical oxygen demand (BOD) and the suspended solids (SS) concentrations with the input variables turbidity (NTU), conductivity (μS), redox potential (mV) and the outflow water temperature (°C).

Filter no.	*No. of cases*	*Mean concentration$_{measured}$*	*Mean concentration$_{predicted}$*	*Error (%)*	*Correct prediction of compliance (%)*[a]	a[b]	b[b]	r^{2b}
			CBR for BOD (mg/L)					
1–2	109	35.52	35.16	1.0	68.81	−0.005	1.282	0.249
3–11	528	11.66	11.31	3.1	87.88	−0.001	1.083	0.423
			CBR for SS (mg/L)					
1–2	132	304.60	243.40	25.1	84.85	0.000	0.968	0.510
3–11	527	106.60	98.80	7.9	85.96	−0.001	1.202	0.724

[a] The likelihoods of correct predictions, if the effluent concentrations are either below or above the thresholds for secondary wastewater treatment. The BOD and SS concentrations for compliance are 20 and 30 mg/L, respectively. [b] $\text{concentration}_{\text{predicted}} = a \times \text{concentration}_{\text{measured}}{}^2 + b \times \text{concentration}_{\text{measured}} + c$, where $c = 0$ and r^2 = coefficient of determination.

Table 2. The BOD and SS concentrations for compliance are 20 and 30 mg/L, respectively. The correct predictions of compliance were all >77%. The probabilities are therefore all at least by 0.27 higher in comparison to pure guessing. The predictions are encouraging and support the potential for future use of CBR as a management tool for the day-to-day process control.

F2

Figure 2 shows the distribution of prediction errors for a selected CBR result (Filter 8; typical UK reed bed; Cooper *et al.* (1996)) visualized in Fig. 1. Where the distribution is clustered, the errors are least. This is promising, as it appears that if the density of the case base can be increased, then the error can be reduced further (see below). More cases are likely to lead to a better CBR performance.

Moreover, research has shown that new case-based methods that utilize the cluster information of data sets are likely to be superior to conventional CBR systems (Verdenius and Broeze, 1999; Roh *et al.*, 2003; Yang *et al.*, 2004). Despite the greater variability of SS in contrast to BOD (Lee *et al.*, 2005), SS has smaller absolute differences between measured and predicted concentra-

Table 4. Optimization of input variable combinations for Filters 3 to 12: Case-based reasoning (CBR) applied to predict the 5-day at 20°C N-Allylthiourea biochemical oxygen demand (BOD) and the suspended solids (SS) concentrations.

Input variables	*No. of cases*	*Mean concentration$_{measured}$*	*Mean concentration$_{predicted}$*	*Error (%)*	*Correct prediction of compliance (%)*[a]	a[b]	b[b]	r^{2b}
			CBR for BOD (mg/L)					
1	640	11.60	11.48	1.0	82.66	−0.005	0.874	0.092
1+2	640	11.60	14.33	19.1	80.94	0.000	0.543	−0.015
1+2+3	640	11.60	11.24	3.2	84.84	−0.006	0.997	0.236
1+2+3+4	640	11.60	11.65	0.4	87.03	−0.007	1.125	0.396
1+2+3+4+5	640	11.60	11.39	1.8	85.47	−0.006	1.059	0.413
			CBR for SS (mg/L)					
1	637	111.50	124.71	10.6	54.47	−0.001	0.964	−0.092
1+2	637	111.50	102.76	8.5	85.56	−0.001	1.174	0.685
1+2+3	637	111.50	97.15	14.8	85.71	−0.001	1.160	0.714
1+2+3+4	637	111.50	100.52	10.9	84.46	−0.001	1.140	0.732
1+2+3+4+5	637	111.50	98.80	16.9	82.52	−0.001	1.149	0.706

The case base contained the following input variables: 1 = turbidity (NTU); 2 = conductivity (μS); 3 = redox potential (mV); 4 = outflow water temperature (°C); 5 = dissolved oxygen (mg/L); [a] the likelihoods of correct predictions, if the effluent concentrations are either below or above the thresholds for secondary wastewater treatment. The BOD and SS concentrations for compliance are 20 and 30 mg/L, respectively; [b] $\text{concentration}_{\text{predicted}} = a \times \text{concentration}_{\text{measured}}{}^2 + b \times \text{concentration}_{\text{measured}} + c$, where $c = 0$ and r^2 = coefficient of determination.

tions (Fig. 2) than BOD. It follows that relatively high raw data variability is not necessarily an indication for an underperforming CBR analysis as can be intuitively expected.

The system typically achieved an 85% success rate for predicting whether or not the water samples met regulatory requirements (Table 2). The theoretical probability of the system predicting a correct answer is 0.5 (right or wrong), based on the number of cases below or above the threshold in the case base and the actual concentration of the target variable of the test cases used. In comparison, prediction errors of up to 54% were recorded for an unrelated project previously, compared to 17% corresponding to pure guessing (Wendler and Bach, 2004).

Table 2 suggests that the system is fit for purpose considering relatively high coefficients of determination r^2, particularly for measured and predicted SS. However, the system requires optimization to further increase the accuracy. Optimization measures would be certain to include the selection of cases with greater process control and data availability. The aim would be to reduce the number of unknown variables.

Extended storage vs. constructed wetlands

Table 3 shows a CBR comparison between extended storage (Filters 1 and 2) and constructed treatment wetlands (Filters 3 to 11) characterized by similar retention times. CBR was applied to predict the BOD and SS concentrations. The case base contained turbidity, conductivity, redox potential, and outflow water temperature as input variables. Concerning the prediction of BOD, relatively high likelihoods of correct predictions were achieved for the wetlands but not for extended storage. This is surprising, considering that extended storage is a much simpler process with a reduced number of unknown variables. However, the buffering capacity of the system is low, and data variability is subsequently higher than for the wetlands. T3

Table 5. Unbiased assessment of data subsets for Filters 3 to 12: Case-based reasoning (CBR) applied to predict the 5-day at 20°C N-Allylthiourea biochemical oxygen demand (BOD) and the suspended solids (SS) concentrations with the input variables turbidity (NTU), conductivity (μS), redox potential (mV), and the outflow water temperature (°C).

Selection	*No. of cases*	*Mean concentration$_{measured}$*	*Mean concentration$_{predicted}$*	*Error (%)*	*Correct prediction of compliance (%)*[a]	a[b]	b[b]	r^{2b}
			CBR for BOD (mg/L)					
1 out of 15	40	11.05	13.52	18.3	83.33	−0.008	1.106	0.052
1 out of 10	58	11.05	11.15	0.9	95.00	−0.005	0.976	0.476
1 out of 6	90	11.05	10.46	1.7	90.00	−0.011	1.216	0.363
1 out of 5	116	11.05	10.87	1.7	91.67	−0.010	1.157	0.231
1 out of 4	144	11.05	10.45	5.7	93.33	−0.003	0.826	0.552
1 out of 3	194	11.05	12.30	10.2	93.33	−0.010	1.399	0.658
1 out of 2	290	11.05	10.48	5.4	90.00	−0.007	1.060	0.585
2 out of 3	385	11.05	9.28	19.1	90.00	0.011	1.208	0.564
3 out of 4	435	11.05	10.15	8.9	93.33	−0.012	1.369	0.655
4 out of 5	464	11.05	10.99	0.5	95.00	−0.010	1.359	0.805
1 out of 1	580	11.05	11.05	0.0	95.00	−0.013	1.509	0.751
			CBR for SS (mg/L)					
1 out of 15	39	91.76	85.76	7.0	69.4	−0.001	0.952	0.092
1 out of 10	57	91.76	68.84	33.3	75.8	0.000	0.628	0.400
1 out of 6	97	91.76	84.07	9.1	79.0	0.000	0.885	0.761
1 out of 5	115	91.76	86.99	5.5	77.4	0.000	0.869	0.686
1 out of 4	144	91.76	79.28	15.7	85.5	−0.001	1.061	0.810
1 out of 3	192	91.76	101.97	10.0	85.5	0.000	0.956	0.872
1 out of 2	288	91.76	85.59	7.2	90.3	0.000	0.990	0.923
2 out of 3	383	91.76	86.83	5.7	87.1	0.000	0.973	0.925
3 out of 4	433	91.76	89.10	3.0	83.9	0.000	1.022	0.936
4 out of 5	461	91.76	81.18	13.0	88.7	0.000	1.041	0.939
1 out of 1	575	91.76	86.18	6.5	87.1	0.000	0.970	0.960

[a] The likelihoods of correct predictions, if the effluent concentrations are either below or above the thresholds for secondary wastewater treatment. The BOD and SS concentrations for compliance are 20 and 30 mg/L, respectively; [b] $\text{concentration}_{\text{predicted}} = a \times \text{concentration}_{\text{measured}}^2 + b \text{ x } \text{concentration}_{\text{measured}} + c$, where $c = 0$ and r^2 = coefficient of determination.

Optimization of the numbers and type of input variables

T4

Table 4 summarizes the findings of an input variable combination optimization exercise. The purpose was to estimate both BOD and SS with as few input variables as possible to reduce costs and effort. The case base therefore contained the following input variables in order of priority (Table 1): turbidity, conductivity, redox potential, outflow water temperature and dissolved oxygen. All predictions of compliance were high except for the prediction of SS with turbidity alone despite a relatively high corresponding correlation coefficient of 0.531 (Table 1). It follows that the first two variables (turbidity and conductivity) that are inexpensive to obtain are sufficient to predict the most important, expensive, and time-consuming to obtain target variables BOD and SS.

Optimization of the size of the input database

T5

The data set of this case study has been described in detail by Lee *et al.* (2005). Table 5 shows an unbiased assessment of data subsets to optimize the size of the input database. The case base contained turbidity, conductivity, redox potential, and outflow water temperature as input variables, because the combination of these variables has the highest mean prediction compliance percentage (Table 4). The data subsets were selected systematically (in sequence; x out of y, where x ≤ y), but technically at random. The probabilities of all filters and input variables to contribute to any calculation were statistically the same; 0.1 and 0.25, respectively.

In contrast to traditional curve fitting techniques, the CBR system is capable of picking up rapidly fluctuating trends among the different input variables (Lee *et al.*, 2005), because the distribution of cases is relatively dense (Table 5). Only neighboring cases will be picked up for relatively small case bases. In comparison, a large case base is likely to be beneficial, if data are sparse and not erratic. This may be the case for most data sets in physics and mechanical engineering but not environmental engineering and science. It follows that the distribution and density of cases, and the relationships between the variables (gradual or erratic trends) should be considered when selecting the optimum number of cases.

CBR is well suited for relatively highly variable water quality data sets such as those from constructed treatment wetlands. Little domain knowledge is required, and the optimum number of cases can be selected by trial and error (Table 5). Findings show that the case study data set could be reduced by 75%, and that BOD and SS can still be predicted reasonably well with four inexpensive variables measured only every 2 weeks. Nevertheless, the CBR should be calibrated with cases of known output to minimize the error.

CONCLUSIONS

CBR was successfully applied to predict BOD and SS, but there is room for improvement by applying optimization techniques to control the variances of the input variables. This would lead to a relatively accurate data set that should be used to calibrate the system. BOD and SS are expensive to estimate, and can be cost-effectively controlled by applying CBR with the input variables turbidity and conductivity and possibly also redox potential.

The CBR system showed better performance for constructed wetlands ("buffers" due to biomass between aggregates) in comparison to extended storage (no "buffer" capacity). Small data sets based on 2-week sampling were sufficient to monitor the water quality.

This paper demonstrates to the reader the successful application of CBR to typical "black box" systems such as constructed wetlands governed by biochemical processes. This paper may also find use as a learning aid for water and environmental engineers and managers.

ACKNOWLEDGMENTS

The authors acknowledge support provided by Mr. J. Glancey and Dr M. Crapper. Sponsor: The University of Edinburgh.

REFERENCES

AAMODT, A., and PLAZA, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications* **7,** 39–59.

AMERICAN PUBLIC HEALTH ASSOCIATION (APHA). (1995). *Standard Methods for the Examination of Water and Wastewater*, 19th ed. Washington, DC: author.

ARDITI, D., and TOKDEMIR, O.B. (1999). Comparison of case-based reasoning and artificial neural networks. *J. Comp. Civ. Eng*. **13,** 162–168.

COLILLA, M., FERNANDEZ, C.J., and RUIZ-HITZKY, E. (2002). Case-based reasoning (CBR) for multicomponent analysis using sensor arrays: Application to water quality evaluation. *Analyst* **127,** 1580–1582.

COOPER, P.F., JOB, G.D., GREEN, M.B., and SHUTES, R.B.E. (1996). *Reed Beds and Constructed Wetlands for Wastewater Treatment*. Swindon, UK: WRc plc.

GERNAEY, K.V., VAN LOOSDRECHT M.C.M., HENZE, M., LIND, M., and JØRGENSEN, S.B. (2004). Activated sludge wastewater treatment plant modeling and simulation: state of the art. *Environ. Model. Software* **19,** 763–783.

LEE, B.-H., SCHOLZ, M., and HORN, A. (2005). Constructed wetlands: Treatment of concentrated storm water runoff (Part A). *Environ. Eng. Sci.*

NUNEZ, H., SANCHEZ-MARRE, M., CORTES, U., COMAS, J., MARTINEZ, M., RODRIGUEZ-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. *Environ. Model. Software* **19,** 809–819.

POLICASTRO, C.A., CARVALHO, A.C.P.L.F., and DELBEM, A.C.B. (2004). A hybrid case-based reasoning approach for monitoring water quality. *Lect. Notes Comput. Sci.* **3029,** 492–501.

RODA, I.R., SANCHEZ-MARRE, M., COMAS, J., CORTES, U., and POCH, M. (2001). Development of a case-based system for the supervision of an activated sludge process. *Environ. Technol.* **22,** 477–486.

RODRIGUEZ-RODA, I., SANCHEZ-MARRE, M., COMAS, J., BAEZA, J., COLPRIM, J., LAFUENTE, J., CORTES, U., and POCH, M. (2002). A hybrid supervisory system to support WWTP operation: implementation and validation. *Water Sci. Technol.* **45,** 289–297.

ROH, T.H., OH, K.J., and HAN, I. (2003). The collaborative filtering recommendation based on SOM cluster-indexing CBR. *Expert System Appl.* **25,** 413–423.

SANCHEZ-MARRE, M., CORRES, U., RODA, I.R., and POCH, M. (1999). Sustainable case learning for continuous domains. *Environ. Model. Software* **14,** 349–357.

SCHOLZ, M. (2003). Case study: Design, operation, maintenance and water quality management of sustainable storm water ponds for roof runoff. *Bioresource Technol.* **95,** 269–279.

SCHOLZ, M., HÖHN, P., and MINALL, R. (2002). Mature experimental constructed wetlands treating urban water receiving high metal loads. *Biotechnol. Prog.* **18,** 1257–1264.

VERDENIUS, F., and BROEZE, J. (1999). Generalized and instant-specific modeling for biological systems. *Environ. Model. Software*. **14,** 339–348.

WATSON, I. (1997). *Applying Case-based Reasoning Techniques for Enterprise Systems*. San Francisco, CA: Morgan Kauhmann.

WENDLER, J., and BACH, J. (2004). Recognizing and predicting agent behavior with case-based reasoning. *Lect. Notes Comput. Sci.* **3020,** 729–738.

YANG, B.S., HAN, T., and KIM, Y.S. (2004). Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis. *Expert Syst. Appl.* **26,** 387–395.