ELSEVIER

Artificial Intelligence in Engineering 14 (2000) 307–317

ARTIFICIAL INTELLIGENCE *in* ENGINEERING

www.elsevier.com/locate/aieng

Prediction of the bulking phenomenon in wastewater treatment plants

L. Belanche[a,*], J.J. Valdés[a], J. Comas[b], I.R. Roda[b], M. Poch[b]

[a]*Secció d'Intel·ligència Artificial., Departamento de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, c/Jordi Girona 1-3, 08034 Barcelona, Spain*

[b]*Laboratori d'Enginyeria Química i Ambiental, Facultat de Ciències, Universitat de Girona, Campus de Montilivi s/n, 17071 Girona, Spain*

Received 6 July 1999; revised 6 June 2000; accepted 6 June 2000

Abstract

The control and prediction of wastewater treatment plants poses an important goal: to avoid breaking the environmental balance by always keeping the system in stable operating conditions. It is known that *qualitative* information — coming from microscopic examinations and subjective remarks — has a deep influence on the activated sludge process. In particular, on the total amount of effluent suspended solids, one of the measures of overall plant performance. The search for an input–output model of this variable and the prediction of sudden increases (*bulking* episodes) is thus a central concern to ensure the fulfillment of current discharge limitations. Unfortunately, the strong interrelation between variables, their heterogeneity and the very high amount of missing information makes the use of traditional techniques difficult, or even impossible. Through the combined use of several methods — rough set theory and artificial neural networks, mainly — reasonable prediction models are found, which also serve to show the different importance of variables and provide insight into the process dynamics.

Keywords: System identification; Wastewater treatment plants; Neural networks; Rough sets; Environmental modeling; Bulking

1. Introduction

The world is an enormously complex system, where the environmental problems are essentially multi-faceted and demand at least a nodding acquaintance with many previously separated specialisms — ecology, economics, sociology, technology, physics, chemistry, and so on [1]. In particular, dirty water is both the world's greatest killer and its biggest single pollution problem [2]. The large amount of wastewater generated in industrialized societies is one of the main environmental pollution aspects that must be seriously considered. New directives and regulations (for EU countries, the European directive of the Council 91/271/EEC) have guaranteed the construction of specific plants to treat these wastewaters. *Activated sludge* systems are the most extensively used in wastewater treatment plants (WWTP). In an activated sludge process, the wastewater, which contains organic matter, suspended solids and nutrients, enters an aerated tank where it is mixed with biological floc particles. After a sufficient contact time, this mixture is discharged into a settler that separates the suspended biomass from the treated water. Most of the biomass is recirculated to the aeration tank, while a little amount is purged daily (see Fig. 1).

Activated sludge is a clear example of an environmental process that is really difficult to understand, and thus difficult to be correctly operated and controlled. The inflow is variable (both in quantity and in quality); not only is there a living catalyst (the microorganisms) but also a population that varies over time, both in quantity and in the relative number of species; the knowledge of the process is scarce, there are few and unreliable on-line analyzers, and most of the data related to the process are subjective and cannot be numerically quantified. It is known that most of the problems of poor activated sludge effluent quality result from the inability of the secondary settler to efficiently remove the suspended biomass from the treated water. When the biomass is strongly colonized by long filamentous bacteria, holding the flocs apart and hindering sludge settlement, the amount of *Total suspended solids* (TSS) at the outflow of the plant increases seriously. Although this phenomenon, called *bulking*, has been extensively studied, the interrelations and diversity of the many bacterial species

* Corresponding author. Tel.: +34-93-401-5644; fax: +34-93-401-7014.
E-mail addresses: belanche@lsi.upc.es (L. Belanche), valdes@lsi.upc.es (J.J. Valdés), quim@lequia1.udg.es (J. Comas), ignasi@lequia1.udg.es (I.R. Roda), manel@lequia1.udg.es (M. Poch).

0954-1810/00/$ - see front matter
PII: S0954-1810(00)00012-1

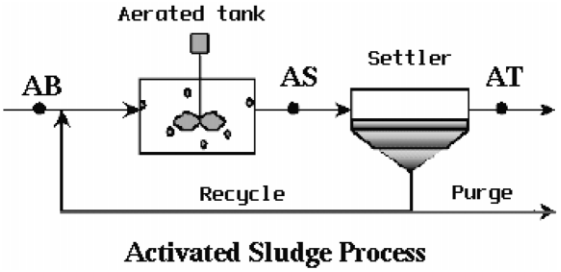

Fig. 1. Flow diagram of the activated sludge process. The influent stream (sample point AB) is combined with the sludge return stream *Recycle* and sent to the aerated tank (sample point AS) for biological oxidation of the organic matter. A settler is then used to remove treated water (sample point AT) from biomass and to thicken it. The withdrawn sludge *Purge* is concentrated to a higher solids content in the sludge line of process.

involved, and the uncertainty about the factors triggering their growth constitute obstacles to a thorough and clearcut understanding of the problem [3].

2. Previous work and objectives

Operation, control and supervision of WWTPs have been approached from many different points of view, including classical control methods [4–7], mechanistic models [8,9], knowledge-based systems [10–12], case-based reasoning [13], neural nets [3,14] and hybrid approaches [15,16]. However, a direct cause–effect relationship for WWTP performance has been established only in a few cases and, even in those, experimental results could lead to contradictory conclusions, avoiding the formulation of deterministic cause–effect relationships that could be used as prediction models. The identification of a model to predict in real-time — with reasonable accuracy — the appearance of sludge bulking is therefore of great practical importance, in view of the potential for improvement in plant efficiency and cost savings [17]. This model should allow to obtain an accurate estimation of TSS ranges at the outflow of the plant, based on the relationship among the most relevant variables of the process, both quantitative (e.g. flow rates and analytical results) and qualitative (biomass microscopic examinations and process observations), in order to know whether the plant is meeting the discharge permit requirements.

To achieve this, several staged studies have already been performed towards the development of input–output behavior models for WWTPs, in which the temporal behavior of two outgoing variables (COD-AT and BOD-AT, see below) was correctly captured and reproduced [18]. The following step has been taking into account *qualitative* information — not considered in the previous studies — and exploring how it affects the formation of a predictive model for the TSS in the outflow of the plant. As mentioned above, it is known that qualitative features — including microscopic examinations of microfauna and bacteria, and some subjective information — are useful indicators of overall process performance, and strongly influence the activated sludge process. Qualitative information is usually put aside because of its nature and the high levels of missing values that it brings along, both being a nuisance — if not a problem — for many learning algorithms and models, which have to accommodate qualitative and missing information in a deformative preprocessing.

Previous works [3] have applied stochastic models and neural networks to accurately predict the occurrence of future bulking episodes. This study uses 14 months of complete daily measurements of quantitative data only, from the Jones Island WWTP in Milwaukee, WI. Although the study is based on real data, it is not common (at least in Europe) to make daily analytical measurements of all process variables (e.g. organic matter and total suspended solids are typically measured twice or three times a week). As a result, the databases are full of missing values, evenly distributed all over time. Incidental equipment failures also bring along compact chunks of missing data. This high incidence of missing information is the main reason why most of the other studies are based on simulated data.

There is also a second need to handle uncertain or imprecise information, a characteristic present in all kinds of variables, especially in all numeric measurements coming from on-line analyzers, but also in analytical determinations and qualitative observations.

The *aim of the work* can therefore be regarded as threefold:

1. The development of simple prediction models for effluent TSS, as an indication of plant performance. By simple we mean taking into account only those variables that are really useful.
2. The study of the relative influence of qualitative information, which is usually ignored.
3. The analysis of several learning approaches, in a difficult task characterized by imprecision, heterogeneity and high incidence of missing information.

The *methods* used throughout the work fall within what is nowadays labeled as *soft computing*, among which we find rough sets, fuzzy sets, evolutionary search methods and artificial neural network learning. In particular, time-delay neural networks of three kinds are used: classical (trained with simulated annealing plus the conjugate gradient), probabilistic (trained as a Bayes–Parzen classifier) and heterogeneous (trained with genetic algorithms).

The *results* show that qualitative information exerts a considerable influence on plant output, although quite variable, since high degrees of information redundancy are discovered. Comparable predictive capabilities are obtained when working with a much-reduced set of variables, which coincide with those highly rated by WWTP experts. Also, a common upper bound in classification accuracy is discovered, in light of the coherent results yielded by methods that

Table 1
Selected variables characterizing the behavior of the studied WWTP

Sample point	On-line data (flow rates)	Analytical data	Qualitative data
AB (inflow)	Q-AB (inflow)	COD-AB, BOD-AB (organic matter) TSS-AB (suspended solids)	
	Q-R (biological recycle)		Presence of foam
AS (bioreactor)	Q-P (biological purge)		Microfauna (*Aspidisca*, *Vorticella*, etc.)
	Q-A (biological aeration)		Filamentous bacteria (*Nocardia*, *Thiothrix*, etc.)
AT (outflow)		COD-AT, BOD-AT, TSS-AT	Look (appearance)

are very different in nature. In addition, despite the high levels of missing information, very reasonable prediction models are found.

3. The WWTP database

The historical database used throughout the work corresponds to a WWTP of a touristic resort in the Costa Brava (Catalonia, Spain). This plant removes organic matter and TSS contained in the raw water of about 30,000 inhabitants-equivalents in winter and about 150,000 in summer. This database comprises a large amount of quantitative and qualitative variables corresponding to an exhaustive characterization of the main points of the plant, such as the inflow, the bioreactor and the outflow, indicated in Table 1 as -AB, -AS and -AT, respectively (see also Fig. 1).

Quantitative information includes analytical results of water quality: organic matter, measured as chemical (COD) and biochemical (BOD) oxygen demand, and the studied total suspended solids, measured as TSS, together with on-line signals coming from sensors: inflow or Q-AB, recycle or Q-R, purge or Q-P and aeration or Q-A flow rates. Qualitative data include information about the presence of foam in the bioreactor ("Presence-foam"), the subjective appearance of outflow ("Look") and daily microscopic examinations (basically, presence of microfauna, e.g. *Aspidisca*, *Vorticella*, and some filamentous organisms, e.g. *Nocardia*, *M. Parvicella*) [20].

The final data set covers a homogeneous representative period of 609 consecutive days, where each day is considered as a new sample. Basic statistical descriptors of the variables included in the database are shown in Table 2 (quantitative variables) and Table 3 (qualitative ones). The relative abundance of qualitative variables is categorized in three different levels: *none*, *some* and *many*, with the exception of the outflow appearance (i.e. "Look-AT"), categorized as *poor*, *fair* and *good*.

Table 2
Basic statistical descriptors for selected quantitative WWTP variables (in 609 days)

Variable	Unit	Missing	Mean	St. dev.
Q-AB	m^3/d	18	10,707.0	3634.0
COD-AB	mg/l	380	795.8	198.0
BOD-AB	mg/l	480	390.7	95.7
TSS-AB	mg/l	380	315.9	91.4
Q-R	m^3/d	1	5597.7	2287.1
Q-P	kg TSS/d	11	771.6	756.6
Q-A	kg O_2/d	61	4138.6	1878.4
COD-AT	mg/l	380	55.8	18.5
BOD-AT	mg/l	480	9.0	4.9
TSS-AT	mg/l	376	9.6	5.8

The most relevant feature of the database is the extremely high incidence of missing values (between 60 and 80%, approximately). This is specially true in the case of the outflow variables COD-AT, BOD-AT and TSS-AT — more suitable as targets for developing prediction models — the variables characterizing water quality at the inflow COD-AB, BOD-AB and TSS-AB, and qualitative variables characterizing the microorganisms. Because of this, the final database processed includes only those days with a recorded value in the studied variable TSS-AT, causing the initial data matrix to shrink from 609 to 233 days (Table 2, last row). Nevertheless, the rate of missing values is still extremely high among potential predictor variables. In addition, the complexity of the WWTP behavior problem is reflected in the frequency distribution of its variables. As an example, Kolmogorov–Smirnov tests applied to the incoming TSS-AB and outgoing TSS-AT variables confirm what direct inspection suggests: whilst the first variable is distributed normally, the second is not. Actually it has a right-skewed distribution, reflecting strong non-linear distortions introduced by the WWTP dynamics (see Fig. 2). All these features make the search for models to characterize this aspect of WWTP behavior considerably hard. They must be always taken into account when evaluating the quality of the learned models.

4. Description of the methods

Four techniques are studied in this work to investigate the influence and classification ability of qualitative variables on TSS-AT ranges: heterogeneous neural networks [21,22], classical neural networks [23], probabilistic networks [24] and the *k*-nearest neighbors algorithm. Rough set theory [25] is also used to perform a reduction of dimension. A brief description is given on heterogeneous and probabilistic networks. Rough set methodology, as used in this work, is also outlined.

Table 3
Basic statistical descriptors for qualitative WWTP variables. The last three columns show the number of days for each variable and category

Variable (609 days)		Category		
	Missing	None	Some	Many
Presence-foam	394	17	153	45
Zooglea	394	117	69	29
Nocardia	399	90	51	69
Thiothrix/021N	396	112	85	16
Type 0041	397	140	44	28
M. Parvicella	395	156	23	35
Aspidisca	503	8	82	16
Euplotes	438	154	16	1
Vorticella	501	4	89	15
Epistylis	501	9	81	18
Opercularia	450	126	27	6
Carniv. ciliates	394	160	48	7
Flagell. >20 μm	394	184	23	8
Flagell. <20 μm	394	176	24	15
Amoebae	394	173	38	4
Testate *amoebae*	394	206	8	1
Rotifer	394	117	97	1
		Poor	Fair	Good
Look-AT	394	9	168	38

4.1. Heterogeneous neural networks

These artificial networks are neural architectures built out of neuron models allowing for heterogeneous and imprecise inputs, defined as a mapping $h : \hat{\mathscr{H}}^n \rightarrow \mathscr{R}_{out} \subseteq \mathbb{R}$. Here $\mathbb{R}$ denotes the reals and $\hat{\mathscr{H}}^n$ is a cartesian product of an arbitrary number, n, of *source sets*. These source sets may be extended reals $\hat{\mathscr{R}}_i = \mathbb{R}_i \cup \{\mathscr{X}\}$, extended families of (normalized) fuzzy sets $\hat{\mathscr{F}}_i = \mathscr{F}_i \cup \{\mathscr{X}\}$ and extended finite sets of the form $\hat{\mathscr{O}}_i = \mathscr{O}_i \cup \{\mathscr{X}\}$, $\hat{\mathscr{M}}_i = \mathscr{M}_i \cup \{\mathscr{X}\}$, where each of the $\mathscr{O}_i$ has a full order relation, while the $\mathscr{M}_i$ do not. The special symbol $\mathscr{X}$ extends the source sets and denotes the unknown element (missing information), behaving as an incomparable element w.r.t. any ordering relation. According to this definition, neuron inputs are vectors composed of n elements among which there might be reals, fuzzy sets, ordinals, categorical and missing data.

A heterogeneous neuron computes a *similarity index*, or proximity relation, followed by the familiar form of a squashing non-linear function with domain in [0,1]. We use in this work a *Gower-like* similarity index [26] in which the computation for heterogeneous entities is constructed as a weighted combination of partial similarities over single variables. This coefficient has its values in the real interval [0,1] and for any two objects i, j given by tuples of cardinality n, is given by the expression:

$$s_{ij} = \frac{\sum_{k=1}^{n} g_{ijk}\delta_{ijk}}{\sum_{k=1}^{n} \delta_{ijk}}$$

where g_{ijk} is a similarity *score* for objects i, j according to their value for variable k. These scores are in the interval [0,1] and are computed according to different schemes for numeric and qualitative variables. In particular, for a continuous variable k and any two objects i, j the following similarity score is used:

$$g_{ijk} = 1 - \frac{|v_{ik} - v_{jk}|}{\text{range}(v_{\cdot k})}$$

Here, v_{ik} denotes the value of object i for variable k and

$$\text{range}(v_{\cdot k}) = \sup_{i,j} |v_{ik} - v_{jk}|$$

The similarity measure used for categorical variables is the overlap:

$$g_{ijk} = \begin{cases} 1 & \text{if } v_{ik} = v_{jk} \\ 0 & \text{if } v_{ik} \neq v_{jk} \end{cases}$$

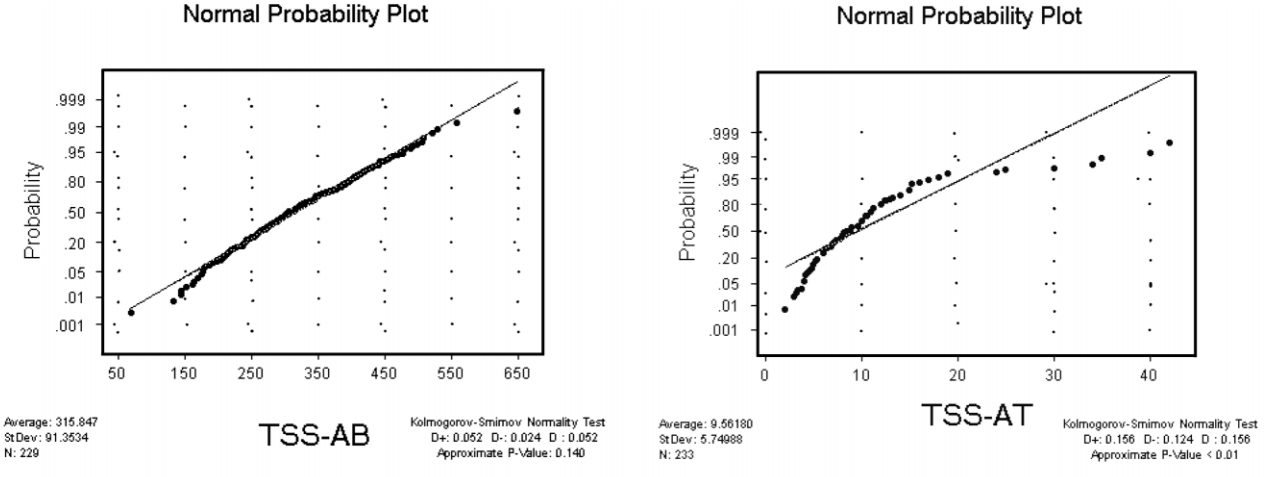

Fig. 2. Normal probability plots of total suspended solids for the Kolmogorov–Smirnov test, for incoming (TSS-AB) and outgoing (TSS-AT) total suspended solids.

The δ_{ijk} is a binary function expressing whether both objects are *comparable* or not according to their values w.r.t. variable k, as follows:

$$\delta_{ijk} = \begin{cases} 1 & v_{ik} \neq \mathscr{X} \text{ and } v_{jk} \neq \mathscr{X} \\ 0 & \text{otherwise} \end{cases}$$

For variables representing fuzzy sets, similarity relations from the point of view of fuzzy theory have been defined elsewhere [27,28], and different choices are possible. In the present case, the situation is not that of a fuzzy similarity relation defined on real values, but a crisp relation between fuzzy entities.

In possibility theory two functions, called *possibility* (Π) and *necessity* (N), measure events by the degree of *unsurprisingness* and *acceptance*, respectively [29]. In particular, the first of these functions expresses the possibility of co-occurrence or simultaneity of two vague propositions, with a value of one standing for absolute certainty. For two fuzzy sets $\tilde{A}, \tilde{B}$ of a reference set X, possibility is defined as:

$$\Pi_{\tilde{A}}(\tilde{B}) = \sup_{u \in X} (\mu_{\tilde{A} \cap \tilde{B}}(u))$$

where $\mu_{\tilde{A} \cap \tilde{B}}(u) = \min(\mu_{\tilde{A}}(u), \mu_{\tilde{B}}(u))$. Notice that this measure is reflexive in the strong sense and also symmetric. In our case, if $\mathscr{F}_i$ is an arbitrary family of fuzzy sets, and $\tilde{A}, \tilde{B} \in \mathscr{F}_i$, the following similarity relation is used:

$$s(\tilde{A}, \tilde{B}) = \Pi_{\tilde{A}}(\tilde{B})$$

As for the activation function, a modified version of the classical logistic is used [21], which is an automorphism (a monotonic bijection) in [0,1].

The framework has provision for other types of variables, as ordinal or linguistic, and other kinds of combination for the partial similarities [22]. The resulting heterogeneous neuron is sensitive to the *degree of similarity* between its weight vector and a given input, both composed in general by a mixture of continuous and discrete quantities — possibly with missing data — and can be used for configuring heterogeneous artificial neural networks. A layered, feed-forward architecture, with a hidden layer composed of heterogeneous neurons and an output layer of classical ones is a basic straightforward choice, thus confirming a *hybrid* structure. The general training procedure for the heterogeneous neural network (HNN for short) is based on genetic algorithms, due to the presence of missing information and the eventual non-differentiability of the similarity measure.

4.2. Rough sets

An important issue in the analysis of dependencies among variables is the identification of information-preserving reduction of redundant variables. In particular, the task is to find a minimal subset of interacting variables having the same discriminatory power as the original ones, which would lead to the elimination of irrelevant or noisy variables, without the loss of essential information. The rough set theory exploits the idea of approximating a set by other sets. Given a finite set of objects U (the universe of discourse), a set $X \subseteq U$ and an equivalence relation R, two subsets can be associated, called the *lower* (R_{L}) and *upper* (R_{U}) approximation, respectively, as follows:

$$R_{\mathrm{L}} = \bigcup_{Y \in U/R} Y \subseteq X$$

$$R_{\mathrm{U}} = \bigcup_{Y \in U/R} Y \cap X \neq \varnothing$$

where U/R is the equivalence class (partition) induced by R. The lower approximation, also called the *positive region* $\mathrm{POS}_R(X)$, is the set of elements which can be certainly classified as elements of X, whereas the upper approximation is the set of elements which can be possibly classified as elements of X. The *dependency coefficient* is defined as the ratio between positive region size and universe size.

A set of variables P is independent w.r.t. the set of objects Q if for every proper subset R of P, $\mathrm{POS}_P(Q) \neq \mathrm{POS}_R(Q)$; otherwise P is said to be dependent w.r.t. Q. Moreover, the set of variables R is a minimal subset or *reduct* of P, if R is an independent subset of P w.r.t. Q, such that $\mathrm{POS}_R(Q) = \mathrm{POS}_P(Q)$. A variable $a \in P$ is superfluous if $\mathrm{POS}_P(Q) = \mathrm{POS}_{P-\{a\}}(Q)$; otherwise a is said to be *indispensable* in P. The set of all indispensable relations is the *core*. An important property of the core is that it is equal to the intersection of all reducts.

Rules of the form $\langle \text{condition} \rangle \Rightarrow \langle \text{decision} \rangle$ can be generated by using the information contained in the reducts and the objects, concerning their condition and decision attributes. The condition part of the rule is a conjunction of attribute-value pairs. The decision part, in this study, is a single pair composed of the object's decision attribute. Three different strategies were used for rule generation from reducts, as follows:

Strategy 1. For each object, this strategy finds a single relative optimal reduct (in the sense of its length), using heuristics for preserving the dependency coefficient. This strategy is usually the fastest.
Strategy 2. For each object, the shortest relative reduct (in the explicit sense) is computed and used for constructing the rule.
Strategy 3. This strategy operates in a classwise manner by finding all shortest relative reducts whose rules cover some element of the corresponding class.

In all cases, repeated rules are not included. Criteria for matching objects with rules are based on a notion of distance, defined as the number of unmatched attributes taken from the set of predictor variables appearing in the rule. Missing attributes are considered in an optimistic sense, i.e. always matching. In this study, two classification

methods were used for testing the performance of the rule sets generated.

Method 1. Find the most frequent decision among rules with minimum distance from a given sample object.
Method 2. Select first all the rules with minimum distance from a given sample object and then, for every selected rule, count the number of matched objects, choosing as decision the one corresponding to the rule with the highest such number.

4.3. Probabilistic neural networks

This learning model [24] is a reformulation of the Bayes–Parzen classifier, a classical pattern recognition technique [30], in the form of a neural network. The fact that the Bayes classifier is optimal in the sense of the expected misclassification cost makes the use of this kind of network very attractive, specially for smooth classification problems. Besides the input layer, there is a so-called *pattern* layer with as many neurons as patterns are included in the training set. Next, a *summation* layer contains one neuron for each class, then leading to the output layer. Each pattern-layer neuron computes a distance measure between the input and the training sample associated with the neuron. The activation functions of these neurons are Parzen windows used to collectively approximate the probability density functions required by the classifier. The cornerstone of this method lies in its approximation of the multivariate population density function, estimated from the training set as the average of separate multivariate distributions, each centered in a sample from the training set. The main drawbacks are the curse of dimensionality, like all kernel-based methods, and the limited ability to ignore irrelevant variables, which may be a cause of poor generalization.

4.4. Setup and specification of the methods

If some fixed-length segment of the most recent input values is considered enough to perform a task successfully, then a temporal sequence can be turned into a set of spatial patterns on the input layer of a multi-layer feed-forward net. These architectures are called *time-delay neural networks* (TDNN), since several values from an external signal are presented simultaneously at the network input using a moving window. Their main advantage in front of recurrent networks is their lower cost of training, which is very important in the case of long training sequences.

Three different TDNN approaches that differ in the neuron model and training method have been tested, as follows.

1. A classical multi-layer perceptron where the neuron model is the usual scalar-product plus bias term, followed by the hyperbolic tangent [23]. This network (which we will call TD_{MLP}) is trained by means of a hybrid procedure composed of repeated cycles of simulated annealing coupled with the conjugate gradient algorithm [31].
2. The hybrid HNN model, incorporating heterogeneous and classical neurons as described, and trained by means of genetic algorithms (id. TD_{HNN}).
3. The probabilistic neural network, using distance-based neurons and trained as a Bayes–Parzen classifier (id. TD_{PNN}).

Four architectures formed by a hidden layer of 2, 4, 6 and 8 neurons and an output layer of a linear neuron were studied. The TD_{HNN} was trained using a standard genetic algorithm [32] with $p_{cross} = 0.6$, $p_{mut} = 0.01$, two population sizes of 26 and 52 individuals, linear scaling with factor 1.5 and stochastic universal selection. The algorithm was allowed 5 runs for each population size and architecture, and stopped after 1000 generations, unconditionally. The TD_{MLP} was trained in one long run for every architecture, where the number of annealing restarts was fixed to 50. In all cases, average and best results found across the architectures are shown.

The TD_{PNN} was used here with a gaussian kernel. During training, each variable and class units were allowed to have their own variance, with values optimized during the process (possible values ranged from 0.001 to 10). Also, the k-nearest neighbors (KNN) algorithm (with $k = 3$) was tested against the data as a further reference (recall that this algorithm has no training phase). The TD_{HNN} treats qualitative and missing information directly, and original real values as triangular fuzzy numbers, by considering a $\pm 5\%$ of imprecision w.r.t. the reported value. The other two neural approaches codify all information as real-valued and a missing input as zero (no input).

5. Description of the experiments

The effluent quality of the WWTP process given by the TSS-AT was discretized by categorizing the original continuous values into three classes {[0,5), [5,13.5], (13.5,∞)}, expressing *low*, *normal* and *high* values. Four main sets of experiments were performed, all in accordance with the general model:

$$y(t) = F\{x_1(t-2), x_1(t-1), \ldots, x_m(t-2), x_m(t-1), y(t-2), y(t-1)\} \ \forall t \geq 3$$

where m is the number of input variables, for a total of $\hat{m} = 2m + 2$ model input variables. Each $x_i(t)$ denotes the value of the ith input variable and $y(t)$ the value of the target TSS-AT output variable, at time t. The number m varies and will be specified accordingly.

For each experiment, a preliminary study of the training data matrices via rough set analysis is first presented, with the aim of evaluating the actual predictive capacity of the considered model and thus what can be expected on its

Table 4
Rough set approach: correct classification percentages for experiment 1 (top two rows) and experiment 2 (bottom two rows), along with the number of rules needed

	Str. 1 (70 rules) (%)		Str. 2 (72 rules) (%)		Str. 3 (18 rules) (%)	
	Met. 1	Met. 2	Met. 1	Met. 2	Met. 1	Met. 2
Train	75	74	79	74	69	74
Test	73.3	73.3	73.3	73.3	73.3	73.3
Train	78	74	79	74	67	74
Test	73.3	73.3	73.3	73.3	73.3	73.3

influence in the output. Next, the matrices are processed by using the three different strategies for rule generation, and the generated rules, using the two classification methods, are applied to the test matrix, yielding corresponding percentages of correct classification. For the training set, the number of generated rules in each case is shown too. In addition, the results obtained by training and testing the three neural methods (classical, heterogeneous and probabilistic) and the *k*-nearest algorithm are collectively shown and discussed. The advantage of this fanning out of methods is that, being so different in nature, they are able to analyze the data from very different perspectives, allowing to draw more general conclusions. It has to be noted that, throughout all the experiments, the methods are applied to the data in constant conditions (with the same described experimental setup). A description of the four groups of experiments follows.

Experiment 1: qualitative. Oriented to reveal the influence of qualitative variables when studied per se; in particular, to reveal their predictive ability on the TSS classes, taking as inputs x_i the qualitative variables of Table 3 ($m = 18$, $\hat{m} = 38$). This leads to a matrix of qualitative information 145 days long, split into a balanced (in the sense of class frequencies) training part (the first 115 days, 79.3%) and test part (the subsequent 30 consecutive days, 20.7%) to be forecast. It should be noted that the initially formed matrix (232 days long) had a portion of missing information so severe that entire rows had to be removed because *all* information was missing. After that, figures for missing information still are 57.8% in training and 56.9% in test. As a further reference, the percentage of *normal* days (the majority class) in the test matrix is 73.3%.

Experiment 2: reduced-qualitative. The previous results via rough set analysis are used in an attempt to reduce the number of model input variables. This, besides being beneficial for the majority of learning methods, will shed some light on the relevance of variables in relation to the TSS-AT. The new matrices consist of the same days as in experiment 1, though only 12 of the original 38 model variables are to be used.

Experiment 3: combined. Aims at discovering how the qualitative variables in Table 3 behave when joined to five selected quantitative variables: those corresponding to inflow characteristics (Q-AB, COD-AB, TSS-AB) and control actions (Q-P and Q-R). These last variables are counted among the most relevant of the overall process, according to their linear intercorrelation structure [19]. Model parameters are thus ($m = 23$, $\hat{m} = 48$). The heterogeneous data matrix generated covers the whole period of days since this time none had to be removed from the matrix, although figures for missing information were 64.2% in training and 63.4% in test. It was split into a training part (the first 191 days, 82.3%) and a test part (the subsequent 41 days, 17.7%) to be forecast. The percentage of *normal* days in the test matrix is 70.7%.

Experiment 4: reduced-combined. The model of experiment 3 is reduced, again via rough set analysis, leading to a model with less variables and to much lesser missing information percentages, of 31.6% in training and 29.8% in test.

6. Experimental results

The information displayed includes average and best *predictive* accuracies obtained with each method. Training information is also shown. For the rough set approach, this information is given for every strategy and method, along with the number of rules generated.

6.1. Experiment 1: qualitative

Beginning with the preliminary analysis, under the rough set approach, the relative reducts and the core were computed. The *dependency coefficient* between the 38 model variables and the predicted TSS-AT in the training set was found to be zero, indicating that no element can be classified with absolute security and, therefore, that the set of variables is rather incomplete. A total of 68 relative reducts were found, with a core composed of 11 variables. The frequency distribution of variables in the reducts reveal that 12 do appear in 75% or more of all the reducts; specifically, the 11 of the core plus an extra variable. On the other hand, another 14 variables from the original set of 38 are superfluous (they occur in no reduct). All this means that information dependency is unevenly distributed in the set of variables, as 32% of them is conveying the major part, while another 37% is carrying no information at all.

The results of the rule generation process are given in Table 4 (top two rows), as percentages of correct classification. All the methods and strategies are signaling the same prediction ability, 73.3%, which coincides with the majority

Table 5
Neural approaches and KNN: correct classification percentages for experiment 1 (top two rows) and experiment 2 (bottom two rows)

	TD_{HNN} (%)		TD_{MLP} (%)		TD_{PNN} (%)	KNN (%)
	Best	Avg.	Best	Avg.		
Train	87.0	82.2	86.9	82.2	76.5	–
Test	80.0	76.3	73.3	47.5	73.3	76.7
Train	85.2	81.5	82.6	81.3	83.5	–
Test	76.7	75.4	76.7	70.2	16.7	73.3

class. This poor performance is nonetheless reflecting the complexity of the data set, with a high rate of missing values affecting all variables, and classes showing severe overlappings, revealed by the null dependency coefficient. It is interesting to observe that strategy 3 for rule generation needed only 23% of the rules required by the other two while keeping the same effectiveness. The results obtained with the three neural approaches and the KNN are shown in Table 5 (top two rows). Several aspects are noteworthy. First, the results are quite similar and consistent both for training and test sets. In other words, no method clearly outperforms the others. Second, there seems to be a limit in training set accuracy around 87.0 and at 80.0% in test, which is not a bad result for such messy data. Also interesting to note are the solid results achieved by the TD_{HNN}, the poor average achieved by the TD_{MLP} and the comparatively good KNN performance.

6.2. Experiment 2: reduced-qualitative

In order to assess the viability of smaller models, a new data matrix was constructed as in experiment 1, but now using only those model variables (12, see Table 6) occurring most frequently (in 75% or more) in the collection of reducts. Note that selected variables include all the filamentous bacteria — dominant in situations regarding poor sludge settleability, making solids more likely to escape the settler —, and also the presence of the predicted variable in the two previous days. Moreover, and due to the frequency of analysis and observations, the 2-day lag variables dominate over 1-day ones, a consistent result. Also, with this variable selection, figures for missing information drop to 25.8% in training and 19.2% in test.

Table 6
Reduced set of *qualitative* variables for experiment 2

Variable	Delay
Presence-foam	$t-2$
Look-AT	$t-2$
Zooglea	$t-2$
Nocardia	$t-2$
Thiothrix/21N	$t-2$
Type 0041	$t-2$
M. Parvicella	$t-2$
Carnivorous ciliates	$t-2$
Rotifer	$t-2$
Aspidisca	$t-1$
TSS-AT	$t-2$
TSS-AT	$t-1$

The predictive power of this reduced data set was computed as before and the new results are shown in Tables 4 and 5 (bottom two rows), shown against those of experiment 1 for ease of comparison. As could be expected, overall training and predictive performance is less and performance of some methods (the TD_{PNN} and the KNN) has fallen — the latter slightly, the former abruptly. The other two neural architectures still keep a decent classification ability, slightly above the 73.3% limit imposed by the major class. Moreover, the results are quite balanced between the training and test sets, and what is more important, almost identical w.r.t. those obtained for the model having all qualitative variables, thus showing that a shorter model with only less than one third of the original variables says the same about TSS-AT than the whole set. If this behavior is confirmed by future investigations, it might have important practical consequences.

6.3. Experiment 3: combined

The preliminary analysis via rough sets was performed on the new combined set of variables. To this end, the continuous process represented by numerical data was transformed into a discrete one by expert introduction of cut-point values. In particular, the following were set: Q-AB (6000; 14,800), COD-AB (560; 1000), TSS-AB (210; 420), Q-R (3500; 10,300) and Q-P (100; 1400). From the total of 48 model variables (10 numeric and 38 qualitative), 325 relative reducts and a core composed by 12 variables were computed. The *dependency coefficient* between the 48 model variables and the predicted TSS-AT category in the training set now rose to 0.22. This shows a gain in secure classification ability due to the addition of the new information given by the set of 10 numerical variables. However, the value of this coefficient is rather low, indicating that the new variable set, although enlarged, is still incomplete. Frequency distribution of variables among the reducts reveals that only 13 variables, from the set of 48, appear in 75% or more of all the reducts (actually, again the core plus an extra variable). Moreover, 15 variables are completely superfluous (they occur in no reduct) and, as in the case of quantitative information only, information dependency is unevenly distributed in the set of variables (27% taking the major part and 31% taking no part at all).

The results (Tables 7 and 8, top two rows) show that, with a single exception, classification performance via rough sets has increased in the training set w.r.t. experiment 1, while it has decreased slightly in the test set (70 vs. 73%). This indicates that the gain effect of the new variables was not enough, as classification performance remains about the same, and new informative *model* variables should be included. For the neural methods, the generalized lower performance (w.r.t. Table 5, top two rows) is at first glance

Table 7
Rough set approach: correct classification percentages for experiment 3 (top two rows) and experiment 4 (bottom two rows), along with the number of rules needed

	Str. 1 (108 rules) (%)		Str. 2 (104 rules) (%)		Str. 3 (31 rules) (%)	
	Met. 1	Met. 2	Met. 1	Met. 2	Met. 1	Met. 2
Train	83	80	82	80	43	80
Test	70	70	70	70	24	70
Train	81	80	81	80	41	79
Test	70	70	70	70	41	70

surprising but can be explained with the sudden increment in parameters while keeping a very small training set. Also noteworthy is the 100% training accuracy achieved by the TD_{PNN}, although test set performance equals the majority class limit, possibly indicating again severe overlappings among the classes.

6.4. Experiment 4: reduced-combined

To evaluate the viability of smaller models, a new data matrix was constructed as in experiment 3. This time only those 13 (out of 48) predictor variables (Table 9) are occurring in 75% of the reducts or more. This reduced set provides much information and deserves careful attention. First, numerical variables are predominant, despite their lower number w.r.t. the qualitative ones. Among them, the physico-chemical inflow characteristics (Q-AB, COD-AB and TSS-AB) and the control actions (purge and recycle flow rates). Second, we can see how this information is needed at *both* delays for the inflow rate and the control actions. Third, the three qualitative variables include the most commonly filamentous bacteria found in this plant (*Nocardia* and *Thiothrix* or type 021N) causing bulking sludge, and a protozoa (*Aspidisca*) whose absence may indicate a decrease in plant performance and poor settling characteristics. It is also remarkable that these three variables also appeared in the previous reduced set of qualitative information, and are the sole survivors when mixed with the numerical information. And fourth, again, the predicted variable itself (TSS-AT) (at both delays) is considered amongst the most informative. The behavior of this model (Tables 7 and 8, bottom two rows) is similar to that of the previous, in the sense that classification performances for training and test sets are slightly less, showing that the effect of the 35 discarded variables was in fact small.

Table 8
Neural approaches and KNN: correct classification percentages for experiment 3 (top two rows) and experiment 4 (bottom two rows)

	TD_{HNN} (%)		TD_{MLP} (%)		TD_{PNN} (%)	KNN (%)
	Best	Avg.	Best	Avg.		
Train	84.3	81.7	81.7	80.6	100	–
Test	75.6	73.2	70.7	70.2	70.7	61.0
Train	83.8	81.2	80.6	78.3	100	–
Test	73.2	71.6	70.7	70.1	70.7	63.4

Turning the attention to the neural models (Table 8, bottom two rows), it is interesting to observe that the overall results are consistent with those obtained in the different experiments, specially in what concerns to the test set. Moreover, since the TD_{PNN} is asymptotically optimal in the sense of the Bayes classifier, this might indicate a limit in what is achievable with the available information. Also, the fact that the TD_{HNN} model gives slightly, but consistently, higher results and a more balanced training/test ratio than all of the other methods has been observed in other application contexts [33,34] and, in this case study, can be attributed to its better treatment of missing values and qualitative information.

7. Conclusions

The influence of qualitative information on WWTPs has been studied, in what concerns the quality of effluent suspended solids, one of the measures of plant performance. We found that qualitative information exerts a considerable influence on plant output, although very unevenly. A high degree of information redundancy was discovered, since comparable predictive capabilities are obtained when working with much-reduced subsets of variables, obtained by rough set analysis. However, it should be noted that this redundancy refers only to the prediciton of *bulking* episodes in the process, and the use of these variables is necessary to guarantee the global performance of the entire activated sludge process.

The analysis produces homogeneous groups of variables; for qualitative variables only, it signals the greater importance of 2-day delayed data in the process dynamics, instead of 1-day data. When qualitative and numerical information are collectively considered, the latter are found to be amongst the more informative, always in both delays. Nonetheless, there are certain qualitative variables (the intersection of Tables 6 and 9) playing a significant role in the process. In both cases, these selected variables are highly rated by WWTP experts. They also tend to be the ones that show the lower amount of missing values, thus reducing the relative overall amount.

In addition, a common upper bound in predictive

Table 9
Reduced set of *combined* variables

Variable	Delay
Q-AB	$t-2$
Q-AB	$t-1$
COD-AB	$t-2$
TSS-AB	$t-2$
Q-R	$t-2$
Q-R	$t-1$
Q-P	$t-2$
Q-P	$t-1$
Nocardia	$t-2$
Thiothrix/21N	$t-2$
Aspidisca	$t-1$
TSS-AT	$t-2$
TSS-AT	$t-1$

classification accuracy has been discovered, located around an 80%, which is a very good result for such messy data. In this respect, our conclusion is that the generalized and (relatively) poor performance can be attributed almost entirely to the data, besides the problem complexity, in light of the consistent results yielded by methods that are so different in nature; the fact that they are based on very different principles allows to derive broader conclusions from the available data. The possibilities of some of these methods (especially the TD_{HNN}) are also noteworthy, provided they can handle heterogeneity, imprecision and missing values, aspects that characterize the data in a real WWTP process.

In conclusion, the observed patterns of behavior are very coherent. The next step should be oriented towards adding information in the form of better delays (e.g. the weekly effect) and a more accurate selection of variables, always taking into account the findings reported herein. Ulterior studies with data coming from other plants will be needed to determine whether these patterns are specific or whether they represent a more general property of WWTPs. A further goal in the future is the development of a predictive model for control variables (Q-P and Q-R). These models will supply the plant manager with a useful tool to improve plant control and operation.

Acknowledgements

The authors wish to thank the *Consorci de la Costa Brava* for the data and information provided. This work has been supported by Spanish CICYT Projects AMB-97/0889, TIC-96/0878 and TAP-99/0747.

References

[1] Porteous A. Dictionary of environmental science and technology. 2nd ed. New York: Wiley, 1996.
[2] Lean G, Hinrichsen D. Atlas of the environment. 2nd ed. New York: Harper, 1994.
[3] Capodaglio AG, Jones HV, Novotny V, Feng X. Sludge bulking analysis and forecasting: application of system identification and artificial neural computing technologies. Water Research 1991;25(10):1217–24.
[4] Ayesa E, Oyarbide G, Larrea L, García-Heras JL. Observability of reduced order models. Application to a model for control of alpha process. Water Science and Technology 1995;31(2):161–70.
[5] Dochain D. Design of adaptative controllers for non-linear stirred tank bioreactors: extension to the MIMO situation. Journal of Process Control 1991;1:41–8.
[6] Moreno R, de Prada C, Lafuente J, Poch M, Montague G. Non-linear predictive control of dissolved oxygen in the activated sludge process, ICCAFT 5/IFAC-BIO 2 Conference, Keystone, 1992. p. 289–93.
[7] Nejjari F, Benhammou A, Dahhou B, Roux G. Nonlinear multivariable control of a biological wastewater treatment process, ECC 97, Brussels, Belgium, 1997.
[8] Henze M, Grady CPL Jr, Gujer W, Marais GR, Matsuo T. Activated sludge model no. 1, IAWPRC Scientific and Technical Reports, vol. 1, IAWPRC, London, 1987.
[9] Henze M, Gujer W, Mino T, Matsuo T, Wentzel MC, Marais GvR. Activated sludge model no. 2, IAWPRC Scientific and Technical Reports, vol. 3, IAWPRC, London, 1995.
[10] Okubo T, Kubo K, Hosomi M, Murakami A. A knowledge-based decision support system for selecting small-scale wastewater treatment processes. Water Science and Technology 1994;30(2):175–84.
[11] Serra P, Sànchez M, Lafuente J, Cortés U, Poch M. ISCWAP: a knowledge-based system for supervising activated sludge processes. Computers in Chemical Engineering 1997;21(2):211–21.
[12] Zhu XX, Simpson AR. Expert system for water treatment plant operation. Journal of Environmental Engineering 1996:822–9.
[13] Sànchez M, Cortés UR, Roda I, Poch M, Lafuente J. Learning and adaptation in WWTP through case-based reasoning. Microcomputers in Civil Engineering 1997;12(4):251–66 (Special issue: Machine Learning).
[14] Côté M, Grandjean BPA, Lessard P, Thibault J. Dynamic modeling of the activated sludge process: improving prediction using neural networks. Water Research 1995;29(4):995–1004.
[15] Sànchez M, Cortés U, Lafuente J, Roda IR, Poch M. DAI-DEPUR: a distributed architecture for wastewater treatment plants supervision. Artificial Intelligence in Engineering 1996;10(3):275–85.
[16] Zhao H, Hao OJ, McAvoy TJ. Modeling nutrient dynamics in a sequencing batch reactor. Journal of Environmental Engineering 1997;123:311–9.
[17] Novotny V, Jones H, Feng X, Capodaglio AG. Time series analysis models of activated sludge plants. Water Science and Technology 1990;23:1107–16.
[18] Belanche Ll, Valdés JJ, Comas J, Roda I, Poch M. Modeling the input–output behaviour of wastewater treatment plants using soft computing techniques. Papers from the ECAI'98 Workshop Binding Environmental Sciences and AI, European Conference on Artificial Intelligence, Brighton, UK, 1998.
[19] Belanche Ll, Valdés JJ, Comas J, Roda I, Poch M. Towards a model of input–output behaviour of wastewater treatment plants using soft computing techniques. Environmental Modeling and Software 1999;14:409–19.
[20] Standard methods for the examination of water and wastewater, 16th ed., Water Environment Federation, Washington APHA, 1992.
[21] Valdés JJ, García R. A model for heterogeneous neurons and its use in configuring neural networks for classification problems. Procs. of IWANN'97, Intl. Conf. on Artificial and Natural Neural Networks, Lecture Notes in Computer Science 1240. Berlin: Springer, 1997. p. 237–46.
[22] Belanche Ll. A theory for heterogeneous neuron models based on similarity. LSI Research Report LSI-00-06-R, Univ. Politècnica de Catalunya, 2000.
[23] Hertz J, Krogh A, Palmer RG. Introduction to the theory of neural computation. Redwood City, CA: Addison-Wesley, 1991.

[24] Specht D. Probabilistic neural networks. Neural Networks 1990;3:109–18.
[25] Pawlak Z. Rough sets: theoretical aspects of reasoning about data. Dordrecht: Kluwer Academic, 1991.
[26] Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971;27:857–71.
[27] Dubois D, Prade H, Esteva F, García P, Godo Ll, López de Mántaras R. Fuzzy set modeling in case-based reasoning. Lecture Notes in Artificial Intelligence 1266. Berlin: Springer, 1997. p. 599–610.
[28] Zimmermann HJ. Fuzzy set theory and its applications. Dordrecht: Kluwer Academic, 1992.
[29] Dubois D, Prade H. The three semantics of fuzzy sets. Fuzzy sets and Systems 1997;90(2):00.
[30] Fukunaga K. Introduction to statistical pattern recognition. Orlando, FL: Academic Press, 1972.
[31] Ackley D. A connectionist machine for genetic hillclimbing. Dordrecht: Kluwer Academic, 1987.
[32] Goldberg DE. Genetic algorithms for search, optimization and machine learning. Reading, MA: Addison-Wesley, 1989.
[33] Belanche Ll, Valdés JJ. Fuzzy inputs and missing data in similarity-based heterogeneous neural networks. Procs. of the 5th Intl. Work-Conference on Artificial and Natural Neural Networks, IWANN'99. Engineering applications of bio-inspired artificial neural networks (Lecture Notes in Computer Science 1607). Berlin: Springer, 1998. p. 863–73.
[34] Valdés JJ, Belanche Ll, Alquézar R. Fuzzy heterogeneous neurons for imprecise classification problems. International Journal of Intelligent Systems 2000;15(3):265–76.