A Study of Qualitative and Missing Information in Wastewater Treatment Plants

Ll. Belanche and **J.J. Valdés**
Secció d'Intel·ligència Artificial.
Dept. de Llenguatges i Sistemes Informàtics.
Universitat Politècnica de Catalunya.
c/Jordi Girona 1-3. 08034 Barcelona, Spain.
{belanche, valdes}@lsi.upc.es

J. Comas and **I.R.Roda** and **M. Poch**
Laboratori d'Enginyeria Química i Ambiental.
Facultat de Ciències.
Universitat de Girona.
Campus de Montilivi s/n 17071 Girona, Spain.
{quim, ignasi, manel}@lequia1.udg.es

Abstract

The correct control and prediction of Wastewater Treatment Plants poses an important goal in order to avoid breaking the environmental balance and to always keep the system in stable operating conditions. In this respect, it is known that *qualitative* information —coming from microscopic examinations and subjective remarks— has a deep influence on the activated sludge process, especially in the total amount of effluent suspended solids (TSS), one of the measures of overall plant performance. The strong interrelation between variables, their heterogeneity, and the very high amount of missing information make the use of traditional techniques difficult, or even impossible. Despite this problems, and through the use of several soft computing methods —rough set theory and artificial neural networks, mainly— acceptable prediction models are found that show the interplay between variables and give insight to the dynamics of the process.

INTRODUCTION

Dirty water is both the world's greatest killer and its biggest single pollution problem (Lean and Hinrichsen 1994). 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 have guaranteed the appearance of specific plants to treat these wastewaters, being activated sludge the system most extensively used in Wastewater Treatment Plants (WWTP). In an activated sludge process, the wastewater (mainly organic matter, suspended solids and nutrients) goes into an aerated tank where it is mixed with biological floc particles. After enough contact time, this mixture is discharged to a settler that separates the suspended biomass from the treated water. Most of the biomass is recirculated to the aeration tank again, while a little amount is purged daily (WEF 1996).

Activated sludge is a clear example of an environmental process which 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 there is 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 is subjective and cannot be numerically quantified.

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

Research contributions in this field have been formulated from many different points of view. However, a direct cause-effect relationship for WWTP performance has been established only in few cases and, even in those, the experimental results could lead to contradictory conclusions (Capodaglio 1991), avoiding the formulation of deterministic cause-effect relationships that could be used as prediction models. The identification of a model that could predict in real-time and with reasonable accuracy the appearance of sludge bulking is thus of great practical importance because of the potential improvement of treatment plant efficiency and cost savings (Novotny et al. 1990). This model should let 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 accomplishing the discharge permit limit.

To tackle such a task, different, interrelated studies have been performed towards the development of a model of input-output behaviour of WWTP using soft computing techniques (Belanche et al. 1998); (Belanche et al. 1999). The next natural step is to take into ac-

AB (inflow)	Q-AB (inflow)	COD, BOD (organic matter) TSS (suspended solids)	-
AS (bioreactor)	Q-R (biological recycle) Q-P (biological purge) Q-A (biological aeration)	-	Presence-foam Microfauna (*Aspidisca, Vorticella*, ...) Filamentous bacteria (*Nocardia, Thiothrix*, ...)
AT (outflow)	-	COD, BOD, TSS	Look (appearance)

Table 1: Selected variables characterizing the behaviour of the studied WWTP.

count qualitative information –which had not been considered in the previous studies– and to explore how it affects the formation of predictive models. This 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. There is also a need to handle uncertain or imprecise information, a characteristic present in all kinds of variables.

Variable	Unit	Missing	Mean	StDev
Q-AB	m^3/d	18	10707.0	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

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

The *purpose* of this paper is to present several experiments performed using qualitative information, either *per se* or together with quantitative information, such as influent characteristics and control actions. Specifically, the influence on effluent TSS levels is studied, as an indication of plant performance and fulfillment of regulations. The final aim of the work is to find a model capable of short-term prediction, which takes into account only really relevant variables and accommodates characteristics of real WWTP data: imprecision, heterogeneity, and high incidence of missing information.

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

A WWTP CASE STUDY

The historical database used throughout the work corresponds to a WWTP of a touristic resort in the Costa Brava (Catalonia). 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). Quantitative information includes analytical results of water quality –organic matter, measured as chemical (COD) and biochemical (BOD) oxygen demand, and Total Suspended Solids, measured as TSS (WEF 1992)–, 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*).

The final data set covers an homogeneous representative period of 609 consecutive days, considering each day as a new sample. Basic statistical descriptors of the variables comprised in the database are shown in table 2 (for quantitative variables) and table 3 (for qualitative ones). The relative abundance of qualitative variables is categorized in three different levels: *none*, *some* and *many*, with the exception of outflow appearance (that is, "Look-AT"), categorized as *poor*, *fair* and *good*. The most relevant feature of the database is the extremely high incidence of missing values (between 60-80%, approximately). This is specially true in the case of outflow variables COD-AT, BOD-AT and TSS-AT –more suitable as targets for developing prediction models– variables characterizing water quality at the inflow COD-AB, BOD-AB and TSS-AB, and qualitative variables characterizing the microorganisms. For this reason, the final database processed includes only those days with a recorded value in the target 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 po-

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. >20μm	394	184	23	8
Flagell. <20μm	394	176	24	15
Amœbae	394	173	38	4
Testate *amœbae*	394	206	8	1
Rotifer	394	117	97	1
		poor	*fair*	*good*
Look-AT	394	9	168	38

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

tential 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 distributes normally, the second does not. Actually, it has a right-skewed distribution, reflecting strong non-linear distortions introduced by the WWTP dynamics (see figures 1, 2). All these features make considerably hard the search for models to characterize WWTP behaviour and must be always taken into account when evaluating the quality of the learned models.

DESCRIPTION OF THE METHODS

Four techniques were employed in this work to study the influence and classification ability of qualitative variables: fuzzy heterogeneous neural networks, classical neural networks, probabilistic networks and the $k-$nearest neighbours algorithm. Rough set theory was also used to perform a reduction of dimension.

Heterogeneous neural networks (HNNs for short) are neural architectures built out of neuron models which allow heterogeneous and imprecise inputs, defined in (Valdés and García 1997); (Valdés, Belanche and Alquézar 1999) as a mapping $h : \hat{\mathcal{H}}^n \to \mathcal{R}_{out} \subseteq \mathbb{R}$. Here $\mathbb{R}$ denotes the reals and $\hat{\mathcal{H}}^n$ is a cartesian product

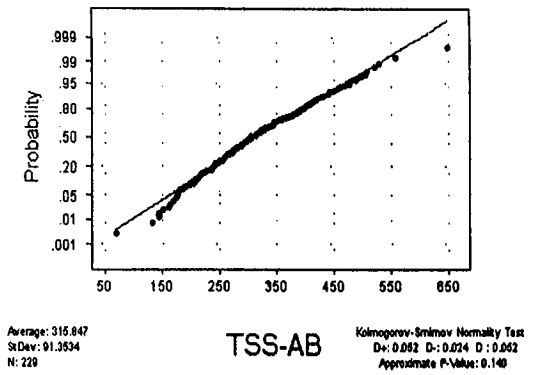

Figure 1: Kolmogorov-Smirnov test for incoming Total Suspended Solids (TSS-AB).

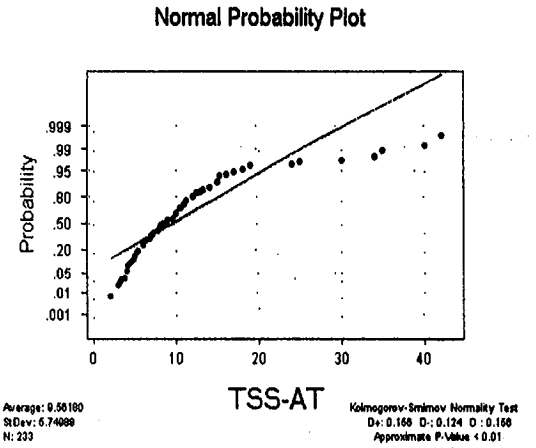

Figure 2: Kolmogorov-Smirnov test for outgoing Total Suspended Solids (TSS-AT).

of an arbitrary number n of *source sets*. These source sets may be extended reals $\hat{\mathcal{R}}_i = \mathbb{R}_i \cup \{\mathcal{X}\}$, extended families of (normalized) fuzzy sets $\hat{\mathcal{F}}_i = \mathcal{F}_i \cup \{\mathcal{X}\}$, and extended finite sets of the form $\hat{\mathcal{O}}_i = \mathcal{O}_i \cup \{\mathcal{X}\}$, $\hat{\mathcal{M}}_i = \mathcal{M}_i \cup \{\mathcal{X}\}$, where each of the $\mathcal{O}_i$ has a full order relation, while the $\mathcal{M}_i$ have not. The special symbol $\mathcal{X}$ 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, nominals and missing data.

An heterogeneous neuron computes a *similarity index*, or proximity relation, followed by the familiar form of a squashing non-linear function with domain in $[0, 1]$. Thus, the neuron is sensitive to the degree of similarity between its input and its weights, both composed in general by a mixture of continuous and discrete quan-

similarity index (Gower 1971) in which the computation for heterogeneous entities is constructed as a weighted combination of partial similarities over subsets of 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 range $(v_{\cdot k}) = \max_{i,j} (|v_{ik} - v_{jk}|)$ (see (Gower 1971) for details on other kinds of variables). The δ_{ijk} is a binary function expressing whether both objects are comparable or not according to their values w.r.t. variable k. It is 1 if and only if both objects have values different from $\mathcal{X}$ for variable k, and 0 otherwise. For variables representing fuzzy sets, similarity relations from the point of view of fuzzy theory have been defined elsewhere (Zimmermann 1992), and different choices are possible. In our case, if $\mathcal{F}_i$ is an arbitrary family of fuzzy sets from the source set, and $\tilde{A}$, $\tilde{B}$ are two fuzzy sets such that $\tilde{A}$, $\tilde{B} \in \mathcal{F}_i$, the following similarity relation is used:

$$g(\tilde{A}, \tilde{B}) = \sup_x \{\min (\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x))\}$$

As for the activation function, a modified version of the logistic is used (Valdés and García 1997), that maps the real interval $[0, 1]$ on $(0, 1)$. The resulting heterogeneous neuron can be used for configuring artificial neural networks, of which 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 conforming a *hybrid* structure. The general training procedure for the HNN is based on genetic algorithms, due to data heterogeneity, missing data and the eventual non-differentiability of the similarity function.

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. Rough Sets (Pawlak 1991) exploit 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 and *upper* (R_U) approximation, respectively, as follows:

$$R_L = \bigcup \{Y : Y \in U/R : Y \subseteq X\}$$
$$R_U = \bigcup \{Y : Y \in U/R : Y \cap X \neq \emptyset\}$$

where U/R is the equivalence class (partition) induced by R. The lower approximation, also called the *positive region* $POS_R(X)$, 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, $POS_P(Q) \neq POS_R(Q)$; 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 $POS_R(Q) = POS_P(Q)$. A variable $a \in P$ is superfluous if $POS_P(Q) = POS_{P-\{a\}}(Q)$; 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 <*condition*> $\Rightarrow$ <*decision*> 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.

model (Specht 1990) is a reformulation of the Bayes-Parzen classifier –a classical pattern recognition technique (Fukunaga 1972)– in the form of an artificial 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 classification problems.

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 feedforward net. These architectures, regardless of the training method and the neuron model, are called *time-delay neural networks* (TDNNs), since several values from an external signal are presented simultaneously at the network input using a moving window (Hertz, Krogh and Palmer 1991). Their main advantage in front of recurrent architectures 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: a hybrid procedure (Ackley 1987) composed of repeated cycles of simulated annealing coupled with the conjugate gradient algorithm (which we will call TDNN-AC), our HNN model (*id.* TDNN-HG), incorporating heterogeneous neurons and trained by means of genetic algorithms, and the probabilistic neural network (TDNN-PR). 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 TDNN-HG was trained using a standard genetic algorithm (Goldberg 1989) with $p_{cross} = 0.6$, $p_{mut} = 0.01$, population size=26, 52 individuals, a linear rank scaling with factor 1.5, and stochastic universal selection. The algorithm was allowed 5 runs for each population size and stopped after 1,000 generations unconditionally. The TDNN-AC uses the hyperbolic tangent instead of the logistic, and is trained in one long run for every architecture, in which the number of annealing tries was fixed to 50.

The probabilistic network TDNN-PR uses a gaussian kernel. During training each variable and class was allowed to have its own variance, with values optimized during the process (possible values ranged from 0.001 to 10). Also, the k–nearest neighbours (KNN) algorithm (with $k = 3$) was tested against the data as a further reference (recall that this algorithm has no training phase). The TDNN-HG treats qualitative and missing information directly and original real values as triangular fuzzy numbers in the form of a ±5% of imprecision w.r.t. the reported value, while the other two neural approaches code a missing input as zero (no input) and discrete values as real numbers.

EXPERIMENTS

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

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

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 influence in the output. Next, the matrices are processed by using the three different strategies for rule generation, and the generated rules, along with 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. Then, 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, 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, all the methods are applied to the data in exactly the same conditions.

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, 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.

Previous results via rough set analysis are used in an attempt to reduce the number of model input variables.

Met. 1	Met. 2	Met. 1	Met. 2	Met. 1	Met. 2	Best	Avg.	Best	Avg.		
75%	74%	79%	74%	69%	74%	87.0%	82.2%	86.9%	82.2%	76.5%	-
73.3%	73.3%	73.3%	73.3%	73.3%	73.3%	80.0%	76.3%	73.3%	47.5%	73.3%	76.7%
78%	74%	79%	74%	67%	74%	85.2%	81.5%	82.6%	81.3%	83.5%	-
73.3%	73.3%	73.3%	73.3%	73.3%	73.3%	76.7%	75.4%	76.7%	70.2%	16.7%	73.3%

Table 4: Rough set approach, Neural approaches and KNN: correct classification percentages for **Experiment 1** (top two rows) and **Experiment 2** (bottom two rows), along with the number of rules needed.

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 part of the original 38 model variables are used.

Experiment 3: Combined.

Aims at discovering how qualitative information behaves 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 variables are counted among the most relevant of the overall process, according to their linear intercorrelation structure (Belanche et al. 1999). 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, 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 lesser variables and to much lesser missing information percentages of 31.6% in training and 29.8% in test.

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.

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 0.0, 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 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, the three neural approaches and the KNN are given in table 4 (top) as percentages of correct classification. All the methods and strategies are signaling the same prediction ability, 73.3%, which coincides with the majority 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.

As for the neural methods, several aspects are noteworthy. First, the results are quite similar and consistent for both training and test sets. In other words, no method clearly outperforms the rest. Second, there seems to be a limit in training set accuracy around 87.0% and at 80.0% in test, which is a not so bad result for such messy data. Also interesting to note are the solid results achieved by the TDNN-HG, the poor average achieved by the TDNN-AC and the comparatively good KNN performance.

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 (twelve, see table 5) 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 4 (bottom). As could be expected, overall training and predictive performance is less and performance of some methods (the TDNN-PR and the KNN) has fallen –slightly the latter, abruptly the former–. On the other hand, 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.

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	$t-1$

Table 5: Reduced set of *qualitative* variables for **Experiment 2**.

Experiment 3: Combined

The preliminary analysis via rough sets was again performed on the new 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 (6,000; 14,800), COD-AB (560; 1,000), TSS-AB (210; 420), Q-R (3,500; 10,300) and Q-P (100; 1,400). 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 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 (table 6, top) show that, with a single exception, classification performance via rough sets has increased in the training set w.r.t. to 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 (see table 4, top) is at first glance 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 TDNN-PR, although test set performance is below average.

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	$t-1$

Table 7: Reduced set of *combined* variables.

Experiment 4: Reduced-Combined

This time only those 13 (out of 48) predictor variables (table 7) are occurring in 75% of the reducts or more. This reduced set is giving 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 Q-R and Q-P (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*), the absence of which may indicate a decrease in plant performance and poor settling characteristics. It is also remarkable the fact that these three variables also appeared in the previous reduced set of qualitative information, and are

Met. 1	Met. 2	Met. 1	Met. 2	Met. 1	Met. 2	Best	Avg.	Best	Avg.		
83%	80%	82%	80%	43%	80%	84.3%	81.7%	81.7%	80.6%	100%	-
70.0%	70.0%	70.0%	70.0%	24.0%	70.0%	75.6%	73.2%	70.7%	70.2%	70.7%	61.0%
81%	80%	81%	80%	41%	79%	83.8%	81.2%	80.6%	78.3%	100%	-
70.0%	70.0%	70.0%	70.0%	41.0%	70.0%	73.2%	71.6%	70.7%	70.1%	70.7%	63.4%

Table 6: Rough set approach, Neural approaches and KNN: correct classification percentages for **Experiment 3** (top two rows) and **Experiment 4** (bottom two rows), along with the number of rules needed.

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 behaviour of this model (table 6, bottom) 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.

Turning the attention to the neural models, 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 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 TDNN-HG 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 (Valdés, Belanche and Alquézar 1999); (Belanche, Valdés and Alquézar 1998) and, in this case study, can be attributed to its better treatment of missing values and qualitative information.

CONCLUSIONS

The influence of qualitative information in WasteWater Treatment Plants (WWTP) has been studied, in what regards to effluent total suspended solids quality, one of the measures for plant performance. Summarizing the results, it was 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 severed subsets of variables, obtained by rough set analysis. This analysis produces homogeneous groups of variables; for qualitative variables only, it signals the greater importance of 2-day delayed data in the process dynamics, as opposed to 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. In both cases, selected variables are highly rated by WWTP experts. They also tend to be the ones with less amount of missing values, thus reducing the relative overall amount.

In addition, a common upper-bound in classification accuracy is discovered, located around an 87% accuracy in the model search process (the *training*) and an 80% of predictive accuracy (that is, using the learned model). In this respect, the generalized and (relatively) poor performance can be attributed almost entirely to the data –besides, of course, to 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 on the available data. The possibilities of these methods are also noteworthy, provided they can handle heterogeneity, imprecision and missing values, aspects that characterize the data in a WWTP process.

In conclusion, the observed patterns of behaviour are very promising and deserve ulterior studies to determine whether these patterns are specific or else they represent a more general property of WWTPs. The future work being done is oriented in this direction, adding information in the form of better delays (*e.g.* the weekly effect) and a more accurate selection of variables, taking into account the findings reported herein. An additional goal is the development of a predictive model for control variables (Q-P and Q-R). These models will hopefully supply the plant manager with a useful tool to improve plant control and operation.

Acknowledgments. The authors wish to thank the *Consorci de la Costa Brava* for the data and information provided. This work has been supported by CICYT Projects AMB-97/889 and TIC96-0878.

References

Ackley, D. 1987. A connectionist machine for genetic hillclimbing. Kluwer Acad. Press.

Belanche, Ll., Valdés, J.J., Comas, J., Roda, I., Poch, M. 1998. Modeling the Input-Output Behaviour of Wastewater Treatment Plants using Soft Computing Techniques. In Procs. of *BESAI'98. Binding Environmental Sciences and AI.* Workshop held as part of *ECAI'98: European Conf. on Artificial Intelligence*, Brighton, UK, pp. 81-94.

Belanche, Ll., Valdés, J.J. Alquézar, R. 1998. Fuzzy Heterogeneous Neural Networks for Signal Forecasting. In Procs. of *ICANN'98, Intl. Congress on Natural*

Verlag, Perspectives in Neural Computing Series.

Belanche, Ll., Valdés, J.J., Comas, J., Roda, I., Poch, M. 1999. Towards a Model of Input-Output Behaviour of Wastewater Treatment Plants using Soft Computing Techniques. *Environmental Modeling and Software* (in press).

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.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties, *Biometrics* 27: 857-871.

Goldberg, D.E. 1989. Genetic Algorithms for Search, Optimization & Machine Learning. Addison-Wesley.

Fukunaga, K. 1972. Introduction to statistical pattern recognition. Academic Press, Orlando.

Hertz, J., Krogh, A., Palmer R.G. 1991. Introduction to the Theory of Neural Computation, Addison-Wesley, Redwood City.

Lean, G., Hinrichsen, D. 1994. Atlas of the Environment. (2nd ed.), Harper Perennial Publ.

Novotny, V., Jones, H., Feng, X., Capodaglio, A.G. 1990. Time series analysis models of activated sludge plants. *Water Science & Technology* 23: 1107-1116.

Pawlak, Z. 1991. Rough Sets: Theoretical aspects of reasoning about data. Kluwer Academic Publ.

Specht, D. 1990. Probabilistic neural networks. Neural Networks, 3: 109-118.

Valdés, J.J., García, R. 1997. A model for heterogeneous neurons and its use in configuring neural networks for classification problems, *Procs. of IWANN'97, Intl. World Conf. on Artificial and Natural Neural Networks*. Lecture Notes in Computer Science 1240, Springer-Verlag, pp. 237-246.

Valdés J.J, Belanche, Ll., Alquézar, R. 1999. Fuzzy Heterogeneous neurons for Imprecise Classification Problems. *Intl. J. of Intelligent Systems* (in press).

WEF: Standard Methods for the Examination of Water and Wastewater, 16th ed., 1992. Water Environment Federation, Washington APHA.

WEF: Operation of Municipal Wastewater Treatment Plants. Manual of Practice No. 11, 5th Edition, 1996. Water Environmental Federation, Alexandria.

Zimmermann H.J. 1992. Fuzzy set theory and its applications. Kluver Academic Publishers.