Identifying characteristic situations in wastewater treatment plants

Karina Gibert
karina@eio.upc.es
Dep. of Statistics and Operation Research.
Universitat Politècnica de Catalunya.[1]
C. Pau Gargallo, 5. Barcelona. 08028. Spain

Ignasi R. Roda
eqairr@fc.udg.es
Lab. of Chemical and Environmental Engineering
Universitat de Girona
Campus Montilivi sn. Girona. 17971

Abstract.

In this paper the use of a new methodology which is a mixture of AI techniques and statistical ones for characterizing the typical situations in a wastewater treatment plant is presented.

This methodology includes several steps which come from the initial facing to data to the automatic generation of interpreting results.

The main goal of this paper is to share an intuitive approach to this methodology rather than going into technical details, as well as to show how it performs when applied to a wastewater treatment plants environment.

Keywords: Statistical Clustering, Knowledge Base, Wastewater Treatment Plant, Conceptual Interpretation of the results, Rules generation.

1 Introduction

The main goal of wastewater treatment plants is, in general terms, to guarantee the outflow water quality (referred to certain legal requirements), in order to restore the natural environmental balance which is disturbed by industry wastes, domestic waste-waters, etc.

The process used to achieve this goal is really complex and delicate; on the one hand, because of the intrinsic features of wastewater; on the other hand, because of the bad consequences of an incorrect management of the plant [10].

Just to have a general idea of what is happening in the plant, we provide a very brief description of the process: the water flows sequentially through three or four processes which are commonly known as pretreatment, primary, secondary, and advanced treatment (see [12] for a detailed description of the process). Figure 1 depicts its general structure:

- In the *pretreatment*, an initial separation of solids from wastewater is performed.
- *Primary* treatment consists of leaving the wastewater in a settler for some hours. Solids will deposit down the settler and could be sent out.
- *Secondary* treatment occurs inside a biological reactor. A population of microorganisms (biomass) degrades the organic matter solved in the wastewater.
- In the *advanced* treatment another settler is used to separate the water from the biomass.

The settler output (solids or biomass) produces a kind of mud which is the input of another process called *sludge line*.

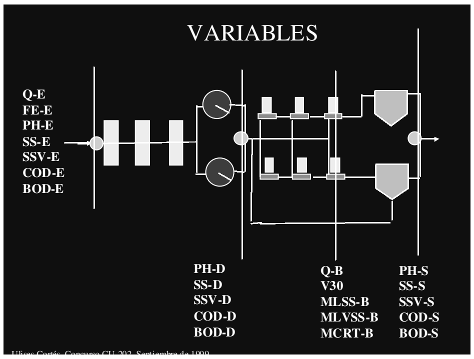

Figure 1. Inflow wastewater *versus* biological reactor-flow.

When the plant is not operating under normal conditions, decisions have to be taken to modify some parameters of the wastewater treatment process in order to reestablish the normality as soon as possible.

The aim of this paper is to present a methodology which combines statistical tools with inductive learning ones in such a way that, on the basis of several (numerical) measures, can identify which are the characteristic situations which can be found in the plant, and produce a *conceptual* description of them. Once identified and *understood* by the user, this typical situations may be used afterwards for supporting the decision making process. This decision making may be either automatic or not.

The methodology includes several steps which will be described along the paper. The more relevant ones are those of identifying typical situations (which is based on a new inductive learning– statistical clustering mixed strategy called *rule-biased clustering*[2] and generating conceptual description of the identified situations (which is especially focussed on this paper and presented in §5).

There are other ways to find a characterization of these situations. Frames, cases, or Bayesian networks could be used as well [1]. In the field of wastewater treatment plants other work has been done for finding characteristic situations [13], citemiquelkb…Using rule-biased clustering to achieve this objective is an effort to better understand the very nature of the classification process and to clarify the ways in which decision-making is performed in *ill-structured domains* [2].

[1] This research has been partially financed by the project *TIC'96*.

[2] The method proposed for that was originally called rule-biased clustering, but in this paper can be named as rule-biased clustering, in order to avoid confusion with other techniques used in the context of AI.

The paper is organized as follows: First of all, the data is presented in §2. Next, in section §3 a general view of the proposed methodology is presented. In §4 application of this methodology to the wastewater treatment plant data is detailed. In §5 different ways to produce objective characterization of the resulting classes are discussed. Finally, conclusions and future work come in §6.

2 The target domain

Data analyzed in this paper comes from the Waste Water Treatment Plant of Girona (in Spain). It is a sample of 396 observations taken from September the first of 1995 to September the 30th of 1996. Each observation refers to a daily mean, and it is identified by the date itself.

The state of the Plant is described through a set of variables which can be grouped as:

- Input (measures taken at the entrance of the plant)
  - Q-E: Inflow wastewater (daily m^3 of water)
  - FE-E Iron pre-treatment (g/l)
  - PH-E: pH
  - SS-E: Suspended Solids (mg/l)
  - SSV-E: Volatile suspended solids (mg/l)
  - COD-E: Chemical organic matter (mg/l)
  - BOD-E: Biodegradable organic matter (m/l)
- After Settler (measures taken when the wastewater comes out of the first settler):
  - PH-D: pH
  - SS-D: Suspended Solids (mg/l)
  - SSV-D: Volatile suspended solids (mg/l)
  - COD-D: Chemical organic matter (mg/l)
  - BOD-D: Biodegradable organic matter (m/l)
- Biological treatment (measures taken in the biological reactor)
  - Q-B: Biological reactor-flow
  - V30: Index at the biological reactor (measures the sedimentation quality of the mixed liquor, in ml/l)
  - MLSS-B: Mixed liquor suspended solids at the biological reactor
  - MLVSS-B: Mixed liquor volatile suspended solids
  - MCRT-B: Mean cell residence time at the biological reactor
- Output (measures taken when the water is meeting the river):
  - PH-S: pH
  - SS-S: Suspended Solids (mg/l)
  - SSV-S: Volatile suspended solids (mg/l)
  - COD-S: Chemical organic matter (mg/l)
  - BOD-S: Biodegradable organic matter (m/l)
- Other variables:
  - QR-G: Recirculated Flow
  - QP-G: Purged flow
  - QA-G: Air inflow

Figure 1 shows where all the measures are taken along the plant. It is easy to see that this set is heterogeneous and usually there are missing values and the sensors may provide noisy data.

3 The Methodology

The methodology proposed in this paper consists of the following steps:

1. Descriptive statistics of every variable: Use classical descriptive techniques to know the nature of the data.
   Very simple statistical techniques may be used to describe the data and to get some preliminary information about it; histograms may be used to make graphical representations of the variability of measures; time series plots may be used to observe evolution along time; plots and multiple box-plots may be used to observe the relationship between some pairs of variables [15].
2. Cluster-classify data (depending on the situation):
   The proposed method for that is *rule-biased clustering*. The original development of this method is in [2] and it was presented with details in [4]. It consist of a mixture of an AI process and a clustering one and it has been successfully used in several real applications previously . Our experience from previous applications is that using this *rule-biased clustering* use to be better than using any statistical clustering method by itself, since *rule-biased clustering* produces *understandable* classes. Also, it use to be better than pure inductive learning methods since it reduces the effects of missing some implicit knowledge in the Knowledge Base.
   As said before, the main goal of this paper is not going into specific details, but to give an intuitive idea of the method performance. For a detailed presentation see [5]. Anyway, the general idea is:
   - Use a (partial) Knowledge Base provided by the expert to induce an initial partition on part of the data; put the data not included in this partition in the *residual class*.
   - Generate prototypes of each *rule-induced class*.
   - Use a hierarchical clustering method to cluster the prototypes of rule-induced classes together with the residual class data.

   The result of *rule-biased clustering* is a dendrogramme which graphically represents the clustering process in form of a binary tree (see Figure **??**) where the height of each node indicates the distance between their children and the leafs of each node are the set of elements included in the class represented for this node.
3. Determine the final classes. This requires no new developments and consists in analyzing the dendrogramme to choose the better horizontal cut (where the largest gap is). This identifies a partition of the data in a set of classes.
4. Built a conceptual description of the resulting partition. This step requires automatic classes interpretation and is especially emphasized in §**??** and it is mainly based on the use of some statistical tools to extract qualitative information from the classes. The aim is to obtain a description directly understandable to the user and which indicates particularities of every class with respect the others.

In the next section the application of this methodology to the wastewater treatment plants is described.

4 The analysis

4.1 First contact to data

Just in order to show how descriptive statistics provides a powerful initial approach to data, main results of descriptive analysis are presented here. In conclusion we observed that:

- There is a group of variables with an approximate 25% of missing values (PH, SS and COD). In the rest of variables less than a 3% of missing value is observed.
- There is a short period (between 30-I-96 and 4-II-96) with outliers values for a number of variables. This period is likely to be carefully analyzed.
- In Figure 2 a *letter-plot* of Q-E and Q-B is shown, distinguishing the observations taken after the 8-II-96. Different performances of the plant before and after the 8-II-96 can be seen on the graphic. Experts were consulted about this matter and they confirmed a change on the politics of the plant: before 8-II-96 a maximum of 39000 m^3 of wastewater flowed into the biological reactor; if the inflow wastewater was greater than 39000 m^3, the excedent bypassed this phase of the treatment process, as a protection measure for the system; this explains the right hand side branch of the graph. After the 8-II-96, almost all the inflow wastewater is introduced in the biological reactor.

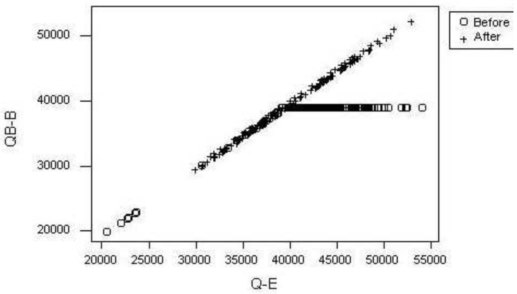

Figure 2. Inflow wastewater *versus* biological reactor-flow.

- All the parameters that denote contamination decrease effectively along the treatment, either in mean as in variability, as expected. As an example, Figure 3) depicts a multiple box-plot for the variable SS (suspended solids) distinguishing three moments of the treatment (input, after settler and output). Distribution of SS is displayed by means of a simple box-plot for each of input, after settler and output moments. It can be seen that SS is around lower values as treatment advances (boxes go down).

4.2 Rule-biased clustering

This methodology was previously applied to several real cases [4], [6], [5], [9], [8].

Briefly, it consist in a biased clustering process. The information used to bias the process is a small set of rules provided by the expert which contains some additional knowledge on this particular domain:

$$\begin{array}{llcll} r_1: & SS-S > 20 & \wedge & BOD-S > 35 & \rightarrow \mathcal{S} \\ r_2: & SS-S > 20 & \wedge & BOD-S < 35 & \rightarrow \mathcal{P} \\ r_3: & SS-S < 20 & \wedge & BOD-S > 35 & \rightarrow \mathcal{Q} \end{array}$$

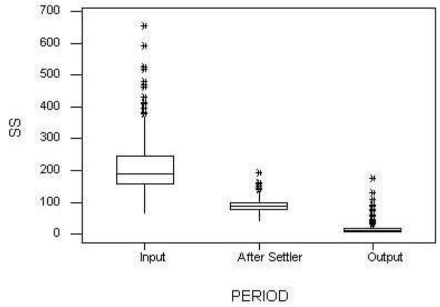

Figure 3. Suspended solids along the treatment

where $\mathcal{S}$ is the group called wrong-processing, $\mathcal{P}$ is solids-excess and $\mathcal{Q}$ organic-matter-excess. Rules are expressed as logic predicates and they are very easy to formulate. For instance, r_1 is expressing that the experts consider that wrong performance of the plant is taking place if the suspended solids at the output are over 20 mg/l and biodegradable organic matter are over 35 m/l.

The set of rules is kept small to show the power of the methods and to avoid evaluation conflicts, which usually appear as the set of rules grows.

This kind of information is hardly taken into account by pure statistical methods, and this may produce a set of classes which have not a clear semantics interpretation. In fact, if data is clustered without considering rules, the resulting classes are not very clear, form the meaning point of view. This is one of the reasons why we decided to consider additional expert knowledge for biasing the clustering process and to use *rule-biased clustering*.

Using these rules, *rule-biased clustering* was performed on several samples of different sizes (between 50 and 250 days each) from the data and a common structure of four classes was detected for all of them. One of the resulting dendrogrammes is shown in figure 4. It is very similar to the other ones, including the position of rules-selected objects (which in this case are the days themselves). The steps of *rule-biased clustering* are detailed bellow:

- First of all, the set of rules were evaluated on all the objects. This determined three groups of objects: those satisfying only rule r_1, those satisfying only rule r_2, and those satisfying only rule r_3. This is called the *rules-induced partition*.
  For every one of these three rules-induced parts, a classical process of statistical clustering was performed (in this case, reciprocal neighbors algorithm is used). This produced three small dendrogrammes, which can be seen with their corresponding labels ($\mathcal{S}$, $\mathcal{P}$ and $\mathcal{Q}$) in figure 4. The days which satisfied those rules can be seen from the labels of the tree leafs.
  For the height of those nodes, it is seen that the rules identified homogeneous groups of observations indeed. In fact, all the days involved with small dendrogramme P, for instance, are really similar among them (nodes with very low level). This means that the rules identified groups of days in which the plant presented very similar behavior.
  However, there were lots of days in which no one of the rules were satisfied (they were included in the residual class for the moment). It will be long and likely difficult to complete the Knowledge Base provided by the experts to correct this situation, since the process we are analyzing is very complex.

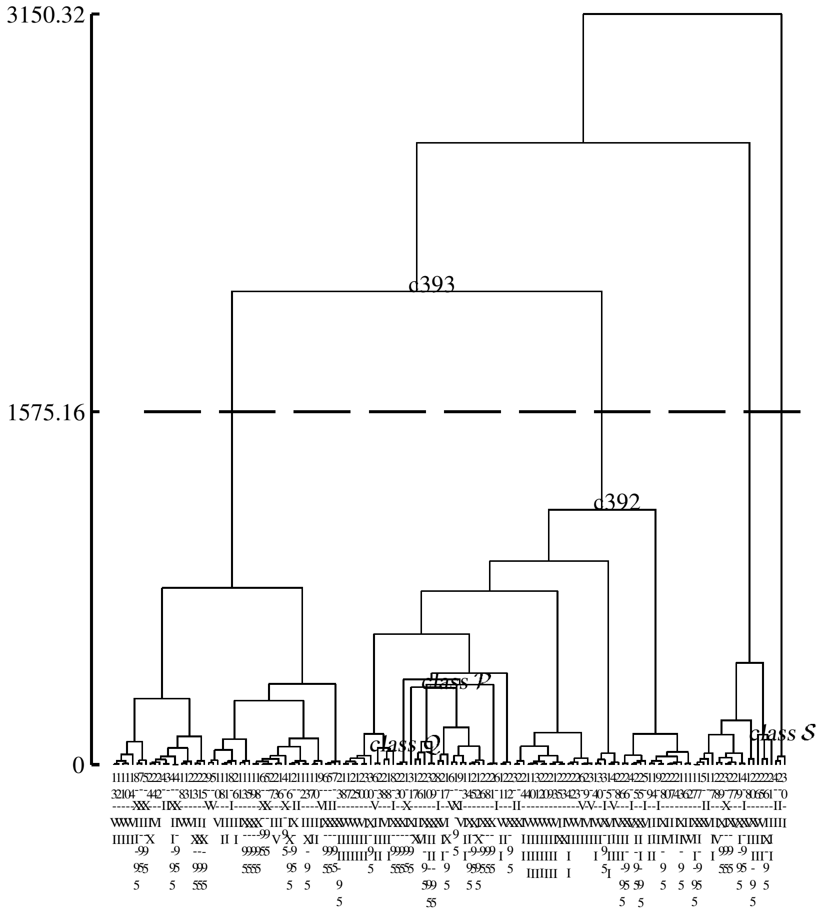

Figure 4. Dendrogramme for rule-biased clustering.

- *Rule-biased clustering* pretends to combine the information provided by the experts (even if it is partial —*i.e.* no complete—, as usually occurs in complex real domains) with a statistical processing of data, in order to integrate all the elements (the days for this application) in a unique hierarchical structure. In other words:

  > *"Let the clustering look for the structure where the experts did not say anything."*

  Cooperative interaction between reasoning techniques (from Artificial Intelligence, even if very simple) and clustering ones (from Statistics) is focussed to find a consistent and interpretable partition of data which can really support decision making.
  The final dendrogramme from Figure 4 was obtained performing a reciprocal neighbors algorithm with the days included in the residual class together with the prototypes of rules-induced classes $\mathcal{S}, \mathcal{Q}, \mathcal{P}$.
- Deciding the number of classes: First point to highlight is that main structure of four classes seems to be clear on the dendrogramme, since the gap between classes $c392$ and $c393$ indicates a big change on the homogeneity of the classes formed before and after that point. So, the resulting classes are those determined by the horizontal line in Figure **??**. The days included in each class are the set of leafs of each subtree.
  These classes includes small kernels determined by the expert rules (nodes labeled$\mathcal{S}$, $\mathcal{Q}$, $\mathcal{P}$) . In this specific application, the result proposed by the automatic process seems to indicate that group $\mathcal{P}$ and $\mathcal{Q}$ could be dealed together, since they are in the same class. This indicates that considering together all the days and making a global view, there are other situations much more different than those represented by $\mathcal{P}$ and $\mathcal{Q}$. Let us see, in the next sections, if the proposed classes really make sense from a semantics point of view.
  Of course, taking a low number of classes will produce a general classification of situations in the plant, as it will be seen bellow. More particular situations will be identified, taking a lower horizontal cut of the tree, that is taking a highest number of classes. Interpreting general classes is always more difficult than finding special features in particular ones. Thus, for the purpose of this paper, a general classification of four classes was preferred.

5 Interpreting the results

In fact, one of the main problems of clustering techniques is that validation of the results is still now an open problem, since there is not an objective criteria to determine the quality of a set of classes. It is easy to evaluate a set of classes in terms of accuracy if there exists a reference partition of the data and comparison is possible. But, unfortunately, the most situations where clustering is required are those in which no good knowledge of the domain exists, unless partial, and this approach is not useful at all. Only the *usefulness* of a classification can be used to decide if it is correct or not [6].

Evaluating the usefulness of a given classification requires a mechanism to *understand* the meaning of the identified classes in order to finally decide if they are useful or not.

This process, usually known as *interpretation of the resulting classes*, use to be done by the user, in a non-systematic way, using its own experience, and becomes difficult as the number of resulting classes increase, as well as the number of variables used for describing data also increase.

If the classes are useful, then the discovered structure can be used as new knowledge for decision making.

In [2] and [3] a first approach for an efficient algorithm which generates automatic interpretations of the classes is introduced. In those documents, the main work is about the interpretation of the classes on the basis of categorical (qualitative) variables.

In this paper, we are interested on the use of numerical variables for interpreting classes. The real application we presented here is especially indicated for this goal, since categorical variables are not present in the data matrix at all and only numerical measures are useful for describing data.

In this phase, we want to identify the issues that allow conceptual interpretations of classes, on the basis of numerical variables, in order to establish a formal methodology in future phases of the research, actually in progress.

The analysis was iterated on several samples, and no special differences were found among the results, referring to the final interpretation. Of course, from one sample to another, different number of objects was included in the classes. However, as said before, this does not imply major changes in the *meaning* of the resulting classes. For simplicity, the results of one of the samples is presented in this paper.

The analysis of the meaning of the classes was inspired in the ideas used in [2] and [3] for qualitative variables. In that work, the main goal was to identify the variables with exclusive values in a given class. To extend these ideas to numerical variables, we started to study multiple box-plots.

5.1 Multiple box plots through classes

For short, a *multiple box plot* can show the relationship between a common numerical variable and some groups. For each class the interval of values taken by the variable is visualized and rare observations (outliers) are marked as "*". A box is displayed from Q1 (first quartile) to Q3 (third quartile) and the Median is marked with an horizontal sign. Boxes include, then, the 50% of the elements of the class. In this case, we could use the multiple box-plot to visualize the distribution of a given variable in all the classes[3] (see figure 5).

With this representation it is easy to see if there is a class in which some variable has special values. In this particular application, this is found, for instance, in the *inflow wastewater (Q-E)* for class 4 (see figure 5).

This fits in the definition of *characterizing variable* for class $\mathcal{C}$originally given in [2]:

A variable X_k is a characterizing variable of a given class $\mathcal{C}$ if

$$\exists \Lambda_{\mathcal{C}}^{k} : \forall i \in \mathcal{C}, x_{ik} \in \Lambda_{\mathcal{C}}^{k} \cap \forall i \notin \mathcal{C}, x_{ik} \notin \Lambda_{\mathcal{C}}^{k}$$

This means that the values taken for X_k in objects of class $\mathcal{C}$ are not taken by any other objects from other classes and viceversa.

In the example, the special values of *Q-E* for class 4 are the set: $\Lambda_{\mathcal{C}_4}^{Q-E} = [20000, 25000]$. All the days included in class 4 satisfy that their Inflow wastewater level is between [20000,25000]. So the value of variable $Q - E$ can be used to identify if a gived day is from class 4 or not.

The *status* of characterizing variable of a given class is relative to a given partition of the objects and it may be or not, depending on the variable values for the other classes.

On the other hand, a class may have several characterizing variables, and all of them can identify the class by itself.

In the application, the variable *biological reactor-flow* (QB-B) is also a characterizing variable for class 4 (see figure 6) with

[3] The fifth box-plot (on the very right) refers to the days excluded from the sample. It must not be taken into account in this discussion.

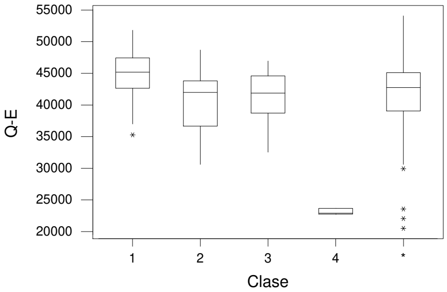

Figure 5. Multiple box-plot of variable *(inflow wastewater) Q-E.*

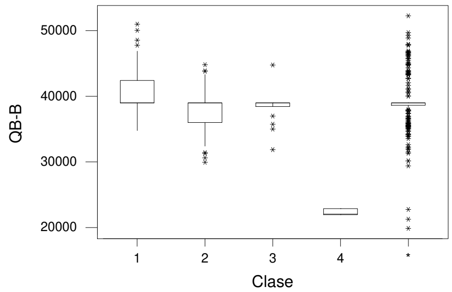

Figure 6. Multiple box-plot of variable *biological reactor-flow (Q-B).*

$$[\Lambda_{\mathcal{C}_4}^{Q-B} = [20000, 25000]$$

So, class 4 can be equally identified as the class with lowest values of *Q-E* (20000-25000), or as the class with lowest values of *Q-B* (the same values).

Given a partition of the individuals set $\mathcal{P} = \{\mathcal{C}_1, \mathcal{C}_2, \dots \mathcal{C}_n\}$ a *characterization system of* $\mathcal{P}$ can be defined as:

$$\mathcal{S}_\mathcal{P} = \{(\mathcal{C}, X_k, \Lambda_\mathcal{C}^k) : \mathcal{C} \in \mathcal{P} \wedge \Lambda_\mathcal{C}^k \neq \emptyset\}$$

which is the set of variables that can be used to identify some of the classes.

$\mathcal{S}_\mathcal{P}$ is a *complete characterization system of* $\mathcal{P}$ if

$$\forall \mathcal{C} \in \mathcal{P}\ \exists k : (\mathcal{C}, X_k, \Lambda_\mathcal{C}^k) \in \mathcal{S}_\mathcal{P}$$

In other words, all the classes can be identified by, at least, one variable.

Finally, $\mathcal{S}_\mathcal{P}$ is minimal if given $\mathcal{C}, \mathcal{C}' \in \mathcal{P},\ \exists k, k'$ such that

$$(\mathcal{C}, X_k, \Lambda_\mathcal{C}^k), (\mathcal{C}', X_{k'}, \Lambda_{\mathcal{C}'}^{k'}) \in \mathcal{S}_\mathcal{P}, \text{ then } \mathcal{C} \neq \mathcal{C}'$$

(there is only one variable to identify every class).

If a minimal and complete characterization system is found for a given partition it is possible to determine the class of a new object on the basis of the values it takes just for a reduced number of variables. This is, in fact, one of the problems faced in he field of supervised machine learning, from which ID3 and its successors are characteristic exponents. Building decision trees use to have exponential cost, although there are techniques for shorting the search space. In this paper, we explore a cheaper possibility, which also is useful for numerical variables.

The above definitions apply, without difference, for numerical and qualitative variables, although different formulations apply to numerical variables.

Looking at the multiple box-plot, it is clear that X_k is a characterizing variable of class $\mathcal{C}$ if it satisfies the following condition:

$$\forall \mathcal{C}' \in \mathcal{P}, \mathcal{C}' \neq \mathcal{C} : [min_\mathcal{C}^k, max_\mathcal{C}^k] \cap [min_{\mathcal{C}'}^k, max_{\mathcal{C}'}^k] = \emptyset$$

which can be computed independently of the graphical representation. However, when numerical variables are considered, it is rather rare to find complete characterization systems of a given $\mathcal{P}$, especially when analyzing real data.

In the example, only class 4 generates characterizing variables *strictu sensu*.

So, we propose to *relax* the concept of characterizing variable for obtaining complete characterization systems easily.

Instead of looking for a characteristic interval

[*minimum, maximum*]

,

we could look for a less restrictive one, for example

[Q1, Q3]

(which is represented in the graph 5 by the boxes). This is equivalent to consider the central 50% of the class.

So, a *0.5-characterizing variable of* $\mathcal{C}$ satisfies the following:

$$\forall \mathcal{C}' \in \mathcal{P}, \mathcal{C}' \neq \mathcal{C} : [Q1_\mathcal{C}^k, Q3_\mathcal{C}^k] \cap [Q1_{\mathcal{C}'}^k, Q3_{\mathcal{C}'}^k] = \emptyset$$

The main idea is to look for boxes [Q1-Q3] which do not intersect with the boxes of other classes. This is a first approach which is very easy to evaluate either from the graphics or *by calculations*. However, other possibilities in between are also considered (ε*-characterizing system*, with a variable $\varepsilon \in [0, 1]$) and other proposals are actually in progress at present on the basis of these ideas [14].

For the moment, an algorithm to find the ε-characterizing variables of a set of classes was derived; of course it is not consulting the multiple-boxplots themselves, but using the formal definitions of characterixing variables dervied from the observation of graphical representations.We can say that multiple-boxplots were used to design the algorithm for finding the characteristic variables of a set of classes.

For the specific application we are dealing with in this paper, the 0.5-characterizing system automatically found is (not minimal):

$(\mathcal{C}_4, Q-E, [20000, 25000]),$
$(\mathcal{C}_1, Q-B, [39000, 43000])_{0.5},$
$(\mathcal{C}_4, Q-B, [20000, 25000]),$
$(\mathcal{C}_4, QR-G, [11000, 12000]),$
$(\mathcal{C}_4, QP-G, [0, 0.]),$
$(\mathcal{C}_4, QA-G, [100000, 150000]),$
$(\mathcal{C}_4, FE-E, [0, 0]),$
$(\mathcal{C}_3, SS-E, [300, 400])_{0.5},$
$(\mathcal{C}_4, SS-E, [50, 100]),$
$(\mathcal{C}_3, SSV-E, [300, 400])_{0.5},$
$(\mathcal{C}_4, SSV-E, [50, 100]),$
$(\mathcal{C}_3, COD-E, [1500, 700])_{0.5},$
$(\mathcal{C}_4, COD-E, [0, 250]),$
$(\mathcal{C}_3, BOD-E, [200, 450])_{0.5},$
$(\mathcal{C}_4, BOD-E, [0, 0]),$
$(\mathcal{C}_4, SS-D, [0, 75]),$
$(\mathcal{C}_4, SSV-D, [0, 40]),$
$(\mathcal{C}_4, COD-D, [0, 150]),$
$(\mathcal{C}_4, BOD-D, [50, 50]),$
$(\mathcal{C}_2, SS-S, [15, 30])_{0.5},$
$(\mathcal{C}_4, SSV-S, [0, 5]),$
$(\mathcal{C}_1, SSV-S, [5, 15])_{0.5},$
$(\mathcal{C}_2, SSV-S, [15, 22])_{0.5},$
$(\mathcal{C}_4, BOD-S, [0, 0]),$
$(\mathcal{C}_2, BOD-S, [18, 30])_{0.5},$
$(\mathcal{C}_4, MLSS-B, [2800, 3000]),$
$(\mathcal{C}_4, MLVSS-B, [1900, 2000]),$
$(\mathcal{C}_4, MCRT-B, [280, 350]),$

It can be appreciated that almost all the variables can identify class 4 *strictu sensu*, since they take very extremely small values for this class, except for variables *MLSS-B, MLVSS-B* and *MCRT-B*. Those variables take values extremely high in C4. The elements that are only 0.5-characterizing variables are indicated with a subindex 0.5.

With a simple processing of these results it is easy to generate a set of characterizing rules for identifying the class of a given new object:

if variable Q-E $\in$ *[20000, 250000] for object* $i \longrightarrow i \in \mathcal{C}_4$
if variable MLSS-B $\in$ *[1900, 2000] for object* $i \longrightarrow i \in \mathcal{C}_4$
if variable SS-E $\in$ *[300,400] for object* $i \longrightarrow i \in \mathcal{C}_1$,
with some error[4] *related to* ε.

Or, in other words:

if variable Q-E is very low *for object* $i \longrightarrow i \in \mathcal{C}_4$
if variable MLSS-B is very high *for object* $i \longrightarrow i \in \mathcal{C}_4$
if variable SS-E is high *for object* $i \longrightarrow i \in \mathcal{C}_1$,
with some error related to ε.

In terms of practice, actions can be associated to each class[5] (performing some analytical test, close input door, increment the biomass of the biological reactor,…). Given the class of a new day, what to do will be determined by the rules.

In fact, more accurate satisfactory results can be obtained if ε is allowed to be variable along a Characterization system. This is, also, treated in another paper, actually in press [14].

Although in this particular application, all the classes can be characterized with, at least, one variable, existence of characterizing variables for every classes cannot be assured unfortunately, and more work is needed to determine the way for completing a characterization system.

5.2 Considering time

However, in this concrete application, data also describes the dynamics of the process, since the same variables (Q, pH, SS, …) are measured in several points of the plant, *i.e.*, entrance, after settler, biological reactor, output, as indicated in figure 1. This allows dynamical analysis based on comparing the values of those variables in different points of measuring. In this case multiple box-plots through different points of the process were used.

As an example, figure 7 shows the evolution of variable *suspended solids* along the plant in four multiple box-plots: one for each class. Each box-plot represents the distribution of suspended solids at the entrance (number 1 in the graph), after Settler (number 2) and at the output (number 3) locally to each class. The interpretation is that class 3 has more suspended solids than the rest at the entrance. Reduction of suspended solids at the settler is much more important for this class than for the others.

Using this kind of graphics, expert concluded that:

- Class3: Great effectiveness of the Settler in reduction of suspended solids and volatile suspended solids.
- Class4: Very lower values of the variables listed bellow at the input, and, as a consequence, after settler, in the biological reactor and also at the output:
  - Suspended solids
  - Volatile suspended solids
  - Chemical organic matter
  - Biodegradable organic matter
  - Flows

Considering that data are sequential along time, time series have been also considered, but it seems to be difficult to involve them in an automatic interpretation process at the moment. More work has to be done on this direction. That is why we focus here only on the multiple box-plots.

5.3 Introducing supplementary variables

Richer interpretations can be achieved if supplementary variables are used at this stage. The concept of *supplementary variable* is taken from the Multivariate Data Analysis environment [11]. A supplementary variable is a variable which is not included in the analysis (in this case, it has not been involved into the rule-biased clustering process), but it is used for interpretation. In our case, there are some other variables not used for the clustering, mainly because of the great number of missing values they presented as said in §4.1, which can be graphed according to the classes defined by the rule-biased clustering.

This variables refers to the result of some analytical test performed under certain conditions in the plant. From the multiple box-plots of those variables, we concluded that:

[4] The relationship between ε and the risc of miss-classification is the object of another paper.
[5] with some degree of error, if needed.

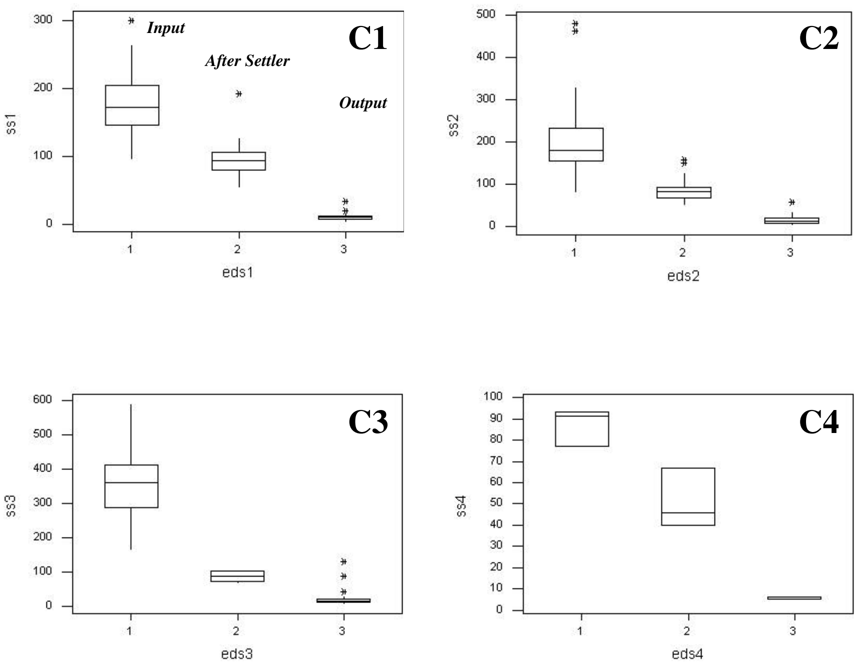

Figure 7. Dynamical evolution of suspended solids (SS) along the treatment.

- C1 has an spectacular reduction of ammonium, NKT between points after Settler and output, what means that ammonium is reduced in the biological reactor (see 8).

5.4 Final interpretation

Using the multiple box-plots through the classes (and the algorithm derived from them) or through different pionts of the wastewater treatment either for active variables (those included in the clustering process) or supplementary ones, the following interpretation is produced. Experts opinion is also included in the discussion bellow, as they confirmed the understandability of the discovered classes.

- Class1: The output water is clean. The input water is medium dirty (intermediate values in almost all the variables). It is remarkable that NKT[6], is highly reduced at the biological reactor.
  Experts identified this class with the class of those days with very good plant performance. As a consequence of the good conditions, even ammonium is reduced, although ammonium control is not among the control parameters of this plant (it is not supposed to be reduced by this kind of plants).
- Class2: The water comes in and out with intermediate values of all the variables. Nonetheless, the performance of the plant is not so good. This class contains some days in which isolated control parameters overcome the permited values. This class includes the $\mathcal{P}$-like and $\mathcal{Q}$-like situations and, from the expert point of view, it is interpreted as the class of days with general good performance, but in which punctual problems can produce some isolated parameter with high value (SS-S, BOD-S).
- Class3: Water comes in very dirty. High number of suspended solids at the entrance. Settler is making high effect (levels of suspended solids and volatile suspended solids... are significantly reduced at the primary treatment, *i.*. the settler.
- Class4: Almost all the variables take very low values. Few input, few output. V30, SSLM, but cells age very high. Experts could tell us that days included in this class correspond to storm periods, and they coincide with those outlier days detected at the very beginning with the descriptive tools §4.1.
  To face the storm the decision of plant head was to minimize the water inflow, closing the input valves, and to maintain the microorganisms (by minimizing the purged flow) in order to protect the system. This produce increasing of biomass in the reactor. As a consequence, the age cells — which is computed as an inverse function of the purged flow (but Q-P$\rightarrow 0$)—, appears with enormous values which, in fact, makes no-sense (they represent ∞).

6 Conclusions and future work

Application of rule-biased clustering to the classification of situations in a Wastewater Treatment Plant is presented. The aim of this paper is to present a complete methodology for extracting usefull and directly understandable knowledge from a set of numeric data comming from complex real domains using a combination of inductive learning and statistical tools.

In section §4 rule-biased clustering was focussed and §5 shows how to use some statistical graphs for extracting conceptual interpretations of the identified classes. in fact, the features observed on the graphs can be automatically found using a non-graphical algorithm, like finding characterizing variables. In future works the final implementation of this algorithm will be applied to this data.

This work constitutes a positive experience which is the first step of establishing a formal methodology to automatically obtain conceptual interpretations of classes, on the basis of numerical variables used to describe objects (days is this case).

From a very small and partial set of rules, the system created a new level of abstraction that catches the nature of a wastewater treatment plant for the given set of data (noisy, incomplete and heterogeneous), producing a set of identified classes as well as a conceptual interpretation of them, which was directly understood by the experts on this matter.

At the moment, we are working in the search of *maximum intervals* exclusive for a given class, instead of using a fixed interval such as $[minimum, maximum]$ or $[Q1, Q3]$, used in this paper. This generalization implies generation of concepts with variable degree of accuracy, but produces even reacher interpretations (observe for instance that 75% of elements in Class 1 satisfies that $Q - B \geq 39000$ and in this paper, only the 50% interval [39000, 43000] was used).

We realized that, in fact, it is easier to calculate the accuracies when ε is variable than when it is fixed. Interpretations of the presented classes using this new (and more flexible) method is advanced and will be presented in [7].

In the future we want to explore the use of time series plots to include sequential information in the interpretations.

ACKNOWLEDGEMENTS

We are grateful to Dr. Manel Poch for advising this paper as an expert in wastewater treatment plants. His comments helped to improve this paper.

We want to thank especially Daniel Rodríguez Cano and Silvia Bayona Huguet for his participation in processing data.

REFERENCES

[1] U. Cortés, M. Sânchez-Marr, L. Ceccaroni, I. R-Roda, and M. Poch, 'Artificial intelligence and environmental decision support systems', *Applied Intelligence*, (2000).

[2] K. Gibert, *L'ús de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats*, Ph.D. dissertation, Politechnical University of Catalonia, 1994.

[3] K. Gibert, T. Aluja, and U. Cortés, 'Knowledge discovery with clustering based on rules. interpreting results', in *Principles of Data Mining and Knowledge Discovery*, Springer-Verlag, (1998).

[4] K Gibert and U. Cortés, *Combining a knowledge based system with a clustering method for an inductive construction of models*, Springer-Verlag, 1994.

[5] K. Gibert and U. Cortés, 'Clustering based on rules and knowledge discovery in ill-structured domains', *Computación y Sistemas*, (1998).

[6] K. Gibert, M. Hernández, and U. Cortés, 'Classification based on rules : an application to astronomy', in 5^{th} *Conference of International Federation of Classification Societies*, (1996).

[7] K. Gibert and A. Salvador, 'Aproximación difusa a la identificación de situaciones caracter´ísticas en el tratamiento de aguas residuales', in *Congreso español sobre tecnologías y lógica fuzzy*, (2000).

[8] K. Gibert and Z. Sonicki, 'Classification based on rules and medical research', *Journal of Applied Stochastich Models and Data Analysis*, (199).

[9] K. Gibert and Z. Sonicki, 'Classification based on rules and medical research', in *VIII Int'l Symposium on Applied Stochastic Models and Data Analysis*. Rocco Curto, (1997).

[10] J.M. Gimeno, Béjar J., and U. M. Sànchez-Marrè, Cortés, 'Discovering and modelling process change: An application to industrial processes', in *Practical Applications of Data Mining and Knowledge Discovery*.

[11] A.and Morineau Lebart and Fenelon, *Traitement Statistique des données*, DUNOD, 185.

[6] To be exact, NKT measures the sum of ammonium and organic nitrogen, although its value is mainly caused by the ammonium level.

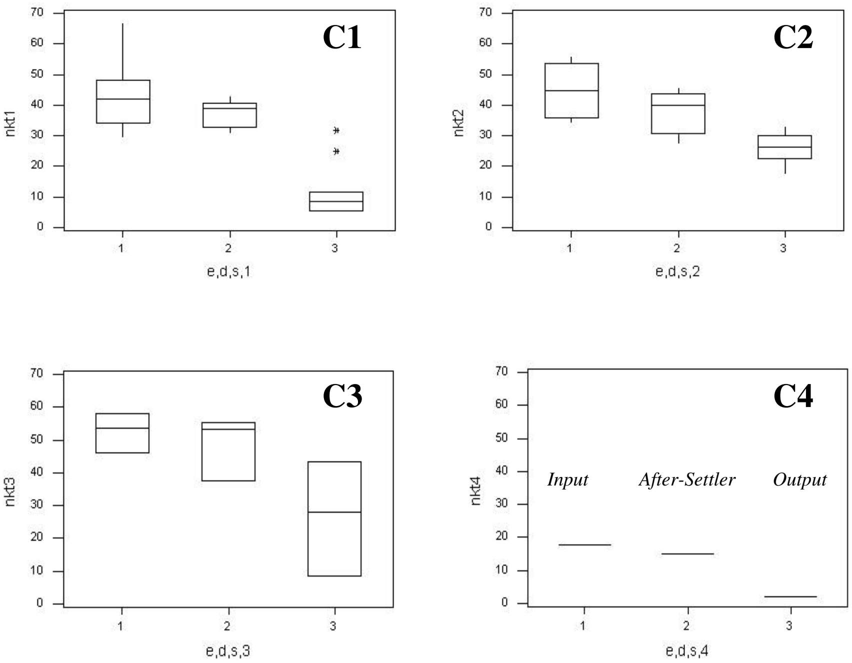

Figure 8. Evolution of NKT along the treatment.

[12] Metcalf and Eddy Inc., *Wastewater engineering: treatment/disposal/reuse*, McGraw-Hill, 1991.

[13] I. R-Roda, M. Poch, M. Snchez-Marr, U. Corts, and J. Lafuente, 'Consider a case-based system for control of complex processes', *Chemical Engineering Progress*, (1999).

[14] A. Salvador, *Identificació de situacions caracteristiques en plantes depuradores*, Master's thesis, Facultat de Matemàtiques i Estadistica, 2000.

[15] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.