EOLI
WP 7 : Supervision System development
Deliverable D 7.3 : Knowledge based system able to recognize and classify a given "expected" but abnormal situation
Francesco Ciappelloni, Djalel Mazouni, Jérôme Harmand

1. Introduction

This document presents the DSS developed for the EOLI solution. Because it is strongly related to the DSS structure, the same remark than in the Deliverable D 7.1 can be made here, that is although the proposed solution is "open" in the sense it can be applied to any SBR process equipped with the minimal required on-line sensors, it was especially developed with respect to the pilot plant of the INRA-LBE, Narbonne and it should be kept in mind that its use for supervising another plant would require some adaptation.

This report presents MATLAB version of the software developed for analyzing on-line data of an aerobic SBR, detecting deviations from normal operation and, in this case, proposing the most probable causes of this deviation. It is a supervision system instead of a diagnosis system since it acts on the long-term, detects problems in using a complete phase data (and does not work on-line but a posteriori from one cycle data to the next cycle). The system uses a statistical method divided in two main phases: off-line model building and "on-line" data diagnosis. As explained above, the term "on-line" refers here to the use of the data of the entire last cycle for decisions concerning the next one. The off-line model identifies the correct working conditions of the system once (standard operative conditions) at the very beginning of the process work (just after startup). It includes the characterization of the possible deviations of the system from these standard conditions in the case of changes in the biomass properties or carbon and nitrogen load characteristics. The on-line diagnosis aims at collecting and analyzing all the data available through industrial sensors, and at classifying the behavior of an entire treatment cycle.

Remark : At this date (T0+22), the software sensors to be developed within the EOLI project have not yet been implemented on the LBE plant (the POLIMI hardware sensor for the monitoring of the nitrification should be installed in a few months on T0+27). Thus, the version of the above presented DSS does not take into account the possible information that such additional sensors would provide. However, it is expected that only small changes would be necessary to include them in the system and get more inside, especially in the determination of the most probable cause of a detected abnormal situation.

2. The supervision system : classification of reaction phases

The developed system aims at computing a "distance" of a given cycle (one anoxic + one aerobic period) from a reference called "standard". In particular, the objective is to detect and isolate a COD or TKN overload, a problem in the oxygen transfer rate (Kla) or related to the heterotrophic or autotrophic bacteria. The fact that the detection is realized a posteriori (once the cycle is finished) is not a problem because the consequences of an overloading or of a problem on the oxygen transfer or on the biomasses are rather long term consequences rather than short terms ones. For instance, a "durable" COD overload will lead in problems on sludge settleability several cycles after this overload has began. The most probable cause of the identified problem is then displayed to the operator who can adjust the time of the

subsequent reaction phases or take any other appropriate decision (calibration, withdrawing of the sludge, etc).

Regarding the objectives of the study, the aerobic phase is the most important period to investigate the process performance. In fact, during this phase, significative information can be extracted from the dissolved oxygen, pH and Redox curves. In particular, using either the pH or the DO curve (as shown in figure 1), it is possible to detect when each substrate is completely degraded. The first point corresponds to the end of the COD consumption, the second corresponds to the end of the nitritation while the last one corresponds to the end of the nitratation. Based on this information, notice that an interesting optimization strategy is to switch to the next phase (settle in this case) as soon as the COD and NH4 are degraded.

The detection of the critical points defined hereabove in figure 1 is very useful for classifying a given cycle. These points define three distinct periods : the first one comprised between the beginning of the aerobic period and T1, the ending time of the COD removal (between 2 and 3 hours in figure 1), the second comprised between T1 and T2 defining the end of the nitritation (between 3 and 5 hours in figure 1) and the third one comprised between T2 and T3 defined as the end of the nitration (between 5 and 7 hours in figure 1). In particular, using the temperature to compute the oxygen saturation, notice that it is possible to determine the quantity of oxygen respectively needed in each of these periods (noted H1, H2 and H3). While the Ti parameters give us immediate information about the degradation rates, the Hi parameters are proportional to pollutant quantity. Of course, these values are not exact values since biological reactions evolve simultaneously (for instance both the COD removal and the nitrification begin when the dissolved oxygen differs from zero), however, the oxygen consumption rates significantly differ between the different phases in order this uncertainty to be neglected.

The Ti and Hi parameters are then used as descriptors in the framework of a Principal Component Analysis (PCA) approach. In the proposed software, two different approaches are proposed. The first one is a model-based approach where the projection basis for the PCA procedure is calculated using data generated by a mathematical model of the SBR plant. This approach is particularly interesting because it allows to simulate - in an acceptable time - a large number of simulations corresponding to a large spectrum of environmental conditions (different C/N ratios, different oxygen transfer rates, heterotrophic and autotrophic biomass concentrations, etc). However, it also means that a validated model of the plant is available : such a model is usually difficult to obtain and its validation is time consuming. The other available approach makes use of a limited number of real cycles for which the functioning conditions are well known provided that the available data have been obtained in a large range of functioning conditions. This approach has the advantage of being rather more robust than the model based approach because it does not use any model of the process but only the available on line measurements.

3. The model-based approach

The model used for the simulations is EM2a model (cf. Deliverable D 2.3) for which parameters have been optimized for fitting LBE data. The data to be used in the PCA procedure can be generated using this model. A large number of simulations can then be run around a standard cycle defined by the following operating parameters : Kla in aerobic and in anoxic conditions, concentrations of COD and NH4 in the input, concentrations of biomass (XH and XA) in the tank, volume of the input and initial volume of the tank, duration of

anoxic and aerobic periods, duration and frequency of air pulses during the anoxic phase, parameters Ti imposed as "optimal", parameters successively used to vary the conditions of the simulations (Kla, CODin, NH4in, XH, XA), ranges of parameters variation.

In order to be as open and as generic as possible, the software leaves the possibility to automatically adapt the values of the maximum growth rates model parameters. To do so, it is assumed that the operator is able to enter the theoretical values (defined as "optimal") of the duration of each cycle for the standard conditions. In the previous list, the first seven parameters are used to identify the μ_{maxH}, μ_{maxNS}, μ_{maxNB} (maximum growth rates for heterotrophic, nitrifiant and nitratant biomasses). The last two parameters are needed to compute the operating parameters for each simulation. Usually the most important "process faults" to be detected are related to the Kla, the CODin and/or the NH4in. The operator can define the range of variations of these parameters from their standard value. The possible variations of a parameter are:
- << : very low (e.g. -50% of nominal value);
- < : low (e.g. -25% of nominal value);
- > : high (e.g. +25% of nominal value);
- >> : very high (e.g. +50% of nominal value).

After the compilation and the execution of all the simulations, a database is created and will be used for the next step : the application of a PCA to reduce the space of analysis from six parameters to less.

4. The PCA study

From now, the procedure is the same for the model-based approach and for the data-based approach. The Principal Component Analysis (PCA) is one of the most popular methods for extracting information from data, and has found application in a wide range of disciplines : data rectification in chemical process operation and control, disturbance detection and isolation, statistical process monitoring and fault diagnosis. PCA transforms the data matrix in a statistically optimal manner by diagonalizing the covariance matrix and by extracting the cross correlation or relationship between the variables in the data matrix. If the measured variables are linearly related and are contaminated by errors, the first few components capture the relationship between the variables, and the remaining components are comprised only of the errors. Thus, eliminating the less important components reduces the contribution of errors in the measured data and represents it in a compact manner. The popularity of PCA relies on its high ability to reduce the dimension of any data matrix while capturing the underlying variations and relationships between the related variables.

In our particular case, the data matrix includes the Ti and Hi of all the available experiments (for the data based approach) or simulations (for the model based approach). Once the projection basis has been computed, the original matrix (six parameters for each considered cycle) is transformed into another matrix in which the columns are the principal components. In our case (in both the model based approach and in the data based approach), it was noticed that more than 90 % of the data can be explained in only looking at what happens in the first two principal components. In other words, most of the time, it will be possible to discriminate two different cycles in projecting its six associated descriptors Ti and Hi onto a plan.

4. The Decision Table for the identification of the most probable cause of a problem

Once the characteristics of a cycle have been projected onto the discriminating plan P via the PCA, the next step is to identify a possible fault affecting the process. It is clear that if two cycles are very similar (similar Ti, Hi), then their parameters projected onto P will be very close one from each other. Then, once projected, the underlying idea is to search for the closest points. Since this plan has precisely been designed in order to discriminate different data, it is expected that two points that are close reflect the same process behavior. Assuming experiments or simulations used for computing the Principal Components (coordinates that are used to project any new parameters extracted from a given cycle) have covered a range sufficiently large of functioning conditions, it becomes possible to associate the state of a specific point to the state of the closest point.

At this step, a number of decision algorithms can be used. One of the most popular procedure is the k-closest neighbors that consists in associating the actual point to the majority class of its k-closest neighbors. If this algorithm is particularly attracting in a supervised context (all points have been associated to a fixed number of predefined known number of classes), its use is less obvious in the present case. Indeed, only the point corresponding to standard conditions can be classified in a "normal class". The other points correspond to specific conditions that differ from the standard conditions by several parameters. The way we have proceeded is as follows. A parameter x is first defined. If there is a point defined as "normal" (corresponding to the standard conditions) in the vicinity x of the point p, then the point p is assigned to be itself in this set. If it is not the case, a decision table (such as the one pictured in Table 1) is built. In each column, the sum of the inverse of the squared distances to all classified points of the plan are specified. The explanation corresponding to the point p corresponds to the case of the table with the greatest value. This procedure is a kind of "weighted explanation table".

It should be noticed here that the developed tool is very generic and not only valid for decisions in a plan. It seems that when the process faults the user is interested in is sufficiently small – as in our case where we are only interested in identifying overloads and problems in the oxygen transfer rate - two dimensions in the discriminating space are sufficient. However, the number of PCs to be finally retained remains under the responsibility of the user : providing that a large number of experiments covering a large range of functioning conditions (including possible simultaneous combination of explanation variables) are available (or a large number of simulations can be run in the model based approach), it would be possible to identify any problem as long as its "signature" in the discriminating plan is known.

5. Annex : Figures and Tables

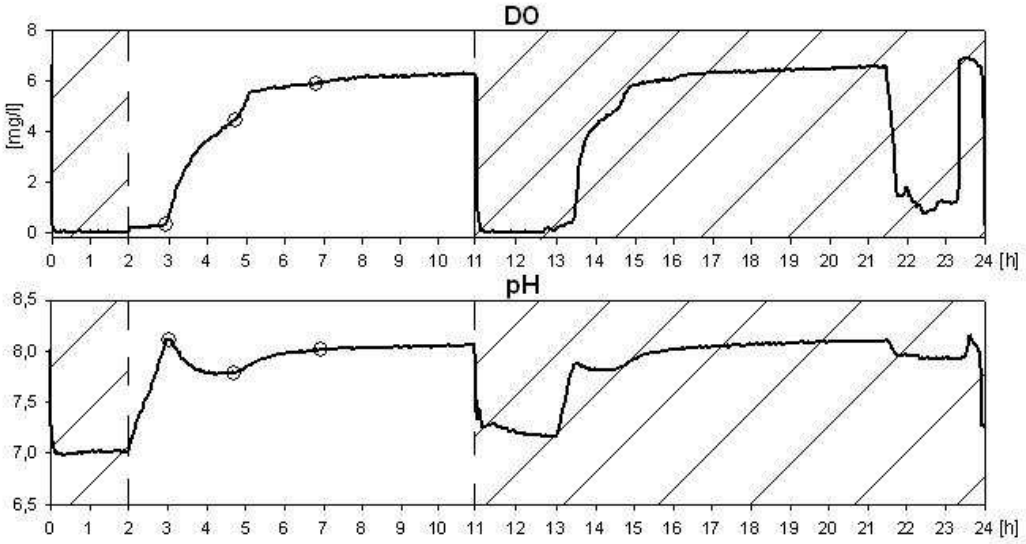

Figure 1 : pH and Dissolved oxygen curves (25/01/04)

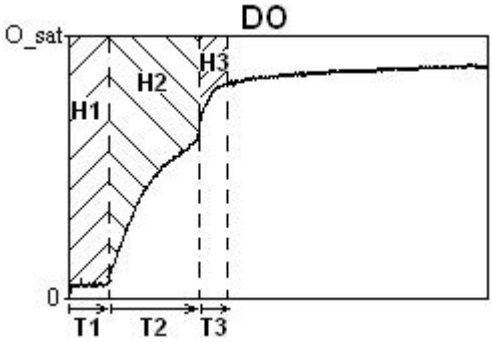

Figure 2 : Definition of parameters Ti and Hi

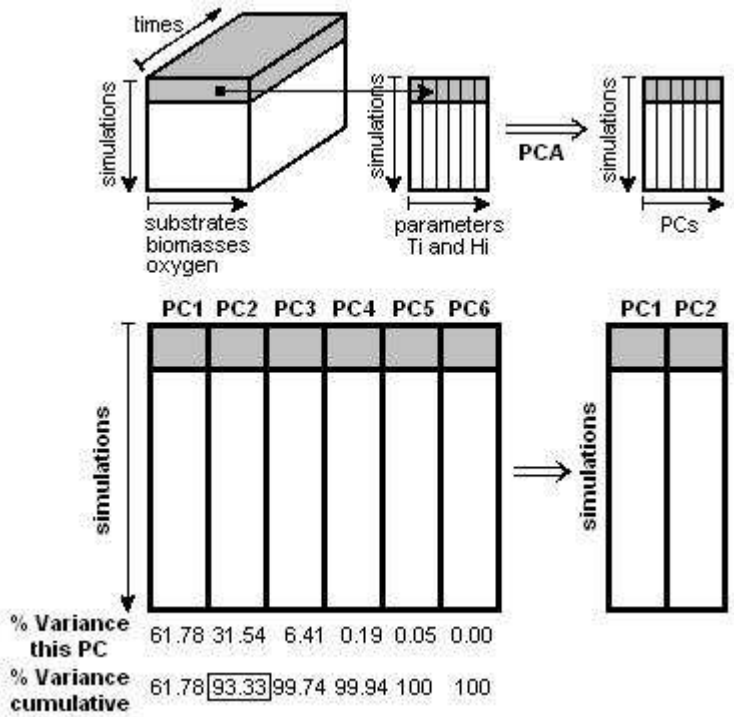

Figure 3 : Reduction of the data matrix by a PCA

Table 1 : Table for fault causes definition

	Kla	COD_{in}	$NH4_{in}$	X_H	X_A
<<	0	0	0	0	0
<	0	0	0	0	0
>	0	0	0	0	0
>>	0	0	0	0	0