Jul 1st, 12:00 AM

Trajectories’ Mining Between Subprocesses in a Wastewater Treatment Plant

Karina Gibert
G. Rodriguez Silva
Ignasi Rodríguez Roda

Follow this and additional works at: http://scholarsarchive.byu.edu/iemssconference

Gibert, Karina; Silva, G. Rodriguez; and Roda, Ignasi Rodriguez, "Trajectories’ Mining Between Subprocesses in a Wastewater Treatment Plant" (2008). International Congress on Environmental Modelling and Software. 90.
http://scholarsarchive.byu.edu/iemssconference/2008/all/90

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.
Trajectories’ Mining Between Subprocesses in a Wastewater Treatment Plant

K. Gibert a, G. Rodríguez Silva a and I. Rodríguez Roda b

a Universitat Politècnica de Catalunya, Campus Nord Ed.C5, Jordi Girona 1-3, Barcelona (karina.gibert@upc.edu, gustavo.rodriguez-silva@upc.edu)

b Universitat de Girona, Campus Montilivi, Girona (ignasi.rodriguezroda@udg.edu)

Abstract: In this work the Trajectories’ Mining between subprocesses is presented in regards to the dynamics of a Wastewater Treatment Plant (WWTP). This kind of Environmental Systems involves a high complexity inherent to its own characteristics. Intelligent Environmental Decision Support Systems (IEDSS) can improve the decision making process, assisting decision makers in the evaluation of alternatives and improving management and control of Environmental Systems in general and for this particular application of WWTP. Our line of work is based on the development of methodologies of Artificial Intelligence and Statistics to solve problems of Knowledge Discovery of Data as is shown by Fayyad [1996]. In this work, knowledge extraction is approached with a methodology named Clustering Based on Rules by States (CIBRxE) formally presented by Gibert and Rodríguez [2007]. Basically, consists of analyzing a process that can be divided in $S = \{e_1, \ldots, e_E\}$ states or subprocesses. After dividing variables based on the subprocesses to which they refer, knowledge discovered is integrated from each subprocess into a unique model of global operation of the phenomenon. Once the data of the WWTP has been analyzed, it was empirically verified that probability of a transition between two consecutive states depends on not only on the previous state but on the whole sequence of previous states.

Keywords: clustering; rules; states; trajectories; wastewater

1 INTRODUCTION

To correctly treat wastewater different operations and unique processes are required. A mixture of physics, chemical and biological agents is needed to form the diagram of the process of each wastewater station. The global process always follows a logical sequence of treatment divided in different stages that can varied according to the structure and objectives of the plant as is shown in Metcalf and Eddy [2003]. The Data presented in this work has been previously clustered and interpreted by Gibert and Roda [2000]. In our current work knowledge discovery is approached with Clustering Based on Rules by States (CIBRxE). Once the data has been analyzed, probability of transition between states is verified to model global operation of the phenomenon.

2 APPLICATION DOMAIN AND PREVIOUS WORK

A sample of 396 observations taken from September the first of 1995 to September the 30th of 1996 from a Catalan WWTP is used. The plant is described daily with measures taken in the following stages of the depuration process: Input(E), Primary Settler(D), Bioreactor (B) and Output (S). In the database 4 different stages of the depuration process are identified. A selection of 25 variables was considered the most relevant by the opinion of an expert and indicated which variables correspond to each stage of the depuration process as is shown in table 1. This data has been previously clustered using Clustering Based on Rules (CIBR) with a knowledge base that collects the legal limits of certain physics and biological parameters that classify the quality of wastewater at the plant’s exit. The rules base $\mathcal{R}$ is in figure 1, Right.
Table 1: Variables used in the Clustering.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>C192</td>
<td>Hydrogen Potential (pH-E)</td>
</tr>
<tr>
<td></td>
<td>Suspended Solids (SS-E)</td>
</tr>
<tr>
<td></td>
<td>Volatile Suspended Solids (SSV-E)</td>
</tr>
<tr>
<td></td>
<td>Chemical Organic Matter (DQO-E)</td>
</tr>
<tr>
<td></td>
<td>Biodegradable Organic Matter (DBO-E)</td>
</tr>
<tr>
<td>C392</td>
<td>Hydrogen Potential (pH-E)</td>
</tr>
<tr>
<td></td>
<td>Suspended Solids (SS-E)</td>
</tr>
<tr>
<td></td>
<td>Volatile Suspended Solids (SSV-E)</td>
</tr>
<tr>
<td></td>
<td>Chemical Organic Matter (DQO-E)</td>
</tr>
<tr>
<td></td>
<td>Biodegradable Organic Matter (DBO-E)</td>
</tr>
<tr>
<td>C383</td>
<td>Hydrogen Potential (pH-E)</td>
</tr>
<tr>
<td></td>
<td>Suspended Solids (SS-E)</td>
</tr>
<tr>
<td></td>
<td>Volatile Suspended Solids (SSV-E)</td>
</tr>
<tr>
<td></td>
<td>Chemical Organic Matter (DQO-E)</td>
</tr>
<tr>
<td></td>
<td>Biodegradable Organic Matter (DBO-E)</td>
</tr>
<tr>
<td>C389</td>
<td>Hydrogen Potential (pH-E)</td>
</tr>
<tr>
<td></td>
<td>Suspended Solids (SS-E)</td>
</tr>
<tr>
<td></td>
<td>Volatile Suspended Solids (SSV-E)</td>
</tr>
<tr>
<td></td>
<td>Chemical Organic Matter (DQO-E)</td>
</tr>
<tr>
<td></td>
<td>Biodegradable Organic Matter (DBO-E)</td>
</tr>
<tr>
<td>C390</td>
<td>Hydrogen Potential (pH-E)</td>
</tr>
<tr>
<td></td>
<td>Suspended Solids (SS-E)</td>
</tr>
<tr>
<td></td>
<td>Volatile Suspended Solids (SSV-E)</td>
</tr>
<tr>
<td></td>
<td>Chemical Organic Matter (DQO-E)</td>
</tr>
<tr>
<td></td>
<td>Biodegradable Organic Matter (DBO-E)</td>
</tr>
</tbody>
</table>

ClIBR is implemented on KLASS software as is shown by Gibert and Roda [2000], this can take advantage of a prior knowledge base to bias classes construction to improve interpretability. Details on this technique and comparison with other clustering techniques are shown in Gibert [1996]. As usual in hierarchical clustering, the final partition is the horizontal cut of the dendrogram $\tau_{En,G}$ that maximizes the ratio between heterogeneity within classes and respect to homogeneity between classes, what guarantees the distinguishability between classes. Following this classical criterion and validated by the experts, a partition in four clusters was performed $P_4 = \{c383, c389, c390, c392\}$. Experts of this environmental process associate concepts to each resulting class, where statistics from each class and Class Panel Graph, figure 2 helped us in the interpretation as is shown by Gibert and Roda [2000].

Class C192: This class contains many variables with specific values of this class, most of them at low values. Verifying the corresponding dates assigned to C192, allowed the experts identify this class as the storm days. To face the storm, the decision is to protect the system minimizing inflow (Q-E), closing input valves and maintaining bioreactor microorganism’s trough minimizing purged flow (QP-G). This produces an increment of biomass in the reactor and justifies the high values of (MCRT-B).

Class C383: This class has input inflow, recirculated inflow and purge inflow at non-low values, cellular age is not high, on the contrary of C392. These days the plant works very well even reduces ammonium (NH4-D y NKT-D lows) in spite of not being a plant specifically designed for this. In addition, the reactor is working a full yield (QB-B high) and the degradation is good (MCRT-B low), as a result the water is so clean. The experts identify this class with a profile of excellent working of the plant.

Class C389: In this class, the input inflow, bioreactor inflow, recirculated inflow and purge inflow are lows, cellular age is non-high, ammonium reduction is lower, output flow is no so clean (DBO-S non-low). Contains days where water is very dirty (SS-E, SSV-E, DQO-E and DBO-E highs) as in organic material as in suspended solids. According to the experts exists a shock of load (organic material) of solids in the input of the process, as indicates the volatile suspended solids inside bioreactor are very lows (MLVSS-B). In addition, the reactor is working a full yield (QB-B high) and the degradation is good (MCRT-B low), as a result the water is so clean. The experts identify this class with a profile of excellent working of the plant.

Class C390: The assigned days in this class present some punctual problem, the inflow is not so dirty (as indicated by the low or medium values of SS-E, SSV-E, DQO-E and DBO-E), but suspended solids in biological reactor are higher than expected (MLVSS-B), indicating a non-optimal depuration.

\[
\mathcal{R} = \{r_1 : \text{If } (SS - S) > 20 \& (DBO-S) < 35 \rightarrow P, r_2 : \text{If } (SS - S) < 20 \& (DBO-S) > 35 \rightarrow Q\}
\]

S: states for (abnormal operation in general)
P: states for (failure in suspended solids treatment)
Q: states for (failure in organic matter treatment)
3 CLUSTERING BASED ON RULES BY STATES (CLBRxE)

Given an environmental domain in which a process is taking place in such a way that it can be divided in $S = \{e_1 \ldots e_E\}$ states or subprocesses, with $I = \{i_1 \ldots i_n\}$ observations described by $X_1 \ldots X_K$ variables and given an a priori knowledge base $R$, which can be partial, containing logic rules as described by Gibert [1996], our proposal is:

1. Calculate a rules-induced partition on $I$ by evaluating $R$ over $X_1 \ldots X_K$ ($P_R$).
2. Divide the variables based on the subprocess to which they refer (let $X_{1e}^i, \ldots, X_{Ke}^i$ be the set of variables referring to the subprocess $e$).
3. For each subprocess $S = \{e_1 \ldots e_E\}$:
   a. Do $IDe = \{i \in I : x_{1e}^i = x_{2e}^i = \ldots = x_{Ke}^i = *\}$, where * indicates missing value.
   b. Do a CBR, as described by Gibert [1996], over $I \setminus IDe$ with variables referring to state $e$ $X_{1e}^i, \ldots, X_{Ke}^i$ but using $P_R$ as a rules-induced partition for all the states of the process. Elements of $ID e$ can not be used in state $e$ since they miss useful information.
   c. Analyze the resulting dendrogram $\tau_e$, perform a horizontal cut to obtain $P_e^+$ and associate concepts to each resulting class $C \in P_e^+$. 
   d. Construct $P_e = P_e^+ \cup \{ID e\}$
4. Construct cross tables $P_e \times P_{e+1}$, applying required delay, if any, between observations and analyze most probable trajectories of individual through classes of sequential states. Crossing can be done at different stages of the process by considering that the individual that is a certain stage delay some time till the next stage. Thus, cell $C, C'$ of the cross table contains the number of elements that $x_{(i-d)Ke} = C$ and $x_{Ke+1} = C'$ instead of $x_{Ke} = C \land x_{Ke+1} = C'$ where $d$ is the delay of the individual from state $e$ to $e + 1$. Delays between steps are to be provided by experts. This information allows wider analysis of the dynamics of the process and it is being studied if this approach can give richer models than the instantaneous conjoint global analysis of all phases performed previously.
5. Identify trajectories related to informations flow between the $S = \{e_1 \ldots e_E\}$ subprocess.
6. Select the most typical trajectories and interpret.
4 KNOWLEDGE DISCOVERY IN WWTP WITH CLBRxE

4.1 ClBR Local to Each State In our methodology $R$ is evaluated on $I$ and a partition is induced $P_R = \{S, P, Q, Residual\}$. $S$ contains 11 elements of $I$ that satisfy $r_1$, $P$ contains 40 elements of $I$ that satisfy $r_2$, $Q$ contains 10 elements of $I$ that satisfy $r_3$ and Residual contains 335 elements of $I$ that do not satisfy any rule or many at the same time which are contradictory. Each stage is treated separately then the relationship between subprocesses is analyzed. The number of classes is identified a posteriori for each subprocess upon the resulting dendrogram.

Input Subprocess. 7 variables are available: Q-E, FE-E, PH-E, SS-E, SSV-E, DQO-E, DBO-E. All prototypes will be calculated for each rules-induced partition in the previous step. We will manually construct an extended residual class that includes those prototypes in addition to residual objects. Finally, a clustering without rules of the extended residual class is performed with the following criteria: Normalized Euclidean Squared Metric, Linkage Wards Criteria, since all the variables are numerical in this application, which is equivalent to perform a ClBR of the Input variables using a knowledge base that includes variables from another state. This is a technical solution to be used in the meanwhile, statistical software KLASS is modified to include this case. The tree $\tau_E$ of clustering process is shown in figure.

Settler Subprocess. 5 variables are available: PH-D, SS-D, SSV-D, DQO-D, DBO-D. As in the Input stage, in Settler is proceed on the same way. The tree $\tau_D$ is the partitioning in the followings.

Bioreactor Subprocess. 8 variables are available: QB-B, QR-G, QP-G, QA-G, V30-B, MLSS-B, MLVSS-B, MCRT-B. As in the Input stage, in Bioreactor is proceed on the same way. The tree $\tau_B$ is the clustering process is shown in figure 3.

Output Subprocess. 5 variables are available: PH-S, SS-S, SSV-S, DQO-S, DBO-S. The rules provided by the expert affected the variables that are used in this stage, therefore we directly realized ClBR with the following criteria: Normalized Euclidean Squared Metric, Linkage Wards Criteria. The tree $\tau_S$ is the clustering process is shown in figure 3.

Analyzing dendrograms $\tau_E$, $\tau_D$, $\tau_B$ and $\tau_S$, the corresponding partitions of 5 classes and Class Panel Graphs were performed, see figure 4. With these elements some concepts can be related to each class as are shown in the right side of figure 3.

4.2 Trajectories Analysis Once the typical situations in each stage are identified, we proceed to study the transition of the wastewater from one stage to the next. The starting point will be the crossing between the partitions of two consecutive stages. On this specific application daily averages were used while wastewater keeps only few hours at each stage. According to this is not required a delay between stages before performing the crossing. To study transitions between states, frequencies of one stage conditioned by the previous stage are shown on table 2, for example $P^S \mid P^B$ shows that poorest quality water hardly ever comes out of the plant when the Bioreactor is functioning correctly ($P_{C324\_0.360\_2} = 0.0632$). In figure 5, Transition Diagram derived from tables of conditional frequencies of subsequent states by ($P^D \mid P^E$), ($P^B \mid P^D$) and ($P^S \mid P^B$) is built. Classes identified for each subprocess are aligned from highest quality water at the top to most polluted water at the bottom. Transition probability is visualized with the thickness of arrows. Transitions between two consecutive states, for instance, from Input Subprocess (E) to Settler Subprocess can be arranged in a transition matrix, which provides transition probabilities between any pair of states. For example, $P_{ED}$ provides probabilities from Input Subprocess to Settler Subprocess. The sequence of classes of wastewater through the states of the process is denoted as trajectory: $T : \{C_E, C_D, C_B, C_S\}$, where $C_E \in P_E$, $C_D \in P_D$, $C_B \in P_B$ and $C_S \in P_S$. For instance, for $T_1 : \{C_{331\_E}, C_{328\_D}, C_{324\_B}, C_{375\_S}\}$ water flows from $C_{331\_E}$ at the Input to $C_{328\_D}$ at the Primary Settler to $C_{324\_B}$ at the Bioreactor and $C_{375\_S}$ at the Output, what means that wastewater has normal levels of pollutants at the input, still keeps average levels of solids in the primary settler, underload wastewater with average waste purged flow rate in the bioreactor and exits with average quality. The 12-IX-1995 was one of the days that wastewater followed this path. Finally, a probability should be associated to every trajectory to identify the most typical patterns and caracteristical dynamics observed in WWTP.
4.3 Trajectories Probability Induction  Association of probability to the trajectories can be approached from different perspectives and in this paper two of them are presented.

**Markov Process Assumption:** The probability of a given trajectory would be the product of corresponding transition probabilities if wastewater treatment was a Markov process and satisfies the property “for any given time instant (say \( t_n \)) the future behaviour, for instance the value of \( X_{t_n+1} \), is totally independent of its history, i.e., the values of \( X_{t_{n-1}}, X_{t_{n-2}} \) and so on. It only depends on the state occupied at the current time instant \( t_n \), given by the value of \( X_{t_n} \).”, Holger [2002]. Consider trajectory \( T : (C_E, C_D, C_B, C_S) \). Assuming that a WWTP is a Markov Process:

\[
p(T) = P_E \cdot P_{ED} \cdot P_{DB} \cdot P_{BS} \tag{1}
\]

**Frequentist Non Parametric Assumption:** In this approach, probabilities at a given stage will be conditioned by the whole path previously followed by the water. \( P(T) \) is then calculated as:

\[
p'(T) = p(C_E) \cdot p(C_D|C_E) \cdot p(C_B|C_D \land C_E) \cdot p(C_S|C_B \land C_D \land C_E) \tag{2}
\]

This is equivalent to direct frequentist estimation of the trajectories along the whole sample.

As the WWTP process really was a Markov process and satisfies the non-memory property announced above, both approaches would produce same results. However, in this application it is empirically verified that WWTP is not a Markov process and results differ on both approaches. As an example consider \( T_1 : (C_{331E}, C_{328D}, C_{324B}, C_{375S}) \)
**Figure 4: Class Panel Graph of \( \mathcal{P}^E \), \( \mathcal{P}^D \), \( \mathcal{P}^B \) and \( \mathcal{P}^S \)**

**Figure 5: Transition Diagram between states**
Table 2: Frequencies of state conditioned by \( P^D \mid P^E, P^B \mid P^D \) and \( P^S \mid P^B \).

<table>
<thead>
<tr>
<th>( P^D )</th>
<th>( C_{327} )</th>
<th>( C_{312} )</th>
<th>( C_{328} )</th>
<th>( C_{329} )</th>
<th>( d_2 )</th>
<th>tot.</th>
<th>useful</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P^D )</td>
<td>( C_{326} )</td>
<td>( C_{312} )</td>
<td>( C_{328} )</td>
<td>( C_{329} )</td>
<td>( d_3 )</td>
<td>tot.</td>
<td>useful</td>
</tr>
<tr>
<td>( 0.129 )</td>
<td>0.087</td>
<td>0.048</td>
<td>0.037</td>
<td>0.150</td>
<td>1</td>
<td>61</td>
<td>0.1333</td>
</tr>
<tr>
<td>( 0.037 )</td>
<td>0.096</td>
<td>0.014</td>
<td>0.067</td>
<td>0.100</td>
<td>1</td>
<td>44</td>
<td>0.1111</td>
</tr>
<tr>
<td>( 0.0654 )</td>
<td>0.037</td>
<td>0.096</td>
<td>0.014</td>
<td>0.100</td>
<td>1</td>
<td>44</td>
<td>0.1111</td>
</tr>
<tr>
<td>( 0.0635 )</td>
<td>0.096</td>
<td>0.014</td>
<td>0.067</td>
<td>0.100</td>
<td>1</td>
<td>44</td>
<td>0.1111</td>
</tr>
<tr>
<td>( 0.0667 )</td>
<td>0.096</td>
<td>0.014</td>
<td>0.067</td>
<td>0.100</td>
<td>1</td>
<td>44</td>
<td>0.1111</td>
</tr>
</tbody>
</table>

### 4.4 Typical Trajectories

In the presented case study, 139 different trajectories are observed (from the 1296 possibilities of different trajectories). The frequencies of those 139 trajectories are shown in figure 6, and for this particular application trajectories with \( p > 0.025 \) are selected and interpreted by the experts.
normal levels of pollutants at the input, high levels of SS-D and DBO-D in the primary settler, underload WW with average waste purged flow rate and exits with average quality

underload WW and high inflow rate at the input, high levels of SS-D and DQO-D in the primary settler, underload WW with low aeration in the bioreactor and exits with high quality and high pH-S

underload WW and high inflow rate, high levels of SS-D and DQO-D in the primary settler, underload WW with average waste purged flow rate and exits with poorest quality

underload WW and high inflow rate, average levels of solids in the primary settler, underload WW with average waste purged flow rate and exits with high quality

underload WW and high inflow rate, average levels of solids in the primary settler, underload WW with high aeration in the bioreactor and exits with high quality and high pH-S

underload WW and high inflow rate, average levels of solids in the primary settler, underload WW with high aeration in the bioreactor and exits with average quality

underflow WW at the input, average levels of solids in the primary settler, underload WW with low aeration in the bioreactor and exits with high quality and high pH-S

5 CONCLUSIONS

In this work CIBRxE was used to identify typical situations in the states of the WWTP. From the results, a transition diagram displays the more frequent transitions along the process allowing to face the analysis of trajectories as a whole. To identify the more typical trajectories, probabilities associated to each one is required. From this work it has been shown that Markov process is not a valid model for WWTP and direct frequentist estimation is more suitable for this purpose. So identification of more typical trajectories has been done using this criteria. The more probable trajectories with probability between 0.025 and 0.1136 have been selected identifying different processes in the WWTP. Currently, a methodology for automatically produce the interpretation of selected trajectories is being developed.

ACKNOWLEDGMENTS

This research has been partially financed by TIN 2004-01368.

REFERENCES

Fayyad, U. e a. Advances in Knowledge Discovery and Data Mining, chapter From Data Mining to Knowledge Discovery: An overview. AAAI/MIT Press, 1996.


Gibert, K. and I. Roda. Identifying characteristic situations in wastewater treatment plants. In Workshop in Binding Environmental Sciences and Artificial Intelligence, volume 1, pages 1–9, August 2000. ECAI.

