Identification of More Characteristic Dynamic Patterns in a WWTP by ClBR×E

Karina Gibert and Gustavo Rodríguez Silva

Technical University of Catalonia, Department of Statistics and Operations Research, c. Jordi Girona 3-1, 08034 Barcelona, Spain
{karina.gibert,gustavo.rodriguez-silva}@upc.edu

Abstract. In this work, advances in the design of an hybrid methodology that combines tools of Artificial Intelligence and Statistics to extract a model of explicit knowledge are presented in regards to the dynamics of a Wastewater Treatment Plant. Our line of work is based in the development of methodologies of AI & Stats to solve problems of Knowledge Discovery of Data where an integral vision of the pre-process, the automatic interpretation of results and the explicit production of knowledge play a role as important as the analysis itself. In our current work the identification of more characteristic dynamic patterns is approached with Clustering Based on Rules by States, which consists in the analysis of the stages that the water treatment moves through, to integrate the knowledge discovered from each subprocess into a unique model of global operation of the phenomenon.

Keywords: clustering, rules, dynamics, states, wastewater.

1 Introduction

This research focus the qualitative analysis of the dynamics of a Wastewater Treatment Plant *(WWTP)* in order to identify typical patterns in the treatment process which could support decision making. Knowledge Discovery of Data *(KDD)* techniques are a good framework to face this goal allowing to extract the knowledge from data. Application of *KDD* techniques has been used in every component of the modeled systems [1]. Automatization and semi-automatization of the knowledge extraction can improve the decision-making process in an Intelligent Environmental Decision Support System *(IEDSS)* reducing the decision time and assisting decision makers in the evaluation of alternatives. The automatic interpretation of results and the explicit production of knowledge play a role as important as the analysis itself [2]. Analysis of characteristic situations in a *WWTP* was done in qualitative terms by global cluster with all variables in previous works, see [3] for details. This approach just permits to identify situations in which the plant may be at specific moment but does not permit to study the evolution of the wastewater treatment process itself, for this reason, the steps to be followed are focused to include dynamics in the decision making support. Dynamics of Wastewater treatment is very difficult to model by

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 372–380, 2008.

classical methods. In this work, an hybrid tool of AI and Stats is applied as an alternative for knowledge discovery of a *WWTP*. Clustering Based on Rules by States *(ClBR×E)* is designed in such a way that prior knowledge of the experts is integrated into the analysis to improve the quality of the clustering processes. Analysis is performed step by step and final integration permits construction of models for dynamics behavior of the plant in qualitative terms.

2 Data Presentation and Previous Work

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 [4]. The main goal of a *WWTP* is 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., When the plant is not on normal operation, which is extremely difficult to model by traditional mechanicistic models [4], decisions have to be taken to modify some parameters of the wastewater treatment process in order to reestablish the normality as soon as possible. This process is very complex, 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.

In this work a sample of 396 observations taken from September the first of 1995 to September the 30th of 1996 from a Catalan *WWTP* are used, corresponding to 40 variables for each daily observation with missing values in some of them [3]. Each observation refers to a daily mean and it is identified by the date itself. The plant is described daily with measures taken in the following stages of the depuration process: Input(E), Settler(D), Biologic Treatment (B) and Output (S). According to this, in the database we identified 4 different stages of the depuration process. Table 1 represents a selection of 25 variables considered the most relevant by the opinion of an expert and indicated which variables correspond to each stage of the depuration process. Some other variables are available like ammonium concentration but experts recommended not to include them to find clusters according to their background knowledge. To interpret and conceptualize all variables are considered. This data has been previously clustered globally using Clustering Based on Rules *(ClBR)* 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, see [5] for details on this technique and for comparison with other clustering techniques. A partition in four clusters was performed $P_4 = \{c383, c389, c390, c392\}$. It was seen that classes identified different typical situations in the plant [3]. This interpretation was supported by the rules-induced with *CCCE* methodology, that allows to automatically identify one class from the others [6]. After careful analysis of results experts could conceptualize classes as:

Table 1. Variables used in the Clustering

VARIABLES	"E" (Input)	"D" (Settler)	"B" (Bioreactor)	"S'" (Output)
Inflow(Q)	Q-E		QB-B	
Iron Pre-treatment(FE)	FE-E			
Hydrogen Potential(PH)	PH-E	PH-D		PH-S
Suspended Solids(SS)	SS-E	SS-D		SS-S
Volatile Suspended Solids(SSV)	SSV-E	SSV-D		SSV-S
Chemical Organic Matter(DQO)	DQO-E	DQO-D		DQO-S
Biodegradable Organic Matter(DBO)	DBO-E	DBO-D		DBO-S
Index 30 at the Biological Reactor(V30)			V30-B	
Recirculated Inflow (QR)			QR-G	
Purge Inflow (QP)			QP-G	
Aeration (QA)			QA-G	
Mixed Liquor Suspended Solids(MLSS)			MLSS-B	
Mixed Liquor Volatile Suspended Solids(MLVSS)			MLVSS-B	
Mean Cell Residence Time(MCRT)			MCRT-B	

Class C_{392}: underflow Q-E.
Class C_{389}: High input flow rate, low values of Biodegradable Organic Matter at Settler
Class C_{383}: High input flow rate, high values of Biodegradable Organic Matter at Settler and low values of Biodegradable Organic Matter at the Input.
Class C_{390}: Average-High flow rate and high values of Biodegradable Organic Matter at the Input and high values of Biodegradable Organic Matter at Settler.

In this work, the same data base is analyzed with the methodology $ClBR{\times}E$ which use the same prior knowledge to bias classes construction at every stage.

3 Clustering Based on Rules by States (CLBR×E)

Given an environmental domain in which a process is taking place in such a way that it can be divided in $\mathcal{S} = \{e_1, \ldots, e_E\}$ states or subprocesses, with $\mathcal{I} = \{i_1, \ldots, i_n\}$ observations described by $X_1, \ldots, X_K$ variables and a knowledge base $\mathcal{R}$, containing logic rules as described in [5], our proposal is:

1. *Phase of initial Knowledge Base construction*:
   The main idea is to allow the expert to introduce the prior knowledge on the formation of classes. This allow to introduce constraints on the formation of classes in a declarative way determining a first set of logic rules $\mathcal{R}^0$ containing part of the expert knowledge on the studied domain.

- Step ξ ($\xi = 1$): Start iteration process

2. *Phase of Background Knowledge acquisition*:
   (a) Determine the rule-induced partition $\mathcal{P}^{\xi}_{\mathcal{R}}$ on $\mathcal{I}$ by evaluating $\mathcal{R}$ over $X_1, \ldots, X_K$. Include a residual class $\mathcal{C}^{\xi}_0$ on $\mathcal{P}^{\xi}_{\mathcal{R}}$ with those objects for which no information is provided.
   (b) Conflict solving phase. Analyze objects of $\mathcal{C}^{\xi}_0$ selected by rules.
      (1) If satisfactory, proceed to the next step.
      (2) Otherwise, return to the background knowledge of acquisition phase and reformulate $\mathcal{R}^{\xi}$.
3. *Phase of analysis by states*:
   (a) Divide variables upon the subprocess to which they refer (let e be a subprocess and $X^e_1, \ldots, X^e_{Ke}$ be variables referring to the subprocess e).

- Select $e = e_1$:
  (b) Build $\mathcal{I}De = \{i \in \mathcal{I} : x_{i1}^e = x_{i2}^e = \ldots = x_{iKe}^e = *\}$
  (c) Perform a *ClBR* over $\mathcal{I} \setminus \mathcal{I}De$ with variables $X_1^e, \ldots, X_{1Ke}^e$ referring to state e but using $\mathcal{P}_{\mathcal{R}}^{e^\xi}$ as a rules-induced partition:
    (1) Clustering within expert constraints: $\mathcal{P}_{\mathcal{R}}^{e^\xi}$ will satisfy the expert requirements. Perform the clustering for each $\mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{e^\xi}$. Notice that every $\mathcal{C} \in \mathcal{I} \setminus \mathcal{I}De$ makes cheaper the construction of classes. Elements of $\mathcal{I}De$ cannot be used in state e since they do not provide any useful information and would generate missing values in the distance matrix. Determine:
      (*i*) The corresponding hierarchical trees $\tau_{\mathcal{C}}^{e^\xi}$,
      (*ii*) Their prototypes $\overline{i}_{\mathcal{C}}^{e^\xi}$, by summarizing the class,
      (*iii*) Their masses $m_{\mathcal{C}}^{e^\xi} = card\, \mathcal{C}$ and
      (*iv*) Their indexes of level $h_{\mathcal{C}}^{e^\xi}$
    (2) Extend the residual class: Add the prototypes $\overline{i}_{\mathcal{C}}^{e^\xi}$ to the residual class, as if they were ordinary objects, but taking into account their masses. The new data set is: $\mathcal{I}^{e^\xi} = \left\{ (\overline{i}_{\mathcal{C}}^{e^\xi}, m_{\mathcal{C}}^{e^\xi}) : \mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{e^\xi} \right\} \cup \left\{ (i, 1) : i \in \mathcal{C}_0^{e^\xi} \right\}$
    (3) Do the integration: Classify $\tilde{\mathcal{I}}^{e^\xi}$ to integrate all the trees $\tau_{\mathcal{C}}^{e^\xi}$, $(\mathcal{C} \in \mathcal{P}_{\mathcal{R}}^{e^\xi})$ in the sole τ^{e^ξ}.
  (d) Local interpretation phase:
    (1) Determine the final number of classes: Analyze the resulting dendrogram τ^e to obtain the best horizontal cut, using heuristic criteria or automatic tools [6]. Construct the cut of τ^e identifies a data partition in a set of classes, $\mathcal{P}_*^{e^\xi}$.
    (2) Between the k best cuts, choose the one to allows a better interpretation and associate concepts to each resulting class $\mathcal{C} \in \mathcal{P}_*^{e^\xi}$.
    (3) Interpretation: Use Class Panel Graph to associate conceptual labels to the classes.
    (4) Build $\mathcal{P}^{e^\xi} = \mathcal{P}_*^{e^\xi} \cup \{\mathcal{I}De\}$
  (e) Set $e = e + 1$ and repeat.

4. *Global Evaluation phase*:
   The expert has to confirm that partitions $\mathcal{P}^{e_1{}^\xi}, \mathcal{P}^{e_2{}^\xi}, \ldots, \mathcal{P}^{e_E{}^\xi}$ obtained with $\mathcal{R}^\xi$ improve the partitions $\mathcal{P}^{e^{\xi-1}}$ obtained with $\mathcal{R}^{\xi-1}$ in the desired way. Tables for comparing different classifications, or terms with major contributions to the distance between them can be used. In addition, is possible to test the significance of the differences using a non-parametric test (δ-test) particularly designed [6]. This step could produce the termination criteria:
   (a) If the improvement is not significant, stop the iteration and assume previous iteration results as the best.
   (b) If not, analyze the results to reformulate the rules set. Build $\mathcal{R}^{\xi+1}$, set $(\xi = \xi + 1)$ and repeat.
5. *Analysis of Trajectories Phase*:
   (a) Apply the MPT (Most Probable Trajectories) Algorithm:
     (1) Build variables $E_1, \ldots, E_E$, where: $E_e = \xi$, if $\mathcal{P}^e = \mathcal{C}_\xi^e$ and $E_e = 999$, if $\mathcal{P}^e = \mathcal{C}_{de}^e$
     (2) Build a variable "T_{cod}" as $T_{cod} = \sum_{e=1}^{E} E_e 10^{4(e-1)}$

(3) Build a Sorted Frequencies Table of the variable T_{cod} with their relative and absolute frequencies.
(4) Build a variable T to identify trajectories with frequency $> \gamma$:
$T = T_{cod}$, if relative frequencies of $T_{cod} > \gamma$ and $T = T_{others}$, otherwise

(b) Build a Trajectories Diagram:
Build a Colored Labeled Transition Diagram as a display of the trajectories selected by the MPT Algorithm ($T_{cod} \neq others$) with:
(1) E columns of nodes. Column e displays the elements of $\mathcal{P}^e$ labeled according to Class Panel Graph of step 3.
(2) Link the nodes of every trajectory τ with arrows of the same color.
This displays the more characteristic trajectories in the process together with their conceptualization.

4 Knowledge Discovery in WWTP with ClBRxE

In our methodology the following set of rules is used:

$\mathcal{R} = \{r_1 : \text{If}(SS - S) > 20(DBO - S) > 35 \longrightarrow S(\text{abnormal operation in general})$
$r_2 : \text{If}(SS - S) > 20(DBO - S) < 35 \longrightarrow P(\text{failure in suspended solids treatment})$
$r_3 : \text{If}(SS - S) < 20(DBO - S) > 35 \longrightarrow Q(\text{failure in organic matter treatment})\}$

As $\mathcal{R}$ is evaluated on $\mathcal{I}$, a partition is induced $\mathcal{P}_{\mathcal{R}} = \{S, P, Q, Residual\}$ where S contains 11 elements of $\mathcal{I}$ that satisfy r_1, P contains 40 elements of $\mathcal{I}$ that satisfy r_2, Q contains 10 elements of $\mathcal{I}$ that satisfy r_3 and Residual contains 335 elements of $\mathcal{I}$ that do not satisfy any rule or do simultaneous by satisfying many with contradictory right parts which are contradictory. Each stage of WW Treatment is treated separately (by repeated use of ClBR), and then the relationship between subprocesses is analyzed: It is important here to remark that number of classes is identified a posteriori for each subprocess upon the resulting hierarchical dendrogram and considering classical criteria of maximizing homogeneity intraclasses and heterogeneity between classes, the local results are:

Input Subprocess. 7 variables are available: Q-E,FE-E,PH-E,SS-E,SSV-E,DQO-E,DBO-E. *ClBR* is performed with Normalized Euclidean Squared Metric, Linkage Wards Criteria. The hierarchical dendrogram τ^E is analyzed and a partition of 5 classes $\mathcal{P}^E$ is performed. Some concepts are related to each class to $\mathcal{P}^E$. As in the Input stage, in Settler and Bioreactor is proceed on the same way. Class Panel Graphs are shown in figure 1.

$\mathcal{C}^E_{329}$ low addition of FE-E [(FE-E)Low].
$\mathcal{C}^E_{318}$ underflow WW(Q-E) [(Q-E)Average].
$\mathcal{C}^E_{328}$ underload WW and high inflow rate [(Q-E)High], [(SS-E)(SSV-E)(DQO-E)(DBO-E)Low].
$\mathcal{C}^E_{331}$ WW at normal level [(DQO-E)(DBO-E)(SS-E)(SSV-E)Average].
$\mathcal{C}^E_{330}$ overload WW and high inflow rate [(Q-E)(DQO-E)(DBO-E)(SS-E)(SSV-E)High].
d_1 unknown.

Settler Subprocess. 5 variables are available: PH-D, SS-D, SSV-D, DQO-D, DBO-D. $\mathcal{P}^D$:

$\mathcal{C}^D_{327}$ low level of SS-D [(SS-D)Low].
$\mathcal{C}^D_{312}$ low pH-D [(pH-D)Low].

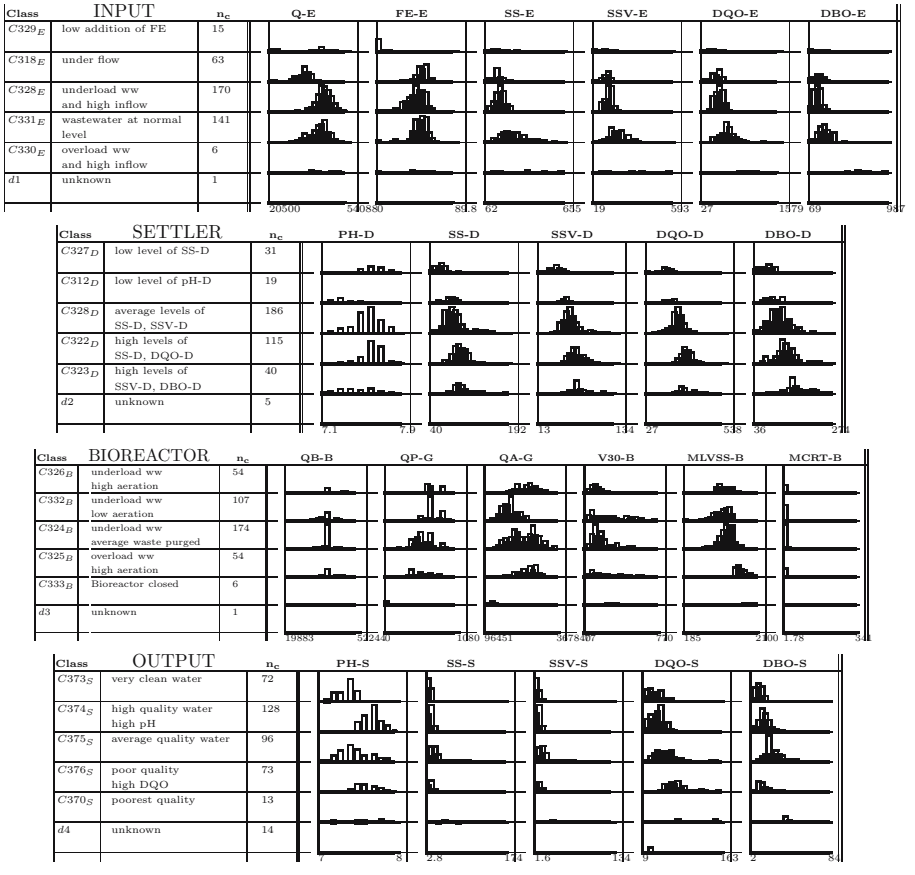

Fig. 1. Class Panel Graph of $\mathcal{P}^E$, $\mathcal{P}^D$, $\mathcal{P}^B$ and $\mathcal{P}^S$

$\mathcal{C}^D_{328}$ average levels of SS-D, SSV-D [(SS-D)(SSV-D)Average].
$\mathcal{C}^D_{322}$ high levels of SS-D, DQO-D [(SS-D)(DQO-D)High].
$\mathcal{C}^D_{323}$ high levels of SSV-D, DBO-D [(SSV-D)(DBO-D)High].
d_2 unknown.

Bioreactor Subprocess. 8 variables are available: QB-B, QR-G, QP-G, QA-G,V30-B, MLSS-B, MLVSS-B, MCRT-B. $\mathcal{P}^B$:
$\mathcal{C}^B_{326}$ underload WW, high aeration [(QA-G)High], [(MLSS-B)Low].
$\mathcal{C}^B_{332}$ underload WW, low aeration [(QA-G)Average], [(MLSS-B)Low].
$\mathcal{C}^B_{324}$ underload WW, average waste purged [(QP-G)Average], [(MLSS-B)Low].
$\mathcal{C}^B_{325}$ overload WW, high aeration [(QA-G)(MLSS-B)High].
$\mathcal{C}^B_{333}$ Bior. closed [(QB-B)(QR-G)(QP-G)(QA-G)Low], [(MLSS-B)(MLVSS-B)(MCRT-B)High].
d_3 unknown.

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 *ClBR* is realized directly with the following criteria: Normalized Euclidean Squared Metric, Linkage Wards Criteria. Some concepts could be related to each class of $\mathcal{P}^S$:

$\mathcal{C}^S_{373}$ high quality water [(DQO-S)(DBO-S)Low].
$\mathcal{C}^S_{374}$ high quality, high pH-S [(pH-S)High], [(DQO-S)Low].
$\mathcal{C}^S_{375}$ average quality water [(DQO-S)(DBO-S)Average].
$\mathcal{C}^S_{376}$ poor quality, high DQO-S [(DQO-S)High].
$\mathcal{C}^S_{370}$ poorest quality water [(DQO-S)(DBO-S)High].
d_4 unknown.

5 Identifying Typical Trajectories

A trajectory is the sequence of classes of a certain day for all stages of the process [7]. Indeed, a certain day can be located in its corresponding class regarding Input, Settler, Bioreactor or Output. 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 2. In [7] it is show that the process is not well modeled under Markov assumptions, so direct frequentist analysis is done. For this particular application trajectories with $\gamma > 0.025$ are selected and interpreted by the experts. In figure 2 trajectories are graphed. Classes identified for each subprocess are aligned in a relative way from cleaner water at the top to more polluted water at the bottom. Trajectories are represented with arrows linking the consecutive classes between stages. For example, wastewater flows from $\mathcal{C}^E_{331}$ to $\mathcal{C}^D_{328}$ to $\mathcal{C}^B_{324}$ and $\mathcal{C}^S_{375}$ on 12-IX-1995, where $\mathcal{C}^E_{331} \in \mathcal{P}^E$, $\mathcal{C}^D_{328} \in \mathcal{P}^D$, $\mathcal{C}^B_{324} \in \mathcal{P}^B$ y $\mathcal{C}^S_{375} \in \mathcal{P}^S$. What means that WW has normal levels of pollutants at the input, still keeps average levels of solids in the settler, underload WW with average waste flow rate in the bioreactor and exits with average quality. This information, regarding the dynamics of the process, is richer than static information originally obtained from a global *ClBR* without considering the structure in subprocesses of the target phenomena.

$\tau_3 = \{c^E_{331}, c^D_{323}, c^B_{3324}, c^S_{375}\}$ normal levels of pollutants at the input, high levels of SSV-D and DBO-D in the primary settler, underload WW with average waste purged flow rate and exits with average quality

$\tau_1 = \{c^E_{331}, c^D_{328}, c^B_{324}, c^S_{375}\}$ normal levels of pollutants at the input, still keeps average levels of solids in the primary settler, underload WW with average waste purged flow rate and exits with average quality

$\tau_2 = \{c^E_{328}, c^D_{328}, c^B_{326}, c^S_{374}\}$ 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

$\tau_6 = \{c^E_{328}, c^D_{328}, c^B_{326}, c^S_{375}\}$ 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

$\tau_4 = \{c^E_{328}, c^D_{328}, c^B_{324}, c^S_{373}\}$ 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

$\tau_5 = \{c^E_{328}, c^D_{322}, c^B_{324}, c^S_{370}\}$ 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

$\tau_7 = \{c^E_{328}, c^D_{322}, c^B_{332}, c^S_{374}\}$ 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

$\tau_8 = \{c^E_{318}, c^D_{328}, c^B_{332}, c^S_{374}\}$ 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

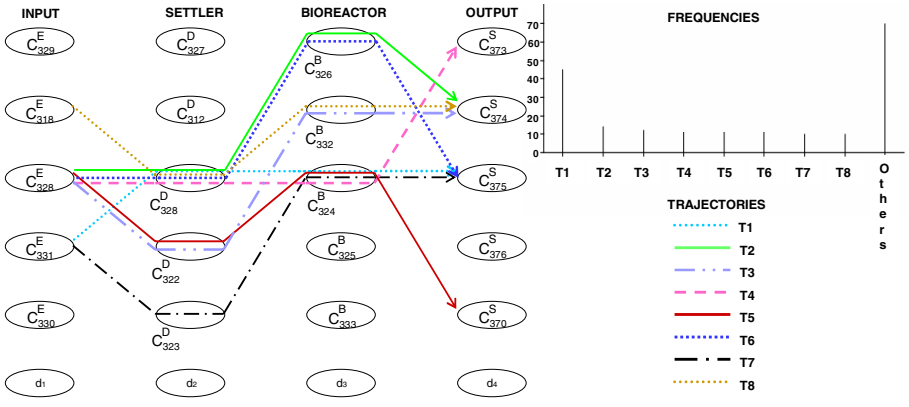

Fig. 2. Transition Diagram Frequencies of Trajectories and between states

6 Conclusions

In this work, an hybrid tool of AI and Stats is applied as an alternative to identify more characteristic dynamic patterns of wastewater in a *WWTP*. *ClBR×E* is designed in such a way that prior knowledge of the experts is integrated into the analysis to improve the quality of the clustering processes. Analysis is performed step by step and final integration permits construction of models for dynamics behavior of the plant in qualitative terms. Dynamics of wastewater treatment is very difficult to model by classical methods and here a combination of AI and Stats provides a nice frame to face modeling. The design of a methodology for dynamic analysis by subprocesses to obtain knowledge about the process is complementary to the identification of characteristic situations in a *WWTP* which was done by global cluster with all variables in previous works. *ClBR×E* allows to analyze trajectories of wastewater along the treatment process as well as to predict the plant's evolution, in the short or mid-term, identifying the most typical trajectories to be observed. Linking this to the characteristic situations previously identified provides a knowledge that improves the decision-making support. This methodology is being formalized and a technique focused on most significative trajectories is being developed to the automatic generation of the corresponding characterization. In the future, the effect of conditioning a class by a trajectory on the global label provided by the expert will be analyzed.

Acknowledgments. This research has been partially financed by TIN2004-01368.

References

1. Struss, P.: Artificial Intelligence Methods for Environmental Decision Support. In: e-Environment: Progress and Challenge, pp. 1–14 (2004)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthursamy, R.: Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park (1996)

3. Gibert, K., Roda, I.: Identifying characteristic situations in wastewater treatment plants. In: Workshop BESAI, vol. 1, pp. 1–9 (2000)
4. Metcalf, Eddy, Inc.: Wastewater engineering treatament. Disposal and reuse. 4th edn. George Tchobanoglous McGraw Hill (revised) (2003)
5. Gibert, K.: The use of symbolic information in automation of statistical treatment for ill-structured domains. AI Communications 9, 36–37 (1996)
6. Gibert, K., Aluja, T., Cortés, U.: Knowledge Discovery with Clustering Based on Rules. In: Interpreting Results. LNCS (LNAI), vol. 1510, pp. 83–92. Springer, Heidelberg (1998)
7. Gibert, K., Rodríguez Silva, G., Rodríguez Roda, I.: Trajectories' Minning Between Subprocesses in a Wastewater Treatment Plant. iEMSs (in press, 2008)