DECISION TREE AND EXPERT SYSTEM FOR IMPROVE THE MANAGEMENT OF WATER

COSTICĂ NITU, EUSEBIU PRUTEANU[*], DAN DĂSCĂLIȚĂ[**],

*University " Polytechnic" of Bucharest, *University Bacău, ** Department of Water Management Siret, Bacău*

Abstract: In this work the main goal is to design and develop a expert system, with special focus on water management and quality. It is remarkable the high quantity of information and knowledge patterns implicit in large databases coming from the monitoring of any system or dynamical environmental process. For instance, historical data collected about meteorological phenomena in a certain area, about the performance of a wastewater treatment plant, about characterizing environmental emergencies (toxic substances wasting).

Keywords: Environmental decision-support system, river, rule-based expert system, water management, Knowledge Acquisition and Management, Data Mining, Environmental Databases

1. INTRODUCTION

An intelligent information system for decreasing the decision-making time and improving consistency and quality of decisions in Environmental Systems can be defined as an Environmental Decision Support System (EDSS) with is an ideal decision-oriented tool for suggesting recommendations in an environmental domain. The main outstanding feature of EDSS is the knowledge embodied, which provides the system with enhanced abilities to reason about the environmental system in a more reliable way.

A common problem in their development is how to obtain that knowledge. Classic approaches are based on obtaining the knowledge with manual interactive sessions with environmental experts. But when there are available databases summarizing the behaviors of the environmental system in the past, there is a more interesting and promising approach: using several common automated techniques from both Statistics and Machine Learning fields. These joint techniques are usually named as data mining or knowledge discovery technologies.

All this information and knowledge is very important for prediction tasks, control, supervision and minimization of environmental impact either in Nature and Human beings themselves. The project is involved with building an Intelligent Data Analysis (IDA) tool to provide the support to these kinds of environmental systems. This tool is basically composed by several statistical data analysis methods, such as one-way and two-way descriptive statistics, missing data analysis, clustering, and relations between variables. Also, several machine learning techniques will be integrated, coming from Artificial Intelligence, such as clustering, classification rule induction, decision tree induction, case-based reasoning techniques, reinforcement learning, and dynamical analysis.

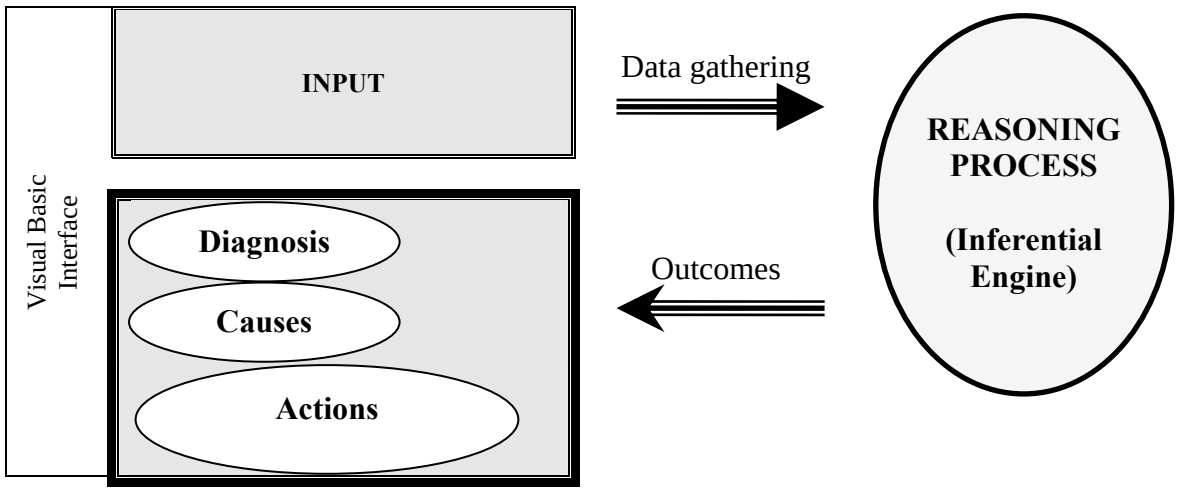

Fig. 1. From data to outcomes (simplified)

In the system, artificial intelligence techniques are applied to the water-management field in the form of an environmental decision-support system.

Some examples of EDSSs developed recently and applied to the water domain are described by, among others, [1], [4], [9], [2], [8].

The development of the EDSS has been carried out following a methodology composed of a series of phases, each with its own inputs, activities and outputs [10]:

1. environmental problem analysis;
2. data collection and knowledge acquisition;
3. system analysis and design;
4. *problem-solving method* (PSM) selection[1];
5. PSMs integration;
6. system implementation;
7. validation;
8. maintenance;

The rules of the expert system are grouped into four modules.

2. REALISATION AND IMPLEMENTATION

In this paper, we analyze the implementation of the *rule-based expert system* (RBES) and of the graphical user interface. RBESs are mainly composed of a knowledge base (KB) and an inferential engine (IE).

2.1. Inferential engine

The inferential engine (IE) works with rules and provides the reasoning mechanism. In case of reuse in different domains, the KB would need to be redeveloped with new knowledge-components (as described in section 2.2). The current implementation of the IE is Java-based (platform independent) and is integrated with a friendly user-interface in Visual Basic (VB) (see Figure 1).

Once the system is started the user has to fill in different forms of data-input that the system presents via the user-interface. In the IE, the rules correspond to the decision trees and the facts correspond to the data introduced

[1] The term *PSM* corresponds to the term *model* used by Poch et al 2002 [10].

by the user.

Then the inference process starts, trying to find out if the facts match some of the antecedents of each one of the active rules. If a rule is triggered, new facts can be introduced into the facts base, as a result of the inference. This process finishes when the IE has tested all the facts with all the active rules. Afterwards, the IE delivers the results to the interface component that parses and shows the results to the user in an appropriate format.

2.2. Knowledge components

For building and validating the KB of a decision-support system for our given practical domain, four knowledge components (KC) are needed [4], [9].

Table 1. Decision trees (DT) and related diagnosed problems.

Decision tree (DT)	Decision tree name	Represented problems
DT1	***Nitrogen***	• *Excess of ammonium, nitrate, nitrite*
DT2	***Phosphorous***	• *Eutrophication*
DT3	***Organic matter***	• *Excess of organic matter*
DT4	***Suspended solids***	• *Excess of suspended solids*

KB is codified by means of rules, which are sets of *conditions and conclusions*. As a prior step to build the KB, knowledge is structured and represented in decision trees (DTs) [4], [2]. Every DT refers to a set of specific problems (shown in Table 1) and is composed of two modules: one for problem diagnosis and one for cause detection.

The developed DTs correspond to those problems for which water managers and environment experts expressed a greater interest and preoccupation. Four DTs have been developed: one for nitrogen-related problems, one for eutrophication[2], one for organic-matter problems, one for suspended solids. There are related to physics-chemical elements in the water, and We tray next time to focus on the physical, biological and morphological characteristics of the river ecosystem (riparian zone and streambed), which can affect the river's functionality and self-purification capacity. The self-purification capacity is in turn an important aspect to be taken into account in water pollution problems.

2.3. Rule modules

The KB, rules are grouped into four modules, or *steps*:

1. *Symptom discovery*.
2. *Problem diagnosis*.
3. *Cause detection*.
4. *Actuation*.

For a full understanding of the KB implementation and functioning, as well as the interaction with the user, we present a complete use case of the EDSS.

The process starts with the selection by the user of one of the following two options:
1. evaluation of possible stream problems;
2. assessment of the alteration degree of the stream.

In the following, we consider the first option because it is the one related to the implementation of the RBES.

Symptom discovery. The system begins to gather data, asking questions to the user about groups of *significant descriptors* (DS), or quality elements for the classification of ecological status. Some of these DS are in accordance with the WFD; other ones have been defined by the authors according to their experience and the

[2] Eutrophication problem is evaluated by means of the N:P molar-ratio calculation.

knowledge acquired from diverse sources (e.g., EPA manuals by Barbour et al. 1999) [1]:

1. River basin DS. These elements are related to the location of the river in its river catchments, to the characterization of the basin and to the identification of diffuse pollution sources (e.g., geology, predominant land use).

2. Hydro morphological DS supporting the biological DS. Examples of these elements are: stream width, water velocity and, in general, the hydrological regime and the river continuity.

3. Water quality DS. Examples of these elements are: nitrogen and phosphorous data, water odor, conductivity, water color, water temperature, pH.

4. Point nutrient-source DS. Identification, location and characterization of the existing point sources of nutrients in the river catchment's, e.g.: input of wastewater, ammonium.

5. Riparian DS. These elements characterize the riparian zone and help to estimate the quality of the river in relation to it. Examples are: types of riparian vegetation, soil permeability.

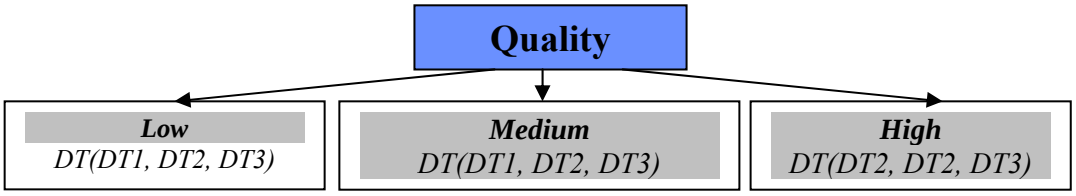

Fig. 2. Symptom-discovery meta-rules

These data and a set of meta-rules representing domain requirements are used to select the DTs to be activated (see Figure 2 for an example of these rules).

Problem diagnosis. When, for instance, DT2 (*phosphorous*) is selected, its problem-diagnosis module is activated. Part of the problem-diagnosis rule-inference is shown in Figure 3.

IF	Total phosphorus concentration<1.mg P/L Geology=Calcareous Phosphate concentration>0.150mg P/L	AND AND AND	**THEN**	Problem eutrophication (hyper)
IF	Total phosphorus concentration>1.8 mg P/L Nitrogen/phosphorus ratio <16 mg pH>5 Phosphate concentration<0.050 mg P/L	AND AND AND AND	**THEN**	Problem eutrophication (low)

Fig. 3. Problem-diagnosis rules for the *phosphorous* decision tree.

In the same way, inference is carried out in the rest of DTs activated by the meta-rules.

- Cause detection**.** For each problem diagnosed, the cause-detection module of the corresponding DT is activated. Part of the cause-detection rule-inference is shown in Figure 4.
- Actuation**.** Once the system executed all triggered rules in activated DTs, it shows the user a set of <diagnosis, cause> pairs (DCPs), for him to analyze.

IF	WasteWater origin=industrial Concentration of total nitrogen <15 mg N/L Concentration of total phosphorous <2 mg P/L Watershed nonpoint-source pollution=due to urban area	AND AND AND AND	**THEN**	(Cause) = urban area
IF	WasteWater origin=industrial Concentration of total nitrogen >15 mg N/L Concentration of total phosphorous >2 mg P/L	AND AND AND	**THEN**	(Cause) =Factory

Fig. 4. Cause-detection rules for the *phosphorous* decision tree.

The user chooses the DCPs he is interested in and, for each one, the actuation category (hydro morphology,

chemistry, biota, best practices, hydrology) and the actuation geographical-scope (river basin, riparian zone, river body). With these data, the system is able to offer an ordered list of recommended courses of action to carry out (see an example in Figure 5), as well as, when possible, a series of complementary parameters, such as: chances of success, feasibility, response time, effort vs. environmental benefit, references.

IF	Problem=eutrophication (hyper)	AND	**THEN**	***Action 1***	Construction of riffles and small dams (EB: increase of DO)
	Category=hydro-morphology	AND		***Action 2***	Construction of man-made steps (EB: increase of DO and reduction of erosion processes)
	Geographical-scope=riparian zone	AND		***Action 3***	Laying rocks in the riverbed (EB: increase of DO)

Fig. 5. Recommended actions in the *actuation* step (simplified[3]).

Forecast**.** The system forecasts what improvements would take place in the river if one of the actions suggested were carried out. As outcome, the system shows the user a comparison of the current problematic state versus the state after the application of the action, as well as a measure of the improvement in the quality of water.

2.4. Decision support

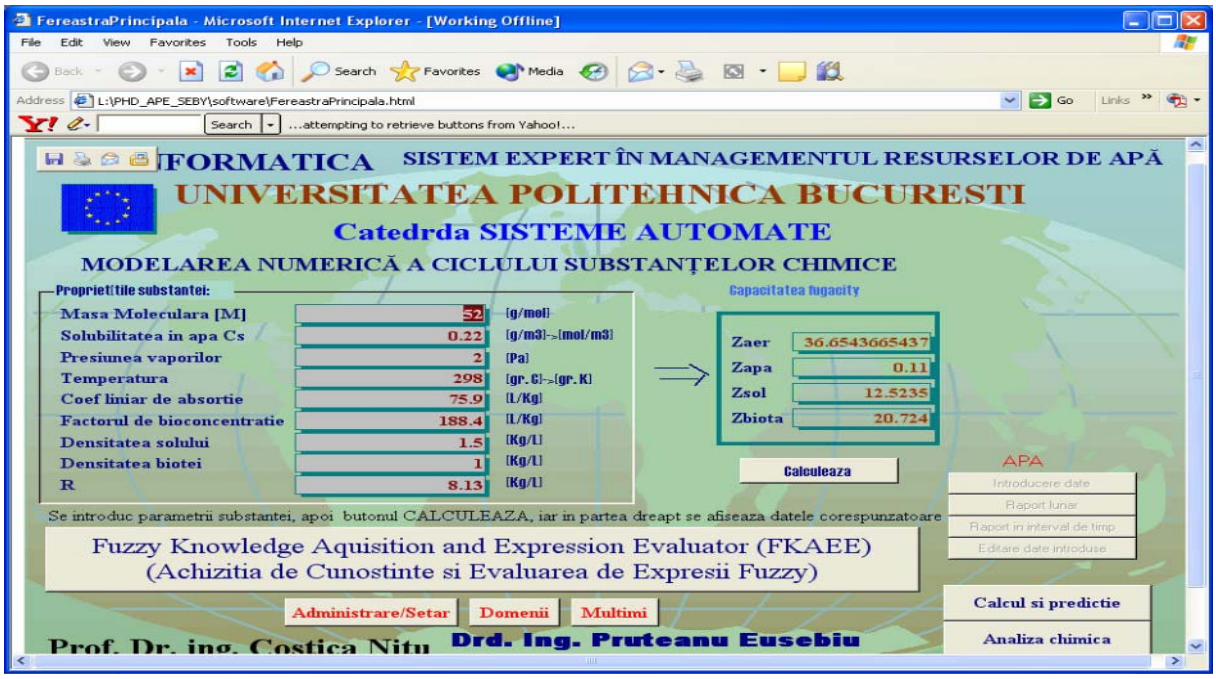

Fig. 6. The architecture of the knowledge management

In summary, the decision support supplied by the system consists of providing:
1. ***Diagnosis***: inferring possible stream problems, assessing the alteration degree of the stream, and evaluating the source and magnitude of nutrient loads.
2. ***Actions***: offering alternative, ranked courses of action to solve possible problems.
3. ***Forecast:*** providing several scenarios to simulate the effect of the different actions proposed as solutions.

The module of Knowledge Management are the following:

[3] ***EB***-Environment Benefit; ***DO***-Dissolved Oxygen; (***DCPs***) <diagnosis, cause> pairs

- Integration of different knowledge patterns for a predictive task, or planning, or system supervision;
- Validation of the knowledge pattern acquired;
- Knowledge utilization by end-users;
- User interaction;

Database management allows adding a new variable to the database, deleting one variable from the database, and modifying the characteristics of a variable such as its relevance or the range of values. Figure 7 depicts a new variable addition in the database.

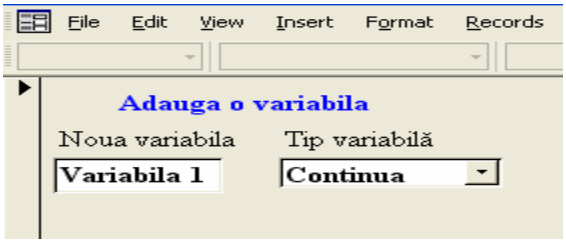

Fig 7. New variable addition in a database

Descriptive statistical analysis is composed by basic statistical analysis such as computation of mean, standard deviation, median value or correlation coefficient. One-way and two-way analysis of both variables and classes are also provided. Graphical representations of analysis results are implemented through both one-way plots and two-way plots, as well as histograms or letterplots for class distribution visualisation.

3. CONCLUSIONS

A system is being developed with the objective of capturing knowledge from water managers and environmental-science experts, regarding nutrients-excess effects in streams and of combining this knowledge into a user-friendly tool to assist water managers. The expert system provided by the system to water managers consists of: (1) diagnosis: inferring possible stream problems, assessing the alteration degree of the stream, and evaluating the source and magnitude of nutrient loads; (2) actions: offering alternative, ranked courses of action to solve possible problems; (3) forecast: providing several scenarios to simulate the effect of the different actions proposed as solutions.

REFERENCES

[1] Barbour, M.T., Gerritsen, J., Snyder, B.D. and Stribling, J.B., Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. Washington, D.C., 1999.

[2] Ceccaroni, L., Cortés U., and Sànchez-Marrè, M., OntoWEDSS: augmenting environmental decision-support systems with ontologies 2004. *Environmental Modelling & Software, in press*.

[3] Chang, N. B., Chen, Y. L. and Chen, H.W., A fuzzy regression analysis for the construction cost estimation of wastewater treatment plants. *Journal of Environmental Science and Health. Part A - Environmental Science and Engineering and Toxic and Hazardous Substance Control*, *32*(4), 885-899, 1997.

[4] Comas, J., Llorens E., Martí, E., Puig, M. A., Riera, J. L., Sabater, F. and Poch, M., Knowledge acquisition in the STREAMES project: the key process in the Environmental Decision Support System development, 03

[5] Davis, J. R., Farley, T. F. N., Young, W. J., and Cuddy, S. M., The experience of using a decision support system for nutrient management in Australia, *Water Science and Technology*, *37*(3), 209-216, 1998.

[6] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy, *Official Journal L327*, *2000-10-23*, *P0001*.

[7] Great Lakes Commission for the Great Lakes States and Provinces. *Toward a Water Resources Management Decision Support System for the Great Lakes-St. Lawrence River Basin*. Great Lakes Commission, Eisenhower Corporate Park, 2805 S. Industrial Hwy.

[8] Rodríguez-Roda, I., Comas, J., Colprim, J., Poch, M., Sànchez-Marrè, M., Cortés, U., Baeza, J. and Lafuente, J., A hybrid supervisory system to support wastewater treatment plant operation: implementation and validation, *Water Science and Technology*, *45*(4-5), 2002.

[9] Rousseau, A. N., Mailhot, A., Turcotte, R., Duchemin, M., Blanchette, C., Roux, M., Etong, N., Dupont, J., and Villeneuve, J. -P., GIBSI–An integrated modelling system prototype for river basin management. *Hydrobiologia* 422/423: 465–475, 2000.