Expert Systems With Applications, Vol. 5, pp. 403–410, 1992
Printed in the USA.

0957-4174/92 $5.00 + .00

Design of an Expert System for Water Pollution Determination / Prevention

METKA VRTAČNIK, DANICA DOLNIČAR, ANDREJA CIZERLE, PAVEL ČOK, SASA A. GLAŽAR, AND RADOJKA OLBINA

University of Ljubljana, Ljubljana, Yugoslavia

Abstract—*An initial model of an expert system was designed to provide quick responses for Slovenian river water pollution emergency situations by: identifying selected pollutants; planning immediate action to cope with urgent hazardous situations; identifying actual pollution source(s); and providing advice to reduce and prevent pollution at the source. The model knowledge base is implemented on the Knowledge.Pro communication system and is derived from two relational databases using the Informix.Sql relational database management system.*

1. INTRODUCTION

IN THE LAST FEW DECADES, protection of surface and groundwaters from pollution has received a high priority since many countries have suffered from considerable deterioration in the quality of their water resources. This pollution is caused by discharges of untreated industrial and municipal wastewaters, agricultural runoffs, spills of hazardous substances, or "night dumpings" of industrial (hazardous) and municipal waste.

Data on river water quality parameters are gathered by laboratories worldwide. However, an effective method of extracting value-added information from these data to facilitate decisions on the implementation of cost-effective pollution prevention measurements to safeguard people, livestock and industrial development in a region remains a problem.

While not always sufficiently recognized, artificial intelligence (AI) approaches based on inductive and deductive learning and pattern recognition methods, represent a great potential in searching for regularities/patterns in large multiparametric data sets. Commercially available AI systems (e.g., expert system shells) are designed to support searches for solutions in various classes of problems, e.g., classification, based on shallow, as well as deep knowledge, process design, process optimization, diagnosis, analysis, planning/scheduling, monitoring, process control as presented by W.B. Gevarter (1987).

In searching for cost-effective solutions to water and wastewater treatment and pollution prevention, the following computer-supported methods are extensively emerging to extract value-added information from existing experimental data, often organized as global relational databases (Iwata, 1989; Bengtson & Parsmo, 1990):

- Computer supported water quality modelling and pattern recognition approaches based on different classification models (James, 1984; Fu, 1985; Juricskay & Veress, 1986; Wienke & Danzer, 1986; Van der Voet, Coenegracht, & Hemel, 1987; Saaksjarvi et al., Khalighi, & Minkkinen, 1989; Tiwari & Ali, 1989; Gao & Xu, 1989; Frank & Lanteri, 1989; Rump 1990);
- Induction systems based on inference theory of learning (Barnwell, Brown & Marek, 1986; Hushon, 1990; Miller, 1989; Venkataramani, Bamopoulos, Forman, & Bacher, 1989; Perman, 1990; Wable, Brodard, Duguet, Mallevialle, & Roustan, 1990; Wood, Houck, & Bell, 1990; Seys, Duran, & Sivitz, 1989; Ouzilleau, Serodes, Beriault, & Suel, 1990).

2. PROBLEM DEFINITION

The classification of Slovenian rivers according to their pollution level is defined by four categories (Yugoslav Federal Register, No. 6, 1978):

Category I (drinking water or drinking water after treatment),

Category II (water that can be used for bathing, water sports or fish farming, and could be used for drinking after adequate treatment),

Revised version of Vrtačnik, M., Dolničar, D., Cizerle, A., Čok, P., Glažar, S. A., & Olbina, R., Design of an expert system for water pollution determination/prevention, pp. 917–928, from Liebowitz: *Expert Systems World Congress Proceedings*, copyright 1991, with permission from Pergamon Press Ltd., Headington Hill Hall, Oxford, OX3 0BW, United Kingdom.

Requests for reprints should be sent to Metka Vrtačnik, University of Ljubljana, Faculty of Sciences and Technology, Department of Chemical Education and Informatics, Vegova 4, 61000 Ljubljana, Yugoslavia.

Category III (water that could be used for irrigation or for industrial purposes after adequate treatment except for the food industry),
Category IV (polluted water).

The criteria for assigning a polution category/level of a particular water sample at sampling site are based on the results of:

- selected physical and chemical analyses,
- specific analyses of heavy metals and selected organic pollutants,
- microbiological analyses.

The overall pollution category of water samples is assigned by an expert panel which includes chemists, biologists and hydrologists.

Baseline data on water quality collected at 131 sampling sites for three consecutive years show that the quality of water in the majority of Slovenian rivers is decreasing. In most cases this is due to uncontrolled discharges of untreated industrial wastewaters and agricultural runoffs (Research of Slovenian surface water quality for the year 1986).

Thus, an attempt has been made to design an expert system to support a quick response to a water pollution emergency situation through:

a. *identification* of selected pollutants with probable damaging effects, and of immediate action in a given environmental situation with a hierarchical sequence of proposed tests to be carried out, taking into account the cost–benefit factor;
b. *planning* immediate action to cope with an urgent hazardous situation;
c. *identification* of the actual pollution source(s), e.g., individual enterprises, farms, etc.;
d. *providing advice* on ways to reduce pollution at the source in order to prevent a recurrence of the hazardous situation.

Such an expert system attempts to provide an efficient support of water pollution control at lower cost by providing sufficient reliability to define the water pollution levels on a reduced number of analyses needed. The system also attempts to provide advice for the most efficient wastewater treatment technology for a given stream.

3. RESEARCH HYPOTHESIS

Knowledge in a given domain consists of descriptions that characterize objects, relationships and procedures for manipulating these descriptions. In order to make expert systems useful tools to handle problems:

- A most appropriate knowledge representation must be searched for in a given problem domain,
- An effective control of the reasoning process, which depends on the type of the domain and knowledge representation, must be implemented.

In order for a system to reason, it must be able to manipulate the data of a specific problem using the general facts of the domain (declarative knowledge) and the rules (procedural knowledge) until a satisfying solution is found or no more rules can be applied (Klaessens & Kateman, 1987; Levine, Drang, & Edelson, 1986).

One of the ways of representing knowledge of a domain is by means of relational or semantic diagrams. Thus, a relational database with general facts of the problem domain (declarative knowledge) was designed. This relational database can be viewed as a template of a knowledge base of an expert system.

4. DESIGN OF A DECLARATIVE KNOWLEDGE OF AN EXPERT SYSTEM— AS A RELATIONAL DATABASE

4.1. Criteria for Pollutant Selection

Water pollutants included in the relational database were selected from the list of 299 compounds considered in Section 311 of the USA Water Pollution Control Act as compounds having a potential impact on the environment based on their:

- toxicity,
- degradability in an aqueous environment.

Approximately 250 compounds will eventually be included in the database. At present the database includes 128 compounds.

4.2. Sources of Information on Selected Pollutants

Several computerized as well as primary literature sources have been used in the collection of reliable information on:

1. physical and chemical properties data, industrial occurrence and treatability data, data on biological effects; Ljubljana, 1989;
2. analytical methods for pollutant determination in fresh and drinking water;
3. site-specific data sources:

- *Register of Water Bodies and Monitoring Points,* The Slovenian Water Works Association, Hydrometeorological Institute of Slovenia;
- *Time Series Data on Chemical, Biological and Hydrological Aspects of Water Bodies,* Research on Slovenian surface water quality, The Slovenian Water Works Association, Hydrometeorological Institute of Slovenia, Biological Institute of Ljubljana University, Boris Kidric Chemical Institute, Institute for Public Health, Maribor (1986, 1987, 1988);
- *Register of Slovenian Industry: Types and Location Coordinates,* Statistical Bureau of Slovenia; Ljubljana, 1989.

4.3. Defining Entity Sets and Their Attributes

The mathematical concept underlying the relational model is the set-theoretic relation, which is a subset of the Cartesian product of a list of domains. A domain is a set of values for a given attribute. A relation is any set of the Cartesian product of one or more domains. In order to develop the conceptual scheme of the relational database model, first the entity-relationship diagrams are designed and then the real database systems are built on them.

The main elements of the entity-relationship diagrams are:

Entity sets—groups consisting of all "similar" objects (or actions, or concepts);

Attributes—(also called variables) used for describing properties of entity sets;

Relationship—a relationship among entity sets is an ordered list of entity sets.

In the process of the design of the relational database on water pollutants, the following entity sets have been identified:

- chemical substances (pollutants),
- synonyms,
- water quality standards,
- reactions in water,
- possible industrial sources of pollutants,
- other possible sources,
- wastewater treatment methods,
- cell multiplication inhibition test,
- test on acute toxicity,
- odor test,
- MERCK rapid tests,
- UV and visual spectroscopy,
- atomic absorption spectroscopy (AAS),
- thin-layer chromatography (TLC),
- high-performance liquid chromatography (HPLC),
- gas chromatography (GC).

Properties of each entity set are described with a selected set of attributes.

4.4. Entity-Relationship Diagram

The conceptual structure of the relational database model on "water pollutants" is graphically summarized in the entity-relationship diagram (Fig. 1). Rectangles are used to represent entity sets, circles to represent attributes, and diamonds to show relationships. The model is designed in such a way that a series of general and specific questions can be answered by a proposed conceptual structure of the relational database:

4.4.1. *General questions: What* are the probable sources of the pollutants?

What are the observable biological effects of pollutants on aquatic organisms?

Which are the main chemical/biochemical transformations of a particular pollutant in water?

How can classes of pollutants be identified on the basis of the characteristic symptoms they cause, e.g. algal bloom, dead fish, odor, etc.?

How can a hypothesis on the possible cause of pollution be confirmed with the choice of a cost/effective analytical procedure for qualitative and quantitative determination of pollutants in river water?

How can pollution be reduced at its source?

4.5. Design of a Site-Specific Database

The conceptual scheme of the relational database on potential pollution sources in Slovenia, and time series of data on water bodies monitoring is designed in such a way that the following main questions can be answered:

What are the levels of pollution for a particular Slovenian river body?

How does the pollution level of a particular river body change over time?

Which enterprises/companies can be identified as possible sources of pollution at a particular sampling site?

How are pollution levels of Slovenian rivers distributed in regions/communities?

Where has the impact of river water pollution in Slovenian regions already caused severe damage to the environment and to industrial development?

The entity-relationship diagram is presented in Fig. 2.

5. IMPLEMENTATION OF A RELATIONAL DATABASE MODEL

Commercial database management systems which will fully support the entity-relationship approach are still under development. Thus, for the implementation of the entity-relationship schemes for "water pollutants" as well as "Slovenian rivers and industries" database models, the relational database management system Informix.Sql, version 2.10.06C was chosen. Informix.Sql is implemented on the industry standard definition for SQL and the proposed ANSI standard for relational database systems. The Informix.Sql system consists of usual programs or modules that perform data-management tasks.

The entity-relationships diagram for the relational database "water pollutants" model was transformed with the application of Informix.Sql system into 18 tables. Each table corresponds to one entity set. The name of each column in the table represents one attribute. Tables were then joined logically into five multiple-table screen forms for data entry and data query by example:

- (PROPERTIES) properties of pollutants, potential industrial sources, biological effects on aquatic organisms and possible fate of pollutants in water,
- (ODOR) odor quality and hedonic tone,
- (SOURCES) potential natural and man-made sources of pollutants in water,

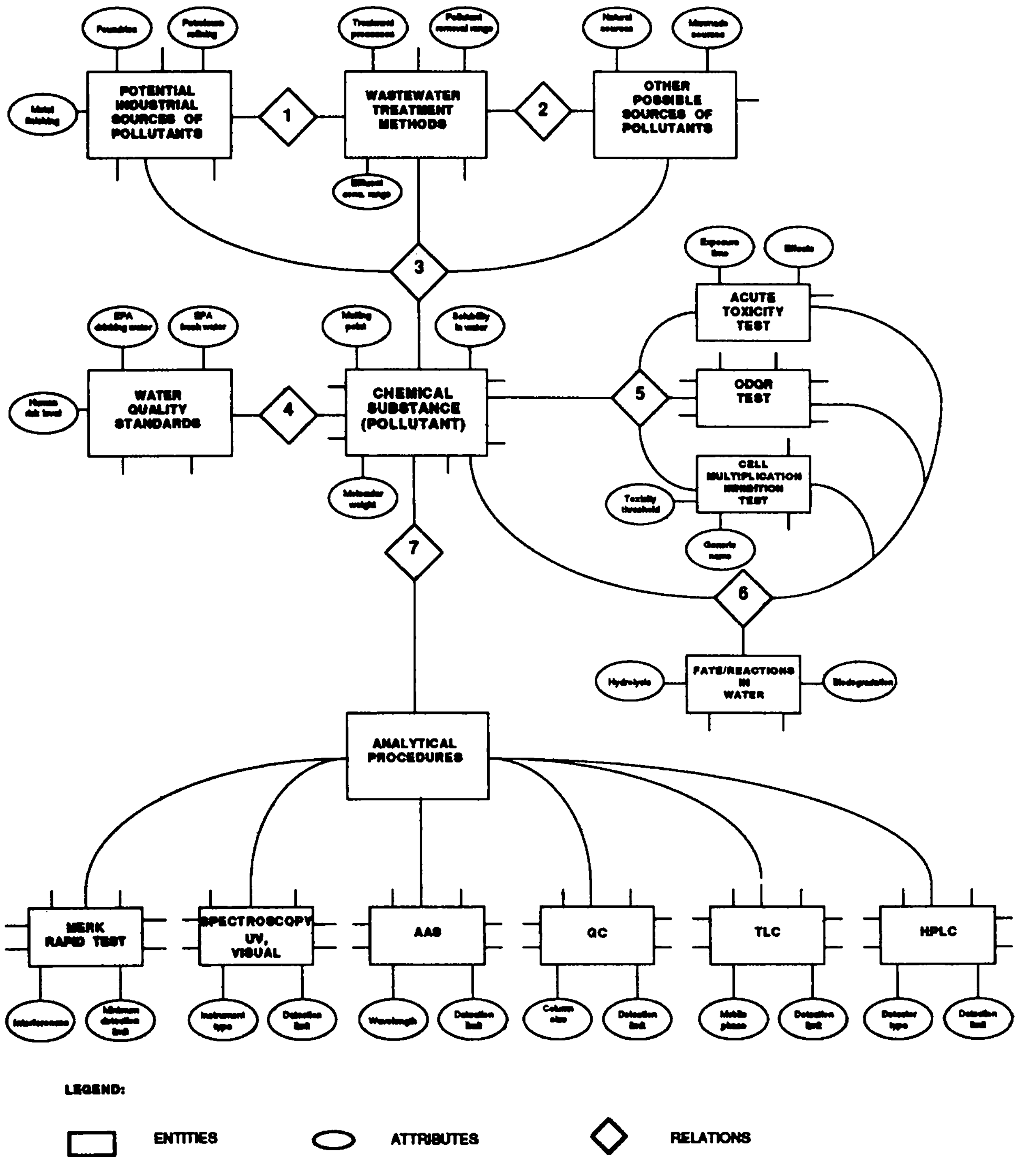

FIGURE 1. Entity-relationship diagram for conceptual level of database "water pollutants."

- (ANALYST) analytical methods for pollutants determination in water,
- (TREATMENT) wastewater treatment methods.

6. DESIGN OF AN EXPERT SYSTEM KNOWLEDGE BASE

On the expert system level, facts in the relational database represent declarative knowledge, or a *global database,* while relationships embedded in the model enable recognition of elements of the procedural part of the knowledge base. In interdisciplinary multiparametric problems, relations among parameters cannot always be explained by the deep knowledge of the domain.

Thus, processing data in the relational database and graphical display of relationships supports the development of empirical rules. The level of generalization of the rule depends on the number and representability of examples in the database.

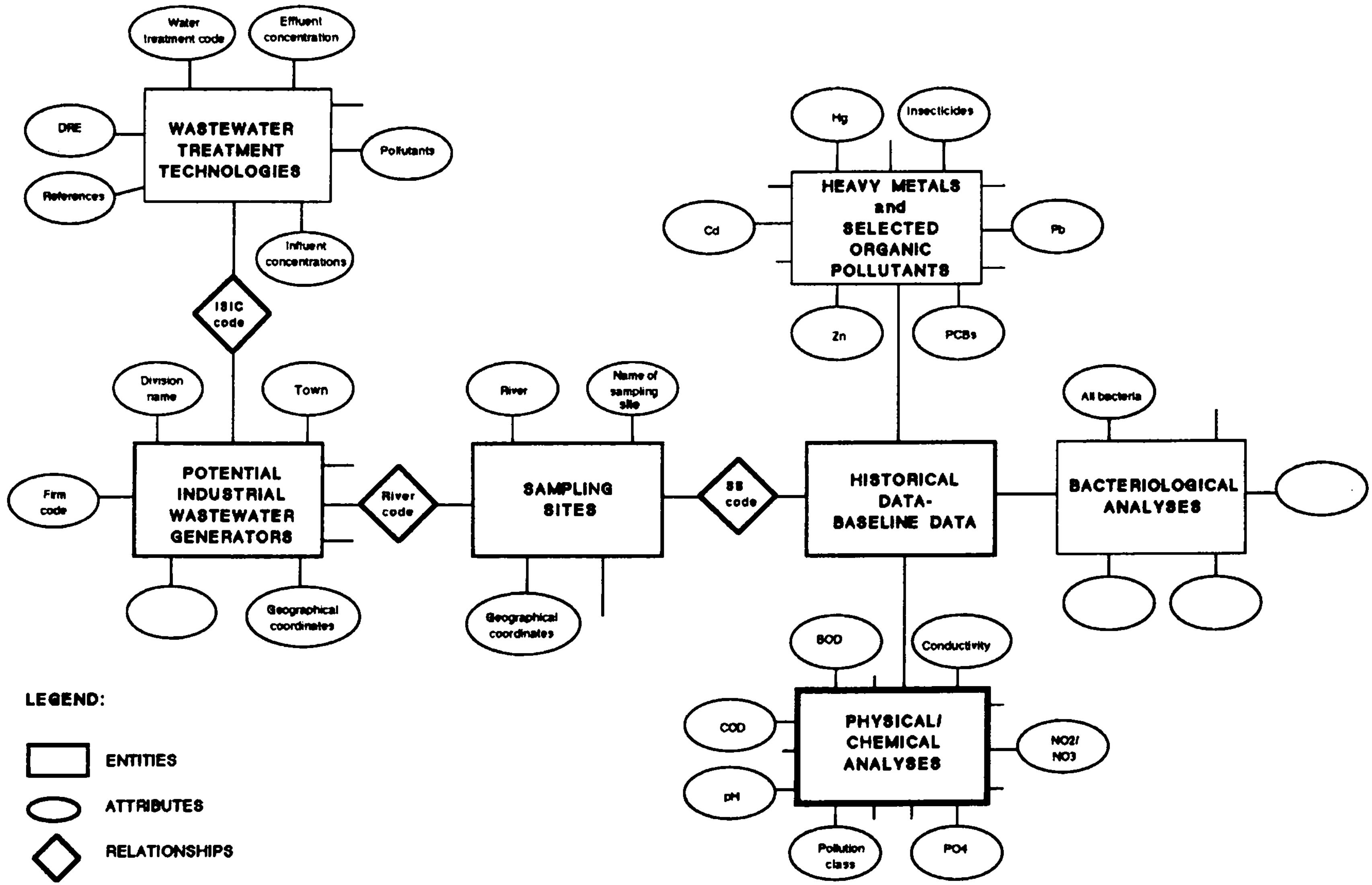

FIGURE 2. Entity-relationship diagram "Slovenian rivers and industries."

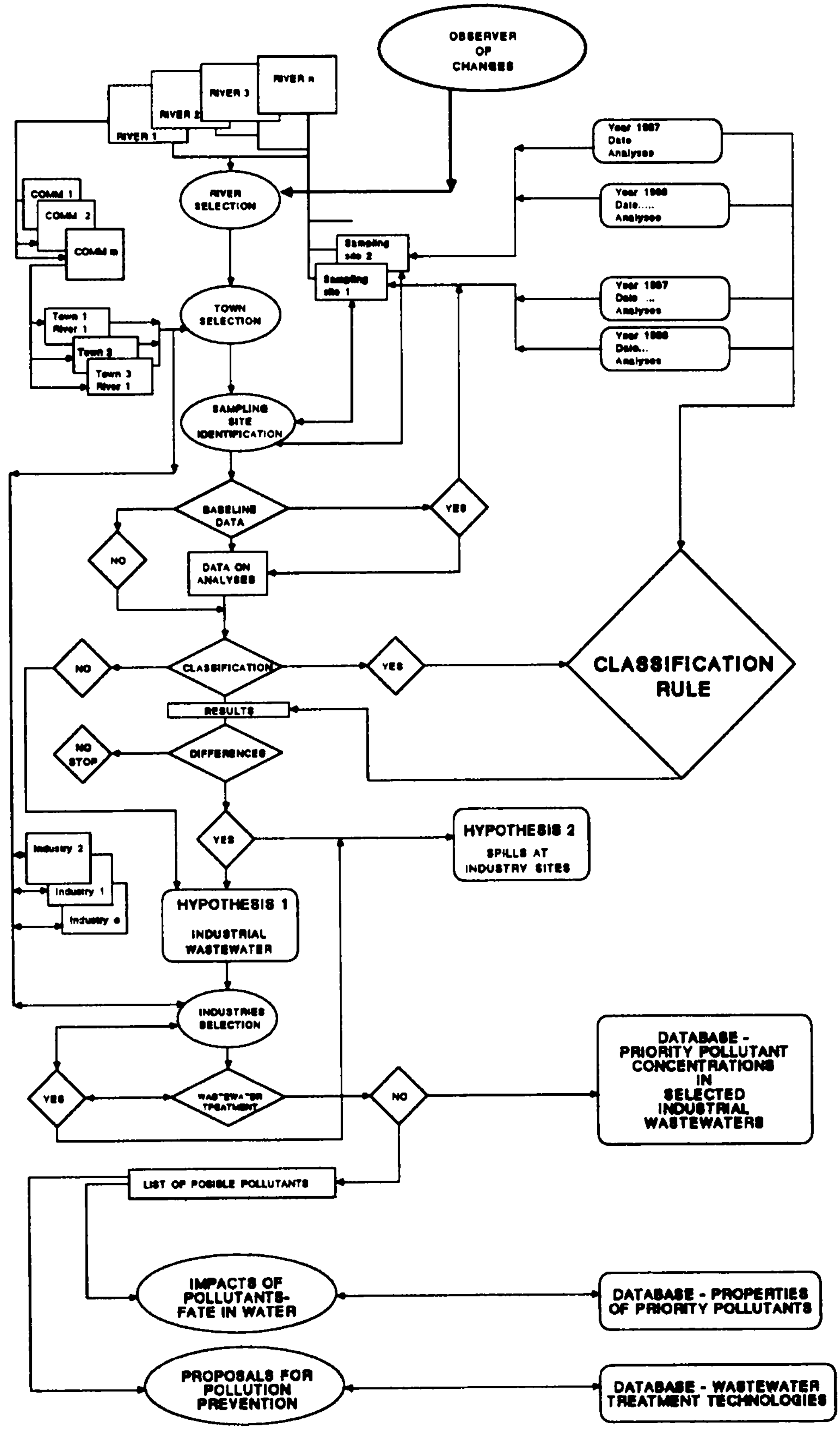

FIGURE 3. Logical relationships of the knowledge base "expert system for water pollution determination/prevention."

Generalization of the rules can be accomplished to a certain extent by a regular and reliable data input. The numeric data compilation is a limiting factor in quick developing relational databases. To overcome this problem, a subscription was made to the TDS-Numerica database. TDS-Numerica gives immediate access to up-to-date, reliable and evaluated data on properties of chemical substances with literature references. The most outstanding feature of the Numerica is that properties can be calculated, based on other known properties or molecular structure, in cases where experimental data are not available.

The model of the expert system knowledge base is given in (Fig. 3). The line of reasoning of an expert system is not yet completely fixed, but at the present developmental stage it enables the user to check two hypotheses:

Hypothesis 1: The pollution is caused by discharges of untreated wastewaters from industries located along the river sites;

Hypothesis 2: The pollution is caused by spills of hazardous substances, which entered the river by an accident or intentionally.

The goal of the consultancy is to identify, within a given region of river water pollution, the most probable industry or industries as pollution sources and to give proposals for the implementation of cost/effective wastewater treatment technology in order to minimize the river water pollution load. The logical relationships are implemented on the KnowledgePro communication system.

An integral part of the expert system is the classification rule, which enables the user of the system to compare the baseline data of the given river pollution levels with the actual results of analysis based on reduced number of measurements. This part is still under development.

Value-added information from the existing baseline data on monitoring Slovenian river water bodies was extracted to reduce the number of measurements for determining water pollution levels. For achieving this goal, discriminant analyses were tested in order to develop a reliable classification rule. The number of parameters which are used normally for the assignment of the pollution class could be reduced from 39 to as few as 21 predictor variables, without significant decrease in the reliability of the classification score (Vrtacnik, Cok, Olbina, Zupan, 1991). Results were compared with the classification rule derived from the inductive expert system shell Assistant.Pro. Assistant is a member of the family of learning systems also called TDIDT (Top Down Induction of Decision Trees). This knowledge representation is in the form of a decision tree. The basic algorithm was developed from Quinlan's ID3, but improved to handle incomplete and noisy data, continuous and multivalued attributes and incomplete domains (Cestnik, Kononenko, & Bratko, 1987).

The classification tree developed by Assitant.Pro enables a determination of Slovenian river pollution levels derived from baseline data for physical–chemical and microbial analyses. Due to the complexity of the classification tree, further work is oriented towards tuning the major criteria for the tree development to simplify its structure.

The classification rule will be evaluated and, if necessary, customized within a group of experts. The customized rule will be implemented on the expert system shell Knowledge.Pro.

7. CONCLUSIONS

This paper describes a prototype expert system for river water pollution determination and prevention. The declarative part of the knowledge is represented in the form of two relational databases "water pollutants" and "Slovenian rivers and industries". The reasoning mechanism of the expert system user interface is designed to analyze two hypotheses of possible causes of river water pollution (untreated industrial wastewater discharges and spills of hazardous substances). Further work is oriented towards the updating of existing relational databases and extraction of value-added information from the existing data by the application of commercially available induction systems, e.g., Assistant.Pro and Knowledge Maker.

Acknowledgements—The authors wish to express their appreciation to the Ministry of Research and Technology of Slovenia and the U.S. Environmental Protection Agency, Cincinnati, OH, for their grants in partial support of this work. The authors thank Aleksandra Kornhauser, University of Ljubljana, Faculty of Sciences and Technology, for her professional support and help. The authors also wish to thank the Slovenian Water Works Association, who allowed them to use data collected in their study.

REFERENCES

Barnwell, T.O. Jr., Brown, L.C., & Marek, V. (1986). *Development of a prototype expert advisor for the enhanced stream water quality model QUAL2E* (Internal Report, 87). Environmental Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Athens, GA.

Bengston, U., & Parsmo, D. (1990). Materials data handling: Interfaces between an expert knowledge system and user oriented problem solving. *Proceedings of the International CODATA Conference*, 11[th] Sci. Tech. Data New Era, 102–106.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In: I. Bratko, & N. Lavrac (Eds.), *Progress in machine learning* (pp. 31–45). Wilmslow: SIGMA Press.

HYDRO. Learned Information (Europe) (1986). *The CRI directory of expert systems* (p. 116). Oxford.

Fu, G. (1985). Optimum estimation of river water quality model parameters using computerized scanning calculation graphic analysis gradient searching method. *Huanjing Kexue Xuebao,* **5**(4), 373–385.

Frank, I.E., & Lanteri, S. (1989). Classification models: Discriminant analysis, Simca, Cart. *Chemometrics and Intelligent Laboratory Systems,* **5**, 247–256.

Gao, H., & Xu, Z. (1989). Pattern recognition, method for water quality evaluation for the Songhua River. *Huanjing Kexue,* **10**(1), 75–77.

Gevarter, W.B. (1987). The nature and evaluation of commercial expert systems building tools. *Computer,* **20**(5), 24–41.

Hushon, J.M. (Ed.) (1990). *ACS Symposium Series,* Vol. 431: Expert systems for environmental applications. Washington, D.C.: American Chemical Society.

Iwata, S. (1989). *Expert systems interfaces for materials data bases.* ASTM Spec. Tech. Publ. 1017 Comput. Networking Mater. Data Bases, 175–184.

James, A. (Ed.) (1984). *An introduction to water quality modelling* (pp. 20–45). New York: John Wiley & Sons.

Juricskay, I., & Verss, E.G. (1986). Prima a supervised classification method and its application in analytical chemistry. *Magyar Kemiai Folyoirat, Budapest,* **82**(6), 255–261.

Klaessens, J., & Kateman, G. (1987). Problem solving by expert systems in analytical chemistry. *Fresenius Z. Anal. Chem.,* **326** (203).

Levine, R.I., Drang, D.E., & Edelson, B. (1986). *A comprehensive guide to AI and expert systems.* New York: McGraw-Hill.

Miller, M.J. (1989). Verification and validation of decision support expert systems. Chemical process risk management in international operations. ACS Symp. Ser. 408. *Expert Syst. Appl. Chem.,* 126–146.

Ouzilleau, F., Serodes, J.B., Beriault, A., & Suel, P. (1990). Application of expert systems to the management of accidental spills. *Sciences et Techniques de l'eau,* **23**(2), 153–159.

Perman, C.D. (1990). Improving the performance of wastewater treatment plants: An expert systems approach. *Dissertation Abstract International: Sciences and Engineering,* **B**, **51**(1), 407.

Rump, H.H. (1990). Matrix values in place of individual-substance values—Possibility of evaluation of problem substances in wastewater. *Gewaesserschutz, Wasser, Abwasser, 112 (Abwasser-Abfaltech: Konzepte Prax.)* 331–349.

Saaksjarvi, E., Khalighi, M., & Minkkinen, P. (1989). Wastewater pollution modeling in the southern area of Lake Saimaa, Finland, by SIMCA pattern recognition method. *Chemometrics and Intelligent Laboratory Systems,* **7**(1–2), 171–80.

Seys, S., Duran, M., & Sivitz, M. (1989). Inteleau. Expert system for assistance in the treatment of sanitary hot water. *Eau, Ind., Nuisances,* **125**, 35–37.

Tiwari, T.N., & Ali, M. (1989). Ground water of Nuzvid Town—Regression and cluster analysis of water quality parameters. *Indian J. Environ. Prot.,* **9**(1), 13–18.

Van der Voet, H., Coenegracht, P.M.J., & Hemel, J.B. (1987). New probabilistic versions of the Simca and Classy classification methods. *Anal. Chim. Acta,* **192**(1), 63–75.

Venkataràmani, E.S., Bamopoulos, G., Forman, A.L., & Bacher, S. (1989). Design of an expert system for environmental assessment of manufacturing processes. *Proceedings of the Industrial Waste Conference,* 43rd, 425–433.

Vrtacnik, M., Cok, P., Olbina, R., & Zupan, M. (1991). Application of baseline data for classifying river water samples—A cost-effective solution or just an illusion? *Journal of Environmental Engineering.*

Wable, O., Brodard, E., Duguet, J.P., Mallevialle, J., & Roustan, M. (1990). An expert system for the design and sizing of ozonization reactors used in water treatment. *L'Eau L'Industrie Les Nuisances,* **136**, 37–38.

Wienke, D., & Danzer, K. (1986). Evaluation of pattern recognition methods by criteria based on information theory and Euclidean geometry. *Anal. Chim. Acta,* **184**, 107–116.

Wood, D.M., Houck, M.H., & Bell, J.M. (1990). Automated calibration and use of stream-quality simulation model. *Journal of Environmental Engineering* (N.Y.), **116**(2), 236–249.