AN EXPERT SYSTEM FOR ASSESSING SAFETY AND SECURITY OF HETEROGENEOUS PUBLIC WATER SOURCES

D. A. SWAYNE,^{1,2} C.W. WANG,³ JOHN STOREY,² D. C. L. LAM,⁴ and JANE KERBY¹

University of Guelph, Guelph, Ontario NIG 2W1, Canada 2 ES Aquatic Inc., Kitchener, Ontario, Canada 3 Universiti Malaya, Kuala Lumpur 59100, Malaysia 4 National Water Research Institute, Burlington, Ontario L7R 4A6, Canada

(Received May 1991)

Abstract. We have implemented a diagnostic system designed to advise on the likely causes of sanitary problems with public water sources. The approach to the problem makes extensive use of rule based expert systems and multi media information (maps, data, text, expert knowledge), The rules were based on actual water survey data, with a weighting scheme designed to highlight causes of health risks approximately in reverse order of importance (where such order may be presumed to exist). Output is available in one of several languages.

I. Introduction

The safety and security of public water supplies is a critical issue in all parts of the world. Heterogeneous public water sources can pose risks for consumers unless timely testing, reporting, and record-keeping of water-borne diseases and likelihood of contamination are controlled. On one hand, problems may be isolated to one household or neighborhood. On the other hand, common trends such as faulty design or unsafe bathing or laundry practices or even routine maintenance may be representative of generic flaws in water use practice which could be widespread. Since expertise in such areas is often scarce or costly, it is desirable to distribute such knowledge to a broad range of users for whom it might not otherwise be accessible. This project has attempted to demonstrate the feasibility of the so-called *expert systems approach* in developing the knowledge in a useable form for distribution.

In 1988 a prototype system called μ RAISON (for Malaysian Regional Analysis by Intelligent Systems ON computer) was developed at the University of Malaya, based on the RAISON system at the National Water Research Institute. The International Development Research Centre sponsored this information system for use by both developed and emerging nations for the purpose of data collection and processing. This development was part of an effort to improve coliphage testing procedures for public water supplies (to replace the more costly and difficult coliform analysis). The Malaysian team implemented a superb prototype system, working closely with NWRI and using the toolkit provided by an earlier version of RAISON (as of 1989).

Concurrent with the development of μ RAISON, we designed an expert system advisor. The objective was to implement a heuristic approach to the evaluation of water supplies. The system had to perform analysis on a per well basis and provide a dialogue with the

user describing the findings and their potential impact on public health. *Expert systems,* which are loosely defined as computer programs which achieve a level of performance comparable in some way to a human expert, are well suited to address our requirements. It is in an expert system's *knowledge base* that any data required for intelligent or expert decisions is encoded. This is analogous to the database in the classical dataprocessing model, except that the knowledge base contains both data (information) and knowledge. In a *rule based expert system*, all knowledge is encoded in the form of rules: IF \leq premise THEN <conclusion>. Consequently, for a rule based system, we will sometimes refer to the knowledge base as the rule base.

It is the *inference engine* in an expert system that initiates any action or computation. The inference engine uses the knowledge base, and any external information entered by the user or from other sources (eg: a database) to offer expert advice, or reach an expert conclusion. In general, all the problem solving capability of an expert system is built into the inference engine. The current nomenclature puts inference engines into several operational classes. One class is *the forward chaining* or data driven engines which work from a set of data reaching a conclusion, for example the R 1 (XCON) system (McDermott *et al.* 1982) which takes a customers computing requirements and builds a hardware configuration. Conversely, *backward chaining* systems take a current state of affairs and works backwards to determine the underlying cause (e.g. MYCIN (Buchanan *et al.* 1984), which accepts a set of patient symptoms and tries to determine what is causing them).

Our expert system prototype was designed to have the attributes of simplicity, ease of maintenance, updateability, upgradeability, and traceability. One of the requirements was a minimal computer configuration: an IBM PC or clone with an enhanced graphics (EGA) capability. As well, the system must make use of existing RAISON geographical and database software in its dialogue with a user. We felt that these objectives would be met if we used a simple, market-tested expert system shell and bridged it to μ RAISON. Modifications to the RAISON design to enhance the bridging process would be an important by-product of this experiment.

2. Knowledge Acquisition and Analysis

2.1. THE DATA

Data which were available to one of the authors (Wang) concerning the type and condition of water supplies were automated (Appendix A). These data gave information such as: type of well (eg. dug, pipe, city water supply, open watercourse), state of repair (as evidenced by cracking or crumbling casing or presence of rubbish), surrounding ground type (clay, sand, etc.), population density, presence of small children, agricultural activity (particularly with the presence of animals). Together with the data of coliform and diarrhea incidence, a geographical perspective of the data is maintained by a graphical display of locations of data sources (Figures 1 and 2).

Additonally, data on the type, availability and status of adjacent latrines were included. These data are relevant since many examples of well contamination can be related to an inferior latrine. Examples of such data are: whether toilets are flush, pit or bushes;

Fig. 1. High I evel (state) map.

numbers of individuals using them; proximity to wells; gradient from wells (upslope or downslope); and their general state of repair / cleanliness.

2.2. KNOWLEDGE ACQUISITION

Our approach to the design of the system was to determine a ranking scheme which would survey the data and rate a water source using multiple categories (well rank, latrine rank and overall rank). These ranks are used as a preliminary assessment of a site.

Approximately 1000 data records were used in the development of the ranking scheme and the knowledge base. Well and latrine characteristics variables were first classified into 'good' and 'bad' according to whether their presence or degree would decrease or increase the incidence of disease. These evaluations were purely subjective. A cracked well covering, for instance, would be 'bad', and would therefore downgrade the quality of a well which had originally been given a good evaluation for being constructed with a proper cover. The origins of this first classification are purely 'common sense'. Some responses were dependent on prior answers, so that relationships had to be encoded in a rule base to represent these dependencies. Using the well cover as an example, 'IF the well cover is cracked THEN...' depends on the well being a type requiring a cover. Once the factors considered to contribute to well contamination (either positive or negative) were isolated, a weight was assigned to each. Wells could then be ranked by collecting data for a site and totalling the factor weights. The rank generated was used by the system to trigger the expert advisor. That is, water supplies which scored low in any of the ranking categories were flagged as potential problem sites, and were then subject to the scrutiny of the expert system. From here, potentially problematic conditions were highlighted, and remedial text or dialogue was generated.

Fig. 2. Well locations showing available data.

2.3. METHODS

The analysis was performed using a combination of RAISON tools, i.e. spreadsheet, production ('IF... THEN...') rules, the interpreted programming language RPL and the commercial package 1stCLASStm. The preliminary work, which involved the determination of factors and the establishment of a ranking scheme, was an iterative process.

The well and latrine rankings were assigned originally by qualitative estimation of their relative merits in preventing/promoting disease. Proper coverage of the observations was then tested by comparing our rating scheme on the original data and other, larger datasets not used in the original rule formulation.

To perform analysis comparing the incidence of coliphage with the factors presumed to lead to coliphage contamination, is was necessary to assign numeric values to different attributes of a well *(e.g.: 20 points for a dug well, -2 points for a damaged well*). A coarser rating for each well was assigned, translating the original assessment into a numerical ranking from 0 (excellent) to 4 (poor). A regression analysis was then performed, comparing a measure of coliphage for a given well with the numerical rating, to see if the magnitude and the trend (slope) of the rating coefficient was indicative of the observations. Figure 7 is one instance of this comparison. Subsequent 'tuning' of the scoring system, and pruning of the dataset of redundant observations improved the correlations. We treated this as an interactive tool and not as a measure of the validity of the scoring, so long as a contradiction was not manifesting itself. This rank was used in the first phase of the expert system remedial advisor to determine which aspects of the well/latrine configuration to highlight in the detailed (advice) phase of the expert system consultation. In a sense this classification becomes a certainty factor in assessing the likely causal agent in promoting disease. It is followed up, though, by a detailed assessment of the factors leading up to the numerical ranking, in a subsequent consultation with the rule base.

The expert system employed *induction* on the data to determine presence of conditions of both 'good' and 'bad' types which might cause sickness. That is, a collection of examples are generalized into a rule base which satisfactorily covers the complete collection. The induction process is carried out by the so-called 'inference engine' part of the lstCLASS system. This system uses the forward chaining process. This is a decision tree capable of explaining the outcomes of the examples which were used in building it. Inconsistencies in outcomes are flagged during the compilation process. If some branches in the tree are unfinished, other possible causes have to be included to complete the explanation facility. We used this occasion to hypothesize the existence of other sorts of health risks (such as alternate disease sources or misstatements during the interviews). During this inductive analysis, the data set was reduced in size. Well/latrine/usage configurations with no coliforms were excluded from rule formulation.

When a particular well configuration was an uncertain source of disease organisms, we tried to incorporate this uncertainty into our initial screening. The following example is illustrative. If the data indicated that in 30% of those configurations the presence of coliforms was observed, an inference engine might offer a measure of certainty of 30%. The size of sample was thought to be too small to assign accurate numerical values to heuristic statements. As well, the presence of numerical weights and certainty factors in the dialogue was felt to lead to unnecessarily complex interactions with end-users. (For systematic long-term studies, certainty and/or probability estimates could be incorporated for trained health personnel less likely to misinterpret numerical results.) Also, the induction facility of IstCLASS (at that time, at least) did not handle uncertain or conflicting evidence in its rule building. Were we to give two instances of a single well, one with coliforms and one without, we would get a conclusion 'either the well has or has not shown presence of coliforms'. Since our system is a simple classifier system, we are concerned with instances of coliform reinforcing our initial conjecture of a problem with that configuration.

It is partly the presence of uncertainty which inspired the two phase design of the expert system. Phase I was used to collect the data for a well site and assign a rank. These ranks were then used to determine the area of focus for phase II. If a well ranked GOOD in the well category and POOR in the latrine category, data would be collected with respect to the latrine, and the system will branch to the LATRINE advisor.

3. Classification Algorithm

It was necessary to segregate the survey data since, initially, it was in the form of strings (Figure 3). Once the data strings were separated, values were assigned to each indicator. These values were then used as the weights in the calculation of the well rank, latrine rank, and total rank. With the data prepared, each well was ranked based on the weighing

Fig. 3. Survey data format.

scheme described above. These ranks were then classified and adjusted qualitatively by the domain experts (Table I).

3.1. THE EXPERT ADVISOR

Using the factors expected to contribute to well contamination (as identified by the domain experts), rule bases were developed to trigger advice. The first rule base assesses the well characteristics (e.g.: too close to the latrine, close to agricultural activity...) and the second rule base tests the properties of the latrine (e.g.: poor technology, too many families, etc.). This prototype system was designed with the intention of providing field personnel with an encapsulated summary of well characteristics and possible problem sources. Analysis is done on a well by well basis. A session begins (Figure 4) by the user entering a language choice (the rulebases were converted to other languages due to the variability of languages spoken by the intended audience).

The user must then select a well number for analysis, and an initial ranking is performed using the criteria shown in Figure 5. Using ranks as indicators (i.e. over 2 (fair)) the system will call either the well advisor, or the latrine advisor, or both. The advisors will then scan the database, looking for problematic conditions. This will trigger the generation of remedial advice (Figure 6), based on any problems discovered. The basic

TABLE I

Fig. 4. Language choice screen.

functions of the rule bases are (1) to assess various aspects about the usage of the well, and (2) to analyze any relationships between the latrine and the well (e.g.: latrine location relative to the well, etc.).

In the final stages of the development, the rule base maintenance was performed almost entirely by a domain expert. In this way, a scientist can experiment freely with rudimentary knowledge representation schemes. This has proved to be very beneficial since slight changes could be made to the rule base, and the effect realized immediately.

The data structure for the inference engine allowed us to assign a narrative to each rule, so that when they fire a meaningful narrative is presented. For example:

IF <factor1>AND<factor2>AND<factor3>... THEN <RESULT>

The result is

IThe latrine is located less than 15 metres from the water source. It **is** ___advisable that the latrine be located at least 15 metees from the **source** HAs well, the latrine is located uphill from the water source. Ideally, the
|latrine should be located downhill from or level with the water to avoid **~conta~ination**

Les conseils qui suivent concernent l'usage des latrines ...

Le nombre de maisons qui utilisnet les latrines est acceptable
Cependant, on baigne aux endroits où on trouve de l'eau à cet emplacement.
Ca peut causer la contamination du puit. Conseiller en conséquence

is paraphrased as 'Is \langle factor \rangle present?'. In the same fashion the \langle result \rangle field is explained as to why it can contribute to a conclusion. IF cracked_casing THEN wastewater₋ intrusion is then replaced by a more natural 'Is the casing cracked...' ... 'Waste water may be returned to the system'.

The system was developed in four languages, with no real impact on the complexity of the design. A language choice query determines the language preference at the beginning of the session, and the appropriate textual database was swapped in. The logic of the

Fig. 7. Regression analysis (log_{10} mfc vs. total rank).

advisor is language independent, and therefore could support any number of languages desired. There is, however, a maintenance problem, since any change must be propagated to all the databases. The dialogue and the responses are available within the system in English, Bahasa Malaysia and Mandarin (pin-yin) and French (see Appendix B). The lstCLASS shell offers us this facility almost for free. We merely load an alternate text file containing the responses in the alternate language, as part of an initial query. This parallel dialogue, although consisting of 'canned text' segments, could provide an interesting example of computer-supported cooperative work in multilingual societies where language barriers might otherwise handicap widespread use of information systems.

With all the functional parts of the system completed, it was then possible integrate the geographical information system (G.I.S.) with the both the information system and the expert advisor. The result was a system which allowed the user the option of either entering a well number (each well in the system was assigned an alphanumeric code) or pointing to a well on the map for advice. (Figure 8)

One potential major use of this system, other than diagnosis, comes from the computer system which was used for its implementation. The RAISON system, developed at National Water Research Institute, has its own generic database, spreadsheet and programming shell. It also has a built-in geographical information system. When this diagnostic tool is used in the field, maps of water supply and sanitation facilities are continuously collected and updated. Proximity of disease and possible types of disease sources are available on short notice to field workers. This extra geographical component, which is absent from many commercial packages, gives fast visual feedback to operators.

Stell): Use cursor keys to point to well

Fig. 8. Expert advisor integrated into G.I.S.

Also, parts of RAISON have been constructed for classification and display of data and testing hypotheses (concusions). Uses of RAISON spreadsheet and database formats which are compatible with many commercial varieties offer acceptable data interchange formats for transmission of data throughout governmental public health structures.

4. Conclusions

This system uses the data gathered to elicit from a diagnostic expert the likely causes of well-water contamination in a medium-sized sample of wells and sanitation facilities. Conclusions are based on real data and a systematic abstraction from these observations. Response is fast and the diagnosis is simple and unequivocal, partially because the query is organized in two tiers. Outcomes are completely specified by the induction facility, although the explanation 'another source of contamination is suspected' is a useful caveat. Analysis, although rudimentary, is accurate and capable of providing remedial advice. Delivery of advice is possible in any of several languages. The geography of the situation offers possible cues for corrective procedures. Maintainability and upgradeability are facilitated through simple text-based communication using a commercial shell.

Interaction with end-users is in text form. It carries with it the 'commonsense' built into the initial analysis. Its dialogue is less technical than a statistical interpretation, and it is tailored to a specific situation. It represents a constructive 'proof of concept' which generates plausible explanations in an important problem domain. Coverage is complete, since all possible prevention steps will be enumerated for a substandard well/sanitation configuration.

APPENDIX A: WATER SUPPLY QUESTIONNAIRE

APPENDIX B: SAMPLE ADVISOR SESSIONS

The dialogue exhibited concerns a well which was already in the training dataset, and is provided for exemplary purposes.

The respondent is asked by the system to choose mother tongue: English, Bahasa Malaysia, or Mandarin (pin-yin). A well is selected by pointing to it on the map.

English Dialogue

Water Quality Advisor

Steps:

1) Use cursor to point to well

2) Press *<i>* to identify the well

3) Press <Enter> to analyse

You picked well STGH3

The Well Class is : Excellent The Latrine class is : Poor The Overall Rating Class is : Fair The Overall rating for the well is good...checking the Latrine... The Latrine Class is unacceptable...calling the advisors...

Advice:

The latrine is located less than 15 metres from the water source. It is advisable that the latrine be at least 15 metres from the source

The Latrine is situated uphill from the water source. Ideally, the latrine should be downhill from, or level with the water soume

and in

The latrine is neither the flush nor bucket type. Either is an improvement on the current system

The following advice applies to the usage of the latrine ...

It is possible that the number of households using the latrine exceeds capacity ... this should he investigated

As well, bathing at the water point is carried out at this location. Such practices can lead to well contamination...advise accordingly.

Bahasa Malaysia Dialogue

PENASIHAT KUALITI AIR

Kaedah:

1) Guna kunci kursot untuk menunjukkan perigi

2) Tekam <i> untuk menentukan perigi

3) Tekam <Enter> untuk menganalisikan perigi

Anda memilih perigi <STCH3>

Kelas perigi adalah terbaik

Kelas tandas adalah buruh

Kelas semula adalah sederhana

Kelas semula untuk perigi adalah baik...memeriksakan tandas....kelas tandas adalah burah.,..memanggil penasihat....

Nasihat-nasihat:

Tandas adalah ditempatkan kurang 15 meter dari sumber air. Dinasihatkan bahawa tandas dibinakan sekurangkurangnya 15 meter dari sumber air.

 \sim

Tandas adalah diletakkan di tempat tinggi dari sumber air. Biasanya, tandas mesti dibinakan di tempat rata atau rendah dari sumber air.

 $\overline{}$

Jenis tandas - bukan jenis cucur/curah atau tong. Kedua-dua jenis tandas adalah lebih baik dibandingkan dengan jenis tandas sekarang.

 \sim

Nasihat-nasihat yang berikut dipakai kepada penggunapengguna tandas.....

Jumlah pengguna-pengguna tandas melebihi sangat ... keadaan itu mesti diperiksakan.

 \overline{a}

Juga, pengguna-pengguna air mandi di tempat sumber air, Aktiviti itu mungkin mengkotorkan perigi....manasihatkan demikian.

Mandarin (pin-yin) Dialogue

Shui zhi gu **wen**

Fang **fa:**

(1) Yong yi dong jian (-5) zhi zhuo jing. (2) An \leq *lai biao ji zhe ge jing.*

(3) An <ENTER> iai fen xi zhe ge jing.

Hi suo xuan zhe de jing [STGH3] Jing de deng ji: **Hen hao Ce suo de deng ji: Cha** Zong ti ping jia den ji: pu tong Jing de zeng ti ping jia ban hao..Jian cha ce suo... Ce suo deng ji bu neng bei jie shou...qing jao gu wen...

Jian yi:

Ce suo jian zai liao li shui yuan bu dao 15 mi de di fang. Jian yi shi ce suo ying gai jian zai ii shui yuan 15 mi yi wai. _..

Ce suo jian zai liao shui yuan de shang po.

Li xiang shang, ce suo ying gai jian zai shui yuan xia po **huo**

yu shui yuan ping xing zhi chu.

Ce suo bu shi la shui huo ce tong shi.

La shui huo ce tong shi bi xian you de yiao hao.

Yi xia de jian yi shi guan yu ce suo de shi yong...

You ke neng jia ting ren shu yi jing chao guo...ying gai diao cha.

Ye you ke neng, xi zao chu zai shui yuan huo gong shui chu de fu jin. Zhe zhong xi guan ke neng dao zhi jing de wu ran...qing zuo ge hie **de tao** lun.

References

Buchanan, B. G. and Shortliffe, E. H. (eds.): 1984, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA.

Harmon, P. and King, D.: 1985, *Expert Systems,* John Wiley & Sons, New York.

Lam, D. C. L., Swayne, D. A., and Storey, J.: 1990, 'An Introduction to the RAISON System', in: *The Use of Simple Inexpensive Microbial Water Quality Tests: Results of a Three Continent, Eight Country Research Project,* IDRC-MR247e, B.J. Dutka and A. H. El-Shaarawi (eds,), pp. 144-155.

McDermott, J.: 1982, 'RI: A rule-based configurer of computer systems', *Artificiallntelligence,* 19, 39-88.

Swayne, D.A., Storey, J., Fraser, A.S. and Lam, D.C.L.: 1989 'A Microcomputer-Based System for Environmental Modeling', *Microcomputer Applications* 8 (2), 35-38.

Tong, S. L., Wang, C. W., Chin, S. T., Loh, C. L., Lim, C. K., and Ho, P. Y. C.: 1990, 'RAISON-MALAYSIA: A User-Friendly Package for Water Quality Data Management', In: *The Use of Simple, Inexpensive Microbial Water Quality Tests: Results of a Three Continent, Eight Country Research Project,* I DRC-MR247e, B.J. Dutka and A. H. EI-Shaarawi (eds.), pp. 156-183.

Trademark

1stCLASS^(tm) is a registered trademark of AI Corporation, MA.