



Environmental monitoring, modeling, and management and Bayesian Belief Networks

M. Kemblowski¹ & P.C. Johnson²

¹*Utah State University*, ²*Arizona State University, USA*

Abstract

The paper addresses the problem of rational environmental monitoring, modeling, and management (Em^3), and associated with it decision-making processes that are characterized by a high degree of uncertainty with respect to the input data (such as contaminant loads), fate and transport parameters (such as degradation rates), process models, and connections between such models. We propose to deal with these complex issues in the framework of probabilistic expert systems, and more specifically, Bayesian Belief Networks (BBNs). We discuss the applicability of BBNs to Em^3 via a simple example related to groundwater monitoring. This is followed by an application of BBNs to a forensic subsurface contamination problem.

1. Introduction

In this paper, we would like to focus on the interplay among environmental monitoring, modeling, and management (Em^3), with particular emphasis on subsurface contamination. The environmental community has been working on subsurface contamination assessment and cleanup procedures for over two decades. The physical, chemical, and biological principles of the relevant fate and transport processes and cleanup technologies (such as pump and treat, venting, bioventing, thermal venting, air sparging, enhanced biodegradation, and co-solvent flushing) have been described in the literature, and are reasonably well understood. One could argue that there is always room for improvement in our understanding, and we do not disagree with this argument. Our professional experience tells us, however, that even when the principles are well understood,



134 *Computer Techniques in Environmental Studies*

the practical applications are frequently flawed and characterized by the lack of rigorous methodology for site assessment and cleanup. In particular, we have often encountered the following problems:

2. Monitoring by checklist. Most of the time, the one-size-fits-all philosophy is used, i.e., the assessment monitoring parameters do not reflect specific site features, and there is no clear connection between these activities and specific environmental/health objectives.
3. Engineering by Rules of Thumb. One famous example is “one upgradient, three downgradient wells”. These Rules of Thumb are clearly easy to enforce and report, but there is no mechanism, or effort, to demonstrate that their fulfillment will actually effectively help in reaching environmental objectives.
4. Lack of logical decision framework. It is very seldom that we see an effort to anticipate the decision pathways based on the future assessment or cleanup performance data. The typical practice is that data is first collected, and then one tries to figure out what to do with it.
5. Lack of data analysis/evaluation framework. Even the collected data are not properly analyzed and/or visualized, since it is generally not known how, and particularly why to use it.
6. Lack of modeling framework. Various models are used to simulate water flow and contaminant transport in the subsurface without explicit, or even implicit, consideration given to the true modeling objectives. Furthermore, the model choice is more often based on its availability or familiarity rather than on the project objectives.
7. Lack of methodology for assessing the value of information.

As a result of these problems, in the realm of Em^3 , we are faced with significant waste of resources. We believe that there is a strong need for a framework that would enable us to address *simultaneously* the following questions:

- Why, what, where, when, and how to monitor?
- What models to use?
- When are sufficient data for decision-making available in the context of the environmental objective and site specifics?
- How does one evaluate the value of information?
- How should cleanup efficiency be evaluated?
- When can you terminate subsurface mitigation?

In this paper, we approach to some of these problems using the general framework of Bayesian Belief Networks (BBNs).

2. Bayesian belief networks in the context of Em^3

We consider BBNs from the perspective of probabilistic expert systems, utilized in this study for the purpose of rational Em^3 . The aim of expert systems is to provide a computational model of intelligent behavior, most importantly, commonsense reasoning. The aim of probability theory is to provide a coherent

account of how belief should change in light of partial or uncertain information. Since commonsense reasoning always applies to incomplete information, one might naturally expect the two disciplines to share language, goals, and techniques.

It is important to remember that probability is not really about numbers but about the structure of reasoning. We emphasize, for example, that when an expert asserts, "The chances that Site A with leaking tanks will contaminate groundwater (symptom S) is p ," the thrust of the assertion is not the precise magnitude of p so much as the specific reason for the expert's belief, the context or assumptions under which the belief should be firmly held, and the sources of information that would cause this belief to change. Probability theory, which is the theoretical foundation of BBNs, is unique in its ability to process context-sensitive beliefs, and what makes the processing computationally feasible is that the information needed for specifying context dependencies can be represented by graphs and manipulated by local propagation.

2.1. Definition of BBNs

Bayesian networks (Pearl, 1988) are graphical, probabilistic models that allow the structured representation of a cognitive or decision process. We will introduce the concept via a simple example of environmental monitoring.

There are two known potential sources of groundwater contamination. One of them, Source I, is believed to be more likely to leak than the other one (Source II). Specifically, the probability of Source I leaking is $P(\text{Source I} = \text{yes}) = 0.4$, whereas $P(\text{Source II} = \text{yes}) = 0.3$. These two sources are initially monitored by a single monitoring well, Well I. In order to estimate the probability of Source I or II leaking given that the contaminant is detected in Well I, we first need to estimate the local conditional probability distribution (CPD) of detecting the contamination in Well I, conditioned on the behavior of Sources I and II. This CPD may be estimated in many ways, for example by using expert judgement or

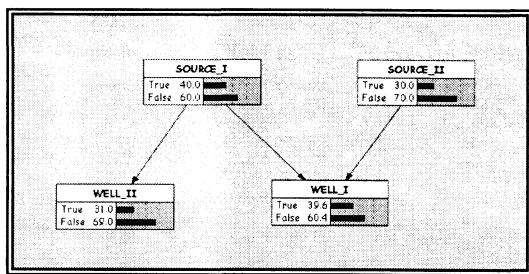


Figure 1. Initial forward probability propagation

We note that $P(\text{Well I} = \text{no} \mid \text{Source I}, \text{Source II}) = 1 - P(\text{Well I} = \text{yes} \mid \text{Source I}, \text{Source II})$. At this point, we can calculate a number of important statistics. Firstly, using the independence of Source I and Source II, we can now calculate the prior joint probability $P(\text{Well I}, \text{Source I}, \text{Source II})$, where:

a Monte-Carlo model similar to the one described by A. Warrick et al. (1998). The Monte-Carlo approach was used in this case, and the resulting CPD for observing contamination in Well I, i.e., $P(\text{well I} = \text{yes} \mid \text{Source I}, \text{Source II})$ is given in Table 1.



136 Computer Techniques in Environmental Studies

$$P(\text{Well I, Source I, Source II}) = P(\text{Well I} \mid \text{Source I, Source II}) P(\text{Source I}) \\ P(\text{Source II})$$

Table 1. Conditional probability $P(\text{Well I} = \text{yes} \mid \text{Source I, Source II})$.

	Source I = yes	Source I = no
Source II = yes	0.8	0.5
Source II = no	0.6	0.1

Secondly, by marginalizing Source I and Source II out of $P(\text{Well I, Source I, Source II})$, we can estimate $P(\text{Well I}) = (0.396, 0.604)$ (see Figure 1). Introducing the new evidence (Well I = yes), and back-propagating this evidence via the

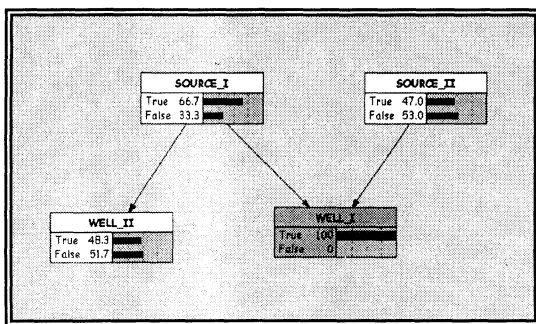


Figure 2. Propagation of Well I = yes

Bayesian network, leads to new posterior probabilities $P^*(\text{Source I} = \text{yes} \mid \text{Well I} = \text{yes}) = 0.667$ (compared to the prior $P(\text{Source I} = \text{yes}) = 0.4$), and $P^*(\text{Source II} = \text{yes} \mid \text{Well I} = \text{yes}) = 0.47$ (compared to 0.3). The new evidence has appreciably changed the probabilities of the potential causes.

At this point one might want to collect more data in order to better distinguish between Source I and Source II. Let us assume that we can install another monitoring well, Well II. This additional well is insensitive to Source II. The conditional probability $P(\text{Well II} \mid \text{Source I})$, where Well II = yes indicates that contamination has been detected in Well II, is given in Table 2.

Table 2. Conditional probability $P(\text{Well II} \mid \text{Source I})$

	Source I = yes	Source II = no
Well II = yes	0.7	0.05
Well II = no	0.3	0.95

We thus have an option of installing and monitoring Well II. Whether we will or will not do that depends on the cost of the information, as well as the cost of any actions that we may take based on the probabilities of Source I and Source II. This decision-making problem is well addressed within the framework of the statistical decision analysis combined with utility theory.

In order to use these approaches, however, one needs to be able to estimate the posterior marginal probabilities $P^{**}(\text{Source I} \mid \text{Well I, Well II})$ and $P^{**}(\text{Source$



II | Well I, Well II). This is again the time when Bayesian nets provide the answer. To demonstrate it, let us assume that Well II = no. Introducing this new

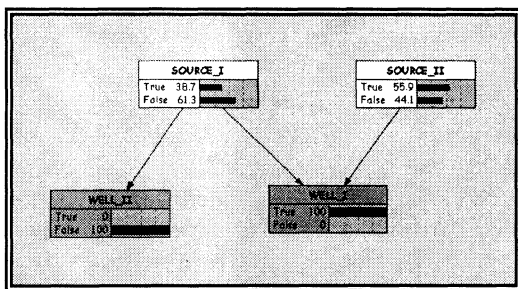


Figure 3. Propagation of Well II = no

evidence to the Bayesian Network depicted in Figure 3 leads to the following estimates of new posterior probabilities: $P^{**}(\text{Source I} = \text{yes} \mid \text{Well I} = \text{yes}, \text{Well II} = \text{no}) = 0.387$, and $P^{**}(\text{Source II} = \text{yes} \mid \text{Well I} = \text{yes}, \text{Well II} = \text{no}) = 0.559$.

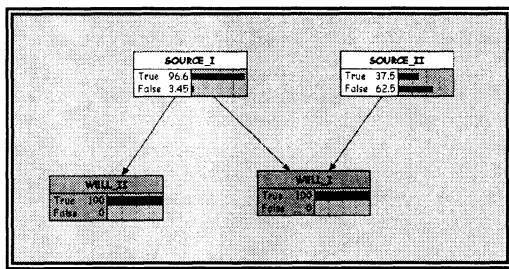


Figure 4. Propagation of Well II = yes

The probabilities $P^{**}(\text{Source I} \mid \text{Well I} = \text{yes}, \text{Well II} = \text{yes}) = 0.966$, and $P^{**}(\text{Source II} \mid \text{Well I} = \text{yes}, \text{Well II} = \text{yes}) = 0.375$, can be estimated in a similar manner by back-propagation of evidence, and are given in Figure 4.

A Bayesian network consists of the following components:

- A set of variables and a set of directed edges between variables.
- Each variable has a finite set of mutually exclusive states.
- The variables together with the directed edges form a directed acyclic graph (DAG). (A directed graph is acyclic if there is no directed path $A_1 \rightarrow \dots \rightarrow A_N$ such that $A_N = A_1$)
- To each variable A with parents $B_1, \dots, B_N$ there is attached a conditional probability table $P(A \mid B_1, \dots, B_N)$.

2.2. Building models

Bayesian networks create a very efficient language for building models of domains with intrinsic uncertainty. However, it is a tedious job to perform evidence transmission even for very simple Bayesian networks. Fortunately, software tools, which can do the calculations for us, are available. Several commercial products exist that contain both an editor for Bayesian networks and a runtime module that takes care of evidence transmission. Therefore we can focus on the practically most important part of environmental probabilistic expert system development: how to use Bayesian networks in model building.



2.2.1. Catching the structure - qualitative modeling

The first thing to keep in mind when organizing a Bayesian model for a decision support system is that its purpose is to give estimates of certainties for events which are *not observable* (or only observable at an unacceptable cost). Thus, the initial task in model building is to identify these events. We call them hypothesis events. In the simple example discussed above, the hypothesis events are the states of potential Sources I and II.

In order to come up with probability estimates for the hypothesis events, we need to provide some information channels. Thus, the task is to identify the types of (potentially) available information (data and models) which may reveal something about the state of a hypothesis variable. Typically, the statement will be that a particular information variable is in a particular state, but soft (probabilistic) statements are also allowed. In the example, the models are used to estimate the CPDs for monitoring wells, and the monitoring data are subsequently used to estimate the probabilities of Source states.

Part of the network structure development is the definition of conditional independence among the variables. If all the variables were directly dependent, the propagation of evidence in the network would be computationally very intensive. It is the conditional independence between the variables that allows us to significantly simplify the evidence propagation, as well as the process of estimating the input conditional probabilities.

The variables A and B are independent given variable C if

$$P(a_i|c_j) = P(a_i|c_j, b_k) \text{ for all } i, j, \text{ and } k.$$

This means that if the state of C is known, than no knowledge of B will alter the probability of A.

2.2.2. Estimation of Conditional Probabilities

The basic concept in the Bayesian treatment of uncertainties in causal networks is conditional probability. Whenever a statement of the probability, $P(A)$, of an event A is given, it is given conditioned by other known factors. These factors may be explicit or implicit. If the factors are presented in a BBN as the parent nodes of node A, we need to specify the probability of A conditioned on the states of the parent nodes, $p(A|P_i)$. Sometimes the conditioning factors are not included explicitly in the network; in this case the probability of A reflects their implicit existence. In the example, the CPD for Well I includes the following: $P(\text{Well I} = \text{yes} \mid \text{Source I} = \text{no}, \text{Source II} = \text{no}) = 0.1$. This conditional probability value implicitly includes our knowledge of potential contamination sources, other than Source I and Source II, in the area, although these sources do not appear explicitly in the model.

The conditional probabilities can be estimated via the frequentist approach, i.e., using the available models and data to estimate $p(A|P_i)$, or by utilizing the expert's belief. The frequentist approach may be direct or indirect. The direct approach utilizes the database of events that include the states of a given node and its parents to estimate the frequency of occurrence for any combination of

states. The indirect approach uses Monte-Carlo simulations of the parent nodes with the parent-child link models to come up with the child conditional probabilities.

In addition to the Monte-Carlo estimations of conditional probabilities, we can also use subjective (expert) estimations whenever the Monte-Carlo estimations are not practical, either due to the complexity of the link models, or to the lack of link model input data. It is well known that people are notoriously bad processors of numerical, and particularly statistical, information. They find it easier to approximately assess local, conditional probabilities than to process this information to estimate marginal probabilities. This is particularly evident when backward propagation is called for. Still, that does not preclude the use of the laws of statistics, along with local conditional probability estimates, when it comes to deciding which state of a hypothesis variable is more probable, and how this may change if new information becomes available. The absolute estimates of the local conditional probabilities can be off, but as long as they reflect genuine experience, the global conclusions will still be useful. In other words, if we strongly believe in the rules by which exact quantities combine, we can use the same combination rules on the rough estimates at hand.

2.3. Evidence propagation.

After the model has been built, we need a computational engine for the initial and additional evidence propagation. The methods for forward and backward evidence propagation are well presented in the literature (for example in Castillo et al., 1997). Several commercial products exist that contain both an editor for Bayesian networks and a runtime module which takes care of evidence propagation, such as: Netica (for Windows and Macintosh, <http://www.norsys.com>), HUGIN (for Windows, <http://www.hugin.dk>), and X-Pert Networks (for Macintosh, <http://ccaix3.unican.es/~AIGroup>).

3. Forensic application of BBNs

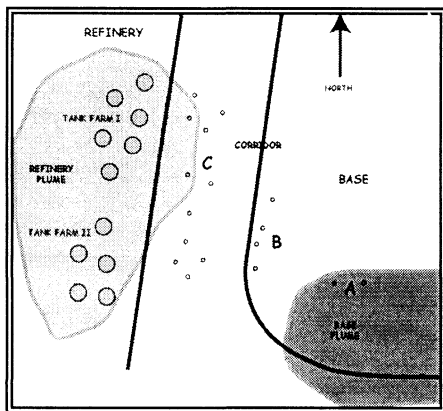


Figure 5. Schematic site map

The authors were asked to review existing data and suggest an additional data collection program to determine if the separate phase plume floating on the water in the southern portion of a military base has originated from a refinery located west of the base (Figure 5). The site is located on the East Coast of the U.S.A. In this section, due to the shortage of space, we present a cursory description of the application of BBNs to this problem.

140 *Computer Techniques in Environmental Studies*

The base has been in use since the 1800's, although the official construction date of the facility is 1917. It has primarily served as a central facility that prepares, stockpiles, and distributes equipment and supplies to all branches of the military. It appears that the facility was at the peak of its activities during World War II. From the early 1940's until at least 1960 the primary activity was clothing manufacturing and treatment. The facility also stored and distributed gasoline and diesel fuel at a number of locations beginning in the 1920's, and perhaps earlier than that. In 1987 the Army Corps of Engineers discovered NAPL on the water table in the vicinity of a base gas station. Subsequent investigations showed the NAPL lens was widespread underneath the southern portion of the facility.

The refinery was constructed and rendered operational in 1879. Initially, the operational units were located in the western portion of the site. It is understood, and accepted by the refinery management, that there have been spills and leaks at the refinery since the beginning of the operations. The earliest records of subsurface hydrocarbon recovery operations date back to the 1930's. These operations were located in the eastern portion of Tank Farm #2. In the 1960's,

hydrocarbon recovery operations were initiated along the eastern side of Tank Farm #1. These operations were updated in the 1980's, and continue today. The question was: what is the likelihood that the significant portion of the base plume has migrated via subsurface from the refinery. In order to address this question, we first looked at the transport potential between the refinery and the base.

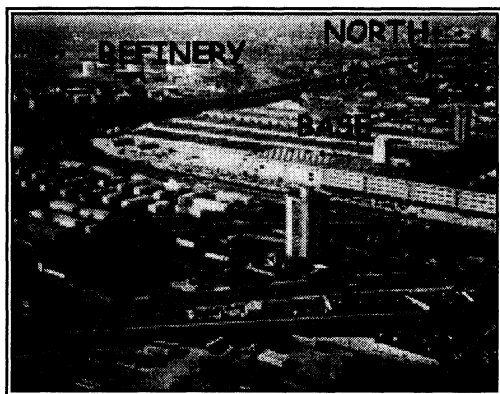


Figure 6. Site view

3.1. Initial data analysis

We have initially considered a number of historical data types and sources, including the following major data areas:

1. Hydrologic and geologic data
2. NAPL character and spatial distribution

3.1.1. Hydrologic and geologic data

The two parties (the base and the refinery) have spent a considerable effort attempting to reconstruct the historical groundwater flow regime. The principal goal of this effort was to determine whether there was a period of time in the past during which the groundwater flow in the upper aquifer could potentially facilitate the transport of separate phase plume from the refinery towards the southern aspect of the base.



In order for this transport process to have taken place, the groundwater flow direction would have to have a strong easterly component. The data from the most recent years show the general flow direction from N to S, with local easterly and westerly components near recovery wells and sewers. The historical data on groundwater regime in the area is rather sparse and inconsistent in terms of temporal and spatial scales. The analysis of the data presented by the two parties, as well as our review of available historical information, did not exclude the possibility of groundwater flow in the E to SE direction in the past. This is reflected in the initial marginal probability of the flow direction (Figure 7).

Based on existing information, it is apparent that the aquifer, due to its high permeability, is theoretically capable of transporting the NAPL plume from the refinery to the base, providing that the source and appropriate groundwater flow conditions exist. This fact is reflected in the high initial marginal probability assigned to high permeability (Figure 7).

3.1.2. Spatial distribution of NAPL

The spatial distribution of NAPL plume has not changed appreciably over the monitoring time. Considering the high permeability of the aquifer, this relative immobility of the plume indicates that the plume has been somehow stabilized. The plume is currently elongated in the direction perpendicular (E-W) to the groundwater flow direction (N-S). This plume shape indicates that other sources of separate phase have likely contributed to the base plume.

3.1.3. Initial BBN

The initial BBN is depicted in Figure 7. This BBN assumes the presence of

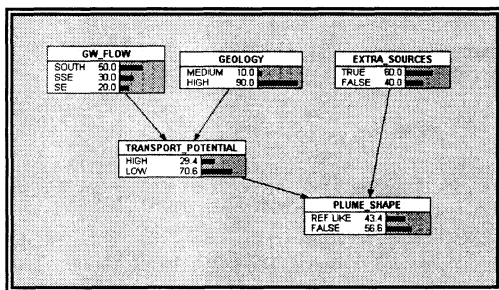


Figure 7. Initial BBN

other sources of contamination on the base and outside the base with $P = 60\%$. Without using the plume shape evidence we see that the probability of the high potential for transport between the refinery and the base is approximately 30%. After introducing the plume shape evidence, i.e.,

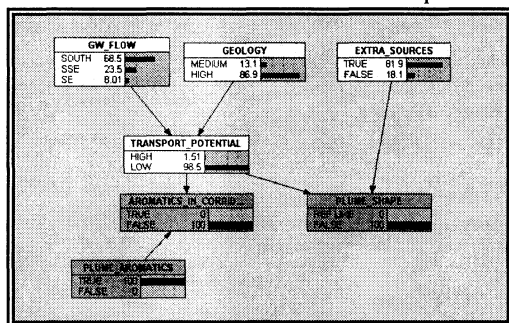
recognizing that the plume shape indicates the contribution of other sources, the high transport potential probability decreased to some 18%. It was felt that this probability was too high to entirely exclude the refinery as a significant contributor to the base plume. A data collection program, focused on groundwater and soil chemical data, was designed to further reduce the information entropy.



142 *Computer Techniques in Environmental Studies*

3.2. Additional data

Additional groundwater and soil samples were collected at a number of multi-level sampling points (see sampling point clusters A, B, and C, Figure 5). It was determined that the water and soil samples located near the base plume (Cluster



A) were characterized by a relatively high level of aromatic compounds. Introduction and propagation of this evidence in the BBN resulted in a relatively high probability of high level of aromatics in the corridor samples (Clusters B and C, outside the refinery plume), $P = 0.42$.

Figure 8. Final Bayesian Belief Network

At this point the high transport potential probability was still about 18%. The field data showed, however, quite low level of aromatics in the corridor samples. The propagation of this evidence via the network resulted in the posterior probability for high transport potential between the refinery and the base $P^* = 1.5\%$. This practically removed the refinery from the list of significant contributors to the base plume.

4. Conclusions

The potential for using Bayesian Belief Networks in Em^3 was explored. Two examples of applying BBNs to environmental problems were presented. The discussion and the examples demonstrate that BBNs are a powerful and promising tool for addressing environmental issues.

5. References

- [1] Castillo, E., J. Gutierrez, and A. Hadi, 1997, *Expert Systems and Probabilistic Network Models*, Springer, New York.
- [2] Cowell, R., P. Dawid, S. Lauritzen, and D. Spiegelhalter, 1999, *Probabilistic Networks and Expert Systems*, Springer, New York.
- [3] Jensen, F. V., 1998, *An introduction to Bayesian Networks*, Springer, New York.
- [4] Pearl, J., 1988, *Probabilistic reasoning in intelligent systems*, Morgan Kaufman Publisher, San Francisco, CA.
- [5] Warrick, A. W., M. H. Young, and P. J. Wierenga, 1998, Probabilistic analysis of monitoring systems for detecting subsurface contaminant plumes, *Ground Water*, Vol. 36, No. 6, pp. 894-900.