
Probabilistic versus Regression Modeling for Disinfection Byproducts

Zoe J. Y. Zhu and Edward A. McBean

Probabilistic network approaches, including Bayesian networks (BN) and decomposable Markov networks (DMN) are graphical models in which a problem is structured as a set of parameters and probabilistic relationships between them. Probabilistic analyses have been effectively used to incorporate expert knowledge and historical data for revising the prior belief in the light of new evidence. In this chapter, the probabilistic approach is compared to traditional methods such as regression analysis for water quality predictions. The capabilities and advantages/disadvantages of DMN approach are described. A DMN through machine learning on the basis of historical data from the experiments is constructed. The results indicate that DMN is a better prediction model than multiple regression both theoretically and experimentally for these applications.

24.1 Introduction

Drinking water utilities face a challenge in recognizing, characterizing, and responding to potential and actual contamination events, while simultaneously minimizing the formation of disinfection byproducts (DBPs) which could result in adverse human health and safety impacts. Because of the major benefits of water disinfection and due to outcomes associated with DBPs, a risk trade-off analysis between microbial and chemical risks becomes necessary. However, in practice, conditions leading to better

disinfection efficiency also lead to higher occurrence of DBPs. The regulatory regime must establish the acceptable levels of risk for both microbial and chemical agents (Sadiq, 2004). Real-time analytical tools for characterizing many of the contaminants and DBPs in-situ do not currently exist, or are prohibitively expensive (ASCE, 2004). In this response, we design herein a monitoring system using surrogate water quality measures such as total organic carbon (TOC), turbidity, pH, chlorine concentration and others which may be correlated to “fingerprint” disinfection product formation. A DMN and machine-learning are used with the results compared with the traditional multiple regression method.

Probabilistic networks are able to assist in building practical systems capable of handling uncertain information. For example, Pearl (1988) published a book on Bayesian networks followed by Castillo et al. (1997); Jensen (1996); Neapolitan (1990) and others. The result is a rich array of high level publications wherein probabilistic networks are cross-disciplinary, ranging from natural sciences to social sciences, from physics to engineering and space, from environmental to medical applications and from ecological modeling to water quality prediction. For example, Siber et al. (1999) developed Bayesian networks coupling an expert knowledgebase with process models to evaluate the potential of naturally occurring reductive dechlorination at sites contaminated with TCE (trichloroethylene). Marcot (2001) combined expert knowledge with ecological data within a Bayesian network to model the causal relationships between planning decisions and impacts on at-risk wildlife species habitats. Stow et al. (2003) compared a BN approach with two deterministic models for predicting the effect of nitrogen loading on estuarine chlorophyll ‘a’ concentrations. Substantial research in the environmental engineering domain has utilized Bayesian approaches for a number of applications, but few studies have utilized DMN and machine learning specifically as illustrated here.

24.2 Decomposable Markov Network and Machine Learning

Decomposable Markov networks, as per Figure 24.1 involve characterization of each node in the network corresponding to a domain variable, and each link identifying a probabilistic dependence or correlation between the corresponding variables. In Figure 24.1, 'a' and 'b' are two causes of 'c' and are not themselves directly connected. This means that 'a' and 'b' are normally independent. However, when 'c' or 'd' is observed, 'a'

and 'b' compete to explain the observations. Hence 'a' and 'b' becomes dependent when 'c' or 'd' is observed. Such dependence is referred to as induced dependence, (Pearl, 1988, 1991).

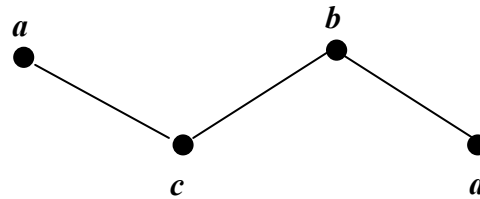


Figure 24.1 An example of Markov Network.

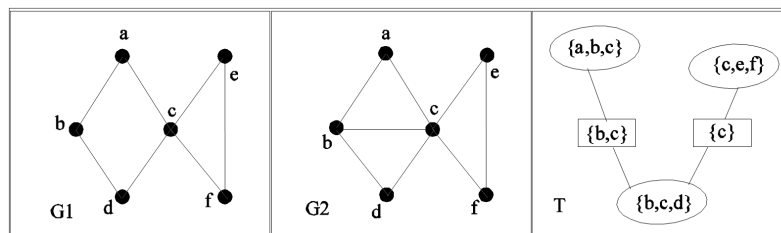


Figure 24.2 G1
A Non-Chordal Graph.

Figure 24.3 G2
A Chordal Graph.

Figure 24.4 T
A Junction Tree.

A set of nodes is complete in a graph if they are pairwise connected. For example, G1 in Figure 24.2, the set of nodes {c, e, f} is complete, and so is the set {b, d}. However, the set {a, b, c} is not complete since 'b' and 'c' are not connected. A (maximum) set of nodes which are pairwise-linked (complete) is called a clique. G1 in Figure 24.2, contains five cliques, they are {a, b}, {b, d}, {a, c}, {c, d}, and {c, e, f}.

A path in a network is a sequence of nodes such that there is a link in the graph between each pair of adjacent nodes in the sequence. A cycle is a path where the first node is identical to the last node. A path in an undirected network has a chord if there is a link between two nonadjacent nodes on the

path. For example, the path (a, b, d, c, a) in G2 has a chord {b, c}. A graph is chordal if every cycle of length > 3 has a chord. Graph G1 is not chordal because the path (a, b, d, c, a) of length 4 has no chord. If we add the chord (b, c) to G1, it becomes G2 which is a chordal graph as per Figure 24.3. The graphical structure of a Decomposable Markov graph is a chordal graph.

Let G be a connected chordal graph. A junction tree (JT) can be derived from the chordal graph, G, where JT is a tree whose nodes are labeled by cliques of G. The tree is organized such that for each pair of cliques in the tree, their intersection is contained in every clique on the path between them. In a junction tree, each link is labeled by the intersection of the two cliques being connected, and is called a separator. A connected graph G has a JT if and only if (iff) G is chordal. Graph G2 has 3 cliques. They are organized into a tree in Graph T as per Figure 24.4, where each clique is shown as a large oval. Each separator is shown as a box. The tree is a JT because, for example, the intersection of clique {c, e, f} and clique {a, b, c} is {c} which is contained in the clique {b, c, d} between them.

Consider now the task of specifying a joint probability distribution (JPD), $P(x_1, x_2, x_3, \dots, x_n)$, for n variables. If each variable may take one of two values, 2^n entries are needed. For the situation of 100 variables, this translates to about $2^{100} \approx 1030$ entries or possible values of concentrations. As an alternative, substantial economies can be achieved when each variable depends on just a small subset of variables (Pearl 2000). Graphical models allow a greatly reduced number of parameters to be specified and allow a more concise representation of the JPD. For a DMN, its JPD can be specified based on its JT. Each clique in the JT is associated with a probability distribution over its member variables and so is each separator. The JPD is then the product of clique distributions divided by the product of the separator distributions.

From the JPD the probability of the target variable is determined, given the values of other variables. Direct probabilistic inference computation of the JPD rapidly becomes computationally intractable. The junction tree representation of a DMN allows the computations to be carried out by passing probability distributions over clique separators along the tree structure. With DMN, probabilistic inference using a JT is efficient as long as the largest clique is not too large, e.g. <16 variables.

A probabilistic network combines a qualitative graphic structure which encodes domain dependencies, with a quantitative probability distribution which encodes the strength of dependencies. The network structure can be a directed or undirected graph. A Bayesian network (BN) structure is a directed acyclic graph and a decomposable Markov network (DMN)

structure is an undirected chordal graph. Many effective probabilistic inference techniques have been developed (e.g. Pearl, 1986; Henrion, 1988; Lauritzen et al., 1988; Jensen, 1990) and the applicability of probabilistic networks have been amply demonstrated in many domains (Charniak, 1991).

When DMN and BN models are developed, the qualitative structure of the domain (the graph which encodes the dependence and independence relations among variables) is specified, as well as the influences (the probability parameters) are quantified. Typically, the most difficult and time-consuming part of the task in building a DMN or BN model is deriving the structure. Probabilities can be derived from various sources: they can be obtained by interviewing domain experts to elicit their subjective probabilities. They can be gathered from published statistical studies or can be derived analytically from the combination of some problems, for example, transmission of 'genes' from 'parents' to 'children'. Finally, the probabilities can be quantified from raw data. It provides an automatic way to obtain such models. Learning probabilistic networks from data has received attention and researched actively, very recently, by many as an alternative to elicitation in knowledge acquisition (Herskovits and Cooper, 1990; Heckerman et al., 1995; Lam and Bacchus, 1994; Xiang, 1997).

Chow and Liu (1968) pioneered learning of probabilistic networks. They developed an algorithm to approximate a joint probability distribution by a tree-structured BN. Rebane and Pearl (1987) extended their method to learn a polytree-structured BN. However, many real world domain models cannot be represented adequately with a tree structured network. The algorithms that follow are all applicable to learning a multiply connected network. Herskovits and Cooper (1990) developed the Kutato algorithm to learn a BN from a database of cases by minimizing the entropy of the distribution defined by the BN. Their method starts with an empty graph (no links) and adds one link at each pass during the search. Later, Cooper (1992) proposed the K2 algorithm that learns a BN based on a Bayesian method which selects a BN with the highest posterior probability given a database. A similar algorithm was independently developed by Buntine (1994). Recently, Heckerman et al. (1995) applied the Bayesian method to learning a BN by combining prior knowledge and statistical data. Spirtes and Glymour (1991) developed the PC algorithm that learns a BN by deleting links from a complete graph. The belief-scoring function (Cooper and Herskovits, 1992; Heckman et al., 1995) is also proposed. Lam and Bacchus (1994) applied a minimal description length (MDL) method to learning a BN. A BN is evaluated as the best if it has the minimal sum of its own encoding length and the encoding length of the data given the BN. The MDL scoring

function prefers networks that fit the data well and that are simple. Instead of learning a BN, Fung and Crawford (1990) developed the Constructor algorithm that learns a DMN. Dawid and Lauritzen (1993) studied ‘hyper Markov laws’ in learning numerical parameters of a DMN with a given decomposable graph. Madigan and Raftery (1994) proposed algorithms for learning a set of acceptable models expressed as BNs or DMNs.

The common approach to developing the learning structure from data is to introduce a scoring metric, and a search procedure. The search procedure generates alternative graphical structures that encode alternative sets of dependence and independence relations. The scoring metric evaluates each structure with respect to the training data and selects the best structure. This research applies an algorithm for learning belief networks from a probabilistic domain model (PDM) where proper subsets of a set of collectively dependent variables may display marginal independence (after Xiang et al., 1997). The algorithm focuses on learning decomposable Markov networks and a multi-link (ML) search.

In practice, the learning must be established using a finite database. Such a database may contain false correlations that do not exist in the underlying problem domain. They cause the generation of a third type of superfluous link, which we refer to as false links. The probability values associated with false links tend to encode noise contained in the database. A threshold is used in the algorithm to control false links as well as typically redundant links.

Learning probabilistic domain models from data provides an alternative, to elicitation from domain experts, for obtaining probabilistic networks. It is particularly useful in problem domains where no domain experts are easily available. In this research, we explore the learning of probabilistic networks to tackle issues arising in water quality analysis.

24.3 Probabilistic Network Modeling

Application of DMN to water quality prediction was divided into two stages: a *learning* stage and an *inference* stage. In the learning stage, a DMN model was developed over a set of variables from the *training* data. For the *training*, every variable has a value. To learn the structure from training data, a scoring metric and a search procedure was employed. The search procedure generated alternative graphical structures that encode alternative sets of dependence and independence relations. The scoring metric evaluates

each structure using 'training' data and selected the structure with the highest score.

The data for each variable in the training were discretized and the correlation structure between the variables is *learned* and encoded as a DMN, specifying the JPD over the domain through its chordal graphs and the local probability distribution associated with each clique. Figure 24.5 (ignore the histograms for now) shows the graphical structure of a DMN model learned from data, which reveals the dependency relations among variables Doc (dissolved organic carbon), Chlo (Chlorine Dose), Temp (temperature), pH, Alka (alkalinity), and HAAs (haloacetic acids).

In the inference stage, the values of some variables are observed, while values of other variables are uncertain or not observable (hidden). To infer the unobservable value from the observed values of other variables using the learned DMN, the observed values were entered into the corresponding cliques of the DMN. After probabilistic reasoning in the DMN, the posterior probability distribution of each unknown variable was retrieved from the DMN. The second task of the inference stage involves exploring the causal relationship between the target variable and other variables, a procedure which identifies the constituents which lead to increases in HAAs. For example, we could enter Chlo = A and DOC = A (see Table 24.2 for definition of the value of bins 'A') to the model in Figure 24.5. In the Figure, the *bins* of each histogram are labeled by the possible values of the corresponding variable. The height of each bin indicates the probability that the variable takes the corresponding value. Entering Chlo = A and DOC = A is equivalent to saying $P(\text{Chlo} = A) = 1$ and $P(\text{DOC} = A) = 1$, as shown in the figure. The posterior probability distribution of HAAs is then computed (Zhu and McBean, 2004). As shown by the histograms in Figure 24.5, when Cl₂ and DOC dose are at the lowest levels, HAAs will be in the lowest level. The concentration of the HAAs is shown in Table 1. DMN can perform forward and backward inference so that if a level of HAAs is known, we can predict the independent variables, such as Cl₂ or DOC.

Table 24.1 Predict HAAs value from given Cl₂ and DOC values.

Updating Observed Value		Updating Predicted Value
Cl ₂	DOC	HAAs
0.32-0.71 mg/L	0.9-2.7 mg/L	10.12 µg/L
0.71-1.1 mg/L	2.7-4.5 mg/L	43.84 µg/L
1.1-1.49 mg/L	4.5-6.3mg/L	49.12 µg/L

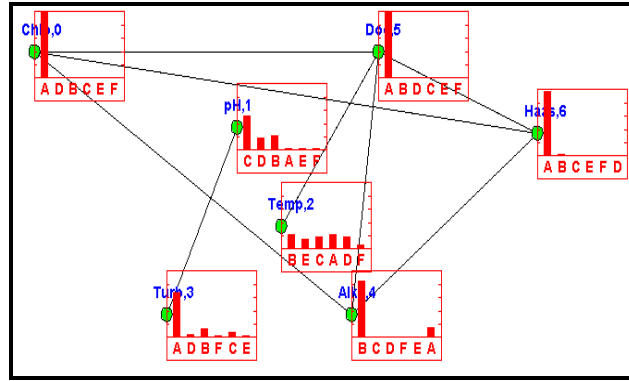


Figure 24.5 Predict HAAs from Cl₂ and DOC.

The orders of bins are determined by the learning software depending on the order in which different values of each variable appear in the training data.

24.3.1 Dataset Training

One hundred and sixty-two datasets were employed as the training datasets; data was discretized for each variable indicated in Table 24.2 and 24.3. DOC in raw water (DOC R), DOC in treated water (DOC T), pre-chlorine dose (PreChlo), post-chlorine dose (PostChlo), temperature (Temp), pH in raw water (pH R), pH in treated water (pH T), HAAs and Total Trihalomethanes (TTHMs) are included for analysis. Data were provided by the Drinking Water Surveillance Program (DWSP) of the Ontario Ministry of the Environment (MOE).

Table 24.2 Discretized intervals or bins for individual variables for HAAs.

Interval or Bin	RawDO mg/L	TreatedDO mg/L	PreChlo	PostChlo	TEMP °C	pH R	pH T	HAAs µg/L
A	1.30-3.68	0.80-1.76	0.15-1.06	0.03-0.86	1.00-6.02	6.36-6.8	6.92-7.2	2.5-29.6
B	3.68-6.06	1.76-2.72	1.06-1.97	0.86-1.69	6.02-11.04	6.8-7.24	7.2-7.488	29.6-56.7
C	6.06-8.44	2.72-3.68	1.97-2.88	1.69-2.52	11.04-16.06	7.24-7.68	7.488-7.77	56.7-83.8
D	8.44-10.82	3.68-4.64	2.88-3.79	2.52-3.35	16.06-21.08	7.68-8.12	7.77-8.06	83.8-110.9
E	10.82-13.2	4.64-5.6	3.79-4.7	3.35-4.18	21.08-26.1	8.12-8.56	8.06-8.34	110.9-138

Table 24.3 Discretized intervals or bins for individual variables for TTHMs.

Interval or Bin	RawDO Mg/L	TreatedDO Mg/L	PreChlo	Postchlo	Temp °C	pH R	pH T
A	1.30-3.68	0.80-1.76	0.15-1.06	0.03-0.86	1.00-6.02	6.36-6.8	6.92-7.2
B	3.68-6.06	1.76-2.72	1.06-1.97	0.86-1.69	6.02-11.04	6.8-7.24	7.2-7.488
C	6.06-8.44	2.72-3.68	1.97-2.88	1.69-2.52	11.04-16.06	7.24-7.68	7.488-7.77
D	8.44-10.82	3.68-4.64	2.88-3.79	2.52-3.35	16.06-21.08	7.68-8.12	7.77-8.06
E	10.82-13.2	4.64-5.6	3.79-4.7	3.35-4.18	21.08-26.1	8.12-8.56	8.06-8.34

The learning tool, Webweaver III (after Xiang et al, 1997), was applied to extract the DMNs from the training data as summarized. The learned model was used to predict the 'hidden' values of the posterior probability and expected value of the target HAAs and TTHMs variables, given the other variables. The prediction ability of the DMN to predict the 'hidden' data is illustrated in Figure 24.6 and Figure 24.7.

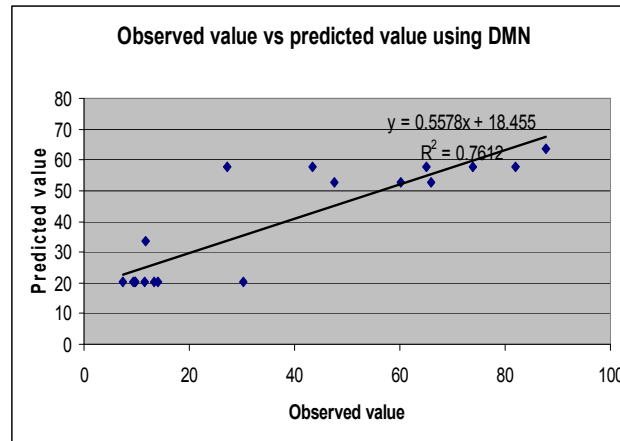


Figure 24.6 Results of DMN for HAAs.

Note that we have discretized each variable in Figure 24.5 into six intervals. In general, the more intervals, the more refined the representation of the model. On the other hand, the computational cost of inference

computation is proportional to k^q , where k is the maximum number of intervals per variable and q quantifies the graph density of the DMN. Therefore, as the number of intervals increases, the inference computation using the model will be more expensive. Furthermore, when the number of intervals increases, the number of data records in each interval decreases indicating that the probability parameters of the model derived will be less accurate, and the graphical structure (unless more data are available).

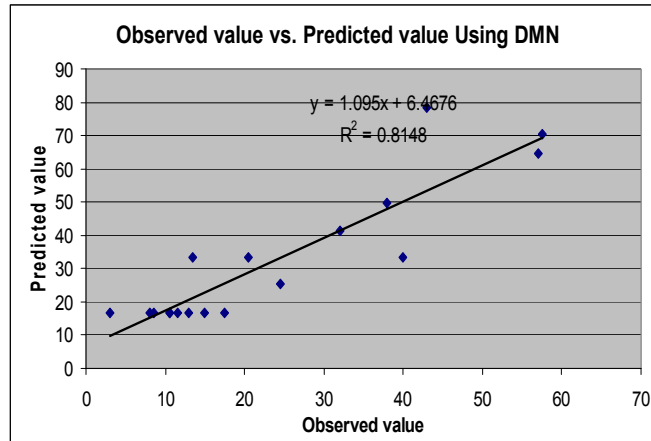


Figure 24. 7 Results of DMN for TTHMs.

24.4 Regression

Multiple regression analysis has been one of the most heavily-used techniques in statistics wherein problems involving more than two variables can be treated in a manner analogous to problems with two variables (McBean & Rovers, 1998). Example applications include use for control charts (Mandel, 1969), calibration (Mendenhall and Ott, 1971), biology and medicine (Armitage, 1971), time series data (Cohen et al., 1969) and water quality prediction (Clark et al., 2001; Golfinopoulos and Arhonditsis, 2002 and Villanueva et al., 2003.).

Nevertheless, multiple linear and nonlinear regression approaches do not perform well in all cases. The problems may arise due to the assumptions; for example, the assumption of normality. The assumed functional form

(linear) may be false. Non-linear multiple regressions have been widely employed by log-transforming the variables data, as per Box and Cox (1964). Such analyses assume that the data have a lognormal distribution, i.e. the logarithm of the response is normally distributed. The most commonly used transformations are modified power transformations of Box and Cox (1964), but these may produce transformation errors. In addition, traditional statistical models, such as multiple regression, suffer a number of issues. The first is that the hypothesis test is based on a test statistic that is at best indirectly related to the quantity of interest - the truth (or probability of truth) of the null hypothesis. The p-value commonly reported in hypothesis testing is the probability, given that the null hypothesis is true, of observing values for the test statistic that are as extreme, or more extreme, than the value actually observed. The scientist, however, is interested in the probability of the correctness of the hypothesis, given that he/she has observed a particular value for the test statistic. A second problem relates to the issue of *conditioning*, which concerns the nature of the sample information in support of the hypothesis. Classical hypothesis tests are conditioned on other hypothetical samples in the sampling distribution and the sampling distribution is a probability density function that is hypothetical in nature.

In this research, we apply nonlinear regression with the backward elimination method. The backward elimination method begins with the largest regression using all independent variables, and subsequently reduces the number of variables in the equation until a decision is reached on the equation to employ. It is an economical approach in the sense that it tries to examine only the 'best' regressions containing a certain number of variables. The basic steps in the backward elimination procedure are:

1. A regression equation containing all variables is computed.
2. The P-value of the t-test is calculated for every predictor variable treated as if it was the last variable to enter the regression equation.
3. The lowest P-value of each variable, say P_L , is compared with a pre-selected significance level P_0 , say 0.05.
 - a. If $P_L > P_0$, remove the variable X_L from consideration and re-compute the regression equation in the remaining variable.
 - b. If $P_L < P_0$, adopt the regression equation as calculated.

This is a satisfactory procedure, especially for the purpose of seeing all the variables in the equation, in order 'not to miss anything' (Draper and Smith, 1981). It also can prevent regression equations relying too heavily on the automatic selection preformed by the computer.

24.5 Comparing the DMN Approach with Regression Analysis

The Ontario Ministry of the Environment (MOE) has collected water quality data, including DBP concentrations, from 28 Ontario communities with sampling encompassing high and low temperature periods. Samples were collected from three sampling points within the water treatment plants, which are raw water, finished water and within the distribution system. All these facilities draw water from surface water supplies and employed conventional treatment procedures: pre-chlorination, conventional coagulation & flocculation, clarification, filtration and post-chlorination to treat the water, and all used alum as the primary coagulant. To assess the distributional assumptions for the individual constituents, and validate the model accuracy and capacity, seventeen individual monitoring rounds were randomly selected from 180 data sets and used as the 'hidden' values.

The data selected for the histogram were subjected to a logarithmic transformation as per Figure 24.8 to 24.10, since the observed data are skewed and some of the data distributions are close to bimodal (e.g. HAAs). The log transformation was performed on all the data sets before developing the regression model to approach the normality assumption (although it doesn't guarantee normality conditions will exist. It is also noted that this data transformation does not eliminate the bimodality).

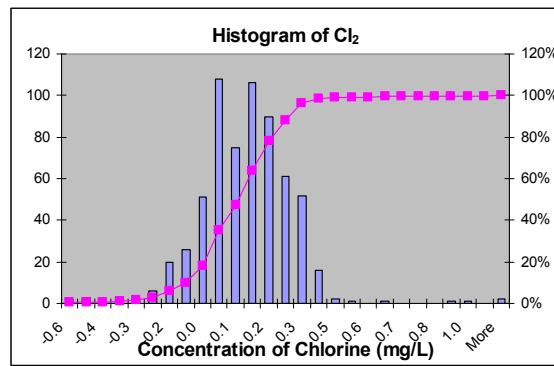


Figure 24.8 Histogram of Chlorine.

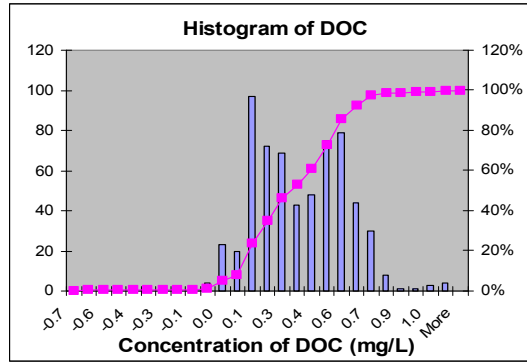


Figure 24.9 Histogram of Dissolved Organic Carbon.

24.5.1 Non-linear Regression of HAAs

We present non-linear regression of HAAs as an example of regression analysis. Table 24.4 summarizes regression results determined by using Microsoft Excel, Data Analysis Toolpack, including goodness-of-fit of the multiple regression model (R^2), the total F-value for the model, the least-square regression coefficients, the standard errors, the t-value and the P-value of significance for rejecting the null hypothesis for each variable. The result shows some of the independent variables, e.g. raw dissolved organic carbon (DOC R), temperature (Temp) and pH (pH R) in raw water are not statistically significant. Therefore, the back-elimination procedure is used for saving the computation.

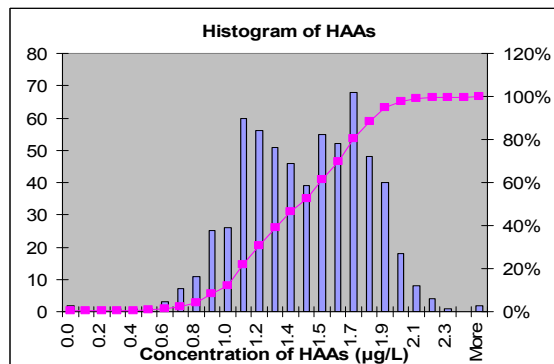


Figure 24.10 Histogram of HAAs.

Table 24.4 Results of regression analyses on HAAs (log-transformed data).

<i>Regression Statistics</i>						
Multiple R	0.84566					
R Square	0.71514					
Adjusted R Square	0.70219					
Standard Error	0.18913					
Observations	162					
<i>ANOVA</i>						
	df	SS	MS	F	Significance F	
Regression	7	13.8291	1.97559	55.23226	7.36E-39	
Residual	154	5.50839	0.03577			
Total	161	19.3375				
	Coefficients	Standard Error	T Stat	P-value	Lower 95%	Upper 95%
Intercept	4.7867	1.1569	4.1374	0.0001	2.5012	7.0722
DOC R	0.2420	0.1520	1.5918	0.1135	-0.0583	0.5424
DOC T	0.8174	0.1820	4.4909	0.0000	0.4578	1.1769
Cl ₂ Pre	0.1279	0.0441	2.9021	0.0043	0.0409	0.2150
Cl ₂ Pos	0.1037	0.0499	2.0769	0.0395	0.0051	0.2023
Temp	-0.0453	0.0419	-1.0815	0.2812	-0.1282	0.0375
pH T	-2.5324	0.7775	-3.2571	0.0014	-4.0683	-0.9965
pH R	-1.5827	1.0844	-1.4595	0.1465	-3.7250	0.5595

Table 24.5 summarizes the results of the nonlinear regression for HAAs with BE procedure. The results indicate that all the constituents, dissolved organic carbon (DOC T), pre-chlorine dose (Cl₂ Pre), post-chlorine dose (Cl₂ Pos), pH in raw water (pH R) are statistically significant.

From the regression results per Table 24.5, we derive the mathematical function for HAAs as:

$$\text{HAAS} = 10^{4.09} * (\text{DOC T})^{1.124} * (\text{Cl}_2\text{Pre})^{0.123} (\text{Cl}_2\text{Pos})^{0.1248} (\text{pHR})^{-3.34} \dots \dots \dots (24.1)$$

where the independent parameters DOC T, Cl₂Pre, Cl₂Pos, and pHR correspond to the dissolved organic carbon (mg/L) in treated water, pre- and

post- chlorine dose (mg/L), and pH in raw water. Equation 24.1 is applied to predict 17 hidden values of HAAs. The predicted values are compared with observed value as per Figure 24.11

Table 24.5 Summary output of HAAs by Backward Elimination Method.

Regression Statistics						
Multiple R						0.83757
R Square						0.70153
Adjusted R Square						0.69393
Standard Error						0.19174
Observations						162
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	4	13.56584	3.391461	92.253956	3.37635E-40	
Residual	157	5.771669	0.036762			
Total	161	19.33751				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	4.0982	0.6467	6.3373	0.0000	2.8209	5.3756
DOC T	1.1247	0.0777	14.4779	0.0000	0.9713	1.2782
Cl ₂ Pre	0.1233	0.0446	2.7667	0.0063	0.0353	0.2113
Cl ₂ Pos	0.1248	0.0480	2.5996	0.0102	0.0300	0.2196
pH R	-3.3414	0.7120	-4.6931	0.0000	-4.7477	-

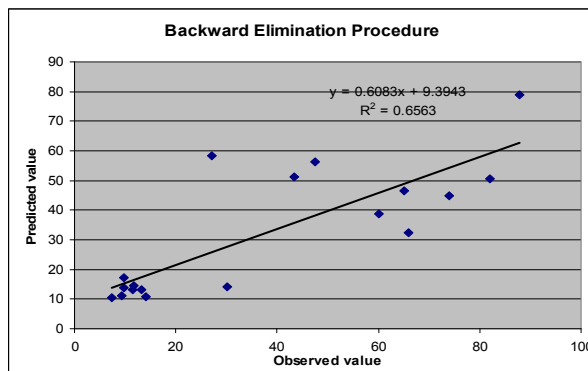


Figure 24.11 Result of Multiple Regression for HAAs.

Comparing Figure 24.6 and Figure 24.11 with the same datasets, the r^2 of predicted value versus observed values using DMN and multiple regression are 0.76 and 0.65 respectively. DMN is 10% better in terms of r-square than the multiple regression.

24.5.2 Non-linear Regression of TTHMs

Table 24.6 is the summary of regression results of TTHMs completed using Excel, Data Analysis Toolpack. As was the case for HAAs, goodness-of-fit of the multiple regression model (R^2) are obtained, the total F-value for the model, the least-square regression coefficients, the standard errors, the t-value and the P-value of significance for rejecting null hypothesis for each variable. The results show that a different set of independent variables are not statistically significant. For example, raw dissolved organic carbon (DOC R), and pH (pH T, pH R) in treated and raw water are not statistically significantly different from zero. Therefore, the BE method is applied.

Table 24.7 gives the results of non-linear regression for TTHMs with the BE method. It indicates that from all the constituents, intercept, dissolved organic carbon in treated water (DOC T), pre chlorine dose (Cl_2 Pre), post chlorine dose (Cl_2 Pos), and temperature (temp) are statistically significant.

From the regression results per Table 24.7, we derive the mathematical function for TTHMs is denoted as:

$$TTHMs = 10^{0.78} * (DOC T)^{1.149} (Cl_2 Pre)^{0.1767} (Cl_2 Pos)^{0.2247} (Temp)^{0.1861} \dots \dots \dots \quad (24.2)$$

Where the independent parameters DOC T, Cl_2 pre, Cl_2 pos, and Temp correspond to the dissolved organic carbon (mg/L) in treated water, pre- and post- chlorine dose (mg/L), and temperature ($^{\circ}C$).

Equation 24.2 is applied to predict the 17 hidden values of TTHMs, with the same hidden value of independent variables for HAAs. The predicted values are compared with the observed values in Figure 24.12.

Comparing Figures 24.7 and 24.12 with the same datasets, the R^2 of predicted value vs observed value using DMN, and multiple regression, are 0.81 and 0.52 respectively. i.e. DMN is 29% better than multiple regression.

It is concluded that DMN performs better on prediction of uncertain data in terms of accuracy over regression. Regression performed poorly because of the normal distribution assumption implicit in regression, which is

explained in the next section. Hence, DMNs promise a newer generation of tools in prediction of water quality parameters and water quality analysis.

Table 24.6 Results of regression analyses on TTHMs (log-transformed data)

Regression Statistics						
Multiple R	0.8683					
R Square	0.7539					
Adjusted R Square	0.7427					
Standard Error	0.1796					
Observations	162					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	7	15.22428622	2.17489803	67.394225	1.07786E-	
Residual	154	4.969777405	0.03227128			
Total	161	20.19406363				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
intercept	2.0026	1.0989	1.8223	0.0703	-0.1683	4.173
DOC R	0.1689	0.1444	1.1696	0.2440	-0.1164	0.454
DOC T	0.8916	0.1729	5.1576	0.0000	0.5501	1.233
Cl ₂ pre	0.2104	0.0419	5.0230	0.0000	0.1276	0.293
Cl ₂ pos	0.2384	0.0474	5.0276	0.0000	0.1448	0.332
Temp	0.1993	0.0398	5.0052	0.0000	0.1206	0.278
pH R	-1.3699	0.7385	-1.8549	0.0655	-2.8288	0.089
pH T	-0.0055	1.0300	-0.0054	0.9957	-2.0404	2.029

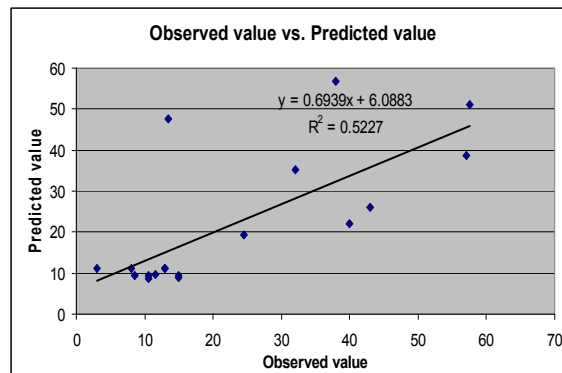


Figure 24.12 Result of Multiple Regression for TTHMs.

Table 24.7 Summary Output of TTHMs By BE Procedure

Multiple R	0.8529					
R Square	0.7274					
Adjusted R Square	0.7204					
Standard Error	0.1902					
Observations	162					
ANOVA						
	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	4	15.15765215	3.78941304	104.727524	2.8474E-43	
Residual	157	5.680816519	0.03618354			
Total	161	20.83846866				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.7814	0.0512	15.2511	0.000000	0.6802	0.8826
DOC T	1.1490	0.0764	15.0462	0.000000	0.9981	1.2998
Cl ₂ pre	0.1767	0.0429	4.1162	0.000062	0.0919	0.2615
Cl ₂ pos	0.2247	0.0479	4.6966	0.000006	0.1302	0.3193
Temp	0.1861	0.0413	4.5017	0.000013	0.1044	0.2677

24.6 DMN is Robust Compared with Regression

The regression approach assumes a Gaussian distribution around a predicted value of a target variable. DMN instead provides the posterior distribution of the target variable without presumption of the shape of the distribution. That is, the DMN can represent distributions of arbitrary shape (subject to proper discretization). For instance, suppose the target variable x has a posterior probability distribution $p(x_1) = 0.38$, $p(x_2) = 0.2$, $p(x_3) = 0.42$, where $x_1 < x_2 < x_3$. Since regression assumes a Gaussian distribution, it cannot fully represent a bimodal distribution. Either it will center the 'bell-shape' at the most likely value x_3 , which grossly under-estimates the possibility that the true value may be x_1 , or it will center the bell-shape at x_2 , which grossly over-estimates the possibility that the true value is x_2 . This example demonstrates that DMN and regression will differ to the greatest extent in their predictions when the actual distribution is skewed or bi- or multi-modal. In water quality prediction, many constituents have skewed and multimodal distributions. For example, HAAs sometimes exhibits a bimodal distribution, illustrated in Figure 24.13 by the curve with two peaks.

Suppose we are interested in predicting the probability of the variable taking its center value (labeled 0 horizontally). A DMN model with proper discretization will predict the variable at its center value correctly with the valley probability. However, regression will increase the width of its bell-shape (through larger variance value) to adapt to the bimodal situation, as shown in the figure by the curve of a single peak. As a result, it will predict the variable at its center value mistakenly with the peak probability. On the other hand, regression will indicate the trend correctly in the tail regions. Although it may still differ from DMN on the actual height of the tail distribution, this difference is at most quantitative as schematically depicted in Figure 24.13. Note that the difference of the two models at the center value is qualitative, not just quantitative. As shown in the experiments using the same datasets, R^2 for HAAs is 0.65 in regression and is 0.76 in DMN. For TTHMs, R^2 is 0.52 in regression and 0.81 in DMN.

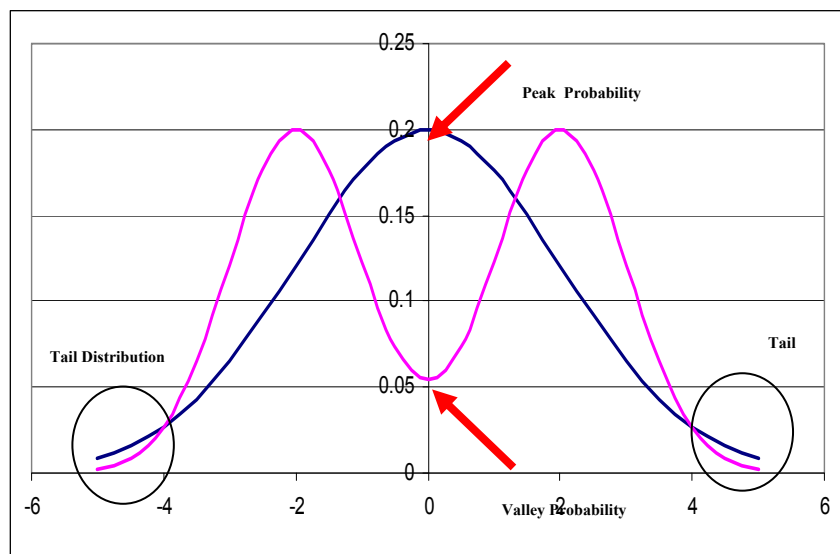


Figure 24.13 Indications why regression is less robust.

DMN provides not only the most likely value of a target variable, but also its posterior distribution, which allows decision-theoretic trade-offs of *possibilities* with *desirabilities*. As described in (Ravindran 1986), decision analysis provides a rich set of concepts and techniques to aid the decision

maker in dealing with complex decision problems under uncertainty. The decision analysis formulation differs from classical statistical inference. The regression results as a deterministic function for the inference. DMN comes out with a probabilistic distribution, which is encoded with the uncertainty.

24.7 Conclusion

This research compared DMN predictive model and classical multiple regression predictive models, explaining why DMN is a potentially better prediction model than multiple regression both theoretically and experimentally for this application. The statistical assumptions required for this application are less restrictive in this application, for DMN than multiple regression. Although both models can be used to identify the relative significance of water quality (NOM indicators, pH, etc.) and operational variables (disinfectant dose, water temperature, contact time, etc.) responsible for the formation of DBPs, multiple regression models have their limitations, not only in terms of accuracy, but also in terms of model capacity. DMN can perform bi-directional inference, but multiple regression cannot. It means that DMN can be a powerful tool for dealing with short term real-time control. With improved usability, the DMN technology will play a larger role in maximizing the predictive potential that multiple regression cannot offer.

References

- American Society of Civil Engineers (ASCE), Interim Voluntary Guidelines for Designing an Online Contaminant Monitoring System, Reston VA, 2004.
- Armitage, P., "Statistical methods in medical research", Blackwell Scientific Publication, London, pages, 504, 1971.
- Box, G.E.P., and Cox, D.R., "An analysis of transformation". Journal of Royal statistical Society, Ser. B, 39, pp. 211-252, 1964.
- Buntine, W. Operations for learning with graphical models. J. of Artificial Intelligence Research 2, 159-225. 1994
- Castillo, E. Gutierrez, J.M. and Hadi, A.S., "Expert systems and probabilistic network models", New York: Springer Verlag. 1997.
- Charniak, E., "Bayesian networks without tears", AI Magazine, vol. 12 no. 4: 50-63, 1991.
- Chow, C.K. and Liu, C.N., "Approximating discrete probability distributions with dependence trees", IEEE Trans. on Information Theory, (14):462-467, 1968.

- Cohen, J., "Statistical power analysis for the behavioral sciences", New York: Academic Press, 1969.
- Clark R. M., Thurnau R.C., Sivaganesan M. and Ringhand P., "Predicting the formation of chlorinated and brominated by-products", *J Environ Eng* 127(6):493 –501, 2001.
- Cooper, G.F. and Herskovits, E., "A Bayesian method for the induction of probabilistic networks from data, *Machine learning*", (9):309-347, 1992.
- Dawid, A.P. and Lauritzen, S.L., "Hyper Markov laws in the statistical analysis of decomposable graphical models", *Annals of Statistics*,21(3):1272-1317,1993
- Draper N.R. and Smith H. "Applied regression analysis", Second edition, John Wiley & Sons, Inc., New York, 1981.
- Fung, R.M. and Crawford, S.L., "Constructor: A system for the induction of probabilistic Models", In *Proc. of AAAI*, pages 762-769, Boston, MA, MIT Press, 1990.
- Heckman , D, Geiger, D. and Chickering D.M., "Learning Bayesian networks: the combination of knowledge and statistical data", *Machine Learning* 20:197-243, 1995.
- Golfinopoulos S.K, Arhonditsis G.B., "Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics", *Chemosphere*, 47:107 –1018, 2002.
- Henrion, M., "Propagating uncertainty in Bayesian networks by probabilistic logic sampling", In J.F. Lemmer and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence 2*, pages 149-163, Elsevier Science Publishers, 1988.
- Herskovits, E.H. and Cooper, G.F., "Kutato: an entropy-driven system for construction of probabilistic expert systems from database", In *Proc. 6th Conf. on Uncertainty in Artificial Intelligence*, Pages 54-62, Cambridge, 1990.
- Jensen, F.V., Lauritzen, S.L., and Olesen,K.G., "Bayesian updating in causal probabilistic networks by local computations", *Computational Statistics Quarterly*, (4):269-282, 1990.
- Jensen, F. V., "An introduction to Bayesian networks", New York, NY: Springer-Verlag, New York Inc. 1996.
- Lam, W., and Bacchus, F. "Learning Bayesian belief networks: An approach based on the MDL principle", *Computational Intelligence* 10(4): 269-293. 1994.
- McBean, E.A. and Rover, F.A., "Statistical procedures for analysis of environmental monitoring data and risk assessment", Prentice Hall PRT, Upper Saddle River, NJ 07458, 1998.
- Madigan, D. and Raftery, A.E., "Model selection and accounting for model uncertainty in graphical models using Occam's window". *Journal of American Statistical Association*, 89(428):1535-1546, 1994.
- Mandel, B. J., "The regression control chart", *Journal of Quality Technology*, Vol. 1, No. 1, 1969.
- Marcot, B. G., Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement, *Forest Ecology and Management*, v 153 (1-3), n 1-3, pp. 29-42, 2001.
- Mendenhall, W. and Ott, L., "A method for the calibration of an on-line density meter", *Journal of Quality Technology*, Vol. 3, No. 2, pages. 80-86, 1971.

- MOE., "Drinking water surveillance program summary report for 2000, 2001 and 2002", <http://www.ene.gov.on.ca/envision/water/dwsp/0002/>, 2004.
- Neapolitan, R. E., "Probabilistic reasoning in expert systems", New York: Wiley, 1990.
- Pearl, J., "Fusion, propagation, and structuring in belief networks, *Artificial Intelligence*", (29):241-288, 1986.
- Pearl, J., "Evidential reasoning using stochastic simulation of causal models". *Artificial Intelligence*, 32:247-257, 1987.
- Pearl, J., "Probabilistic reasoning in intelligent systems". 2nd. ed. San Francisco, Calif., Morgan Kaufmann, 1988 and 1991.
- Ravindran A., Phillips D. T., Solberg J. J., "Operations Research", 2nd, John Wiley & Sons, 1986.
- Sadiq R. and Rodriguez M. J., "Disinfection by-products in drinking water and predictive models for their occurrences: a review". *Science of the Total Environment*, 321, 21-46, 2004.
- Siber, N. A., M. Pantazidou, and M. J. Small, Expert system methodology for evaluating reductive dechlorination at TCE sites, *Environmental Science and Technology*, v 33(17), pp. 3012-3020, 1999.
- Spirte, P. and Glymour, C., "An algorithm for fast recovery of sparse causal graphs", *Social Science Computer Review*, 9(1):62-73, 1991.
- Stow, C. A., C. Roessler, M. E. Borsuk, J. D. Bowen, and K. H. Reckhow, Comparison of Estuarine water quality models for total maximum daily load 12 development in Neuse River Estuary, *Journal of Water Resources Planning and Management*, v 129 (4), pp. 307-314, 2003.
- Villanueva C.M., Kogevinas M., Grimalt J.O., "Haloacetic acids and trihalomethanes in finished drinking waters from heterogeneous sources", *Water Res*, 37:953 -958, 2003.
- Xiang, Y. Wong, S.K.M and Cercone, N., "A Microscopic study of minimum entropy search in learning decomposable Markov networks", *Machine Learning*, Vol.26, No.1, 65-92, 1997.
- Zhu Z. J and McBean E. A., "Estimation of censored data water quality values using decomposable Markov networks", *Journal of Environmental Informatics*, 4 (2) 48-5, 2004.