

28

Dynamic Bayesian Network Approach for Modeling Trihalomethanes from Ontario Water Supply Systems

Zoe J. Y. Zhu, Jeff Kroes and Edward A. McBean

A dynamic Bayesian network (DBN) approach is used to quantify relational knowledge for modeling relations or dependencies between variables from the dynamic system of disinfection byproduct (DBP) formation which changes over time. The DBN framework is used to assess causality between constituent parameters of water supply quality, based on data from communities in Ontario which rely on groundwater as their source of supply. The DBN models are used to assess probabilistic dimensions and to assist decision-making by identifying control options to decrease DBP formation.

28.1 Introduction

Chlorination has been the major strategy for the disinfection of drinking water in Ontario due to favorable economics and its effectiveness in killing water-borne pathogens. However, concerns with disinfection by chlorination have been the subject of considerable scrutiny since the 1970s when halogenated disinfection by-products (DBPs) such as total trihalomethanes (TTHMs) were discovered due to the interaction of chlorine with organic matter like humic substances present in water (Rook, 1974). As a result, methods to reduce the formation rates of DBPs have become a focus in water treatment. Several factors including, but not limited to, chlorine dose, pH, temperature, dissolved

Zhu, Z.J., J. Kroes and E. McBean. 2010. "Dynamic Bayesian Network Approach for Modeling Trihalomethanes from Ontario Water Supply Systems." *Journal of Water Management Modeling* R236-28. doi: 10.14796/JWMM.R236-28.

© CHI 2010 www.chijournal.org ISSN: 2292-6062 (Formerly in Dynamic Modeling of Urban Water Systems. ISBN: 978-0-9808853-3-0)

organic carbon (DOC), and contact time, have been reported to significantly affect the formation of TTHMs (McBean et al., 2008). Current research also shows that algae contributes to increases in formation of DBPs upon chlorination (Plummer and Edzwald, 2001; Nguyen et al., 2005). Krasner et al. (1989) have researched 35 water treatment plants in the U.S. and found that the mean TTHM value was 34 µg/L in spring, 44 µg/L in summer, 40 µg/L in fall and 30 µg/L in winter and they were the largest class of DBPs detected on a weight basis. Initial epidemiologic studies have found associations between TTHMs and increased risk of cancer, with the most consistent findings being for bladder cancer (Morris et al., 1992). Several epidemiologic studies have also examined the associations between TTHMs and adverse pregnancy outcomes including spontaneous abortion (Waller et al., 1998; Savitz et al., 1995), pre-term delivery (Wright et al., 2003), and stillbirth (Dodds et al., 1999; Toledano et al., 2005).

Due to the adverse health effect of TTHMs, the United States Environmental Protection Agency (USEPA) promulgated the Stage 1 and Stage 2 disinfectant/disinfection byproducts (D/DBPs) rule to cope with TTHMs and some other harmful DBPs in drinking water (USEPA, 1999). The maximum contaminant levels (MCLs) under the Stage 1 D/DBP rule include 80 µg/L for TTHMs, with the maximum residual disinfectant concentration of chlorine as 4.0 mg/L. In Austria, Switzerland and Luxembourg the TTHMs regulatory limits in drinking water are 30 µg/L, 25 µg/L and 50 µg/L respectively. A survey by Arora et al. (1997) of TTHMs in 100 water treatment plants respectively in the U.S.A. found that 20% and 60% of water treatment plants could not meet the requirements for TTHMs in DBP Rule I and DBP Rule II respectively.

In this research, the authors have investigated the levels of TTHMs at 176 water treatment plants in Ontario, finding that 102 plants are <40 µg/L, 34 plants are >40 µg/L and <80 µg/L, and 40 plants are >80 µg/L, as summarized in Table 28.1. Given these findings, the more stringent TTHMs controls need to be a high priority among issues of drinking water quality in Ontario.

Table 28.1 Level of TTHM in Ontario treatment plants.

Water Treatment Operation in Ontario				
Water Treatment Plant	Total plants in each region	TTHM <40µg/L	TTHM >40 but <80 µg/L	TTHM >80µg/L
Central	22	20	2	0
Eastern	42	28	10	14
Northern	42	12	9	21
SouthWest	43	29	11	3
West Central	27	23	2	2
Total	176	112	34	40
Percentage	100%	64%	19%	23%

Numerous studies have used multivariate models or forward regression and backward regression methods to correlate DBP formation with various combinations of explanatory variables with varying levels of success (e.g. McBean et al., 2008). Sadiq and Rodriguez (2004) have summarized a number of such predictive models for TTHMs.

While the models derived from laboratory data have demonstrated good predictive characteristics of TTHMs (e.g. Amy et al., 1987; Amy et al., 1998; Adin et al., 1991; Rathbun, 1996; Chang et al., 1996; Clark, 1998; Rodriguez et al., 2000), the models developed from field conditions have much lower prediction capabilities for TTHMs formation (e.g. Golinopoulos and Arhonditsis, 2002). However, the above-mentioned models have limitations because the models assume the data are normally distributed. In reality, while typically the data may be normally distributed at a specific location, when the data are pooled over many locations, the data fail to be well described by normality, which reduces the performance of the prediction models. As well, these techniques neglect probabilistic and temporal dependencies between water quality measures. In this study, a DBN prototype is developed which does not require the normality assumption and consequently, more accurate TTHM predictions are possible.

A DBN is used here to assess changes in temporal ground water quality and DBP formation. The resulting network model is developed in two steps. In the first step, data resulting from the pre-processing model are examined, which were organized as yearly measurements covering facilities using groundwater as a source of water supply in Ontario. The first task was to identify dependencies between the variables of water quality in order to detect useful information on process dynamics.

In the second step, the constructed DBNs are used for predicting the values of the variables in the future. The resulting networks were investigated using WebWeavr-III (Xiang 1999). The Bayesian network inference tool is used for analyzing measurement data from three or more successive time periods. Use of DBNs to predict future values requires not only discovery of a dependency model between the variables but also relate together, variables from successive time periods to embed temporal features into the model. In Bayesian reasoning, the marginal probability distribution of any variable may be updated upon acquiring evidence for other variables. In response, the DBN can provide detailed predictions about not only the DBP contamination, but also other variables of interest, such as how COD affects DBP formation. The results of the prediction will assist decision making on treatment options.

28.2 Bayesian and Dynamic Bayesian Network

A Bayesian network (BN) is a probabilistic graphical model for knowledge representation in intelligent systems. Dependence relations in the problem domain are represented as a directed acyclic graph. Nodes represent variables and arcs represent direct causal dependence relations. The strength of a dependency is expressed by a probability. A BN allows for exact probabilistic inference. Observations can be made on some variables and then states of other variables inferred, as in Figure 28.1. If we have observed the level of total dissolved solid and electronic conductivity, the pH level can be inferred and further, the level of the TTHMs can also be inferred.

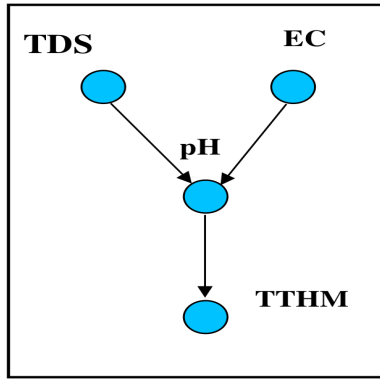


Figure 28.1 A simple static Bayesian network.

A static Bayesian network consists of a set V of domain variables, a directed acyclic graph, and a distribution $P(x|\text{par}(x))$ for each x in V with parents $\text{par}(x)$. $P(x|\text{par}(x))$ encodes the strength of causal dependence. The joint probability distribution (JPD) is the product of distributions associated with nodes in the graph. Given a BN (V, G, P) , the following holds:

$$P(V) = \prod_{v \in V} P(v | \text{par}(v)) \quad (28.1)$$

A DBN is a Bayesian network for non-static domains. The state of a variable may change over time. The value of total dissolved solids (TDS) or electrical conductivity (EC) may increase or decrease each year. The previous values of variables, such as the first instance of TDS and EC, are referred to as the forward interface (FI). In a DBN, the Markov assumption is: given the pre-

sent, the future is independent of the history. The history is captured by the FI. The network template includes the BN for a time instant plus FI of the previous instance. Figure 28.2 shows a simple dynamic Bayesian network.

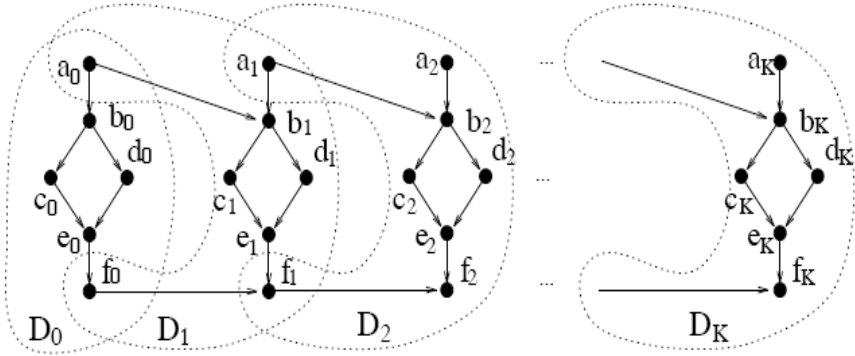


Figure 28.2 A dynamic Bayesian network (Xiang, 1999).

At any time $t = j < k$, the slices D_0, \dots, D_{j-1} represent the domain history and D_{j+1}, \dots, D_k predict the future. Evidence (observations obtained in the past and present) may be entered into D_0, \dots, D_j . The entire network will be updated automatically to its posterior distribution.

As mentioned, DBNs are the temporal extension of BNs. DBNs deal with problems of a temporal nature. It reasons over the set of random processes $V = \{V[t] : t \in T\}$ instead of random variables. The JPD of a DBN can be defined as:

$$P(V) = \prod_{t \in T} \prod_{v(t) \in V(t)} P(v(t) | par(v(t))) \tag{28.2}$$

The JPD over multiple time instances is the product of the JPD over each time instance. These JPDs for each time instance are multiplied together through the outer product operator. In this research, the focus is on discrete-time and discrete-space random processes.

- A prior model

$$P(V(0)) = \prod_{v(0) \in V(0)} P(v(0) | par_0(v(0))) \tag{28.3}$$
- A transition model

$$P\left(V(t) \mid V(t-1) = \prod P(v(t) \mid \text{par}_t(v(t)))\right) \quad (28.4)$$

Specifying evolution of the process as it moves from time $t-1$ to time t for $t \in \{1, 2, 3, \dots\}$.

28.3 Application of DBN to Prediction of Water Quality and DBP

The drinking water surveillance program (DWSP) is a voluntary monitoring program operated by the Ontario Ministry of the Environment (MOE) in cooperation with municipalities to gather scientific data concerning drinking water quality in Ontario, Canada. The DWSP provides field data from which predictive models can be generated. Table 28.2 shows the location of data sets assembled for drinking water.

Table 28.2 Study areas in Ontario.

Datasets from the following locations in Ontario
Aurora, Bolton, Bourget, Cambridge, Clarence, Capreol, Deloro, Guelph, Havelock, Port Perry, Waterloo

The MOE identified that the datasets collected from these monitoring wells are important in assessing the groundwater quality and in the prediction of the effect of certain water quality parameters on drinking water. The period covered in these locations is 1998–2004. Each site has several monitoring wells and water samples were collected periodically from these wells and the concentrations in these water samples were used in this study.

In this research, we also found nitrate, TDS and electrical conductivity level will affect the DBP. The plot of TTHM vs EC, TDS and nitrate is shown in Figure 28.3, where the data are from Havelock, Ontario; when TTHMs are in the highest level, 94.5 $\mu\text{m/L}$, the EC, TDS and nitrate are also at their highest levels, 709 $\mu\text{S/cm}$, 496 mg/L and 0.948 mg/L respectively. High EC, TDS and nitrate levels also happened in the summer and fall, which is consistent with Krasner et al.'s survey (1989). The higher chlorine dosage and pre-chlorination were responsible for the higher DBP concentrations (McBean et al., 2008).

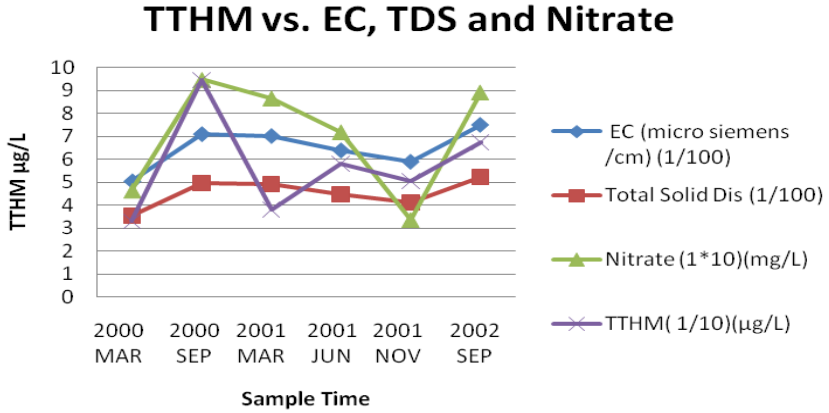


Figure 28.3 Effects of EC, TDS and nitrate to TTHM over time.

28.4 Steps for Developing DBN Model

A forward interface of water quality parameters is created for propagation of information and structure, as shown in Figure 28.4.

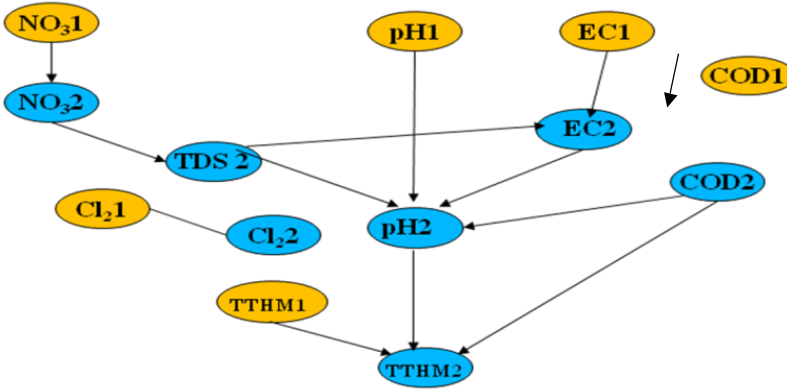


Figure 28.4 DBN structure of water quality.

- Step 1. Choose the water quality parameters, TDS, EC, pH, Cl₂, COD and TTHMs.
- Step 2. Obtain the model structure by combining learning and water expert knowledge.

- Step 3. Each node in the graph represents a variable and directed arcs represent causal dependence between variables. For instance, *TDS* depends on the level of NO_3 (represented by the arc from NO_3 to *TDS*), as well as the previous time value of *TDS* (represented by the arc from *TDS* 1 to *TDS* 2).

The variable *TTHMs* depends on its past state and the current states of pH, COD, and Cl_2 . *TTHMs* is chosen as a target variable.

The previous values of variables, such as the first instance of *TDS* 1 and *TTHM* 1, are referred to as the forward interface. The forward interface allows the propagation of information from a previous time interval into the next time interval. The network template is drawn with WebWeavr-III using the Bayesian network editor utility, as in Figure 28.5.

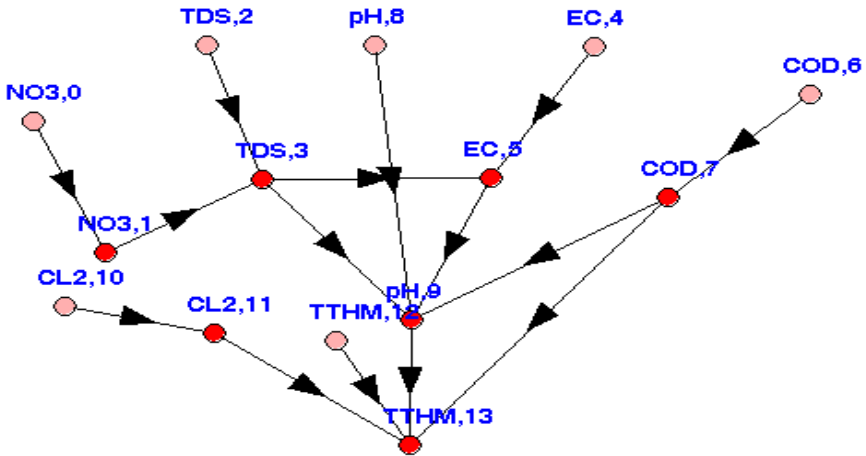


Figure 28.5 The network template.

The conditional probability distributions (CPTs) are assigned to all variables. For instance, the CPT assigned to variable Cl_2 at time i is:

$$\begin{aligned} P(\text{Cl}_2,i <2 \mid \text{Cl}_2,i-1 <2) &= 0.96 \\ P(\text{Cl}_2,i \geq 2 \mid \text{Cl}_2,i-1 <2) &= 0.04 \\ P(\text{Cl}_2,i <2 \mid \text{Cl}_2,i-1 \geq 2) &= 0.00 \\ P(\text{Cl}_2,i \geq 2 \mid \text{Cl}_2,i-1 \geq 2) &= 1.00 \end{aligned}$$

The above CPT encodes how Cl_2 changes over time. The variable Cl_2 has a binary domain $\{<2, \geq 2\}$. It consists of two conditional probability distributions, one for each possible value of Cl_2 at time $i-1$. The top distribution says that if $\text{Cl}_2 < 2$ at $i-1$, then it most likely remains so, with a small chance to increase to

≥ 2 at time i . The bottom distribution says that if $Cl_2 \geq 2$ at time $i-1$, then it will definitely be ≥ 2 at time i .

A dynamic Bayesian network compiler is used to compile the network. The DBN must be compiled into a moral graph and then junction tree developed to allow us to perform inferences. This procedure is described in more detail in Jensen et al.,1990, and Xiang, 1999. Figure 28.6 shows the moral graph and Figure 28.7 the junction tree during the compilation.

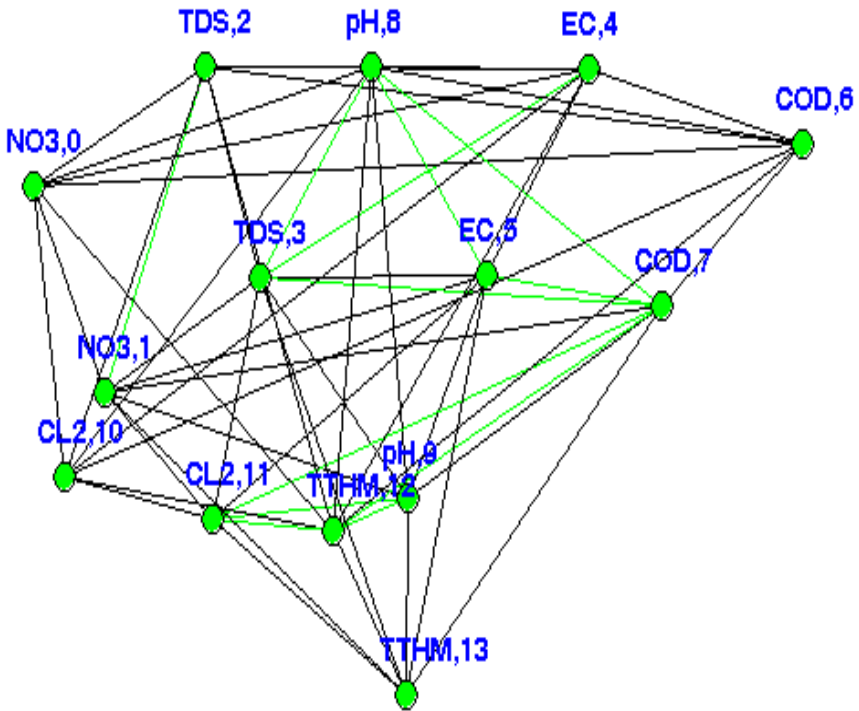


Figure 28.6 Moral graph and junction tree.

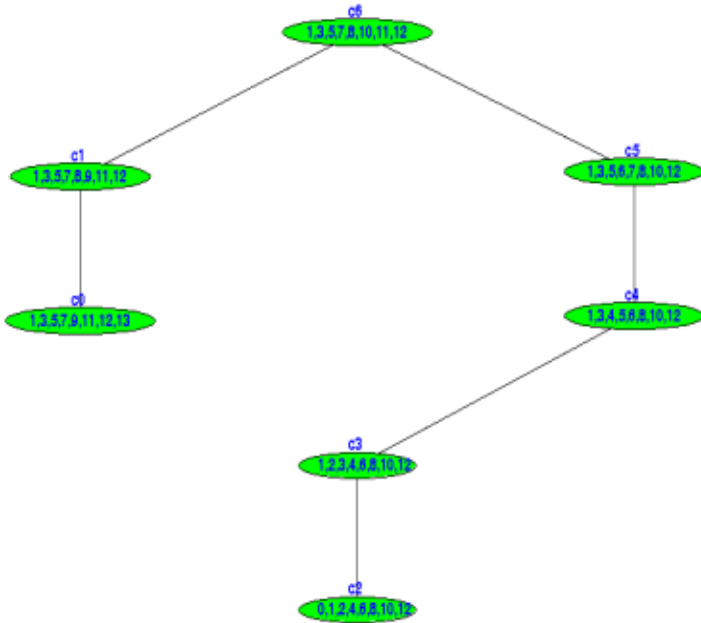


Figure 28.7 Junction tree.

28.5 Application Results

The DBN is run for ten time intervals to identify the extent to which TTHMs change as time evolves. There are two trials that are conducted, each with different initial observations. The data are pooled and used to extract the prior probability distribution and is combined with expert opinion to produce the conditional probability tables. After the conditional probabilities are determined, the DBN is constructed to allow the inference on states of variables after observations are made. The trials show the following experimental results:

- In the first trial, the observation is made that $Cl_2 < 2$ mg/L and that $COD < 3$ mg/L. When the DBN is run, we find that at the end of the first time instance, TTHMs have a high probability of being < 80 $\mu\text{g/L}$. In the next time instance, TTHMs have a slightly lower probability of being < 80 $\mu\text{g/L}$. This shows that TTHMs are increasing as time progresses. This trend continues

for the rest of the ten time intervals where the probability of TTHMs being $\geq 80 \mu\text{g/L}$ continues to increase.

- In the next trial, we observe that $\text{Cl}_2 \geq 2$ and that $\text{COD} \geq 3$ in the initial time instance to find out how these observations will effect the TTHMs concentration. After running the DBN, the probability of TTHMs being $\geq 80 \mu\text{g}$ starts very high after the first time instance. As time evolves, TTHMs increase a small amount, and start to level off at a high probability of being $\geq 80 \mu\text{g/L}$.

The network models for predicting TTHMs are developed. Probabilistic and temporal dependencies between water constituents are addressed and two simulations are run. The results show that NO_3 , the COD, Cl_2 , and pH are significant in the assessment of the TTHMs formation in the drinking water. COD and Cl_2 are the most significant precursors for formation of TTHMs, as shown in Figure 28.8.

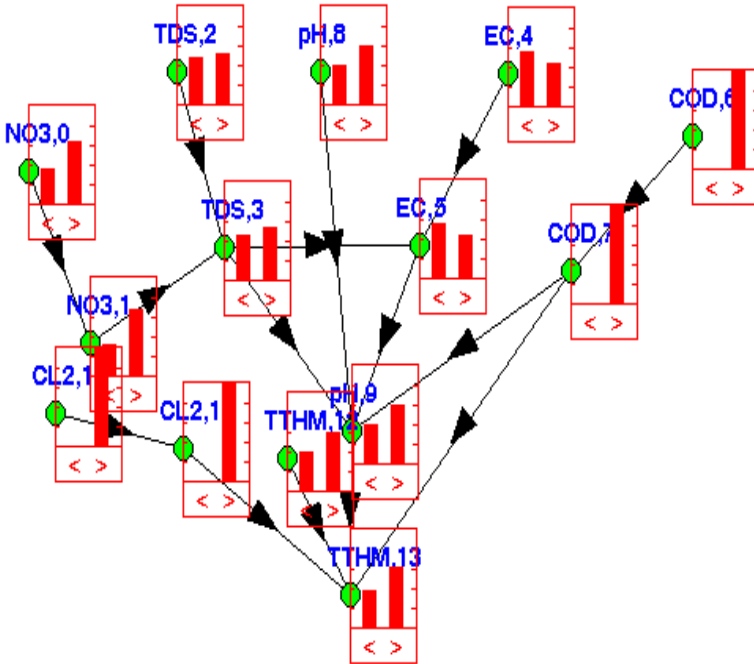


Figure 28.8 Simulation results.

As the probability distribution of COD and Cl_2 increases to the highest level, the probability distribution of TTHMs reach the highest level. This indicates

that the treatment plants need to use options such as pre-ozonation, enhanced coagulation, or activated carbon filtration treatment as effective methods for reducing TTHMs precursors, COD, NO₃, etc. A reduced chlorine dose may be used for the pre-chlorination until TTHMs lower to acceptable levels.

28.6 Conclusion

A DBN model is created using the water quality dataset in Ontario for ground-water sources of supply. The probabilistic dependencies between the constituents of water quality are reflected in DBN. The model can be used to identify TTHMs and represent the constituents of water quality and anticipate a potential problem over MCL, that is, 80 µm/L. DBN offers considerable potential for use in drinking water quality prediction. It allows for the incorporation of the causal and temporal nature of water quality dataset.

References

- Adin, A., Datzhendler, J., Alkaslassy, D., and Rav-Acha, C. Trihalomethane formation in chlorinated drinking water: a kinetic model, *wat. res.* vol. 25, no. 7, pp. 797-805, 1991.
- Amy, G.L., Chadik, P.A., and Chowdhury, Z.K. Developing models for predicting trihalomethane formation potential kinetics. *Journal of AWWA*, 79, 89-96. 1987.
- Amy, G.L., Siddiqui, M., Ozekin, K., Zhu, H., Wang, C. Empirical based models for predicting chlorination and ozonation byproducts: haloacetic acids, chloral hydrate, and bromate. EPA Report CX 819579, 1998.
- Arora H., LeChevallier, M.W., Dixon K.L. DBP occurrence survey, *Journal of the American Water Works Association*; 89(6): 60-68, 1997.
- Chang, E.E., Chao, S., Chiang, P., and Lee, J. Effects of chlorination on THM formation in raw water. *Toxicological and Environmental Chemistry*, 56, 211-225, 1996.
- Clark, R. M. and Sivaganesan. Predicting Chlorine Residuals and formation of TTHMs in Drinking Water, *Journal of Environmental Engineering*, 1998.
- Dodds, L., King, W., Woolcott, C., and Pole, J. Trihalomethanes in public water supplies and adverse birth outcomes. *Epidemiology* 10, 233-237. 1999.
- Golfinopoulos, S.K. and Arhonditsis, G.B. Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. *Chemosphere*, 47, 1007-1018, 2002.
- 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.
- Krasner, S., Krasner, M.J., McGuire, J.G., Jacangelo, N.L., Patania, K.M., Reagan, K., and Aieta, E. The occurrence of disinfection by-products in US drinking water, *J AWWA* 81 (8) (1989), pp. 41-53, 1989.

- McBean, E., Zhu, Z., and Zeng, W., *Systems Analysis Models for Disinfection by-product Formation*, Civil Engineering and Environmental Systems, Taylor & Francis, 2008.
- Morris, R.D., Audet, A.M., Angelillo, I.F., Chalmers, T.C., and Mosteller, F. Chlorination, chlorination by-products, and cancer: a meta analysis. *Am. J. Public Health* 82, 955–963, 1992.
- Nguyen, M.L., Westerhoff, P., Baker, L., Hu, Q., Esparza-Soto, M., and Sommerfeld, M., Characteristics and reactivity of algae produced dissolved organic carbon. *J. Environ. Eng.* 131 (11), 1574–1582, 2005.
- Plummer, J.D. and Edzwald, J.K. Effect of ozone on algae as precursors for trihalomethane and haloacetic acid production. *Environ. Sci. Technol.* 35 (18), 3661–3668, 2001.
- Rathbun, R.E. Regression equations for disinfection by-products for the Mississippi, Ohio and Missouri rivers. *Science of Total Environment*, 191, 235–244, 1996.
- Rodríguez, M.J., Serodes, J.B., and Morin, M. Estimation of water quality compliance with trihalomethane regulations using modeling approach. *Journal of Water Supply: Research and Technology – AQUA*, 49, 57–73, 2000.
- Rook, J.J. Formation of haloforms during chlorination of natural water, *Water Treat Exam* 23 (1974), pp. 234–236, 1974.
- Sadiq, R. and Rodríguez, 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.
- Toledano, M.B., Nieuwenhuijsen, M.J., Best, N., Whitaker, H., Hambly, P., de Hoogh, C., Fawell, J., Jarup, L., and Elliott, P., Relation of trihalomethane concentrations in public water supplies to stillbirth and birth weight in three water regions in England. *Environ. Health Perspect.* 113, 225–232, 2005.
- USEPA. Microbial and disinfection by-product rules simultaneous compliance guidance manual, EPA 815-R-99-015, 1999.
- Waller, K., Swan, S.H., DeLorenze, G., and Hopkins, B., Trihalomethanes in drinking water and spontaneous abortion. *Epidemiology* 9, 134–140, 1998.
- Wright, J.M., Schwartz, J., and Dockery, D.W. Effect of trihalomethane exposure on fetal development. *Occup. Environ. Med.* 60, 173–180, 2003
- Xiang, Y. Temporally invariant junction tree for inference in dynamic Bayesian network. Invited contribution in M. Wooldridge and M. Veloso, editors, *Artificial Intelligence Today: Recent Trends and Developments*, page 473–487, Springer, 1999.