Communicating human health risks associated with disinfection byproducts in drinking water supplies: a fuzzy-based approach

Sadiq, R.; Rodriguez, M.; Imran, S. A.; Najjaran, H.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1007/s00477-006-0069-y
Stochastic Environmental Research and Risk Assessment, 21, April 4, pp. 341-353, 2007-04-01

Communicating human health risks associated with disinfection byproducts in drinking water supplies: a fuzzy-based approach

NRCC-45607

Sadiq, R.; Rodriguez, M.; Imran, S.A.; Najjaran, H.

A version of this document is published in / Une version de ce document se trouve dans: Stochastic Environmental Research and Risk Assessment (SERRA), v. 21, no. 4, April 2007, pp. 341-353 doi: 10.1007/s00477-006-0069-y

Communicating human health risks associated with disinfection byproducts in drinking water supplies: a fuzzy-based approach

Rehan Sadiq[1*], Manuel J. Rodriguez[2], Syed A. Imran[3], and Homayoun Najjaran[1]

[1]Institute for Research in Construction, National Research Council, Ottawa, ON, Canada
[2]École Supérieure d'Aménagement du Territoire, Université Laval, Québec City, QC, Canada
[3]Centre of Sustainable Infrastructure Research, Institute for Research in Construction, National Research Council, Regina, SK, Canada

Abstract: Chlorine used for the disinfection of water supplies can react with naturally occurring organic compounds and form potentially harmful disinfection byproducts (DBPs). A risk index for two regulated groups of chlorinated DBPs - trihalomethanes (THMs) and haloacetic acids (HAAs), using fuzzy *C*-means (FCM) clustering algorithm and fuzzy rule-based modeling is proposed for risk communication. The proposed index evaluates the cancer and non-cancer risks individually for THMs and HAAs using the FCM algorithm. Subsequently, two different fuzzy rule-bases were used to evaluate the overall risk-index based on cancer and non-cancer risks. The overall risk-index will provide drinking water utilities with an effective communication tool for communicating aggregated water quality compliance. Simulated DBP occurrence data obtained from the City of Quebec, Canada is used to demonstrate the application of this methodology.

Keywords: Water quality, DBPs, fuzzy *C*-means, fuzzy rule-base, quality ordered weights, cancer and non-cancer risk.

---

1. INTRODUCTION

Chlorination has been used since the early 1900s as a primary disinfectant to inactivate pathogenic microorganisms in drinking water and as a secondary residual in the distribution networks to prevent regrowth or counter effects of intrusion of pathogens prior to consumption. However, chlorine reacts with the natural organic matter in the distributed water and forms disinfection byproducts (DBPs). More than 600 different types of DBPs have been identified. DBPs concentrations may vary by orders of magnitude depending on the disinfection practice. Some of the DBPs reported have not been identified in field-scale studies, however, they have been observed in laboratory-scale studies (Richardson, 1998).

Trihalomethanes (THMs) and haloacetic acids (HAAs) are a class of DBPs that have been implicated in a number of human health risks (cancer and non-cancer risks). THMs are formed through the reactions of hypochlorous acid (HOCl) with natural organic matter (NOM) in the presence or absence of bromide. THMs consist of chloroform (TCM), dichlorobromoform (DCBM), dibromochloroform (DBCM) and bromoform (TBM). In addition, HAAs can also form in the presence of NOM and bromide ions (from the source water) during chlorination (US EPA, 1999).

In temperate environments, THM levels in drinking water are significantly affected by seasonal conditions (Singer *et al.*, 1995; Health Canada, 1996; Arora *et al.*, 1997; Chen and Weisel, 1998; Rodriguez and Sérodes, 2001; Rodriguez *et al*, 2004). During winter and in some cases where ice cover protects surface raw waters, THMs concentrations are low due to low water temperatures and low levels of NOM. This situation arises as the chlorine demand is much lower and consequently lower chlorine doses are applied. Higher DBPs concentrations have been observed in the extremities of water distribution networks, especially in the summer season due to higher temperatures (Health Canada, 1996).

1.1. DBPs regulatory framework

The DBP regulations are promulgated based on evidence of their potential adverse human health effects, in particular cancer and reproductive disorders (Cantor *et al.*, 1998; Graves *et al.*, 2002). Routine water quality sampling helps to identify whether regulatory thresholds (guideline or standard) of DBPs are violated or not. The threshold values are based on potential toxicity of DBP indicators. A wealth of literature reporting the adverse health effects through toxicological laboratory studies is available. Some of the adverse health effects of THMs and few HAAs are summarized in Table 1.

The US EPA does not regulate individual THM levels in the drinking water. However, the maximum contaminant level (MCL) for THM stipulates that the sum of concentration of all THMs (Total Trihalomethanes or TTHMs) should not exceed 0.080 mg/L based on an annual running average of measured TTHMs in the distribution system. Similarly, HAAs in drinking water are regulated as the sum of concentrations of 5 specific haloacetic acids termed as HAA_5 (dichloroacetic acid (DCAA), trichloroacetic acid (TCAA), dibromoacetic acid (DBAA), monochloroacetic acid (MCAA) and monobromoacetic acid (MBAA)). The MCL for HAA_5 is 0.06 mg/L based on an annual running average. The upcoming Stage 2 Disinfectant/ Disinfectant Byproduct Rule (D/DBPR) by the US EPA proposes that the DBPs be evaluated for compliance based on a locational running average to ensure equal protection of all consumers regardless of their location on the distribution network. Details of these regulations can be found at US EPA website (http://www.epa.gov/).

The World health Organization (WHO, 1993) published drinking water guidelines for a few common DBPs including THMs and HAAs. In addition to guidelines for THMs, the WHO has also suggested that *the sum of the ratios of the THM levels to the guideline values should not exceed 1* (Table 2). Health Canada (2001) has set total THM levels of 0.1 mg/L as an interim maximum acceptable concentration, which serves as a guideline for Provincial regulations. No Canadian drinking water quality guideline exists for other DBPs for the time being. The Australian-New Zealand (2000) and UK (2000) drinking water standards are also summarized in Table 2 for comparison.

1.2. Water quality indices

Water quality indices (WQI) play very important role in risk communication, which are used to summarize a large quantity of data in a meaningful way (Ott, 1978). The WQI provides a communication tool for drinking water community to describe the overall status of drinking water to public. Significant literature is available on describing the overall (aggregate) water quality by an index using various statistical and mathematical techniques. Swamee and Tyagi (2000) have discussed in detail the pros and cons of different techniques and approaches available for evaluating the water quality index (WQI). Sinha *et al.* (1994) used pH, chloride concentration, turbidity, residual chlorine, conductivity and MPN (Most probable number – a bacterial counting technique) into a WQI through a weighting scheme, which can represent an overall water quality at various nodes in the distribution system. The normalized WQI (0-100) defines the overall water quality in each segment of the distribution system.

The WQI are derived based on routine water quality monitoring data, which are crucial for implementing regulatory framework effectively. From a regulatory compliance viewpoint, threshold level of contaminant (or a water quality indicator) concentration in the drinking water is established in the context of possible adverse human health impacts. For this reason it is extremely useful to relate WQI to some sort of 'acceptability' measure for drinking water, which can be interpreted as the *membership* of a fuzzy set.

Silvert (2000) argued that the concept of 'acceptability' is fuzzy in the colloquial rather than mathematical sense, since health effects can be measured far more accurately than we can evaluate their significance. A general lack of consensus exists on the definition of acceptability, which makes it more pragmatic to convert water quality monitoring data into a fuzzy membership representing the dissonance of acceptability of those water quality indicators under given conditions.

2. FUZZY SETS

Fuzzy set theory was proposed to interpret the uncertainties of the real situations carrying vagueness. The idea of fuzzy sets was coined by Zadeh (1965) and was predominantly used by researchers in electronics /computer engineering and artificial intelligence, however it became extremely popular in other disciplines in the last two decades. Recently, an immense number of applications in environmental / civil engineering have been reported in the literature where advanced assessment methods such as fuzzy synthetic evaluation are being employed for developing environmental indices.

The term synthetic is used to connote the process of evaluation whereby several individual elements and components of an evaluation are synthesized into an aggregate form; the whole is a synthesis of the parts (Liu and Lo, 2005). Simple fuzzy classification, fuzzy similarity method and fuzzy comprehensive assessment, are all subtitles of fuzzy synthetic evaluation and have been used recently by a number of researchers in various environmental areas (Sadiq and Rodriguez, 2004; Lu *et al*., 1999; Chang *et al*., 2001; Lu and Lo, 2002; Tao and Xinmiao, 1998). Fuzzy synthetic evaluation methods process all the components according to predetermined weights and decrease the fuzziness by using membership functions therefore sensitivity is quite high compared to other index evaluation techniques.

Swamee and Tyagi (2000) discussed in detail the advantages and shortcomings of different aggregation techniques available for developing environmental indices. In the aggregation process, recognition of two potential pitfalls, namely *exaggeration* and *eclipsing*, is important. Exaggeration occurs when all water quality indicators individually posses lower value (meaning in acceptable range), yet the index comes out unacceptably high. Eclipsing is the reverse phenomenon, where one or more of the indicators are of relatively high value (meaning in an unacceptable range), yet the estimated index comes out as unacceptably low. These phenomena are typically affected by the method of aggregation. Thus the challenge is to determine the best aggregation method that will simultaneously reduce both exaggeration and eclipsing. Lo and Liu (2005) pointed out that the fuzzy synthetic evaluation cannot perform well in case critical water quality indicators are leaning towards two extremes, therefore, it may not be very helpful to handle the issues if eclipsing or exaggeration occurs.

2.1 Fuzzy *C*-means algorithm

Fuzzy clustering analysis categorizes the water quality into different classes to evaluate the overall water quality. Various studies in water quality arena have been conducted using fuzzy clustering analysis, which include Kung *et al*. (1992), Melchers and Matthies (1996), and more recently Liuo and Lo (2005). The purpose of clustering is to partition the space of a given data samples. A clustering approach involves *minimization* of some objective function, or error criterion. Hard *k*-means clustering considers the elements of partition matrix using binary logic, which assigns a data point crisply either belonging to a cluster or not, i.e. {1, 0}, and therefore this technique exhibits sharp classification results. In contrast, Bezdek (1981) proposed fuzzy *C*-means (FCM) algorithm as an extension of hard *k*-means, which requires *minimization* of an error function, often called *C*-means, where "*C*" is the number of classes or clusters. In case of FCM, the inclusion and the exclusion decision is shifted with the transition region, by means of gradual fuzzy membership measure on the continuous interval [0, 1]. This is the similarity view of the membership functions, which represents a notion of being similar of the cluster or a class. Therefore in FCM the membership function measures the *degree of similarity* of an element to the classes in consideration.

For a dataset comprising of N observations, each described by vector of P, $\vec{x}_j = (x_{j1}, x_{j2},\ldots, x_{jP})$, the FCM algorithm classifies the data sets into C clusters on the basis of measured similarities among the data set. The centroid (prototype) of each cluster is the vector most representative of the cluster, i.e., it is the geometric mean of all the vectors belonging to that cluster, denoted as

$\vec{v}_i$ (Hammah and Curran, 1998). The solution of the problem of classifying data sets into C clusters can be achieved by *minimizing* the following objective function (Bezdek, 1981):

$$J_m(U,V) = \sum_{j=1}^{N} \sum_{i=1}^{C} (\mu_{ij})^m \; d^2(\vec{x}_j, \vec{v}_i); \qquad C \le N \tag{1}$$

The quantity $d^2(\vec{x}_j, \vec{v}_i)$ is a distance between observation vector $\vec{x}_j$ and the cluster centroid $\vec{v}_i$. This distance represents a measure of dissimilarity between two points, therefore when two points overlap each other this distance approaches to zero. There are various types of *distance measures* available in the literature, but the most common one is Euclidean distance, which has the following form in case of $\boldsymbol{R}^P$ space

$$d^2(\vec{x}_j, \vec{v}_i) = \sum_{p=1}^{P} (x_{jp} - v_{ip})^2 \tag{2}$$

The μ_{ji} is the membership of observation $\vec{x}_j$ in cluster i and is related to distance $d^2(\vec{x}_j, \vec{v}_i)$ and can be computed by formula

$$\mu_{ji} = \left[\left(\frac{1}{d^2(\vec{x}_j, \vec{v}_i)} \right)^{\frac{1}{m-1}} \right] \left[\sum_{k=1}^{C} \left(\frac{1}{d^2(\vec{x}_j, \vec{v}_k)} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{3}$$

The parameter $m \in [1, \propto]$ is the *degree of fuzzification* or *fuzziness index*, which regulates the degree between *similarity measures* and *distance measures* (commonly m between 1.5 and 4 are selected). Therefore the *similarity measures* between two sets can be written as

$$\left.\begin{aligned}
&\mu_{ji} = \frac{1}{\sum_{k=1}^{C} \left[\left(\vec{x}_j - \vec{v}_i \right)^2 \Big/ \left(\vec{x}_j - \vec{v}_k \right)^2 \right]^{1/m-1}}; \\
&\mu_{ji} = 1; \quad 1 \le i \le C \quad if \quad \left(\vec{x}_j - \vec{v}_i \right)^2 = 0; \\
&\mu_{ji} \in [0, 1] \quad and \quad \sum_{k=1}^{C} \mu_{jk} = 1.
\end{aligned}\right\} \tag{4}$$

To develop an index based on FCM algorithm, we define a transformation function $F(\bullet)$ on a set of measurements ($\vec{x}_j$), which maps values over an interval $\vec{w}_j$ = [0, 1] describing an "acceptability" risk criteria. The vector $\vec{w}_j$, called quality memberships, is derived by transformation function $F(\bullet)$ using threshold values (x_T) of the actual measurements $\vec{x}_j$, which can be regulatory standards, guidelines, or self-imposed limits (Liou *et al*., 2003). The *distance* *measure* is a Euclidean distance between a vector $\vec{w}_j$ (a transformed value of $\vec{x}_j$ measurements) and the prototype vector $\vec{s}_i$. Similarly, the *similarity measure* represents the resemblance between observation vector $\vec{w}_j$ and the centroid of pre-defined risk clusters $\vec{s}_i$ (Liou and Lo, 2005). The formulation for *similarity measure* (μ_{ji}) remains the same as of Equation (4) except that x is replaced by w, and v is replaced by s to account for non-commensurate units of vector $\vec{x}_j$. Therefore the *similarity measure* (μ_{ji}) can be written as

$$\mu_{ji} = \frac{1/\left(\vec{w}_j - \vec{s}_i\right)^{\left(\frac{2}{m-1}\right)}}{\sum_{k=1}^{C}\left(1/\left(\vec{w}_j - \vec{s}_k\right)^{\left(\frac{2}{m-1}\right)}\right)} \tag{5}$$

2.2 Fuzzy rule-base

The relationships between fuzzy variables can be represented by *if-then* rules of the form "*If antecedent proposition then consequent proposition*". In a linguistic model with both antecedent and consequent as fuzzy propositions, the rule can be written as follows:

R_i: *If* x_1 is A_i and x_2 is A_j *then* y is B_k; $i = 1, 2, \ldots, K$ (6)

The antecedents x_1 and x_2 (inputs) and the consequent y (output) are linguistic variables whereas their respective counterparts A_i, A_j and B_k are qualitative linguistic constants e.g., *low* (L), *medium* (M) and *high* (H), representing various clusters. The truth-value (real number) of these propositions depends on the *degree of similarity* between the variables and the constants. To understand the inferencing in fuzzy systems, consider the following rule-base

If x_1 is *low*	and	x_2 is *low*	then	y is *low*
If x_1 is *low*	and	x_2 is *medium*	then	y is *low*
If x_1 is *low*	and	x_2 is *high*	then	y is *low*
If x_1 is *medium*	and	x_2 is *low*	then	y is *medium*
If x_1 is *medium*	and	x_2 is *medium*	then	y is *medium*
If x_1 is *medium*	and	x_2 is *high*	then	y is *high*
If x_1 is *high*	and	x_2 is *low*	then	y is *medium*
If x_1 is *high*	and	x_2 is *medium*	then	y is *high*
If x_1 is *high*	and	x_2 is *high*	then	y is *high*

For fuzzy composition, various aggregation operators can be used in sequences (Klir and Yuan, 1995) to make inferencing. Based on the above rules, the fuzzy composition for *y* can be written as

$$\left.\begin{aligned}
\mu_y^L &= \left(\mu_{x1}^L \otimes \mu_{x2}^L\right) \oplus \left(\mu_{x1}^L \otimes \mu_{x2}^M\right) \oplus \left(\mu_{x1}^L \otimes \mu_{x2}^H\right) \\
\mu_y^M &= \left(\mu_{x1}^M \otimes \mu_{x2}^L\right) \oplus \left(\mu_{x1}^M \otimes \mu_{x2}^M\right) \oplus \left(\mu_{x1}^H \otimes \mu_{x2}^L\right) \\
\mu_y^H &= \left(\mu_{x1}^M \otimes \mu_{x2}^H\right) \oplus \left(\mu_{x1}^H \otimes \mu_{x2}^M\right) \oplus \left(\mu_{x1}^H \otimes \mu_{x2}^H\right)
\end{aligned}\right\} \tag{7}$$

Therefore, output *y* can be written as a fuzzy set (μ_y^L , μ_y^M , μ_y^H). The signs $\otimes$ and $\oplus$ represent *and*-type (intersection-based) and *or*-type operators (union-based) operators, respectively. The commonly used *and*-type operators are *product* and *minimum*. Similarly, the *or*-type operators are *sum* and *maximum*. Therefore, the common types of inferencing in fuzzy composition are *sum-product* and *max-min* rules. For sake of simplicity we use *sum-product* fuzzy composition.

2.3 Defuzzification

A process known as *defuzzification* can be used to calculate the crisp value of a fuzzy set. Defuzzification is an important step in fuzzy rule-based modelling. Many defuzzification techniques are available (Chen and Hwang, 1992). Cheng and Lin (2002) used the *maximum* operator to determine classification of fuzzy subsets from a fuzzy set. The other common defuzzification methods in practice are centre of area method (Yager, 1980), first of maximum, last of maximum, and middle of maximums. A weighted average approach (scoring method) can also be used to determine a crisp index value by assigning specific quality-ordered weights *q* to *similarity measures* or memberships (Lu *et al*., 1999; Silvert, 2000; Sadiq and Rodriguez, 2004). Therefore, the index can be derived as

$$Index = \left(\sum_{i=1}^{C} \mu_{ji} * q_i\right) \tag{8}$$

The final index value can be adjusted by specific quality-ordered weights or utility values $q_i \in [0, 1]$. For example for three pre-defined clusters, *low*, *medium* and *high*, specific quality-ordered weights of $q_1 = 0$, $q_2 = 0.5$ and $q_3 = 1$ can be used, respectively to evaluate final crisp index value.

3. PROPOSED APPROACH FOR THE DEVELOPMENT OF RISK INDEX

The fuzzy algorithms briefed in the previous section are used to develop risk index for 6 DBPs given in Table 1. The proposed methodology is a five-step process, which is shown schematically in Figure 1. A step-by-step methodology is explained in this section as follows:

Step 1 - Transformation

Step 1 involves transformation of DBPs concentrations $\vec{x}_j$ into quality memberships $\vec{w}_j$. A simple transformation $F(\bullet)$ using linear function is proposed (Figure 2). The transformation function $F(\bullet)$ requires threshold values (x_T) for each DBP compound for both cancer and non-cancer effects, which are derived based on toxicological data provided in Table 3. These threshold values are also provided in the table of Figure 2. This step converts various non-commensurate concentrations and of various significance into the unit interval [0, 1] which will be used for clustering.

Step 2 – Fuzzy clustering

In Step 2, the centroids (s_i) of three clusters *low* (L), *medium* (M) and *high* (H) risks are defined as $s_1 = 0$, $s_2 = 0.5$, and $s_3 = 1$, respectively for both cancer and non-cancer effects. The *similarity measures* (memberships) to these clusters can be determined using Equation (5) for each DBP. The similarity measure or fuzzy membership for cancer and non-cancer effects can be written as $\left(\mu_j^k\right)_i$; where k = L, M, or H; j = cancer or non-cancer; and i = 1 (for THMs) or 2 (for HAAs).

Step 3 – Stage 1 Fuzzy rule-base

A similar fuzzy rule-base algorithm explained earlier in section 2.2 can be used separately for both cancer and non-cancer effects to combine the effects of both DBPs groups - THMs and HAAs. The rule bases are provided in Figure 3. The rules can be interpreted as "*if* THMs are of level 'column' *and* HAAs are of level 'row' *then* effects (cancer or non-cancer) are of level 'intersection of column and row'. For example, for cancer risk rule-base (stage 1), the first rule can be read as "***if*** THMs are at *low risk* ***and*** HAAs are at *low risk* ***then*** the overall risk of cancer is *low*", and so on. This step helps us to determine memberships functions for cancer $\left(\mu_{cancer}^k\right)$ and non-cancer risks $\left(\mu_{non-cancer}^k\right)$, where k = L, M, or H.

Step 4 – Stage 2 Fuzzy rule-base

This step is similar to Step 3, except that the third rule-base (stage 2) given in Figure 3 is used to derive the memberships or similarity measures for overall risk. The first rule in this rule-base can be interpreted as "***if*** cancer risk is *low* ***and*** non-cancer risk is *low* ***then*** the overall risk is *low*", and so on. This step evaluates the final similarity measures $\left(\mu_{risk}^L, \mu_{risk}^M, \mu_{risk}^H\right)$ for three pre-defined clusters.

Step 5 – Developing risk index using defuzzification

The crisp value of risk index *I*, can be determined by Equation (8), which can also be expanded as:

$$I = 0 \times \mu_{risk}^L + 0.5 \times \mu_{risk}^M + 1 \times \mu_{risk}^H \qquad (9)$$

In this study, the coefficients (quality ordered weights) in Equation 9 were assigned arbitrarily based on the authors' experience and recent research (Sadiq and Rodriguez, 2004). However, guidelines may be established for the risk index based on expert opinion (Lu *et al*., 1999). The higher value of *I* represents higher risk. Assigning larger weights to the memberships of "*high*" risk similarity measure represents a *risk-averted* attitude of the decision-makers. If larger quality-ordered weights are assigned to the memberships of "*low*" risk, then it represents a *pro-risk* attitude of the decision-makers. If equal weights are assigned to the similarity measures, it implies that the decision-maker is indifferent and represents a *compromising or normative* *attitude*. The details on attitudinal decision-making are beyond the scope of this paper. But in this study, a *risk-averted* approach is used and weights are assigned in the descending order of risk, i.e., the largest weight is assigned to μ^{H}_{risk} and the smallest weight is assigned to μ^{L}_{risk} .

4. APPLICATION

The algorithms developed in the previous sections were applied to a case study in the Quebec City region of Canada. The application relies on data on chlorinated DBPs generated under experimental chlorination of water of a Quebec City utility. The data represent the simulated occurrence of DBPs in the distribution system of the utility. The Quebec City utility takes the water from a river with high colour content. The water treatment process involves pre-chlorination, coagulation, sedimentation, filtration, disinfection with ozone, and post-disinfection with chlorine.

4.1. Data collection

Eighteen samples of post-ozonated water were chlorinated from Feb. 2001 to Jan. 2002. Once each sample was collected, it was then subjected to bench-chlorination with 2.5 mg Cl_2/L dose (using sodium hypo-chlorite). This represents the maximum applied post-chlorination dose. The water temperature measured in the field was reproduced in the laboratory and the pH was set to 7.5 (the average value found in distribution systems for both utilities). After chlorination, THMs and HAAs were measured following six different contact times varying from 0.25, 1, 2, 6, 24 and finally to 48 hrs.

Table 4 presents the statistical summary of the chlorinated DBPs identified. Predominantly non-brominated DBPs were identified and attributed to low bromide levels in the source water (<10 μg/L). As expected, the concentrations of THMs and HAAs increased with contact time. However, the levels of most species stabilized after 24 hrs of contact. The data are reported with respect to season and chlorine contact time. The contact time is used as a surrogate for water residence time in the distribution system, which shows spatial variability in DBP concentration (Rodriguez *et al*., 2004).

THMs and HAAs were analysed according to the US EPA methods 551.1 and 552.2, respectively (US EPA, 1990; US EPA, 1995). Analysis for THMs and HAAs was conducted using two Perkin Elmer auto-system XL gas chromatographs with electron capture detectors (Rodriguez and Sérodes, 2001; Sérodes *et al*., 2003). For THM species, analytical protocols ensured detection limits of 0.5 μg/L for TCM and of 0.3 μg/L for BDCM, DBCM, and TBM. Detection limits for DCAA and TCAA were 1.1 and 0.6, respectively. The concentrations of other HAAs were always below the detection limits.

4.2. Evaluating risk index

To determine the risk index, two parameters *m* (fuzziness index) and *q* (quality ordered weights are predefined. We used m =2.5 and q = (0, 0.5, 1) for memberships to *low*, *medium* and *high* risk to carry out analysis. Figure 4 shows the variations in risk index for 18 individual samples corresponding to each of the 6 contact times (i.e. 0.25, 1, 2, 6, 24 and 48 hrs). Values of risk index ware observed consistently high for contact (residence) time more than 1 day. Figure 5 demonstrates this aspect more clearly by plotting mean risk index (μ) for 18 sample data points for given contact time. A confidence band using standard deviation (σ) is also shown to view the variations. The ($\mu \pm \sigma$) is a confidence band developed using actual concentration of 18 data samples corresponding to each contact time.

It is important for water utilities to minimize the contact time of chlorine in water distribution system to reduce the DBPs production and comply with regulatory limits (by reducing residence time in storage tanks, reducing residence time by stagnation of water in dead-ends, changing the position of chlorine boosters, etc.). For a given distribution system, a threshold value of risk index can be established by regulators, based on which they can determine what is the maximum allowable residence time. This threshold level of risk index represents an acceptable risk, which should be economically and technically justifiable.

4.3 Sensitivity analyses

Sensitivity analyses are performed in two steps. In the first step, the values of fuzziness index (m) and quality ordered weights (q) are varied to see the variability in mean value of risk index using DBPs data provided in Table 4. The results of sensitivity analysis are provided in Table 5. It can be noticed for high value of $m = 3.5$, the variation in mean risk index value is limited. For example for q (0, 0.5, 1), the variability was between 0.17 to 0.37 at contact hours of 0.25 and 48 hrs, respectively. In comparison, for $m = 1.5$ and 2.5, this variability was 0.08 to 0.36 and 0.12 to 0.36, respectively. Similar observations can be made for other combinations of quality ordered weights. Therefore, a value of fuzziness index $m < 2.5$ can be a good choice to make the risk index sensitive. In addition for quality ordered weights, the advantage of selecting q (0, 0.5, 1) is that it helps scaling risk index estimates in the interval of [0, 1], which makes interpretation easy.

In the second step, we fixed $m = 2.5$ and q (0, 0.5, 1) and varied the concentrations of different DBPs to observe the sensitivity of risk index. Table 6 shows an increase (decrease) in risk index by increasing (decreasing) the concentration of individual DBPs. To observe this variability, the concentration of each DBP is increased from a predefined base value, which is selected as half of the threshold estimates used for cancer risk (see table of Figure 2). The base estimate of risk index is calculated as 0.51. First, the concentration of TCM is increased to its threshold estimate (i.e., 57 ppb), the risk index changed to 0.77, with a percentage change of more than 50%. It shows that risk index is very sensitive and can recognize an unusual increase (decrease) in the DBP concentrations. We changed the values of each DBP to their threshold values stepwise and observed a continuous increase in the risk index. It can also be noticed that when concentrations of all DBPs reached to their threshold values, the risk index still did not reach to the maximum value of 1. This happened because non-cancer risk thresholds are different from cancer risk thresholds. If the DBPs concentrations are high enough to surpass both thresholds, the risk index will become exactly 1. As mentioned earlier that in the aggregation process two pitfalls *exaggeration* and *eclipsing* are common encountered. The proposed risk-index is very sensitive and has a capability to address *exaggeration* and *eclipsing* effectively.

4.4 Risk index as function of contact time

The calculation of the risk index can be simplified by using commonly available spreadsheet applications. For similar initial water quality conditions, a relationship between contact time and risk index is developed. Figure 6 shows three simple models, which are fitted to 108 (18×6) data points and superimposed on the mean risk index curve. The advantage of developing these models is that for any given residence time, risk index can be evaluated. These fitted models include exponential, uni-variate and multi-variate (second degree polynomial) models, which can be written as following:

Exponential model: $I_1 = 1 - e^{-0.35T}$ (10)

Uni-variate model: $I_2 = 0.16 + 0.006T$ (11)

Multi-variate model: $I_3 = 0.12 + 0.019T - 0.0003T^2$ (12)

For example, consider a water utility attempting to maintain drinking water supply at a residence time of 12 hrs. By using T = 12 hrs in above three models, the estimated risk indices are $I_1 = 0.985$, $I_2 = 0.232$, and $I_3 = 0.305$. The Exponential model is very conservative and predicts very high estimated risk-index after certain contact hours (Figure 6) while the Uni-variate model does not consider the inherent non-linearity present in the risk-index development. We recommend multi-variate model, which fits best to the data and takes into account the non-linearity in the risk-index development.

Other generalized models can be developed to relate the effects of seasonal changes, chlorine dose, and water quality on the risk index. Using the aggregate risk-index utilities can proactively control conditions that might lead to unacceptable DBP risks as well as effectively communicate this risk to consumers.

5. CONCLUSIONS

In this paper, a new methodology is proposed to develop an index that provides a surrogate for human health risk associated with the two major groups of chlorinated DBPs present in drinking water, i.e., THMs and HAAs. The risk index is developed using fuzzy *C*-means (FCM) clustering algorithm and fuzzy rule-based model, which is sensitive and capable of capturing unusual increase (or decrease) in the individual DBPs concentrations. Therefore, the proposed risk index adequately addresses the issues of *exaggeration* and *eclipsing* in the aggregation process.

The methodology was applied for a case where THMs and HAAs occurrence are experimentally simulated with treated water of the Quebec City water utility. Results are very encouraging and suggest that the approach can be applied to full-scale data and can be extended to other chlorinated and non-chlorinated DBPs.

The approach used for developing risk index in this paper is generic and it can also be used for aggregating physico-chemical and microbiological water quality data regularly monitored. The advantage of using this approach is that it provides a single index value, which is a surrogate of overall water quality to make risk-informed decisions. The temporal and spatial variation of risk index in the distribution system can help to locate most sensitive points for monitoring/ sampling of water quality, which are particularly important for epidemiologists and regulators.

Indexing results obtained from the proposed methodology significantly depends on model and parameter choices. In future work, it will be important to improve the applicability of the methodology based on expert panels constituted by toxicologist, epidemiologist, environmental engineers and regulators.

6. REFERENCES

Arora, H, LeChevallier, M.W., and Dixon, K.L. 1997. DBP occurrence survey, *Journal of the American Water Works Association*, 89(6): 60-68.

Aus-NZ 2000. Australian drinking water guidelines, Australian National Health and Medical Research Council, http://www.health.gov.au/nhmrc/publications/synopses/eh19syn.htm.

Bezdek J.C. 1981. *Pattern recognition with fuzzy objective function algorithms*, New York: Plenum.

Cantor, K.P., Lynch, C.F., Hildesheim, M.E., Dosemeci, M., Lubin, J., Alavanja, M., and Craun, G. 1998. Drinking water source and chlorination by-products I. risk of bladder control, *Epidemiology*, 9(1): 21-28.

Chang, N.B., Chen, H.W., and Ning, S.K. 2001. Identification of river water quality using the fuzzy synthetic evaluation approach, *Journal of Environmental Management*, 63:293–305.

Chen, S. J., and Hwang, C.L. 1992. *Fuzzy Multiple Attribute Decision-Making*, Springer-Vaerlag, NY.

Chen, W.J., and Weisel, C.P. 1998. Halogenated DBP concentrations in a distribution system, *Journal of the American Water Works Association*, 90(4): 151-163.

Cheng, C-H, and Lin, Y. 2002. Evaluating the best main battle tank using fuzzy decision theory with linguistic criteria evaluation, *European Journal of Operation Research*, 142: 174-186.

Graves, C.G., Matanoski, G.M., and Tardiff, R.G. 2002. Weight of evidence for an association between adverse reproductive and developmental effects and exposures to disinfection by-products: a critical review, *Regulatory Toxicology and Pharmacology*, 34: 103-124.

Hammah, R.E., and Curran, J.H. 1999. On distance measures for the fuzzy k-means algorithm for joint data, Rock Mechanics and Rock Engineering, 32(1): 1–27.

Health Canada 1996. A one-year survey of halogenated disinfection byproducts in the distribution system of treatment plants using three different disinfection processes, *Minister of Public Works and Government Services Canada*, H46-2/96-206E.

Health Canada 2001. *Summary of guidelines for drinking water quality*, Federal-Provincial subcommittee, www.hc-sc.gc.ca/ehd/bch/water_quality.htm.

Klir, G.J., and Yuan, B. 1995. *Fuzzy sets and fuzzy logic - theory and applications*, Prentice- Hall, Inc., Englewood Cliffs, NJ, USA.

Kung, H.T., Ying, L.G., Liu, Y.C. 1992. A complementary tool to water quality index: fuzzy clustering analysis. *Water Resources Bulletin*, 28 (3): 525–533.

Liou, S.M., Lo, S-L., and Hu, C.Y. 2003. Application of two-stage fuzzy set theory to river quality evaluation in Taiwan, *Water Research*, 37: 1406–1416.

Liou,Y-T, and Lo, S-L. 2005. A fuzzy index model for trophic status evaluation of reservoir waters, *Water Research*, 39(7): 1415-1423.

Lu, R-S., and Lo, S-L. 2002. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory, *Water Research*, 36: 2265-2274.

Lu, R-S., Lo, S-L, and Hu, J-Y. 1999. Analysis of reservoir water quality using fuzzy synthetic evaluation, *Stochastic Environmental Research and Risk Assessment*, 13: 327-336.

Melcher, D., and Matthies, M. 1996. Application of fuzzy clustering to data dealing with phytotoxicity. *Ecological Modeling*, 85: 41–49.

Ott, W.R. 1978. *Environmental indices: theory and practice*, Ann Arbor Science Publishers Inc., pp. 371.

Richardson S.D. 1998. *Drinking Water Disinfection Byproducts*, John Wiley's encyclopedia of environmental analysis and remediation, Ed. Robert, A.M. and Richard, S.D., John Wiley's, 3: 1398-1421.

Rodriguez, M.J., and Sérodes J.B. 2001. Spatial and temporal evolution of trihalomethanes in three water distribution systems, *Water Research*, 35(6): 1572-1586.

Rodriguez, M.J., Sérodes, J.B., and Levallois, P. 2004. Behavior of trihalomethanes and haloacetic acids in a drinking water distribution system, *Water Research*, 38: 4367-4382.

Sadiq, R., and Rodriguez, M.J. 2004. Fuzzy synthetic evaluation of disinfection by-products – a risk-based indexing system, *Journal of Environmental Management*, 73(1): 1-13.

Serodes, J-B, Rodriguez, M.J., Li, H. and Bouchard, C. 2003. Occurrence of THMs and HAAs in experimental chlorinated waters of the Quebec City area (Canada), *Chemosphere*, 51: 253-263.

Silvert, W. 2000. Fuzzy indices of environmental conditions, *Ecological Modelling*, 130(1-3): 111-119.

Singer, P.C., Obolensky, A., and Greiner, A. 1995. DBPs in chlorinated North Carolina drinking water, *Journal of the American Water Works Association*, 87(10): 83-92.

Sinha, R., Gupta, P., and Jain, P.K. 1994. Water quality modeling of a city water distribution system, *Indian Journal of Environmental Health*, 36(4): 258-262.

Swamee, P.K., and Tyagi, A. 2000. Describing water quality with aggregate index, *Journal of Environmental Engineering* ASCE, 126(5): 451-455.

Tao, Y., and Xinmiao, Y. 1998. Fuzzy comprehensive assessment, fuzzy clustering analysis and its application for urban traffic environment quality evaluation, *Tansportation Research*, 3(1): 1–57.

U.S. EPA 2006. *Integrated Risk Information* (IRIS). Online database at http://www.epa.gov./ngispgm3/iris.

United Kingdom 2000. *Water supply regulations for England and Wales*, www.dwi.detr.gov.uk/regs/si3184/3184.htm.

US EPA 1990. *Determination of chlorination disinfection by-products and chlorinated solvents in drinking water by liquid-liquid extraction and gas chromatography with electron-capture detection*, Environmental monitoring systems laboratory office of research and development, Method 551.1, United States Environmental Protection Agency, Cincinnati, Ohio.

US EPA 1995. *Determination of haloacetic acids in drinking water by liquid-liquid extraction and gas chromatography with electron-capture detection*, National exposure research laboratory office of research and development, Method 552.2, United States Environmental Protection Agency, Cincinnati, Ohio.

US EPA 1999. *Microbial and disinfection by-product rules – simultaneous compliance guidance manual*, United States Environmental Protection Agency, EPA 815-R-99-015.

US EPA 2001. *National primary drinking water standards*, United States Environmental Protection Agency, EPA 816-F-01-007.

WHO 1993. *Guidelines for drinking-water quality*, 2nd Edition, Volume 1, Recommendations, World Health Organization, Geneva.

Yager, R.R. 1980. A general class of fuzzy connectives, *Fuzzy Sets and Systems*, 4: 235-242.

Zadeh, L.A. 1965. Fuzzy sets, *Information and Control*, 8 (3): 338-353.

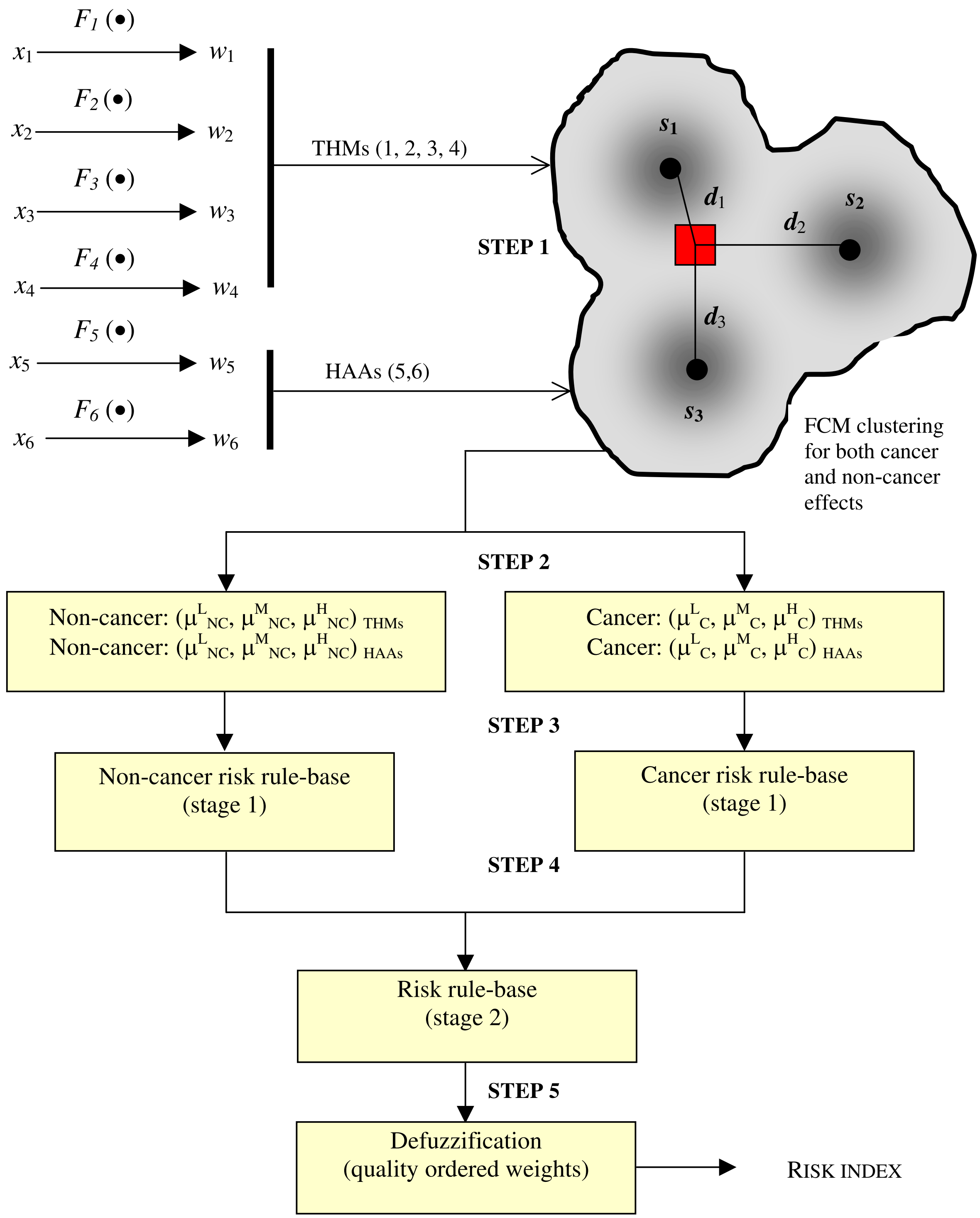
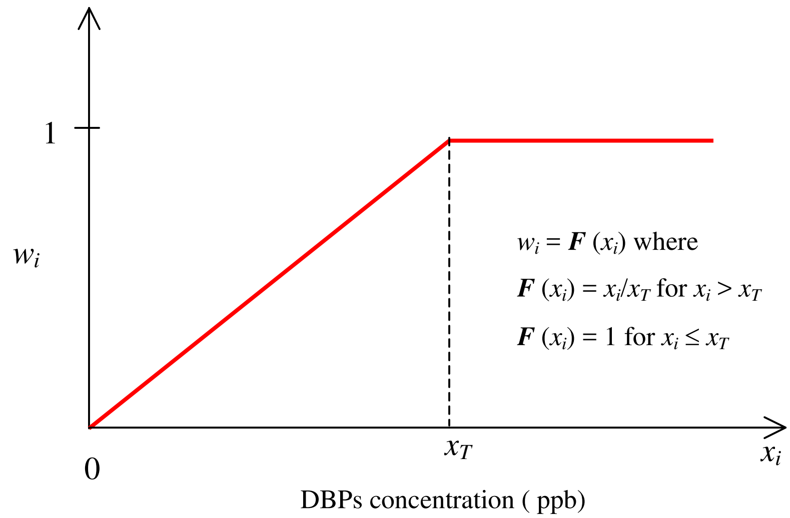


Figure 1. Proposed approach for developing risk index for DBPs



	Threshold concentration (x_T)	
DBPs	Cancer (ppb)	Non-cancer (ppb)
TCM	57	35
DBCM	4	70
BDCM	6	6
TBM	40	70
DCAA	20	20
TCAA	40	40

Figure 2. Threshold concentration (x_T) for DBPs used for transformation w_i [0, 1]

Cancer risk rulebase

THMs ↓ / HAAs →	L	M	H
L	L	L	M
M	M	M	M
H	H	H	H

Non-cancer risk rulebase

THMs ↓ / HAAs →	L	M	H
L	L	L	L
M	M	M	H
H	H	H	H

Stage 1

Overall risk rulebase

Cancer ↓ / Non-cancer →	L	M	H
L	L	L	M
M	M	M	H
H	M	H	H

Stage 2

Figure 3. Fuzzy rule-bases for estimating overall risk

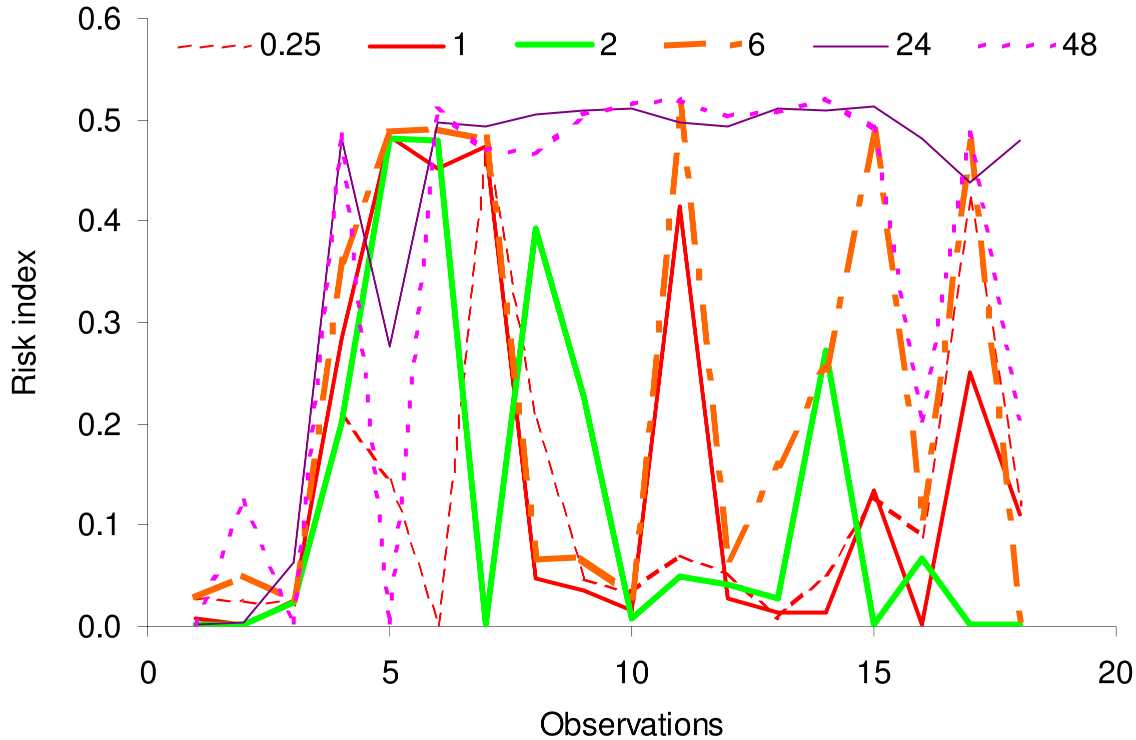
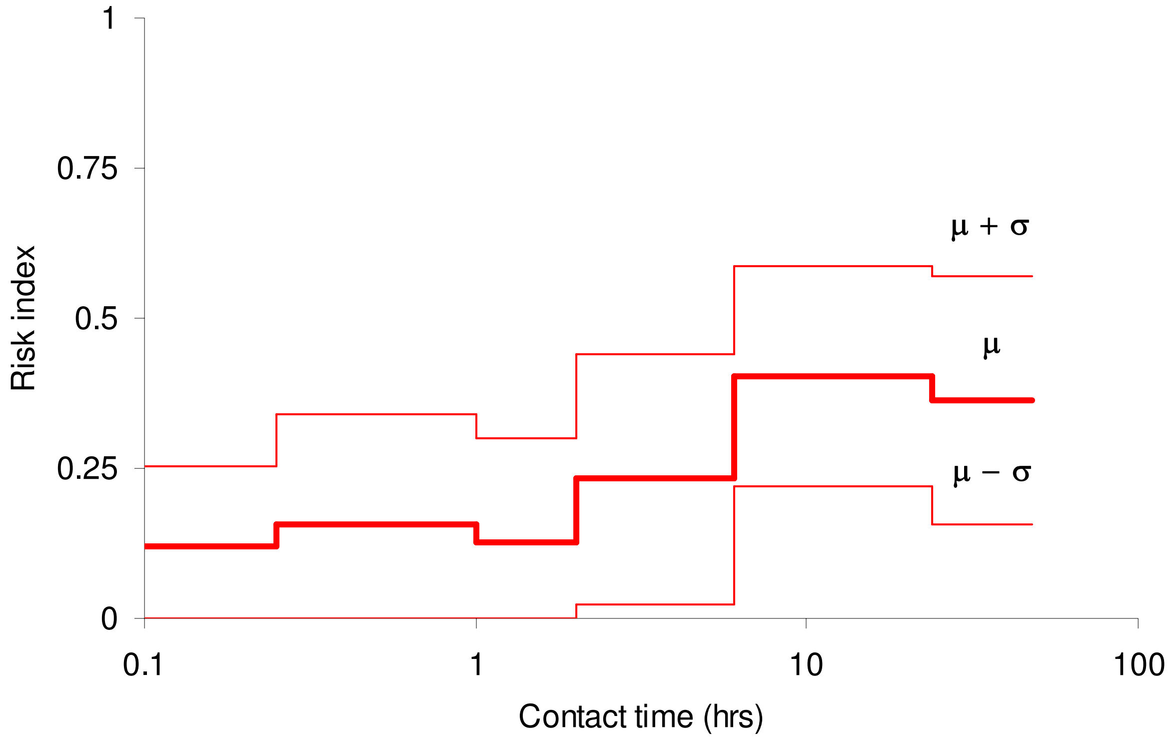


Figure 4. Variations in risk index for individual samples with respect to contact time



Contact time (hr)	0.25	1	2	6	24	48
μ	0.119	0.155	0.127	0.232	0.404	0.363
σ	0.136	0.185	0.172	0.208	0.184	0.206

Figure 5. Statistical variation of risk index with respect to contact time

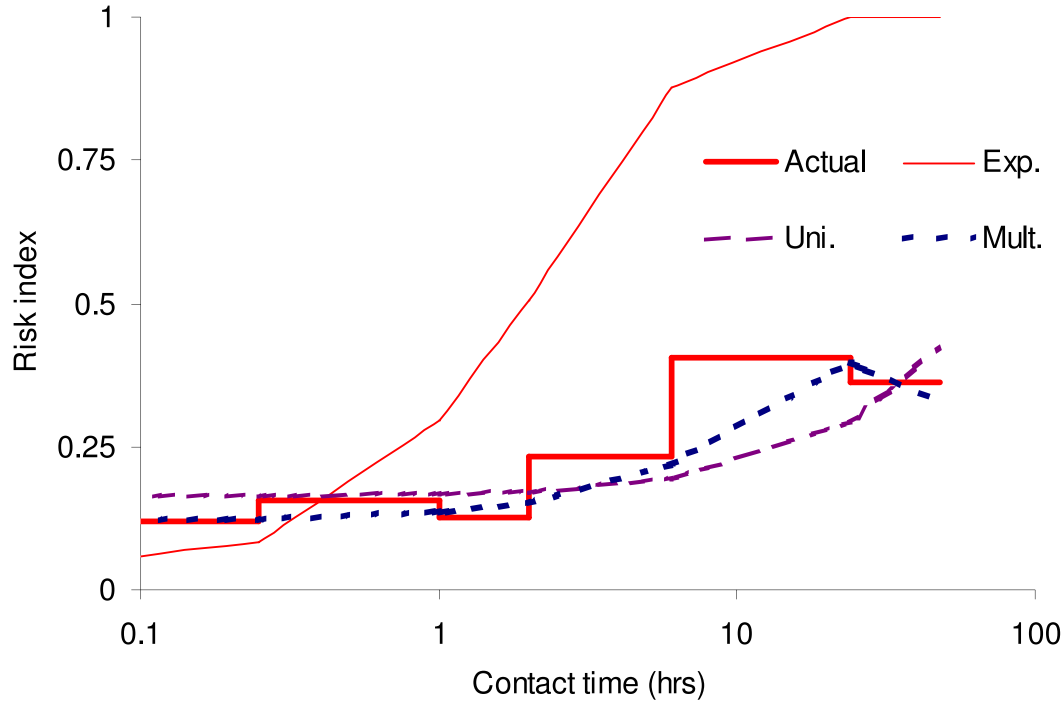


Figure 6. Fitted models to risk index as a function of contact time

Table 1. Toxicological information on DBPs (Sadiq and Rodriguez, 2004))

DBPs	Compound	Detrimental effects
Trihalomethanes (THMs)	Chloroform	Cancer, liver, kidney, and reproductive effects
	Dibromochloromethane	Nervous system, liver, kidney, and reproductive effects
	Bromodichloromethane	Cancer, liver, kidney, and reproductive effects
	Bromoform	Cancer, nervous system, liver and kidney effects
Haloacetic acids (HAAs)	Dichloroacetic acid	Cancer, reproductive and developmental effects
	Trichloroacetic acid	Liver, kidney, spleen and developmental effects

Table 2. DBPs (ppb) regulations in various jurisdictions of the world

Compound	Acronym	WHO (1993)	US EPA (2001)	Canada (2001)	Aus-NZ (2000)	UK (2000)
Trichloromethane (chloroform)	TCM	200	0[2]			
Dibromochloromethane	DBCM	100	0[2]			
Bromodichloromethane	BDCM	60	60[2]			
Tribromomethane (bromoform)	TBM	100	0[2]			
Total trihalomethanes	TTHM	$\sum_{i=1}^{4} \frac{THM}{WHO} \leq 1$	80	100	250	100
Dichloroacetic acid	DCAA	50			100	
Trichloroacetic acid	TCAA	100			100	
Haloacetic acids	HAA_5		60	[1]		

1: Under consideration

2: Maximum contaminant level goals (MCLG)

Table 3. Carcinogenic and non-cancer risk information for selected DBPs

Disinfection by-products (DBPs)	Carcinogenic potential	[1]Risk (1×10^{-6}) (μg/L)	[2]UF	[3]MF	RfD (mg/kg/day)	[4]RQ (μg/L)
†Chloroform	Probable (B2)	[5]5.7	10^3	1	1×10^{-2}	350
†Dibromochloromethane	Possible (C)	0.4	10^3	1	2×10^{-2}	700
†Bromodichloromethane	[6]Not classifiable (D)	[8]0.6			No data	[7]60
†Bromoform	Probable (B2)	4	10^3	1	2×10^{-2}	700
*Dichloroacetic acid	Probable (B2)	[8]0.5			No data	[9]50
*Trichloroacetic acid	Possible (C)	[8]1			No data	[9]100

1: Concentration corresponding to unit risk of 1 in a million
2: Uncertainty factor; 3: Modifying factors - these are factors used to convert NOAEL and LOAEL into RfD
4: Concentration estimated at risk quotient (RQ) = 1;
where RQ = Dosage/RfD and Dosage = [Concentration × intake rate/body weight]
5: Estimated from slope factor (SF) of 0.0061 $(mg/kg/day)^{-1}$; where Risk = Dosage × SF
6: Also classified as B2 (see IRIS, US EPA, 2006)
7: US EPA (2001) and WHO (1993)
8: Derived from RQ/100
9: WHO (1993)
†: Included in THMs (a group of 4 trihalomethanes, the US EPA (2001) recommended the value of 80 μg/L)
*: Included in HAA_5 (a group of 5 haloacetic acids, the US EPA (2001) recommended the value of 60 μg/L)

Table 4. Statistical summary (average and standard deviation) for DBPs species for waters of Quebec City for around the year sampling

Contact time (hr)	DCAA (ppb)	TCAA (ppb)	TCM (ppb)	DBCM (ppb)	BDCM (ppb)	TBM (ppb)
0.25	13.7 (9.9)*	11.4 (7.9)	19.3 (15.6)	1.1 (2.2)	1.7 (1.2)	n.d**
1	12.7 (10.5)	10.5 (9.5)	24.8 (18.3)	0.7 (1.7)	2.1 (1.5)	n.d
2	10.9 (10)	9.2 (9.2)	26.6 (18.6)	0.6 (1)	2.4 (1.2)	n.d
6	17.7 (13)	15.6 (12.1)	30.7 (24.4)	1.1 (2.1)	2.6 (1.6)	n.d
24	25.1 (11)	22.8 (12)	32.0 (26.1)	1.1 (1.7)	2.6 (1.6)	n.d
48	25.0 (13.8)	22.9 (14.9)	39.4 (34.8)	0.8 (1)	2.6 (1.6)	n.d
Overall	17.5 (12.6)	15.4 (12.3)	28.8 (24.1)	0.9 (1.6)	2.3 (1.5)	n.d

* Values in the parenthesis show standard deviation of 18 samples collected over a year for each contact time

** Values always below the detection limit

Table 5. Variation in the mean value of risk index by changing *m* and *q*

	q (0, 0.5, 1)*			*q* (0.25, 0.5, 0.75)			*q* (0.1, 0.3, 0.6)		
m →	1.5	2.5[$]	3.5	1.5	2.5	3.5	1.5	2.5	3.5
T ↓	Mean value of risk index								
0.25	0.08	0.12	0.17	0.29	0.31	0.34	0.13	0.15	0.17
1	0.17	0.16	0.18	0.33	0.33	0.34	0.17	0.16	0.18
2	0.13	0.13	0.16	0.31	0.31	0.33	0.15	0.15	0.17
6	0.21	0.23	0.27	0.35	0.37	0.38	0.18	0.19	0.21
24	0.40	0.40	0.40	0.45	0.45	0.45	0.26	0.26	0.27
48	0.36	0.36	0.37	0.43	0.43	0.43	0.25	0.25	0.25

* Represents quality ordered weights assigned to *low*, *medium* and *high* risk in defuzzification (step 5)

[$] The risk index developed in this paper is based on *m* = 2.5 and *q* (0, 0.5, 1)

Table 6. Sensitivity analysis for risk index by changing individual DBPs concentrations

*TCM	DBCM	BDCM	TBM	DCAA	TCAA	**Risk index (I)	$\Delta^{\$}$
28	2	3	20	10	20	0.51	0.0
57	2	3	20	10	20	0.77	51.0
57	**4**	3	20	10	20	0.85	66.7
57	**4**	**6**	20	10	20	0.97	90.2
57	**4**	**6**	**40**	10	20	0.97	90.3
57	**4**	**6**	**40**	**20**	20	0.98	92.2
57	**4**	**6**	**40**	**20**	**40**	0.99	94.1

* concentration of DBPs are in ppb

$ $\Delta = \frac{|0.51 - I| \bullet 100}{0.51}$ where 0.51 is the base value of risk index for the first scenario (row 1)

** The risk index developed in this paper is based on m = 2.5 and q (0, 0.5, 1)