NRC Publications Archive Archives des publications du CNRC

Interpreting drinking water quality in the distribution system using Dempster-Shafer theory of evidence

Sadiq, R.; Rodriguez, M. J.

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version
acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien
DOI ci-dessous.

Publisher's version / Version de l'éditeur:

<https://doi.org/10.1016/j.chemosphere.2004.11.087>

Chemosphere, 59, April 2, pp. 177-188, 2005-04-01

NRC Publications Record / Notice d'Archives des publications de CNRC:

<https://nrc-publications.canada.ca/eng/view/object/?id=a2fbe1ef-6dcd-4a42-a74d-c10d37a65bb0>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=a2fbe1ef-6dcd-4a42-a74d-c10d37a65bb0>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the
first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la
première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez
pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research
Council Canada

Conseil national de
recherches Canada

Canada

National Research
Council Canada

Conseil national
de recherches Canada

NRC - CNRC

Interpreting drinking water quality in the distribution system using Dempster-Shafer theory of evidence

Sadiq, R.; Rodriguez, M.J.

NRCC-47672

A version of this document is published in / Une version de ce document se trouve dans:

Chemosphere, v. 59, no. 2, April 2005, pp. 177-188

Doi: [10.1016/j.chemosphere.2004.11.087](http://dx.doi.org/10.1016/j.chemosphere.2004.11.087)

<http://irc.nrc-cnrc.gc.ca/ircpubs>

Interpreting drinking water quality in the distribution system using Dempster-Shafer theory of evidence

^{1*}Rehan Sadiq and ²Manuel J. Rodriguez

¹Institute for Research in Construction, National Research Council, Ottawa, ON, Canada, K1A 0R6

²Département d'Aménagement, Université Laval, Québec City, QC, Canada, G1K7P4

Abstract

Interpreting water quality data routinely generated for control and monitoring purposes in water distribution systems is a complicated task for utility managers. In fact, data for diverse water quality indicators (physico-chemical and microbiological) are generated at different times and at different locations in the distribution system. To simplify and improve the understanding and the interpretation of water quality, methodologies for aggregation and fusion of data must be developed. In this paper, the Dempster-Shafer theory also called theory of evidence is introduced as a potential methodology for interpreting water quality data. The conceptual basis of this methodology and the process for its implementation are presented by two applications. The first application deals with the interpretation of spatial water quality data fusion, while the second application deals with the development of water quality index based on key monitored indicators. Based on the obtained results, the authors discuss the potential contribution of theory of evidence as a decision-making tool for water quality management.

Keywords: water quality, data fusion, theory of evidence, aggregation operators, water distribution system.

*Corresponding author

Dr. Rehan Sadiq
Research Officer
Urban Infrastructure Rehabilitation Program, Institute for Research in Construction (IRC)
National Research Council (NRC), 1200 Montreal Road, M-20
Ottawa, Ontario, Canada K1A 0R6
Email: Rehan.sadiq@nrc-cnrc.gc.ca
Phone: 1-613-993-6282

INTRODUCTION

Monitoring and inspection of a system or a process may use more than one type of measurements and/or observations to describe the overall *Condition State*. The credibility of measurements to assess overall *Condition State* is important to be quantified for reliable decision-making. The data fusion is useful for an objective aggregation that can be reproducible and interpretable. Many infrastructure engineering problems, e.g., condition assessment of assets, production process quality control, and water quality monitoring require more than one performance indicator to define the *Condition State*. In addition, the aggregation of spatial or temporal observations of one (or more) performance indicator(s) is generally performed for reliable predictions.

The data fusion refers to the scientific aggregation of the observations and measurements. In some cases, different data sets (e.g., measured by different types of sensors and probes, various water quality indicators) give information on various aspects of the system or a process by complementing each other. Therefore, the motivation is to collect more information for accurate prediction of *Condition State*. It is also possible that the information collected by various data sets can also be redundant if it deals with the same aspect of the problem, but it improves the reliability as one measurement/observation is confirmed by the other. Complementing information and redundancy of data sets are the basis of data fusion applications in condition assessment of assets and water quality monitoring.

Regular monitoring of raw water quality, treatment processes and water quality in the distribution systems are integral parts of *total drinking water quality management* for the implementation of a *multi-barrier approach* for maintaining high-quality tap water for consumers. Water distribution systems are subjected to adverse reactions and events that can change the high-quality water to unpalatable and unsafe for human consumption by the time it arrives at the tap of the consumer (LeChevallier *et al.*, 1996). As water quality can change significantly in the distribution system, regular monitoring is even more essential to ensure that high-quality drinking water reaches the consumer.

To monitor the quality of water in the distribution system, physical, chemical, and biological indicators are recorded from routine grab sampling, followed by an analysis in the

laboratory or using portable kits in the field (APHA, AWWA, WPCF, 1995). Sensor technology exists that enables capturing some indicators through online monitoring rather than grab samples. This technology is continually evolving to encompass more types of water quality indicators. Some common water quality indicators used for water distribution are turbidity, residual disinfectant, pH, nitrates, phosphates, organic compounds, total/fecal coliforms, and heterotrophic bacteria (HPC) (Coulibaly and Rodriguez, 2003; Hunsinger and Zioglio, 2002; Clark, 1994).

Water uses generate a large amount of water quality data by routine sampling to control and maintain the acceptable *Condition State* of water quality in the system. Information is gathered on diverse water quality indicators using different techniques (manual sampling or auto-samplers and subsequent laboratory analysis, or online monitoring with automatic analyzer equipment). To better understand and interpret the water quality data, the use of novel techniques that favour the fusion and the aggregation of data is required to be explored.

In this paper, the application of Dempster–Shafer (D–S) theory or theory of evidence for interpretation of water quality in the distribution system is demonstrated with the help of two examples. The first example discusses the application of theory of evidence for water quality data fusion for the case of water samples collected at different locations in the distribution system at a given time (interpreting spatial information), which is equally valid for fusion of temporal data or combining both. The second example briefly discusses the application of D–S theory for developing water quality index (WQI) that helps in aggregating and interpreting water quality linguistically, but in a rational manner.

DEMPSTER–SHAFFER THEORY FOR INTERPRETING WATER QUALITY MONITORING DATA

There are numerous techniques available for conducting data and knowledge and information fusion, and most common among them are Bayesian inference, Dempster–Shafer rule of combination, fuzzy rule-based inference, and neural networks (Roemer *et al.*, 2001). The idea of evidence integration and accumulation of beliefs are commonly used in Bayesian inference, which implies that $p(A) + p(\neg A) = 1$, i.e., the belief in a hypothesis A can be used to derive the belief in its complement (Alim, 1988). But “NOT A ” is the missing evidence (lack of

knowledge) that is dealt as equal noninformative priors (*Principle of Insufficient Reason*) in Bayesian inference instead of ignorance. Alim (1988) argued that “No evidence” is different from having the same degree of confidence in all hypotheses, which is the basic motivation behind D–S theory.

Dempster–Shafer theory is a theory of evidence, which is based on classic work by Dempster (1968) and Shafer (1976). The D–S theory can be interpreted as a generalization of probability theory where probabilities are assigned to *subsets* as opposed to mutually exclusive *singletons*. The probability theory can associate evidence to only one possible event, whereas D–S theory determines the evidence to sets of events, i.e., if the evidence is sufficient enough to permit the assignment of probabilities to single event (singleton), the D–S theory inference reduces to the probabilistic formulation (Sentz and Ferson, 2002).

The D–S theory applications in civil and environmental engineering vary from slope stability (Binaghi and Luzi, 1998), environmental decision-making (Attoh-Okine and Gibbons, 2001; Chang and Wright, 1996), seismic analysis (Alim, 1988), failure detection (Tanaka and Klir, 1999), biological surveillance of river water quality (Boyd *et al.*, 1993), and remote sensing (Wang and Civco, 1994) to climate change (Luo and Caselton, 1997). Many more applications of D–S theory can be seen in detailed bibliography reported by Sentz and Ferson (2002). However, the potential for application of D–S theory in the drinking water industry, in particular for fusion and aggregation of water quality monitoring data in the distribution system, has not been investigated until now.

In the following section, the concepts of D–S theory application will be introduced by means of an example of data fusion of monitoring information on water quality in the distribution system.

Basic Concepts of Dempster–Shafer Theory and Application

The *frame of discernment* Θ (also called *universe of discourse*) is defined as a set of mutually exclusive alternatives, which has 2^{Θ} subsets in the domain. For example, if the *frame of discernment* Θ is a set $\{L, M, H\}$ it may have $8 (= 2^3)$ subsets. Three important concepts, namely, *basic probability assignment* (m or bpa), *belief* (bel), and *plausibility* (pl) functions are used in D–S theory. Alim (1988) summarized some basic features of the D–S theory as follows:

- ♦ Evidence in the form of belief (or disbelief) is attributed to subsets in Θ ;
- ♦ As evidence accumulates, the hypothesis set tends to narrow down toward precise estimation of probability; and
- ♦ Ignorance does not assume equal priors or uniformly distributed, rather it is assigned to *frame of discernment* Θ . For example, if some evidence “ a ” is attributed to subset “L” in Θ , the ignorance “ $1-a$ ” will not be equally distributed to “M” and “H”, rather it is assigned to $\Theta = \{L, M, H\}$.

Example 1: In this example, it is assumed that water quality in the distribution is reported qualitatively using three risk levels — *low* (L), *medium* (M) and *high* (H) from consumption viewpoint based on compliance of drinking water regulations. The *frame of discernment*, $\Theta = \{L, M, H\}$ contains 8 subsets ϕ (a null set) $\{L\}$, $\{M\}$, $\{H\}$, $\{L, M\}$, $\{M, H\}$, $\{L, H\}$, and $\{L, M, H\}$. Therefore, depending on the evidence, water could be rated as *low*, *medium*, *high*, *low* or *medium*, *low* or *high*, *medium* or *high*, and *low* or *medium* or *high* (in case of complete ignorance).

Basic Probability Assignment

The *basic probability assignment* (*bpa* or m) is different from classical definition of *probability* and is defined by mapping over the interval $[0, 1]$, where the null set $m(\phi)$ is “0” and the sum of the *basic probability assignments* $m(A)$ in a given set A is “1”. The $m(A)$ expresses the proportion of all relevant and available evidence that supports the claim that a particular element of Θ belongs to the set A but to no particular subset of A (Klir, 1995). For a given *basic probability assignment* m , every set for which $m(A) \neq 0$ is called *focal element*. Formally, this description of m can be represented with the following equation:

$$\left. \begin{array}{l} m(A) \rightarrow [0, 1] \\ m(\phi) = 0 \\ \sum_{A \subseteq \Theta} m(A) = 1 \end{array} \right\} \quad (1)$$

Example 1 (Contd.): If the water utility manager reports with 60% confidence that water is of *low* risk quality and with 30% confidence that it is *low* or *medium* risk, the ignorance is therefore 10%. The *focal elements* of hypothesis A can be written as

$$m(A)_L = 0.6 \quad m(A)_{L,M} = 0.3 \text{ and therefore } m(A)_\Theta = 0.1 \text{ because } m(A)_L + m(A)_{L,M} + m(A)_\Theta = 1$$

The *basic probability assignments* for remaining subsets will be zero.

Belief Function

The lower and upper bounds of an interval can be determined from the *basic probability assignment*, which contains the probability set bounded by two nonadditive measures *belief* and *plausibility*. The lower bound belief (bel) for a set A is defined as the sum of all the *basic probability assignments* of the proper subsets (B) of the set of interest A , i.e., $B \subseteq A$. The general relation between *bpa* and belief can be written as

$$bel(A) = \sum_{B \subseteq A} m(B) \quad (2)$$

The belief functions also follow these relationships

$$\left. \begin{aligned} bel(\phi) &= 0 \\ bel(\Theta) &= 1 \end{aligned} \right] \quad (3)$$

Example 1 (Contd.): The belief functions can be derived as

$$bel(A)_L = m(A)_L = 0.6; \quad bel(A)_M = m(A)_M = 0.0; \quad bel(A)_H = m(A)_H = 0.0$$

$$bel(A)_{L,M} = m(A)_L + m(A)_M + m(A)_{L,M} = 0.6 + 0.0 + 0.3 = 0.9$$

$$bel(A)_{L,H} = m(A)_L + m(A)_H + m(A)_{L,H} = 0.6; \quad bel(A)_{M,H} = m(A)_M + m(A)_H + m(A)_{M,H} = 0.0$$

$$bel(A)_{L,M,H} = m(A)_L + \dots + m(A)_\Theta = 1.0$$

Plausibility Function

The upper bound, *plausibility*, is the summation of *basic probability assignment* of the sets B that intersect with the set of interest A , i.e., $B \cap A \neq \phi$, and therefore it can be written as

$$pl(A) = \sum_{B \cap A \neq \phi} m(B) \quad (4)$$

The *plausibility function* can be related to *belief function* through a function called *doubt*, which is defined as the compliment of belief

$$\left. \begin{aligned} pl(A) &= 1 - bel(\neg A) \\ pl(A) &= 1 - doubt(A) \end{aligned} \right] \quad (5)$$

In addition, the following relationships for belief and plausibility functions hold true in all circumstances

$$\left. \begin{aligned} pl(A) &\geq bl(A) \\ pl(\phi) &= 0 \\ pl(\Theta) &= 1 \\ pl(\neg A) &= 1 - bel(A) \end{aligned} \right] \quad (6)$$

Example 1 (Contd.): Continuing on the example, the *plausibility function* can be derived as follows

$$pl(A)_L = m(A)_L + m(A)_{L,M} + m(A)_{L,H} + m(A)_\Theta = 1.0$$

$$pl(A)_M = m(A)_M + m(A)_{L,M} + m(A)_{M,H} + m(A)_\Theta = 0.4$$

$$pl(A)_H = m(A)_H + m(A)_{L,H} + m(A)_{M,H} + m(A)_\Theta = 0.1$$

$$pl(A)_{L,M} = m(A)_L + m(A)_M + m(A)_{L,M} + m(A)_{L,H} + m(A)_{M,H} + m(A)_\Theta = 1.0$$

$$pl(A)_{L,H} = m(A)_L + m(A)_H + m(A)_{L,M} + m(A)_{L,H} + m(A)_{M,H} + m(A)_\Theta = 1.0$$

$$pl(A)_{M,H} = m(A)_M + m(A)_H + m(A)_{L,M} + m(A)_{L,H} + m(A)_{M,H} + m(A)_\Theta = 0.4$$

$$p(A)_{L,M,H} = m(A)_M + m(A)_M + m(A)_H + m(A)_{L,M} + m(A)_{L,H} + m(A)_{M,H} + m(A)_\Theta = 1.0$$

Belief Interval

The belief interval (U) represents a range in which true probability may lie. It can be determined by subtracting *belief* from *plausibility*. The narrow uncertainty band represents more precise probabilities. The probability is uniquely determined if $bel(A) = pl(A)$ and for classical probability theory all probabilities are unique (Yager, 1987). If $U(A)$ has an interval $[0, 1]$, it means that no information is available, but if the interval is $[1, 1]$, then it means that A has been completely confirmed by $m(A)$.

Example 1 (Contd.): The uncertainty interval for the case at hand is

$$U(A)_L = [0.6, 1.0]; U(A)_M = [0.0, 0.4]; U(A)_H = [0.0, 0.1]$$

$$U(A)_{L,M} = [0.9, 1.0]; U(A)_{L,H} = [0.6, 1.0]; U(A)_{M,H} = [0.0, 0.4]; \text{ and}$$

$$U(A)_\Theta = [1.0, 1.0]$$

Dempster–Shafer Rule of Combination

The purpose of data fusion is to summarize and simplify information rationally. The D-S theory assumes sources of information are independent. The multiple sources of information in our context could be water quality samples collected at various points S_i s in the distribution system at a given time “ t_j ”. The D–S rule of combination can help in providing an overall picture of water quality at a given time “ t_j ” in the distribution system. Similarly, evidences about the water quality can be aggregated temporally (samples collected at various times t_j s) at a given sampling point S_i using D–S rule of combination.

Alim (1988) described that the “combined” belief represents not only the total belief in a set A and all of its subsets but also takes into account the contribution of different sources of evidence that focus on A . The D–S inference uses trade-off type combination operators and less information is assumed than that of Bayesian inference by compromising on precision, but Bayesian theory does not express any uncertainty associated with it and uses *Principle of Insufficient Reason* for inference (Sentz and Ferson, 2002).

The D–S rule of combination strictly emphasizes on the agreement between multiple sources and ignores all the conflicting evidence through *normalization*. A strict *conjunctive logic* through AND operator (estimated by a product of two probabilities) is employed in combination of evidence. The D–S combination rule determines the joint m_{1-2} from the aggregation of two *basic probability assignments* m_1 and m_2 by following equation:

$$m_{1-2}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K} \quad \text{when } A \neq \phi; \quad \text{and} \quad m_{1-2}(\phi) = 0 \quad (7)$$

where

$$K = \sum_{B \cap C = \phi} m_1(B) m_2(C) \quad (8)$$

where K is the *degree of conflict* in two sources of evidences. The denominator $(1-K)$ in Equation (7) is a normalization factor, which helps aggregation by completely ignoring the conflicting evidence. The above equations can also written as

$$m_{1-2}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \phi} m_1(B) m_2(C)} \quad (9)$$

Example 1 (Contd.): Water quality is monitored at two locations S_i ($i = 1, 2$) in the distribution system at a given time t_j . The utility manager is interested in overall water quality in the distribution system at t_j based on these two observations S_1 and S_2

$$\begin{aligned} m_1(B)_L &= 0.6 & m_2(C)_M &= 0.4 \\ m_1(B)_{L,M} &= 0.3 \text{ and} & m_2(C)_{M,H} &= 0.2 \\ m_1(B)_\Theta &= 0.1 & m_2(C)_\Theta &= 0.4 \end{aligned}$$

By applying D–S rule of combination on sources of information B and C , the following data is generated:

	$m_1(B)_L$	$m_1(B)_M$	$m_1(B)_H$	$m_1(B)_{L,M}$	$m_1(B)_{L,H}$	$m_1(B)_{M,H}$	$m_1(B)_\Theta$
	0.6	0.0	0.0	0.3	0.0	0.0	0.1
$m_2(C)_L = 0.0$	{L} 0.0	{ ϕ } 0.0	{ ϕ } 0.0	{L} 0.0	{L} 0.0	{ ϕ } 0.0	{L} 0.0
$m_2(C)_M = 0.4$	{ ϕ } 0.24	{M} 0.0	{ ϕ } 0.0	{M} 0.12	{ ϕ } 0.0	{M} 0.0	{M} 0.04
$m_2(C)_H = 0.0$	{ ϕ } 0.0	{ ϕ } 0.0	{H} 0.0	{ ϕ } 0.0	{H} 0.0	{H} 0.0	{H} 0.0
$m_2(C)_{L,M} = 0.0$	{L} 0.0	{M} 0.0	{ ϕ } 0.0	{L,M} 0.0	{L} 0.0	{M} 0.0	{L,M} 0.0
$m_2(C)_{L,H} = 0.0$	{L} 0.0	{ ϕ } 0.0	{H} 0.0	{L} 0.0	{L,H} 0.0	{H} 0.0	{L,H} 0.0
$m_2(C)_{M,H} = 0.2$	{ ϕ } 0.12	{M} 0.0	{H} 0.0	{M} 0.06	{H} 0.0	{M,H} 0.0	{M,H} 0.02
$m_2(C)_\Theta = 0.4$	{L} 0.24	{M} 0.0	{H} 0.0	{L,M} 0.12	{L,H} 0.0	{M,H} 0.0	$\Theta = 0.04$

Degree of conflict = $K = 0.24 + 0.12 = 0.36$, therefore

Normalization factor = $1 - K = 0.64$

$$m_{1-2}(A)_L = 0.24/0.64 = 0.38 ; m_{1-2}(A)_M = (0.12 + 0.06 + 0.04)/0.64 = 0.34$$

$$m_{1-2}(A)_H = 0.0; m_{1-2}(A)_{L,M} = 0.12/0.64 = 0.19; m_{1-2}(A)_{L,H} = 0.0$$

$$m_{1-2}(A)_{M,H} = 0.02/0.64 = 0.03; m_{1-2}(A)_{\emptyset} = 0.04/0.64 = 0.06$$

Similarly, *belief* and *plausibility functions* and belief interval can be determined by using corresponding equation described earlier.

Subsets	$m_{1-2}(A)$	$bel_{1-2}(A)$	$pl_{1-2}(A)$	$U_{1-2}(A)$
$\emptyset$	0.0	0.0	1.0	[0.0, 1.0]
{L}	0.38	0.38	0.63	[0.38, 0.63]
{M}	0.34	0.34	0.62	[0.34, 0.62]
{H}	0.0	0.0	0.09	[0.0, 0.09]
{L, M}	0.19	0.91	1.0	[0.91, 1.0]
{L, H}	0.0	0.38	0.66	[0.38, 0.66]
{M, H}	0.03	0.34	0.62	[0.34, 0.62]
$\emptyset$	0.06	1.0	1.0	[1.0, 1.0]

From the above analysis it can be noticed that based on evidence from two samples, the water quality can be rated as *low* or *medium*.

Modified Combination Rules

Serious drawbacks have been identified in D–S rule of combination. Zadeh (1984) presented an intriguing example of a patient who is diagnosed by two physicians *A* and *B*. The physician *A* diagnosed that the patient has a disease *x* with the 99% probability (confidence) and has only 1% probability of disease *y*. The physician *B* diagnosed that the patient has a disease *z* with the 99% probability and has only 1% probability of disease *y*. The *frame of discernment* for the diseases is $\Theta = \{x, y, z\}$. Using D–S rule of combination, following results will be obtained:

$$\text{Degree of conflict} = K = 0.9999 \therefore \text{Normalization factor} = 1 - K = 0.0001$$

$$m_x(\text{disease}) = 0.0; \quad m_y(\text{disease}) = 1.0; \quad \text{and} \quad m_z(\text{disease}) = 0.0$$

These results are counterintuitive, as 99.99% evidence was neglected due to conflict. Sentz and Ferson (2002) have provided an excellent review of various methods and techniques to resolve this discrepancy. Most common methods are Yager's modified Dempster's rule (1987),

Inagaki's Unified Combination rule (1991), and Zhang's Center Combination rule (Zhang, 1994).

Aggregation Operators

The triangular norms (t -norms) are a class of operators introduced for the development of a probabilistic generalization of the theory of metric spaces (Ramik and Vlach, 2001). The t -norms are used extensively in *fuzzy set theory*. They provide a tool for defining various types of *intersection* of fuzzy sets and expressing conjunctive logic. The t -norms, satisfy the axioms of commutativity, associativity, monotonicity, and boundary condition (Ramik and Vlach, 2001). Triangular conorms (t -conorms) provide a tool for defining various types of *union* of fuzzy sets and expressing conjunctive logic. These operators also satisfy all the axioms of commutativity, associativity, monotonicity, and boundary condition (Ramik and Vlach, 2001). The t -norms and t -conorms provide a range of operations for the aggregation of fuzzy sets (and probability theory).

Figure 1. Aggregation operators (after Larsen, 2002)

Aggregation or fusion is done through satisfying *several* or *few* criteria (performance indicators). When the requirement is such that *all* (or *several*) criteria have to be met, *t*-norms (*and*-type operators) are typically used; but when the requirement is such that only *few* criteria have to be met (out of many), *t*-conorms (*or*-type operators) are typically used. Consequently, on the scale of strictness of criteria, the *t*-norms represent the more *strict* criteria because being *intersection-based* they require *conjunction* (*and*-type operator) of aggregation, while the *t*-conorms represent more *relaxed* criteria, as being *union-based* they require *disjunction* (*or*-type operator) of aggregation (Sentz and Ferson, 2002). Figure 1 illustrates the entire range of aggregation operators from *very strict* to *very relaxed*. Note that *t*-norms and *t*-conorms are only two classes out of an entire range of aggregation operations. Average-type (e.g., arithmetic mean, ordered weighted average (OWA) operators) or compromising/compensatory operators lie in between two extremes.

Disjunctive Operator for Dempster–Shafer Rule

Traditional D-S rule of combination does not all allow to fuse the information from completely conflicting sources because the normalization factor $(1-K)$ becomes zero in Equation 7. Yager (2004) addressed this issue and proposed the use of disjunctive operators. Equation 9 can be modified as

$$m_{1-2}(A) = \frac{\sum_{B \cap C = A} \max[m_1(B), m_2(C)]}{\sum_{B \cap C \neq \phi} \max[m_1(B), m_2(C)]} \quad (10)$$

Other disjunctive operators (see Figure 1) than “*max*” can also be used in Equation 10. In the physician-patient example discussed by Zadeh (1984), the new diagnosis will be $m_x(\text{disease}) = 0.497$, $m_y(\text{disease}) = 0.005$, and $m_z(\text{disease}) = 0.497$.

Example 1 (Contd.): The disjunctive (*maximum*) operator (Equation 10) is used in modified combination rule. After estimating the *basic probability assignments*, the *belief* and *plausibility functions* are determined as described before in Equations 2 and 4, respectively.

Subsets	$m_{1-2}(A)$	$bel_{1-2}(A)$	$pl_{1-2}(A)$	$U_{1-2}(A)$
ϕ	0.0	0.0	1.0	[0.0, 1.0]
{L}	0.35	0.35	0.53	[0.35, 0.53]
{M}	0.28	0.28	0.50	[0.28, 0.50]
{H}	0.10	0.10	0.28	[0.10, 0.28]
{L, M}	0.09	0.72	0.90	[0.72, 0.90]
{L, H}	0.05	0.50	0.72	[0.50, 0.72]
{M, H}	0.09	0.47	0.65	[0.47, 0.65]
Θ	0.04	1.00	1.00	[1.0, 1.0]

From the above analysis it can be noticed that based on evidence from two samples, the water quality can be rated as *low* or *medium*.

Combining Sources of Varying Credibility

The approaches described before implicitly assume that all sources of information are equally credible. Sampling locations for monitoring water quality may be representative of a part of water distribution system, e.g., if one sample is collected from main distribution line and the other is collected from a minor line, the influence zones of both samples are different. Similarly, if the samples are collected at the same point when two different flow conditions prevail, the evidence of water quality also needs to be adjusted based on flow conditions. Similarly, if water utility staff with different levels of expertise collects water samples, the observations need to be adjusted based on their credibility.

Yager (2004) discussed the credibility issue in detail and suggested a *credibility transformation function*. This approach discounts the evidence with a credibility factor (α) and distributes remaining evidence ($1-\alpha$) equally among elements (n) of *frame of discernment*.

$$m(A)_a = m(A) \bullet \alpha + \frac{1-\alpha}{n} \quad (11)$$

where

α is the *Credibility factor* and n is the *focal elements* in *frame of discernment*.

Example 1 (Contd.): Assume that credibility adjustment factors assigned to two samples collected at different locations in the distribution are $\alpha_B = 1.0$ and $\alpha_C = 0.5$. These factors represent the confidence of the collected information. The modified evidences will be

$$\begin{aligned}
 m_1(B)_L &= 0.6 & m_2(C)_M &= 0.36 \\
 m_1(B)_{L,M} &= 0.3 \text{ and} & m_2(C)_{M,H} &= 0.1 \\
 m_1(B)_\Theta &= 0.1 & m_2(C)_L &= 0.17 \\
 & & m_2(C)_H &= 0.17 \text{ and} \\
 & & m_2(C)_\Theta &= 0.2
 \end{aligned}$$

As credibility of first evidence is 100%, therefore no adjustment is required for $m_1(B)$, but evidence $m_2(C)$ is only 50% credible, the evidence is adjusted as below:

$$m_2(C)_M = 0.4 \bullet 0.5 + (1-0.5)/3 = 0.36$$

$$m_2(C)_L = 0.0 \bullet 0.5 + (1-0.5)/3 = 0.17$$

$$m_2(C)_H = 0.0 \bullet 0.5 + (1-0.5)/3 = 0.17$$

$$m_2(C)_{M,H} = 0.2 \bullet 0.5 = 0.1$$

$$m_2(C)_\Theta = 0.4 \bullet 0.5 = 0.2$$

It is important to note that as $\alpha \rightarrow 0$ (i.e., confidence for given evidence), the inference tends to become Bayesian, i.e., *Principle of Insufficient Reason* is applied. A limiting case for evidence is $m_2(C)_\Theta = 1.0$, i.e., complete ignorance in D–S framework. If the credibility factor $\alpha = 0$, the adjusted evidence will become

$$m_2(C)_L = 0 \bullet 0 + (1-0)/3 = 0.33,$$

similarly

$$m_2(C)_M = 0.33 \text{ and } m_2(C)_H = 0.33$$

The adjusted evidences can be combined using modified D–S rule of combination as described earlier.

Subsets	$m_{1-2}(A)_a$	$bel_{1-2}(A)_a$	$pl_{1-2}(A)_a$	$U_{1-2}(A)_a$
ϕ	0.0	0.0	1.0	[0.0, 1.0]
{L}	0.42	0.42	0.61	[0.42, 0.61]
{M}	0.26	0.26	0.42	[0.26, 0.42]
{H}	0.13	0.13	0.24	[0.13, 0.24]
{L, M}	0.08	0.76	0.87	[0.76, 0.87]
{L, H}	0.03	0.58	0.74	[0.58, 0.74]
{M, H}	0.05	0.44	0.58	[0.44, 0.58]
Θ	0.03	1.00	1.00	[1.0, 1.0]

As noticed from the above analysis that *belief* of water quality being *low* is the highest among other *Condition States* (*medium* and *high*), therefore based on the available information the utility manager (or decision-maker) may conclude that water quality in the distribution is acceptable. But if the utility manager wants to be more confident about his judgement, he (she) will conclude that water quality is *low* or *medium* because the *belief* of subset {L, M} is 76%. In the above example, we allowed a subset {L, H}, that does not contain two contiguous states. But in reality, generally only two contiguous states are possible, i.e., in our case {L, M} or {M, H}.

If the decision-maker wants to increase the confidence for his (her) judgement concerning water quality *Condition State* he (she) will collect more sample (evidence). In this way he (she) can narrow down the uncertainty and increase the confidence in his (her) judgment.

DEMPSTER–SHAFFER THEORY APPLICATION FOR DEVELOPING WATER QUALITY INDEX

Water quality is generally defined by a collection of upper and lower limits on selected possible contaminants (Maier, 1999). Water quality indicators can be classified into three broad categories: physical, chemical, and microbiological contaminants. Within each class, a number of quality indicators are considered. The acceptability of water quality for its intended use depends on the magnitude of these indicators (Swamee and Tyagi, 2000) and is often governed by regulations (US EPA, 2001).

The physical, chemical, and microbiological processes occurring in drinking water distribution pipes are numerous and complex. A wealth of literature is available on water quality represented by an aggregate 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 overall water quality index (WQI). Sinha *et al.* (1994) combined pH, chloride concentration, turbidity, residual chlorine, conductivity, and MPN (most probable number – a bacterial counting technique) into a single water quality index through a weighting technique to represent an overall water quality at various nodes in the distribution system. Sadiq *et al.* (2004) have suggested a fuzzy-based framework for aggregative risk analysis of water quality failure in the distribution system. Recently, Sadiq and Rodriguez (2004a) proposed a risk-based fuzzy synthetic evaluation technique for aggregating effects of disinfection byproducts found in drinking water.

The WQI is a systematic way of interpreting measurements and (or) observations of water quality, which helps managers to describe a *Condition State* or to share and communicate with the public in a consistent manner. The WQI provides a general means of comparing and ranking water quality. Traditionally, WQI encompasses factors like *number* of indicators not meeting the regulation, *frequency* of a particular indicator by which it is not meeting the requirement in a given sampling protocol, and *amount* by which indicators are violating the regulatory requirements. These three factors are combined to form the WQI, which can be interpreted by predefined qualitative ranking system.

For overall water quality based on various indicators, credibility adjustment is required for each indicator for its contribution. For example, if the water quality is defined by turbidity, total coliforms, residual chlorine, and aesthetic indicators (taste, odour, colour), the violation of turbidity from its threshold value has lesser consequences and impacts with respect to microbial violations. Different credibility weights need to be defined for each indicator representing its body of evidence in defining overall water quality. Another useful application of D–S rule of combination is to develop a WQI that integrates various water quality indicators (of non-commensurate units) as a single entity. The example 2 will illustrate such application for the case of aggregation of three important physico-chemical and microbiological indicators of water quality in the distribution system.

Example 2 The application of disinfection agents in drinking water reduces the microbial risk but poses chemical risk in the form of their byproducts. A *risk-risk* tradeoff is required to optimize the dose and type of disinfection practices. Three water quality indicators — trihalomethanes (THMs), residual chlorine (RC), and heterotrophic plate counts (HPCs) (indicator for microbial presence) — are identified for evaluating the overall water quality in the distribution system. The water quality is defined by five risk classes — *very low* (VL), *low* (L), *medium* (M), *high* (H), and *very high* (VH). Therefore, *the frame of discernment* is $\Theta = \{VL, L, M, H, VH\}$. These water quality indicators are defined by these five classes of risk (Figure 2). The thresholds shown in Fig. 2 for Example 2 were established based on water quality standards and based on authors' experience with the water quality in Canadian distribution systems.

- ♦ The *bpa* for a given water quality indicator is determined by mapping on corresponding triangular functions as shown in Figure 2. The qualitative scale is defined in such a way that *bpa* for only two risk classes are obtained. Therefore, for any value of water quality indicator, maximum two *focal elements* are possible. In this setting, subsets with two or more elements are not allowed. For a water quality indicator, *bpa* is represented by a 5-tuple set $\{VL, L, M, H, VH\}$.

- ♦ For a given water sample, the *bpa* for three indicators are represented as follow:

$m(RC)_{VL}$	$m(HPC)_{VL}$	$m(THM)_{VL}$
$m(RC)_L$	$m(HPC)_L$	$m(THM)_L$
$m(RC)_M$	$m(HPC)_M$	$m(THM)_M$
$m(RC)_H$	$m(HPC)_H$	$m(THM)_H$
$m(RC)_{VH}$	$m(HPC)_{VH}$	$m(THM)_{VH}$

- ♦ The credibility factors α are assigned to these indicators based on expert judgement

$$\alpha_{RC} = 0.9 \qquad \alpha_{HPC} = 0.5 \qquad \alpha_{THM} = 0.8$$

- ♦ The *bpa* for each water quality indicator is adjusted by credibility factors α using Equation 11. The adjusted *bpa* for water quality indicators are aggregated using modified disjunctive operator D–S rule of combination.
- ♦ The *belief* and *plausibility functions* and belief interval are determined.

Figure 2. *Basic probability assignments* for water quality indicators

Example 2 (Contd.): A water sample was collected from distribution system and tested for residual chlorine, THMs, and HPCs.

RC = 0.09 mg/L; HPC = 62 /100ml; and THM = 118 ppb

The *bpa* for each water quality indicator is derived from Figure 2

$m(RC)_{VL} = 0.0$	$m(HPC)_{VL} = 0.0$	$m(THM)_{VL} = 0.0$
$m(RC)_L = 0.0$	$m(HPC)_L = 0.48$	$m(THM)_L = 0.0$
$m(RC)_M = 0.0$	$m(HPC)_M = 0.52$	$m(THM)_M = 0.0$
$m(RC)_H = 0.87$	$m(HPC)_H = 0.0$	$m(THM)_H = 0.0$
$m(RC)_{VH} = 0.13$	$m(HPC)_{VH} = 0.0$	$m(THM)_{VH} = 1.0$

The *bpa* is adjusted with respect to their credibility factors. The evidence is modified to

$m(RC)_{VL} = 0.02$	$m(HPC)_{VL} = 0.10$	$m(THM)_{VL} = 0.04$
$m(RC)_L = 0.02$	$m(HPC)_L = 0.34$	$m(THM)_L = 0.04$
$m(RC)_M = 0.02$	$m(HPC)_M = 0.36$	$m(THM)_M = 0.04$
$m(RC)_H = 0.80$	$m(HPC)_H = 0.10$	$m(THM)_H = 0.04$
$m(RC)_{VH} = 0.14$	$m(HPC)_{VH} = 0.10$	$m(THM)_{VH} = 0.84$

The adjusted *bpa* for water quality indicators can be aggregated using disjunctive operator D–S rule of combination.

<i>bpa</i>	<i>belief</i>	<i>Plausibility</i>
$m(WQ)_{VL} = 0.04$	$bl(WQ)_{VL} = 0.04$	$pl(WQ)_{VL} = 0.04$
$m(WQ)_L = 0.14$	$bl(WQ)_L = 0.14$	$pl(WQ)_L = 0.14$
$m(WQ)_M = 0.15$	$bl(WQ)_M = 0.15$	$pl(WQ)_M = 0.15$
$m(WQ)_H = 0.33$	$bl(WQ)_H = 0.33$	$pl(WQ)_H = 0.33$
$m(WQ)_{VH} = 0.34$	$bl(WQ)_{VH} = 0.34$	$pl(WQ)_{VH} = 0.34$

The probability mass function of risk can be plotted using *belief function*. The *universe of discourse* of risk scale is *soft* in nature.

Figure 3. Probability mass function of risk

Utility values can be assigned to soft items to determine the water quality index as a crisp output. Yang and Xu (2002) discussed a probabilistic method to determine the utility values for soft items in a heuristic way. These values can also be determined through linear optimization based on expert judgement. Here, an arbitrary linear function is proposed to estimate the crisp WQI (a surrogate for representing risk) and all five classes of risk are assigned utility values as follow:

$$WQI = \frac{u}{2^0} [bl(WQ)_{VH}] + \frac{u}{2^1} [bl(WQ)_H] + \frac{u}{2^2} [bl(WQ)_M] + \frac{u}{2^3} [bl(WQ)_L] + \frac{u}{2^4} [bl(WQ)_{VL}] \quad (12)$$

where utility coefficient u is assumed ≈ 1.3 .

New regulations for the allowable concentrations of disinfection byproducts are being developed in the U.S. and elsewhere for drinking water supplies. Disinfection reduces the risk from microbial infections, but may pose cancer and other risks from the DBPs (THMs are the most commonly identified DBPs). Many other DBPs, however, remain to be identified and the public health significance of these is unknown. Society is facing a difficult tradeoff between established (known) microbial risks due to pathogens and more uncertain (unknown) risks from DBPs. In the case of evaluating the *risk–risk* trade-offs in drinking water, the competing risks must be assessed within a common framework.

Example 2 (Contd.): The *risk–risk* trade-off for HPCs (a microbial indicator) and THMs (representative DBP) is established at different levels of residual chlorine concentration in Figures 4a–4d. The WQI is used as a surrogate for risk, estimated using Equation 12.

(a) residual chlorine = 0 mg/L

(b) residual chlorine = 0.2 mg/L

(c) residual chlorine = 0.5 mg/L

(d) residual chlorine = 4 mg/L

Figure 4. Water quality index (WQI) representing risk profiles at various residual chlorine levels

The analysis is performed for 0, 0.2, 0.5, and 4 mg/L residual chlorine concentrations. When levels of residual chlorine are not detectable, the WQI varied approximately from 0.6 to 1.0. Higher risks were observed for even very low HPC and THM concentrations (Figure 4a), because the minimal levels of residual chlorine are necessary to provide safeguard against microbial contamination. But when the residual concentration is increased to 0.2, 0.5, and 4.0 mg/L, the WQI varied from 0.2 to approximately 0.8 (Figures 4b–4d), which is comparatively lower than the first case.

The three-dimensional characteristic risk curves (e.g. Figure 4) can be established for various water quality indicators, which are able to predict levels of any particular indicator (e.g., HPCs) that are required to achieve *acceptable risk* under given conditions. For example, for an *acceptable risk* (WQI) of 0.25, the residual chlorine in the distribution system is reported to be in the range of 0.2–0.5 mg/L and THM potential is estimated (using regression or kinetic models, see Sadiq and Roriguez, 2004b) to be in the range of 25–50 ppb, and the HPC levels should not exceed 200/100ml. This concept can be extended to more water quality indicators.

SUMMARY AND CONCLUSIONS

In this paper, the evidence theory was introduced as an innovative methodology that can be used for simplifying and improving the understanding of data generated through routine water quality monitoring in distribution systems. Two examples were presented that support the potential application of theory of evidence for data fusion, namely, interpretation of overall water quality in the distribution system based on spatial data collected at different sampling locations and development of WQI.

For the first example, additional aspects should be investigated in the future, such as the impact of the uncertainty on the confidence of the decision-maker's judgement (according to the amount of information available, in this case the number and the frequency of spatial distribution of samples collected). For the second example, additional information should be considered in the future to develop more robust indices, i.e., additional water quality indicators (e.g., pathogenic indicators such as coliforms and other disinfection byproducts like haloacetic acids), operational parameters (e.g., pressures, flow rates, reservoir level control, etc.), and data on the distribution system infrastructure (e.g., pipe breakage rate and replacement, pipe flushing etc.).

Theory of evidence can efficiently deal with the difficulties related to host of indicators describing water quality, with spatial and temporal dimensions of distribution system, where redundancy of information is routinely observed as well as the credibility of available data is varied. Future research must focus on the implementation of decision-making tools using theory of evidence that can be adapted to specific water utility conditions and manager's needs. The potential combination of theory of evidence with modeling techniques, such as linear and nonlinear time-series analysis, neural networks, and genetic algorithms, to predict the *condition state* of water quality must also be evaluated through future research efforts to implement more powerful decision-making tools.

REFERENCES

- Alim, S. 1988. Application of Dempster-Shafer theory for interpretation of seismic parameters, *ASCE Journal of Structural Engineering*, 114(9): 2070-2084.
- APHA, AWWA, WPCF. 1995. Standard methods for the examination of water and wastewater. 19th edition, Washington, DC.
- Attoh-Okine, N.O., and Gibbons, J. 2001. Use of belief function in brownfield infrastructure redevelopment decision making, *ASCE Journal of Urban Planning and Development*, 127(3): 126-143.
- Binaghi, E., Luzi, L., Madella, P., Pergalani, F., and Rampini, A. 1998. Slope instability zonation: a comparison between certainty factor and fuzzy Dempster-Shafer approaches, *Natural Hazards*, 17: 77-97.
- Boyd, M., Walley, W.J., and Hawkes, H.A. 1993. Dempster-Shafer reasoning for the biological surveillance of river water quality, *Water Pollution 93*, Milan, Italy.
- Chang, Y.C., and Wright, J.R. 1996. Evidential reasoning for assessing environmental impact, *Civil Engineering Systems*, 14(1): 55-77.
- Clark, R.M. 1994. Modelling water quality changes and contaminant propagation in drinking water distribution systems: a US perspective, *Journal Water SRT-Aqua*, 43(3): 133-143.
- Coulibaly, H., Rodriguez M.J. 2003. Spatial and temporal variation of drinking water quality in ten Quebec small utilities, *Journal of Environmental Engineering & Science*, 2(1): 47-61.
- Dempster, A. 1968. A generalisation of Bayesian inference, *Journal of Royal Statistical Society, Series B* 30, 205-247.
- Hunsinger, R.B., and Zioglio, G. 2002. Rationale for online monitoring, In: *Online monitoring for drinking water utilities co-operative research report*, Ed. Hargesheimer, E., Conio, O., and Popovicova, J., American Water Works Association Research Foundation, CO.
- Inagaki, T. 1991. Interdependence between safety-control policy and multiple sensor scheme via Dempster-Shafer theory, *IEEE Transactions on Reliability*, 40(2): 182-188.
- Klir, J.G. 1995. Principles of uncertainty: what are they? why do we need them?, *Fuzzy Sets and Systems*, 74: 15-31.
- Larsen, H.L. 2002. *Fundamentals of fuzzy sets and fuzzy logic*, <http://www.cs.aue.auc.dk/~legind/FL%20E2002/FL-01/FL-01%20Introduction.pdf>.
- LeChevallier MW, Welch NJ, Smith DB. 1996. Full-scale studies of factors related to coliform regrowth in drinking water, *Applied and Environmental Microbiology*, 62(2): 201-2211.
- Luo, W.B., and Caselton, B. 1997. Using Dempster-Shafer theory to represent climate change uncertainties, *Journal of Environmental Management*, 49(1): 73-93.
- Maier, S.H. 1999. *Modeling water quality for water distribution systems*, Ph.D. thesis, Brunel University, Uxbridge.
- Ramik, J., and Vlach, M. 2001. *Generalized concavity in fuzzy optimization and decision analysis*, Kluwer Academic Publishers, Boston.

- Roemer, M.J., Kacprzyński, G.J., and Scholler, M.H. 2001. Improved Diagnostic and prognostic assessments using health management information fusion, *2001 IEEE*, 365-377.
- Sadiq, R., and Rodriguez, M.J. 2004a. Fuzzy synthetic evaluation of disinfection by-products – a risk-based indexing system, to appear *Journal of Environmental Management*, 73(1): 1-13.
- Sadiq, R., and Rodriguez, M.J. 2004b. Disinfection by-products (DBPs) in drinking water and the predictive models for their occurrence: a review, *The Science of the Total Environment*, 321(1-3): 21-46.
- Sadiq, R., Kleiner, Y., and Rajani, B.B. 2004. Aggregative risk analysis for water quality failure in distribution networks, *AQUA - Journal of Water Supply: Research & Technology*, 53(4): 241-261.
- Sentz, K. and Ferson, S. 2002. Combination of evidence in Dempster-Shafer theory, SAND 2002-0835.
- Shafer, G. 1976. *A mathematical theory of evidence*, Princeton University Press, Princeton, N.J.
- 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, *ASCE Journal of Environmental Engineering*, 126(5): 451-455.
- Tanaka, K. and Klir, G.J. 1999. Design condition for incorporating human judgement into monitoring systems, *Reliability Engineering and System Safety*, 65: 251-258.
- US EPA 2001. *National primary drinking water standards*, United States Environmental Protection Agency, EPA 816-F-01-007.
- Wang, Y., and Civco, D.L. 1994. Evidential reasoning-based classification of multi-source spatial data for improved land cover mapping, *Canadian Journal of Remote Sensing*, 20: 381-395.
- Yager, R.R. 1987. On the Dempster-Shafer framework and new combination rules, *Information Sciences*, 41: 93-137.
- Yager, R.R. 2004. On the determination of strength of belief for decision support under uncertainty – Part II: fusing strengths of belief, *Fuzzy Sets and Systems*, 142: 129-142.
- Yang, J-B., and Xu, D-L. 2002. On the evidential reasoning algorithm of multiple attribute decision analysis under uncertainty, *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 32(3): 289-304.
- Zadeh, L.A. 1984. Review of books: A mathematical theory of evidence, *The AI Magazine*, 5(3): 81-83.
- Zhang, L. 1994. Representation, independence, and combination of evidence in the Dempster-Shafer theory, *Advances in Dempster-Shafer theory of evidence*, Ed. Yager R.R. Kacprzyk, J., and Fedrizzi, M., NY, John Wiley and Sons, Inc., pp. 51-69.