Knowledge-Based Systems 85 (2015) 159–169

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Integrating failure prediction models for water mains: Bayesian belief network based data fusion

Golam Kabir *, Gizachew Demissie, Rehan Sadiq [1], Solomon Tesfamariam [2]

School of Engineering, University of British Columbia (UBC), Kelowna, BC V1V 1V7, Canada

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 12 March 2015
Accepted 3 May 2015
Available online 9 May 2015

Keywords:
Bayesian belief network (BBN)
Bayesian regression
Correlation
Data fusion
Failure rate
Corrosion
Soil resistivity
Soil corrosivity
Geographic information system (GIS)

ABSTRACT

Due to incomplete and partial information, data/information from multiple sources with different credibility or confidence, and the involvement of human (expert) judgment for the interpretation and integration of data/information, uncertainties become a major concern for the development of water main failure prediction model. To reduce these uncertainties, a new Bayesian belief network based data fusion model is developed for the failure prediction of water mains. To accredit the proposed framework, it is implemented to predict the failure of CI and DI pipes of the water distribution network of the City of Calgary. Analysis results indicate that ~6.16% and 8.20% of the total 18,762 CI and DI pipes are at *high* and *very high* failure rates, respectively. The proposed model can be integrated with the geographic information system of the utilities and capable of identifying the most 'vulnerable' and 'sensitive' pipes within the distribution network as well as estimate the total number of pipes that need maintenance/rehabilitation/replacement (M/R/R) actions.

1. Introduction

For last few decades, concerns have been repeatedly raised about deteriorating water/wastewater infrastructure that may lead to catastrophic failures. The American Society of Civil Engineers ASCE [2] Infrastructure Report Card has given a poor grade of "D" to the nation's public drinking and wastewater systems. A substantial amount of drinking-water, wastewater and stormwater infrastructure of Canada are in "fair" to "very poor" condition, according to the assessment reported by the Infrastr [8]. According to the watermainbreakclock.com, there have been more than 4 million breaks in the United States and Canada since January 2000, with an average of 850 water main breaks every day and that lead to annual repair cost of more than 3 billion dollars (2012 U.S. dollar).

Based on the survey of 188 utilities in USA and Canada, Folkman [16] reported that circumferential crack and corrosion induced failures are the most common failure mode of water mains. Metallic pipe corrosion is associated with low resistivity of soil, presence of anaerobic bacteria, chlorides, sulfate and sulfides, lower pH, differential aeration of soil around metallic pipes, difference in soil composition, and stray direct current from external forces [15,38,19,3]. The United States drinking water and wastewater systems are greatly affected by corrosion which costs over 50.7 billion dollars annually, or more than 1 trillion dollars over the next twenty years (watermainbreakclock.com). It is necessary to identify the potential corrosive environment and take appropriate action to avoid pipe failures, and to save significant future costs [38].

Uncertainties become an integral part of the water main failure prediction model due to incomplete and partial information, and the involvement of human (expert) judgment for the interpretation of data/information [20,21,36,37,42]. Moreover, due to unreliable recording of actual failure times or inaccurate measurements of the variables, data quality also become a major concern and source of uncertainties [21]. It is very difficult to fully understand the complex nature of soil environment and its contributions to metallic pipe corrosions. To account for this complexity and prevalent uncertainties, Demissie et al. [13] have developed a Bayesian belief network (BBN) model to predict soil corrosivity index (SCI) combining *in situ* collected data, experimental data, and expert opinion, and considering the inter-dependencies between soil parameters. These information can aid utility managers and other authorities as they need location-specific pipe failure information to develop

* Corresponding author. Tel.: +1 250 862 0733.
E-mail addresses: golam.kabir@ubc.ca (G. Kabir), gizachew.demissie@ubc.ca (G. Demissie), rehan.sadiq@ubc.ca (R. Sadiq), solomon.tesfamariam@ubc.ca (S. Tesfamariam).
[1] Tel.: +1 250 807 9013.
[2] Tel.: +1 250 807 8185.

http://dx.doi.org/10.1016/j.knosys.2015.05.002
0950-7051/

an effective pipe failure prediction model for the successful implementation of short- and log-term preventive management plans. For the failure prediction model, to deal with these uncertainties, Kabir et al. [21] utilized Bayesian regression model to consider the model parameters as random variables and incorporate external information, etc. (e.g. relevant historical information, elicited expert opinions) into the model. Kabir et al. [21] have also shown that Bayesian regression models provide improved performance in predicting future observations compared to the normal regression models.

Data/information from multiple sources (e.g. soil resistivity and SCI based failure models) with different credibility or confidence of the information can be integrated using a data fusion technique [18,4]. Data fusion can be defined as a combination of multiple sources to obtain improved information; in this context, improved information means less expensive, higher quality, or more relevant information [10,18]. Data fusion techniques can also be regarded as a mathematical techniques used to combine multiple values of a feature into single value [44]. The goal of using data fusion in this analysis is to obtain a lower prediction error probability and a higher reliability by using data from multiple distributed sources or models [24,4]. An appropriate data fusion methods are required to develop an effective water main failure prediction model. Objective of this research is thus to utilize data fusion technique for new and effective failure rate prediction of water mains. Then BBN model is used as a data fusion method to combine the failure rates from multiple water main failure prediction models. Such data fusion techniques have ability to address scarcity of data/information, and proactively predict the structural failure of water mains.

2. Methodology

The framework of the proposed study is shown in Fig. 1. The first step entails gathering pipe characteristics data, soil information and pipe breakage data from the water utility's Geographic Information System (GIS). In the second step, based on the soil information, SCI will be developed using the BBN model (developed by [13] and considering soil resistivity as a surrogate measure of SCI. In the third step, correlation analysis will be performed to measure the dependence between the observed breakage rate, soil resistivity and SCI. In the fourth step, Bayesian regression models will be developed using pipe characteristics data with soil resistivity (Model 1) and pipe characteristics data with SCI (Model 2). In the fifth step, the BBN-based data fusion model will be developed to combine the failure rates obtained from Model 1 and Model 2. Finally, the result from the data fusion model can be integrated with the GIS of the utilities for decision making. The following subsections briefly discuss Bayesian belief network, Bayesian linear regression, and data fusion methods.

2.1. Bayesian belief network

BBN is a graphical model that permits a probabilistic relationship or causal dependencies among a set of variables [32]. BBN is based on the Bayes' theorem that manage uncertainty by explicitly representing the conditional probability dependencies between variables [11,40]. In a BBN analysis, for n number of mutually exclusive parameters $M_{i\ (i=1,2,\ldots,n)}$, and a given observed data N, the updated probability is computed by:

$$p(M_i|N) = \frac{p(N|M_i) \times p(M_i)}{\sum_j p(N|M)p(M_j)} \tag{1}$$

where $p(M)$ denotes the prior occurrence probability of M, $p(N)$ denotes the marginal (total) occurrence probability of N and is effectively constant since the obtained data is in hand, $p(N|M)$ refers to the conditional occurrence probability of N given that M occurs too, and $p(M|N)$ represents the posterior occurrence probability of M given the condition that N occurs [12,28,32].

The conditional probabilities can be obtained through expert knowledge elicitation [28], or training from data [40]. BBN is flexible to capture both top-down inference (predictive analysis) and bottom-up inference (diagnostic analysis) [12]. BBN is used to update probabilities when new data/information is available [13,32]. For more detail information about BBN, interested readers are referred to Kabir et al. [20], Demissie et al. [13], Tesfamariam and Liu [41], Cockburn and Tesfamariam [12], Cinar and Kayakutlu [11], Tang and McCabe [40], Nadkarni and Shenoy [28], and Pearl [32].

2.2. Bayesian linear regression

In Bayesian regression, the model parameters are treated as random variables rather than fixed (unknown) constants [17,6]. For this external information can be incorporated into the model by constructing a probability distribution (prior) that describes the uncertainty in the model parameters [9,17]. In many real life situations, a prior distribution contains virtually no information (noninformative prior) but can be updated whenever new information/data is available which is the main advantages of Bayesian analysis [23,17]. In the linear multiple regression problem, the mean value of the response variable y_i, $E(y_i|\beta, X)$, can be expressed as:

$$E(y_i|\beta, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = x_i\beta, \quad i = 1, 2, \ldots, n \tag{2}$$

where $x_i = (x_{i1}, \ldots, x_{ik})$ are predictor values for the ith individual and $\beta = (\beta_1, \ldots, \beta_k)$ are unknown regression parameters or coefficients [21,1]. Let $\theta = (\beta_1, \ldots, \beta_k, \ \sigma^2)$ indicates the vector of unknown parameters. Equal variances are assumed in the normal linear regression, where $var(y_i|\theta, X) = \sigma^2$ [1,5]. For the Bayesian formulation of the normal regression model, it can be assumed that (β, σ^2) has the typical noninformative prior [21,23]:

$$g(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \tag{3}$$

For the Bayesian regression model, the posterior joint density of (β, σ^2) can be represented as [21,1]:

$$g(\beta, \sigma^2|y) = g(\beta|y, \sigma^2)g(\sigma^2|y) \tag{4}$$

For the prediction of a future observation, the posterior predictive density of $\tilde{y}$, $p(\tilde{y}|y)$, can be represented as [21,1]:

$$p(\tilde{y}|y) = \int p(\tilde{y}|\beta, \sigma^2)g(\beta, \sigma^2|y)d\beta d\sigma^2 \tag{5}$$

Detailed discussions are referred to Kabir et al. [21], Lu et al. [23], Albert [1], Bolstad [5], Casella and Berger [9], Gelman et al. [17], and Box and Tiao [6].

2.3. Data fusion

The data fusion techniques can be classified into three nonexclusive categories: (i) data association (identity level fusion), (ii) state estimation (feature level fusion), and (iii) decision level fusion [44,10]. The data association level fusion must determine the set of measurements that correspond to each target [18,24]. State estimation (feature level fusion) techniques aim to determine the state of the target under movement (typically the position) given the observation or measurements [46,24]. On the other hand, decision fusion techniques aim to make a high-level inference based on the knowledge of the perceived situation, events and activities that are produced from the detected targets [46,18]. These techniques often