Reliability Engineering and System Safety 130 (2014) 1–11

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Bayesian Belief Networks for predicting drinking water distribution system pipe breaks

Royce A. Francis [a,*], Seth D. Guikema [b], Lucas Henneman [b]

[a] Department of Engineering Management and Systems Engineering, 1776 G St., NW #159, The George Washington University, Washington, DC 20006, USA
[b] Department of Geography and Environmental Engineering, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form
25 April 2014
Accepted 27 April 2014
Available online 10 May 2014

Keywords:
Bayesian Belief Networks
Drinking water distribution systems
Expert systems
Graphical models

ABSTRACT

In this paper, we use Bayesian Belief Networks (BBNs) to construct a knowledge model for pipe breaks in a water zone. To the authors' knowledge, this is the first attempt to model drinking water distribution system pipe breaks using BBNs. Development of expert systems such as BBNs for analyzing drinking water distribution system data is not only important for pipe break prediction, but is also a first step in preventing water loss and water quality deterioration through the application of machine learning techniques to facilitate data-based distribution system monitoring and asset management. Due to the difficulties in collecting, preparing, and managing drinking water distribution system data, most pipe break models can be classified as "statistical–physical" or "hypothesis-generating." We develop the BBN with the hope of contributing to the "hypothesis-generating" class of models, while demonstrating the possibility that BBNs might also be used as "statistical–physical" models. Our model is learned from pipe breaks and covariate data from a mid-Atlantic United States (U.S.) drinking water distribution system network. BBN models are learned using a constraint-based method, a score-based method, and a hybrid method. Model evaluation is based on log-likelihood scoring. Sensitivity analysis using mutual information criterion is also reported. While our results indicate general agreement with prior results reported in pipe break modeling studies, they also suggest that it may be difficult to select among model alternatives. This model uncertainty may mean that more research is needed for understanding whether additional pipe break risk factors beyond age, break history, pipe material, and pipe diameter might be important for asset management planning.

1. Introduction

While U.S. drinking water distribution system integrity may be compromised through a number of mechanisms, pipe breaks and their associated repair and renewal activities are among the most important causes of drinking water distribution system contamination [1,2], and pipe breaks may be associated with increased health risk [3–8]. However, developing statistical models to predict pipe breaks is a difficult activity. For example, Yamijala et al. [9] compare the predictive accuracy of several statistical regression models that have previously been used in the literature for estimating the probability or number of pipe breaks and/or leaks on individual pipe segments. Their results indicate that more research is required to improve pipe break modeling predictive accuracy.

The present study extends this modelling work by studying the use of Bayesian Belief Networks (BBNs) to model pipe breaks in drinking water distribution systems. This is the first, to the authors' knowledge, attempt at using BBNs to model drinking water distribution system pipe breaks. BBNs are a flexible and powerful technique for structuring knowledge bases into joint probability distributions factored according to their causal relationships. These factored, causal structures are directed acyclic graphs that facilitate diagnostic and predictive evidence synthesis [10]. BBNs support patterns of natural human reasoning [11], and may provide a robust mathematical platform for development of real-time drinking water distribution system monitoring and management decision-making.

2. Purpose and objectives

The objective of this paper is to explore the usefulness and practicality of modelling drinking water distribution system pipe breaks with Bayesian Belief Networks [BBNs]. The focus of this

* Corresponding author. Tel.: +1 412 780 6453.
E-mail addresses: seed@gwu.edu (R.A. Francis), sguikema@jhu.edu (S.D. Guikema), lhenneman@gmail.com (L. Henneman).

http://dx.doi.org/10.1016/j.ress.2014.04.024
0951-8320/

paper will be a summary of BBN structure learning in the context of drinking water distribution system pipe break modelling, the cross-validation of BBN structures learned in this case study, and discussion of future research related to the application of BBN models in drinking water systems.

3. Evolution of statistical models of drinking water pipe breaks

Modeling of drinking water distribution system pipe breaks has been an active area of research, especially during the last 10–15 years. Modeling efforts might be classified as either physically-based or statistically-based approaches [12]. The former approach provides insights into distribution system pipe breaks by developing mathematical models based on the fundamental physics governing pipe breaks. While these models are fundamentally compelling, they rely on data that is impractical or impossible to collect in the field. The latter approach provides insights into distribution system pipe breaks by developing statistical models. Statistical models have become widely used in practice because they can often be intuitively interpreted, and span a range of complexity to accommodate a variety of utility data and resource availabilities.

Most statistical models focus on either individual pipes or aggregate pipe groupings. This seems to indicate a practical trade-off of strategic flexibility and model generalizability. Most investigators focusing on evaluating pipe replacement strategies for asset management [13–17] develop statistical models for individual pipe breaks, while investigators focusing on understanding the broader infrastructural impacts on drinking water infrastructure maintenance develop statistical models for aggregate pipe groupings [9,18–20].

It is worthwhile to review a few of the reported model approaches. The intention of this brief exposition is not comprehensiveness, but to illustrate the progression of the role of model uncertainty in pipe break modeling. The basic model applied across a variety of studies is the Weibull-Exponential model. A version of this model was developed by Mailhot et al. [13] for inclusion into an optimization model for evaluating water pipe replacement strategies. In their investigation, Mailhot et al. propose a two-stage stochastic model where a time-dependent breakage risk stage is merged with a constant exponential risk function. Their main contribution is a derivation of the probability distribution for breaks and time between breaks. The focus of their investigation, however, was not prediction of future pipe breaks. Rather, the goal was to characterize the uncertainty in pipe breakage rates for inclusion into asset management models. Their modeling approach is typical of those used in simulation studies used to evaluate pipe replacement and renewal asset management strategies.

Most statistical models developed and tested on field data employ generalized linear models (GLMs) [21]. For example, in their individual water main renewal planner (I-WARP), Kleiner and Rajani [15] developed a GLM based on the non-homogeneous Poisson process. They have included dynamic (e.g., time-dependent) and static (e.g., pipe-dependent) variables while also implementing the zero-inflated Poisson process (ZIP) approach. In their model, the most important risk factors were pipe length, known previous pipe failures, rain deficit, and pump failures. While pipe age and freezing index were also included, their influences were modest when compared with those of the variables above. Kleiner and Rajani conclude that I-WARP is good at estimating the total number of breaks, but not the number of breaks per pipe. However, they suggest I-WARP is useful for prioritizing renewal activities since pipes can be ranked based on the number of predicted pipe breaks.

Similarly, Clark et al. [16] fit a Cox proportional hazards model (Cox-PH) incorporating a shared frailty term. This shared frailty term was represented by a gamma random variate that represented unexplained frailty factors that may cause pipe breaks. Clark et al. classified the pipe segment data into two categories, metallic and non-metallic pipes, and fit the Cox-PH shared frailty model to both classes. Clark et al. conclude that the most important variables are pipe diameter and pipe material. Additionally, their results indicate much more uncertainty in the hazard predictions for the metallic pipe segments compared with the non-metallic segments. They suggest that breakage rates in metallic pipes may be more influenced by unknown random factors than non-metallic pipes.

Berardi et al. [17] utilize evolutionary polynomial regression (EPR) to model pipe breaks based on pipe age, pipe diameter, pipe segment length, and population (locations/properties) serviced. They obtain a model for annual pipe break rate and present a decision support model incorporating previous break history. This model can then be used to predict the number of individual pipe breaks in a given planning horizon by integrating the decision support equation over the time horizon. Although their model is elegant in its simplicity and predictive power, one advantage of using EPR is the opportunity to compare a relatively large family of models, even when the dataset includes a small number of covariates. This diversity in potential model choices indicates model uncertainty might require expert judgment in the interpretation of pipe break models.

With exception of Yamijala et al. [9] and Berardi et al. [17], a survey of statistical models for drinking water distribution system pipe breaks reveals that the predominant statistical approaches are either the Cox proportional hazards model [22] the Weibull–Exponential process, or the non-homogeneous Poisson process to predict time to first break and time between breaks. In addition, with the exception of Yamijala et al. [9] and Vanrenterghem-Raven [23], these models did not include broader environmental and infrastructural information that might help explain pipe breakage rates. These models might be classified as “statistical–physical” in the sense that the risk factors considered directly correspond to the laboratory-based physical models. Consequently, the most important explanatory variables identified in statistical analyses have been pipe age, pipe material, failure history, and pipe length. Although most models do not include broader environmental information and data on proximate infrastructure services and systems, the prevailing approaches in the literature have shown useful for a focus on deploying existing information in financial asset management, renewal, and replacement plans.

On the other hand, some investigators have sought approaches which might comparatively be described as “hypothesis-generating” in the sense that the most often discussed risk factors may not adequately explain the variability in pipe breakage rates for decision making. These hypothesis-generating studies might be further described as having two orientations: predictive management and diagnostic management. The predictive management studies suggest pipe breakage rate models should be augmented with additional environmental and infrastructural data to improve the efficiency of the resultant asset management plans. Consequently, their statistical goal remains parametric prediction of aggregate pipe breaks while including these additional data [9,23]. The diagnostic management studies demonstrate clustering approaches that might be used to further investigate causes of infrastructure failure. The goal is not necessarily to predict pipe breaks in the sense of the prevailing studies, but to identify clusters of susceptible pipes that may be targeted for further surveillance, investigation, or hypothesis generation [18,19].