I^{2}SDS - Decision Sciences
THE GEORGE WASHINGTON UNIVERSITY
SCHOOL OF BUSINESS

The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2013-3
February 11, 2013

Bayesian Belief Networks for Predicting Drinking Water Distribution System Pipe Breaks

Royce A. Francis
Department of Engineering Management and Systems Engineering
The George Washington University, USA

Seth D. Guikema
Department of Geography and Environmental Engineering
The Johns Hopkins University, USA

Lucas Henneman
Department of Geography and Environmental Engineering
The Johns Hopkins University, USA

Bayesian Belief Networks for Predicting Drinking Water Distribution System Pipe Breaks

Royce A. Francis (seed@gwu.edu)[a*], **Seth D. Guikema** (sguikema@jhu.edu)[b], **Lucas Henneman** (lhenneman@gmail.com)[b]

[a] Department of Engineering Management and Systems Engineering, The George Washington University, Washington, DC, USA

[b] Department of Geography and Environmental Engineering, The Johns Hopkins University, Baltimore, MD, USA

Abstract: In this paper, we use Bayesian Belief Networks (BBNs) to construct a knowledge model for pipe breaks in a water zone. Development of expert systems such as BBNs for analyzing real-time 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 real-time distribution system monitoring and management. 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.

Keywords: Bayesian Belief Networks, drinking water distribution systems, expert systems, graphical models.

Corresponding Author: Royce A. Francis, Phone: +1-412-780-6453, E-mail: seed@gwu.edu

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]. However, developing statistical models to predict pipe breaks is a difficult activity. For example, Yamijala et al [3] 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. 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 [4]. BBNs support patterns of natural human reasoning [5], 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 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 [6]. 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 [7-11] 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 [3], [12-14].

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, Poulin and Villenueve [7] 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. They found that the probability density function for the average number of breaks per year is:

$$\tilde{N}(t) = \frac{\tilde{\lambda}_{m+1}}{\alpha}\exp[\alpha t]\left(\exp[\alpha]-1\right) \approx \tilde{\lambda}_{m+1}\exp[\alpha t]$$

where: α is the slope of the linear relationship between hazard values and break orders; λ_j is the parameter of the exponential distribution used to describe time between the $(j\text{-}1)^{th}$ break and the j^{th} break; and m is the m^{th} break. The average number of pipe breaks during the interval [t, t+dt] conditional on the occurrence of k breaks before time t' where ($t' > t$) is:

$$N\left(t \middle| k;t'\right) = \frac{\lambda_{k+1}}{l}\exp\left[\alpha\left(t-t'\right)\right]\cdot dt$$

where l is the length of the pipe segment. 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) [15]. For example, in their individual water main renewal planner (I-WARP), Kleiner and Rajani [9] 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. Their ZIP GLM has the following form:

$$\Pr\left[k_{i,t}\right]=\begin{cases} G_{i,t}+\left(1-G_{i,t}\right)\exp\left[-\lambda_{i,t}\right] \text{ for } k_{i,t}=0 \\ \left(1-G_{i,t}\right)\dfrac{\lambda_{i,t}^{k_{i,t}}}{k_{i,t}!}\exp\left[-\lambda_{i,t}\right] \text{ for } k_{i,t}>0 \end{cases}$$

$$i=1,2,...,N \text{ is the number of pipes}$$

$$t=1,2,...,T \text{ is the number of years of breakage data available}$$

where: $k_{i,t}$ is the number of breaks at year t for pipe i; $G_{i,t}$ is the parameter of the Bernoulli process indicating nonzero breaks (e.g., $k_{i,t}$=0 with probability $G_{i,t}$); and λ is the breakage rate. $G_{i,t}$ is usually incorporated into the ZIP through its own logit link function:

$$G_{i,t}=\frac{\exp\left[f\left(\lambda_{i,t}\right)\right]}{1+\exp\left[f\left(\lambda_{i,t}\right)\right]}$$

Additionally, $\lambda_{i,t}$ takes a general linear form:

$$\lambda_{i,t}=\exp\left[\alpha_0+\theta\cdot\tau\left(g_{i,t}\right)+\alpha\mathbf{z}_i+\beta\mathbf{p}_t+\gamma\mathbf{q}_{i,t}\right]$$

where: a_0 is the intercept; θ is the time-function coefficient; $\tau(g_{i,t})$ is the function used to incorporate pipe age; z_i is a vector of pipe-dependent covariates; p_t is a vector of time-dependent covariates; and $q_{i,t}$ is a vector of covariates dependent on both pipe and time. 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. [10] 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. The metallic (steel, cast-iron, ductile-iron) shared frailty model found to predict breaks/year/pipe section is:

$$h\left(t, material, diameter\right) = \exp\left[7.53 \cdot DIP + diameter\left(-0.309 + 0.152 \cdot DIP\right)\right] \cdot h_{0,DIP}\left(t\right) \cdot W_{PipeID}^{DIP}$$

where: diameter is the pipe diameter (in.); material is ductile iron (DIP), or cast-iron/steel (CIP); DIP is the ductile-iron indicator variable (DIP=1 if DIP, DIP=0 if CIP); t is the years from installation; $W^{DIP}{}_{PipeID}$ is the gamma random variate (mean = 1, variance = 11.43) shared frailty term; and the hazard baseline function for breaks/year/pipe section is $h_{0,DIP}\left(t\right) = \left(5.100 \cdot 10^{-8}\right) t^{3.82}$. The non-metallic (PVC) shared frailty model found is:

$$h\left(t, material = PVC, diameter\right) = \exp\left[-0.202 \cdot diameter\right] \cdot h_{0,PVC}\left(t\right) \cdot W_{PipeID}^{PVC}$$

where: $W^{PVC}{}_{PipeID}$ is the gamma random variate (mean = 1, variance = 0.94) shared frailty term; $h_{0,PVC}\left(t\right) = \left(2.098 \cdot 10^{-8}\right) t^{3.62}$. 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. [11] 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. The first step is to use (EPR) to predict the number of break events recorded in a 14 year monitoring survey:

$$BR_{class} = 0.084904 \cdot \frac{A_{class} \cdot L_{class}}{D_{class}^{1.5}}$$

where: *BR*=number of pipe break events in monitoring period; *A* is the equivalent pipe age within the selected pipe class, $A_{class} = \frac{\sum_{class} pipe\_length \cdot pipe\_age}{\sum_{class} pipe\_length}$; *L* is the sum of the pipe lengths in the class; and *D* is the equivalent pipe diameter within the selected pipe class, $D_{class} = \frac{\sum_{class} \left(pipe\_length \cdot pipe\_diameter \right)}{\sum_{class} pipe\_length}$. This model can be incorporated into a model of the break rate (breaks/year), given an observation period *T*:

$$\lambda_i^{EPR} = \frac{BR_{class}}{T} \cdot \frac{L_i}{L_{class}}$$

where: λ_i is the break rate for pipe *i*; L_i is the length of pipe segment *i*; and L is the sum of the pipe lengths within the selected class. To incorporate this model into decision support, Berardi et al. have obtained the following relationship for incorporating previous break history:

$$\lambda_i(t) = \begin{cases} \frac{a_1}{T} \cdot \frac{Lp_i \left(A_{0,class} + t \right)}{D_{class}^{1.5}} & \text{if } Brp_i = 0 \\ \frac{a_i}{T} \cdot \frac{Lp_i \left(A_{0,class} + t \right)}{D_{class}^{1.5}} & \text{if } Brp_i > 0 \end{cases}$$

where: a_1 is the corresponding EPR model coefficient (here, 0.084904); a_i is given by $a_i = Brp_i \cdot \frac{D_{class}^{1.5}}{A_{0,class}} \cdot \frac{1}{Lp_i}$; Brp_i is the number of previously recorded breaks on segment *I*; $A_{0,I}$ is the equivalent age of the pipe class at the end of the monitoring period *T*; and *t* is the time since the end of the monitoring period. This model can then be used to predict the number of individual pipe breaks in a given planning horizon, *h*, by integrating the decision support equation over the time horizon. Although the model presented here 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. [3] and Berardi et al. [11], 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 [16] 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 [3] and Vanrengenthem [17], 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 [3], [17]. 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 [12], [13].

This paper builds upon this knowledge base by exhibiting features of both hypothesis-based approaches using BBNs to construct a graphical representation of the relationships among auxiliary variables and pipe characteristics. This knowledge discovery technique has the potential to combine predictive management and diagnostic management. The graphical structure of the BBN may facilitate predictive management by allowing distribution system operators and managers to use auxiliary variables to prioritize surveillance in areas indicated by the BBN model, while the inference capabilities of the BBN model permit prediction of pipe breaks using a variety of combinations of data type and quality.

4. METHODS

4.1. Overview of Bayesian Belief Network Modeling

Bayesian Belief Networks (BBNs) [4], [18], [19] have become a popular and effective method for making inferences about relationships between several variables through a causal and probabilistic structure. Because information flows through BBNs according to a causal structure, BBNs are often referred to as "expert" systems (e.g., [20]). BBNs are joint probability distributions between multiple variables expressed as directed acyclic graphs consisting of a set of nodes, a set of edges, and conditional probability tables. In the BBN framework, there are two types of nodes: parent nodes and child nodes. BBNs are attractive for probabilistic reasoning for their decomposability: the probabilities of events represented by child nodes are conditioned only on the events represented by their parent nodes.

Bayesian Belief Networks may be especially well suited for risk assessment and environmental applications. BBNs are quite flexible, due to their non-parametric nature, and can deal with relationships between variables that may be otherwise difficult to identify. Furthermore, the probabilistic language of the BBN is extended by causal language through the directed arcs of the network. The direction of causality may be encoded through expert knowledge, or it may be learned using algorithms available in various software implementations.

The causal reasoning facilitated by BBNs is an important advantage in the potential use of this tool as the basis for a real-time dynamic risk prediction framework, as opposed to rule-based or other data mining approaches. Not only does the BBN allow for direct representation of the "world" the model will be used in, but it also permits diagnostic and predictive (e.g., bidirectional) inference in the event that only subsets of variables of interest may be observed [4]. This feature of BBNs is a principal distinguishing feature, setting it apart as a unique candidate for practical application in drinking water distribution system management. Although a full set of variables may be initially needed to construct the network and learn its structure and parameters, operational use of the network may be applied in situations where the full set of variables may not be observed. Sparse sampling of important variables in time and space, difficulty in measurement, and irregular patterns in relevant data are important features of real-time drinking water network data that may be amenable to modeling through BBNs.

In this project, we use freely available software (e.g., The R Project for Statistical Computing [21], The 'bnlearn' Package for R [22]) to learn the structure and parameters of a BBN that might be used to predict the number of pipe breaks in a water zone. This learning will be based on raw data collected from a mid-eastern US utility, and it will use the constraint-based and score-based methods to learn the BBN structure. To learn the BBN parameters, the EM algorithm is applied to the cross-validated rsmax2 structure in Netica.

The BBN modelling described is a critical step towards real-time distribution system management. Development of an expert system for analyzing real-time 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 real-time distribution system monitoring and management.

4.2. Structure Learning in BBNs

Structure learning is a difficult task in BBN modeling due to its computational complexity (NP-hard) and memory intensiveness [23]. The structure in a BBN model has historically been studied using constraint-based algorithms which first construct the Markov blankets corresponding to each variable, then identifying the independence mappings (I-maps) that correspond to the factorization of the joint probability distribution [24-26]. The I-maps and Markov blankets at this point are typically undirected: the undirected networks are called Markov random fields or Markov networks. These I-maps and Markov blankets are used to direct the links in the network, creating a directed, or partially directed, acyclic graph (DAG or PDAG). The resultant DAG is what is known as the Bayesian Network.

Compared to the Markov network, BBNs are more restricted in the probabilistic relationships they can represent. While all directed relationships shown in a BBN could be encoded in a Markov network, BBNs cannot contain the cycles and certain types of cliques that may usefully be studied using Markov networks. BBNs have remained a bit more popular in practice than Markov networks, however, because of their semantic ability to represent locally-based causal reasoning similar to the type of reasoning humans find natural. In fact, the DAG uniquely represents causal relationships in the knowledge base, and facilitates diagnostic and predictive reasoning [5].

Several constraint-based learning algorithms have found application in recent development of BBN structure learning methods: the grow-shrink algorithm (GS) [27], the incremental association Markov blanket (IAMB), the interleaved incremental association Markov blanket (inter.IAMB), and the fast incremental association Markov blanket (fast.IAMB) [28], [29]. Score-based learning algorithms include the optimization-based hill-climbing (HC) and tabu (tabu) search methods which assume the underlying network follows an approximately multivariate normal (continuous) or multinomial (discrete) distribution in the joint variable space and evaluate the networks using the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). Implementations of these methods may be found in various BBN software packages. Koller and Friedman [30] and Scutari [31] provide helpful summaries of these computational methods.

In this paper, we employ the rsmax2 combined constraint-score-based approach to structure discovery [22], [31]. This algorithm allows the user to specify both the algorithm used to construct the equivalence class of perfect maps and the score used to identify the joint probability distribution that best fits the data. This integrated approach to discovery helps avoid overfitting by maximum likelihood score-based methods by using the equivalence classes identified through the constraint-based algorithms to restrict the score search space [30].

4.3 Parameter Learning in BBNs

While theoretically elegant, learning the structure and parameters simultaneously of a Bayesian Belief Network is considerably more computationally difficult than learning its structure or parameters alone. In addition, the number of parameters required increases greatly as the complexity of the network increases. In even the simplest cases where every variable is binary in the possible states it may take, the number of parameters to be learned increases exponentially as the number of variables in the knowledge domain increases. Even so, for complex knowledge bases, the possibility of learning BBNs from data represents an attractive possibility to augment existing expert elicitation-based approaches. BBN parameters may be learned using maximum likelihood (MLE) or Monte Carlo based methods. For BBN-specific implementations of these methods, the reader is again referred to Scutari [31]. In this paper, we use the EM algorithm implemented in Netica [32].

5. CASE STUDY

5.1. Data Set and Summary Statistics

We have collected pipe break and location data for Baltimore County, MD, USA, during 2010-2011. Our dataset includes 3686 records, augmented with GIS data from the United States Geological Survey (USGS) and the United States Census Bureau. A list of the variables included in our dataset appears in Table 1. In concurrent research we are investigating the use of other data mining and statistical learning techniques, including multivariate adaptive regression splines (MARS), generalized additive models (GAMs), and random forests to predict drinking water distribution system pipe breaks in this case study [33].

The pipe break data included in our case study was recorded by the street address nearest the location of the break, but their pipe locations had not been geocoded as of the time of submission of this paper. Consequently, our analysis is based on pipe break aggregation at the census tract level. As discussed above, most pipe break models are developed at the level of pipe segments including such characteristics as pipe age, pipe diameter, and pipe material and find that these are statistically significant risk factors with respect to pipe break predictive accuracy. However, data on these risk factors were not available for the Baltimore County system, even at an aggregated level. We instead collected publicly available proxies for some of this information. For example, we included the average house age at the census tract level, hypothesizing that census tracts with older houses would tend to have older distribution system pipes. Similarly, we included population density as a proxy for intensity of water use.

Soil characteristics were shown to be important for estimating pipe break risk by Yamijala et al. [3]. We used the United States Department of Agriculture National Resource Conservation Service's (NRCS) Soil Data Mart to capture soil characteristics. We first calculated the percentage of each census tract composed of each of the different NRCS-classified soil types. The 204 original soil types were reduced to 13 in the final model by grouping similar soil types and summing their respective percentages for each tract. The 13 final covariates are: silt loam (SL), urban land complex (ULC), gravelly loam (GL), channery loam (CL), loam and mucky loam (ML), gravelly sandy loam (GSL), quarry, gravel pits, and rock outcrop (ROC), sandy loam and loamy sand (SLS), complex (C), tidally flooded area (TF), udorthents (udorthents), other non-classified soils (other), and water (W).

We also explore the possibility of correlations in population characteristics (e.g., age and ethnic and racial composition of the census tracts) with pipe age. The population data used is from the United States Census Bureau and includes racial and housing occupancy statistics. The variables we consider are: population per square mile (pop), population 18 or older (plus18), percent white (white), percent black or African-American (BLA), percent American Indian and Alaska Native (AAN), percent Asian (Asian), percent Native Hawaiian and other Pacific Islander percent (HIP), percent reporting as two or more races percent (Multi), percent Hispanic or Latino of any race (HL), and percent other. The variables used for housing characteristics were housing units per area (HUPA), occupied houses per area

(OPA), percent of total houses occupied (OP), vacant houses per area (VPA), percent of total houses vacant (VP), and average house age.

Because weather has been shown to strongly influence pipe break risk, particularly winter weather in areas experiencing freeze-thaw conditions, we include a number of weather characteristics calculated on a monthly basis. These were the average daily minimum and maximum temperatures, average daily precipitation, the percentage of days with low temperatures above 32F, the number of days with low temperatures below 32F, and the number of days for which the the high temperature was above 32F and the low temperature was below 32F.

5.2. Results

5.2.1. BBN Learned Structure

The model structures presented in this paper were obtained through constraint-based methods, and hybrid constraint-based/score-based combinations, but not score-based methods alone. While we do calculate the likelihood score for a score-based model for comparison purposes in the results section, score-based methods tend to overfit the training data [30], so in some cases it may be preferred to use constraint-based models. Although constraint-based models may be sensitive to false negative independence tests, they can serve as the starting point for the score-based algorithms in order to avoid overfitting.

The simplest of the constraint-based methods is the grow-shrink algorithm. The structure learned by this algorithm is shown in Figure 1. In the grow-shrink (GS) BBN, “breaks” mediates between two causal basins. One causal basin contains the variables dealing with temperature and precipitation, while the other causal basin includes general infrastructure characteristics, soil types, and other physical variables. Most of the demographic data are not connected to the overall BBN, and these nodes have been omitted from the Figure. This model structure suggests that the temperature and physical relationships studied in the literature to date may be part of the equivalence class in the knowledge structure explaining pipe breaks, but these results suggest that many more variables in the context could have a meaningful influence on the prediction or explanation of pipe breaks.

The other method illustrated is the hybrid score-based/constraint-based approach. The combination of constraint-based and score-based methods may potentially improve the fit and interpretation of the BBN. The BBN structure learned using the rsmax2 algorithm by combining the grow-shrink constraint based algorithm with the tabu search score based method is shown in Figure 2. The structures indicate some differences between the structure learned using the grow-shrink algorithm and the structure learned using the rsmax2 algorithm. In fact, the primary relationships obtained in the grow-shrink BBN between "Breaks" and the knowledge base are not preserved. While breaks is no longer related to the population in each census tract area nor the minimum temperature in the rsmax2 BBN structure, it seems there is a relationship between the housing units per area and vacant housing units per area that was not characterized in the same way by the grow-shrink algorithm.

5.2.2. BBN Model Evaluation

Our BBN models have been evaluated using their negative log-likelihood score. The models have not undergone cross-validation for this article because we are not demonstrating BBN predictive accuracy in this paper. In Table 2, we report the log-likelihood scores computed during evaluation of the network structures obtained using different the grow-shrink and rsmax2 hybrid tabu-grow-shrink methods. These results show that, while the rsmax2 BBN structure represents more dependence relationships, these additional relationships may be difficult to quantify as more data is required, thus resulting in a lower likelihood score.

5.2.3 Variable Importance

An analysis of the sensitivity of pipe break predictions to the findings at other nodes in the network reveals an important challenge remaining in developing BBNs for pipe break modelling and infrastructure rehabilitation planning. In Table 3, we report the sensitivity of findings at "Breaks" in the discrete case to findings at other nodes for the grow-shrink algorithm. This table corresponds to the networks and joint probability distributions shown graphically in Figures 3 and 4. In Table 4, the sensitivity of findings at "Breaks" is reported for the rsmax2 algorithm. To perform the sensitivity analysis in Netica, the data have been discretized via the following approach. First, the data are standardized. In fact, because the learning algorithms seem quite sensitive to scaling, our learned BBN models have been obtained by implementing the learning algorithms on the standardized data. Second, for nodes with large proportions of "ZERO" observations, a discrete state "ZERO" was created. Next, the non-zero data were discretized by quartile, and discrete states for the four quartiles were created. Tables 3 and 4 shows that except observing "Breaks" itself, modest gains in the resolution of uncertainty at "Breaks" may be obtained by resolving uncertainty in other variables. The most important variables in the sensitivity analysis for the GS algorithm were the population serviced, average age of structures in the census tract, the average monthly minimum temperature, and several soil type categories (ML, ULC, silt_loam). In the rsmax2 algorithm, the variables that have the largest influence on pipe breaks were housing units per area, occupied houses per area, vacant houses per area, average maximum and minimum temperature, average age of structures in the census tract, daily precipitation, and the remaining temperature variables. These results agree somewhat with prior pipe break models as several of these variables has been discussed in other statistical modelling investigations of pipe breaks. For example, average age of structures might be thought of as a proxy variable for pipe age; population served was included in the EPR investigation of Berardi et al. [11]; and average monthly minimum temperature was included in Yamijala et al. [3].

On the other hand, the modest information gains reflected in Table 3 illustrates the difficulty in learning BBNs network from our pipe breaks data. The nature of the dataset may pose several challenges to any statistical learning algorithm. First, our pipe breaks dataset is zero-inflated not only in the variable of interest, monthly pipe breaks within census tracts, but also in other variables within the knowledge base. This characteristic may make dependence relationships among variables difficult to ascertain. Second, because the pipe breaks dataset is zero-inflated, discretization of the dataset for the purpose of fitting a discrete BBN may bias the results. While there has been some discussion of the role of discretization techniques for fitting discrete BBNs, these techniques are mathematically sophisticated and not widely employed in BBN software. For example, one approach that may merit discussion in future work is the use of minimal description length (MDL) in the BBN structure learning algorithm [34]. While using the MDL could improve the performance of the BBN approach described in this paper with respect to the sensitivity of findings at "Breaks" to findings at other nodes, it is unclear how the quality of subsequent insights might be improved. Third, while most pipe breaks modelling studies were performed on datasets with access to individual pipe locations with detailed pipe-level characteristics and breakage histories, we do not have access to such detailed data. Consequently, our results may be affected by confounding attributable to unaccounted risk factors indicated in prior research, including breakage history and pipe materials. This idea might be drawn from the BBN results we have presented: only four variables have any influence on the findings at the "Breaks" node when discretizing the nonzero data by quartiles, and each of these variables has been identified as an important risk factor (or proxy for an important risk factor) in previous research. If our dataset was augmented with detailed pipe level characteristics, it is possible that the knowledge map represented by the BBN could indicate more causal paths inducing a higher sensitivity in the findings at the "Breaks" node. Finally, our results indicate agreement with findings in the literature suggesting considerable model uncertainty in the study of drinking water distribution system pipe breaks. While the strongest example of the possibility that a family of models could obtain similar predictions is given by Berardi et al. [11], the inclusion of a "shared frailty" term in the Clark et al. [10] model also suggests that statistical learning algorithms that examine a family of comparable models may be useful. Although the results presented above suggest that our dataset has relationships between breaks and other variables that are not well-fit when conditioning only on the basis of the independence relationships in the data, alternative nonparametric modelling approaches may be appropriate for improving the accuracy of pipe break predictions.

6. DISCUSSION

The BBN approach seems to have some value for knowledge discovery in the overall knowledge base. The BBN approach has the capability to produce a graphical map of the relationships among variables and risk factors in the knowledge base. In addition, the BBN approach has the capacity to facilitate diagnostic (e.g., reasoning based on observations of Breaks' children) and predictive (e.g., reasoning based on observations of Breaks' parents) belief updating. Such capability is nontrivial, especially since such reasoning can proceed with imperfect or incomplete data, conditional on the knowledge map having been constructed.

Despite these potential advantages, our results indicate that BBNs currently do not seem to be the best tool for predicting pipe breaks in this dataset. The research reported here shows that there remain some important obstacles and challenges to be addressed before BBNs might be used to learn both structure and parameters of a knowledge base from data. On the basis of the results of this case study, there may be two specific obstacles worthy of attention: most learning algorithms assume multivariate normality or multinomiality when learning continuous BBNs; and, learning algorithms may be sensitive to the discretization of raw data for learning discrete BBNs.

In this study, both continuous and discrete BBN models were attempted. In the continuous case, as with many statistical learning approaches, the BBN model was sensitive to scaling. In addition, the multivariate normal or multinomial assumption places too much rigidity on structure learning, since the constraint-based approaches implemented in bnlearn require computation of correlation coefficients. This assumption may not sufficiently differentiate BBNs in practice from such approaches as generalized linear models (GLMs), generalized additive models (GAMs), penalized (e.g., lasso and ridge) regression models, or principal components and partial least squares regression models. In fact, we argue that in the case that multivariate normality or multinomiality is crucial to implementation of the structure learning algorithm, BBN models learned from data are advantageous only when evidence propagation from disparate parameter spaces in the joint probability distribution are of interest. In other words, under strict multivariate normality local inference in the BBN is equivalent to the various regression approaches; while global inference in the BBN is equivalent to jointly interpreting the results of several regression models simultaneously. This latter insight is similar to causal chain modeling [35], an approach the authors have not seen used in application to pipe break models. However, BBNs might be superior for integrating various pieces of evidence in real-time inference not necessarily germane to regression or classification models. Perhaps the main consideration limiting the pragmatism of BBN modeling is the multivariate normality assumption for continuous data.

In the discrete case, the most important consideration affecting the validity of model structure and interpretation may be the discretization of continuous data for use with discrete probabilistic models. In some domains, discrete observations and variable states prevail, and may be natural knowledge bases for BBN deployment. Otherwise, care should be taken to justify the discretization of each variable, as the discretization choices may significantly affect the results and interpretation of the resultant BBN model. Alternatively, the copula Bayesian network (CBN) technique might be used to facilitate distribution-free inference via the BBN modelling approach [36-39]. This is the focus of ongoing research conducted by this study's authors.

Another important consideration for the construction and deployment of BBNs for infrastructure management purposes is the causal interpretation of the model. While the model assumes that parents are directed towards children according to their causal relations, the equivalence class approach to structure building may obfuscate these relationships because an undirected link between a pair of variables may sometimes appear in the equivalence class for a DAG. In other words, the models may be score-equivalent when either the link directed from parent to child or child to parent is included in the model. As a result, the learned network may sometimes include directed arcs that do not have a mechanistic justification. Moreover, some directed arcs might be learned from the data according to the independence structure encoded in the data, although these arcs may simply reflect correlations and not causal relationships. If the causal interpretation of the model is paramount, construction of the BBN from expert knowledge may be the best approach.

Finally, this modelling study demonstrates that construction of a BBN from data for the purpose of prediction is a complex and challenging task. More flexibility may be required to construct models of either continuous data that do not fit the multivariate normal assumption or hybrid continuous and discrete data that adhere to a mixture of underlying distributional assumptions. More research into the generalization of BBNs to address these challenges is needed.

ACKNOWLEDGMENTS

The authors acknowledge NSF Grant #1031046, and the George Washington University School of Engineering and Applied Science for supporting this research.

TABLES AND FIGURES

Table 1. Variables included in the pipe breaks dataset. Census and demographic data from the US Census Bureau. Soil Type and Land Cover Variable from the US Geological Survey.

Census, Demographic, and Weather Variables	*Soil Type and Land Cover Variables*
tract: the census tract number	water:
Breaks: the number of breaks in the month-tract pair	silt_loam: silt loam
Month	ULC: urban land complex
Area: census tract area	GL: gravelly loam
pop: the population per area in each census tract	CL: channery loam
plus_18: the percent of the population above 18	ML: loam/mucky loam
white: percent of the population that is white	GSL: gravelly/sandy loam
BLA: percent of the population that is black	ROC: quarry gravel pits/rock outcrop
AAN: percent of the population that is American Indian and Alaska Native percent	SLS: sandy loam and loamy sand
Asian: percent of the population that is Asian	C: complex
HIP: percent of the population that is Hawaiian or Pacific Islander	TF: tidally flooded
Other: percent of the population that is other races	soils: sassafras and croom
Multi: percent of the population that is multiple races	Ubrthents: loamy fill
Latino: percent of the population that is Hispanic or Latino of any race	
HUPA: housing units per area	
OPA: occupied houses per area	
OP: percent of total houses occupied	
VPA: vacant houses per area	
VP: percent of total houses vacant	
AA: average house age	
maxT: average max temperature for the month	
mint: average minimum temperature for the month	
daily_precip: total precipitation for the month divided by area	
greater_32: the percent of days the temperature is above 32F	
less_32: the percent of days that are below 32F	
straddle_32: the percent of days that straddle 32F	

Table 2. Negative Log-likelihood scores of models learned using indicated algorithms. Model likelihood estimated in Netica via the EM algorithm using the complete discretized dataset (N=3686).

Score	GS	Rsmax2, GS
Negative Log-Likelihood	16.38	20.70

Table 3. Sensitivity of findings at "Breaks" node to findings at other nodes using the mutual information criterion. Results shown are for the grow-shrink constraint based method only.

Node	Mutual Information	Percent	Variance of Beliefs
Breaks	0.82885	100	0.1498667
pop	0.0818	9.87	0.0015869
AA	0.01998	2.41	0.0002148
ML	0.01855	2.24	0.0003578
soils	0.00258	0.311	0.0000325
maxT	0.00244	0.295	0.0003051
minT	0.00244	0.295	0.0003051
ULC	0.00223	0.269	0.0000266
silt_loam	0.00198	0.239	0.000022
plus_18	0.00097	0.118	0.0000077
GSL	0.00091	0.11	0.0000081
SLS	0.00058	0.0704	0.000006
HIP	0.00026	0.0315	0.0000016
Latino	0.00006	0.00688	0.0000005
CL	0.00004	0.00512	0.0000006
TF	0.00002	0.00268	0.0000002
Other	0.00001	0.00122	0.0000001
ROC	0.00001	0.000938	0.0000001
GL	0.00001	0.000846	0
Multi	0	0.000566	0
C	0	5.41E-05	0

Table 4. Sensitivity of findings at "Breaks" node to findings at other nodes using the mutual information criterion. Results shown are for the rsmax2 hybrid method using tabu search and the grow-shrink constraint based method.

Node	Mutual Information	Percent	Variance of Beliefs
Breaks	0.82495	100	0.1489128
HUPA	0.07876	9.55	0.0015991
OPA	0.06751	8.18	0.0010228
VPA	0.04668	5.66	0.0006463
maxT	0.02703	3.28	0.0040797
minT	0.02451	2.97	0.0037078
AA	0.01981	2.4	0.0002076
straddle_32	0.01917	2.32	0.0029637
less_32	0.01871	2.27	0.0028761
daily_precip	0.0153	1.85	0.0023565
greater_32	0.01344	1.63	0.0021125
GSL	0.0004	0.0491	0.0000022
ROC	0.00009	0.0109	0.0000005
HIP	0.00007	0.00833	0.0000006
VP	0.00001	0.000805	0
GL	0	8.07E-05	0

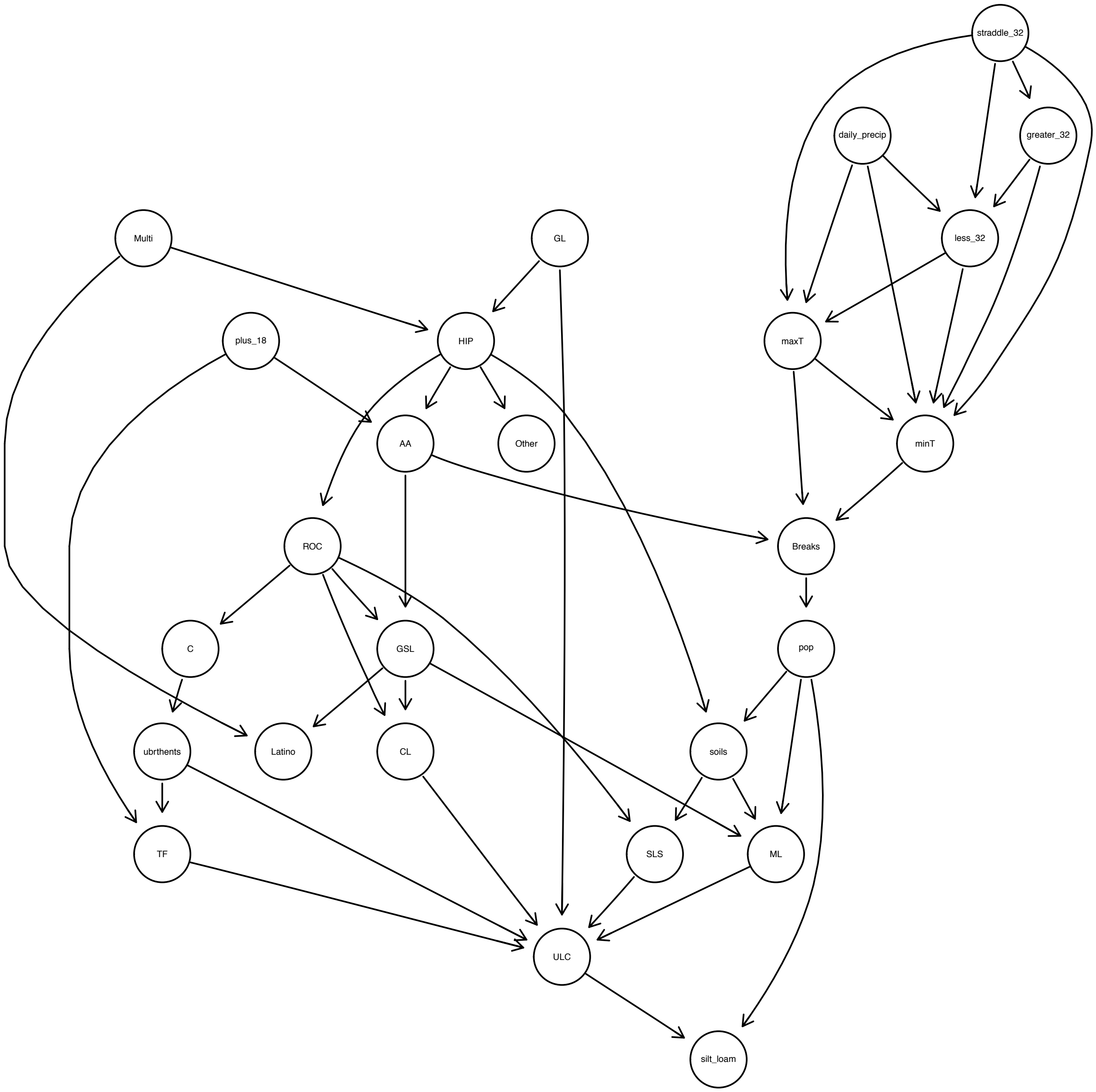

Figure 1. BBN learned from the grow-shrink (GS) algorithm. The variable of interest, "Breaks," appears to separate two "causal basins" dealing with precipitation and temperature (right) and site-specific physical conditions (left).

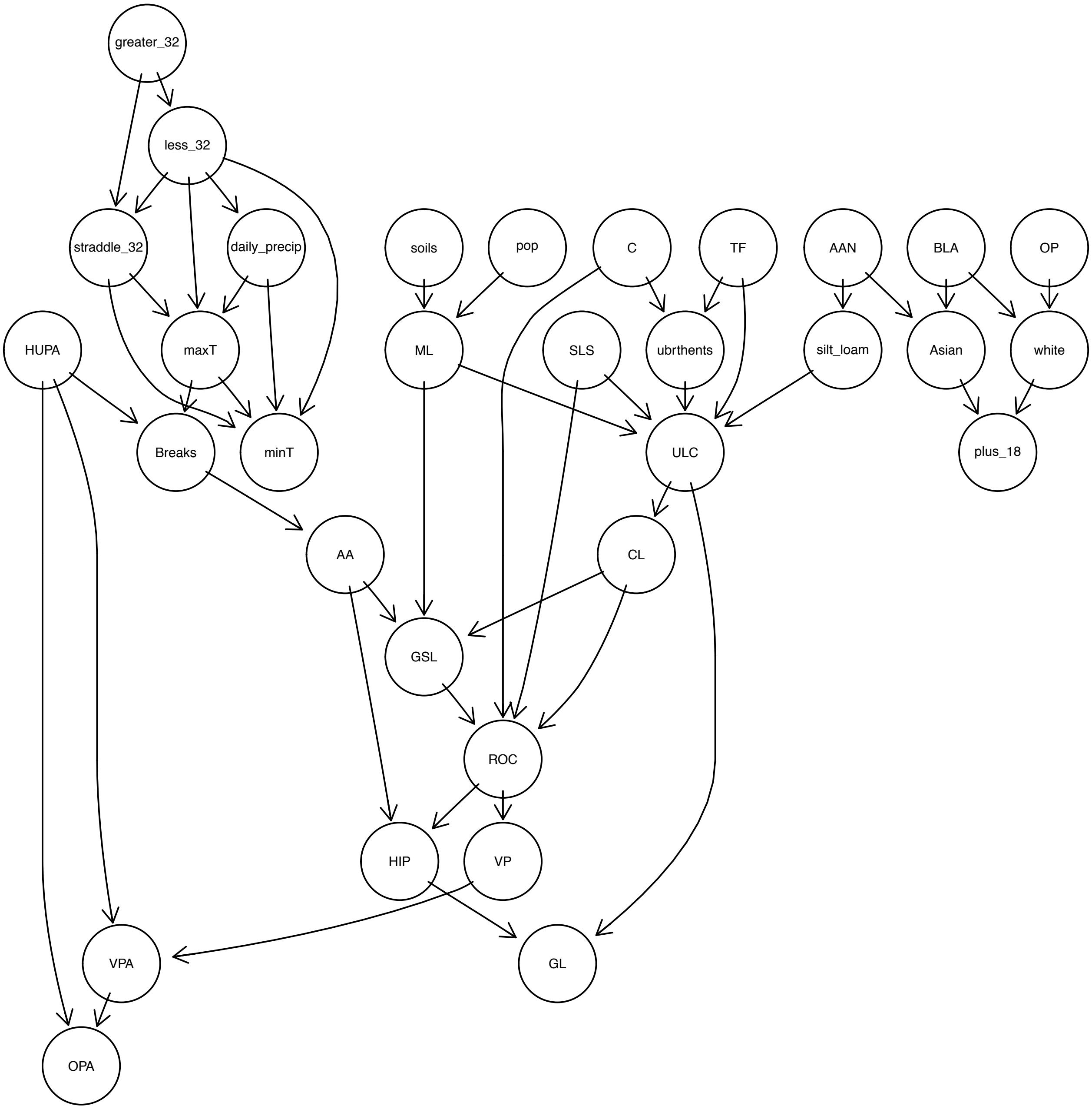

Figure 2. BBN learned by the rsmax2 algorithm with the grow-shrink constraint-based algorithm combined with tabu search.

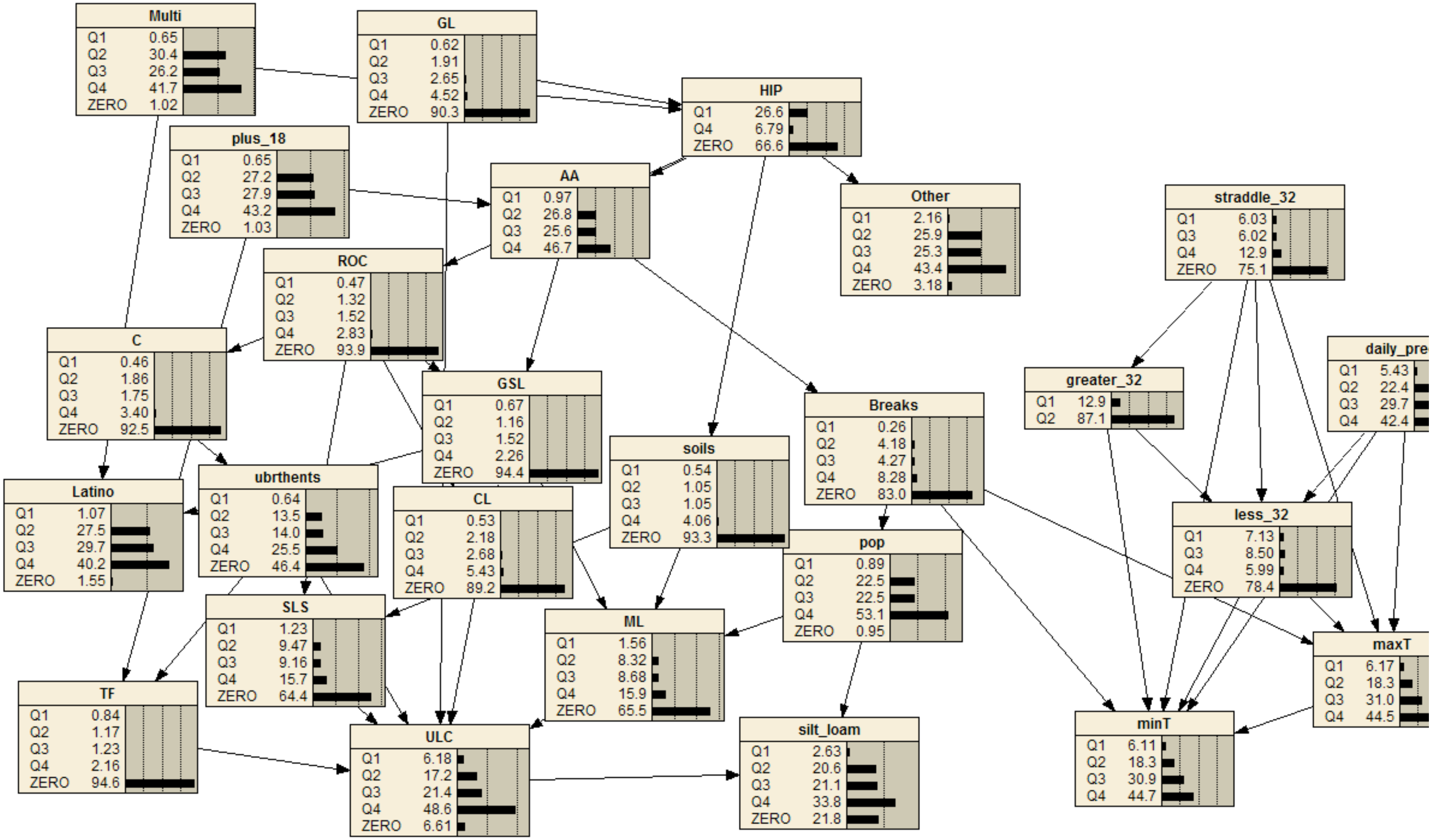

Figure 3. BBN learned using the gs algorithm with parameters learned in Netica using the Gradient Descent algorithm.

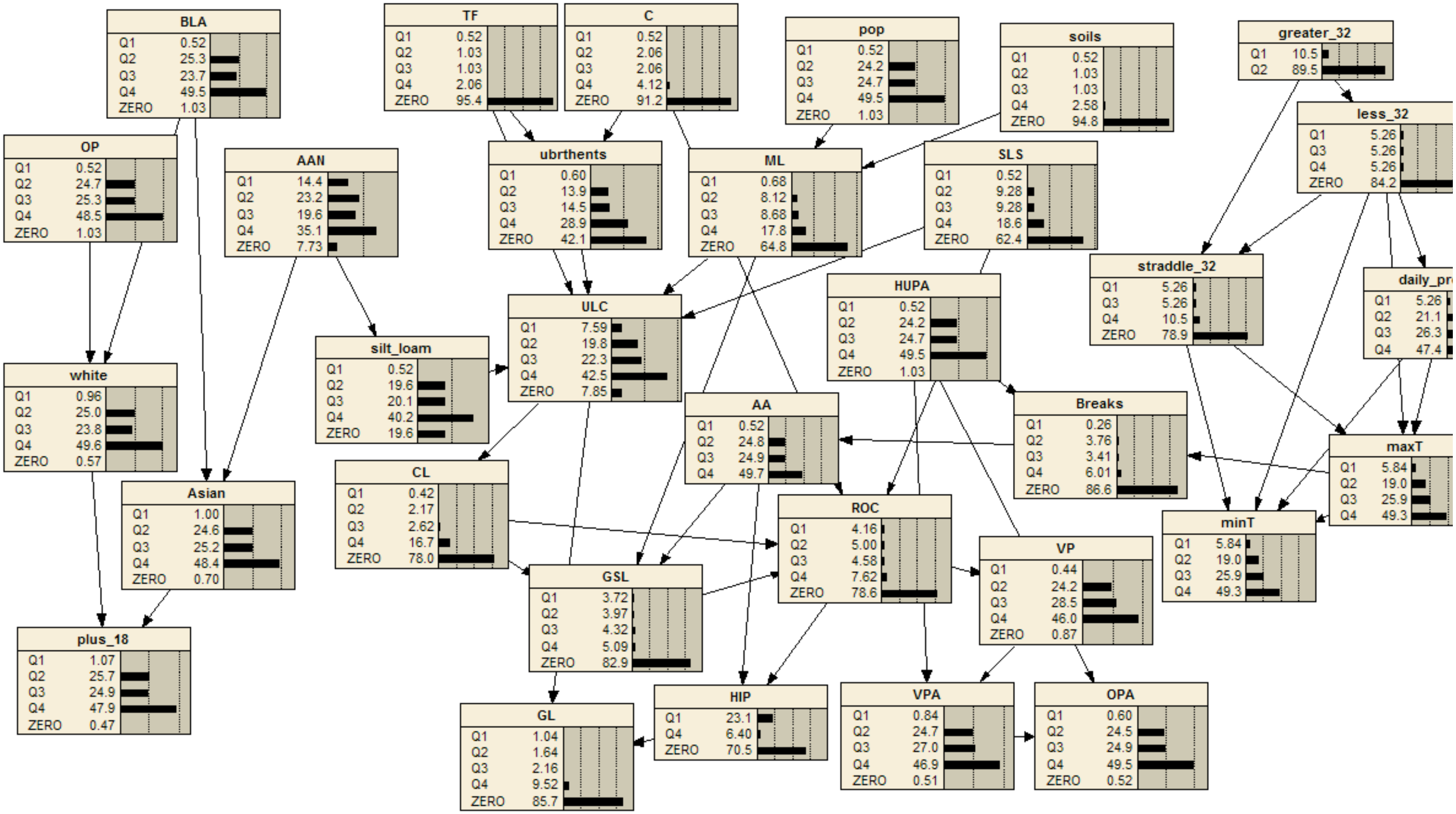

Figure 4. BBN learned using the rsmax2 algorithm (score-learning using tabu search, constraint-based learning using grow-shrink algorithm) with parameters learned using the Gradient Descent algorithm.

REFERENCES

[1] K. Nygard, E. Wahl, T. Krogh, O. A. Tveit, E. Bohleng, A. Tverdal, and P. Aavitsland, “Breaks and maintenance work in the water distribution systems and gastrointestinal illness: a cohort study,” *International Journal of Epidemiology*, vol. 36, no. 4, pp. 873–880, Aug. 2007.

[2] M. W. LeChevallier, R. W. Gullick, M. R. Karim, M. Friedman, and J. E. Funk, “The potential for health risks from intrusion of contaminants into the distribution system from pressure transients,” *Journal of Water and Health*, vol. 1, no. 1, pp. 3–14, 2003.

[3] S. Yamijala, S. D. Guikema, and K. Brumbelow, “Statistical models for the analysis of water distribution system pipe break data,” *Reliability Engineering and System Safety*, vol. 94, no. 2, pp. 282–293, Feb. 2009.

[4] J. Pearl and S. Russell, “Bayesian Networks,” in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. Cambridge: MIT Press, 2001.

[5] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Revised Second Printing. San Francisco: Morgan Kaufmann, 1988.

[6] Y. Kleiner, B. Adams, and J. Rogers, “Water distribution network renewal planning,” *Journal of Computing in Civil Engineering*, vol. 15, no. 1, pp. 15–26, 2001.

[7] A. Mailhot, A. Poulin, and J.-P. Villeneuve, “Optimal replacement of water pipes,” *Water Resour. Res.*, vol. 39, no. 5, p. 1136, 2003.

[8] L. Dridi, A. Mailhot, M. Parizeau, and J.-P. Villeneuve, “Multiobjective Approach for Pipe Replacement Based on Bayesian Inference of Break Model Parameters,” *Journal of Water Resources Planning and Management*, vol. 135, no. 5, pp. 344–354, 2009.

[9] Y. Kleiner and B. Rajani, “I-WARP: Individual Water Main Renewal Planner,” *Drink. Water Eng. Sci.*, vol. 3, no. 1, pp. 71–77, 2010.

[10] R. M. Clark, J. Carson, R. C. Thurnau, R. Krishnan, and S. Paguluri, “Condition Assessment Modeling Using Frailty Modeling,” *Journal AWWA*, vol. 102, no. 7, pp. 81–91, 2010.

[11] L. Berardi, O. Giustolisi, Z. Kapelan, and D. A. Savic, “Development of pipe deterioration models for water distribution systems using EPR,” *Journal of Hydroinformatics*, vol. 10, no. 2, p. 113, Apr. 2008.

[12] D. P. de Oliveira, J. H. Garrett Jr, and L. Soibelman, “A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage,” *Advanced Engineering Informatics*, vol. 25, pp. 380–389, 2011.

[13] D. de Oliveira, D. Neill, and J. Garrett Jr, “Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network,” *Journal of Computing in Civil Engineering*, vol. 25, no. 1, pp. 21–30, 2011.

[14] A. Debón, A. Carrión, E. Cabrera, and H. Solano, “Comparing risk of failure models in water supply networks using ROC curves,” *Reliability Engineering and System Safety*, vol. 95, no. 1, pp. 43–48, Sep. 2009.

[15] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models,” *Journal of the Royal Statistical Society, Series A*, vol. 135, no. 3, pp. 370–384, 1972.

[16] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society, Series B*, vol. 34, no. 2, p. 187, 1972.

[17] A. Vanrenterghem-Raven, “Risk factors of structural degradation of an urban water distribution system,” *Journal of Infrastructure Systems*, vol. 13, no. 1, pp. 55–64, 2007.

[18] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Mach Learn*, vol. 29, pp. 131–163, 1997.

[19] J. Pearl, “Bayesian networks, causal inference, and knowledge discovery,” Computer Science Department, UCLA, Los Angeles, CA, R-281, 2001.

[20] S. K. Andersen, S. K. Andersen, S. K. Andersen, S. K. Andersen, K. G. Olesen, K. G. Olesen, K. G. Olesen, K. G. Olesen, F. V. Jensen, F. V. Jensen, F. V. Jensen, and F. V. Jensen, "HUGIN--A shell for building Bayesian belief universes for expert systems.," in *Readings in Uncertain Reasoning*, San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1990, pp. 332–337.

[21] R Development Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2011.

[22] M. Scutari, "'bnlearn'-an R package for Bayesian network learning and inference," UCL Genetics Institute, University College, London, London, UK, 2011.

[23] D. M. Chickering, "Learning Bayesian networks is NP-complete," in *Learning from Data: AI and Statistics V*, D. Fisher and H. J. Lenz, Eds. Springer-Verlag, 1996, pp. 121–130.

[24] R. D. Schachter, "Model Building with Belief Networks and Influence Diagrams," in *Advances in Decision Analysis: From Foundations to Applications*, no. 10, W. Edwards, R. F. Miles, and D. von Winterfeldt, Eds. New York: Cambridge, 2007, pp. 177–201.

[25] C. K. Chow and C. N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 1–6, 1968.

[26] J. Pearl and T. Verma, "A theory of inferred causation," presented at the Second International Conference on the Principles of Knowledge Representation and Reasoning, Cambridge, MA, 1991, pp. 1–12.

[27] D. Margaritis, "Learning Bayesian Network Model Structure from Data," PhD Thesis. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2003.

[28] I. Tsamardinos, C. F. Aliferis, A. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery," *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference*, pp. 376–381, 2003.

[29] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of Markov blankets and direct causal relations," presented at the KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003.

[30] D. Koller and N. Friedman, *Probabilistic graphical models: Principles and techniques*. Cambridge: MIT Press, 2009.

[31] M. Scutari, "Learning Bayesian Networks with the bnlearn R Package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.

[32] Norsys Software Corporation, "Netica." Norsys Software Corporation, Vancouver, BC, Canada.

[33] L. Henneman, S. D. Guikema, and R. A. Francis, "Predicting Water Distribution System Pipe Break Risk to Support Proactive Pipe Management: Area-Based Models." Manuscript in Preparation, 2012.

[34] N. Friedman, "Discretizing continuous attributes while learning Bayesian networks," *Presented at the International Conference on Machine Learning*, 1996.

[35] J. Whittaker, "Regression and Graphical Chain Models," in *Graphical Models in Applied Multivariate Statistics*, New York: John Wiley & Sons, 1990, pp. 300–344.

[36] A. M. Hanea, D. Kurowicka, and R. M. Cooke, "Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets," *Quality and Reliability Engineering International*, vol. 22, no. 6, pp. 709–729, 2006.

[37] A. M. Hanea, D. Kurowicka, R. M. Cooke, and D. A. Ababei, "Mining and visualising ordinal data with non-parametric continuous BBNs," *Computational Statistics & Data Analysis*, vol. 54, no. 3, pp. 668–687, Mar. 2010.

[38] A. Hanea and D. Kurowicka, "Mixed non-parametric continuous and discrete Bayesian belief nets," in *Advances in Mathematical Modeling for Reliability*, no. 1, T. Bedford, J. Quigley, L. Walls, B. Alkali, A. Daneshkhah, and G. Hardman, Eds. Amsterdam, Netherlands: Advances in mathematical modeling for reliability, 2008, pp. 9–16.

[39] A. Hanea and W. Harrington, "Ordinal $PM_{2.5}$ data mining with non-parametric continous bayesian beleif nets," *Information Processes Journal*, vol. 9, no. 4, pp. 280–286, 2009.