Prediction of Water Mains Failure - A Bayesian Approach

Golam Kabir
PhD Candidate, School of Engineering, University of British Columbia, Kelowna, BC, Canada

Solomon Tesfamariam
Associate Professor, School of Engineering, University of British Columbia, Kelowna, BC, Canada

Rehan Sadiq
Professor, School of Engineering, University of British Columbia, Kelowna, BC, Canada

ABSTRACT: To develop an effective preventive or proactive repair and replacement action plan, water utilities often rely on water main failure prediction models. However, in the prediction modeling water mains failure, uncertainty is inherent regardless of quality and quantity of data used in model-data fusion. To improve the understanding of water main failure processes, a new and effective Bayesian framework 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. In this study, Bayesian model averaging method is presented to identify the influential pipe-dependent and time-dependent covariates whereas Bayesian Weibull proportional hazard model is applied to develop the survival curves and to predict the failure rates of CI and DI pipes.

1. INTRODUCTION

It is very difficult to fully understand the processes that cause failure in underground water mains which are repairable components (Le Gat and Eisenbeis 2000). Due to multiple factors affecting these failure processes and data scarcity, it is often complex and hard to develop water main failure models statistically (Røstum 2000). Accurate water main break prediction models are vital for water utilities in terms of budgeting and to prioritize the maintenance, rehabilitation and replacement (M/R/R) of water mains (Kabir et al. 2015). For this, substantial efforts have been made to develop pipe failure prediction using statistical models. Statistical models attempting to predict the behaviour of water pipes are not only affected both by the quantity and quality of available data, but also by the applied statistical techniques (Kleiner and Rajani 2002, 2001).

Survival analysis is a branch of statistics dealing with deterioration and failure over time and involves the modelling of the elapsed time between an initiating event and a terminal event (Dridi et al. 2009; Kleiner and Rajani 2001). Survival analysis incorporates the fact that while some pipes break, others do not and this information has a strong impact on pipe failure analysis (Røstum 2000). The models use covariates to differentiate the pipe failure distributions without splitting the failure data, thereby giving a better understanding of how covariates influence the failure of the pipe (Le Gat and Eisenbeis 2000). Kleiner and Rajani (2001) resented brief review of the survival analysis methods for modelling pipe failure.

Because of the incomplete and partial information, integration of data/information from different sources, involvement of human (expert) judgment for the interpretation of data and observations, uncertainties become an integral part of the water main failure prediction models (Kabir et al. 2015). Moreover, the decision-making problem becomes more complex and uncertain when multiple experts are involved who have different levels of credibility about their knowledge of the problem (Tesfamariam et al. 2006). Data quality also becomes a serious issue

as many data sets contain uncertainties, e.g. due to unreliable recording of failure times or inaccurate measurements of the confounding factors or even the lack of the actual failure times (Economou et al. 2007). Therefore, some researchers presented Bayesian inference or analysis for water main failure model considering the model parameters as random variables and incorporated external information (e.g. elicited expert opinions, relevant historical information) into the model by constructing a probability distribution that describes the uncertainty in the model parameters (prior to the observing data from the experiment) (Dridi et al. 2009; Economou et al. 2007; Watson et al. 2004).

In most of the Bayesian analysis, nonhomogeneous Poisson process (NHPP) (Economou et al. 2007; Watson et al. 2004) and exponential/Weibull models (Dridi et al. 2009) are considered. However, all these studies only consider pipe age for their analysis ignoring other influential physical (i.e., length, diameter, and manufacturing period) and environmental (i.e., soil condition, temperature) factors. On the other hand, most of the water main failure prediction studies did not mentioned any covariate or model selection method that can handle the uncertainties. Therefore, the objective of this study is to develop a new and effective Bayesian analysis framework for failure rate prediction of water mains. For this, Bayesian model averaging (BMA) is used for covariate selection and Bayesian Weibull Proportional Hazard Model (BWPHM) is used for the failure rates prediction of water mains. The proposed framework will enhance the predictive capability of pipe failure models and will assist the utility authorities to better address the structural and hydraulic failure of water mains, proactively.

2. METHODOLOGY

The framework of the proposed study is shown in Figure 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, the influential and significant covariates will be selected using the BMA approach. In the third step, water mains failure prediction model will be developed using BWPHM. Finally, the model will be evaluated using 5-fold cross validation method. The following subsections briefly discuss BMA and BWPHM method.

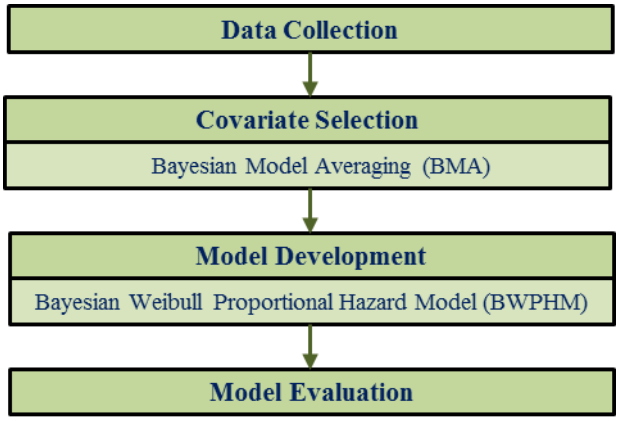

Figure 1: Proposed framework.

2.1. Bayesian Model Averaging (BMA)

BMA is an average of the posterior distributions under each model weighted by the corresponding posterior model probabilities (Leamer 1978). If $M = \{M_1, \ldots, M_K\}$ denotes the set of all models being considered and if Δ is the quantity of interest, then the posterior distribution of Δ given the data D is (Raftery et al. 1997)

$$Pr(\Delta|D) = \sum_{k=1}^{K} Pr(\Delta|M_k, D) Pr(M_k|D) \quad (1)$$

The posterior probability of model M_K is given by (Leamer 1978; Raftery et al. 1997)

$$Pr(M_k|D) = \frac{Pr(D|M_k)Pr(M_k)}{\sum_{l=1}^{K} Pr(D|M_l)Pr(M_l)} \quad (2)$$

where

$$Pr(D|M_k) = \int Pr(D|\theta_k, M_k)\, Pr(\theta_k|M_k)\, d\theta_k \quad (3)$$

is the marginal likelihood of model M_k, θ_k is the vector of parameters of model M_k, $\Pr(D|\theta_k, M_k)$ is the likelihood, $\Pr(\theta_k|M_k)$ is the prior density of θ_k under model M_k, and $\Pr(M_k)$ is the prior probability that M_k is the true model (Raftery et al. 1997).

Averaging over *all* of the models in this fashion provides better predictive ability, as measured by a logarithmic scoring rule, than using any single model M_j:

$$-E\left[log\left\{\sum_{k=1}^{K} Pr(\Delta|M_k, D)Pr(M_k|D)\right\}\right] \leq -E\left[log\{Pr(\Delta|M_j, D)\}\right] \quad (j = 1, \dots, K), \tag{4}$$

where Δ is the observable to be predicted and the expectation is with respect to $\sum_{k=1}^{K} Pr(\Delta|M_k, D)Pr(M_k|D)$(Raftery et al. 1997).

2.1. Bayesian Weibull Proportional Hazard Model (BWPHM)

Weibull Proportional Hazard Model (WPHM) is a parametric version of Cox-PHM but the baseline hazard function is assumed to follow a specific distribution when the model is fitted with data (Le Gat and Eisenbeis 2000). In WPHM, a set of vector of covariates x_1, ..., x_p and error term ϵ, is assumed to be linearly related to the logarithm of time T, and can be expressed as the loglinear model for the *i*th individual

$$log\ t_i = \mu + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \sigma\epsilon_i \tag{5}$$

where β_1, ..., β_p are the regression coefficients of the p covariates and ϵ_i is assumed to have a Gumbel distribution with density $f(\epsilon) = \exp(\epsilon - e^{\epsilon})$. There are p+2 unknown parameters in this WPHM model, the p regression coefficients, the constant term or intercept μ, and the scale parameter σ (Abrams et al. 1996; Albert 2009).

The density of the log time, $y_i = log\ t_i$ is given by (Abrams et al. 1996; Albert 2009)

$$f_i(y_i) = \frac{1}{\sigma} exp(z_i - e^{z_i}) \tag{6}$$

where, $z_i = (y_i - \mu - \beta_1 x_{i1} - \dots - \beta_p x_{ip})/\sigma$. Also, the survival function for the *i*th individual is given by

$$S_i(y_i) = \exp(z_i - e^{z_i}) \tag{7}$$

Then the likelihood function of the regression vector $\beta = (\beta_1, \dots, \beta_p)$, intercept μ and scale parameter σ is given by (Albert 2009)

$$L(\beta, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(y_i)\}^{\delta_i} \{S_i(y_i)\}^{1-\delta_i} \tag{8}$$

where δ_i is a censoring indicator. If $\delta_i = 1$, the observation is not censored and t_i is the actual survival time. Otherwise when $\delta_i = 0$, the observation t_i is the censored time.

If uniform priors are assigned for μ, β and the usual noninformative prior proportional to $1/\sigma$ is assigned for scale parameter σ, the posterior density up to a proportionality constant can be expressed as (Abrams et al. 1996; Albert 2009)

$$g(\beta, \mu, \sigma|data) \propto \frac{1}{\sigma} L(\beta, \mu, \sigma) \tag{9}$$

3. CASE STUDY

The proposed methodology is applied on the water distribution network of the City of Calgary. The City of Calgary is located in Alberta, Canada and has a population of 1.1 million people.

3.1. Data Collection and Preparation

The water distribution consists of 4,281km length of pipe with a total of 49,531 individual pipes. The City of Calgary water network comprises of 21.92% ductile iron (DI), 16.10% cast iron (CI), 3.76% cementitious (asbestos cement concrete and concrete cylinder pipes), 2.72% steel, 0.86% copper, and 54.64% plastic pipes. Figure 2 indicates that among 17,682 pipe breaks majority of breaks occurred in CI (64.37%) and DI pipes (32.08%) whereas very few breaks found for cementitious (2.36%), plastic (1.18%), steel (0.57%), and copper (0.08%) pipes. Only CI and DI pipes are considered for further analysis because of high percentage of breaks.

Pipe characteristics data like age, diameter, length, vintage or manufacturing period, number of connection of each pipes (for land use determination), soil resistivity, and soil corrosivity index are collected from GIS database of the City of Calgary. It has been found that 2,882 CI pipes and 2,067 DI pipes experienced breaks from 1956-2013. The pipes installed during the period 1960-1976 experienced high number (80%) of breaks. Therefore, two groups were considered for vintage (VINT), pipes installed during 1960-1976 and others. Other weather data such as temperature, precipitation and rainfall was acquired from Environment Canada. Data from Calgary International Airport CS weather station (Latitude: 51°06'31.080" N, Longitude: 114°00'52.000" W) was used and daily mean temperature is considered for the analysis. Rain deficit (RD) and freezing index (FI) are calculated according to the equation provided in Kleiner and Rajani (2002). The variables used to represent input data together with summary statistics of these data are given in Table 1.

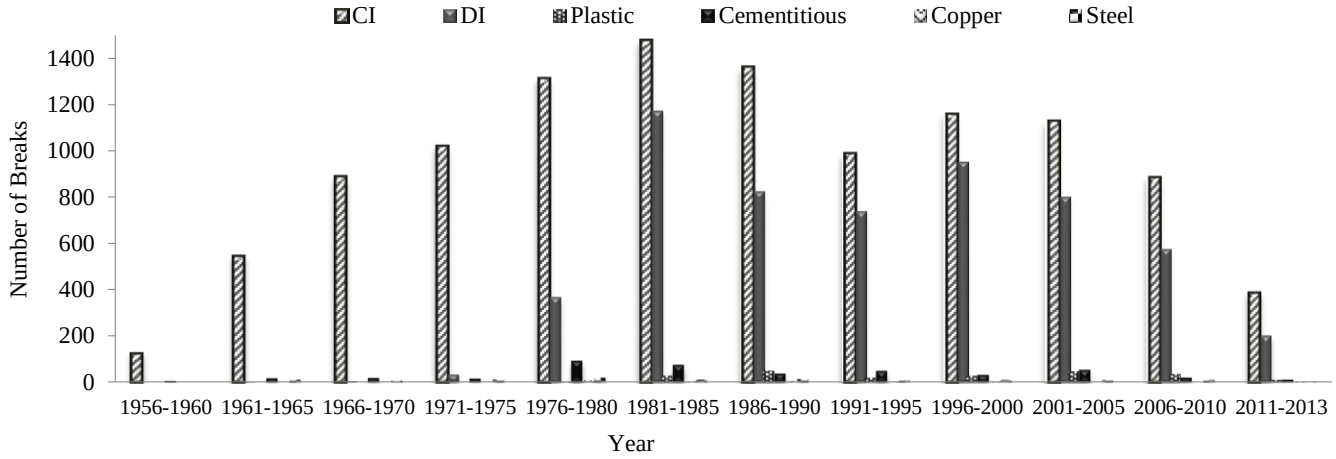

Figure 2: Breaks of the City of Calgary.

Table 1: Summary of control variables used in this study.

Variable	*Description*	*Unit*	*Measured scale*	
			Cast Iron (CI)	*Ductile Iron (DI)*
NBRKS	*Number of previous breaks*	*NA*	*min:1, max:22, mean:2.284, SD:1.86*	*min:1, max:16, mean:2.271, SD:1.85*
DIA	*Pipe diameter*	*mm*	*min:20, max:600, mean:184.3, SD:63.7*	*min:100, max:400, mean:205.3, SD:63.74*
LENGTH	*Pipe length*	*m*	*min:3, max:937, mean:164.3, SD:89.13*	*min:4, max:1691, mean:185.6, SD:102.36*
FI	*Freezing index*	*degrees-days*	*min:-232.96, max:672.86, mean:9.18, SD:67.75*	*min:-232.96, max:622.43, mean -18.614, SD:58.12*
RD	*Rain deficit*	*cm*	*min:-22.226, max:218.8, mean:2.88, SD:21.51*	*min:-21.29, max:218.8, mean:21.65, SD:23.75*
VINT	*Vintage*	*NA*	*min:0, max:1, mean:0.82, SD:0.384*	*min:0, max:1, mean:0.35, SD:0.475*
RESIS	*Soil resistivity*	*Ωm*	*min:0, max:15382, mean:2304, SD :1366.14*	*min:0, max:16666, mean:1894, SD :1170.459*
SCI	*Soil corrosivity index*	*NA*	*min:0, max:20.430, mean:7.293, SD:5.138*	*min:0, max:20.428, mean:10.373, SD:6.269*
LANUSE	*Land use*	*NA*	*min:0 (residential), max:1 (commercial), mean:0.21, SD:0.41*	*min:0 (residential), max:1 (commercial), mean:0.16, SD:0.37*

NA symbolizes that there are no units

3.2. Covariate Selection

The data are stratified according to material type cast iron (CI) and ductile iron (DI) in order to establish the influence of covariates. For both CI and DI, two strata are defined according to the number of observed previous failures (NOPF); with no previous failure (NOPF=0) and with one or more previous failures (NOPF >0) (Le Gat and Eisenbeis 2000). Covariates to be included into the models are selected using BMA. The posterior distributions for the coefficients of CI pipes (NOPF >0) model based on the model averaging results are shown in Figure 3. The posterior distribution for LENGTH, DIA, and RESIS are indeed centered away from 0, and RESIS with a moderate spike at 0 whereas FI, and RD are centred close to zero with a large spike at zero. On the other hand, the posterior distribution of VINT, SCI, and LANUSE are centered at zero. Therefore, LENGTH, DIA, and RESIS are finally selected for CI pipes (NOPF >0) model. Similarly, the other influential covariates are determined for the remaining three strata models.

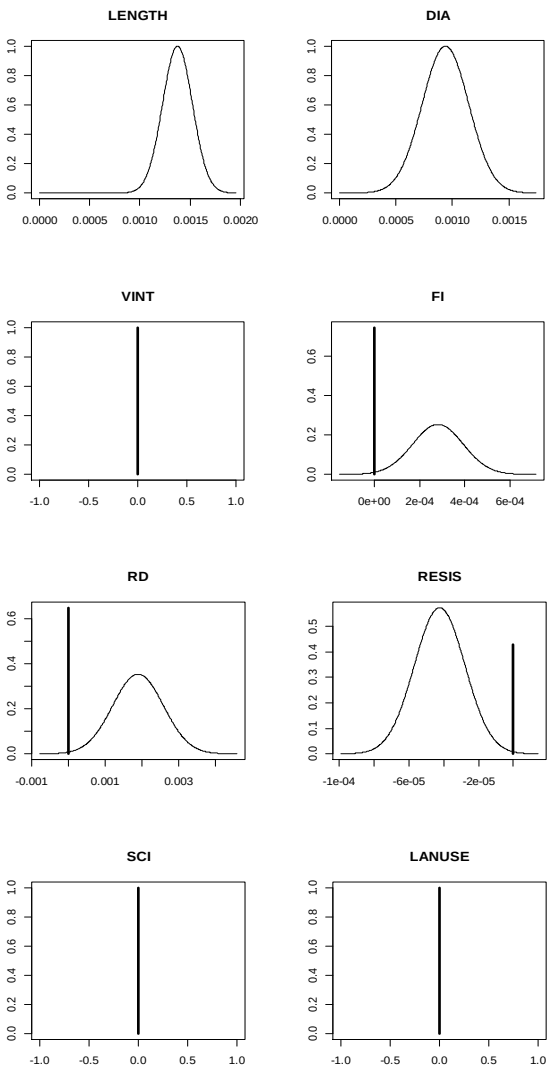

Figure 3: Posterior distribution of CI pipes (NOPF >0) based on BMA.

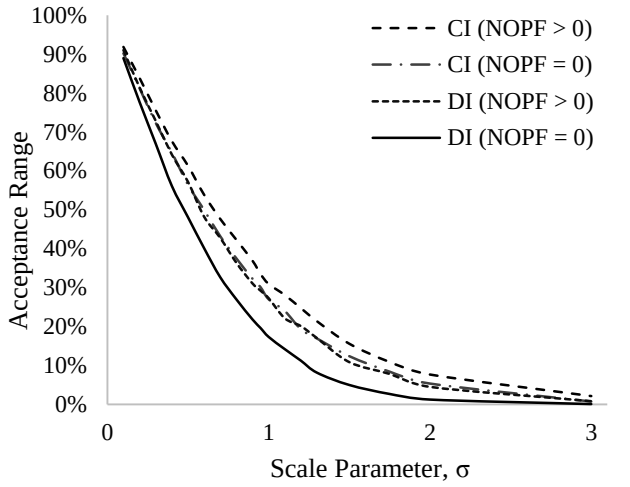

Figure 4: Selection of scale parameters for BWPHMs.

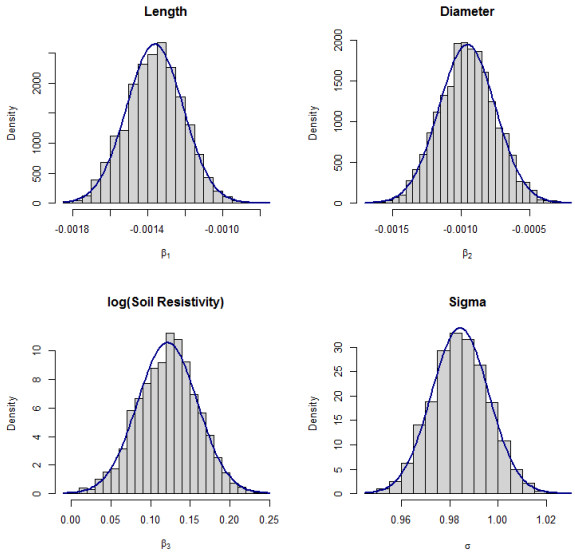

Figure 5: Histogram of coefficients of BWPHM for CI pipes (NOPF > 0).

3.3. Model Development

The influential covariates from the results of BMA are considered for BWPHM development. To find a proposal density, scale parameter (σ) have to choose so that the Metropolis random walk chain has an acceptance range in the 20–40% range (Albert 2009). Figure 4 indicates the relationship between scale parameters and acceptance range for both CI and DI pipe strata. The final scale parameters for the different BWPHMs are chosen for 30% acceptance rate.

The posterior distribution of the parameters or coefficients of BWPHM for CI pipes (NOPF > 0) from the joint posterior distribution is shown in Figure 5. Similarly, the posterior distribution of the coefficients of other BWPHMs are determined. In order to test the normality of the coefficients, Shapiro–Wilk, Anderson-Darling and Kolmogorov-Smirnov normality tests were performed. For all the coefficients of the different BWPHMs, p-value were found greater than 0.05. Thus, we failed to reject the hypothesis that the data is normally distributed (Thode 2002). Table 2 present the mean and standard deviation of the

coefficients of BWPHMs for CI and DI pipe strata with NOPF=0 and NOPF>0.

Table 2: Mean and standard deviation of the coefficients of BWPHM models for CI and DI pipes.

	CI pipes				*DI pipes*			
	NOPF = 0		*NOPF > 0*		*NOPF = 0*		*NOPF > 0*	
	Mean	*Std dv*	*Mean*	*Std dv*	*Mean*	*Std dv*	*Mean*	*Std dv*
Intercept	*7.17E+00*	*3.11E-02*	*4.29E+00*	*2.76E-01*	*6.64E+00*	*3.22E-01*	*4.76E+00*	*9.98E-02*
LENGTH	*-3.46E-03*	*1.35E-04*	*-1.36E-03*	*1.39E-04*	*-2.85E-01*	*2.52E-02*	*-8.81E-04*	*2.45E-04*
DIA	*0*	*0*	*-9.52E-04*	*2.01E-04*	*1.74E-03*	*2.60E-04*	*0*	*0*
VINT	*0*	*0*	*0*	*0*	*-5.62E-01*	*3.72E-02*	*-3.02E-01*	*7.62E-02*
FI	*-1.46E-03*	*1.26E-04*	*0*	*0*	*-2.73E-03*	*1.42E-04*	*0*	*0*
RD	*1.02E-01*	*2.92E-02*	*0*	*0*	*-1.53E-02*	*3.72E-04*	*0*	*0*
log.RESIS.	*0*	*0*	*1.20E-01*	*3.59E-02*	*1.72E-01*	*3.77E-02*	*0*	*0*
SCI	*0*	*0*	*0*	*0*	*0*	*0*	*-1.77E-02*	*3.80E-03*
LANUSE	*1.82E-01*	*4.84E-02*	*0*	*0*	*4.72E-01*	*3.90E-02*	*2.30E-01*	*5.92E-02*

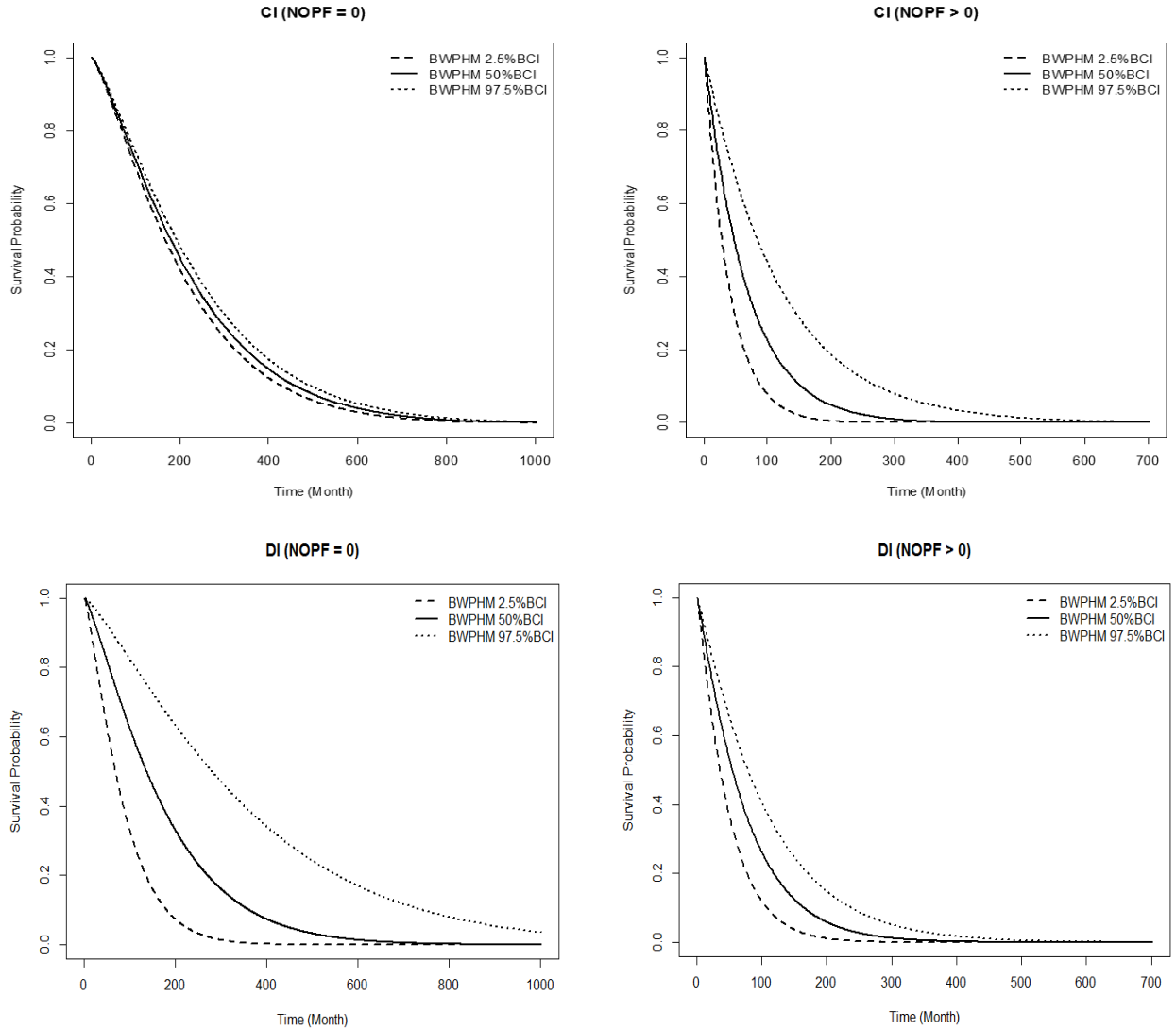

Figure 6: Survival curves from BWPHM for CI and DI pipe strata with NOPF=0 and NOPF>0.

Based on the mean and standard deviation of the coefficients presented in Table 2, the survival curves are determined for CI and DI pipe strata with NOPF=0 and NOPF>0. Figure 6 shows the 2.5th, 50th, and 97.5th percentiles survival curves or posterior median and 95% Bayesian interval estimates for the survival of CI and DI pipe strata with NOPF=0 and NOPF>0. Figure 6 indicates that the survival time of DI and CI for NOPF>0 are almost similar but the survival time of CI is higher compared to DI or CI has a better survival rate than DI pipe for NOPF=0. Figure 7 also shows that, the survival curves for pipes with previous breaks (NOPF>0) are steeper than the curves for pipes with no previous breaks (NOPF=0). For this, the survival time of CI and DI pipes with NOPF=0 is higher than NOPF>0 which is the normal phenomena for the repairable components.

In the past studies, it has been found that the number of breaks increases considerably with the length of the pipes (Le Gat and Eisenbeis 2000; Røstum 2000). Table 2 also indicates that pipe length marginally increase the rate of breakage or longer pipe breaks more than shorter pipes for both CI and DI pipes, either with one or more breaks. According to the Table 2, pipes installed in high soil resistivity are less likely to fail than those installed in soils with low resistivity. After CI pipe has experienced more than one break, soil resistivity dominates the other covariates. Table 2 also reveals that high soil corrosivity index increases the hazard of the pipe while low resistivity decrease it; after DI pipe has experienced more than one break. Table 2 also shows that the freezing index and rain deficit effects more for the occurring of first failure compared to successive failures. The negative coefficients of FI from indicate that more breaks are expected with the decrease of temperature. However, the increase of hazard for FI is marginal due to the 3m installation depths in the city that masked the influence of the temperature by limiting the penetration of frost loads.

3.4. Model Evaluation

In order to evaluate the performance of the BWPHM and Cox-PHM models (Cox 1992; Røstum 2000), 5-fold cross validation method (Efron and Tibshirani 1997) are used. The overall match between observed and predicted values for each of the folds was assessed using mean absolute error (MAE), root mean square error (RMSE), and mean relative absolute error (MRAE) by the following equations.

$$MAE = \frac{\sum_{i=1}^{n} |O - P|}{n} \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (O - P)^2}{n}} \quad (11)$$

$$MRAE = \frac{1}{n} \frac{\sum_{i=1}^{n} |O - P|}{\sum_{i=1}^{n} |O - \bar{P}|} \quad (12)$$

where, O and P are observed and predicted values respectively, $\bar{P}$ is mean of P, and n is number of observations. Table 3 gives the comparison results of BWPHM and Cox-PHM models for CI and DI strata. The results show that the performance of BWPHM is noticeably better compared to Cox-PHM.

Table 3: Performance of BWPHM and Cox-PHM for break prediction.

Model Errors		Models	Fold				
			1	2	3	4	5
MAE	CI	BWPHM	5.95	5.89	5.99	5.35	7.19
		Cox-PHM	5.77	6.17	6.51	7.14	6.99
	DI	BWPHM	5.85	5.77	6.25	5.68	5.12
		Cox-PHM	10.13	10.58	9.91	10.16	9.82
RMSE	CI	BWPHM	7.86	8.54	7.55	6.57	9.27
		Cox-PHM	7.87	8.61	8.82	9.45	9.55
	DI	BWPHM	8.52	10.03	9.30	9.24	8.06
		Cox-PHM	13.56	14.36	13.76	13.77	13.13
MRAE	CI	BWPHM	0.74	0.95	1.23	2. 27	0.87
		Cox-PHM	1.88	1.06	1.58	2. 75	0.59
	DI	BWPHM	0.53	0.72	1.18	0.95	0.44
		Cox-PHM	1.72	2.94	2.57	3.63	1.73

4. CONCLUSIONS

The aim of this study was to develop a Bayesian framework for failure rate prediction of water mains considering uncertainty, since accurate quantification of uncertainty is necessary for improving our understanding of water mains' failure processes. BMA is conducted in this study

to bring insight on selecting appropriate covariates and Bayesian Weibull proportional hazard model is applied to develop survival curves for CI and DI pipes using 57 years of historical data collected for the City of Calgary. The results indicated that the impact of the covariates influencing pipe failure differ according to material type. The results also indicated that the survival time of CI and DI pipes with NOPF=0 is higher than NOPF>0. After experiencing first break, soil resistivity is the most significant or influential parameters for the increases the hazard of the CI pipes. The proposed study can be strengthen a great deal by integrating it with the GIS system of the utilities to develop an effective program of municipalities.

As a future research, the results from this framework can be further integrated with economic assessment model (e.g., life cycle costing) to estimate costs of M/R/R and to develop optimal maintenance or replacement plans. In order to reduce the uncertainty in environmental factors and land use, weather data from multiple sample stations and zoning system of the City of Calgary, and traffic load in the vicinity can be considered.

5. ACKNOWLEDGEMENTS

The financial support to the second and third authors through Natural Sciences and Engineering Research Council of Canada (NSERC) under Discovery and CRD grant programs is acknowledged.

6. REFERENCES

Abrams, K., Ashby, D., and Errington, D. (1996). "A Bayesian approach to Weibull survival models—application to a cancer clinical trial". *Lifetime Data Analysis*, 2(2), 159-174.

Albert, J. (2009). *Bayesian Computation with R*, 2nd Ed., Springer, New York, 304.

Cox, D.R. (1992). Regression models and life-tables. In *Breakthroughs in Statistics* (pp. 527-541). Springer New York.

Dridi, L., Mailhot, A., Parizeau, M. and Villeneuve, J.P. (2009). "Multiobjective Approach for Pipe Replacement Based on Bayesian Inference of Break Model Parameters". *Journal of Water Resources Planning and Management*, 135(5), 344-354.

Economou, T., Kapelan, Z., and Bailey, T.C. (2007). "An aggregated hierarchical Bayesian model for the prediction of pipe failures." In: Proceedings of 9th International Conference on Computing and Control for the Water Industry (CCWI), Leicester, UK.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548-560.

Kabir, G., Tesfamariam, S., Francisque, A., and Sadiq, R. (2015). "Evaluating risk of water mains failure using a Bayesian belief network model," *European Journal of Operational Research*, 240(1), 220-234.

Kleiner, Y., and Rajani, B. (2002). "Forecasting Variations and Trends in Water-Main Breaks." *Journal of Infrastructure Systems*, 8(4), 122–131.

Kleiner, Y., and Rajani, B. (2001). "Comprehensive review of structural deterioration of water mains: statistical models." *Urban Water*, 3(3), 131–150.

Leamer, E.E. (1978), *Specification Searches*, New York: Wiley.

Le Gat, Y., and Eisenbeis, P. (2000). "Using maintenance records to forecast failures in water networks." *Urban Water*, 2(3), 173–181.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models". *Journal of the American Statistical Association*, 92(437), 179-191.

Røstum, J. (2000). Statistical modelling of pipe failures in water networks. PhD Dissertation, Norwegian University of Science and Technology.

Tesfamariam, S., Rajani, B. and Sadiq, R. (2006). "Consideration of uncertainties to estimate structural capacity of ageing cast iron water mains - a possibilistic approach." *Canadian Journal of Civil Engineering*, 33(8), 1050-1064.

Thode, H. C. (2002). *Testing for normality* (Vol. 164). CRC Press.

Watson, T.G., Christian, C.D., Mason, A.J., Smith, M.H., and Meyer, R. (2004). "Bayesian-based pipe failure model". *Journal of Hydroinformatics*, 06(4), 259-264.