Application of genetic programming to modeling pipe failures in water distribution systems

Qiang Xu, Qiuwen Chen and Weifeng Li

ABSTRACT

The water loss from a water distribution system is a serious problem for many cities, which incurs enormous economic and social loss. However, the economic and human resource costs to exactly locate the leakage are extraordinarily high. Thus, reliable and robust pipe failure models are demanded to assess a pipe's propensity to fail. Beijing City was selected as the case study area and the pipe failure data for 19 years (1987–2005) were analyzed. Three different kinds of methods were applied to build pipe failure models. First, a statistical model was built, which discovered that the ages of leakage pipes followed the Weibull distribution. Then, two other models were developed using genetic programming (GP) with different data pre-processing strategies. The three models were compared thereafter and the best model was applied to assess the criticality of all the pipe segments of the entire water supply network in Beijing City based on GIS data.

Key words | GP, pipe failure model, water distribution systems

Qiang Xu
Qiuwen Chen* (corresponding author)
Weifeng Li
Research Centre for Eco-Environmental Sciences,
Chinese Academy of Sciences,
Shuangqing Road 18,
Haidian District,
Beijing 100085,
China
E-mail: *qchen@rcees.ac.cn*
*Also at: Three Gorges University
Daxue Road 8, Yichang 443002
China

INTRODUCTION

Water distribution system represents the arteries of a city, in which pipe failures are a regular occurrence when the residual strength of a deteriorated pipe becomes inadequate to resist the load on it (Skipworth 2002). Pipe failure incurs large economic and social costs both directly and indirectly, such as water and energy lost, repair cost, traffic delays, factory production lost due to inadequate water or service interruptions. Plenty of investments have been made to discover where a break happens.

According to interviews with staff in the water industries, the most commonly used approach for locating breaks is to detect the sound of water leaking using a pole connected to a pipe or to capture the acoustic signal using special devices like a leak noise correlator. Experienced workers can tell where a break happens according to the sound. Alternatively, the leak noise correlator can compute the location of a break according to the captured signal. However, this method is inefficient and labor-intensive. Therefore, reliable and robust pipe failure models are necessary to assess a pipe's propensity to fail, which can assist in positioning breaks (Berardi *et al.* 2008).

A literature survey indicated that a number of research projects have been recently undertaken with the goal of studying pipe break principles and developing predictive models (Davis *et al.* 2007; Berardi *et al.* 2008; Yamijala *et al.* 2009). The methods can be classified into three categories in general: physically based approachs, statistical approachs and data mining approachs.

Physically-based approaches (Rajani & Kleiner 2001; Davis *et al.* 2007) aim at discovering the physical mechanisms underlying pipe failure. But it will take a long time to observe the process, and may be costly. However, the statistical methods can be applied when the input data quality varies and the data quantity is limited (Kleiner & Rajani 2001). For water distribution pipes, statistical models provide a cost-effective means of analysis (Berardi *et al.* 2008).

doi: 10.2166/hydro.2010.189

There are two different statistical models: deterministic and probabilistic models. Kettler & Goulter (1985) developed a time-linear model. Besides the time-linear models, time-exponential models were developed as well (Shamir & Howard, 1979) and were improved by Walski & Pelliccia (1982).

The use of probabilistic models allows for formal measurement of the uncertainty of an individual pipe's failure rate, even though it requires the elicitation of expert knowledge (Berardi *et al.* 2008). The proportional hazards model was first used by Marks & Jeffrey (1985) to predict the breaks by computing the probability of the time duration between consecutive breaks. Inspired by survival analysis, Mailhot *et al.* (2000) developed a rigorous approach to estimate the parameters of statistical models using brief recorded data. The approach was later applied to three municipal water infrastructure systems by Pelletier *et al.* (2003). Some other models were also employed, such as a time-dependent Poisson model (Constantine *et al.* 1996), an accelerated lifetime model (Lei & Saegrov 1998; Le Gat & Eisenbeis 2000), a Bayesian diagnostic model (Watson *et al.* 2004), a logistic generalized linear model (Yamijala *et al.* 2009) and a decision tree method (Chen *et al.* 2008). Despite different variables being considered, all of these models aim to describe pipe break rates by a unique expression in which all pipes share the same explanatory variables. It must be noted that, in order to obtain statistical significance, pipes often need to be aggregated into homogeneous groups, so that effective analysis can be conducted (Shamir & Howard 1979; Kleiner & Rajani 1999).

Recently, data mining techniques such as genetic programming (GP) were introduced to study the resistance coefficient of the water conveyance system (Giustolisi 2004) and to discover the patterns in pipe failure datasets (Babovic *et al.* 2002). The employment of such techniques is required due to the complexity of pipe failure processes. Giustolisi & Savic (2006) proposed a novel hybrid data-driven technique, Evolutionary Polynomial Regression (EPR), which was used by Savic *et al.* (2006) and Berardi *et al.* (2008) to model pipe failures in water distribution systems. Parsimonious symbolic formulae were returned by EPR with higher accuracy in describing failure occurrence in homogeneous pipe groups than in the previous statistical models. This technique was further improved by Savic *et al.* (2009) and Giustolisi & Savic (2009).

This research took Beijing City as an example to develop pipe failure models by applying the GP technique. The pipe properties data, including the age, diameter, length and 19 years of breakage records (1987–2005), of the water distribution system were collected and used. The developed models were verified by part of the collected data. Finally, the models' performances were compared and evaluated.

GENETIC PROGRAMMING (GP)

Pioneered by Koza (1992), genetic programming is an evolutionary algorithm-based methodology which is used to find computer programs that perform a user-defined task. It has been applied successfully to broad fields such as automatic design, pattern recognition, data mining, robotic control, synthesis of artificial neural architectures, bioinformatics, music and picture generation (Langdon & Poli 2002). Compared to black-box data mining methods, GP provides the potential to gain insight into the relationships between the variables.

In this study, the genetic programming was implemented in the C++ language by the authors. The rank selection method was used to choose individuals for genetic operations, which were crossover and mutation. The crossover rate was set to be 0.5, meaning the first half of individuals according to their fitness rank were selected to spawn offspring. The mutation rate was 0.001, indicating that 0.1% of the nodes in an individual would be altered. The goodness-of-fit is evaluated using the coefficient of determination (CoD):

$$\mathrm{CoD} = 1 - \frac{\sum n(\hat{y} - y_{obs})^2}{\sum_n (y_{obs} - \bar{y}_{obs})^2} = 1 - \frac{\mathrm{SSE}}{\sum_n (y_{obs} - \bar{y}_{obs})^2} \qquad (1)$$

where n is the number of samples, $\hat{y}$ is the value predicted by the model, y_{obs} is the observed value, $\bar{y}_{obs}$ is the average of the observed values and SSE is the sum of the squared errors.

It is important that, in a linear regression model, the value of CoD should be between 0 and 1. However, GP is not a strict regression model and the value of CoD may become negative because the value of SSE can be unrestrictedly large. The formulae with smaller SSE values, or CoD values closer to 1, are believed to fit the observed data better. During the iteration, the formulae that have higher CoD values are

preserved and carried out the genetic operators, while the others are eliminated.

CASE STUDY

Data collection

The water distribution system of Beijing City was selected as the case study. Two datasets were collected, which were pipe physical property data based on ArcGIS and the pipe breakage records during 1987–2005. The pipe properties data included pipe diameter, material, year laid, length and user information. The pipe breakage records contained information on pipe diameter, material, failure time, year laid, break causes and a vague address. The common attributes that can link the two datasets together are the diameter and year laid. These then directly determined the analysis level (group/segment level) because breakage records on the pipe segment level were unavailable. There were many reasons that caused pipe breakage according to the records. This study focused on the ageing-induced deterioration process.

The pipe property data showed that more than 70% of the pipes were made from cast iron and ductile iron. The breakage records indicated that 91.1% of the failed pipes were made from cast iron. Therefore, only the cast iron pipes were investigated. Basic statistical information on cast iron pipes buried before 2005 and the corresponding breaks during 1987–2005 are shown in Table 1. The considered factors included diameter, length and age that may directly affected the pipe performance. Other important factors that can cause the spatial difference in pipe failure rates, such as earth density and traffic loads, have not been included due to data unavailability.

Table 1 | Features of cast iron pipes

Features	Values
Years laid	From 1901 to 2005
Diameter/mm	From 75 to 600
Total length/km	3322.5
Number of pipe segments	313,804
Number of deterioration breaks	566

Age-dependent Weibull model

The Weibull distribution is often used in life data analysis, whose three-parameter probability density function is given as

$$f(\alpha) = \frac{\beta}{\eta}\left(\frac{\alpha-\gamma}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{\alpha-\gamma}{\eta}\right)^{\beta}\right] \quad (2)$$

where β is the shape parameter, η is the scale parameter and γ is the location parameter. In such models, age is considered to be the most important factor influencing the pipe break. Thus the pipes are aggregated into groups by age. To discover when a type of pipe is more prone to fail, one way is to track a pipe's performance for its whole lifetime. This approach is obviously time-consuming. The other way is to analyze a group of pipes with different ages and explore their age-dependent behavior.

Chen *et al.* (2008) analyzed the pipe ages when breaks happened and found that the ages of the failed pipes followed the Weibull distribution. The pipes were first grouped by age, and then the total length and the break density (averaged number of breaks per unit length) of each age group were calculated as follows:

$$L_i = \sum_p l_{i,p} \qquad \lambda_a = \frac{\sum_i B_{i,a}}{\sum_i L_i} \quad (3)$$

where i is the pipe age at the end of the observation period (in the following, age i always means the age at the end of the observation period), a is the age when a break occurs, p is the pipe index, l_i,p is the length of the pth pipe of age i, L_i is the total length of age i, $B_{i,a}$ is the number of breaks in pipes whose age at the end of the observation period is i and whose age when the break occurred was a, and $\lambda_a$is the corresponding break density.

In this study, the pipe break density was treated as the pipe age's occurrence frequency. After normalization, the parameters β, η and γ of the Weibull probability density function were estimated to be 1.944, 28.56 and 0, respectively. Figure 1 showed the normalized frequency and the probability density function curve. It was observed from the results that a break is most likely to take place when the pipe age is about 20 years.

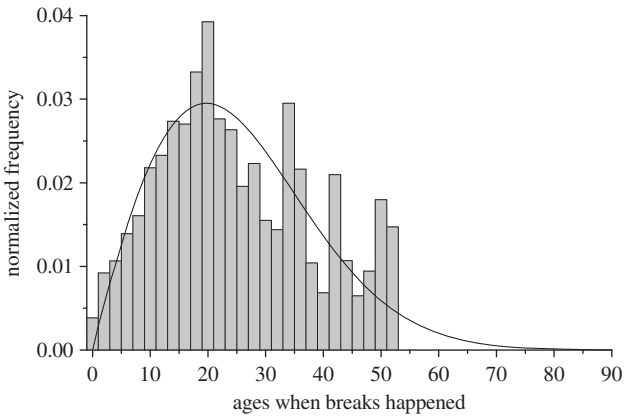

Figure 1 | Age distribution of pipes that have breakage record.

GP models based on diameter aggregation

Data preprocessing

As mentioned in the introduction, the pipes should be aggregated into groups to obtain statistically significant results because pipes of the same attributes (i.e. diameter, age, material) should share the same pipe failure rates (Shamir & Howard 1979). In this study, the pipes were grouped according to their diameter.

First of all, all the data were divided into two parts by randomly selecting pipes out of the whole database. The first part contained 60% of the total data, which was used for model development. The remaining 40% of the data were used for model validation. Then, the first part of the data was grouped by diameter, resulting in nine groups, given in Table 2, where D was the diameter, $\overline{A}$ was the length-averaged age

Table 2 | Results of the data grouped by diameter

D [mm]	L [km]	$\overline{A}$ [yr]	B [times]
75	230.2	28.3	144
100	709.8	13.7	146
150	228.6	13.1	26
200	382.7	11.6	22
250	25.4	36.7	1
300	250.9	17.5	14
400	483.3	11.2	15
500	3.9	27.4	0
600	400.2	12.1	4

similar to Berardi *et al.* (2008), L was the total length of the pipes of a group and B is the total failure of a group between 1987 and 2005. The length-averaged age is given by

$$\overline{A} = \frac{\sum(A_p \cdot l_p)}{\sum l_p} \tag{4}$$

where A_p and l_p were the age at the end of the observation period and the length of the pth pipe in a group. Note that all the ages used in this paper were the ages at 2005.

In the next section, symbolic formulae that mapped the relationship between the three independent variables (D, $\overline{A}$, L) and the dependent variable B were developed by using GP.

Formulae returned by GP

The independent variables were [D, $\overline{A}$, L], the dependent variable was [B] and the function set was [+, *, /, abs(), exp()]. The fitness of an individual formula was evaluated by CoD. When running, the GP frame generated and evaluated the formulae consisting of the variables and functions, and finally the formula that had the highest fitness was returned. Due to the intrinsic algorithm of GP, the returned formula was usually not the same for each run. Six formulas were listed in Table 3. After the formula was obtained from GP, it was validated by the remaining 40% of the data, which were grouped in the same manner as the model construction data.

Taking both the goodness of fit and the parsimony of the formulae into consideration, the third equation was finally selected:

$$B = \frac{\overline{A}(L + 5.198)}{D - 28.147}. \tag{5}$$

Figures 2 and 3 showed how the model fits the observed model development data and model validation data. respectively. The CoD values were 0.994 and 0.951, respectively.

Criticality assignment to individual pipe segments

The obtained GP models were based on the data grouped by diameter: however, what the utility managers were really concerned about was the propensity to fail of an individual

Table 3 | Formulae returned by GP and the corresponding CoD

Formulae	CoD
1) $B = 1.51\frac{\overline{A}L}{D}$	0.982
2) $B = \frac{\overline{A}L}{D}(1 + \frac{1}{\sqrt{D}-7.108})$	0.990
3) $B = \frac{\overline{A}(L+5.198)}{D-28.147}$	0.994
4) $B = 1.741\frac{\overline{A}L}{D}\exp(-8.112\frac{D}{\overline{A}L})$	0.994
5) $B = \frac{\overline{A}L}{D-\frac{L}{2.71\overline{A}}-\overline{A}}$	0.998
6) $B = \frac{\overline{A}L}{D} + \frac{A^2L}{D^2} + \frac{\overline{A}L^2}{D^3} + 17.9\frac{\overline{A}L}{D^2} + 17.9\frac{L^2}{D^3}$	0.998

pipe segment. Therefore, a pipe failure model must be able to assess the criticality of an individual pipe segment.

In the developed GP model, pipe diameter D was selected as the grouping criteria, that is to say, within a group the diameters of all the pipes were the same. If breaks of unit length were considered or the breaks were presumed to distribute according to the length of each pipe, the only difference lay in the pipe's age. Therefore, by combining with the time-dependent Weibull model, the GP model based on diameter aggregation could be used to calculate the break density of an individual pipe segment:

$$\lambda_{j,a} = \frac{B_j}{TL_j} \cdot f(a) \tag{6}$$

where Bj is the breakages of group j calculated by Equation (5), T is the observation duration, which is 19 in this case, L_j is the total pipe length of group j and $f(a)$ is the contribution of pipes at age a calculated by Equation (2).

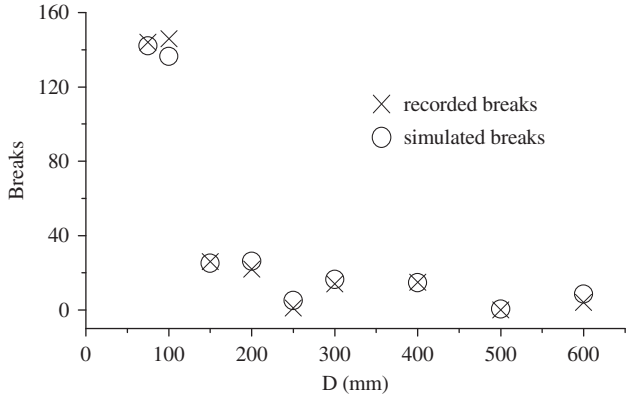

Figure 2 | Fitting for model development data.

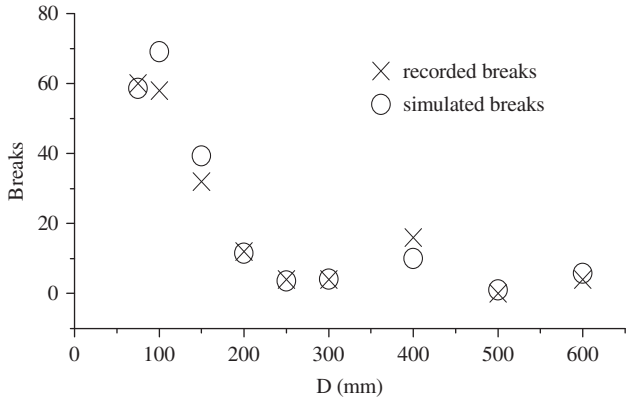

Figure 3 | Fitting for model validation data.

GP models based on aggregation by diameter and age

Data preprocessing

A refined grouping criterion was applied in this subsection. All the pipes were grouped by their diameter and age at the end of the observation period, which resulted in 501 groups. In each group, all the pipes had the same diameter D and pipe age A, and the total pipe length L was obtained by summing up all the pipes' lengths. 300 groups were used for model development and the remaining 201 groups were used for model validation.

Formula returned by GP

Symbolic formulae were obtained by applying the GP approach to the data. The candidate functions were [+, *, /, abs(), exp()]. The independent variable set and dependent variable set were [D, A, L] and [B], respectively. After running the GP program repeatedly with different formulae complexity constraints and different population sizes, many formulae were returned from the GP program, and the following one was selected as the best:

$$B = \left(\frac{40.47 - A}{L + 6.616} + 1\right) \cdot \frac{AL}{D} \tag{7}$$

where B was the total number of breaks of a group during the investigation period, and A, L and D were the age, length and diameter of a group.

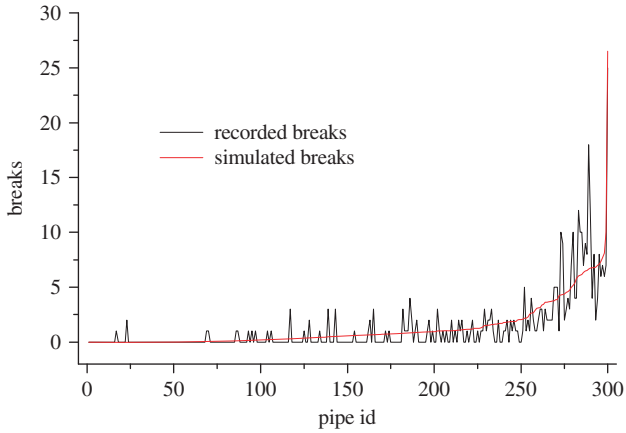

Figure 4 | Fitting for model development data.

To eliminate the negative returned values, the formula was then modified to

$$B = \max\left(0, \left(\frac{40.47 - A}{L + 6.616} + 1\right) \cdot \frac{AL}{D}\right). \quad (8)$$

Figures 4 and 5 showed how Equation (8) fits the observed model development data and model validation data. The corresponding CoD values were 0.741 and 0.657, respectively. To display the difference between the recorded data and the modelled data in a convenient way, the data pairs were sorted in ascending order. Otherwise, they would be hard to distinguish because of the fluctuations. The horizontal axis showed the sorted order of a pipe group.

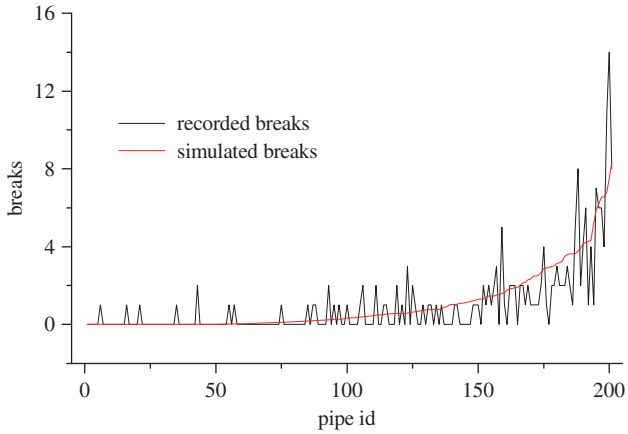

Figure 5 | Fitting for model validation data.

Assuming that breaks are uniformly distributed with the pipe length, the break density of a pipe group would be calculated by

$$\lambda_j = \frac{B_j}{TL_j} \quad (9)$$

where B_j is the breaks of group j calculated by Equation (8), T is the observation duration and L_j is the total pipe length of group j.

Tradeoff between model performance and formulae complexity

In a GP program, model performance usually increases with formulae complexity. However, the rate of increase becomes smaller and smaller when the formulae complexity keeps growing. Another important parameter in the GP program is population size which determines the possibility of finding the optimal solution within a specific search scope. The bigger the population size, the higher possibility of finding the optimal solution. However, the more time the program consumes. Therefore, minimum formulae complexity and population size were investigated in this subsection.

In this study, the best solutions for different population sizes (1000, 2000, 4000, 6000 and 8000) and different formulae complexity (described by tree size, or node number, from 5–50 in intervals of 5) were compared. For each combination of population size and tree size, the program was repeated for 30 runs and the average values of the CoD were shown in Figure 6. The improvement in performance (i.e. CoD) became slow when the tree size reached 30. The population size of 6000 (the cross mark) was seen as optimal because of its high CoD and relatively small population size. Thus, the optimal combination of GP parameters for this case (extracting the equation from 300 items of data) is a tree size of 30 and a population size of 6000.

DISCUSSION

The Weibull model

In this model, only pipe age at the time when the break occurred was considered as the explanatory variable of pipe

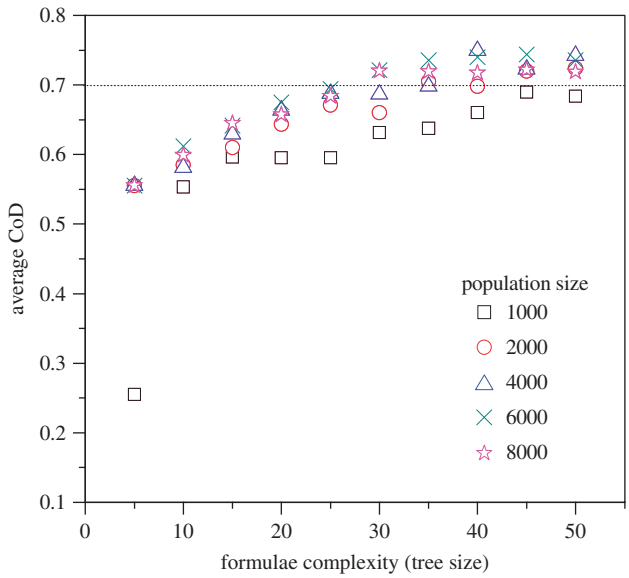

Figure 6 | CoD values with different population and tree sizes.

breakage. To a certain extent, this was reasonable, because deterioration was an ageing-inducing process. The pipes with the same diameter should share the same time-dependent model, and all the pipes should have the same model structure. Differences should exist in the model parameters from diameter to diameter. It was more appropriate to explore a particular Weibull probability density function for each group. However, the data quantity varied from group to group. For example, the number of records for groups of $D = 250$, 500 and 600 were only 8, 0 and 8 during the period 1987–2005. Obviously, the sample sizes were too small to fit a statistically significant curve. As a result, a single Weibull probability density function was used here for all the groups, which was not an optimal choice.

The GP models based on diameter aggregation

The pipe length, diameter and age were selected to build the model. The other influencing factors such as water pressure, traffic loadings and so forth were not included. Although it is known that relevant information before a leakage took place is required to build a leakage model, the water pressure and traffic loadings were not recorded with break development.

The aggregated GP formulae could well explain the pipe breaks by the pipe properties for different groups. The CoD values of the model were fairly high because the data were so highly aggregated that variances between the groups were large and the number of groups was small (see Table 2). The CoD value of model validation was smaller than that of model development. The reason could be overfitting because the RMSE (root mean square error) increased from 4.2 for model development to 4.9 for model validation.

The GP models based on aggregation by diameter and age

In this model, the pipes with the same diameter and the same age were aggregated into one group. The entire pipe network was categorized into 501 groups. 300 groups were applied for model development and the others were used for model validation. The obtained model was more efficient than the previous GP-based model. The possible reason is that the information on age influence was better incorporated.

The CoD values decreased from 0.741 for model development to 0.657 for model validation. When investigating the RMSE values, a decrease from 1.4 to 1.1 was found. Therefore, the decrease in CoD value was not because of overfitting. The main reason is that the variance of the sub-group dataset became small.

The form of the equations

It was noticed that, when the tree size was limited to a small value (e.g. 5 nodes), the returned formulae would easily and consistently converge to $B = ALD^{-1}$ or its product of a constant (such as the first equation in Table 3). As the tree size increased, the independent variables (i.e. A, L and D) were adjusted (e.g. the third equation in Table 3) or the constant was replaced by a term (like the second and fourth equations in Table 3). Therefore, in this study, the formulae can be divided into two parts, the core term ALD^{-1} and the adjustment term. The core term could usually be obtained in each running, while the adjustment terms were usually different. The core term ALD^{-1} could represent the base physical behavior of the phenomenon, while the adjustment terms should not be interpreted from a physical standpoint but served only to improve fitting on the training data.

From the structure of the equations, the pipe breaks were not simply related to the pipe length in a linear way. If divided by the pipe length L, most equations were still a function of L, which meant the break density was related to pipe length. However, the relationship was not consistent among the

generated equations (Table 3). It can monotonically increase (the sixth equation in Table 3) or decrease (the third equation in Table 3) or be consistent (the first and second equations) with pipe length. In particular, the break density calculated by Equation (9) was not a monotonic function of pipe length. In this study, an assumption that the break density of a group was constant was made. It was a simple way to convert the break density from a pipe group level to a pipe segment level, which seemed not to be an optimal one from the results. If data of higher resolution were available, the influence of pipe length on the break density on a segment level would be worth investigating.

Evaluation of the models

In this subsection, the three models were compared, and the advantages of each model were analyzed.

The break densities of the entire pipe network were calculated by the three models (i.e. age-dependent Weibull model, GP model based on diameter aggregation and GP model based on aggregation by diameter and age). The pipes were then sorted according to the break densities in descending order. Thus, three ordered columns were achieved. For comparison, two other columns were introduced that included the ordered data of the recorded break density and randomly ordered data. The relationships between the cumulative proportion of recorded breaks and the cumulative proportion of pipe lengths were investigated, as shown in Figure 7, where GP1 represented the GP model based on diameter aggregation and GP2 was for the GP model based on aggregation by diameter and age. Five curves were obtained from the five ordered columns. It was seen from curves 1–4 that, at the beginning, the proportion of breaks increased rapidly as the proportion of length increased. When the proportion of length kept on increasing, the increase of the proportion of breaks became slower and slower. This meant that, in general, the pipes that were more prone to breaks were sorted to the front. This result was very useful for the water plants because different monitoring priorities can be assigned to the pipes according to this order, which can apparently improve the breakage detection efficiency, especially when labor and device resources are limited.

As illustrated in Figure 7, supposing 30% of the entire pipes were checked for breaks; without the help of models the detection would be random (shown as curve 5) and as a result only about 30% of breaks could be found. However, if the time-dependent Weibull model (curve 4) was employed, the number of breaks that can be found would increase to about 56%. The number would soar to nearly 80% if the combination (curve 2) of GP1 and the Weibull model was used or the GP2 model (curve 3) was used. We were confident to say that the two GP models can save labor and device resources by a large margin. From the perspective of a future pipe renovation plan, using a random replacement of 30% of the pipeline may result in about 30% reduction of breaks. However, replacing the same length of pipes using a GP model would potentially eliminate nearly 80% of the pipe failures.

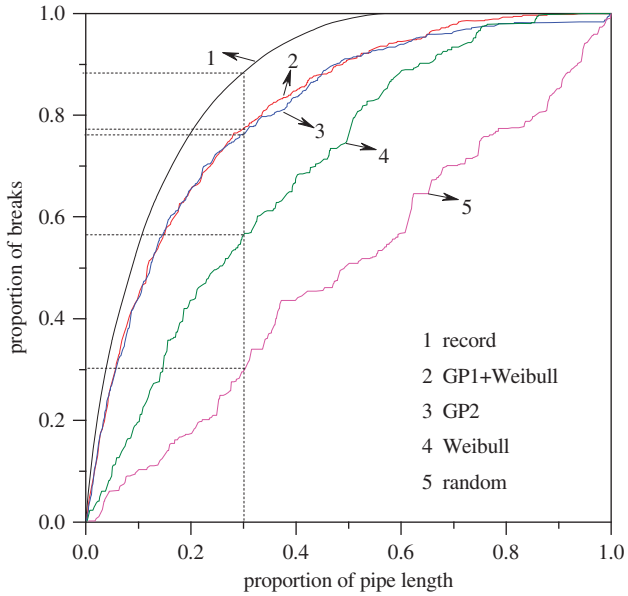

Figure 7 | Efficiencies of the different models.

It was also observed from Figure 7 that curves 2 and 3 were very close, indicating that the performance of the two GP models were almost equivalent, although the model structures and the fitting CoD values were very different. However, Equation (9) was more understandable and parsimonious than Equation (6).

Application of the models for prediction

The developed models could be used to predict break densities of pipes in the future so as to support a monitoring scheme plan and maintenance plan in a similar way as reported in Berardi *et al.* (2008). It was seen from the models

that, for a specific group of pipes, only the variable A changes with time, while the others (i.e. D and L) remain constant. Taking the GP2 model as an example, the break density of a specific group j in a t-year prediction time ($\lambda_{j,t}$) can be expressed as follows:

$$\lambda_{j,t} = \frac{1}{TL_j} \cdot \max\left(0, \left(\frac{40.47 - (A_0 + t)}{L_j + 6.616} + 1\right) \cdot \frac{(A_0 + t)L_j}{D_j}\right) \quad (10)$$

where T is the observation duration, L_j is the total pipe length of group j, D_j is the pipe diameter of group j and A_0 is the age of the group at the end of the observation period.

Once the break densities are calculated, different priorities of break detection can be assigned to the pipes. It is reasonable that high priority should be given to the pipe segments with high predicted break density values.

It is important to notice that, as the prediction time t gets longer, the prediction accuracy of the break density will become lower. To overcome this drawback, data from newly detected breaks should be assimilated to update the model by reconstructing formulae similar to Equation (8).

CONCLUSION

Models to estimate pipe breaks were developed in this research by using genetic programming and statistical techniques. The water distribution system of Beijing City was selected as the case study area, and the pipe information data as well as failure records were collected. Three models were built in the study. The first one was the age-dependent Weibull model, which revealed the relationship between the breakage density and the pipe age. The other two models were built by using GP with two different data pre-processing strategies. In the first strategy, the pipes were aggregated into groups by diameter and then used to develop a pipe failure model. To assign criticality to an individual pipe segment, the age-dependent Weibull model was combined. In the second model, the pipes were aggregated into more refined groups by diameter and age, and then models were obtained through GP. To balance the formulae complexity and the simulation accuracy, the parameter configuration of the GP approach was discussed. According to the validation, both of the models from GP can well estimate the break numbers. The application of the models to guide the future planning on a break detection campaign was finally discussed.

The developed models were applied to the entire water distribution system of Beijing City. From the results, it was concluded that the model developed by GP and limited available data can apparently save labor and instrumentation by a large amount. In particular, the developed model can be applied to optimize the design of a breakage detecting system.

ACKNOWLEDGEMENT

The authors are grateful for the funding from the Ministry of Sciences and Technology of the People's Republic of China (2006BAB17B03) and the support of Chutian Scholarship (KJ2010B002).

REFERENCES

Babovic, V., Drécourt, J. P., Keijzer, M. & Friss Hansen, P. 2002 A data mining approach to modelling of water supply assets. *Urban Wat.* **4**(4), 401–414.

Berardi, L., Kapelan, Z., Giustolisi, O. & Savic, D. A. 2008 Development of pipe deterioration models for water distribution systems using EPR. *J. Hydroinf.* **10**(3), 113–126.

Chen, Q. W., Qu, J. H., Liu, R. P. & Li, W. F. 2008 Rule-based model for aging-induced leakage in water supply network of Beijing City. *China Wat. Waste Wat.* **24**(11), 52–56.

Constantine, A. G., Darroch, J. N. & Miller, R. 1996 Predicting Underground Pipe Failure. *Water (J. Australian Wat. Assoc.)* **23**(2), 9–10.

Davis, P., Burn, S., Moglia, M. & Gould, S. 2007 A physical probabilistic model to predict failure rates in buried PVC pipelines. *Reliab. Engng. Syst. Safety*. **92**(9), 1258–1266.

Giustolisi, O. 2004 Using genetic programming to determine Chezy resistance coefficient in corrugated channels. *J. Hydroinf.* **6**(3), 157–173.

Giustolisi, O. & Savic D. A. 2006 A symbolic data-driven technique based on evolutionary polynomial regression. *J. Hydroinf.* **8**(3), 207–222.

Giustolisi, O. & Savic, D. A. 2009 Advances in data-driven analyses and modelling using EPR-MOGA. *J. Hydroinf.* **11**(3), 225–236.

Kettler, A. J. & Goulter, I. C. 1985 An analysis of pipe breakage in urban water distribution networks. *Can. J. Civil Engng.* **12**(2), 286–293.

Kleiner, Y. & Rajani, B. 1999 Using limited data to assess future needs. *J. AWWA*. **91**(7), 47–61.

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

Koza, J. R. 1992 *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA.

Langdon, W. B. & Poli, R. 2002 *Foundations of Genetic Programming*. Springer, Berlin.

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

Lei, J. & Saegrov, S. 1998 Statistical approach for describing failures and lifetimes of water mains. *Wat. Sci. Technol.* **38**(6), 209–217.

Mailhot, A., Pelletier, G., Noel, J. F. & Villeneuve, J. P. 2000 Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: methodology and application. *Wat. Res. Res.* **36**(10), 3053–3062.

Marks, D. H. & Jeffrey, L. A. 1985 *Predicting Urban Water Distribution Maintenance Strategies: A Case Study of New Haven, Connecticut*. US Environmental Protection Agency, Washington, D.C., USA.

Pelletier, G., Mailhot, A. & Villeneuve, J. P. 2003 Modeling water pipe breaks – three case studies. *J. Wat. Res. Plann. Mngmnt.* **129**(2), 115–123.

Rajani, B. & Kleiner, Y. 2001 Comprehensive review of structural deterioration of water mains: physically based models. *Urban Wat.* **3**(3), 151–164.

Savic, D. A., Giustolisi, O., Berardi, L., Shephard, W., Djordjevic, S. & Saul, A. 2006 Modeling sewer failure using evolutionary computing. *Proc. ICE, Wat. Mngmnt.* **159**(2), 111–118.

Savic, D. A., Giustolisi, O. & Laucelli, D. 2009 Asset deterioration analysis using multi-utility data and multi-objective data mining. *J. Hydroinf.* **11**(3), 211–224.

Shamir, U. & Howard, C. 1979 An analytical approach to scheduling pipe replacement. *J. AWWA*. **71**(5), 248–258.

Skipworth, P. J. 2002 *Whole Life Costing for Water Distribution Network Management*. Thomas Telford, London.

Walski, T. M. & Pelliccia, A. 1982 Economic analysis of water main breaks. *J. AWWA*. **74**(3), 140–147.

Watson, T. G., Christian, C. D., Mason, A. J., Smith, M. H. & Meyer, R. 2004 Bayesian-based pipe failure model. *J. Hydroinf.* **6**(4), 259–264.

Yamijala, S., Guikema, S. D. & Brumbelow, K. 2009 Statistical models for the analysis of water distribution system pipe break data. *Reliab. Engng. Syst. Safety* **94**(2), 282–293.

First received 21 October 2009; accepted in revised form 16 March 2010. Available online 28 October 2010