Comparison of four models to rank failure likelihood of individual pipes

Yehuda Kleiner and Balvant Rajani

ABSTRACT

The use of statistical methods to discern patterns of historical breakage rates and use them to predict water main breaks has been widely documented. Particularly challenging is the prediction of breaks in individual pipes, due to the natural variations that exist in all the factors that affect their deterioration and subsequent failure. This paper describes alternative models developed into operational tools that can assist network owners and planners to identify individual mains for renewal in their water distribution networks. Four models were developed and compared: a heuristic model, a naïve Bayesian classification model, a model based on logistic regression and finally a probabilistic model based on the non-homogeneous Poisson process (NHPP). These models rank individual water mains in terms of their anticipated breakage frequency, while considering both static (e.g. pipe material, diameter, vintage, surrounding soil, etc.) and dynamic (e.g. climate, operations, cathodic protection, etc.) effects influencing pipe deterioration rates.

Key words | heuristics, logistic regression, naïve Bayesian classification, non-homogeneous Poisson process, pipe break forecast, ranking

Yehuda Kleiner (corresponding author)
Balvant Rajani
National Research Council Canada,
Ottawa,
Ontario,
Canada
E-mail: *Yehuda.Kleiner@nrc-cnrc.gc.ca*

INTRODUCTION

The use of statistical methods to discern patterns of historical breakage rates and use them to predict water main breaks has been widely documented. Kleiner & Rajani (2001) provided a comprehensive review of approaches and methods that had been developed prior to their review. Since then, several more methods have been proposed, such as those by Dandy & Engelhardt (2001), Park & Loganathan (2002), Jarrett *et al.* (2003), Mailhot *et al.* (2003), Watson *et al.* (2004), Dridi *et al.* (2005), Giustolisi *et al.* (2005), Boxall *et al.* (2007), Giustolisi & Berardi (2007), Economou *et al.* (2008) and Le Gat (2008) to name but a few. In all these methods, few, if any, heuristic methods have been documented to discern patterns of historical breakage rates.

Many factors, operational, environmental and pipe-intrinsic factors, jointly affect the breakage rate of a water main. While not all pipes are created equal (even pipes of the same material and size), it is normally assumed that pipes that share specific intrinsic properties, such as material, diameter, vintage, etc., can be expected to have the same breakage pattern, all else being equal. However, non-pipe-intrinsic factors may have varying effects on the breakage patterns of different pipes, even if all else is equal. For example, two pipes of the same material, diameter, age, etc. can be impacted differently by climate. We may never have enough data to account for these differences due to variability. At the same time, it is unreasonable to perform a statistical analysis on the breakage pattern of a single pipe because sufficient data on breaks to conduct a credible analysis are not available. For these reasons, the forecasting of breaks in individual water mains has proven to be quite a challenge.

This paper describes four specific models intended to serve as operational tools for water distribution network owners and planners to help them rank individual water mains for renewal, while considering both static (e.g. pipe material, diameter, vintage, surrounding soil, etc.) and dynamic (e.g. climate, operations, cathodic protection, etc.)

doi: 10.2166/hydro.2011.029

effects influencing pipe deterioration rates. It should be noted that risk-based prioritization of water mains for renewal entails the quantification of both likelihood and consequences of failure. However, this research focused only on failure likelihood, and therefore in this paper the word 'ranking' refers strictly to failure likelihood-based ranking.

Four models were developed and compared: a heuristic model named 'ordered lists' (OL), comprising procedures that reflect intuition and observations, a model based on naïve Bayesian classification (NBC), one based on logistic regression and finally a probabilistic model based on the non-homogeneous Poisson process (NHPP). The first three are the so-called 'ranking' models, intended for forecasting, within a homogeneous group, the ranking of individual pipes in terms of their relative failure likelihood, rather than attempting to forecast the actual number of breaks. The rationale was that the existing statistical/empirical models (such as D-WARP – see Kleiner & Rajani 2004) forecast the aggregate breakage rate for a homogeneous group of pipes; the ranking models would be useful to drill down to the individual pipe level within the group (note that D-WARP considers both static and dynamic effects impacting pipe breakage rate). The fourth model is different from the first three in that it actually endeavours to forecast future breakage rates of individual pipes, rather than just rank them on relative failure likelihood. The ranking models were set out to address the following challenge: 'In a homogeneous group comprising N individual pipes, with available breakage history of T years, find which n pipes are expected to have the highest number of breaks in the next y years.'

The rest of this paper is organized as follows. We first describe the covariates that were examined and used for the ranking models. Some of these covariates are new in the context of pipe breakage analysis and are based on practical observations and heuristic arguments. Subsequently, the three ranking models are described with details on how these and other covariates were used to rank individual mains. In the interest of good readability, some of the lengthy formulations are provided in Appendices A, B and C rather than in the main text (available online at http://www.iwaponline.com/jh/014/029.pdf). The last model introduced is based on NHPP. The comparison between the four models is presented next by way of an example dataset, followed by some concluding comments.

Covariates for the ranking models

Pipe length. In all reported analyses of breakage frequency in pipe groups, aggregate length of a group has been used as a normalizing factor (e.g. Shamir & Howard 1979; Clark *et al*. 1982; Walski & Pelliccia 1982; Kettler & Goulter 1985; Kleiner & Rajani 2004 and others). This has the implication that breaks are distributed uniformly along the pipes, which carries the expectation that the number of breaks is directly proportional to the length of pipes. This implication has been questioned by others, e.g. Goulter & Kazemi (1988), Goulter *et al*. (1993), Jacobs & Karney (1994) and Mavin (1996).

The literature reflects that pipe length has frequently been used as a covariate to 'explain' at least some of the variability observed in individual water mains. Andreou *et al*. (1987), Eisenbeis *et al*. (1999), Le Gat & Eisenbeis (2000) and Røstum (2000) and others used the log of pipe length as a covariate in their proportional hazards-based models (PHM).

The researchers cited above reported various outcomes with respect to the 'quality' of pipe length as a covariate. In some water distribution networks, length was found to be statistically significant, while in others it was not. Additionally, in some pipe materials (cast iron (CI), PVC) length was found to be significant when the number of previous breaks was between 1 and 3, while insignificant when the number of previous breaks was zero or equal to or greater than 4. In other materials (asbestos cement (AC)) it was found to be not significant at any level of previous breaks. In some cases the square root of the pipe length was found to be a significant covariate.

The dichotomy of using length as a normalizing factor, as well as the inconclusive results using length as covariate, triggered an investigation, details of which are reported by Kleiner & Rajani (2010). They conclude that pipe length is a surrogate for pipe exposure and higher exposure necessarily leads to more breaks. However, the natural randomness inherent in the relationship between length and breaks is relatively high. Further, pipe break is a discrete entity while pipe length is a continuous physical property.

Individual pipes, whose length might typically vary between a few tens and a few hundreds of meters, typically do not experience too many breaks before they are replaced. This discrete nature of breakage data amplifies the natural randomness in relatively short pipes. Consequently, the randomness or the 'noise' in the data often overwhelms any mathematical relationship that may exist between length and observed breakage rate. However, when aggregated pipe lengths are examined, the aggregate number of breaks becomes continuous-like in its behaviour and the natural randomness produces 'noise' that is much smaller relative to the mathematical relationship and therefore no longer overwhelms this relationship. Furthermore, it appears that the relationship between breakage rates and length of individual water mains can be better characterized as non-parametric but monotone. This was revealed by rank correlation analyses, which consistently yielded better results than linear correlation.

Breakage history: number of known previous failures (NOKPF). The essence of all statistical/empirical pipe failure prediction models is to use failure history to discern failure patterns. Therefore, these models always use the number of previous failures (*NOPF*) either explicitly or implicitly. In PHM it can be used as a covariate or as a stratification criterion (e.g. Andreou *et al.* 1987). In the past, *NOPF* has not been used explicitly as a covariate in NHPP models. Gustafson & Clancy (1999) and Mailhot *et al.* (2003) used break order (e.g. first break, second break since installation, etc.) as a stratification criterion of sorts, upon which distribution parameters of time to failure are dependent.

It should be recognized that, in reality, most water utilities will have left-truncated datasets because they do not have pipe failure data that cover the entire history of pipes since installation (except for recently installed pipes). For this reason, models that use break order as a parameter will find little applicability. Also for this reason we chose to name this candidate covariate *number of known previous failures* (*NOKPF*) as opposed to the more commonly used *number of previous failures* (*NOPF*).

Breakage history: Recency. *Recency* expresses the tendency of historical breaks to have occurred more towards the beginning or towards the end of the observation period. *Recency* is defined with the use of Figure 1. T is

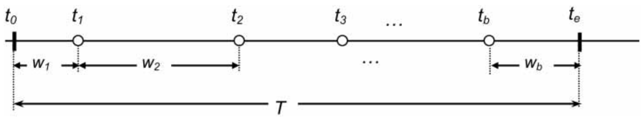

Figure 1 | Timeline for definitions of *Recency* and *Scatter*.

the observation period, which starts at t_0 and ends at t_e. t_1 denotes the time (since t_0) of the first break recorded for this pipe, t_2 is the time of the second break, and so forth. The pipe in Figure 1 experienced a total of b breaks during period T. The *Recency* (Re_i) of breakage pattern of pipe i during a historical observation period T is defined as follows:

$$Re_i = \frac{1}{b_i}\sum_{j=1}^{b_i}\frac{t_{ij}}{T} = \frac{1}{b_i T}\sum_{j=1}^{b_i} t_{ij}; \quad b_i > 0 \tag{1}$$

where b_i is the total number of breaks experienced by pipe i during period T and t_{ij} is the year in which pipe i experienced its jth break (t_{ij} is measured from t_0). Re_i is defined as zero when $b_i = 0$ (no breaks in period T). It is clear that $0 \le Re_i \le 1$. Thus, lower values of *Recency* mean that the observed historical breaks tend to have occurred more towards the beginning of the observation period, whereas higher values indicate breaks have occurred more recently.

Breakage history: Scatter. There are several ways to represent how observed historical breaks are concentrated (or scattered) through the observation period. Two different formulations were examined, one based on the square deviation from the mean breakage year and the other based on Shannon's entropy. No consistent differences in model performance were observed between the two formulations and Shannon's entropy was selected for convenience. The *Scatter* (Sc_i) of breakage pattern of pipe i during a historical observation period T is defined as follows (Figure 1):

$$Sc_i = \frac{-1}{\mathrm{Ln}(b_i+1)}\sum_{j=1}^{b_i+1} w_{ij}\mathrm{Ln}(w_{ij}); \quad b_i > 0$$
$$\text{where} \quad w_{ij} = \frac{t_{ij} - t_{i,j-1}}{T} \quad \text{and} \quad w_{i,b_i+1} = t_e - t_{b_i} \tag{2}$$

It can be shown that when all breaks are evenly spaced along T then $Sc = 1$. Also when all breaks are concentrated at t_0 or at t_e then $Sc = 0$. When all points are concentrated exactly in the middle of historical observation period T, then $Sc = 2/\mathrm{Ln}(b+1)$.

Geographical clustering. Water utilities often lack data that are geographically related (directly or indirectly), such as soil data, overburden characteristics (land development, traffic patterns), historical installation practices, groundwater fluctuations, transient pressures, poor bedding, etc. These data, if available, may sometimes help 'explain' variations in breakage rates among individual water mains in a 'homogeneous' group of pipes. In the absence of such data, the proximity of a pipe to a cluster of historical breaks may serve as a useful surrogate.

The K-Means algorithm (MacQueen 1967) was used to create the *Cluster* covariates (or *Clusters*). Given n data points and K centroids (cluster centres), this algorithm assigns each data point to the nearest centroid, then recalculates and shifts the location of the centroid and again re-assigns each data point to the new location of the centroid, and so forth until equilibrium is reached and the centroids no longer shift their locations. The K-Means algorithm is capable of clustering multi-dimensional (or multi-attribute) data. The application to two-dimensional geographical data (only X and Y coordinates) is therefore relatively simple and fast.

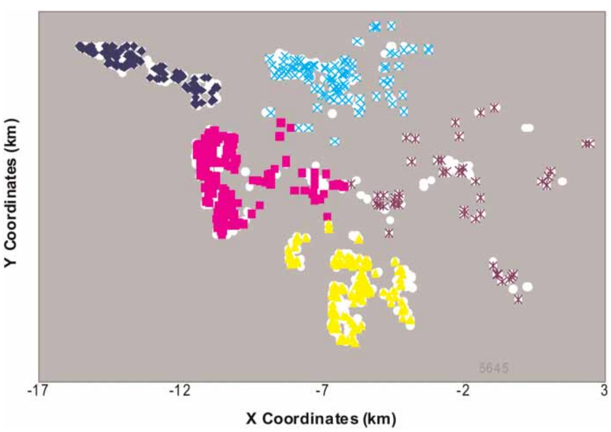

Figure 2 | LogReg example – Spatial distribution and clustering of pipes and breaks.

In anticipation of typical availability of such geographical data, it was deemed sufficiently accurate to assume that the location of a break is always at the centre of the pipe (which can be computed from the pipe-node coordinates). It should be noted that the K-Means clustering algorithm does not determine the optimal number of clusters or their approximate locations, rather the user needs to visually determine those based on observation of historical break locations as well as on prior knowledge.

Figure 2 illustrates an example of clusters. Breaks and pipes were arbitrarily (visually) divided into five clusters, represented by five different patterns. The light-coloured dots in the background represent the centres of all pipes in the group (obviously not all pipes experienced breaks in this period).

Ordered lists (OL) model

The OL model is essentially a weighted, non-parametric heuristic model, premised on the assumption that some monotone relationships exist between a set of covariates and anticipated breakage frequency in individual water mains. The model comprises two sub-procedures, model training (or calibration) and model validation. Model validation is conducted on a 'holdout' sample data.

Procedure for model training

(a) Partition the observation period T into training and validation periods, where the validation period comprises the latest v years in T (Figure 3).

(b) Select *RefYear* and partition the training period into *WinPast* and *WinFuture*. Identify n pipes with the

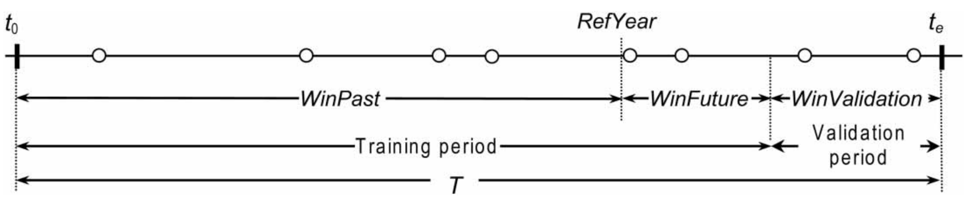

Figure 3 | Timeline for training and validation procedure.

highest number of breaks in period *WinFuture* and *WinValidation*. Record these pipes in lists named L_f and L_v, respectively.

(c) Create an ordered list L_b, containing all pipes sorted by the covariate *NOKPF* (number of known previous failures) observed during *WinPast* in descending order (pipe with highest past breaks ranks first, etc.).

(d) Similar to (c) above, create an ordered list L_l, containing all pipes sorted by length in descending order (longest pipe first, etc.).

(e) Similar to (c) above, create an ordered list L_R, containing all pipes sorted by *Recency* during *WinPast* in descending order (highest *Recency* first, etc.).

(f) Similar to (c) above, create an ordered list L_S, containing all pipes sorted by *Scatter* during *WinPast* in descending order (highest *Scatter* first, etc.).

(g) Assign weights W_b, W_l, W_R, W_S to lists L_b, L_l, L_R, L_S, respectively. Each list represents a covariate and each covariate is associated with a weight.

(h) For each pipe i in the group, compute a composite score Cs_i that combines the rank R_i^L of this pipe in each list and the corresponding weight:

$$Cs_i = f(R_i^{L_b}, R_i^{L_l}, R_i^{L_R}, R_i^{L_S}, W_b, W_l, W_R, W_S) \tag{3}$$

where $R_i^{L_b}$ represents the rank of pipe i in list L_b, etc., and $f(\cdot)$ is an aggregation function. Several different types of aggregation functions were tested as described later.

(i) Rank all pipes by their composite scores in ascending order (lower score corresponds to lower order rank, which means higher breakage potential). The n pipes with the lowest scores are those predicted by the model to have the highest number of breaks in the future window. Record these pipes in a list L_w.

(j) Compare list L_w to list L_f. If a certain pipe i appears in both lists (irrespective of location within a list) this occurrence is defined as a 'hit'. Determine the number of hits, H.

(k) Find a set of weights ($W_b^*, W_l^*, W_R^*, W_S^*$) that maximizes H. We then say that the model has been trained on the group of pipes that meets this condition. We used genetic algorithm (simple, single objective GA, using binary coded genes, crossover and mutation operators and roulette wheel-based selection mechanism) to maximize H.

Note that the *Cluster* covariate, which is a categorical covariate, was not used in the ordered lists model. It can be used, however, by partitioning the pipes in the group into subgroups corresponding to clusters and applying the model separately to each of these subgroups.

Procedure for model validation

For simplicity, we define *WinPastValidation* = *WinPast* + *WinFuture*. T is now partitioned into *WinPastValidation* and *WinValidation* (Figure 3).

(i) Repeat training steps (c) through g. to create new ordered lists – 'validation' lists L'_b, L'_l, L'_R, L'_S. For lists L'_b, L'_R, L'_S use the time period *WinPastValidation*. Note that L'_l, which refers to pipe length, is the same as L_l (pipe lengths do not change).

(ii) Apply the weights ($W_b^*, W_l^*, W_R^*, W_S^*$) to the validation lists and repeat training steps (i) and (j), where the n pipes with the lowest scores are recorded in list L'_w. The list L'_w contains pipes predicted by the model to have the highest number of breaks in the validation time window. Note that the observed breaks in the validation time period (holdout sample data) did not participate in the training of the model.

(iii) Determine H', the number of hits (pipes that appear in both lists) in lists L'_w and L_v.

(iv) Repeat the validation process for various n values where n is the number of pipes with the highest breakage rates – see training step (i). Use the method described in Appendix A to analyse results (see http://www.iwaponline.com/jh/014/029.pdf).

Aggregation functions. As the ordered lists model is a heuristic procedure, it is not possible to determine *a priori* what type of aggregation function will be best suited for a specific dataset. Consequently, seven different published aggregation functions were tested. These functions can be classified into three classes, 'Cardinal', 'Ordinal' and 'Mixed' weights. The Cardinal weights class comprises aggregation functions in which each weight is assigned specifically to a covariate (list), exactly as described in training step (h) above. In this class we included two specific functions, *CardIndepend* and *CardDepend*. The ordinal weights class comprises aggregation functions in which a weight is assigned to a covariate based on the relative ranking that this covariate has for a given pipe. In this class we included four functions: *OrdMaxMin*, *OrdMaxNext*, *OrdAll* and *OrdChoquet*. The Mixed weights class comprises one function *Mixed*, which involves a combination of cardinal and ordinal weights. A description of the aggregation functions, as well as their precise mathematical formulations, are provided in Appendix B (see http://www.iwaponline.com/jh/014/029.pdf).

Table 1 | Summary of ordered lists (OL) model example results (number of hits) – training

Aggregation function		Number of pipes recorded with at least 1 break 138	2 breaks 24	3 breaks 6	4 breaks 0
Cardinal/ independent	Max.	46	3	1	
	Min.	44	1	0	
	Mean	45.2	2.1	0.8	
Cardinal/with dependencies	Max.	48	5	1	
	Min.	44	0	0	
	Mean	45.2	1.9	0.4	
Ordinal/max. & min.	Max.	39	2	1	
	Min.	39	2	1	
	Mean	39.0	2.0	1.0	
Ordinal/max. & next	Max.	42	4	1	
	Min.	42	4	1	
	Mean	42.0	4.0	1.0	
Ordinal/all	Max.	43	4	1	
	Min.	42	2	0	
	Mean	42.3	3.9	0.4	
Ordinal/choquet	Max.	43	5	1	
	Min.	42	4	0	
	Mean	42.4	4.1	0.8	
Mixed weights	Max.	43	5	1	
	Min.	42	4	0	
	Mean	42.7	4.1	0.9	

Example. A homogeneous group of CI pipes from Calgary was analysed, including 1,091 pipes (146.6 km or 91 miles total length), 150 mm (6″) in diameter, installed between 1956 and 1960, and each pipe was at least 20 m long. Full year breakage data were available for the years 1961–2006. The OL model was trained on 40 years' failure data from 1962 to 2001, where *WinPast* was taken as 35 years (1962–1996), *WinFuture* was taken as 5 years (1997–2001) and *WinValidation* was taken as 5 years (2002–2006). *RefYear* is accordingly 1996. The weights obtained from training (calibration) were used to forecast the ranking of pipes in the validation period. The model was trained on several *n* values that were selected to correspond to the number of pipes in the group observed with a precise number of breaks. The results are summarized in Tables 1–4. Note that the results for the same aggregation function can vary through multiple runs (a 'run' refers to a training and validation session) because training is done using genetic algorithm, which is a search heuristic with random elements. Consequently, the model was applied 16 times with each aggregation function. The result tables provide the minimum, maximum and average number of training hits (in 16 runs), as well the minimum, maximum and average number of validation hits (note that in the training period a maximum of three breaks was observed in any given pipe, while in the validation periods up to four breaks were observed). Note further that *P*-values were used (Tables 2 and 4) to assess the ranking ability of the model for a given dataset. Appendix A describes in detail how these *P*-values are computed (see http://www.iwaponline.com/jh/014/029.pdf).

The application of the OL models on the example dataset shows that it could be trained to identify about one-third of all the pipes that experienced at least one break in the *WinFuture* period, as well as about 10 to 15% of the pipes that experienced at least two breaks and about 0 to 15% of

Table 2 | Summary of ordered lists (OL) model example results – training *P*-values (total 1,091 pipes in sample)

# pipes with ≥1 break	Hits	*P*-value	# pipes with ≥2 breaks	Hits	*P*-value	# pipes with ≥3 breaks	Hits	*P*-value
138	35	0.00001	24	3	0.014	6	1	0.033
	40	0.0		4	0.0014		2	0.0004
	45	0.0		5	0.0001			

Table 3 | Summary of ordered lists (OL) model example results (number of hits) – validation

Aggregation function		Number of pipes recorded with at least 1 break 170	2 breaks 30	3 breaks 6	4 breaks 2
Cardinal/ independent	Max.	54	7	2	0
	Min.	49	6	0	0
	Mean	51.4	6.4	1.1	0
Cardinal/with dependencies	Max.	55	7	1	0
	Min.	44	1	0	0
	Mean	49.3	4.2	0.4	0
Ordinal/max. & min.	Max.	52	9	1	1
	Min.	46	5	1	0
	Mean	47.3	5.7	1.0	0.9
Ordinal/max. & next	Max.	52	7	1	1
	Min.	52	7	1	1
	Mean	52.0	7.0	1.0	1.0
Ordinal/all	Max.	53	8	2	1
	Min.	45	5	0	0
	Mean	50.3	7.0	1.3	0.3
Ordinal/ choquet	Max.	55	8	2	1
	Min.	50	6	1	0
	Mean	52.0	7.2	1.8	0.6
Mixed weights	Max.	55	8	2	1
	Min.	50	7	1	0
	Mean	51.6	7.2	1.9	0.2

pipes with at least three breaks. It is also noted that in this dataset there were no significant differences between the training (calibration) success rates of the various aggregation functions.

In the validation session, the OL model only succeeds to identify fewer than one-third of all the pipes that experienced at least one break in the *WinValidation* period, as well as about 20 to 30% of the pipes that experienced at least two breaks, about 0 to 30% of pipes with at least three breaks and 0 to 50% of pipes with at least four breaks. As noted in the training session, no significant difference between the forecasting success rates of the various aggregation functions is seen in this dataset (this could differ in other datasets).

Naïve Bayesian classification (NBC) model

The naïve Bayesian classification method is based on Bayes' rule and uses attributes (or covariates or classifiers) to partition data into pre-defined classes. The underlying assumption in the NBC model is that these covariates are conditionally independent of one another (i.e. if the class is known then the covariates are statistically independent of each other) and ignores possible interactions between attributes (covariates). Denoting $X_1, X_2, \ldots, X_J$ as covariates and Y as the outcome (or class), which in our case is binary (0/1), Bayes' rule can be stated as:

$$\Pr(Y|X_1, X_2, \ldots, X_j) = \frac{\Pr(X_1, X_2, \ldots, X_j|Y)\Pr(Y)}{\Pr(X_1, X_2, \ldots, X_j)} \tag{4}$$

where the expression on the left denotes posterior probability, the numerator on the right is the product of likelihood and prior probability and the denominator denotes evidence. Usually only the numerator is of interest because the denominator does not depend on Y. Using the conditional independence assumption described above, it can be shown that the likelihood function becomes:

$$\Pr(X_1, X_2, \ldots, X_j|Y) = \prod_{k=1}^{j} \Pr(X_j|Y) \tag{5}$$

and therefore the conditional probability of outcome Y becomes:

$$\Pr(Y|X_1, X_2, \ldots, X_j) = \Pr(Y)\prod_{k=1}^{j} \Pr(X_j|Y) \tag{6}$$

Table 4 | Summary of ordered lists (OL) model example results – validation *P*-values (total 1,091 pipes in sample)

# pipes ≥1 break	Hits	*P*-value	# pipes ≥2 breaks	Hits	*P*-value	# pipes ≥3 breaks	Hits	*P*-value	# pipes ≥4 breaks	Hits	*P*-value
170	45	0.00004	30	6	0.0001	6	1	0.033	2	1	0.0037
	50	0.0		7	0.00001		2	0.0004		2	0.0
	55	0.0		8	0.0						

when the response (Y) is binary, as in our case, we can simplify by computing the so-called likelihood ratio (LR).

$$\begin{aligned} \mathrm{LR} &= \frac{\Pr(Y=1|X_1, X_2, \ldots, X_j)}{\Pr(Y=0|X_1, X_2, \ldots, X_j)} \\ &= \frac{\Pr(Y=1)\prod_{k=1}^{j}\Pr(X_j|Y=1)}{\Pr(Y=0)\prod_{k=1}^{j}\Pr(X_j|Y=0)} \\ &= \frac{\Pr(Y=1)}{\Pr(Y=0)}\prod_{k=1}^{j}\frac{\Pr(X_j|Y=1)}{\Pr(X_j|Y=0)} \end{aligned} \tag{7}$$

The likelihood ratio LR_i for each pipe i in the homogeneous group is calculated and subsequently the pipes are sorted in ascending LR order. The n highest ranked pipes on the list are those with the highest likelihood to break. The calculations are simplified by ranking the pipe according to their Ln(LR) since the Ln of LR is monotone, i.e.:

$$\mathrm{Ln(LR)} = \mathrm{Ln}\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] + \sum_{k=1}^{j}\mathrm{Ln}\frac{\Pr(X_j|Y=1)}{\Pr(X_j|Y=0)} \tag{8}$$

Continuous covariates (*Length*, *Recency*, *Scatter*) as well as discrete covariate (*NOKPF*) have to be converted to classes (or bins) in order to evaluate $\mathrm{Ln(LR}_i)$. Consequently, these covariates were converted to classes such as Very high (VH), High (H), Medium (M), Low (L) and Very low (VL), where each class is defined by lower and upper bounds for each covariate. Complete details are provided in Appendix C (see http://www.iwaponline.com/jh/014/029.pdf).

Procedure for model training

The training part of this NBC model is to find lower and upper bounds for covariates that maximize the number of hits. The procedure for training is as follows.

Training steps (a) and (b) are identical to training steps (a) and (b) in OL training.

(c) For every pipe i in the group $Y_i = 1$ if it is among the n highest breaking pipes during *WinFuture* period. Otherwise $Y_i = 0$.

(d) Establish bins (classes), e.g. Very high (VH), High (H), Medium (M), Low (L) and Very low (VL), for every covariate, as is described in Appendix C (see http://www.iwaponline.com/jh/014/029.pdf). These bins are established by specifying their lower and upper bounds for each covariate. Note that the *Cluster* covariate is by nature a class covariate and therefore does not require any transformation.

(e) Transform all covariates (except *Cluster*) to class covariates for every pipe in the group and compute likelihood ratio (LR_i) for every pipe i in the group.

(f) Rank all pipes by LR_i in descending order (i.e. pipes with the highest likelihood of ($Y = 1$) are ranked highest. The n highest ranked pipes are those predicted to have the highest number of breaks in the future window *WinFuture*. Record these pipes in a list L_{w}.

(g) Compare list L_{w} to list L_{f}. If a certain pipe i appears in both lists (irrespective of location within a list) this occurrence is defined as a 'hit'. Determine the number of hits, H.

(h) Find a set of lower and upper bin bounds that maximizes H. We used genetic algorithm to maximize H.

Procedure for model validation

(i) Repeat training step (e), but use *WinPastValidation* instead of *WinPast*.

(ii) Repeat training step (f), but use *WinPastValidation* instead of *WinPast* (use bin boundary values that were found in training step (h) above).

(iii) Repeat training step (g) but use *WinValidation* instead of *WinFuture*. Record these pipes in a list L'_{w}. The list L'_{w} contains pipes predicted by the model to have the highest number of breaks in the validation time window. Note that the validation time period did not participate in the training of the model, hence it is a holdout sample.

(iv) Determine H', the number of hits that appear in both lists L'_{w} and L_{v}.

(v) Repeat the validation process for various n values (where n is the number of pipes with the highest breakage rates – see training step (f). Use the method described in Appendix A to evaluate results (see http://www.iwaponline.com/jh/014/029.pdf).

Example. The same dataset used for the OL model above is employed here to demonstrate the application of the naïve Bayesian classification method. The periods used to define *WinPast*, *WinFuture* and *WinValidation* for the OL model were also used for the NBC model. Three separate training and validation sessions were carried out, namely for identifying pipes with at least one break, at least two breaks and at least three breaks. Note that this example cannot be trained to identify at least four breaks because no pipes were observed with more than three breaks in the *WinFuture* time period. Results are presented in Tables 5–7. It is interesting to note that the lower and upper bounds vary, sometimes significantly, among the three training sessions. It should also be remembered that the bounds are found using GA so as to maximize the number of training hits. Running the same training sessions several times consecutively will produce results that will often vary from each other since GA is a search heuristic with random elements. As can be expected, *P*-values of the training sessions are generally better than validation.

Logistic regression (LogReg) model

Logistic regression investigates how well a set of independent (explanatory) variables can explain the value of a dichotomous dependent variable. In our model the dichotomous dependent variable Y is whether the pipe belongs to the n highest breaking pipes in the next y years ($Y = 1$) or does not belong ($Y = 0$). The independent (explanatory) variables can be *NOKPF*, *Length*, *Cluster*, *Recency* and *Scatter*, where each can be taken in either their parametric form (numerical value) or their non-parametric form (ordinal value or relative rank). *Cluster* covariates can also be considered in the form of binary covariates (i.e. each cluster is represented by a 0/1 value, but the sum of all *Cluster* covariates equals unity). The probability that pipe i belongs to the n highest breaking pipes using the logistic function is expressed by:

$$\Pr_i(Y_i = 1) = \frac{e^{\underline{\beta x}}}{1 + e^{\underline{\beta x}}} \tag{9}$$

where $\underline{x}$ is a vector of covariates (or explanatory variables) and $\underline{\beta}$ is a vector of coefficients to be found by maximum likelihood. The likelihood function l for N observations ($=N$ pipes in the group) can be expressed by:

$$l = \prod_{i=1}^{N} \Pr_i^{Y_i}(1 - \Pr_i)^{1-Y_i} \tag{10}$$

Table 5 | NBC model example results – training and validation for pipes with at least one break

# pipes recorded with ≥ 1 break	Hits	*P*-value	Class	Upper and lower bin bounds for covariates # Breaks	*Length* (m)	*Recency*	*Scatter*
Training 138	46	0.0	VL	0–3	20–99	0–0.047	0–0.25
Validation 170	56	0.0	L	3–4	99–136	0.047–0.195	0.25–0.414
			M	4–9	136–430	0.195–0.5	0.414–0.422
			H	9–11	430–485	0.5–0.852	0.422–0.484
			VH	11–14	485–791	0.852–1.0	0.484–1.0

Table 6 | NBC model example results – training and validation for pipes with at least two breaks

# pipes recorded with ≥ 2 breaks	Hits	*P*-value	Class	Upper and lower bin bounds for covariates # Breaks	*Length* (m)	*Recency*	*Scatter*
Training 24	8	0.0	VL	0–2	20–346	0–0.445	0–0.086
Validation 30	3	0.046	L	2–3	346–394	0.445–0.555	0.086–0.367
			M	3–7	394–479	0.555–0.625	0.367–0.484
			H	7–12	479–599	0.625–0.961	0.484–0.93
			VH	12–14	599–791	0.961–1.0	0.93–1.0

Table 7 | NBC model example results – training and validation for pipes with at least three breaks

# pipes recorded with ≥ 3 breaks	Hits	*P*-value	Class	Upper and lower bin bounds for covariates # Breaks	*Length* (m)	*Recency*	*Scatter*
Training 6	3	0.0	VL	0–4	20–195	0–0.438	0–0.156
Validation 6	1	0.033	L	4–7	195–202	0.438–0.46	0.156–0.195
			M	7–11	202–256	0.46–0.805	0.195–0.297
			H	11–12	256–358	0.805–0.859	0.297–0.664
			VH	12–14	358–791	0.859–1.0	0.664–1.0

and the log-likelihood is therefore:

$$\Lambda = \text{Ln}(l) = \sum_{i=1}^{N} Y_i \text{Ln}(\text{Pr}_i) + (1 - Y_i)\text{Ln}(1 - \text{Pr}_i) \tag{11}$$

As before, the procedure to identify the n highest breaking pipes is divided into two sub-procedures, model training and model validation. Model validation is conducted on a holdout sample.

Procedure for model training

The modelling training process is as follows.

Training steps (a), (b) and (c) are identical to training steps (a), (b) and (c) in NBC training.

(d) For every pipe in the group, establish the covariates (explanatory variables) $\boldsymbol{x}$. These could be for example, actual *NOKPF* during *WinPast*, *Length*, *Recency* during *WinPast* and *Scatter* during *WinPast*. These covariates could be taken at their numerical values or alternatively at their rank value or combinations of ranks and numerical values.

(e) *Cluster* covariate, as described earlier, is taken as a categorical covariate. If the pipes in the group are partitioned into K clusters then K binary (zero/one) covariates are assigned, one to each cluster, with corresponding k coefficients, where the sum of these K covariates must equal unity. For example, a model that has i parametric covariates and $K = 3$ categorical covariates is described by (categorical covariates in the square brackets):

$$\underline{\beta x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + [\beta_{i+1} x_{i+1} + \beta_{i+2} x_{i+2} + \beta_{i+3} x_{i+3}]$$

where

$$x_{i+1}, x_{i+2}, x_{i+3} = 0/1; \; x_{i+1} + x_{i+2} + x_{i+3} = 1$$

(f) Perform the logistic regression and find coefficients $\underline{\beta}$ by maximizing the log-likelihood function.

(g) Apply $\underline{\beta}$ to the covariates in *WinPast* and obtain Pr_i for each pipe i.

(h) Rank all pipes by Pr_i in descending order (i.e. pipes with the highest probability of $(Y = 1)$ are ranked highest. The n highest ranked pipes are those predicted to have the highest number of breaks in the future window *WinFuture*. Record these pipes in a list L_w.

(i) Identical to training step (g) in NBC training.

Procedure for model validation

Recall that for simplicity we define *WinPastValidation* = *WinPast* + *WinFuture*. T is now partitioned into *WinPastValidation* and *WinValidation* (Figure 3).

(i) Repeat training steps (d) and (e), but use *WinPastValidation* instead of *WinPast*.

(ii) Repeat training step (g), but use *WinPastValidation* instead of *WinPast* (use $\underline{\beta}$ values that were found in training step f. above).

(iii) Repeat training step (h), but use *WinValidation* instead of *WinFuture*. Record these pipes in a list L'_w. The list L'_w contains pipes predicted by the model to have the highest number of breaks in the validation time window. Note that the validation time period did not participate in the training of the model, hence it is a holdout sample.

(iv) Determine H', the number of pipes (hits) that appear in both lists L'_w and L_v.

(v) Repeat the validation process for various *n* values (where *n* is the number of pipes with the highest breakage rates – see training step (g). Use the method described in Appendix A to evaluate results (see http://www.iwaponline.com/jh/014/029.pdf).

Example. The same dataset used for testing OL and NBC models earlier is used to demonstrate the LogReg model. However, geographical data for pipes (Figure 2) was added to form the *Cluster* covariate. The periods used to define *WinPast*, *WinFuture* and *WinValidation* for the OL and NBC models were used for the LogReg model. Potential covariates include *NOKPF*, *BreakRate* (=*NOKPF*/*Length*), *Length*, *Recency*, *Scatter* and *Clusters*. Each of the physical covariates (i.e. all except *Clusters*) can be considered either at its numerical value or at its rank value. Therefore, there are altogether 11 potential covariates (5 values, 5 ranks and 1 that corresponds to *Clusters*). The model can be run with numerous different combinations of covariates.

Table 8 provides results for nine different runs with different combinations of covariates. Each run in fact comprises three training sessions, to identify pipes with at least one or two or three breaks in their known history. It can be seen that the highest training results do not necessarily lead to the highest validation results. It can also be seen that none of the covariate combinations are clearly superior or inferior to the rest.

Non-homogeneous Poisson process based model

The non-homogeneous Poisson process (NHPP) has been suggested by several researchers to model and forecast water main breaks (e.g. Constantine & Darroch 1993; Constantine *et al*. 1996; Røstum 2000; Jarrett *et al*. 2003; Economou *et al*. 2008, among others). The approach proposed here differs from others in that it allows for the consideration of dynamic factors (climate, operations, etc.), while existing NHPP approaches consider only pipe-intrinsic, static factors (diameter, length, material, etc.).

Table 8 | LogReg model example results – training and validation with various covariates

Covariates	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9
Breaks (value)	x	x	x				x	x	x
Break rates (value)	x		x	x	x	x	x	x	x
Length (value)	x	x				x		x	x
Recency (value)	x	x	x	x		x	x		x
Scatter (value)	x	x	x	x		x	x	x	
Breaks (rank)				x	x				
Break rates (rank)									
Length (rank)			x	x	x				
Recency (rank)					x				
Scatter (rank)					x				
Clusters	x	x			x	x	x	x	x
Training hits, *H*									
138 pipes observed with ≥1 break	35	35	34	35	39	36	40	36	35
24 pipes observed with ≥2 breaks	4	4	3	4	3	4	3	4	4
6 pipes observed with ≥3 breaks	1	1	1	0	0	1	1	1	1
Validation hits, *H′*									
170 pipes observed with ≥1 break	51	52	54	56	55	46	51	50	47
30 pipes observed with ≥2 breaks	8	9	7	5	5	9	7	9	8
6 pipes observed with ≥3 breaks	1	1	1	0	0	1	1	1	1

The proposed NHPP-based model is in a different class compared to the three ranking models described earlier, in that it endeavours to forecast actual breakage rates in individual water mains, rather than just rank their relative rates.

In the proposed model, breaks at year t for an individual pipe i are assumed to be Poisson arrivals with mean intensity (or mean rate of occurrence) $\lambda_{i,t}$. Therefore, the probability of observing $k_{i,t}$ breaks is given by:

$$\Pr(k_{i,t}) = \frac{\lambda_{i,t}^{k_{i,t}} \cdot \exp(-\lambda_{i,t})}{k_{i,t}!} \tag{12}$$

where $\lambda_{i,t} = \exp[\alpha_o + \theta\tau(g_{i,t}) + \underline{\alpha}\underline{z}^i + \underline{\beta}\underline{p}^t + \underline{\gamma}\underline{q}^{i,t}]$ and where α_o is a constant, $\tau(g_{i,t})$ is the age covariate, and θ is its coefficient, $g_{i,t}$ is the age of pipe i at year t; $\underline{z}^i$ is a row vector of pipe-dependent covariates (e.g. length, diameter, etc.) and $\underline{\alpha}$ is a column vector of corresponding coefficients; $\underline{p}^t$ is a row vector of time-dependent covariates (e.g. climate) and $\underline{\beta}$ is a column vector of corresponding coefficients; $\underline{q}^{i,t}$ is a row vector of both pipe-dependent and time-dependent covariates (e.g. number of known previous failure – *NOKPF*, cathodic protection) and $\underline{\gamma}$ is a column vector of corresponding coefficients. In this paper, the function $\exp[\theta\tau(g_{i,t})]$ is referred to as the 'ageing function' and therefore coefficient θ is called the 'ageing coefficient'. Note that if $\tau(g_{i,t}) = g_{i,t}$ then the ageing is exponential, i.e. λ is an exponential function of pipe age, whereas if $\tau(t) = \text{Ln}(g_{i,t})$ the ageing function becomes a power function, i.e. λ becomes a power function of pipe age. Year t is taken relative to the first year for which breakage records are available. The likelihood function for Equation (12) is:

$$L = \prod_{i=1}^{N}\prod_{t=1}^{T} \frac{\lambda_{i,t}^{k_{i,t}} \exp(-\lambda_{i,t})}{k_{i,t}!} \tag{13}$$

Coefficients $\underline{\alpha}, \underline{\beta}, \underline{\gamma}$ are found by maximizing the log-likelihood function (14):

$$\text{LL} = \sum_{i=1}^{N}\sum_{t=1}^{T} k_{i,t} \cdot \text{Ln}(\lambda_{i,t}) - \lambda_{i,t} - \text{Ln}(k_{i,t}!) \tag{14}$$

Covariates of the NHPP model

Pipe-dependent covariates. Pipe-dependent covariates can be considered explicitly in the probabilistic model or implicitly by partitioning the data into homogeneous populations with respect to these covariates. For example, if pipe diameter is deemed to impact breakage rate then it can be explicitly considered in the $\underline{z}^i$ vector of covariates (Equation (12)) with a corresponding coefficient. Alternatively, the pipe inventory (comprising pipes of diameters, say, 6″, 8″ and 12″) can be partitioned (or grouped) into three groups, each comprising only pipes of a certain diameter and each analysed separately to produce group-specific coefficients. The explicit consideration of a covariate in Equation (12) introduces some limitations. For instance, in the example where the pipe inventory consists of three different diameters, if diameter is included in the $\underline{z}^i$ vector and the diameter coefficient is found to be negative, say, $-\alpha_r$, then the implication is that 12″ diameter pipes for instance are always expected to have a breakage rate that is smaller than the 6″ diameter pipes by a factor of $\exp(2\alpha_r)$. The grouping approach encompasses two advantages: (a) removal of the forced proportionality described above, and (b) obviation of the need to speculate about possible interactions among these covariates. These two advantages outlined above come at the cost of reduced statistical significance due to analysis of smaller pipe populations (groups), as well as the extra pre-processing effort that is needed to form these homogeneous groups.

The model can consider any number of covariates as long as these covariates are supported by available data. Further, covariates can be considered at their physical value (e.g. pipe length, diameter) or configured as categorical covariates (e.g. very long, long, short, large diameter, medium diameter, etc.). Covariates, such as pipe material, soil type, pipe cluster, etc., are inherently categorical and can be considered either as grouping criteria (as described above) or directly in Equation (12). To consider a categorical covariate, say, with m categories, m binary (zero/one) covariates are assigned, one to each category, with corresponding m coefficients, where the sum of these m covariates must equal unity.

Time-dependent covariates. Three climate-related covariates are considered as time-dependent covariates, namely freezing index (FI), cumulative rain deficit (RDc) and snapshot rain deficit (RDs). Kleiner & Rajani (2004) provided a detailed introduction and a rationale for using these covariates. FI is a surrogate for the severity of a winter, RDc is a surrogate for average annual soil moisture and RDs is a surrogate for locked-in winter soil moisture (appropriate for cold regions, where soil can freeze in the winter). Additional phenomena could be considered in the model if they are deemed to contribute to observed variations in breakage rate, provided these phenomena are supported by available data. Time-dependent covariate data are essentially time series describing these phenomena (one time series per phenomenon) over the observed period. Such phenomena can be represented quantitatively or qualitatively. For example, in one of the case studies documented in this research, uncharacteristically elevated breakage rates were observed in a network during two non-contiguous years. A quick inquiry with the utility revealed that the network had experienced pump station failures in those years, which resulted in high breakage rates probably due to transient pressures. A qualitative time series describing this phenomenon was incorporated in the model and the calibration results improved significantly. Other phenomena represented by time-series could include pressure regime changes over time, leak detection campaigns, changes over time of overburden (traffic) intensity, etc.

Note that in general time-dependent covariates such as those related to climate can typically be used to train the model on observed historical breaks but not to forecast (unless one endeavours to forecast climate as well). The rationale for using climate-related covariates is that 'true' background ageing rates (in terms of increase in breakage intensity as a function of time) are more likely to emerge if external effects, such as climate, are considered in the training process.

Pipe- and time-dependent covariates. Such covariates include the number of known previous failures (*NOKPF*), a covariate related to hotspot cathodic protection *(HSCP)* and a covariate related to retrofit cathodic protection (*RetroCP*). It should be noted that it may be beneficial to use the Ln(*NOKPF*) as the covariate in order to ensure stability in the maximum likelihood calculations, especially when discrepancies between breakage rates of individual pipes in the group are substantial.

A hotspot cathodic protection (CP) program is an opportunistic placement of sacrificial anodes, whereby a sacrificial anode is installed every time pipe is exposed for repair. These anodes typically reach full effectiveness some time (typically 1 year) after installation and deplete during operation (typically reaching complete depletion after 15–20 years in the ground), as described in Kleiner & Rajani (2004). Consequently, each pipe i has a discernible number of active anodes protecting it in any given year t. The covariate HSCP in pipe i at year t is taken as a function of the density of the active anodes along pipe i and is expressed as:

$$\mathrm{HSCP}_{i,t} = 0.1(1 - \exp(-30 q_{i,t-1})) \quad (15)$$

where $q_{i,t}$ is the density of active anodes per metre. Note that HSCP tends asymptotically to 0.1 as the number of active anodes increases. This implies that the efficacy of HSCP protection is maximized at one anode per 10 m of pipe length. The coefficients '0.1' and '30' in Equation (15) are chosen so as to assure reasonable values for anode densities found in practice.

Retrofit CP refers to the practice of systematically protecting existing pipes with galvanic cathodic protection. Kleiner & Rajani (2004) provided a detailed explanation of how the *RetroCP* covariate was created. The premise is that once a pipe is retrofitted, the ageing pattern (in terms of growing breakage rate) of a pipe is modified relative to its pre-retrofit ageing. This necessitates the consideration of three distinct phases each that are described by three additional parameters, namely transition period duration t_{tr}, coefficient θ' to describe ageing during the transition period (which is actually 'negative ageing') and post-retrofit ageing coefficient θ''. Accordingly, breakage intensity $\lambda_{i,t}$ in Equation (12) is modified to include these three phases:

$$\left.\begin{aligned}
&\lambda_{i,t} = \exp[\underline{\alpha}\,\boldsymbol{z}^{i} + \alpha_o + \theta\tau(g_{i,t}) + \underline{\beta}\,\underline{p}^{t} + \underline{\gamma}\,\underline{q}^{i,t}] \\
&\qquad \text{for } t \le t_{\mathrm{retrofit}} \\
&\lambda_{i,t} = \exp[\underline{\alpha}\,\underline{z}^{i} + \alpha_o + \theta\tau(g_{i,t_{\mathrm{retrofit}}}) + \theta'(g_{i,t} - g_{i,t_{\mathrm{retrofit}}}) \\
&\qquad + \underline{\beta}\,\underline{p}^{t} + \underline{\gamma}\,\underline{q}^{i,t}] \quad \text{for } t_{\mathrm{retrofit}} < t \le t_{\mathrm{retrofit}} + t_{\mathrm{tr}} \\
&\lambda_{i,t} = \exp[\underline{\alpha}\,\underline{z}^{i} + \alpha_o + \theta\tau(g_{i,t_{\mathrm{retrofit}}}) + \theta'(g_{i,t} - g_{i,t_{\mathrm{retrofit}}}) \\
&\qquad + \theta''\tau(g_{i,t} - (g_{i,t_{\mathrm{retrofit}}} + t_{\mathrm{tr}})) + \underline{\beta}\,\underline{p}^{t} + \underline{\gamma}\,\underline{q}^{i,t}] \\
&\qquad \text{for } t > t_{\mathrm{retrofit}} + t_{\mathrm{tr}}
\end{aligned}\right\} \quad (16)$$

where index t_{retrofit} represents the year in which the pipe was retrofitted with CP, and t_{tr} is the transition time in years. Note that Equation (16) implies that retrofit CP impacts only pipe ageing (i.e. only age covariate is modified), and the impact of all other covariates on pre- and post-retrofit breakage intensity remains the same. This may be a reasonable assumption (with the exception of *NOKPF* covariate); however, the model can be easily modified to incorporate an additional, post-retrofit set of covariates/coefficients. Also, in the situation where a specific pipe has been hotspot-protected in the years before it is retrofitted at year t, all active hotspot CP anodes starting at year t are assumed to be completely overwhelmed by the high density of retrofit anodes, with the consequence that the HSCP covariate is disregarded after year t.

Zero-inflated Poisson (ZIP) process

In reality, most water mains fail relatively rarely, which means that many (if not most) data points in a typical dataset will have the observed value $k_{i,t} = 0$ (i.e. zero breaks observed for pipe i at year t). It has been observed (e.g. Lambert 1992) that a counting process with many zeros (i.e. many more than what is expected from Equation (12)) cannot be adequately represented by a Poisson process. Lambert (1992) proposed a technique referred to as the 'zero-inflated Poisson' (ZIP) regression, for handling zero-inflated count data. In this approach, the counting process at hand is produced simultaneously by two mechanisms, namely a zero-generating process and a Poisson process. Economou *et al.* (2008) used this approach in their model to predict pipe breakage rates, and called the probability of obtaining a zero data point 'the natural tendency of the pipe to resist failure'. ZIP process can be incorporated in the proposed model and it has been observed sometimes (but not always) to improve prediction accuracy. The probability of observing $k_{i,t}$ breaks (at year t for an individual pipe i) when zero inflated count is considered becomes:

$$\Pr(k_{i,t}) = \begin{cases} G_{i,t} + (1 - G_{i,t})e^{-\lambda_{i,t}} & \text{for } k_{i,t} = 0 \\ (1 - G_{i,t})\lambda_{i,t}^{k_{i,t}} e^{-\lambda_{i,t}}/k_{i,t}! & \text{for } k_{i,t} > 0 \end{cases} \quad (17)$$

$$i = 1, 2, \ldots, N; \; t = 1, 2, \ldots, T$$

where N is the number of pipes and T is the number of years of available breakage data, $G_{i,t}$ is the parameter of the second mechanism (the first is the Poisson process) that produces $k_{i,t} = 0$ with probability $G_{i,t}$. It is convenient to formulate $G_{i,t}$ in a *Logit* form because its value must lie in the interval [0, 1], i.e. $\text{Logit}(G_{i,t}) = f(\text{some covariates})$ or $G_{i,t} = e^{f(\cdot)}/(1 + e^{f(\cdot)})$.

It is reasonable to assume that $G_{i,t}$ is generally influenced by the same covariates that influence the mean intensity $\lambda_{i,t}$. Therefore we define $G_{i,t}$ as a function of $\lambda_{i,t}$

$$G_{i,t} = \frac{e^{g_0 - \lambda_{i,t}}}{1 + e^{g_0 - \lambda_{i,t}}} \quad (18)$$

where g_0 is the ZIP coefficient. Note that with this formulation $G_{i,t}$ tends to zero as $\lambda_{i,t}$ increases and $G_{i,t}$ tends to unity as $\lambda_{i,t}$ decreases.

Model training and validation

As mentioned earlier, training of the model (or discerning its coefficients) is done by maximizing Equation (14) on observed data in the training period. The Lipschitz (-continuous) Global Optimizer (LGO) algorithm (Pintér 2005) was used in the implementation of the NHPP model but in principle other alternative algorithms can also be used.

Since numerous candidate covariates can be applicable for a specific pipe group, some of the covariates may not always be significant for all datasets. The likelihood ratio (LR) statistic can be used as a criterion to evaluate the significance of candidate covariates (e.g. Ansell & Phillips 1994). A specific covariate is removed or dropped if its contribution to LR does not exceed the required threshold at the desired confidence level (typically 5 or 1%). It should be noted that, strictly speaking, it is not sufficient to examine the LR of each covariate at a time, but rather all combinations of the candidate covariates should be tested as well because it is possible that a pair of covariates considered simultaneously in the model is statistically significant, even if each of the covariates on its own is not.

The discerned coefficients of the trained model (for a specific dataset) are used to forecast the number of breaks for the validation period and then compare the observed and forecasted number of breaks.

Two criteria need to be examined when evaluating validation results, namely accuracy of the prediction of number of breaks for every pipe at every year in the validation period (point prediction) and ranking ability. Although it is clear that perfect accuracy in point prediction will result in perfect ranking ability, the two parameters should nonetheless be evaluated independently since, in practice, perfect point prediction is unrealistic. In fact, ranking ability is the only criterion upon which the NHPP-based model can be compared to the three ranking models.

Example. The same dataset (from Calgary) used to examine the ranking models is used here to illustrate the application of the NHPP-based model. In addition, Calgary embarked on a (on-going) hotspot cathodic protection program in 1970. The model was trained on 40 years' failure data from 1962 to 2001. The clustering scheme used was the same as illustrated in Figure 2. The coefficients obtained from training were used to forecast breaks for validation for years 2002–2006. The results are illustrated in Figures 4 and 5 and summarized in Tables 9 and 10. The following should be noted:

- The two outliers in 1982 and 1986 were likely the result of pumping station failures that occurred in these years causing a significant spike in the number of pipe failures. A user-defined time-dependent covariate (or time-series) was created to represent these events qualitatively (in this time series, a unity was assigned to years 1982 and 1986 and zeros to all other years). As expected, these outliers were captured by the trained model, likely resulting in more realistic values for the other covariates, especially the ageing rate. The assignment of unity to 1982 and 1986 in the time series implies that the impact of the pump failures was identical in those two years, which of course we have no way of knowing for sure. This example should not be viewed as over-fitting because it is based on relevant available data.

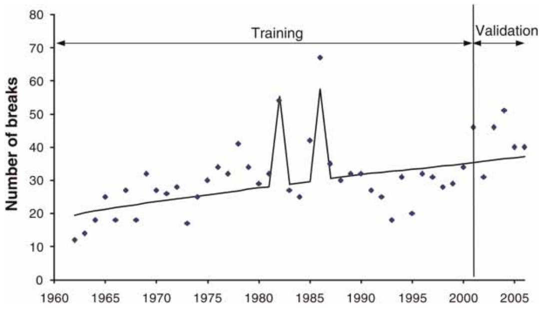

Figure 4 | NHPP model example – training and validation (breaks aggregated by year).

- The ageing covariate $\tau(t) = \mathrm{Ln}(g_{i,t})$ was used in this example. An examination of the coefficients (Table 9) reveals that background ageing was therefore proportional to the fifth root (power of about 0.2) of pipe age.
- Climate covariates as well as the hotspot CP (HSCP) covariate were found to be not significant at the 5% level and were therefore removed. Water mains in Calgary are typically buried at a depth of 2.4 m, which may explain the insignificant impact of FI and RDs covariates.
- The positive sign of NOKPF may point to a 'worse than old' condition (in repairable systems four repair-related conditions are observed, 'good as new', 'good as old', 'better than old' and 'worse than old').
- The length covariate in this case study was taken as the natural log of pipe length, which means that the number of estimated breaks was nearly linearly proportional to the length of the pipe (relatively strong dependency).
- While the NHPP-based model was quite accurate in predicting the total (cumulative) number of breaks that occurred between 2002 and 2006 (Table 11), it was not as successful in estimating the number of breaks per pipe.

The NHPP model applied to this dataset tended to overestimate the number of breaks for pipes that experienced few breaks, while under-estimating the number of breaks for those pipes that experienced a higher number of breaks. A similar tendency has been observed by others, e.g. Røstum (2000). The NHPP model predicts the expected number of breaks, which is a real number, while the observed number of breaks is of course an integer. In this type of comparison it is expected that discrepancies between observed and predicted breaks would be relatively large, especially in individual pipes with few breaks. Moreover, this discrepancy may be even greater where there are many pipes with zero or few breaks and only a few pipes with many breaks.

Comparisons of models

As stated earlier, it is clear that the three ranking models and the NHPP-based model can only be compared on their ability to rank individual mains because only the latter model can actually forecast the expected number of breaks.

Data used for comparisons. Although datasets from several water utilities were made available for this research,

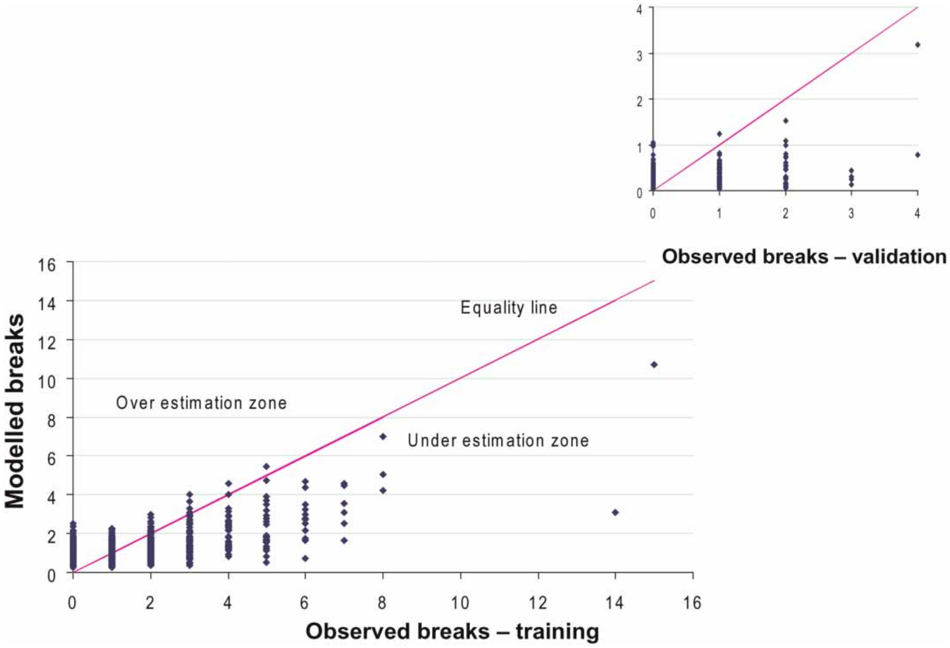

Figure 5 | NHPP model example – training and validation (breaks aggregated by pipe).

Table 9 | Summary of NHPP model example results

Covariates	Coefficient	Significant at 95% confidence[a]
Group constant	–8.20	Yes
Age	0.21	Yes
Length	1.02	Yes
NOKPF	0.73	Yes
HSCP		No
FI		No
RDs		No
RDc		No
Pump failures	0.99	Yes
Cluster 1	–1.33	
Cluster 2	–1.08	
Cluster 3	–0.60	Yes[a]
Cluster 4	–1.08	
Cluster 5	–0.91	
ZIP	1.87	Yes

[a]The cluster covariates were tested for statistical significance as a block of 5 covariates, i.e. the likelihood ratio was compared to $\chi_5^2 \cong 11.07$.

Table 10 | Ranking ability of NHPP model example results

n break(s) in validation period	*n* = 1	*n* = 2	*n* = 3	*n* = 4	*n* = 5
# pipes with at least *n* (observed) break(s) out of 1,091 pipes in group	170	30	6	2	0
# of pipes, *k*, identified correctly	61	8	1	1	n/a
P-value (probability of identifying *k* pipes by pure chance)	0.0	0.0	0.033	0.004	

only datasets from Ottawa, Calgary and Scarborough (all in Canada) were used for the comparisons because of their relative historical depth. Two homogeneous pipe groups from each of the three utilities were selected at random for comparisons, for a total of six datasets. These pipe groups varied in vintage, diameter, material and operational features. Each group was trained on various training periods with various combinations of covariates and various forecast periods for a total of 37 scenarios (runs), as detailed in Table 12. The comparisons were done only on their ranking abilities for the validation period and not on the training results. Results of these comparisons are presented in Tables 13–15. It should be noted that in trial runs, the NBC (naïve Bayesian classification) model produced results that largely mimicked the results obtained from the LogReg (Logistic regression) model. Consequently, the NBC-based model was not included in the comparison tables below.

Table 11 | Predicted vs. observed breaks (cumulative)

	# Breaks Training period	Validation period
Observed (1962–2001)	1,184	208
Predicted (2002–2006)	1,177	182

As noted earlier, the OL model (as well as the NBC model, which was not included in the comparisons below) uses a search heuristic technique with random elements (genetic algorithm or GA) for training. Consequently, repetitions of an identical scenario can yield different results. In the comparisons, each scenario was repeated 16 times for the OL model. In the comparison tables, the results of the OL model are presented as a range of minimum to maximum 'hits' obtained from these multiple scenarios. The LogReg (Logistic regression) model does not use GA for training; however, as noted earlier, there are numerous potential covariates and many possible combinations for training the model. Different combinations of covariates yield different training and validation results for the same scenario. In the comparisons, outcomes were determined from 16 different combinations (scenarios) of covariates in the LogReg model. In the comparison tables, the results of the LogReg model are presented as a range of minimum to maximum 'hits' obtained from these multiple scenarios.

The following are noteworthy comments and observations on the tabulated comparisons of the results obtained from the application of the different models to various groups/runs formed from the datasets:

- The OL and NBC ranking models split the training period into a past window (*WinPast*) and a future window (*WinFuture*), where past observations are adjusted to provide the best possible 'prediction' of the observations in the future window. Subsequently, the knowledge (or parameters) gained from this training is applied to the validation period and the forecasted values are compared to the observed ones. Further, as noted earlier, training is done via a random search algorithm. Extensive experimentation showed that there was no significant correlation between the quality of the training results and the quality of the validation results. In other words, the fact that the training results of a certain realization are better than the rest when the same scenario is run multiple times does not necessarily guarantee that the corresponding validation results will be better as well. This presents a dilemma in the application of these models to real situations, namely which set of parameters to choose, those that yield the best training results or those that yield the best validation results.
- The LogReg model has to be trained on the n pipes with at least k breaks in the training period in order to be able to forecast which pipes will have at least k breaks during the validation period. Training cannot be performed if there are no pipes with a least k breaks in the training period, i.e., when $n = 0$. In such cases 'n/a' was entered in Tables 13–15.
- In the LogReg model we are faced with a similar dilemma to that discussed above with regards to the OL and NBC models. As noted earlier, there are numerous potential covariates and many possible combinations to train the model. Different combinations of covariates yield different training and validation results for the same scenario. As in the OL model, the quality of training results is not significantly correlated to the quality of

Table 12 | Datasets and scenarios used for comparisons of models

Group	City	Material	Diam. (mm)	Vintage	Hotspot CP	Retrofit CP	# of pipes	Length (km)	Training (years)	Forecast (years)
1	Ottawa	CI	150	1951–60	1990	Yes	88	14.20	10, 20, 30	2, 5, 10
2	Ottawa	DI	150	1971–80	1990	No	288	43.35	10, 20	2, 5
3	Calgary	CI	200	1961–65	1970	No	280	27.06	10, 20, 30	2, 5, 10
4	Calgary	PDI	150	1971–75	1970	No	585	48.33	10, 20, 30	2, 5, 10
5	Scarb.	AC	150	1951–60	–	–	107	20.61	10, 20, 30	2, 5, 10
6	Scarb.	DI	200	1966–75	1984	No	128	16.33	10, 20	2, 5

CI: cast iron; DI: ductile iron; PDI: protected ductile iron; AC: asbestos cement.

Table 13 | Comparisons of results obtained from different models: Ottawa dataset

Dataset	Group/run	Training years	Forecast years	Training From	To	Forecast From	To	Number of pipes	Total length (km)
Ottawa	1A	10	2	1995	2004	2005	2006	88	14.202
	1B	10	5	1992	2001	2002	2006	88	14.202
	1C	20	2	1985	2004	2005	2006	88	14.202
	1D	20	5	1982	2001	2002	2006	88	14.202
	1E	20	10	1977	1996	1997	2006	88	14.202
	1F	30	2	1975	2004	2005	2006	88	14.202
	1G	30	5	1972	2001	2002	2006	88	14.202
	2A	10	2	1995	2004	2005	2006	288	43.353
	2B	10	5	1992	2001	2002	2006	288	43.353
	2C	20	2	1985	2004	2005	2006	288	43.353
	2D	20	5	1982	2001	2002	2006	288	43.353

	# Hits with 1 break				# Hits with 2 breaks				# Hits with 3 breaks				# Hits with 4 breaks				# Hits with 5 breaks				# Hits with 6 breaks			
Run	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR
1A	**7**	0	0–1	0–1	**0**				**0**				**0**				**0**				**0**			
1B	**25**	8	6–9	7–11	**4**	1	0–0	0–0	**0**				**0**				**0**				**0**			
1C	**7**	0	0–2	0–0	**0**				**0**				**0**				**0**				**0**			
1D	**25**	6	6–11	4–8	**4**	1	0–1	0–1	**0**				**0**				**0**				**0**			
1E	**30**	20	15–25	14–24	**7**	1	0–3	2–2	**2**	1	0–1	0–0	**0**				**0**				**0**			
1F	**7**	0	0–1	0–0	**0**				**0**				**0**				**0**				**0**			
1G	**25**	7	6–10	2–5	**4**	1	0–1	0–0	**0**				**0**				**0**				**0**			
2A	**17**	7	5–10	5–7	**3**	2	0–2	0–1	**2**	1	0–1	0–1	**1**	1	0–1	n/a	**1**	1	0–1	n/a	**1**	1	0–1	n/a
2B	**29**	9	9–16	11–15	**8**	0	1–6	2–5	**2**	0	0–1	0–0	**2**	0	0–1	n/a	**2**	0	0–1	n/a	**1**	0	0–1	n/a
2C	**17**	7	4–11	6–7	**3**	2	0–2	0–2	**2**	2	0–2	0–1	**1**	1	0–1	n/a	**1**	1	0–1	n/a	**1**	1	0–1	n/a
2D	**29**	15	9–15	11–13	**8**	4	1–5	2–5	**2**	2	0–2	0–0	**2**	2	0–2	n/a	**2**	2	0–2	n/a	**1**	1	0–1	n/a

Numbers in bold font at the left column of each category represent the number of pipes observed to have experienced at least *n* breaks.

Table 14 | Comparisons of results obtained from different models: Scarborough dataset

Dataset	Group/run	Training years	Forecast years	Training From	To	Forecast From	To	Number of pipes	Total length (km)
Scarborough	5A	10	2	1984	1993	1994	1995	107	20.607
	5B	10	5	1984	1993	1994	1998	107	20.607
	5C	20	2	1974	1993	1994	1995	107	20.607
	5D	20	5	1974	1993	1994	1998	107	20.607
	5E	20	10	1974	1993	1994	2003	107	20.607
	5F	30	2	1964	1993	1994	1995	107	20.607
	5G	30	5	1964	1993	1994	1998	107	20.607
	5H	30	10	1964	1993	1994	2003	107	20.607
	6A	10	2	1984	1993	1994	1995	122	16.333
	6B	10	5	1984	1993	1994	1998	122	16.333
	6C	20	2	1981	2000	2001	2002	122	16.333
	6D	20	5	1979	1998	1999	2003	122	16.333

Run	# Hits with 1 break # pipes	I-warp	OL	LR	# Hits with 2 breaks # pipes	NHPP	OL	LR	# Hits with 3 breaks # pipes	NHPP	OL	LR	# Hits with 4 breaks # pipes	NHPP	OL	LR	# Hits with 5 breaks # pipes	NHPP	OL	LR	# Hits with 6 breaks # pipes	NHPP	OL	LR
5A	**5**	0	0–0	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5B	**6**	0	0–1	0–1	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5C	**5**	0	0–1	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5D	**6**	0	0–1	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5E	**11**	1	0–4	0–2	**1**	0	0–0	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**			
5F	**5**	0	0–1	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5G	**6**	0	0–1	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**				**0**			
5H	**11**	0	0–3	1–2	**1**	0	0–0	0–0	**1**	0	0–0	n/a	**0**				**0**				**0**			
6A	**6**	2	1–2	0–2	**2**	1	0–1	0–1	**0**				**0**				**0**				**0**			
6B	**10**	5	3–4	1–4	**3**	2	1–2	0–2	**1**	0	0–0	0–0	**0**				**0**				**0**			
6C	**6**	3	1–4	2–3	**1**	0	0–0	0–0	**0**				**0**				**0**				**0**			
6D	**11**	4	2–6	2–5	**3**	1	0–2	0–1	**2**	1	0–1	0–0	**1**	1	0–1	0–0	**1**	1	0–1	0–0	**0**			

Numbers in bold font at the left column of each category represent the number of pipes observed to have experienced at least *n* breaks.

Table 15 | Comparisons of results obtained from different models: Calgary dataset

Dataset	Group/run	Training years	Forecast years	Training From	Training To	Forecast From	Forecast To	Number of pipes	Total length (km)
Calgary	3A	10	2	1987	1996	1997	1998	280	27.062
	3B	10	5	1987	1996	1997	2001	280	27.062
	3C	20	2	1977	1996	1997	1998	280	27.062
	3D	20	5	1977	1996	1997	2001	280	27.062
	3E	20	10	1977	1996	1997	2006	280	27.062
	3F	30	2	1967	1996	1997	1998	280	27.062
	3G	30	5	1967	1996	1997	2001	280	27.062
	3H	30	10	1967	1996	1997	2006	280	27.062
	4A	10	2	1987	1996	1997	1998	585	48.334
	4B	10	5	1987	1996	1997	2001	585	48.334
	4C	20	2	1977	1996	1997	1998	585	48.334
	4D	20	5	1982	2001	2002	2006	585	48.334
	4E	20	10	1977	1996	1997	2006	585	48.334
	4F	30	2	1976	2004	2005	2006	585	48.334

	# Hits with 1 break				# Hits with 2 breaks				# Hits with 3 breaks				# Hits with 4 breaks				# Hits with 5 breaks				# Hits with 6 breaks			
Run	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR	# pipes	NHPP	OL	LR
3A	**10**	1	0–2	1–1	**0**				**0**				**0**				**0**				**0**			
3B	**30**	6	4–8	2–6	**8**	1	0–1	0–0	**3**	0	0–0	0–0	**1**	0	0–0	0–0	**0**				**0**			
3C	**10**	2	0–3	1–2	**0**				**0**				**0**				**0**				**0**			
3D	**30**	9	4–8	1–2	**8**	1	0–2	0–1	**3**	0	0–0	0–0	**1**	0	0–0	0–0	**0**				**0**			
3E	**45**	16	8–17	14–17	**15**	5	0–6	3–4	**7**	1	0–2	0–1	**3**	0	0–0	0–0	**2**	0	0–0	0–0	**0**			
3F	**10**	2	0–2	1–1	**0**				**0**				**0**				**0**				**0**			
3G	**30**	8	1–8	6–7	**8**	1	0–1	0–1	**3**	0	0–0	0–0	**1**	0	0–0	0–0	**0**				**0**			
3H	**45**	17	8–17	12–17	**15**	5	0–5	3–5	**7**	1	0–2	0–2	**3**	0	0–0	0–0	**2**	0	0–0	0–0	**0**			
4A	**15**	2	0–3	0–0	**0**				**0**				**0**				**0**				**0**			
4B	**40**	10	5–12	10–13	**5**	0	0–0	0–0	**0**				**0**				**0**				**0**			
4C	**15**	0	0–2	0–1	**0**				**0**				**0**				**0**				**0**			
4D	**27**	6	5–9	7–8	**7**	2	0–2	0–0	**2**	0	0–0	n/a	**1**	0	0–0	n/a	**1**	0	0–0	n/a	**0**			
4E	**59**	21	11–25	17–23	**17**	4	0–7	3–6	**4**	0	0–2	0–0	**2**	0	0–0	0–0	**1**	0	0–0	n/a	**0**			
4F	**10**	4	0–2	0–1	**2**	0	0–0	0–0	**0**				**0**				**0**				**0**			

Numbers in bold font at the left column of each category represent the number of pipes observed to have experienced at least *n* breaks.

the validation results. Therefore, again, one is faced with the dilemma: which combination of covariates should be selected for real applications, the one that yields the best training results or the one that yields the best validation results?

- CP can only be considered directly by the NHPP model. The other models do not make use of CP-related information.
- Short available breakage histories limit the options to partition training period into *WinPast* and *WinFuture* time windows in the OL, LR and NBC methods. For example, it is not possible to forecast breaks for more than 5 years (e.g. *WinPast* = 5 years and *WinFuture* = 5 years) when available historical data comprises only 10 years of breakage records. The NHPP model will also benefit from a longer historical dataset; however, it is less limited than the ranking models in applicability to short datasets.
- A close scrutiny of the number of 'hits' in Tables 13–15 suggests that no specific model is consistently superior or inferior to the others, in terms of the number of 'hits' in the validation period. Rather the relative success of the different models varies for the different datasets.

Although no specific model showed a clear superiority in terms of the number of 'hits' in the validation period, our conclusion is that for practical applications the NHPP-based model is the preferred one for the following reasons:

- It is not subject to the dilemmas described above with regards to the ranking models. Training of the NHPP-based model is based on comparison of observed to modelled breaks (rather than on the number of 'hits') and a widely accepted mathematical procedure exists (likelihood ratio test) to identify insignificant covariates.
- NHPP-based model can directly consider time-dependent covariates, including quantitative and qualitative with no restrictions.
- In contrast to the ranking models, which provide only the relative ranking of pipes by their anticipated breakage rate, the NHPP model provides an actual forecast of mean breaks for every pipe in every year in the forecast period. Notwithstanding the obvious uncertainties that are inherent in such forecasts, this form of the results lends itself to a more robust and rigorous decision process that needs to consider expected failure costs.
- The ranking models provide relative ranking within a homogeneous group of pipes and therefore it is not directly possible to rank pipes across groups. In contrast, the NHPP model provides actual forecasts of failures, which are directly comparable across groups.

Finally, a word of caution is warranted here regarding the accuracy of available data on the occurrence of failure. The exact timing of failure occurrence is determined by when the failure was detected rather than when it actually occurred. Some available data (Hughes 2008) indicate that the vast majority of leaks surfaced within a few days from initiation; however, sometimes time lags between occurrence and discovery can be as long as several months. The NHPP model described here uses a time step of one year. With such a time step, the occurrence to discovery time lag is not likely to have any discernable impact. However, this lag phenomenon is likely to have some impact if shorter time steps are used. It is also interesting to note that use of shorter time steps (e.g. 30 or 90 days) is likely to enhance the 'explanatory power' of some temperature-related covariates, such as freezing index, FI (Rajani & Kleiner 2012).

SUMMARY AND CONCLUSIONS

Three different 'ranking models' were introduced to identify, within a homogeneous group of pipes, those individual pipes that are expected to experience the highest number of breaks in the future. The covariates include, for each individual pipe, breakage history (number of known previous failures – *NOKPF*, *Recency* and *Scatter*), pipe length (*Length*) and geographical *Cluster*, which serves as a surrogate for geographically related data that are otherwise unavailable (e.g. soil, traffic, installation practices, etc.).

The ordered lists (OL) is a non-parametric heuristic model that assigns weights to the covariates and tests the aggregated weighted covariates on their ability to forecast pipes that will experience the highest number of breaks. The model is based on the premise that some monotone relationships exist between a set of covariates and anticipated breakage frequency in individual water mains. Model training is achieved by finding weights that maximize the model's ability to forecast pipe breaks.

The naïve Bayesian classification (NBC) model is based on Bayes' rule, where covariates (*NOKPF*, *Length*, *Recency*, *Scatter*) are partitioned to form classes, e.g. very high, high, medium, etc. Subsequently, the membership of each individual pipe of these classes determines its probability to be among a defined group of pipes with the highest number of breaks, i.e. those pipes with *n* or more breaks. For each covariate, class bounds are defined to partition the covariates into classes, e.g. minimum pipe length has to be defined, above which the pipe is classified as 'very long'. Model training is done by finding those class bounds that maximize the model's ability to forecast pipe breaks.

The logistic regression (LogReg) model is used to determine the likelihood of a pipe to belong to a group of pipes with the highest number of breaks, i.e. those pipes with *n* or more breaks. Covariates (*NOKPF*, *Length*, *Recency*, *Scatter*) can be used in their parametric form (value) or non-parametric form (rank). The model is trained by extracting coefficients of the logistic function that maximizes its ability to forecast pipe breaks.

The geographical *Cluster* covariate is categorical by nature. In OL it cannot be used directly as a covariate. Instead, the homogeneous group can be partitioned into clusters and training carried out for each partitioned subgroup. *Cluster* covariate can be considered directly in NBC and LogReg models.

A fourth model, NHPP-based, was also introduced. From the numerous cases examined throughout this research, as well as from the 37 comparison tests presented here, we have observed that covariates *Age*, *Length*, *NOKPF* and *RetroCP* (wherever relevant) appear to be statistically significant in the vast majority of cases, where NHPP-based model was applied.

In many instances, at least some of the climate covariates emerged as significant but not always. This may be because sometimes there is a relatively high correlation between climate time-series but further research may be required to ascertain that. Also, climate covariates are not always statistically significant. This may be for various reasons. Water utilities in cold-climate often bury their water mains quite deep (e.g. 2.4 m in Ottawa), where the potential impact of climate-related factors is greatly dissipated. Further, the predominant failure mode in ductile iron pipes is corrosion holes. Climate-related factors are less likely to impact ductile iron pipes with corrosion holes than, say, cast iron mains, which are prone to circumferential or longitudinal pipe failures. Also, it is worth noting that temperature-related covariates can be constructed in many ways (other than *FI* discussed here) so as to capture not only the intensity of temperature effects but also the rate of change, the amplitude of change, etc. These covariates have been recently explored thoroughly by Rajani & Kleiner (2012).

While the ranking models can only rank individual pipes (within a group) in terms of their anticipated relative likelihood of failure, the NHPP-based model provides an actual forecast of the mean number of breaks anticipated in each pipe at each year of the forecast period. The four models could naturally be compared only on the basis of their ranking abilities due to the limitation of the ranking models. These models were applied to 37 different scenarios extracted from six different datasets that in turn were taken from three different water utilities and results from these scenarios were compared. The success rates of the models varied among these scenarios but no one model was observed to be consistently superior or consistently inferior to the others, in terms of ranking ability. However, it is clear that though the NHPP-based model is not superior in ranking ability, it is superior in its ability to directly consider quantitative and qualitative time-dependent covariates and has the ability to forecast actual number of breaks, which is important to support effective decision making.

ACKNOWLEDGEMENT

The research reported in this paper was co-sponsored by the Water Research Foundation (WaterRF), the National Research Council of Canada (NRC) and in-kind support of water utilities from the United States and Canada.

REFERENCES

Andreou, S. A., Marks, D. H. & Clark, R. M. 1987 A new methodology for modeling break failure patterns in deteriorating water distribution systems: applications. *Adv. Water Resour.* **10**, 11–20.

Ansell, J. I. & Phillips, M. J. 1994 *Practical Methods for Reliability Data Analysis*. Oxford University Press, Oxford, UK.

Boxall, J. B., O'Hagen, A., Pooladsaz, S., Saul, A. & Unwin, D. 2007 Estimation of burst rate in water distribution mains. In

Proceedings of the Institution of Civil Engineers, Water Management I60, Issue WM2, June, pp. 73–82.
Clark, R. M., Stafford, C. L. & Goodrich, J. A. 1982 Water distribution systems: a spatial and cost evaluation. *J. Water Resour. Plann. Manag. Div. ASCE* **108** (3), 243–256.
Constantine, A. G. & Darroch, J. N. 1993 *Pipeline Reliability: Stochastic Models in Engineering Technology and Management* (S. Osaki & D. N. P. Murthy, eds). World Scientific Publishing Co, Singapore.
Constantine, A. G., Darroch, J. N. & Miller, R. 1996 *Predicting Underground Pipe Failure*. Australian Water Works Association, St Leonards, NSW.
Dandy, G. C. & Engelhardt, M. O. 2001 The optimal scheduling of water pipe replacement using genetic algorithms. *J. Water Resour. Plann. Manag.* **127** (4), 214–223.
Dridi, L., Mailhot, A., Parizeau, M. & Villeneuve, J. P. 2005 A strategy for optimal replacement of water pipes integrating structural and hydraulic indicators based on a statistical water pipe break model. In *Proceedings of the 8th International Conference on Computing and Control for the Water Industry*, University of Exeter, UK, September, pp. 65–70.
Economou, T., Kapelan, Z. & Bailey, T. 2008 A zero-inflated Bayesian model for the prediction of water pipe bursts. In *Proc. 10th International Water Distribution System Analysis Conference*, CD-ROM edition, Kruger National Park, South Africa.
Eisenbeis, P., Røstum, J. & Le Gat, Y. 1999 Statistical models for assessing the technical state of water networks – some European experiences. In *Proceedings of the AWWA Annual Conference*, Chicago.
Giustolisi, O. & Berardi, L. 2007 Pipe level burst prediction using EPR and MCS-EPR. In *Proceedings of the Combined International Conference of Computing and Control for the Water Industry (CCWI2007) and Sustainable Urban Water Management (SUWM2007) on Water Management Challenges in Global Change*, Leicester, UK, September, pp. 39–46.
Giustolisi, O., Laucelli, D. & Savic, D. A. 2005 A decision support framework for short time planning of rehabilitation. *Proceedings of Computer and Control in Water Industry (CCWI)* **1**, 39–44.
Goulter, I. C. & Kazemi, A. 1988 Spatial and temporal groupings of water main pipe breakage in Winnipeg. *Can. J. Civ. Eng.* **15** (1), 91–97.
Goulter, I. C., Davidson, J. & Jacobs, P. 1993 Predicting water-main breakage rate. *J. Water Resour. Plann. Manag. ASCE* **119** (4), 419–436.
Gustafson, J.-M. & Clancy, D. V. 1999 Modeling the occurrence of breaks in cast iron water mains using methods of survival analysis. In *Proc. AWWA Annual Conference*, Chicago.
Hughes, D. 2008 Personal communication.
Jacobs, P. & Karney, B. 1994 GIS development with application to cast iron water main breakage rate. In *2nd Int. Conf. on Water Pipeline Systems*, BHR Group Ltd, Edinburgh, Scotland.
Jarrett, R., Hussain, O. & Van der Touw, J. 2003 *Reliability Assessment of Water Pipelines Using Limited Data*. OzWater, Perth, Australia.
Kettler, A. J. & Goulter, I. C. 1985 An analysis of pipe breakage in urban water distribution networks. *Can. J. Civ. Eng.* **12**, 286–293.
Kleiner, Y. & Rajani, B. 2001 Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* **3**, 131–150.
Kleiner, Y. & Rajani, B. 2004 Quantifying effectiveness of cathodic protection in water mains: theory. *J. Infrastruct. Syst. ASCE* **10** (2), 43–51.
Kleiner, Y. & Rajani, B. 2010 Dynamic influences on the deterioration rates of individual water manis (I-WARP). *Water Research Foundation Research Report*, Denver, USA.
Lambert, D. 1992 Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** (1), 1–14.
Le Gat, Y. 2008 Extending the Yule process to model recurrent failures of pressure pipes. Private communication.
Le Gat, Y. & Eisenbeis, P. 2000 Using maintenance records to forecast failures in water networks. *Urban Water* **2**, 173–181.
MacQueen, J. B. 1967 Some methods for classification and analysis of multivariate observations. In *Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, vol. 1, pp. 281–297.
Mailhot, A., Paulin, A. & Villeneuve, J.-P. 2003 Optimal replacement of water pipes. *Water Resour. Res.* **39** (5), 1136.
Mavin, K. 1996 Predicting the failure performance of individual water mains. *Urban Water Research Association of Australia, Research Report No. 114*, Melbourne, Australia.
Park, S. & Loganathan, G. V. 2002 Optimal pipe replacement analysis with a new pipe break prediction model. *Journal of the Korean Society of Water and Wastewater* **16** (6), 710–716.
Pintér, J. D. 2005 LGO: *A Model Development and Solver System for Continuous Global Optimization*. Pintér Consulting Services, Inc., Halifax, NS, Canada.
Rajani, B. & Kleiner, Y. 2012 Exploration of the relationship between water main breaks and temperature covariates. *Urban Water Journal*. Available from: www.urbanwater.net (accessed 21 February 2012).
Røstum, J. 2000 *Statistical modelling of pipe failures in water networks*. PhD Thesis, Norwegian University of Science and Technology, Trondheim, Norway.
Shamir, U. & Howard, C. D. D. 1979 An analytic approach to scheduling pipe replacement. *J. AWWA* **71** (5), 248–258.
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. *Journal of Hydroinformatics* **6** (4), 259–264.

First received 8 March 2011; accepted in revised form 18 August 2011. Available online 8 November 2011