On the prediction of underground water pipe failures: zero inflation and pipe-specific effects

Theodoros Economou, Zoran Kapelan and Trevor C. Bailey

ABSTRACT

The prediction of pipe failures in water distribution systems is an essential planning tool for water companies. Previous methods focus on the prediction of either future failure numbers or aspects of pipe condition. However, most of these only predict at the level of large pipe groups (of similar characteristics) and often cannot provide uncertainty bounds. Here, a new statistical method is developed to predict the probability of failure at the single pipe level. The method extends the Non-Homogeneous Poisson Process (NHPP) in two ways: firstly, it incorporates pipe-specific random effects to account for unmeasured information on the factors affecting the pipe failures. Secondly, the method explicitly accounts for zero inflation, that is the possibility that more zero failures occur than expected from a simple Poisson assumption. This zero-inflated NHPP (ZINHPP) model was applied to two real-life datasets, one from North America and one from New Zealand. The results clearly demonstrate improved prediction capability, especially in the New Zealand data, which contain a much larger percentage of pipes with zero failures.

Key words | Bayesian mixture model, Markov chain Monte Carlo, random effects

Theodoros Economou (corresponding author)
Zoran Kapelan
Trevor C. Bailey
College of Engineering,
Mathematics and Physical Sciences,
North Park Road,
Exeter EX4 4QF,
UK
E-mail: *t.economou@ex.ac.uk*

NOTATION

Variables

$T_{i,0}$	start of observation period for pipe i
$T_{i,\text{end}}$	end of observation period for pipe i
n_i	failure count for pipe i
$t_{i,j}$	time of failure j for pipe i
M	total number of pipes in network
x_i	vector of explanatory variables for pipe i
w_i	0/1 variable where $w_i = 0$ implies pipe i never failed
$[t_a, t_b)$	time interval (left-closed, right-open)
$\Lambda(t_a, t_b)$	expected number of failures in $[t_a, t_b)$
$N(t_a, t_b)$	random variable for the number of failures in $[t_a, t_b)$

Parameters

β_0	intercept in the failure rate
β_1	parameter relating to pipe length
β_2	parameter relating to pipe diameter
β_3	parameter relating to medium pressure
β_4	parameter relating to high pressure
β_5	parameter relating to medium pressure change
$1-p_i$	probability of zero inflation in pipe i
$\gamma_{i,\,0}$	pipe-specific intercept in p_i
γ_1	parameter relating to age in p_i
θ_i	random effect for pipe i
α	parameter for the gamma model for θ_i
β	parameter for the gamma model for θ_i
$\lambda(t_{i,\,j}, x_i)\|\theta_i$	failure rate for pipe i given random effects θ_i

Notation

$\text{Norm}(\mu, \sigma)$	normal distribution with mean μ and variance σ
$\text{Gam}(\alpha, \beta)$	gamma distribution with parameters α and β
$L_{PP}(\tau_i, x_i\|\theta_i)$	NHPP likelihood
$f_{\text{ZIPP}}(n_i;\tau_i,x_i\|\theta_i)$	probability density function for ZINHPP

doi: 10.2166/hydro.2012.144

$l(\tau,x;\ p\|\theta)$	ZINHPP log-likelihood
I_{event}	$I_{\text{event}}=1$ if event is true, $I_{\text{event}}=1$ otherwise
ϕ	a vector of Bayesian model parameters
$p(\phi\|y)$	posterior distribution
$\pi(\phi)$	prior distribution
$l(y\|\phi)$	Bayesian model likelihood
DIC	deviance information criterion
$\hat{D}$	mean posterior deviance
$\tilde{D}$	deviance evaluated at the posterior mean of parameters

INTRODUCTION

The successful prediction of pipe failures is important for water companies in terms of budgeting and for planning replacements or repairs. The complex processes that affect the occurrence of failures in water pipes are not fully understood nor always observed. Factors affecting pipe failures can be related to pipe characteristics (e.g. pipe material, diameter, past failure history, etc.), the environment (e.g. soil characteristics, weather data, corrosivity, traffic conditions, etc.) and the service itself (e.g. water pressure, level/type of maintenance, etc.).

Statistical methodology has been frequently used to model recurrent failures in water pipes. Kleiner & Rajani (2001) overview statistical models for pipe failures and Loganathan *et al.* (2002) summarise statistical pipe replacement analyses. In addition, Engelhardt *et al.* (2000) review rehabilitation strategies for distribution networks from a UK perspective including statistical models that attempt to capture the failure mechanism.

A frequently adopted model to describe failure occurrence in time is the counting process which can also allow for other effects in the form of explanatory variables. In particular, the Non-Homogeneous Poisson Process (NHPP) has often been used due to its flexibility and desirable properties. One such property is non-homogeneity in time which can capture the deterioration mechanism in ageing water pipes, by explicitly modelling failure rate.

Constantine & Darroch (1993) and Constantine *et al.* (1996) were the first to apply a conventional NHPP model where the failure rate depended exponentially on time. Goulter *et al.* (1993) use NHPP to model the recurrence of failures when the failure rate is a function of both time and distance from the previous failure. Saldanha *et al.* (2001) consider an NHPP for failures in service water pumps in a nuclear plant where the failure rate is described by both a power law and a log-linear formulation. The NHPP is also utilised by Kleiner & Rajani (2008) and Kleiner *et al.* (2010) where both pipe-dependent and time-dependent explanatory variables (such as pipe diameter and temperature) are considered in the failure rate. In addition, a Bayesian random effect NHPP model has been applied to a network of water pipes in New Zealand by Watson (2005) and was used in conjunction with cost functions to identify an optimal pipe replacement policy. Park *et al.* (2008) have considered a global (same failure rate assumed for all pipes) NHPP model for a network of 9649 pipes observed over a 70 year period.

Conventionally, the NHPP model is defined by the failure rate $\lambda(t;\ x)$, using a monotonic function of time such as the power law: $\lambda(t;\ x)=\theta\ t^{\theta-1}\exp(\beta x')$ or the log-linear relationship: $\lambda(t;\ x)=\alpha\ \exp\{\gamma t+\beta x'\}$, where x is a vector of possible explanatory variables. Such models are well established and allow analytical integration to obtain the expected number of failures $\Lambda(t_a,\ t_b;\ x)$ over a time interval $(t_a,\ t_b]$ for $0\le t_a<t_b$. For instance, for a power-law-based $\lambda(t;\ x)$:

$$\Lambda(t_a,\ t_b;\ x)=\int_{t_a}^{t_b}\lambda(u;\ x)\ \mathrm{d}u=\left[t_b^{\theta}-t_a^{\theta}\right]\exp\{\beta \mathrm{x}'\} \tag{1}$$

where $\Lambda(t_a,\ t_b;\ x)$ is the mean of a Poisson random variable $N(t_a,\ t_b)$, corresponding to the count of failures in $(t_a,\ t_b]$.

Such conventional NHPP models are useful for predicting pipe failures, however these often lack the flexibility to adequately capture the complex failure mechanism in real-life datasets. For instance, a common assumption when fitting an NHPP is that the failure mechanism in water pipes of identical characteristics (material, diameter, etc.) is the same (the homogeneity assumption) and that the variability within pipes is purely stochastic. In practice, the homogeneity assumption may be too naive as pipes are buried and the actual

condition and factors affecting each pipe are often unknown. This paper advocates the use of pipe-specific random effects to address this issue. Random effects models are well established in statistics (Gelman *et al.* 2004; Faraway 2006) and allow model parameters to vary stochastically. The modelled variability in the parameters can capture the effect of unobserved explanatory variables. To date, relatively little attention has been given to random effects in modelling pipe failure with the exception of Watson *et al.* (2004).

In addition, there are issues with water pipe data that may limit the effectiveness of conventional NHPP models. A particular problem in repairable equipment with relatively long lifetimes is that data often lack past information (left-truncation). Indeed, data involving water pipes are frequently limited to only a few years in relation to the age of pipes in the network (Gat & Eisenbeis 2000). This, added to the fact that, in general, failures are rare over the lifespan of a pipe, results in a large number of zero failures for many datasets. Furthermore, some pipes may have extra resistance to failure, for instance due to better installation where 'extra' implies resistance that is on top of what is expected by empirical knowledge. In other words, an assumption is made here that pipes which are identical in what can be measured as explanatory variables can exhibit different failing behaviour due to unknown/unobserved factors. In this paper, a zero-inflated NHPP (ZINHPP) model is proposed to account for the possibility of more zeros in failure counts than would be expected from the NHPP alone, by allocating extra probability to the possibility of no failure. In addition, this model provides a novel way of accounting for the aforementioned possible extra resistance to failure.

The work presented here is based on Economou *et al.* (2008) where a similar ZINHPP model was first introduced. However, the model had an extra random effect in the failure rate, resulting in loss of model identifiability due to over-parametrisation. The ZINHPP model in Economou *et al.* (2008) was later used in Kleiner & Rajani (2010) to incorporate zero inflation. An improved ZINHPP model is presented here, leading to more reliable results.

METHODOLOGY

Random effects NHPP

Consider that a single water pipe i has been observed in the time period $[T_{i,0},\ T_{i,\text{end}})$ having experienced n_i failures at times $T_{i,0} < t_{i,1},\ t_{i,2}, \ldots,\ t_{i,n} \le T_{i,\text{end}}$. $T_{i,0}$ is not necessarily the time of installation but the starting time of observation. Assuming an NHPP, the likelihood (Rigdon & Basu 2000) of pipe i with failure rate $\lambda(t; x_i|\theta_i)$, given any random effects θ_i is:

$$L_{PP}(\tau_i, x_i|\theta_i) = \left[\prod_{j=1}^{n_i} \lambda(t_{i,j}; x_i|\theta_i)\right]^{w_i} \exp\{-\Lambda(T_{i,0}, T_{i,\text{end}}; x_i|\theta_i)\} \tag{2}$$

where $\tau_i = T_{i,0}, t_{i,1}, \ldots, t_{i,n_i}, T_{i,\text{ end}}$ and x_i are the explanatory variables, $w_i = 0$ if pipe i has never failed and $w_i = 1$ otherwise. The likelihood for a network of M pipes is $\prod_{i=1}^{M} L_{PP}(\tau_i, x_i|\theta_i)$.

The pipe-specific θ_i are different for each pipe but are assumed to arise from a common stochastic model; hence they are random effects, not fixed effects. This reduces the dimensionality of the model while preserving flexibility but at the same time induces a dependency structure between pipes. We follow the logic introduced by Watson *et al.* (2004) where an observed failure in one pipe in the network may provide useful information about the failure rate of another similar pipe. This can be valuable for decision-making where some pipes have no recorded failure data.

As an example, consider a network where only a single explanatory variable x_i is available. Assuming a power-law model for the failure rate of pipe i, then a possible random effects model is:

$$\begin{aligned} &\lambda(t; x_i|\theta_i) = \theta_i t^{\theta_i - 1} \exp\{\beta_0 + \beta_1 x_i\} \\ &\theta_i \sim \text{Gam}(\alpha, \kappa) \end{aligned} \tag{3}$$

The random effects θ_i are modelled using a two-parameter Gamma distribution; hence the failure rate is also stochastic.

With the inclusion of random effects θ_i in the model, more heterogeneity in the pipes is allowed for than just the amount explained by the variability in the explanatory variables x_i. This enables the model to deal with pipes that

exhibit uncharacteristic behaviour relative to other pipes. Pipe-specific models such as Equation (3) are potentially useful. However, relatively little work exists in using such models within the water pipe literature (Watson 2005). Kleiner *et al.* (2010) have recently utilised a time-dependent, pipe-specific Poisson model for failure counts; however, heterogeneity between pipes is modelled using explanatory variables (many of which are not available in the data used in this study) rather than random effects.

Zero-inflated NHPP

An NHPP model is equivalent to a Poisson distribution with a time-varying mean; therefore the total number of failures for each pipe is Poisson-distributed (Cook & Lawless 2007). When a considerable number of pipes have experienced no failures at all, the Poisson assumption and its mean-variance relationship sometimes fail to adequately describe this. This phenomenon appears frequently in many kinds of count data. It is referred to as zero inflation and is a term used when the proportion of zeros in the data exceeds the proportion generated by the fitted model.

To cope with zero inflation in counts of defects in manufactured items, Lambert (1992) introduced a zero-inflated Poisson (ZIP) regression model where extra probability is allowed for zero counts. The ZIP model is a special case of a mixture model with two components: one which generates zeros with probability (1–p) and another which produces counts with probability p (including zero) from a Poisson distribution. Mixture models are well established in statistics (Pawitan 2001; Gelman *et al.* 2004) and are constructed by combining two or more probability models.

Using the same idea, the conventional NHPP model can be extended for zero inflation, where the failure process is either an NHPP or a zero-generating process depending on a mixing probability p. A ZINHPP model may be formulated as a mixture distribution as follows:

$$f_{\mathrm{ZIPP}}(n_i;\tau_i,x_i|\theta_i)=\begin{cases}(1-p_i)+p_i\exp\left[-\Lambda\left(T_{i,0},T_{i,\mathrm{end}}|\theta_i\right)\right] & \text{if} \quad n_i=0\\ p_iL_{\mathrm{PP}}(\tau_i,x_i|\theta_i) & \text{if} \quad n_i=1,2,\ldots\end{cases} \tag{4}$$

where $\exp[-\Lambda(T_{i,\,0},\,T_{i,\,\mathrm{end}}|\theta_i)]$ is the probability of zero failures in $[T_{i,0},\,T_{i,\mathrm{end}})$. Integrating $f_{\mathrm{ZIPP}}()$ over τ_i then summing over n_i results in the value 1; hence it is a proper probability density function. When $n_i=0$, this is either due to the NHPP (reflected in the term $\exp[-\Lambda(T_{i,0},\,T_{i,\mathrm{end}}|\theta_i)]$) with probability p_i or due to a zero-generating process with probability (1–p_i). If enough information exists about pipe i that can be adequately captured by the NHPP, p_i will be close to 1. Otherwise, p_i will be small, giving more probability of zero failures but still allowing a non-zero failure rate.

Conditional on θ_i, the log-likelihood of the ZINHPP model for pipes $i=1,\ldots,M$ is given by

$$l(\boldsymbol{\tau},\boldsymbol{x};\boldsymbol{p}|\theta)=\sum_{i=1}^{N}\left(I_{(n_i=0)}\log\left((1-p_i)+p_i\exp\left\{-\Lambda\left(T_{i,0},T_{i,\mathrm{end}}|\theta_i\right)\right\}\right)\right)$$
$$+\sum_{i=1}^{N}I_{(n_i>0)}[\log(p_i)+\log(L_{PP}(\boldsymbol{\tau}_i,\mathbf{x}_i|\theta_i))] \tag{5}$$

where $I_{(\mathrm{event})}=1$ if the event is true and $I_{(\mathrm{event})}=0$ otherwise. Also, $\boldsymbol{x}=(x_1,\ldots,x_M)$, $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_M)$, $\boldsymbol{p}=(p_1,\ldots,p_M)$ and $\boldsymbol{\tau}=(\tau_1,\ldots,\tau_M)$.

Each pipe has a parameter p_i (the mixing probability) where (1–p_i) may be considered to reflect resistance to failure. Therefore, a small value of p_i indicates that a pipe has extra resistance to failure on top of the underlying breakage resistance modelled by the NHPP.

Bayesian statistics

Many options exist to estimate the models described so far but here the Bayesian framework (Gelman *et al.* 2004) is considered. This framework is particularly flexible because it allows estimation of mixture models including random effects such as the ones described here, in a relatively straightforward manner. In Bayesian statistical modelling, model parameters $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_k)$ are viewed as random variables whose distribution $p(\boldsymbol{\phi}|y)$ called the 'posterior distribution' reflects their uncertainty. Estimation of the posterior distribution involves both the data y but also any *a priori* parameter information expressed in terms of 'prior distributions'. The joint posterior distribution is given by

$$p(\boldsymbol{\phi}|\mathbf{y})=\frac{l(\mathbf{y}|\boldsymbol{\phi})\pi(\boldsymbol{\phi})}{\int_{\boldsymbol{\phi}}l(\mathbf{y}|\boldsymbol{\varphi})\pi(\boldsymbol{\phi})} \tag{6}$$

where $l(y|\backslash\boldsymbol{\phi})$ is the model likelihood. In practise, the integral in the denominator of Equation (6) is often intractable.

Markov chain Monte Carlo (MCMC) methods (Gilks *et al.* 1996; Gamerman 1997) provide a way to numerically simulate samples from $p(\boldsymbol{\phi}|\boldsymbol{y})$ without having to explicitly evaluate the denominator in Equation (6). A commonly used MCMC technique is Gibbs sampling (Gilks *et al.* 1996), utilised in the freely available WinBUGS software (Spiegelhalter *et al.* 2003). It is a method which directly samples from the conditional posterior distribution of each parameter. Both models described in this paper were implemented in WinBUGS.

In Bayesian statistical modelling, inference and point estimates are based mainly on the posterior distribution which directly reflects all uncertainty about each parameter. The posterior means or medians can be used as point estimates whereas an interval estimate (termed a credible interval in the Bayesian context) is simply derived from posterior quantiles. Parameter significance can be assessed, for instance, from a 95% credible interval by examining whether zero is included in the interval. If not, then the probability of the parameter taking the value zero is less than 0.05. This mirrors the p value of a classical significance test of the null hypothesis that the parameter is equal to zero.

Furthermore, a well-established measure for model comparison is the deviance information criterion (DIC), which is an estimate for the expected mean squared predictive error. A model with the lowest DIC is the one with the best out-of-sample predictive power (Gelman *et al.* 2004). It is given by $\mathrm{DIC} = 2\hat{D} - \tilde{D}$, where $\hat{D}$ is the mean of the posterior distribution for the deviance and $\tilde{D}$ is the deviance evaluated using posterior means of parameters. Deviance here is calculated as minus twice the log-likelihood and it is an important model comparison tool in statistics:

$$D = -2\log(L(\mathbf{y}|\boldsymbol{\theta})) \tag{7}$$

where θ are the model parameters and $\boldsymbol{y}$ is the data and $L(y|\theta)$ is the model likelihood. For the NHPP model $L(\boldsymbol{y}|\boldsymbol{\theta})$ is $L_{\mathrm{PP}}(\boldsymbol{\tau_i}, \boldsymbol{x_i}|\theta_i)$ from Equation (2) and for the ZINHPP $\log(L(\boldsymbol{y}|\boldsymbol{\theta}))$ is $l(\boldsymbol{\tau}, \boldsymbol{x}; \boldsymbol{p}|\boldsymbol{\theta})$ from Equation (5). There is a connection of the deviance to the Kullback–Leibler information measure (Gelman *et al.* 2004), which measures the discrepancy between the model distribution and the true distribution of the data. The DIC can also be viewed as the deviance, penalised for the number of model parameters which makes it the Bayesian equivalent of the Akaike information criterion (AIC) (Pawitan 2001). The penalty is based on the fact that more model parameters lead to a better model fit with calibration data but may result in over-fitting, i.e. poor model prediction performance on validation (i.e. unseen data).

CASE STUDIES

Data description

Two real-life datasets were considered. The first is a group of 1,349 pipes forming part of a large North American water system. Failure information is given in Table 1. However, confidentiality prohibits presenting any geographical information. The 'earliest failure on record' refers to the year with the earliest recorded pipe failure and the 'latest installed pipe' refers to the year with the most recently installed pipe. All pipes are made of cast iron and are of the same diameter, and the only explanatory variable available is pipe length in metres. Pipe length is an influential variable for pipe failures (Hall *et al.* 2006; Boxall *et al.* 2007; Kleiner & Rajani 2007; Berardi *et al.* 2008). However, many other important variables are missing here such as system pressures, loadings and water hammer events, soil type and moisture content, pipe and bedding class, soil properties, method of manufacture, etc. Pipe-specific random effects (e.g. θ_i in Equation (3)) allow for such unobserved variables which vary between pipes. Hence they can absorb the extra variability and allow the underlying model (NHPP) to explain the underlying behaviour of the failure rate. Failure

Table 1 | Pipe data description

	North America	New Zealand
Number of pipes	1349	532
Total number of failures	5425	175
Earliest failure on record	1962	1990
Latest failure on record	2003	2001
Earliest installed pipe	1945	1930
Latest installed pipe	1960	1983

(pipe burst) times are given in months and all pipes were observed from 1962 until the end of 2003. All pipes were installed before 1960 so the data are left-truncated.

The second dataset is taken from Watson (2005) and refers to the Howick pressure zone in Manukau City, Auckland, New Zealand. The available data consist of 532 asbestos cement pipes with 175 yearly recorded failures (bursts) in the 11 year period 1990–2001 (Table 1). Note that these data are also left-truncated. It is a thinned dataset in the sense that the author reduced the data by choosing only asbestos cement pipes and by removing pipes less than 5 m in length but also pipes with unknown installation dates.

The available explanatory variables for the New Zealand data (NZD) are pipe length (m), pipe diameter (mm), maximum absolute pressure (i.e. night time pressure) and pressure change (between maximum and minimum 24 h recorded pressure). Pressure is given as a three-level (1 = low, 2 = medium, 3 = high) categorical variable and pressure change as a two-level variable (1 = low, 2 = medium). The diameter variable is defined for specific values, i.e. 25, 40, 50, 100, 150, 200 and 300 mm.

Failure count histograms (Figure 1) show that both distributions are heavily skewed with considerable mass around zero, especially for the NZD. 22% of pipes in the North American data (NAD) have never failed compared to 85% in the NZD, meaning that the proportion of pipes that have never failed is much larger for the NZD. In that respect, it can be argued that the NZD is more compatible with datasets commonly found in real water distribution systems. However, even though a large number of pipes have experienced failure in the NAD, most of the failures were located on a small proportion of pipes (see Figure 1). Based on this, a zero inflation model may be useful for both datasets.

A good indicative measure for pipe failure rate is the average number of failures per unit time (Figure 2). The rate varies considerably between pipes for both datasets. In the rest of this section, the NHPP and ZINHPP models are applied to the NAD and NZD, and results are compared.

NHPP and ZINHPP models

For each dataset, a power-law failure rate is assumed for each pipe i, as in Equation (3). The effect from the explanatory variables in the NAD is given by $\exp\{\beta_0 + \beta_1 \times 1\}$, where $x_1 =$ length and for the NZD by $\exp\{\beta_0 + \beta_1 \times 1 + \beta_2 \times 2 + \beta_3 I_{(x_3=2)} + \beta_4 I_{(x_3=3)} + \beta_5 I_{(x_4=2)}\}$, where $x_1 =$ length, $x_2 =$ diameter, $x_3 =$ pressure and $x_4 =$ pressure change. Recall that $I_{(\text{event})}$ is an indicator function so, for instance, the parameter β_3 is the difference in the failure rate between $x_3 = 2$ and $x_3 = 1$ whose value is incorporated in the intercept β_0.

This formulation of the rate, with a pipe specific 'shape' parameter θ_i for each pipe is a versatile model allowing for each pipe in the system to behave uniquely. The parameters for each model are the intercept β_0, the associated coefficients β_j of the explanatory variables, and the shape and rate parameters α and κ of the model for θ_i. The parameters

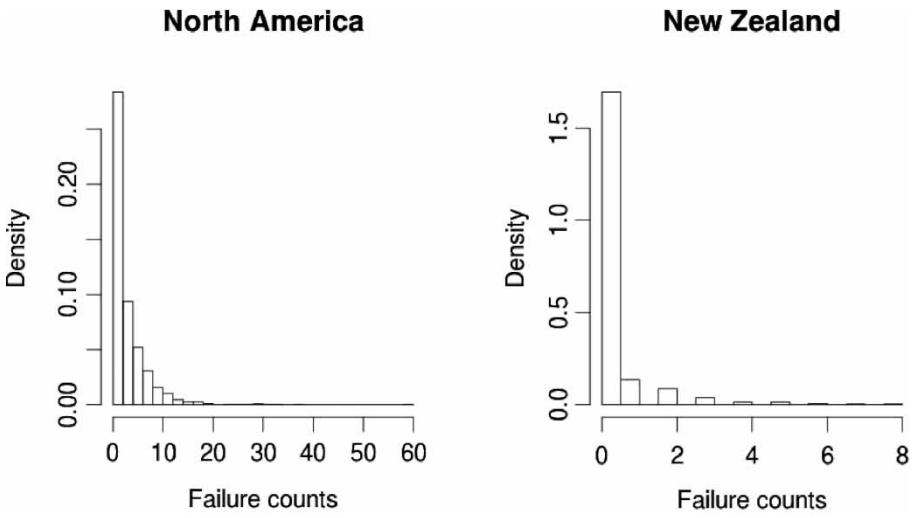

Figure 1 | Failure counts histogram for each dataset.

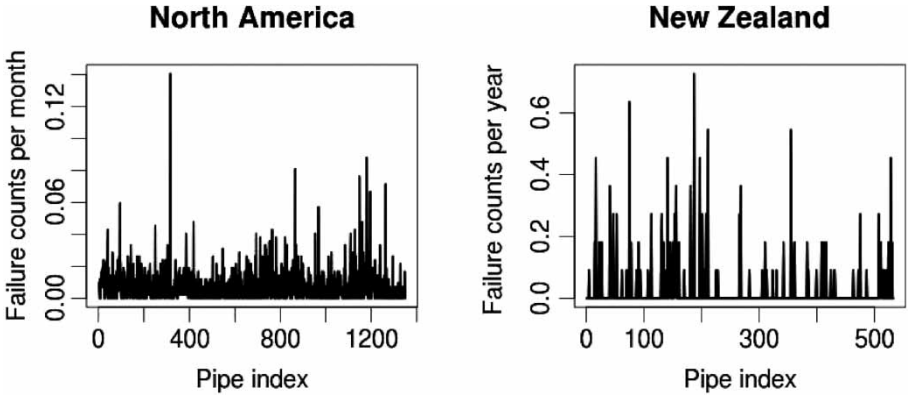

Figure 2 | Empirical failure rate for each pipe.

of the ZINHPP model are the same with the addition of the zero-inflation probabilities p_i for each pipe.

Earlier, the quantity $(1–p_i)$ was referred to as the tendency of a pipe to resist to failure. Therefore, it may be sensible to model p_i in terms of pipe age such that

$$\log\left(\frac{p_i}{1-p_i}\right) = \gamma_{i,0} + \gamma_i \text{Age}_i \tag{8}$$

where Age_i is the pipe age at the end of the observation period. Note that $\gamma_{i,\,0}$ are pipe-specific but are not random effects. The parameters to be estimated in the ZINHPP model are β_0, each β_j, α, κ, each $\gamma_{i,0}$ and γ_1.

Results

The observation period in the NAD is quite long so to assess model performance for more realistic situations with large periods of left-truncation, only data after 1969 were used. Also, no data were included after 1998 so models were calibrated for the period 1969–1998 and then used to predict failures in the validation period 1999–2003. The observation period in the NZD is much shorter so no validation was considered (the same approach was followed in Watson (2005)).

For the NAD, parameter estimates for each model are given in Table 2 which also shows the assumed prior distributions, standard errors (s.e.), 95% credible intervals (Cr.I.) and the DIC. Estimates in the analysis were always taken as means of posterior distributions.

All parameters are significant since zero is not included in the 95% Cr.I. but particular interest lies on β_1 relating to the effect of pipe length where the estimate for both models is positive. That means that the effect of length on the failure rate is scaling by $\exp(0.004) = 1.004$ for every metre of length. Significance is also evident for parameter γ_1 whose estimate is positive, meaning that the effect of age is to increase the odds $p_i/(1–p_i)$. In other words, an older pipe

Table 2 | Parameter estimates (NAD)

Model	Parameter	Prior	Estimate (s.e.)	95% Cr.I.	DIC
NHPP	β_0	Norm(0, 1000)	0.33 (0.027)	[0.28, 0.38]	Calibration: −3028
	β_1	Norm(0, 1000)	0.004 (0.0001)	[0.004, 0.005]	
	α	Gam(0.5, 0.005)	41.1 (12.8)	[25.0, 81.0]	Validation: −3192
	κ	Gam(0.5, 0.005)	38.8 (12.4)	[23.2, 77.9]	
ZINHPP	β_0	Norm(0, 1000)	0.64 (0.03)	[0.58, 0.70]	Calibration: −1544
	β_1	Norm(0, 1000)	0.004 (0.0001)	[0.003, 0.005]	
	α	Gam(0.5, 0.005)	42.8 (11.0)	[25.1, 70.0]	
	κ	Gam(0.5, 0.005)	40.5 (10.0)	[23.3, 66.5]	Validation: −1612
	γ_1	Norm(0, 1000)	19.9 (1.8)	[16.7, 23.6]	

has less (extra) resistance to failure. The NHPP has a smaller DIC for both calibration and validation periods, meaning that the ZINHPP is not a better model in that respect.

Observed (NAD) and predicted cumulative number of failures are plotted for each model for two selected pipes in Figure 3, together with 95% prediction intervals. These figures are plotted using a monthly time step. Note that the vertical line represents the validation period start. The intervals have a 'staircase' shape since the predictive distributions are discrete. An example of how zero inflation is useful is shown in the left-hand panel plots, which refer to a long pipe (pipe X) that has never failed during 1969–1998. The NHPP model clearly overestimates failures driven by the variable length whereas the ZIHNPP accounted for that through the parameter p_i. However, for pipes that did fail in 1969–1998 (i.e. the majority) such plots are similar for both models, as in the ones in the right-hand panel (pipe Y).

Prediction intervals encompass both parameter uncertainty and model variability, unlike confidence intervals which reflect only the former or Monte Carlo integration (Alvisi & Franchini 2010) which reflects only the latter. The intervals in Figure 3 may be too wide for some practical applications but here the goal was to improve the performance of the conventional NHPP model and obtain more realistic prediction intervals. Yearly number of failures for the whole NAD pipe group (Figure 4, top panel) show little difference between the two models, so zero inflation has little effect on the total failure count which is adequately captured by the NHPP except for the last few years. The data exhibit a slow increase in failure numbers followed by a sharp drop after 1995 which the models do not capture. This is because of the monotonic characterisation of the failure rate in both NHPP and ZINHPP which acts as a restriction in capturing such behaviour. No available physical explanation for this sharp drop exists, due to the lack of information about the dataset. Possible reasons include rehabilitation performed on the network or a rise in warm winters over the last decade.

A plot of the total cumulative failures against time (Figure 4, bottom panel) also shows little difference between the two models.

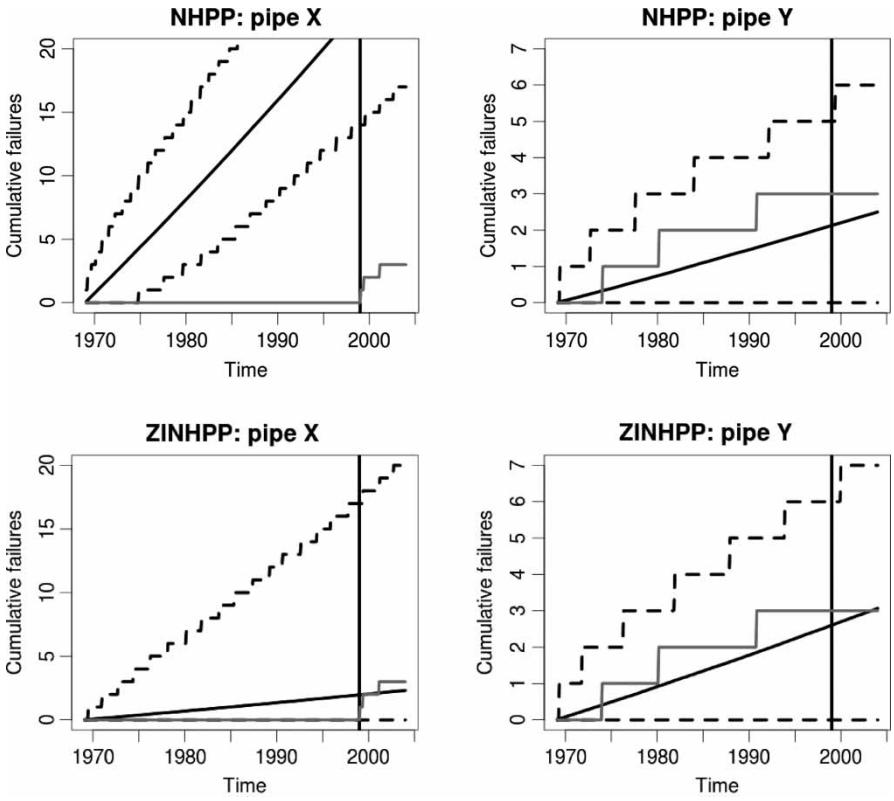

Figure 3 | NAD cumulative failures. Grey: observed, black solid: predicted, black dashed: 95% prediction intervals, black vertical: validation period start.

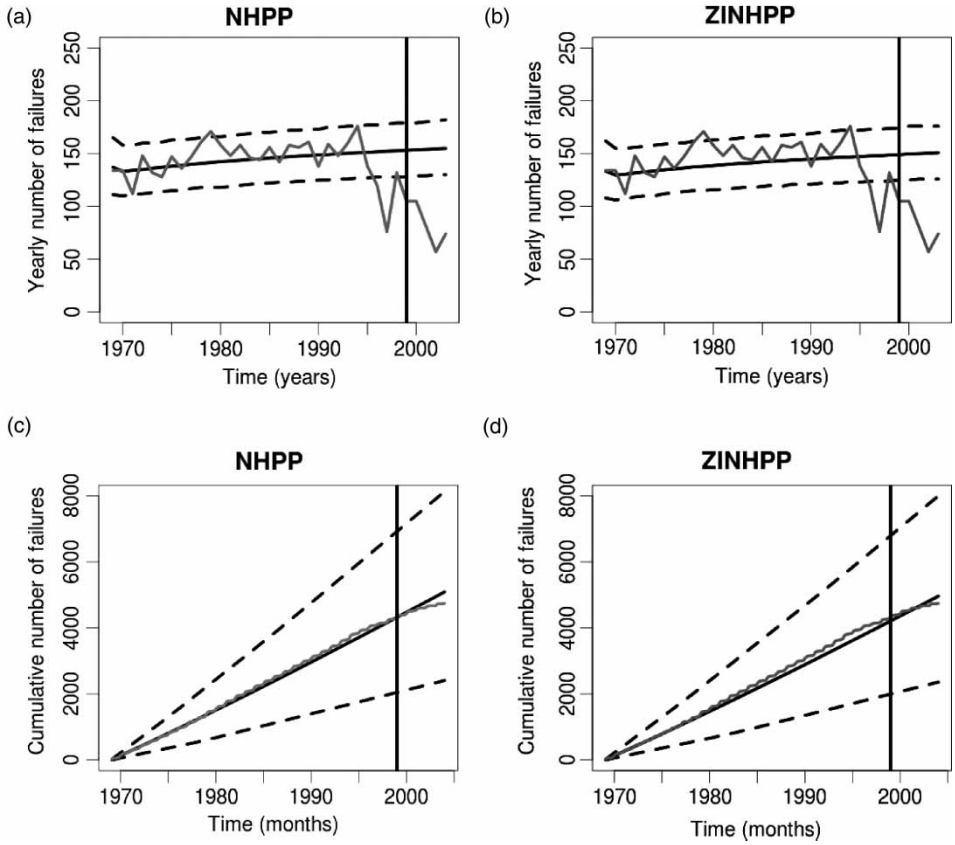

Figure 4 | NAD total and cumulative failures. Grey: observed, black solid: predicted, black dashed: 95% prediction intervals, black vertical: validation period start.

Table 3 | Parameter estimates (NZD)

Model	Parameter	Prior	Estimate (s.e.)	95% Cr.I.	DIC
NHPP	β_0	Norm(0, 1000)	−9.8 (1.6)	[−13.0, −6.9]	515
	β_1	Norm(0, 1000)	0.009 (0.001)	[0.007, 0.001]	
	β_2	Norm(0, 1000)	0.007 (0.003)	[0.0007, 0.01]	
	β_3	Norm(0, 1000)	1.0 (0.7)	[−0.4, 2.7]	
	β_4	Norm(0, 1000)	1.3 (0.9)	[−0.3, 3.3]	
	β_5	Norm(0, 1000)	1.0 (0.3)	[0.4, 1.7]	
	α	Gam(0.5, 0.005)	18.7 (8.1)	[8.2, 38.7]	
	κ	Gam(0.5, 0.005)	13.3 (3.5)	[7.8, 21.8]	
ZINHPP	β_0	Norm(0, 1000)	−2.6 (0.6)	[−3.7, −1.5]	344
	β_1	Norm(0, 1000)	0.005 (0.0005)	[0.004, 0.006]	
	β_2	Norm(0, 1000)	0.009 (0.002)	[0.004, 0.01]	
	β_3	Norm(0, 1000)	0.47 (0.6)	[−0.5, 1.6]	
	β_4	Norm(0, 1000)	0.19 (0.6)	[−0.9, 1.5]	
	β_5	Norm(0, 1000)	0.58 (0.2)	[0.2, 1.0]	
	α	Gam(0.5, 0.005)	11.2 (7.2)	[3.1, 29.8]	
	κ	Gam(0.5, 0.005)	6.2 (4.3)	[1.4, 17.4]	
	γ_1	Norm(0, 1000)	−1.7 (1.9)	[−5.2, 2.4]	

Parameter estimates for the NZD are provided in Table 3. Parameters β_1 and β_2 of the NHPP are positive and significant, which implies that increasing length and diameter give higher failure rate. Empirically, diameter often has a negative effect on rate, but the finding here was also noted by Watson (2005). Parameters β_3 and β_4 are not significant, meaning that there are no differences in the rate for the different levels of pressure. However, β_5 is significant, suggesting that the rate for medium pressure change is significantly more than for low pressure change. Estimates and inference for the ZINHPP model are similar but slightly different in value. Interestingly, γ_1 is not significant so pipe age does not affect zero inflation here. The DIC is lower for the ZINHPP so for this dataset it fits better than the NHPP.

Typical cumulative failure plots for the NZD (based on a yearly time step) are shown in Figure 5 for two selected pipes. As in the NAD, the NHPP overestimates both zero-failing pipes (left-hand panel) and underestimates pipes that did fail (right-hand panel). This is particularly true for pipes where the explanatory variables have values implying a high estimated failure rate and a low observed failure count (and vice versa). The ZINHPP model has the flexibility of adjusting the rate of failures accordingly in these situations, through parameter p_i. Again, there is little difference between the total (and cumulative) yearly failure count plots (Figure 6). However, the models fit the data better than in the case of NAD. There is no 'abnormal' behaviour in the NZD such as the apparent drop in NAD failure counts.

Discussion

The ZINHPP model fits the NZD better in terms of the DIC. For the NAD, the ZINHPP has a higher DIC than the NHPP, meaning that the models have effectively the same fit. The reason the DIC is higher is because the ZINHPP has more parameters and the DIC penalises for the number of parameters. There are two possible reasons for the better fit of the ZINHHP to the NZD. First, the NZD has many more pipes that never failed, hence zero inflation

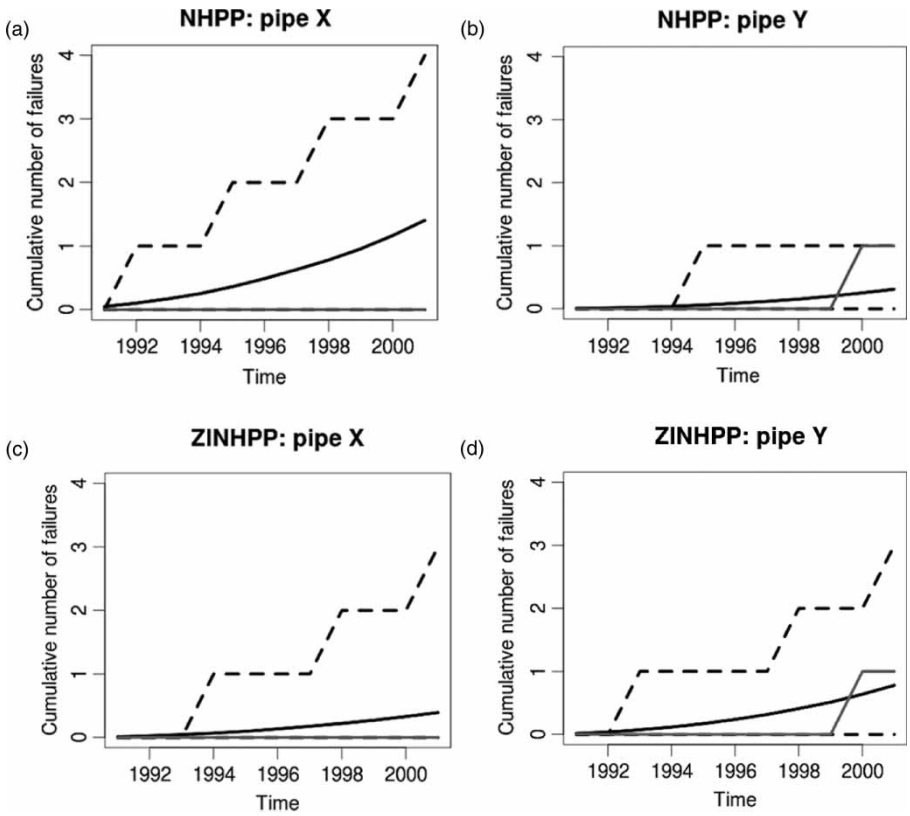

Figure 5 | NZD cumulative failures. Grey: observed, black solid: predicted, black dashed: 95% prediction intervals.

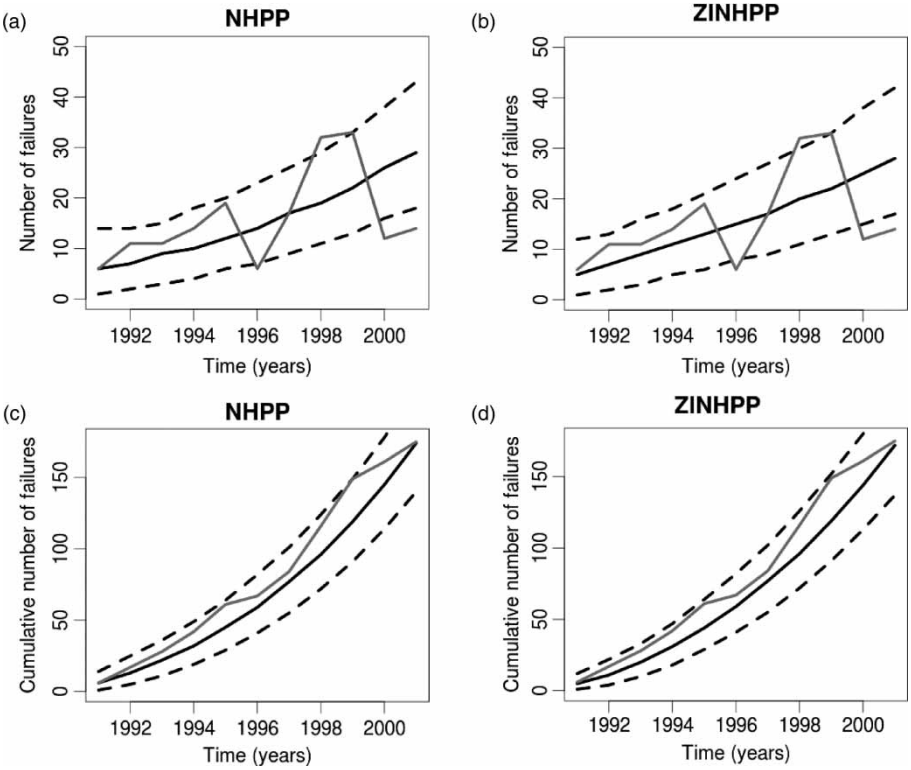

Figure 6 | NZD total and cumulative failures. Grey: observed, black solid: predicted, black dashed: 95% prediction intervals.

works better. Second, the NZD has more explanatory variables driving the failure rate but each variable has a fixed parameter. Those fixed parameters will capture the 'global' relationship of variables and failure rate. However, some pipes that deviate from this global relationship will not be appropriately modelled. The ZINHPP can allow for such pipes by adjusting the behaviour of the failure rate through parameters p_i. As for the NAD, there is only one explanatory variable so the chance of a pipe deviating from this global relationship is smaller.

The total number of failure plots is the same for both models in each dataset. This is because in both models the total failure counts are Poisson. The ZINHPP improves the fit on individual pipes while, on aggregate, it behaves the same as the NHPP.

CONCLUSIONS

A pipe-specific model for the prediction of failures (i.e. bursts) was developed. The conventional NHPP was used to characterise the underlying failure rate with a power-law model, in conjunction with random effects which accounted for pipe heterogeneity due to unobserved variables. A zero-inflated NHPP model was then developed to allow for an excess number of zeros in failure counts, a common phenomenon in pipe data.

A particular issue underlying this study is that of data quality and quantity in pipe datasets. Real systems have complex failure histories and often very little is known/measured about these. Realistically, expectations about statistical model performance on such data is generally low and it is hoped that the models presented in this paper offer an improvement. The ZINHPP model shows its utility, particularly in the quite frequent case where many pipes do not present any failure during the observation period, i.e. when the percentage of zeros is very high.

In addition, zero inflation, as defined in this paper, acts at the pipe level (zero inflation is pipe-specific) but it is constant in time. Thus, the probability of extra zeros is stationary, therefore the effect of zero inflation

on the failure rate is constant scaling in time. In terms of a conceptual plot of the failure rate (curve) against time, the impact of zero inflation is manifested as a simple shift of the curve. Future work should allow for time-dependent zero inflation.

ACKNOWLEDGEMENT

The North American pipe dataset used in this paper was provided by Dr Yehuda Kleiner whom the authors gratefully acknowledge.

REFERENCES

Alvisi, S. & Franchini, M. 2010 Comparative analysis of two probabilistic pipe breakage models applied to a real water distribution system. *Civil Engng Environ. Syst.* **27** (1), 1–22.

Berardi, L., Kapelan, Z., Giustolisi, O. & Savic, D. A. 2008 Development of pipe deterioration models for water distribution systems using epr. *J. Hydroinf.* **10** (2), 113.

Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J. & Unwin, D. M. 2007 Estimation of burst rates in water distribution mains. *Wat. Mngmnt.* **160**, 73–82.

Constantine, A. G. & Darroch, J. 1993 *Pipeline Reliability. Stochastic Models in Engineering Technology and Management*. World Scientific, Singapore.

Constantine, A. G., Darroch, J. & Miller, R. 1996 Predicting underground pipe failure. *Journal of Australian Water Association* **23** (2), 9–10.

Cook, R. J. & Lawless, J. F. 2007 *The Statistical Analysis of Recurrent Events*. Springer, Berlin.

Economou, T., Kapelan, Z. & Bailey, T. C. 2008 A zero-inated Bayesian model for the prediction of water pipe bursts. In: *Proc. 10th International Water Distribution System Analysis Conf.*, Kruger National Park, South Africa. ASCE, New York.

Engelhardt, M. O., Skipworth, P. J., Savic, D. A., Saul, A. J. & Walters, G. A. 2000 Rehabilitation strategies for water distribution networks: a literature review with a UK perspective. *Urban Wat.* **2** (2), 153–170.

Faraway, J. J. 2006 *Extending the Linear Model with R*. Chapman and Hall, London.

Gamerman, D. 1997 *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London.

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

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2004 *Bayesian Data Analysis*. Chapman and Hall/CRC, London.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. 1996 *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Goulter, I., Davidson, J. & Jacobs, P. 1993 Predicting water-main breakage rates. *J. Wat. Res. Plann. Mngmnt* **119** (4), 419–436.

Hall, M., Kapelan, Z., Long, R. & Savic, D. 2006 Deterioration Rates of Sewers. Technical report, UKWIR Rep. No. 06/RG/05/15, 97.

Kleiner, Y., Nafi, A. & Rajani, B. 2010 Planning renewal of water mains while considering deterioration, economies of scale and adjacent infrastructure. *Wat. Sci. Technol: Wat. Supply* **10** (6), 897–906.

Kleiner, Y. & Rajani, B. 2007 Static and dynamic effects in prioritizing individual water mains for renweal. In: *Water Management Challenges in Global Change*. CRC Press, Boca Raton, FL, pp. 61–68.

Kleiner, Y. & Rajani, B. 2008 Prioritising individual water mains for renewal. In: *Proc. World Environmental and Water Resources Congr.*, Honolulu. ASCE, New York, pp. 1–10.

Kleiner, Y. & Rajani, B. 2010 Dynamic Influences on the Deterioration Rates of Individual Water Mains (i-warp). Technical report, Water Research Foundation, Denver, CO.

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

Lambert, D. 1992 Zero-inated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** (1), 1–14.

Loganathan, G. V., Park, S. & Sherali, H. D. 2002 Threshold break rate for pipeline replacement in water distribution systems. *J. Wat. Res. Plann. Mngmnt* **128** (4), 271–279.

Park, S., Jun, H., Kim, B. J. & Im, G. C. 2008 Modeling of water main failure rates using the log-linear rocof and the power law process. *Wat. Res. Mngmnt* **22**, 1311–1324.

Pawitan, Y. 2001 *In all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications, Oxford.

Rigdon, S. E. & Basu, A. P. 2000 *Statistical Methods for the Reliability of Repairable Systems*. John Wiley & Sons, New York.

Saldanha, P. L. C., de Simone, E. A. & e Melo, P. F. F. 2001 An application of nonhomogeneous Poisson point processes to the reliability analysis of service water pumps. *Nucl. Engng Design* **210**, 125–133.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Lunn, D. 2003 *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit. Available from: http://www.mrc-bsu.cam.ac.uk/bugs.

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

Watson, T.G. 2005 *A Hierarchical Bayesian Model and Simulation Software for the Maintenance of Pipe Networks*. PhD thesis, Department of Civil and Resource Engineering, University of Auckland.

First received 23 November 2010; accepted in revised form 2 February 2012. Available online 12 June 2012