An Aggregated Hierarchical Bayesian Model for the Prediction of Pipe Failures

T. Economou[1], Z. Kapelan [2] and T. Bailey[1]

[1]University of Exeter, Mathematics Research Institute, Harrison Building, North Park Road, Exeter EX4 4QF, United Kingdom
[2] University of Exeter, Centre for Water Systems, Harrison Building, North Park Road, Exeter EX4 4QF, United Kingdom

Abstract

The lack in quality and quantity in the data for underground water pipes has been a major obstacle in modelling and predicting pipe bursts. Motivated by this insufficiency of data and in particular the failure to record actual pipe burst times, an aggregated Bayesian non-homogeneous Poisson process (NHPP) model is proposed. The model simply utilises the age of the pipes and the number of bursts within an observation period in order to capture the deterioration process, without the need for individual times of each burst. Modifications are also made to account for right censoring in the data as well as left truncation as it is often the case that a pipe is observed long after it has been laid down. Actual failure data from an underground water pipe network in New Zealand is used to demonstrate the ability of the aggregated model to perform as well as conventional NHPP models. Implementation of the models utilises sampling techniques and in particular Bayesian Markov-chain Monte Carlo (MCMC) methods.

Keywords: NHPP; Bayesian Inference; MCMC; Aggregated data; asset management; water distribution system.

1 Introduction

Forecasting pipe burst behaviour in water distribution systems is essential in terms of scheduling replacements and repairs, and in planning associated budgets. Nevertheless the processes which give rise to pipe failures are complex and affected by a large variety of factors both measurable and immeasurable therefore by simply considering the age of the pipe to explain the deterioration in the pipes is not sufficient (Boxall et al., 2004). In addition, most existing pipe burst data are poor in the sense that it is only of lately that water companies have started recording at least some information on pipe failures resulting in data sets containing burst information on only a small percentage of the lifetime of the pipes (Gat and Eisenbeis, 2000). This results in data sets being left truncated, with respect to time.

Various modelling approaches have been used so far for prediction, depending upon the failures of interest, the nature and complexity of the network and the availability, scope and reliability of relevant data (Kleiner and Rajani, 2001). A repairable system is one that retains the reliability level that it had before a repair. Therefore one approach is to consider the network or even the pipes themselves as repairable systems and thus use models drawn from reliability theory which treat the occurrence of failures as a non-homogeneous Poisson process (NHPP), i.e. a Poisson process with time varying failure rate. This approach has the advantage of explicitly incorporating and characterizing the non-linear relationship between failure rate and pipe deterioration with age as well as allowing for the inclusion of other covariates. Recently, Watson (2005) has demonstrated how this approach in conjunction with Bayesian MCMC methods can be used effectively to develop a pipe replacement policy.

However, one drawback of the NHPP models is that they require detailed data on actual failures times in order to estimate the shape of the underlying reliability curve. In practice, at least in the UK, this level of detail may not be available since many water companies record only total numbers of failures in different parts of the network over time periods of numbers of years, rather than actual failure times resulting in data being aggregated over the observation period. This paper proposes a modified NHPP model which only makes use of numbers of failures in an observed time period and the age of each pipe at the end of this period, but is still able to capture the age deterioration phase of the reliability curve and provide predictive results comparable to those obtained from the conventional NHPP model which requires both numbers and times of failures. The commonly employed power law process (Lee and Lee, 1978; Landers et al., 2001; Sen, 2002) in conjunction with the proportional hazards model (Cox, 1972) is adopted for the intensity of the NHPP.

In section 2 the NHPP model and its aggregated version are described. The likelihood of both models is derived which accounts for right censoring as well as left truncation. The models are tested, in section 3, on some real data of pipe bursts in Manukau City, Auckland, New Zealand, using Bayesian inference and in particular MCMC methods. The implementation of this type of models within the Bayesian framework is quite novice in the water pipe field with the most recent example being a PhD thesis by Watson, 2005. The reason why the Bayesian inference is potentially superior is that it does not only allow the consideration of each pipe individually but also the inclusion of random effects. Section 4 is composed of conclusions and further possible extensions and improvements to the model.

2. Model Specification

2.1 Basic NHPP Model

Suppose pipe i is observed in the time period $[0,T_i]$, n_i represents the number of failures in that time period and $t_{i1}, t_{i2}, \ldots, t_{in_i}$ denote the times of each failure. Assuming these failures occur as a

NHPP with intensity function $\lambda(t_{ij})$, $j = 1,2,\ldots n_i$, then in the conventional reliability context one would normally consider the (time truncated) likelihood function

$$L(\cdot) = \left[\prod_{j=1}^{n_i} \lambda(t_{ij})\right] \exp\left\{-\int_0^{T_i} \lambda(y)dy\right\} = \left[\prod_{j=1}^{n_i} \lambda(t_{ij})\right] \exp\{-\Lambda([0,T_i))\} \qquad (1)$$

which is the joint probability density function (PDF) of the failure times $f(t_{i1},\ldots,t_{in_i})$ where $\Lambda((0,T_i]) = \int_0^{T_i} \lambda(t)dt$ is the expected number of failures in [0,T$_i$) (Meeker and Escobar, 1998). This model is denoted here as Model 1. A wide range of models exist which can be used to represent the intensity function $\lambda(t_{ij})$ (Kleiner and Rajani, 2001). Here we adopt a formulation based on the proportional hazards model $\lambda(t_{ij}) = \lambda_0(t_{ij})e^{\boldsymbol{\beta}\mathbf{x}_i}$ where $\boldsymbol{\beta}\mathbf{x}_i$ is a linear function of possible suitable pipe covariates $\mathbf{x}_i = (x_1, x_2,..., x_k)$ with associated parameters $\boldsymbol{\beta} = (\beta_1, \beta_2,..., \beta_k)$ and where the baseline hazard is parameterised as:

$$\lambda_0(t_{ij}) = \gamma_i \vartheta_i (t_{ij})^{\vartheta_i - 1};\ \gamma_i, \vartheta_i > 0\,.$$

which is sometimes referred to as the power law. An important feature of this model is that the shape parameter, ϑ, can represent both a deteriorating system ($\vartheta > 1$) and an improving system ($\vartheta < 1$). Setting $\gamma_i = e^{\beta_{0i}}$ the overall intensity function is

$$\lambda(t_{ij}) = \vartheta_i (t_{ij})^{\vartheta_i - 1} e^{\boldsymbol{\beta}_i \mathbf{x}_i}$$

where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_1,..., \beta_k)$.

2.2 Aggregated NHPP Model

An important property of the assumption made above, that bursts occur as a NHPP with intensity function $\lambda(t)$, is that the number of failures, $F(t)$, in any time interval $[t_1, t_2]$ follow a Poisson distribution with mean $\int_{t_1}^{t_2} \lambda(t)dt = \Lambda([t_1, t_2])$ (Rigdon and Basu, 2000), i.e.

$$\Pr(F(t_1) - F(t_2) = n) = \frac{e^{-\Lambda([t_1,t_2])}\Lambda([t_1, t_2])}{n!} \qquad (2)$$

As mentioned previously, the likelihood given by (1) involves both the number of failures, n_i and the individual failure times t_{ij}. Suppose, however, that only n_i and not t_{ij} are available. Then using (2), we see that $n_i \sim \text{Poisson}(\Lambda([0,T_i)))$ which allows direct use of the Poisson likelihood when dealing with aggregated data (i.e. data on number of failures and length of period of observation only). We are making the assumption here that the observation period starts at time zero (i.e. the installation time of the pipe) which is rarely the case since typically we will only have failure information for a few recent years and the pipe will usually have been installed long before the start of that period. So if we suppose that we start observing pipe i at $t_{0i} > 0$, then $n_i \sim \text{Poisson}(\Lambda([t_{0i}, T_i)))$. Assuming that we have N pipes and that these pipes are independent, then the overall likelihood for the data on all pipes is:

$$L(\cdot) = \prod_{i=1}^{N} \frac{\exp\{-\Lambda([t_{0i}, T_i))\}[\Lambda([t_{0i}, T_i))]^{n_i}}{n_i!}$$

and since

$$\Lambda([t_{0i}, T_i)) = \int_{t_{0i}}^{T_i} \lambda(t_{ij}) dt_{ij} = \int_{t_{0i}}^{T_i} \vartheta_i t_{ij}^{\vartheta_i - 1} e^{\boldsymbol{\beta}_i \mathbf{x}_i} dt_{ij} = \left[T_i^{\vartheta_i} - t_{0i}^{\vartheta_i}\right] e^{\boldsymbol{\beta}_i \mathbf{x}_i} \qquad (3)$$

we have

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}_i) = \prod_{i=1}^{N} \left(\frac{1}{n_i!}\right) \exp\left\{-\left[T_i^{\vartheta_i} - t_{0i}^{\vartheta_i}\right] e^{\boldsymbol{\beta}_i \mathbf{x}_i}\right\} \left(\left[T_i^{\vartheta_i} - t_{0i}^{\vartheta_i}\right] e^{\boldsymbol{\beta}_i \mathbf{x}_i}\right)^{n_i}$$

Hence the Poisson likelihood can be used to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}_i$ without the need of t_{ij}. This model is denoted here as Model 2.

Going back to (1) we can see that in the case that the likelihood needs to be modified for left truncation as well as the case when $n_i = 0$. Therefore we introduce the binary variable δ_i where

$$\begin{aligned} \delta_i = 0 \quad & \text{if } n_i = 0 \\ 1 \quad & \text{if } n_i > 0 \end{aligned}$$

Using δ_i and (3) the likelihood in (1) now becomes

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}_i) = \left[\prod_{j=1}^{n_i} \vartheta_i t_{ij}^{\vartheta_i - 1} e^{\boldsymbol{\beta}_i \mathbf{x}_i}\right]^{\delta_i} \exp\left\{-\left[T_i^{\vartheta_i} - t_{0i}^{\vartheta_i}\right] e^{\boldsymbol{\beta}_i \mathbf{x}_i}\right\}$$

3. Model Application

In order to assess the performance of Model 2 in relation to Model 1, both were fitted to the Howick Pressure Zone data in Manukau city, Auckland, New Zealand (Watson, 2005) which consists of 532 asbestos cement pipes with 175 recorded failures in the period 1990-2001. Only 81 out of the 532 pipes have reported bursts resulting in 451 pipes that have zero failures. Both Model 1 and Model 2 were implemented in the Bayesian framework using MCMC and in particular Gibbs sampling. These were fitted in WinBUGS (Spiegelhalter et al., 1999) for the first 9 years worth of data and were used to predict the number of failures in each pipe for the remaining two years of the observation period. Two parallel Markov chains were run for each of the models, taking one sample in every 25 iterations collecting a total of 5000 samples from each chain. This was enough to ensure good convergence and rate of mixing for each parameter. Covariates in the model include pipe length, pipe diameter, pressure and absolute pressure. A gamma prior was used for each of the ϑ_i's whose parameters were also given diffuse gamma hyperpriors. Furthermore $\beta_{0i}, \beta_1, ..., \beta_k$ were each assumed to have a normally distributed prior with zero mean and large variance.

The probability that a pipe will fail in those two years was extracted from the samples used to build the posterior predictive distributions of the number of failures in each pipe. Specifically, the probability was calculated by dividing the total number of non-zero elements in each sample by the total number of the elements in the sample. These probabilities were then used to perform a Bernoulli trial for each pipe thus deciding whether a pipe will fail or not. Confusion matrices were then constructed, comparing whether each pipe actually failed in the two years with the predicted cases of failures. For each model, 500 of these matrices were produced and averaged to ensure accuracy in the results.

Model 1	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe failures	Not failed	495.5	15.5	511
	Failed	18.3	2.7	21
	Total	513.8	18.2	

Model 2	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe failures	Not failed	496.6	14.4	511
	Failed	18.5	2.5	21
	Total	515.1	16.9	

Considering that Model 2 was fitted on the same but less informative data, the predictions from it match quite closely those of Model 1 suggesting that Model 2 is a good substitute of model 1 when data does not include times of failures. In addition, the aggregated model was still able to adequately capture the ageing process through the parameter ϑ_i although point estimates of these and the β_{0i}'s did differ between the two models. Coefficient estimates of the covariates were practically the same for the two models. The mean values of the posterior predictive distributions of the number of failures for each model are presented in Figure 1 below, one against the other.

4. Conclusions

In summary, although the data did not contain actual times of failures, the aggregated model was still able to adequately capture the ageing process in individual pipes (a key element of the NHPP), at least as good as Model 1. This fact could prove to be quite important when dealing with data sets of pipe bursts alone which is (at least in the UK) often the case. On the other hand predictions from both models were not quite satisfactory when compared to the actual ones and these could be due to the data being zero-inflated hence a possible extension to the models would be to account for the extra number of zeros in terms of the failures.

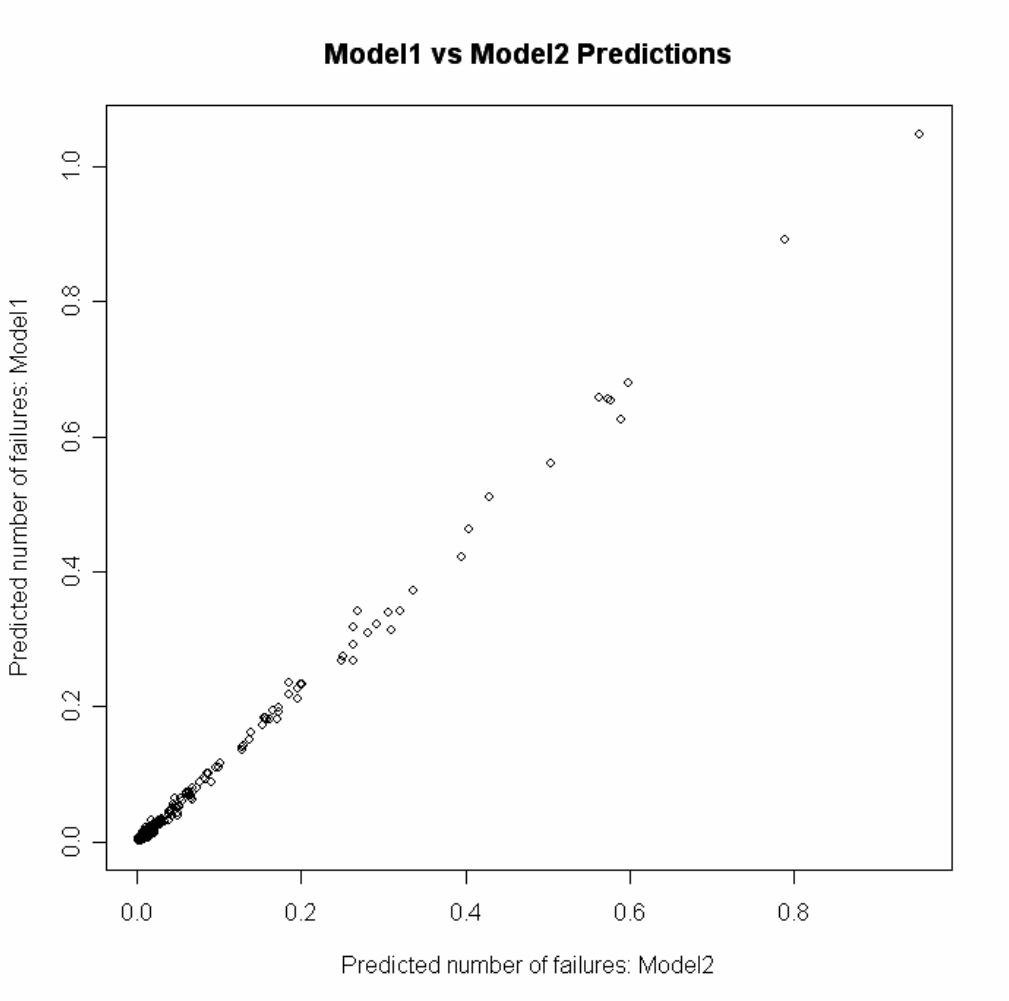

Figure 1

References

Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J. and Unwin, D. M. (2004), "Estimation of burst rates in water distribution mains", *Research Report No. 546/04*, Department of Probability and Statistics, University of Sheffield.

Cox, D. (1972), "Regression models and life-tables", *Journal of the Royal Statistical Society, Series B (Methodological)*. 34, 187-220.

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

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

Landers, T., Jiang, S. and Peek, J. (2001), "Semi-parametric pwp model robustness for log-linear increasing rates of accurrence of failures", *Reliability Engineering and System Safety*, 73, 145-153.

Lee, L. and Lee, S.K. (1978), “Some results on inference for the Weibull process”, *Technometrics*, 20 (1), 41-45.

Meeker, W. Q. and Escobar, L.A. (1998), “Statistical methods for reliability data”, New York, John Wiley and Sons, Inc.

Rigdon, S. E. and Basu, A. P. (2000), “Statistical methods for the reliability of repairable systems”, New York, John Wiley and Sons, Inc.

Sen, A. (2002), “Bayesian estimation and prediction of the intensity of the power law process”, *Journal of Statistical Computation and Simulation*, 72, 613-631.

Spiegelhalter, D., Thomas, A. and Best, N. (1999), “WinBUGS Version 1.2 user manual”, MRC Biostatistics Unit.

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, p. 238.