Bayesian modelling of time aggregated water pipe bursts with a zero-inflated, non-homogeneous Poisson process

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

[1] School of Engineering, Computer Science and Mathematics, University of Exeter, UK.

Abstract: A commonly used approach to modelling recurrent failures is based on a non-homogeneous Poisson process (NHPP) and requires data on actual failure times. Modelling and predicting bursts in underground water pipes is vital to water companies from both an economic and conservation perspective, but often does not allow for use of a conventional NHPP for two reasons. Firstly, because data is commonly only recorded on numbers of failures over a (relatively long) time period and not on exact failure times. Secondly, because failures are usually only observed in a very small proportion of pipes in the network. This paper proposes a model derived from the conventional NHPP 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. The model is then further extended to account for censoring and truncation in the data as well as an excess of zeros. Application of this 'aggregated' model and its zero-inflated extension are illustrated on a data set involving a network of 532 cement water pipes in Manukau City, Auckland, New Zealand.

Keywords: NHPP; Zero-inflated; Aggregated; Censoring; Truncation.

1 Introduction

Predicting pipe failures (i.e. bursts or blockages) in water distribution systems is important in terms of scheduling replacements and repairs, and in planning associated budgets. Various modelling approaches have been used 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, 2001). One approach is to use models drawn from reliability theory which usually 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 Markov Chain Monte Carlo (MCMC) methods can be used effectively to develop a pipe replacement policy.
However, one drawback of NHPP models is that they require detailed data on actual failure times 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 historically many water companies have only recorded 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. Moreover the data are most likely limited to only a few years in relation to the age of the pipes in the network (Gat and Eisenbeis, 2000). In other words the data on the number of failures is left truncated, for instance having only 10 years worth of failures for a 100 year old pipe. This added to the fact that bursts are rare events over the lifespan of a pipe, results in data sets having excess zeros in terms of failures. In this article a methodology is developed which extends conventional NHPP models to not only cope with aggregated numbers of failures but also with zero-inflation in the data. The models are implemented within the Bayesian framework using MCMC methods and applied to data involving a network of 532 cement water pipes in Manukau City, Auckland, New Zealand.

2 Model Specification

2.1 Basic Model

Suppose that pipe bursts occur as a NHPP with time-dependent intensity function $\lambda(t)$. An important property of the NHPP is that the number of failures, $N(t)$, in any time interval $[t_1, t_2]$ follow a Poisson distribution with mean $\int_{t_1}^{t_2} \lambda(t)dt = \Lambda([t_1, t_2])$ (Meeker and Escobar, 1998), i.e.

$$Pr(N(t_1) - N(t_2) = n) = \frac{e^{-\Lambda([t_1,t_2])}\Lambda([t_1, t_2])}{n!} \tag{1}$$

A variety of models exist that can be used to express the intensity $\lambda(t)$ (Kleiner and Rajani, 2001). Here we adopt a formulation based on the power law (e.g. Landers et al., 2001; Sen, 2002) where:

$$\lambda(t) = \gamma\theta(\boldsymbol{x})t^{\theta(\boldsymbol{x})-1}$$

and $\theta(\boldsymbol{x}) = \boldsymbol{\beta x}$ is a linear function of suitable pipe covariates $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_k)$ with associated parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$. Note that the shape function, $\theta(\boldsymbol{x})$, can represent both a deteriorating system ($\theta(\boldsymbol{x}) > 1$) and an improving system ($\theta(\boldsymbol{x}) < 1$).

2.2 Likelihood Function

Suppose we observe pipe i in the time period $[0, T_i)$ and that n_i denotes the numbers of failures each of which occurred at times: $0 < t_1 < t_2 < \cdots < t_{n_i}$. So if $t_{n_i} < T_i$ then the data are time truncated, whereas if $t_{n_i} = T_i$ the data are failure truncated. Then, the conventional NHPP (Rigdon and Basu, 2000) leads to a time truncated likelihood function for the failure times in pipe i as:

$$\begin{aligned} f(t_1, t_2, \ldots, t_{n_i}) &= \left[\prod_{j=1}^{n_i} \lambda(t_{ij})\right] exp\left\{-\int_0^{T_i} \lambda(y)dy\right\} \\ &= \left[\prod_{j=1}^{n_i} \lambda(t_{ij})\right] exp\left\{-\Lambda\left([0, T_i)\right)\right\} \end{aligned}$$

As mentioned previously, this likelihood 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 (1), 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 we have N pipes and that these pipes are independent, then the overall likelihood for the data on all pipes is:

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

and since

$$\Lambda\left([t_{0i}, T_i)\right) = \int_{t_{0i}}^{T_i} \lambda(t_{ij})dt_{ij} = \int_{t_{0i}}^{T_i} \gamma\theta(\boldsymbol{x}_i)t_{ij}^{\theta(\boldsymbol{x}_i)-1}dt_{ij} = \gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right]$$

we have

$$L(\gamma, \boldsymbol{\beta}) = \prod_{i=1}^{N} \left(\frac{1}{n_i!}\right) e^{-\gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right]} \left(\gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right]\right)^{n_i} \qquad (2)$$

Note that in this formulation γ is a global parameter, rather than pipe specific. In our case this is appropriate since pipes are of similar material and the ground they are buried in is comparable. However, γ can easily be made pipe-specific if the application demands it.

2.3 Zero Inflated NHPP

A substantial part of the data sets involving water pipe failures are left truncated due to the fact that it is only lately that water companies have started collecting information on failures. Clearly, there are likely to be very few recorded failures over the most recent 5 or 10 years of a pipe installed up to 100 years ago, given that failures over its whole lifespan are relatively rare. This results in data sets where the majority of the pipes appear to have no failures at all. To cope with this zero-inflation in defects of items in manufacturing, Lambert (1992) introduced the Zero Inflated Poisson (ZIP) model which has also been widely used in many other applications (e.g. Gupta et al., 1996; Zorn, 1998; Bohning et al., 1999; Ghosh et al., 2006). The idea in the ZIP model is that it has two states; the first which produces zeros with probability $(1-p)$ and the second which produces counts from a $Poisson(\mu)$ distribution with probability p. Assuming a random sample $y_1, y_2, \ldots, y_m$ this results in a mixture distribution

$$f(y_k; p_k, \mu_k) = \begin{cases} (1-p_k) + p_k e^{-\mu_k}\,, & \text{if} \quad y_k = 0\,, \\ p_k \dfrac{e^{-\mu_k}\mu_k^{y_k}}{y_k!}\,, & \text{if} \quad y_k = 1, 2, \ldots \end{cases}$$

whose log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{p}; \boldsymbol{y}) &= \sum_{k=1}^{m} I_{(y_k=0)} \ln\left[(1-p_k) + p_k e^{-\mu_k}\right] \\ &+ \sum_{k=1}^{m} I_{(y_k>0)} [\ln(p_k) - \mu_k + y_k \ln(\mu_k) - \ln(y_k!)] \end{aligned}$$

where

$$I_{(\text{event})} = \begin{cases} 1\,, & \text{if event is True}\,, \\ 0\,, & \text{if event is False} \end{cases}$$

Using the same idea we can extend our 'aggregated' model for pipe burst by assuming that each of the N pipes follows a zero-inflated nonhomogeneous Poisson distribution, so that:

$$f(n_i, T_i, t_{0i}) = \begin{cases} (1-p_i) + p_i exp\left\{-\Lambda\left([0, T_i)\right)\right\} & n_i = 0 \\ p_i \dfrac{exp\left\{-\Lambda\left([0, T_i)\right)\right\}\left(\Lambda\left([0, T_i)\right)\right)^{n_i}}{n_i!} & n_i = 1, 2, \ldots \end{cases}$$

with log-likelihood:

$$\begin{aligned} l(\gamma, \boldsymbol{p}, \boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{n}, \boldsymbol{T}, \boldsymbol{t_0}) &= \sum_{i=1}^{N} I_{(n_i=0)} \ln\left[(1-p_i) + p_i exp\left\{-\gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right]\right\}\right] \\ &+ I_{(n_i>0)}\left[[\ln(p_i) - \gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right] + n_i \ln\left(\gamma\left[T_i^{\theta(\boldsymbol{x}_i)} - t_{0i}^{\theta(\boldsymbol{x}_i)}\right]\right) - \ln(n_i!)\right] \end{aligned} \tag{3}$$

where $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$.
Following the lines of Lambert (1992) which suggests that the parameter p can itself be parameterised to be a function of covariates, we set

$$\text{logit}(p_i) = \gamma T_i^{\theta(\boldsymbol{x}_i)}$$

Hence p_i is related to the age of the pipe at the end of the observation period and the characteristics of the reliability curve as determined by $\theta(\boldsymbol{x}_i)$ and γ.

3 Model Application

Here we consider application of the proposed models to the Howick Pressure Zone data in Manukau city, Auckland, New Zealand (Watson, 2005). These data consist of 532 asbestos cement pipes with 175 recorded failures in the eleven year period 1990-2001. Some pipes experienced multiple failures and in actuality only 81 of the 532 pipes had reported failures, so that 451 pipes had zero failures.
Both the aggregated model (2) and the ZIP model (3) were fitted in WinBUGS (Spiegelhalter et al., 1999) for the first nine years of data and used to derive a posterior predictive distribution for the number of failures in each pipe for the remaining 2 years of the observation period. Covariates used in the model include pipe length, pipe diameter, pressure and absolute pressure.
Two parallel Markov chains were run for each version of the model, using a burn in of 10,000 and then sampling every 25 iterations to collect a total of 10,000 samples from each chain. This was enough to ensure good convergence and rate of mixing for each parameter. A Gaussian prior with zero mean and large variance was assumed for each of the parameters γ, and $\beta_0, \beta_1, \ldots, \beta_k$.
Although the data did not contain actual times of failures, both models were still able to capture the ageing process in the vast majority of the pipes through the function $\theta(\boldsymbol{x}_i)$. However, the ZIP model leads to a substantial reduction in the standard errors of the covariate coefficients involved in $\theta(\boldsymbol{x}_i)$, suggesting that the deterioration in the pipes is more precisely estimated when zero-inflation is allowed for.
In reality 26 pipe failures were experienced in the 532 pipes in the last two years of the data collection period. The posterior predictive mean of total number of pipe failures from the zero-inflated model for these years based on the previous nine years of data was 21.5, with associated 95% credible interval $(10, 29)$.

4 Conclusions

In this paper we have considered two modifications (time aggregation and zero-inflation) to the conventional NHPP model previously used to predict

bursts in underground pipes. In summary, when applied to real data, the 'aggregated' zero-inflated model proposed here would appear to be able to adequately capture the ageing process in individual pipes (a key element of the NHPP) and provide usable predictions of numbers of pipe failures, despite the sparsity of failures in the data and the lack of information on actual failure times. On-going work is investigating refinements to the model to incorporate measurement error in covariates, and also to formulate a zero-inflated mixture of both the 'aggregated' and conventional NHPP, so as to take advantage of data sets where exact failure times are available for some pipes in the network, but only total failures for others.

References

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated poisson and the decayed, missing and filled teeth index in dental epidimiology. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. **162**, 195-209.

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

Ghosh, S., Mukhopadhyay, P. and Lu, J. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*. **136**, 1360-1375.

Gupta, P., Gupta, R. and Tripathi, R. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis*. **23**, 207-218.

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

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

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.

Meeker, W. and Escobar, L. (1998). *Statistical methods for reliability data*. New York: John Wiley and Sons Inc.

Rigdon, S. and Basu, A. (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. (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.

Zorn, C. (1998). An analytic and empirical examination of zero-inflated and hurdle poisson specifications. *Sociological Methods Research*. **26**, 368-400.