Bayesian Pipe Burst Models Using Cumulative Failure History

Theodoros ECONOMOU[*], Zoran KAPELAN[**] and Trevor BAILEY[*]
*Mathematics Research Institute, SECaM, University of Exeter, UK
**Centre for Water Systems, SECaM, University of Exeter, UK

Abstract: The processes and mechanisms giving rise to failures in repairable systems such as underground water pipes are quite complex and not quite fully understood yet. Many data sets involving pipe break history are poor in both quantity and quality resulting in data sets in which a significant number of pipes have no failures recorded at all. Conventional point process models such as the NHPP may be inadequately flexible to capture the failure process. In this paper both an NHPP and a zero-inflated version of it are applied on a North American pipe data set, using the cumulative number of failures in the past as an explanatory variable.

Key words: Pipe burst, Bayesian model, NHPP, zero-inflation, asset management, water distribution system

Introduction

The processes that cause bursts in underground water pipes, which are repairable components which may or may not behave independently, are often complex and hard to model statistically. This is mainly due to the fact that many factors affecting these processes are hard or even impossible to observe sometimes and in addition, existing records are often of poor quality. Accurate prediction of pipe bursts is vital to water companies in terms of budgeting and planning replacements or repairs therefore models need to be flexible enough to capture these complex processes as well as allowing for possible heterogeneity between pipes and for un-natural variations in the data (e.g. measurement errors). There are many applications of statistical models on water pipe systems some examples of which are: Lei and Saegrov, 1998; Mailhot et al., 2000; Dridi et al., 2005. A common way to picture pipe failures is points on a time

line so an intuitive and commonly adopted class of models for pipe bursts is point process models. (Kleiner and Rajani, 2001; Gat and Eisenbeis, 2001). Point processes describe the occurrences of events in time according to an intensity function or the occurrence rate. Here the events are pipe failures and the intensity function can be seen as the failure rate. A particular point process is the non-homogenous Poisson process (NHPP) which is useful in the sense that it is flexible enough to capture the (possibly) non-linear relationship of the failure rate with time and at the same time allowing for the inclusion of suitable pipe factors (Loganathan et al., 2002). In addition, unlike the homogenous case, in an NHPP, the times between each failure are not independently distributed so the NHPP is a well established process, able to capture the deterioration (ageing) in water pipes.

One of the main properties of the NHPP is that it can be viewed as a time dependent Poisson distribution. Formally, in any time interval $(t_1, t_2]$, the number of pipe bursts follows a Poisson distribution with mean

$$\Lambda((t_1, t_2], \mathbf{x}) = \int_{t_1}^{t_2} \lambda(t, \mathbf{x}) dt$$

where $\lambda(t, \mathbf{x})$ is the failure rate (failures/unit time) which depends on time t and a vector of pipe factors $\mathbf{x}$. This property was in fact used in (Economou et al., 2007) where an aggregated model was developed when dealing with data that do not include actual burst times. As mentioned earlier, many pipe burst data have a time span which is much less than the actual age of the pipes resulting in the fact that in terms of counts of failures under the Poisson assumption, there will be an excess amount of zeros in the number of failures (a situation commonly known as zero-inflation). To deal with zero-inflation in models that assume a Poisson distribution for counts, (Lambert, 1992) introduced the zero-inflated Poisson (ZIP) model which has been used extensively since then (Angers and Biswas, 2003; Ghosh et al., 2006). The idea here is to utilise the idea of the ZIP to be used in conjunction with the NHPP essentially offering extra flexibility if needed to allow for an excess of zeros in the failures.

In this paper we are proposing a zero-inflated NHPP model to account for the excess of zeros in the data by adopting the idea of the ZIP model. In Section 2 the NHPP model is described and then extended to its zero-inflated version. Both models are applied to a real-life North American data set involving a network of 1349 pipes. The results obtained are presented in Section 3. Section 4 presents an overview with conclusions as well as on-going work. This work is an extension of the work in Economou et al. (2008) where the same data set was analysed without using the past cumulative burst frequency as an explanatory factor.

1. NHPP MODEL SPECIFICATION

Conventionally, the NHPP is modelled through the failure rate (intensity function) $\lambda(t, \mathbf{x})$ and here we assume a parametric model based on the power law:

$$\lambda(t, \mathbf{x}) = \theta t^{\theta-1} e^{\boldsymbol{\beta}\mathbf{x}}; \quad \theta > 0$$

where θ is the shape parameter. Note that $\theta > 1$ implies that the pipe is 'ageing' or getting worse since $\lambda(t, \mathbf{x})$ will be increasing non-linearly with t whereas $\theta = 1$ implies a constant failure rate and the process is HPP. Essentially, $\theta t^{\theta-1}$ is the baseline failure rate of a pipe which is shifted accordingly by the influence from related explanatory variables $\mathbf{x} = (1, x_1, \ldots, x_q)$ with associated parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)$. Water pipes are repairable components which do not necessarily return to a better or the same condition after they are repaired, leading to the idea of including the cumulative number of failures $Z(t)$ as a possible (time dependent) factor. Following the lines of Gat, 2006 the factor is incorporated in the failure rate as such:

$$\lambda(t, \mathbf{x}, Z(t)) = (1 + \alpha Z(t)) \theta t^{\theta-1} e^{\boldsymbol{\beta}\mathbf{x}}$$

where formally, $Z(t)$ is the cumulative number of failures observed up to time t and α is just a parameter.

To write down the likelihood of the NHPP, suppose that a pipe has been observed, in time interval $(t_0, T]$, to fail n times at $t_1, t_2, \ldots, t_n$ where $t_0 < t_1 < \cdots < t_n \leq T$. Given this data, the likelihood of the NHPP with failure rate $\lambda(t, \mathbf{x}, Z(t))$ is given by:

$$L(\cdot) = \left[\prod_{j=1}^{n} \lambda(t_j, \mathbf{x}, Z(t_j)) \right]^{\delta} \exp\{-\Lambda((t_0, T], \mathbf{x})\} \qquad (1)$$

where δ is zero if n is zero and equal to one otherwise. Note that t_0 is not necessarily zero, i.e. the observations starts when the pipe was installed since this is rarely the case as many data sets are left censored in that respect. If that is the case then t_0 in (1) should reflect the time that the pipe was first observed in relation to its installation date. Also, (1) is said to be failure truncated if $T = t_n$ and time truncated if $T > t_n$.

NHPP Likelihood for A Network of Water Pipes

To extend the results above to more than one pipe, suppose that N pipes have been observed each in the intervals $(t_{0i}, T_i]$ where $i = 1,2,...,N$ denotes a single pipe. The failure rate of an individual pipe is:

$$\lambda_i(t, \mathbf{x}_i, Z_i(t)) = (1 + \alpha Z_i(t))\theta_i t^{\theta_i - 1} e^{\boldsymbol{\beta}_i \mathbf{x}_i}; \quad \theta_i > 0$$

where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_1, \ldots, \beta_q)$. By making β_0 pipe specific, we include random effects in the model which will account for possible heterogeneity between the pipes other than that explained by the explanatory variables and θ_i. In other words, these random effects will allow for any 'strange' behaviour in the failure process of the pipe other than that explained by the ageing and the pipe factors. Using (1), the overall likelihood for a network of N pipes, assuming independence, is:

$$L(\cdot) = \prod_{i=1}^{N} \left[\left[\prod_{j=1}^{n_i} \lambda_i(t_{ij}, \mathbf{x}_i, Z_i(t)) \right]^{\delta_i} \exp\{-\Lambda_i((t_{0i}, T_i], \mathbf{x}_i)\} \right]$$

The Zero-inflated NHPP Model

The ZIP model was first introduced by Lambert, 1992 to cope with zero-inflation in defects of items in manufacturing. The underlying idea is that in the ZIP, extra probability is added to the event of zero counts in the Poisson model essentially classifying it as a mixture model: counts are either generated by a Poisson distribution with probability p or by a zero generating process with probability $(1-p)$. Mathematically, y is distributed as a ZIP if

$$f(y) = \begin{cases} (1-p) + pe^{-\mu} & \text{for} \quad y = 0 \\ p\dfrac{e^{-\mu}\mu^y}{y!} & \text{for } y = 1,2,... \end{cases}$$

The idea can be relatively easily extended to the NHPP model where an extra parameter p_i is introduced for each pipe so that failures are either generated by a NHPP with probability p_i or by a process that generates no failures with probability $(1-p_i)$. The likelihood of the zero-inflated NHPP can written down using p_i explicitly but here we prefer to introduce a new variable u_i, a thing which helps in the coding of the model as well :

$$L(\cdot)=\prod_{i=1}^{N}\left[u_i\left(\left[\prod_{j=1}^{n_i}\lambda_i(t_{ij},\mathbf{x}_i,Z_i(t))\right]^{\delta_i}\exp\left\{-\Lambda_i\left((t_{0i},T_i],\mathbf{x}_i\right)\right\}\right)+(1-u_i)(1-\delta_i)\right]$$

$$u_i \sim Bernoulli(p_i)$$

In our point of view, $(1-p_i)$ reflects the natural tendency of the pipe to resist failure. This will of course be different for each pipe which is why it is pipe specific and in addition it is parameterised so that a random effect is also included:

$$\text{logit}(p_i)=\gamma_{0i}+\gamma_1 x_{1i}+\cdots+\gamma_q x_{qi}+\gamma_{\text{age}}T_i$$

Note that in addition to the pipe factors, the age of each pipe at the end of the observation period T_i is also included under the assertion that the resistance to failure of a pipe will probably decrease as it gets older.

3. CASE STUDY

Description of Data

Both models, the NHPP and the zero-inflated NHPP (ZINHPP), were applied to a data set of a network of 1349 underground water pipes of a municipality in North America. All pipes are made of the same material, namely cast-iron and were all installed between 1945 and 1960. The times of failures are given in months of each year but here we choose to work at the yearly level since the monthly level is probably too small to work with sensibly. Table 1 shows some details about the failure data:

Table 1. Description of the Network

Number of pipes	1349
Total number of failures	5425
Earliest failure on record	1962
Latest failure on record	2003

In order to test the predictive accuracy of the models as well as their ability to cope with left-truncated data, the models were calibrated over the 30-year period 1969-1998 and validated over the 5-year period 1999-2003. Details are given in Table 2.

Table 2. Data Set for Model Application

Calibration (observation) period start	Jan1969
Calibration (observation) period end	Dec 1998
Validation (prediction) period start	Jan 1999
Validation (prediction) period end	Dec 2003
Total number of failures (calibration)	4324
Pipes with no failures (calibration)	346
Total number of failures (validation)	422
Pipes with no failures (validation)	1032

Bayesian Estimation of Parameters

The two models described in section two are quite complex in terms of number of parameters recalling that there exist 2 pipe specific parameters for the NHPP model and 3 for the zero-inflated one. These types of models are naturally and perhaps more easily handled within the Bayesian framework and in particular using Markov chain Monte Carlo (MCMC) methods. In the Bayesian paradigm, parameters are seen as random quantities where the randomness represents the uncertainty that we have about them. This uncertainty is expressed in the form of prior distributions which are updated by the likelihood of the data to arrive at the posterior distributions which express the uncertainty of the parameters after the data has been taken into account. In the North American data set, the only available pipe factor is the length of each pipe so:

$$\boldsymbol{\beta}_i \mathbf{x}_i = \beta_{0i} + \beta_1 \text{length}_i \quad \text{and} \quad \text{logit}(p_i) = \gamma_{0i} + \gamma_1 \text{length}_i + \gamma_{\text{age}} T_i$$

The relatively uninformative priors assumed for each parameter are summarized in Table 3.

Table 3. Prior Distributions

θ_i	$\text{Gamma}(a,b)$
α	$\text{Normal}(0,1000)$
β_{0i}	$\text{Normal}(\mu,\sigma^2)$
γ_{0i}	$\text{Normal}(0,1000)$
β_1	$\text{Normal}(0,1000)$
a	$\text{Gamma}(0.01,0.01)$
b	$\text{Gamma}(0.01,0.01)$

μ	$\text{Normal}(0,1000)$
$1/\sigma^2$	$\text{Gamma}(0.01,0.01)$
γ_1	$\text{Normal}(0,1000)$
γ_{age}	$\text{Normal}(0,1000)$

The models were applied using WinBUGS (Spiegelhalter et al., 1999) and samples of the posterior predictive distributions for the actual number of failures as well as the probability of failure were collected for both the calibration and the validation period.

Note that although there is a large number of parameters ($\theta_i, \beta_{0i}, \gamma_{0i}$), these are assigned a common prior distribution (and hence a common posterior), effectively only using information in the data to calibrate a much smaller number of parameters.

Results

The means of the posterior distributions were taken as the point estimates for each parameter and are shown in table 4. Table 5 shows statistics of the pint estimates for the pipe specific parameters.

Table 4. Estimates of Global Parameters

Parameter	NHPP		Zero Inlflated NHPP	
	Estimate	St. Error	Estimate	St. Error
β_1	0.00268	0.00017	0.00234	0.00015
α	0.439	0.042	0.372	0.032

Table 5. Estimates of Pipe Specific Parameters

Para-meter	NHPP			Zero Inlflated NHPP		
	Mean	Min	Max	Mean	Min	Max
θ_i	0.128	0.094	0.182	0.147	0.117	0.192
β_{0i}	1.494	1.460	1.54	1.487	1.407	1.612
p_i	-	-	-	0.902	0.055	1

The variability in the point estimates (means of posteriors) of the probabilities p_i indicates that the ZI-NHPP is likely to be capturing the differences that pipes with zero recorded failures may have with other pipes.

Samples from the posterior distribution of the probability of one or more failures were collected for both the calibration and the validation period. These probabilities were then used in a Bernoulli trial to decide whether a pipe will fail or not, i.e. if zero is the outcome of the trial then pipe does not fail and vice versa. A 2x2 'confusion' matrix could then be constructed whose diagonal entries reflect the number of pipes correctly predicted to fail or not whereas off-diagonal entries are the number of wrongly classified pipes. 500 of these matrices were constructed and averaged for each model. Results are shown in Tables 6-9.

Table 6. NHPP model calibration period

NHPP model	Predicted pipe failures			
		Not failed	Failed	Total
Actual pipe failures	Not failed	101.7(8%)	244.3(18%)	346
	Failed	103.2(8%)	899.8(67%)	1003
	Total	204.9	1144.1	1349

Table 7. Zero-inflated NHPP model calibration period

ZI-NHPP model	Predicted pipe failures			
		Not failed	Failed	Total
Actual pipe failures	Not failed	165.7(12%)	108.3(13%)	346
	Failed	90.7 (7%)	912.3(68%)	1003
	Total	256.4	1020.6	1349

Table 8. NHPP model validation period

NHPP model	Predicted pipe failures			
		Not failed	Failed	Total
Actual pipe failures	Not failed	743.3(55%)	288.7(21%)	1032
	Failed	172.9(13%)	144.1(11%)	317
	Total	916.2	432.8	1349

Table 9. Zero-inflated NHPP model validation period

Zero-Inflated NHPP model	Predicted pipe failures			
		Not failed	Failed	Total
Actual pipe Failures	Not failed	750.3(56%)	281.7(20%)	1032
	Failed	173.9(13%)	143.1(11%)	317
	Total	924.2	424.8	1349

From Tables 5 and 6 it appears that the ZI-NHPP is doing a slightly better job in reproducing the right number of pipes that failed in the sense that the two main diagonal entries add up to 80% comparing to the 75% from the NHPP model. The difference is clearly due to the fact that the zero-inflated model is doing a better job at

fitting pipes with zero failures as should have been expected. This, however, is not true for the validation period where the two models perform similarly.

4. Conclusions

The processes involved in modelling the occurrence of bursts in water pipes are clearly very complex and existing data rarely contain enough information which may be helpful in capturing these processes. Here we have taken the conventional NHPP model and extended it to incorporate random effects as well as utilising the history of cumulative failures as an explanatory factor. The model has sufficiently captured the pipe failures, 75% in the calibration and 66% in the validation period. Furthermore, under the assertion that the excess number of zeros in the failures can be assigned extra probability, a zero-inflated NHPP was fitted to the data which did perform better in the calibration period (80%) but not in the validation period (75%). This may be due to the fact that the extra probability of not failing, namely p_i, was not time dependent but assumed to be constant over time. By making p_i dependant on time may give the right amount of flexibility to the ZI-NHPP to really capture the extra resistance that some pipes may have to failing for reasons other than age.

5. Acknowledgements

The pipe data set used in this paper has been provided by Dr Yehuda Kleiner which is gratefully acknowledged.

References

Angers J. F. and Biswas A. (2003) “A Bayesian analysis of zero-inflated generalized Poisson model.” *Computational Statistics and Data Analysis*, **42**, 37-46.

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, UK.

Dridi L., Mailhot A., Parizeau M. and 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” *Proceedings of the 8th International Conference on Computing and Control for the Water Industry***,** U. of Exeter, UK.

Economou, T., Kapelan, Z. and Bailey, T. C. (2007) “An aggregated hierarchical Bayesian model for the prediction of pipe failures”, *Proc 9th International*

Conference on Computing and Control for the Water Industry (CCWI), Leicester, UK

Economou, T., Kapelan, Z. and Bailey, T. C. (2008) "A zero-inflated Bayesian model for the prediction of water pipe bursts", *Proc 10th International Water Distribution System Analysis Conference (WDSA)*, Kruger National Park, South Africa

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

Gat, Y. (2006) "Stochastic tools based on failure and condition records" *IWA Beijing Conference Sep. 2006*

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

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), 1-14.

Lei, J. and Saegrov S. (1998) "Statistical approach for describing failures and lifetimes of water mains." *Water science and technology*, **38** 6), 209-217.

Loganathan, G. V., Park, S. and Sherali H. D. (2002) "Threshold break rate for pipeline replacement in water distribution systems." *Journal of Water Resources Planning and Management*, **128**(4), 271-279

Mailhot A., Pelletier, G., Noel J.F. and Villeneuve J. P. (2000) "Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: Methodology and Application." *Water Resources Research*, **36**(10), 3053-3062