A ZERO-INFLATED BAYESIAN MODEL FOR THE PREDICTION OF WATER PIPE BURSTS

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, UK

[2]University of Exeter, Centre for Water Systems, Harrison Building, North Park Road, Exeter EX4 4QF, 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. Therefore flexible statistical models that try to explain these processes are necessary. Data available on pipe breaks is usually poor and left truncated resulting in data sets in which a significant number of pipes have no failures recorded at all. Trying to apply flexible point process models such as the NHPP to such data may limit the ability to adequately capture the failure process. In this paper a zero-inflated NHPP is proposed which tries to deal with such problems. The new model is tested on the Canadian pipe data set. The results obtained show that although the zero-inflated NHPP did not outperform the NHPP in general, it provided a better fit to the data and slightly more precise results.

Keywords: Pipe burst, Bayesian model, NHPP, zero-inflation, asset management, water distribution system.

1. INTRODUCTION

The statistical modelling and predicting of pipe bursts in underground water distribution systems is vital for water companies in terms of budgeting and planning replacements or repairs. The complex mechanisms that affect the occurrence of failures in water pipes are not quite fully understood, let alone observed since they are components of a system which is buried underground. Age alone is clearly not enough to reflect the deterioration in the pipes (Boxall et al., 2004) thus flexible probabilistic models that are able to capture the deterioration in the pipes and at the same time account for un-natural variations in the data (e.g. measurement errors). A common and flexible way of modelling the occurrences of pipe failures in time is to view them as stochastic point processes (Kleiner and Rajani, 2001; Gat and Eisenbeis, 2001). A frequently adopted point process is the so called non-homogeneous Poisson process (NHPP) mainly for the reason that it is flexible enough to capture the 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.

A fundamental property of the NHPP is that in any time interval $(t_1, t_2]$, the number of failures follows a Poisson distribution with mean

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

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 is in fact a useful property which is utilized in (Economou et al., 2007) where actual times of failures are not available and an aggregated NHPP model is developed. The fact that the NHPP can also be basically viewed as a Poisson distribution over the observation period would solve the problem of aggregation of the data but would cause concern for another. Failures are usually only observed for a very small proportion of pipes in the network meaning that 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).

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 data set involving a network of 1349 pipes in a Canadian town which presented in Section 3 where results are illustrated and compared. Section 4 presents an overview with conclusions as well as on-going work.

2. MODEL SPECIFICATION

The most intuitive way to apply a NHPP process is by modelling its intensity function (or the failure rate) $\lambda(t, \mathbf{x})$. Here we consider a parametric form based on the power law, namely:

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

Where θ is the shape parameter and in the water pipe context, $\theta > 1$ implies ageing 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. $\mathbf{x} = \left(1, x_1, \ldots, x_q\right)$ is a vector of related explanatory variables and $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_q\right)$ is a vector of parameters. In a few words, $\theta t^{\theta-1}$ can be seen as the baseline failure rate which is affected by the explanatory factors through $e^{\boldsymbol{\beta}\mathbf{x}}$ and so $\boldsymbol{\beta}$ describes the way that each of the variables affects the failure rate.

Assume now that we have data on a pipe observed in $\left(t_0, T\right]$ that failed n times at $t_1, t_2, \ldots, t_n$ where $t_0 < t_1 < \cdots < t_n \leq T$. The likelihood then for a NHPP with failure rate $\lambda(t, \mathbf{x})$ is

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

where δ is zero if n is zero and equal to one otherwise. Note that theoretically, $t_0 = 0$ (i.e. we start observing the pipe when it was installed) but this is not often the case in practice since many data sets on water pipe failures are left-truncated meaning that the pipes were observed some time after being

installed. In that case t_0 in (1) should reflect the time that the pipe was first observed in relation to its installation date.

NHPP Likelihood for A Network of Water Pipes

Extending the results above to more than one pipe, the failure rate $\lambda_i(t, \mathbf{x}_i)$ of pipe i is:

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

where $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_1, \ldots, \beta_q)$. Note that we have taken β_0 to be pipe specific here therefore this parameter can be viewed as a random effect which will account for possible heterogeneity between the pipes because otherwise the difference in the failure rates between each pipe would only come from 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 is:

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

The Zero-inflated NHPP Model

The idea behind ZIP model, originally introduced to cope with zero-inflation in defects of items in manufacturing, is to add extra probability to the event of zero counts in a Poisson model. Essentially a ZIP model is a mixture model: counts are either generated by a Poisson distribution with probability p or by a zero generating process with probability $(1-p)$. Formally, y is generated by a ZIP distribution if

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

Extending this idea to the NHPP model, we introduce an extra parameter p_i 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 $(1-p_i)$. In order to write down the likelihood down as neatly as possible it is worth introducing a new variable u_i which follows a Bernoulli distribution with parameter p_i so that $u_i = 1$ with probability p_i and $u_i = 0$ with probability $(1-p_i)$. The likelihood of the zero-inflated NHPP model can now be written as:

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

since $\delta_i = 0$ if no failures were recorded for pipe i during the observation period.

In our point of view, $(1-p_i)$ reflects the natural tendency of the pipe to resist failure. This of course may vary between pipes which is why we have defined p_i to depend not only on the pipe factors but on a random effect as well. Furthermore we include the age of the pipe at the end of the observation period to affect p_i as we make the assumption that a pipes resistance to failure will probably decrease as it gets older. Since $0 \le p_i \le 1$, the logit of p_i is defined as:

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

3. CASE STUDY

Description

Both the NHPP model and its zero-inflated version were applied to a data set of a network of underground water pipes of a municipality in South-Central Ontario in Canada. All pipes are made of cast iron and were installed between 1945 and 1960. The times of failures are given in months of each year. The pipes failure summary is given in 1.

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 Parameter Estimation

The models described in Section 2 are somewhat involved especially in the actual number of parameters recalling that there exist 2 pipe specific parameters for the NHPP model and 3 for the zero-inflated one. These kinds of models are naturally handled within the Bayesian framework and in particular using MCMC methods. In the Bayesian context, the uncertainty about parameters 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. The data set for the Canadian pipe network includes the length of each pipe and 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	$\text{Gamma}(a,b)$
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.

Results

The obtained parameter estimates, taken as the means of their posterior distributions, for the global parameters are shown in Table 4. Summary statistics for pipe specific parameters are shown in Table 5.

Table 4. Estimates of Global Parameters

Parameter	**NHPP**		**Zero Inlflated NHPP**	
	Estimate	St. Error	Estimate	St. Error
β_1	0.00471	0.00028	0.00372	0.00029
γ_1	-	-	0.223	0.031
γ_{age}	-	-	-0.0167	0.0041

Table 5. Estimates of Pipe Specific Parameters

Parameter	**NHPP**			**Zero Inlflated NHPP**		
	Mean	Min	Max	Mean	Min	Max
θ_i	0.967	0.840	1.13	1.040	0.920	1.107
β_{0i}	-5.743	-6.913	-4.480	-5.640	-6.763	-4.211
γ_{0i}	-	-	-	0.026	-7.017	12.931

Note that the parameters for length (β_1, γ_1) are positive essentially making the point that a longer pipe will have a higher failure rate but also be less prone to resisting failure. The fact that the parameter for age in the mixing probability of the zero-inflated model is negative comes at surprise since it slightly decreases the odds the NHPP process in relation to the zero process. Perhaps this is reflecting the fact that an older pipe is more 'stable' in the sense that younger pipes will have failures induced by external factors (e.g. bad installation). It is also worth noting that standard errors for parameters θ_i are generally smaller for the zero-inflated model meaning that ageing captured more precisely.

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 5-8..

Table 5. NHPP model calibration period

NHPP model	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe failures	Not failed	147.5 (11%)	198.5 (15%)	346
	Failed	137.1 (10%)	865.9 (64%)	1003
	Total	284.6	1064.4	1349

Table 6. Zero-inflated NHPP model calibration period

Zero-Inflated NHPP model	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe failures	Not failed	197.4 (15%)	148.6 (11%)	346
	Failed	125.3 (9%)	877.7 (65%)	1003
	Total	322.7	1026.3	1349

Table 7. NHPP model validation period

NHPP model	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe failures	Not failed	721.5 (53%)	310.5 (23%)	1032
	Failed	171.0 (13%)	146.0 (11%)	317
	Total	892.5	456.5	1349

Table 8. Zero-inflated NHPP model validation period

Zero-Inflated NHPP model	**Predicted pipe failures**			
		Not failed	Failed	Total
Actual pipe Failures	Not failed	718.2 (53%)	313.8 (23%)	1032
	Failed	168.6 (12%)	148.4 (12%)	317
	Total	886.8	462.2	1349

As it can be seen from Tables 5 and 6, the zero-inflated model has fitted the data better than the NHPP model (80% of correctly predicted pipes/events for the Zero-inflated NHPP model and 75% for the NHPP model). However, the validation results shown in Tables 7 and 8 do not show that much difference between the two models (65% of correctly predicted pipes/events for the Zero-inflated NHPP model and 64% for the NHPP model) .

As mentioned earlier, posterior predictive distribution samples of the number of failures were collected for both the validation and the calibration period. Typical plots for the predicted and observed cumulative number of failures against time are shown below for selected pipes. The 95% prediction intervals are also superimposed on the plots. Dots represent the actual cumulative number of failures and the vertical line represents the start of the prediction (i.e. validation) period.

Figures 1 and 2 illustrate cumulative failure curves, for two pipes (1691 and 5282) with above average number of failures for the zero-inflated model (equivalent plots for the NHPP model are almost identical and hence not shown here). A good match between the model predicted mean and the observed data can be seen on both figures. Having said this, note that confidence intervals are still rather wide indicating an uncertain prediction in both cases.

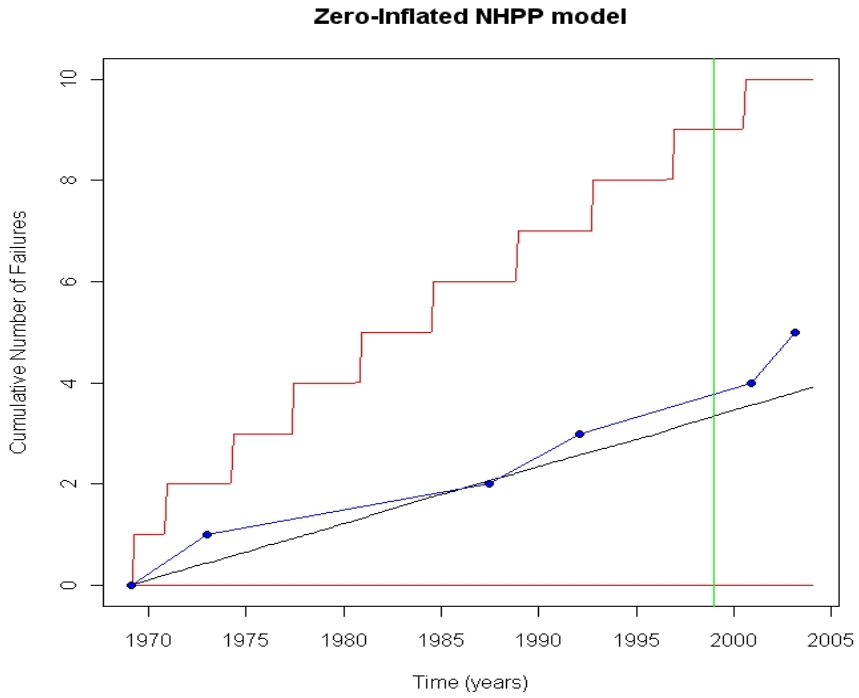

Figure 1. ZI NHPP Cumulative Number of Failures for Pipe 1691

Figure 3 illustrates a pipe where no failures have been observed in the calibration period while two failures occurred in the validation period. The dashed line represents the NHPP model mean while the solid indicates the zero-inflated model. The following can be observed from this figure: (1) both models predict well in term of mean values despite the fact that no failures were recorded in the calibration period; (2) the zero-inflated model is performing a bit better – the reason for this is that when no failures have been observed it enables the model to 'allow' extra probability of zero failures whenever this is required.

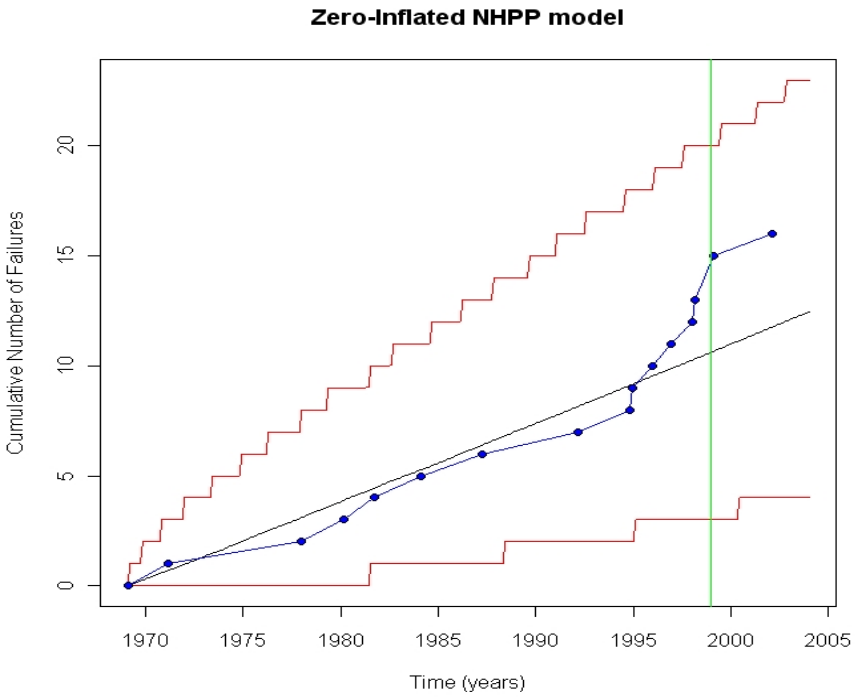

Figure 2. ZI NHPP Cumulative Number of Failures for Pipe 5282

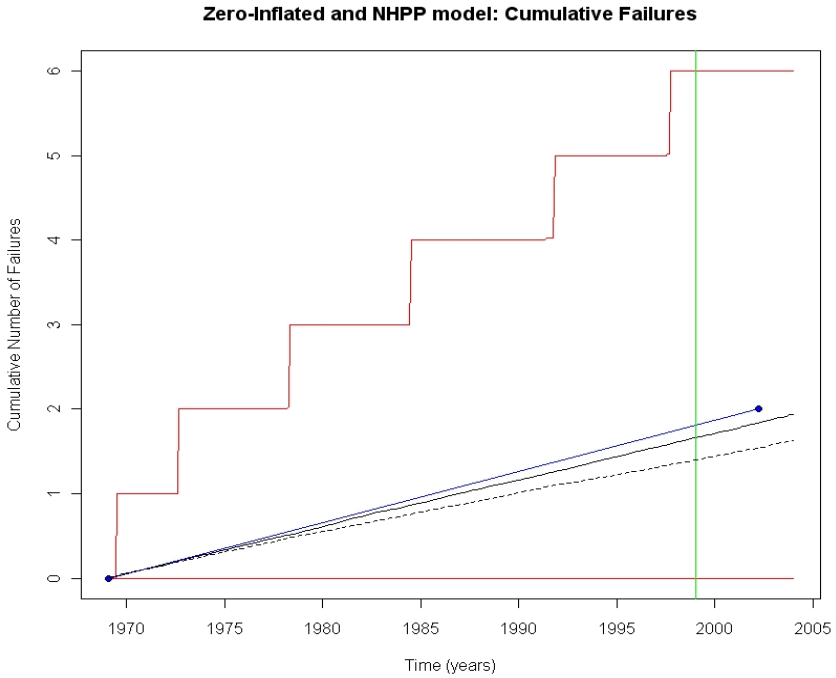

Figure 3. ZI NHPP and NHPP Cumulative Number of Failures for Pipe 3484

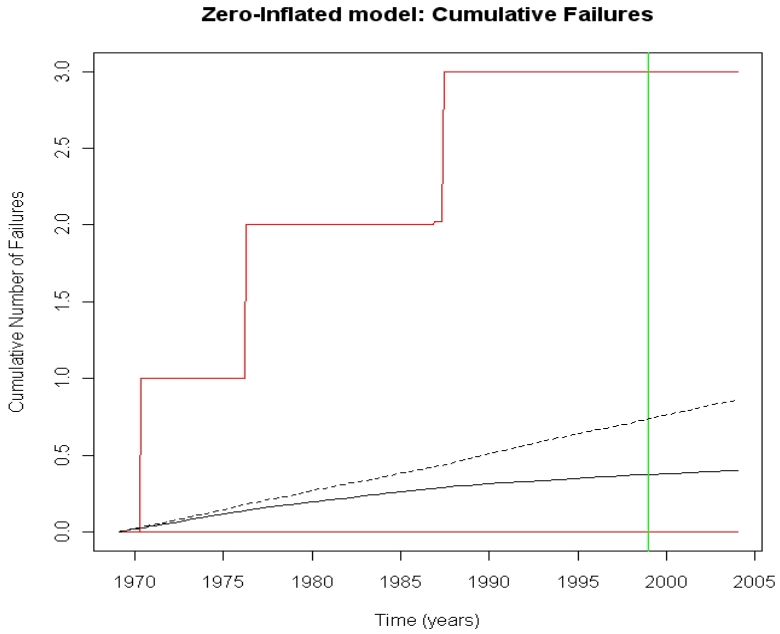

Figure 4. ZI NHPP and NHPP Cumulative Number of Failures for Pipe 1656

Figure 4 shows a plot for a pipe with no failures recorded either in the calibration or the validation period. As it can be seen form this figure, the zero-inflated model captures better the pipe behaviour in this case.

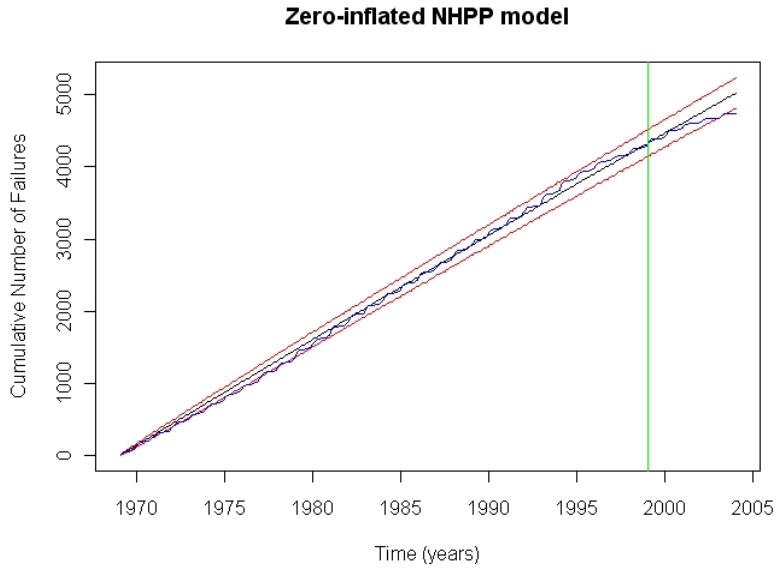

Figure 5. ZI NHPP Cumulative Number of Network Failures

In Figure 5, a plot for the entire network for the zero-inflated NHPP model is shown (equivalent plot for the NHPP model is not shown here as it is similar to the zero-inflated one). As it can be seen for this figure, the model predictions fit well the observations, especially in the calibration period (as expected).

Note that the confidence intervals shown in Figures 1–5 represent 2.5% and 97.5% quantiles of the posterior predictive distribution of the number of failures which are integer values and hence the step-like lines. The prediction mean line, on the other hand, has a continuous value since it is estimated as the mean of this distribution.

4. OVERVIEW AND CONCLUSIONS

In this paper we have considered a frequently used model for describing the occurrence of failures in a repairable system, namely the NHPP model and we have modelled its failure rate to include variables as well as random effects. The model proved flexible enough to capture the deterioration in the pipes as well as any heterogeneity between them. Furthermore, the model was adjusted to account for possible zero-inflation that may well exist in pipe failure data due to the fact that many pipes never experience a break during the observation period. Although the zero-inflated version of the NHPP did not outperform the NHPP in general, it provided a better fit to the data and slightly more precise results. A reason for this similar performance may well be the fact that the data set considered is a detailed one containing adequate information in terms of pipes failures.

The intuitive next step in this model is to allow the mixing probability of the zero inflated model to vary with time thus enabling the model to be flexible enough to allow for zero inflation at each time step. In this paper the mixing probability did depend on the age of the pipe but clearly this is not equivalent to having explicit dependence on time itself.

The fact that a mixture of a NHPP and a zero generating process provided slightly more precise results leads to the conclusion that a next step in trying to improve the NHPP model would be a mixture of two or even more NHPPs to allow an even more flexible model that captures the different behaviours in different pipes. Taking this a step further, it might be worth considering the hidden Markov model concept where the parameters of the failure rate vary stochastically over time according to a Markov chain of 2 or more states. This might enable the model to capture the sort of behaviour of the failures in Figure 3 where there is a clear jump in the cumulative failures curve.

5. ACKNOWLEDGEMENTS

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

6. 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.

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

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

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.

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