Journal of Hydroinformatics | 06.4 | 2004

# Bayesian-based pipe failure model

T. G. Watson, C. D. Christian, A. J. Mason, M. H. Smith and R. Meyer


## ABSTRACT

The efficient long term management of large-scale public funded assets is an area of growing importance. Ageing infrastructure, growth and limited capital all result in the need for a more robust and rigorous methodology to prioritise rehabilitation and renewal decisions and, as importantly, to forecast future expenditure requirements. The overall objective of this research is to develop a Bayesian-based decision support system that will facilitate the identification of efficient asset management policies. The Bayesian approach enables us to formally incorporate, express and update our uncertainty when determining such policies. This is particularly relevant for water utilities that have incomplete or unreliable historical failure data sets and, as a consequence, rely heavily on past engineering experience.

An object oriented discrete event simulation has been developed to analyse existing maintenance policies, test the Bayesian methodology and to develop and identify improved maintenance policies. This paper focuses on the areas of research relating to the long term management of water distribution systems and, in particular, will present: (1) an overview of the Bayesian approach, (2) development and initial results for an object oriented discrete event simulation and (3) proposed future research and development.


**Key words** | Bayesian statistics, discrete event simulation, counting process, hierarchical model, nonhomogenous Poisson process

**T. G. Watson**
National Institute of Water and Atmospheric Research Limited,
Auckland,
New Zealand

**C. D. Christian** (corresponding author)
Department of Civil and Environmental Engineering,
The University of Auckland,
Private Bag 92019,
Auckland,
New Zealand
Tel: +64 9373 7999
Fax: +64 9373 7462
E-mail: *c.christian@auckland.ac.nz*

**A. J. Mason**
**M. H. Smith**
Department of Engineering Science,
The University of Auckland,
Auckland,
New Zealand

**R. Meyer**
Department of Statistics,
The University of Auckland,
Auckland,
New Zealand

## INTRODUCTION

It is generally accepted that water supply infrastructures throughout the world are deteriorating systems. This, coupled with further expansion and renewal, places ever increasing financial stress on the present levels of funding. Water rates in Auckland, New Zealand, for example, are predicted to increase 5–20% per year for the next 10–20 years just to keep pace with maintenance and renewal programmes. The water industry is therefore facing the problem of managing deteriorating networks in the most efficient manner possible to maintain existing and future levels of service.

It is obvious that, for a deteriorating water supply system to meet predetermined levels of service, maintenance actions need to occur. Maintenance is carried out to prevent system failures, as well as to restore the system function when a failure has occurred. The prime objective of maintenance is thus to maintain or improve the system reliability and operation regularity (Hoyland & Rausand 1994). Maintenance typically consists of rehabilitation, repair and renewal. Most maintenance policies combine the making of renewal decisions after one or more failures with planned pipe replacements based on engineering judgement and a knowledge of the system.

In order to enhance maintenance decisions it is therefore essential to improve the understanding of the deterioration process and the evolution of failures of water mains pipes and to develop appropriate predictive models that can assist in the decision-making process. The resulting pipe break models can serve both as a diagnostic tool and an optimisation tool (e.g. for developing best replacement strategies), but also, when coupled with an economic assessment model, they become powerful tools for decision making by water managers (O'Day 1982).

## COUNTING PROCESS MODEL

For a well-designed water supply system the bulk of the pipes can be considered repairable, as the cost of failure is small in comparison to replacement. When a pipe failure occurs a small part of the pipe can be repaired, or possibly replaced, to restore the pipe to its functioning state without replacing the entire pipe. The pipe may be repaired several times before being replaced. Therefore, pipes that are considered repairable cannot be modelled by the conventional fitting of a statistical lifetime distribution, as successive failures are firstly, not identically distributed, and secondly, not independent. The succession of pipe failures can, however, be modelled using a counting process (Lei & Saegrov 1998).

A counting or point process is a stochastic model that describes the occurrence of events in time. These occurrences, in this case failures, can be thought of as points along a time axis. If the times between failures tend to get shorter with age the item is said to be deteriorating. Alternatively, if the times between failures are increasing then the item is improving. Define $N(t)$ to be a random variable denoting the number of failures for a given pipe in the time interval $[0,t]$. The expected number of failures from time 0 through to time $t$ is known as the mean function, $\Lambda(t)$, and can be written as

$$\Lambda(t) = E(N(t)) \tag{1}$$

The intensity function, $\lambda(t)$, of a counting process at time $t$ is defined to be

$$\lambda(t)=\frac{\mathrm{d}}{\mathrm{d}t}\Lambda(t)=\frac{\mathrm{d}}{\mathrm{d}t}E(N(t))=\lim_{\Delta t\to 0}\frac{E(N(t+\Delta t)-N(t))}{\Delta t} \tag{2}$$

The intensity function may be regarded as the mean number of failures, $N(t)$, per unit time. In other words, this describes the rate at which breaks are occurring, i.e. the failure rate. The intensity function is therefore the most important measure of a pipe's reliability. Many repairable systems, and we assume this includes pipes, typically have a 'bathtub' shaped intensity function (Figure 1).

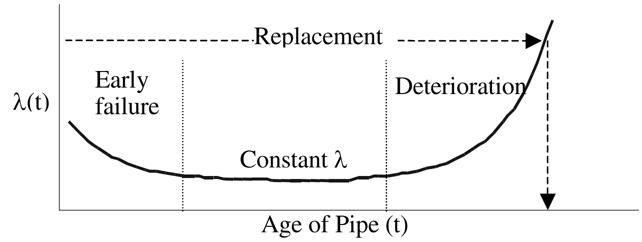


**Figure 1** | Bathtub shaped intensity function.

When a pipe is newly installed the intensity can be high and failures frequent. This can be described as a settling-in period, possibly due to construction practices. After early faults have settled down, the intensity will be smaller and remain relatively constant for long periods of its useful life. Then, as the pipe ages, the intensity will begin to increase and the pipe starts deteriorating. This is the period of most interest, as eventually the failure rate will exceed a certain level and it will become cost-efficient to replace the pipe.

This deterioration phase can be modelled by a Non-homogeneous Poisson Process (NHPP). The NHPP allows the intensity to vary with time. This implies that the times between failures are neither independent nor identically distributed. The NHPP also assumes minimal repair with negligible repair times. Minimal repair assumes that a repair will have no effect on the pipe failure rate and restores the pipe to 'as bad as old' condition. The likelihood of the pipe failing is the same immediately before and after a failure. This is considered a reasonable approximation, as a pipe repair is typically acting on only a small percentage length of pipe compared to the overall length of the pipe. Negligible repair time is considered acceptable when comparing a pipe's lifetime (measured in years) with repair times (counted in hours) (or possibly days).

There are several models that can be used to describe the intensity of a NHPP. Common ones include the power law model, $\lambda(t) = at^{b-1}$, the exponential model $\lambda(t) = a\mathrm{e}^{bt}$ and derivations of these models based on the proportional hazards model $h(t) = h_0(t)\mathrm{e}^{bz}$ (Cox 1972), replacing the baseline hazard function with the baseline intensity function, $\lambda(t) = \lambda_0(t)\mathrm{e}^{bz}$. Once the intensity function has been estimated the number of failures in a given time period is Poisson distributed:

$$P(N(t)=n)=\frac{\Lambda(t)^n}{n!}e^{-\Lambda(T)} \quad (3)$$

The expected number of failures in a given time period is of obvious use for rehabilitation and budgetary considerations.

## BAYESIAN APPROACH

The use of databases, and in particular geographic information systems, for managing infrastructure has only become common in New Zealand within the last 10–15 years. The often poor quality of earlier paper records means it is very rare for water utilities to have an entire history of failures. Recorded failure data is therefore, more often than not, limited to recent periods of time. Additionally, much of the data that has been recorded is of dubious quality, and quite often in an incompatible form or format. Research in European countries has identified that a smaller amount of more accurate data can lead to better results than more complete, but uncertain, data (Gat & Eisenbeis 2000). The unreliable nature of the data creates several problems in modelling the deterioration process of pipes. The most obvious one is that, with a lack of failure history, it becomes very difficult to estimate the failure rates of pipes. This has been a major shortcoming in previous research and has resulted in the dominance of engineering judgement in the decision-making process.

Even when there is a reliable failure history, the variability of pipe failures is large. This means that, even though two pipes may have exactly the same failure rate, the number of observed failures can be quite different. A natural estimator of the failure rate of a pipe can be calculated by dividing the number of failures by the length of the pipe and the observed time period. Basing maintenance on the observed historical number of failures can perhaps unnecessarily bias the decisions in favour of pipes with higher observed breakage rates.

The problem therefore is how to estimate the failure rate given limited reliable data and relying predominantly on engineering knowledge. One methodology that is well suited to this type of problem is Bayesian statistical modelling. A Bayesian model can combine engineering knowledge, in the form of our beliefs about the failure rate (prior distribution), with the data at hand to provide a formal estimate of the likely breakage rate distribution (posterior distribution). Bayes' equation can be written as

$$\text{Posterior}=\frac{\text{Prior}\times\text{Likelihood}}{\sum\text{Prior}\times\text{Likelihood}} \quad (4)$$

where the denominator of the right-hand side is a fixed normalising factor which ensures that the posterior probabilities sum to 1.

The above equation, relating to a failure rate, can therefore be expressed as

$$\begin{aligned}&\text{Posterior}\propto\text{Prior}\times\text{Likelihood}\\&P(\lambda|\text{data})\propto P(\lambda)P(\text{data}|\lambda)\end{aligned} \quad (5)$$

where $\propto$ means 'equal to except for a constant of proportionality' and $|$ is used to express a conditional probability. It is important to note that prior can be based on existing failure data that might be available, or even a combination of knowledge and data. As such, the Bayesian approach can be seen as a formal statistical updating methodology.

Another additional advantage of Bayesian inference is that it also provides a conceptually straightforward statistical procedure for providing management advice under uncertainty, commonly referred to as risk management. The posterior distribution explicitly accounts for the uncertainty of model parameters and structure, indicating the support given to each possibility using the best available information and data. When combined with cost data, the posterior can be seen as a formal measurement of risk (e.g. the expected consequences of each policy option and indications of uncertainty).

## CONSTANT FAILURE RATE MODEL

Let us consider one pipe with a constant failure rate, i.e. $\lambda(t)=\alpha$ for all $t$ and assume we observe $N$ breaks in the time interval $T$. As briefly mentioned above, the natural and obvious estimate of the failure rate $\alpha$ is

$$\hat{\alpha}=\frac{N}{T} \tag{6}$$

However, suppose we have prior information about $\alpha$ (from previous pipe failure studies, engineering knowledge or existing data, for example) and can express this as a Gamma distribution:

$$\alpha \sim \text{Gamma}(a,b) \ (\mu = a/b, \text{var} = a/b^2) \tag{7}$$

where the symbol '~' is used to mean 'distributed by'. The shape and tightness of this distribution, the mean and variance, will reflect our uncertainty about what we believe to be the actual failure rate. From Bayes' equation (5) and the Poisson likelihood function (3),

$$p(\alpha \mid N) \propto p(\alpha)p(N \mid \alpha)$$

$$\propto b^a \alpha^{a-1} e^{-b\alpha} e^{-\alpha T}(\alpha T)^N$$

$$\propto \alpha^{a+N-1} e^{-\alpha(b+T)} \tag{8}$$

or, in other words, the observed breakage information, $N$, updates our belief about the failure rate $\alpha$ to give

$$\alpha | N \sim \text{Gamma}\left(a+N, b+T\right), \left(\mu=\frac{a+N}{b+T}, \text{var}=\frac{a+N}{(b+T)^2}\right) \tag{9}$$

We can see that, with a large quantity of data (i.e. many breaks), the posterior mean tends towards $N/T$, the natural estimator of $\alpha$. Likewise, if there is little failure data available, the mean is dominated by the prior estimate of $\alpha$ ($a/b$). If the pipe has not broken at all, for example, we still have an estimate of $\alpha$ based on our engineering knowledge of similar pipes. It can be seen therefore that the Bayesian approach provides an excellent way of combining engineering knowledge and available data in a robust and formal statistical manner.

In the above example the posterior is of standard form (i.e. a Gamma distribution). In Bayesian terminology this is known as conjugacy. A conjugate prior is a prior that, when combined with the likelihood function, results in the posterior being of a standard recognisable form. However, this is not always the case, and for most problems of higher complexity, and hence useful application, the posterior is not of standard form. To solve such equations one must employ Monte Carlo Markov Chain (MCMC) techniques. This is beyond the scope of this paper and the reader can refer to the following references for further information (Brooks 1998; Migon & Gamerman 1999; Green 2000). Implementation of a more advanced model utilising MCMC techniques is not pursued in this paper and is the topic of a future proposed paper.

## HIERARCHICAL BAYESIAN MODEL

The above example illustrated the Bayesian methodology for one pipe. However, it would make sense to incorporate information and knowledge of all the pipes in the network in making estimates of the failure rate for any one pipe. To achieve this, we use a hierarchical Bayesian model.

Hierarchical models aim to combine the information from various sources of data while exploiting an assumed similarity between parameters. In this model it is assumed that the underlying failure rates $\lambda_i$ of each pipe $i$ are drawn from the same prior distribution, named the hyperprior. This pooling of data greatly improves the precision of the estimates of $\lambda_i$ and can be seen as a compromise between the assumptions that all pipes are identical, and therefore have the same failure rate, and that all pipes are different, in which case all pipes are treated separately using only data from pipe $i$ to estimate $\lambda_i$. In this model, a pipe failure on one side of a network may provide knowledge about the failure rate of a similar pipe on the opposite side of the network. This may be particularly valuable for decision-making if no failure data exists for some pipes.

Our specification of the hyperprior determines how the individual failure information updates other similar pipes in the network (Figure 2). If an informed ('tight') hyperprior, with a small variance, is specified, then *a priori* it is implied that the pipes are very similar. Any failure data relating to one pipe directly updates the belief in the

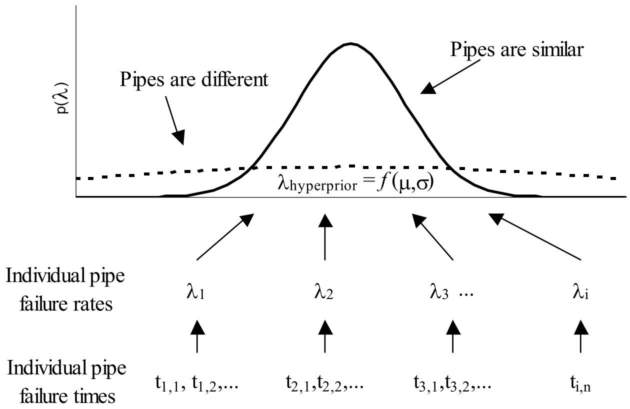


**Figure 2** | Overview of a hierarchical Bayesian model.

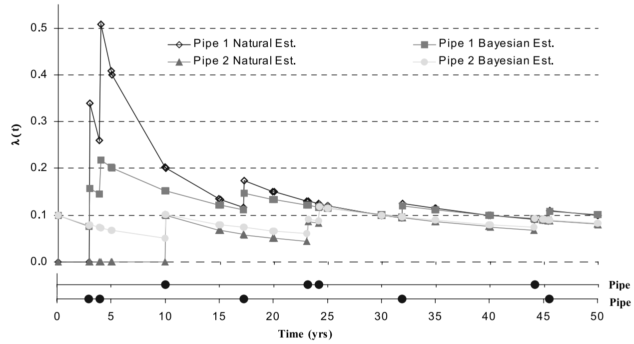


**Figure 3** | Bayesian hierarchical model simulation results.

failure rate of another pipe. In extreme, this is effectively combining all the pipes and failure data into one long length of pipe. Conversely, applying an uninformed ('loose') hyperprior, one with large variance, effectively treats all pipes as different. Typically, one would specify a semi-informed hyperprior, somewhere in between.

## SIMULATION

An object-oriented discrete event simulation with an embedded hydraulic solver has been developed to test various maintenance policies and demonstrate the Bayesian methodology. In particular, the simulation can highlight the Bayesian updating, or learning, process. As more failure data and any other relevant information becomes available, the simulation can formally update estimates of parameters of interest. The estimates, with their corresponding statistical properties, can then be used to update projected costs and hence help identify efficient replacement strategies.

The main advantage of this simulation approach is that it is transparent and easily understood by network management staff, increasing their faith in, and ultimately their use of, the new approach. Our simulation can also be easily adapted to handle a wide range of criteria and then be used to analyse any results that may be of interest. Providing valuable insights into the decision-making process and their effects is particularly important when the costs are not all monetary and a lot of the decisions are based on rules of thumb. The simulation can be used to search for optimal parameters for a predefined policy, or mixtures of policies.

## RESULTS

Breaks were generated randomly for two pipes assuming a constant failure rate $\lambda = 0.1$ break/yr. The failures can be seen as dots along a time axis in Figure 3. The simulation was then run for 50 years with failure rate estimates being updated at every break and at 5 yearly intervals. The prior for $\lambda$ was set with mean 0.1 and standard deviation 0.1. It is therefore implied that something is known about the likely value of $\lambda$.

Results displayed in Figure 3 show how the hierarchical Bayesian model provides better estimates of $\lambda$, especially in the first 25 years. As previously discussed, as more data becomes available the model converges to the natural estimate of $\lambda$.

For this example the natural estimate is taken to be the maximum likelihood factor of $\lambda$ under the assumption that $\lambda$ has a Poisson distribution with a mean of $\lambda T$, where $T$ is the observed time. The main effect of the model can be seen to be to pull the estimates of $\lambda$ together. The amount of this 'shrinking effect' is dictated by the amount of variance we believe exists between the two pipes. The estimates will therefore either be shrunk towards their expected mean or expanded towards their natural estimates, depending on our belief in the similarity of the pipes. However, all information will increase the accuracy of the estimates of $\lambda$ for both pipes. This is the main advantage of the Bayesian approach in that information from one pipe will increase the accuracy of failure estimates for similar pipes.

## CONCLUSIONS

The aim of this paper was to introduce and outline the development of a Bayesian-based decision support system for water distribution systems. In particular, to present a formal statistical methodology that combines engineering knowledge with recorded failure data that is used to provide estimates of the failure rate of pipes that make up a water network.

The Bayesian approach was shown to overcome data problems that commonly affect traditional statistical inference techniques. These problems include missing data, lack of data and truncation of data. Additionally, the Bayesian approach explicitly accounts for parameter and model uncertainty, providing water managers with formal and clear measurements of risk.

The hierarchical nature of the model can be seen as a compromise between the assumptions that all pipes are identical and that all pipes are different. This approach ensures that all data throughout the network is used in the most efficient way possible. Failure, and therefore cost estimates, can be calculated on an individual pipe level. It should also be seen that this approach is well suited, and potentially more beneficial, to other assets in which information can be 'pooled'. These asset types may include network structures such as sewer networks, rail networks and possibly electricity distribution networks.

More generally, it was shown that Bayesian statistical methodology is well suited to engineering type problems and its use should be actively encouraged.

## ACKNOWLEDGEMENTS

The majority of this research was funded by a Technology in Industry Fellowship on behalf of Technology New Zealand, The University of Auckland and Harrison Grierson Consultants Limited.

## REFERENCES

Brooks, S. P. 1998 Markov chain Monte Carlo method and its application. *The Statistician* **47**(Part 1), 69–100.

Cox, D. R. 1972 Regression models and life tables. *J. R. Statist. Soc.* **34**(Series B), 187–220.

Gat, Y. & Eisenbeis, P. 2000 Using maintenance records to forecast failures in water networks. *Urban Water* **2**(3), 173–181.

Green, P. J. 2000 A primer on Markov Chain Monte Carlo. In *Complex Stochastic Systems* (ed. O. E. Barndorff-Neilsen, C. Kluppelberg & D. R. Cox), pp. 1–62. Chapman & Hall, London.

Hoyland, A. & Rausand, M. 1994 *System Reliability Theory*. Wiley, New York.

Lei, J. & Saegrov, S. 1998 Statistical approach for describing failures and lifetimes of water mains. *Wat. Sci. Technol.* **38**(6), 209–217.

Migon, H. S. & Gamerman, D. 1999 *Statistical Inference: An Integrated Approach*. Arnold, London.

O'Day, D. 1982 Organizing and analyzing leak and break data for making main replacement decisions. *J. Am. Wat. Works Assoc.* **74**(November), 588–594.