A hidden semi-Markov model for the occurrences of water pipe bursts

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

[1] School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, UK

Abstract: A frequently used approach when modelling the bursts of underground water pipes is to assume a non-homogeneous Poisson process (NHPP) for the occurrence of failures. This however does not account for possible serial dependence in the failures or that the occurrence of failures may also be affected by some temporal process other than ageing. This paper proposes a hidden semi-Markov model which is NHPP conditional on the states of the hidden process.

Keywords: Hidden semi-Markov; NHPP; Censoring; Truncation.

1 Introduction

Water companies (especially in the UK) have a need for proper and accurate estimations on water pipe bursts or blockages. In addition to the fact that processes driving the occurrences of pipe failures are complex and often unmeasurable, the available historical data is often scarce and unreliable. The statistical models employed would then need to be complex and flexible enough to capture the failure process. It is common practice to consider that the occurrences of pipe failures are generated by a point process in time, specifically a (possibly nonhomogeneous) Poisson process (Kleiner and Rajani, 2001). Recently Economou et al. (2007) considered an aggregated nonhomogeneous Poisson process (NHPP) model with an intensity function $\lambda(t, \boldsymbol{x})$ (essentially the failure rate) dependent on time t and covariates $\boldsymbol{x}$. They further extended the model to account for possible zero-inflation in their pipe burst data. Although the models adequately captured the ageing process and accurately predicted the total number of failures in the network, they performed poorly in predicting at the individual pipe level. This could be because there was nothing in the proposed models to account for serial dependence and possible unobserved covariates or even processes that might have influenced the burst occurrence. In this paper we consider how aspects like these can be incorporated into such models.

Water pipes are degradable components of an ageing system whose failure mechanism may involve several 'states' relating to the 'health' of the pipe. This leads to the idea of incorporating a hidden Markov process to the

NHPP model such that the intensity function $\lambda(t, \boldsymbol{x})$ is different according to which state the hidden process is in at time t. By doing this, the resulting process that generates the failures allows for both serial dependence and overdispersion (MacDonald and Zucchini, 1997). In a Markov process, the times between states are exponentially distributed and in the case of water pipe failures this may well be unrealistic. A semi-Markov process is essentially a Markov process with temporal structures and this not only makes the above model more flexible, but it also allows explicit modelling of the duration time between states (Dong and He, 2007). Following the thinking of Özekici (2003), the hidden process can be seen as an environmental process giving rise to variations in the parameters of the model. One could then say that model for the pipe failures includes time dependent random effects. The formulation of the model is presented in section 2 and model application in section 3.

2 Model Specification

Consider a single water pipe k which has been observed in the time interval $[t_0, t_{end}]$ and has failed n times at $t_1, t_2, \ldots, t_n$ ($t_0 = 0$ implies that observation started at the installation date of the pipe). Suppose that the occurrences of these failures occur as a NHPP that depends on a hidden state S_t of a semi-Markov process where $S_t \in \{1, 2, \ldots, M\}$ is the state space of the process. Here S_t is interpreted as the state of the process at time t. The semi-Markov process is basically defined by an initial distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_M)$, a transition probability matrix $\mathbf{P} = \{p_{i,m}\}$ and a matrix of holding times $\mathbf{H}(\mathbf{t}) = \{f_{i,m}(t)\}$. If at time t the process is in state i, it will decide to move to state m at time $(t + s)$ with probability $p_{i,m}$ and it will do so after holding for a time period with a density function $f_{i,m}(s)$. Here we assume that the resulting process occurs in discrete time, mainly due to the nature of pipe failure data but also because it somewhat reduces the complexity of the model.
For the characterization of the intensity function $\lambda(t, \boldsymbol{x})$ of the NHPP we adopt a power function of time:

$$\lambda_k(t, \boldsymbol{x}|S_t) = \gamma_k \theta^{(S_t)} t^{(\theta^{(S_t)}-1)} \exp\left\{\boldsymbol{\beta}^{(S_t)} \boldsymbol{x}\right\}$$

So given the state S_t the process is NHPP with intensity function $\lambda(t, \boldsymbol{x})$ which depends on state specific parameters θ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ corresponding respectively to the shape parameter and the parameters relating to possible explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_q)$. The scale parameter γ_k is pipe specific and thus constitutes a random effect to account for the between pipe variability. It is worth mentioning here that whilst the process conditional on S_t is a NHPP, the overall failure process risen from this model is not.

Note that from now on, the subscript k will be dropped for clarity until the results are generalized to more than one pipe. Now, in a NHPP the conditional distribution of t_j, the time of the j^{th} failure given the state and that the previous failure occurred at t_{j-1} is

$$\begin{aligned} h(t_j|t_{j-1}, S_{t_j}) &= \lambda(t_j|S_{t_j}) \exp\left[-\int_{t_{j-1}}^{t_j} \lambda(u|S_{t_j})du\right] \\ &= \lambda(t_j|S_{t_j})\mathrm{e}^{-\Lambda\left([t_{j-1},t_j]|S_{t_j}\right)} \end{aligned} \tag{1}$$

A second thing that needs to be considered in order to compute the likelihood of the model is the state sequence of the semi-Markov process $\{S_{t_0}, S_{t_1}, \ldots, S_{t_n}\}$, i.e. the state at the start of the observation period, the state at the time of the first failure t_1, the state at t_2 and so on. Consider a specific realization of this sequence $\{S_{t_0} = z_0, S_{t_1} = z_1, \ldots, S_{t_n} = z_n\}$ where $z_0, \ldots, z_n \in \{1, 2, \ldots, M\}$. The probability associated with this sequence then is

$$\begin{aligned} &\pi_{z_0} p_{z_0,z_1} f_{z_0,z_1}(t_1 - t_0) \cdots p_{z_{n-1},z_n} f_{z_{n-1},z_n}(t_n - t_{n-1}) \\ &= \pi_{z_0} \prod_{j=1}^{n} p_{z_{j-1},z_j} f_{z_{j-1},z_j}(t_j - t_{j-1}) \end{aligned} \tag{2}$$

Here, f_{z_{j-1},z_j} is assumed to follow a negative binomial distribution. Conditional on (2), the joint probability distribution of the data can be computed using (1):

$$\begin{aligned} &\lambda(t_1|S_{t_1} = z_1)\mathrm{e}^{-\Lambda\left([t_0,t_1]|S_{t_1}=z_1\right)} \cdots \lambda(t_n|S_{t_n} = z_n)\mathrm{e}^{-\Lambda([t_{n-1},t_n]|S_{t_n}=z_n)} \\ &= \prod_{j=1}^{n} h(t_j|t_{j-1}, S_{t_j} = z_j) \end{aligned} \tag{3}$$

The multiplication of equations (2) and (3) will result in the likelihood of the data conditional on the fact that $\{S_{t_0} = z_0, \ldots, S_{t_n} = z_n\}$ is known. Since this is not the case, to define the likelihood of the model we also need to sum over all possible values of the states. In other words, the likelihood of the model is:

$$L(\cdot) = \sum_{z_0=1}^{M} \sum_{z_1=1}^{M} \cdots \sum_{z_n=1}^{M} \left[\pi_{z_0} \prod_{j=1}^{n} p_{z_{j-1},z_j} f_{z_{j-1},z_j}(t_j - t_{j-1}) h(t_j|t_{j-1}, S_{t_j})\right]$$

Buried in the likelihood equation above is the assumption that he data is failure truncated, i.e. $t_n = t_{end}$. The situation where $t_n < t_{end}$ is referred to as time truncation and if so, the probability of no failures occurring in $[t_n, t_{end}]$ needs to be incorporated in the likelihood. In an NHPP, the number of events in a time period T is Poisson distributed (Meeker and

Escobar, 1998) with mean $\int_T \lambda(t)dt$. So to account for time truncation in $L(\cdot)$ one also needs to also sum over $\sum_{z_{end}=1}^{M}$ and multiply the product in the sums by $\exp\{-\Lambda([t_n, t_{end}]|S_{t_{end}})\}$.
The second modification to the likelihood that should be considered is the case when the data is left censored meaning that we started observing the pipe after its installation date, so $t_0 = \tau > 0$. In this case, the distribution of the time to the first failure $h(t_1|t_0 = 0)$ should be replaced by $h(t_1|t_0 > 0)$. Formally, we need $\Pr(\tau < t_1 \leq \infty)$ which can be computed by dividing the density function of t_1 by $1 - H(\tau)$ where H is the distribution function of t_1:

$$\begin{aligned} 1 - H(\tau) &= \int_{\tau}^{\infty} \lambda(t_1) \mathrm{e}^{-\int_0^{t_1} \lambda(u)du} dt_1 = \int_{\tau}^{\infty} \gamma\theta t_1^{(\theta-1)} \mathrm{e}^{\boldsymbol{\beta x}} \exp\left\{-\gamma t_1^{\theta} \mathrm{e}^{\boldsymbol{\beta x}}\right\} dt_1 \\ &= \left[-\exp\left\{-\gamma t_1^{\theta}\right\}\right]_{\tau}^{\infty} = \exp\left\{-\gamma\tau^{\theta}\right\} = c(\tau) \end{aligned}$$

This implies that dividing (3) by $c(t_0|S_{t_1} = z_1)$ one can obtain the 'left censored' likelihood $L^*(\cdot)$. Considering now the situation when no failures have been observed, it is fairly easy to see that the contribution to the likelihood will be

$$L^0 = \sum_{z_0=1}^{M} \sum_{z_{end}=1}^{M} \pi_{z_0} p_{z_0,z_{end}} f_{z_0,z_{end}} \exp\{-\Lambda([t_0, t_{end}]|S_{t_0} = z_{end})\}$$

Reintroducing the pipe subscript k and assuming that we have N independent pipes that failed n_k times each, the overall likelihood of the model is

$$\prod_{k=1}^{N} \left[\delta_k L_k^* + (1 - \delta_k) L_k^0\right] \tag{4}$$

where $\delta_k = 0$ if $n_k = 0$ and is equal to 1 otherwise.

3 Model Application

For the sake of neatly deriving and writing the likelihood in (4), we forced the transitions of the hidden chain to actually happen at each failure. This of course is not realistic as transitions could happen at any time, a thing which is taken into account when estimating the parameters. In addition,the likelihood in (4) is very complex to even compute which is why recursive algorithms were used fit the model. These are not mentioned here due to lack of space.
The model is currently being applied to a Canadian distribution network of water pipes. This network consists of 1349 pipes with 5425 recorded failures in the period 1945-2003. The model is implemented within the Bayesian context using MCMC methods.

References

Dong, M. and He, D. (2007). A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical systems and signal processing*, **21**, 2248-2266.

Economou, T., Bailey, T. and Kapelan, Z. (2007). Bayesian modelling of time aggregated water pipe bursts with a zero-inflated, non-homogeneous Poisson process. *Proceedings of the 22nd international workshop on statistical modelling*, 227-232.

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

MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall.

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

Ozekici, S. and Soyer, R. (2003). Bayesian analysis of Markov modulated Bernoulli processes. *Mathematical methods of operations research*, **57**, 125-140.