Stochastic models for prediction of pipe failures in water supply systems

André Damião da Costa Martins
Instituto Superior Técnico, UTL
Lisboa, Portugal
andre.c.martins@ist.utl.pt

ABSTRACT

The failure prediction process plays an important role in infrastructure asset management of urban water systems. This process aims at assessing the future behaviour of a urban water network. However, failure prediction in urban water systems is a complex process, since the available failure data often present a short failure history and incomplete records. In this study, the single-variate Poisson process, the Weibull accelerated lifetime model and the linear extended Yule process, were implemented and explored in order to identify robust and simple models that combine good failure prediction results using short data history. The three models were applied to a Portuguese urban water supply system. The Weibull accelerated lifetime model presented the best results, accurately predicting failures and detecting pipes with high likelihood of failure. Nevertheless, the failure prediction process is based on a Monte Carlo simulation, which can be a time-consuming procedure. The linear extended Yule process could also effectively detect pipes more prone to fail; however, it presented a clear tendency to overestimate the number of future failures. The single-variate Poisson process is the simplest model of the three, and presented lower quality failure predictions. It is noteworthy that no significant difference between the three models results was found when predicting failures in pipes with no failure history.

Keywords

Failure prediction; Infrastructure asset management; LEYP; Poisson process; Water supply networks; Weibull regression.

1. INTRODUCTION

Managing urban water systems is not a simple activity, and, due to several factors, such as climate change, economic restrictions, ageing of the systems, increasing customer demands, it is becoming even more complex. Currently, for environmental and economic reasons, water utilities are increasingly concerned on minimising water losses and in meeting customer demands. Traditional infrastructure asset management methodologies, such as the *like-for-like* strategy, are not adequate for urban water system long-term planning. To help urban water utilities facing these new challenges, new methodologies are being developed, such as the AWARE-P methodology (Alegre *et al.*, 2011).

Failure prediction plays a major role during planning and decision support processes, whether in evaluating different solutions for the identified urban water system problems, whether in evaluating the current system performance under several scenarios. Nevertheless, urban water utilities are only recently becoming aware of the importance of keeping an organised, updated and complete inventory and failure database. Failure data history is very limited in the majority of the urban water systems. For this reason, the task of predicting failures is more difficult than expected. A great challenge is to find a failure prediction model that can produce good quality results, even in the cases of lack of failure data.

The aim of this work is to identify robust and simple models that combine: good predictions for short failure records; robustness when applied to different pipe samples; and simplicity. The studied failure prediction models were: the single-variate Poisson process; the Weibull accelerated lifetime model; and the linear extended Yule process. These models were applied to a Portuguese urban water supply failure data provided by the Serviços Municipalizados de Água e Saneamento de Oeiras e Amadora (SMAS O&A). The prediction results were compared and discussed, regarding the prediction accuracy and the ability of each model to identify the pipes that are more likely to fail; some improvements were also suggested.

All statistical analysis in this work was conducted with the use of the statistical software R (R Development Core Team, 2011).

2. MODELS REVIEW

A variety of failure prediction models have been developed and applied in water research. Deterministic models were developed by Shamir and Howard (1979) and Clark *et al.* (1982), relating the number of failures and the time to next failure, respectively, as functions of explanatory variables. However, to model failures in water systems, stochastic models are believed to be more suitable, since they take into account the random nature of these failures. Hence, two

different types of stochastic models have been developed: single-variate and multivariate models.

In single-variate models pipes are differentiated using grouping criteria rather than covariates. Such models are presented in Herz (1996) and in Gustafson and Clancy (1999). The Poisson process implemented in this work is a single-variate model with a constant failure rate for each group. Single-variate models present the issue of requiring the division of failure data into several homogeneous groups, which can lead to a shortage of records in some groups, where predictions may not be significant.

Multivariate models, which use covariates, allow to differentiate the pipe failure distributions without splitting failure data. In addition, these models allow a better understanding of how the different pipe attributes influence the occurrence of failures. The Proportional Hazard Model (PHM), developed by Cox (1972), is one of the most common multivariate models used in water research.

Accelerated lifetime models define the logarithm of the time to next failure as the linear combination of the covariates vector and a random error variable. A particular case, is the Weibull accelerated lifetime model, applied in Le Gat and Eisenbeis (2000) and studied in this work, when the times between failures are assumed to be Weibull distributed.

New lifetime distributions, with higher complexity, have been developed recently in order to approximate the hazard function to a bathtub curve, such as: the log-extended Weibull distribution (Silva *et al.*, 2009); and the lognormal-power function distribution (Reed, 2011). However, it is not clear that the hazard function for urban water failure data follows a bathtub curve, therefore the higher complexity of these models do not seem to be justified.

Another recent developed approach, suggested by Le Gat (2009) and studied in this work, is the linear extended Yule process. This is a counting process where the rate of the process is given by a linear function of the number of past events.

A more complete description of failure prediction models can be found in Kleiner and Rajani (2001).

3. SINGLE-VARIATE POISSON PROCESS

According to Ross (2006) a Poisson process is a counting process $\{N(t), t \geq 0\}$ with rate γ satisfying the following conditions:

$\forall \gamma \in \mathbb{R}^+$ and $\forall s, t, u, v \in \mathbb{R}^+$, such that $s < t < u < v$,

1. $N(0) = 0$.
2. Independent increments, i.e. $N(t) - N(s) \perp\!\!\!\perp N(v) - N(u)$.
3. $N(t) \sim Poisson(\gamma t)$.

A property of the Poisson process, derived from condition 3, is that the expected number of events is proportional to the observation time, where γ is the coefficient of proportionality and defines the intensity of the process.

When analysing failure data in urban water systems, it is assumed that the number of events (i.e. failures) is also proportional to the length of pipes. The rate of the Poisson process in some pipe i is $\gamma_i = \lambda l_i$, where l_i is the length of the pipe. Therefore, the failure rate per km in the overall system is represented usually by λ (number of failures / km / year), where λ is the proportional coefficient between the rate γ_i of the counting process $N_i(t)$ and the length of the respective pipe, l_i. The failure rate, λ, is estimated using the maximum likelihood method.

Considering a failure data $\boldsymbol{n} = \{n_i\}_{i=1,..,m}$, with n_i = number of observed failures in pipe i during the observation time t_i, the likelihood function (Equation 1) can be written using the Poisson probability function.

$$L(\lambda|\boldsymbol{n}, \boldsymbol{t}, \boldsymbol{l}) = \prod_{i=1}^{m} \frac{e^{\lambda l_i t_i} (\lambda l_i t_i)^{n_i}}{n_i!}, \tag{1}$$

where $\boldsymbol{t} = \{t_i\}_{i=1,..,m}$ and $\boldsymbol{l} = \{l_i\}_{i=1,..,m}$.

The solution of the likelihood maximisation problem (applying the logarithm to Equation (1) and maximising the resulting function) is:

$$\hat{\lambda} = \arg\max_{\lambda \in \Theta} L(\lambda|\boldsymbol{n}, \boldsymbol{t}, \boldsymbol{l}) = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} t_i l_i}. \tag{2}$$

If λ is estimated using the entire data set, then the failure rate will be the same for all pipes, no matter their properties. Nevertheless, the data set can be divided based on the pipe characteristics, such as material and diameter, creating different categories. To create these categories a preliminary analysis of data needs to be performed in order to define categories that gather pipes with similar failure rate. Once the pipe data is categorised, the failure rate λ_k can be estimated for each category C_k using Equation (2).

After λ_k is estimated for each category C_k, probabilities of failure can be calculated for each pipe. The maximum likelihood estimator of the probability of a pipe i, belonging to C_k, to suffer n failures during a time period t is presented in Equation (3) (by the invariance property of maximum likelihood estimators).

$$\widehat{P(N_i(t) = n)} = \frac{e^{\hat{\lambda}_k l_i t} \left(\hat{\lambda}_k l_i t\right)^n}{n!}. \tag{3}$$

The estimator of the expected number of failures of pipe i, during the time period t, is presented in Equation (4); this expected value is used to predict the number of failures for each pipe.

$$\widehat{E[N_i(t)]} = \hat{\lambda}_k l_i t. \tag{4}$$

4. WEIBULL ACCELERATED LIFETIME MODEL

Accelerated lifetime models relate the logarithm of the time to failure with a linear combination of p covariates, $\boldsymbol{x} = [1\ x_1\ x_2\ ...\ x_p]$, and an error term, Z.

$$\ln T = \boldsymbol{x}^\intercal \boldsymbol{\beta} + \sigma Z, \tag{5}$$

where $\boldsymbol{\beta} = [\beta_0\ \beta_1\ ...\ \beta_p]$ are unknown regression parameters and σ is a scale parameter.

Equation (5) shows that the distribution of the random variable Z defines the distribution family of T. In particular, if Z follows the standard Gumbel distribution, then T will be a Weibull random variable, $T \sim \text{Weib}(\sigma^{-1}, e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}})$.

4.1 Estimation of parameters

Parameters σ and $\boldsymbol{\beta}$ were estimated using the maximum likelihood method.

In this model, the sample was composed by the time between recorded failures t_i and the explanatory covariates $\boldsymbol{x}_i$. The likelihood function is the product of the Weibull density function applied to these observed times. Nevertheless, when conducting survival analysis, the decision variable, time between failures, is, in many cases, right censored. Instead of entering the likelihood function with the Weibull density function, right censored times will enter with the Weibull survival function, since the only available information is that the due pipe survived that right censored time. Hence, using a sample of observed times, each associated with p covariates, $\{(t_i, \boldsymbol{x}_i)\}_{i=1,\ldots,n}$, and a sample of censored times, each associated with p covariates, $\{(c_j, \boldsymbol{y}_j)\}_{j=1,\ldots,m}$, the likelihood function can be expressed using the Weibull density and survival functions (Equation 6).

$$\begin{aligned} L(\sigma,\boldsymbol{\beta}|\mathbf{t},\mathbf{c},\boldsymbol{X},\boldsymbol{Y}) &= \prod_{i=1}^{n} f(t_i|\sigma,\boldsymbol{\beta},\boldsymbol{x}_i) \prod_{j=1}^{m} S(c_i|\sigma,\boldsymbol{\beta},\boldsymbol{y}_j) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}} \left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}-1} e^{-\left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}} \\ &\quad \times \prod_{j=1}^{m} e^{-\left(\frac{c_j}{e^{\boldsymbol{y}_j^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}}, \end{aligned} \tag{6}$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \ddots \\ \boldsymbol{x}_n \end{bmatrix}$ and $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \ddots \\ \boldsymbol{y}_m \end{bmatrix}$ are the covariate matrices.

Applying the logarithm, Equation (7) is obtained.

$$\begin{aligned} l(\sigma,\boldsymbol{\beta}|\mathbf{t},\mathbf{c},\boldsymbol{X},\boldsymbol{Y}) &= \sum_{i=1}^{n} \ln \frac{1}{\sigma e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}} + \left(\frac{1}{\sigma}-1\right)\left(\ln t_i - \boldsymbol{x}_i^\intercal\boldsymbol{\beta}\right) \\ &\quad - \left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}} - \sum_{j=1}^{m} \left(\frac{c_j}{e^{\boldsymbol{y}_j^\intercal\boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}. \end{aligned} \tag{7}$$

Unlike the Poisson process, the maximum likelihood estimators can not be analytically expressed. Therefore they must be obtained through numerical maximisation.

In this work, the estimation of parameters was done using function *survreg* of the survival package of the statistical software R.

4.2 Prediction of the number of failures

It is very important to estimate the expected number of failures in a repairable system, which allows calculating the expected total cost of repairs in the repairable system during some time period. One of the biggest drawbacks of the Weibull distributions is that their convolution can not be computed. Thus, the distribution of the number of failures during some time can not be analytically derived. In order to predict the number of failures, Le Gat and Eisenbeis (2000) presented an algorithm based on Monte Carlo simulations, described in this section.

The concept behind Le Gat and Eisenbeis (2000) algorithm is to generate a large number of simulations in each pipe and, consequently, determine the mean number of failures obtained in all simulations for each pipe.

To build simulations over a pipe i, with covariates $\boldsymbol{x}_i$, it is necessary to generate times between failures. The survival function for this pipe is given by Equation (8), where $\eta = e^{\boldsymbol{x}_i^\intercal\boldsymbol{\beta}}$.

$$S(t) = e^{-\left(\frac{t}{\eta}\right)^{\frac{1}{\sigma}}}. \tag{8}$$

Solving the survival function S as a function of t, the expression to generate random times is obtained (Equation 9).

$$t = -\eta\,(\ln S)^{\sigma}. \tag{9}$$

The Monte Carlo simulations in pipe i will be built as follows. Successive times between failures are generated until their sum overlaps the prediction time window. Subsequently, the number of generated times is recorded, ignoring the last one, since it falls outside the prediction window. This experiment is repeated 1,000 times and finally the predicted number of failures will be the mean of all 1,000 simulated numbers of failures. This procedure is repeated for all pipes, obtaining a number of predicted (expected) failures for each one.

4.2.1 Improvements of the WALM prediction process

In this section, it is suggested an improvement of the failure prediction process presented by Le Gat and Eisenbeis (2000). Figure 1 shows the time line of a pipe from the beginning of the failure history until the end of the prediction window.

Figure 1: Time line and failure instants.

Where:
t_0 is the beginning of failure history of the pipe;
t_l is the instant of the last recorded failure in the pipe;
t_s and t_f represent the start and finish instants of the prediction window, respectively;
u is the elapsed time between t_l and t_s.

The approach presented in Le Gat and Eisenbeis (2000) ignores the elapsed time u between the last recorded failure and the beginning of the prediction window. Therefore, the first generated time to failure is counted from t_s, using Equation (9). However, since the times between failures follow a Weibull distribution, the failure counting process does not have the stationary increments property, that is $T|T > u$ and $T + u$ are not equally distributed. The improvement consists in generating a time to failure from t_l using the distribution of $T|T > u$ as described in Equations (10) and (11).

$$P\{T > t|T > u\} = \frac{e^{-\left(\frac{t}{\eta}\right)^{\frac{1}{\sigma}}}}{e^{-\left(\frac{u}{\eta}\right)^{\frac{1}{\sigma}}}} = e^{-\left(t^{\frac{1}{\sigma}} - u^{\frac{1}{\sigma}}\right)\eta^{-\frac{1}{\sigma}}}. \tag{10}$$

$$t = \left(-\eta^{\frac{1}{\sigma}} \ln S + u^{\frac{1}{\sigma}}\right)^{\sigma}. \tag{11}$$

Equation (11) is used only to generate the first time of failure, since u will be 0 when generating the following times to failure. When a pipe has no recorded failures it is assumed that the time of last failure is equal to the beginning of the pipe history, that is $t_l = t_0$ and $u = t_s - t_0$.

4.2.2 *Dynamic variables and prediction of failures*

The number of failures prediction process can become more complex if some of the covariates are dynamic, i.e. they can change during the process, such as pipe age and number of previous failures.

Number of previous failures covariate. This covariate can easily enter the prediction process, being updated in every iteration, whenever a new time to failure is generated. However, when using this covariate there is the risk of the Monte Carlo simulations entering a never ending cycle. Since every covariate acts exponentially on the distribution of the times between failures, each time the number of previous failures is increased the expected time to the next failure drops exponentially. The sum of the expected times to failure constitute a geometric series, with ratio $e^{\beta_{nopf}} < 1$, thus it is a convergent series. Therefore there is no guarantee that the sum of the times will overlap the prediction window and that the simulations will halt.

One way to avoid this issue is to apply the logarithm to the number of previous failures, i.e. to consider the function $\ln(x_{nopf} + 1)$ as a covariate. It can be shown, that for some bounded β_{nopf}, the sum of the expected times between failures constitutes a divergent series, guaranteeing that the simulations will halt for each pipe.

A simpler solution is to use a finite valued covariate. For instance, a binary covariate pf, where $pf = 1$ if the pipe has failed in the past and $pf = 0$ otherwise. The binary covariate may not give the same information that the discrete number of previous failures, nevertheless, it guarantees that the Monte Carlo simulations halt, whatever the estimated $\boldsymbol{\beta}_{pf}$.

Pipe age covariate. Although the pipe age variable is one of the most important explanatory variables, its introduction in the model increases the complexity of the failure prediction process. Unlike the number of previous failures covariate, which only needs to be incremented after each failure time is generated, the age of the pipe variable is continuously increasing with T. So, T can not be generated with a fixed distribution, since this distribution will continuously change throughout the duration of T.

One way to avoid this issue would be to use a fixed covariate that translates the age of the pipe variable, e.g. the installation year or decade. However, the use of a fixed covariate to translate a time dependent covariate may not be realistic. By using the installation year as a covariate, the predicted number of failures from 2005 to 2010 would be the same as the predicted number of failures from 2025 to 2030. Therefore two different approaches are suggested.

The first one is the pipe age at last failure approach. This approach considers the age of the pipe at the last recorded failure as a covariate. This way, the covariate only needs to be updated at every iteration of the failures prediction process, as the number of previous failures covariate. One disadvantage of this approach is that the age of the pipe will not act on the failure distribution as long as the pipe does not fail.

The second one is the age class approach. The aim of this approach is to allow the ageing effect on the failure distribution, independently of the occurrence of previous failures. The age class approach categorises the pipe age variable into different classes, taking into account how this variable influences the failure rate. No many classes should be considered, in order to keep the model's simplicity, e.g. three classes. The approach assumes that the time between failures present a different distribution for each age class. The time between failures in age class k is represented by T_k. In order to estimate the distribution of T_k, the distribution parameters are obtained using the maximum likelihood estimation (Equation 7). However, in this approach, the failure times entering the log-likelihood function will only be the times elapsed in age class k; this means that if some pipe starts the observation in age class k but only fails in class $k+1$, then the time elapsed in class k will enter in the regression of T_k as a right censored time and the time elapsed in class $k+1$ will enter in the regression of T_{k+1} as an observed time.

For each pipe, the prediction window is divided in subwindows, such that the pipe will belong to the same age class in each subwindow. The failure prediction process will behave as before in each prediction subwindow. The predicted number of failures of the pipe is obtained summing the predicted number of failures in each subwindow.

This approach presents some disadvantages. When dividing the failure data into different classes it becomes difficult to understand the age effect on the failure rate. Recent pipe materials are only represented in the first age classes, which makes impossible to predict the number of failures in these pipes when they belong to older classes. This approach is not explored in this study, since there were no significant improvements on the failure prediction results and it is considerably more complex than the pipe age at last failure approach.

5. LINEAR EXTENDED YULE PROCESS

The Linear Extended Yule Process (LEYP) is an extension of the pure birth process, with the intensity function being a linear function of the number of past events and dependent of the age of the process. Let $N(t^-)$ represent the number of events during $[0, t[$ and $dN(t) = N(t + dt) - N(t)$. The LEYP is defined by:

$$\forall_{t\in\mathbb{R}^+}\forall_{j\in\mathbb{N}}\ ,\ \alpha \in \mathbb{R}^+\ ,\ \lambda(t) \in \mathbb{R}^+$$

$$\begin{cases} N(0) = 0 \\ P\{dN(t) = 1 | N(t^-) = j\} = (1 + \alpha j)\lambda(t)dt \end{cases} \tag{12}$$

It is proven by Le Gat (2009) that the probability function of the LEYP is the one presented in Equation (13).

$$P\{N(t) - N(s) = n | N(b) - N(a) = j\} = \\ = \frac{\Gamma(\alpha^{-1} + j + n)}{\Gamma(\alpha^{-1} + j)n!} \frac{(\mu(b) - \mu(a) + 1)^{\alpha^{-1}+j} (\mu(t) - \mu(s))^n}{(\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1)^{\alpha^{-1}+j+n}} \tag{13}$$

where $\mu(t) = e^{\alpha\Lambda(t)}$ and $\Lambda(t) = \int_0^t \lambda(u)du$.

Equation (13) implies that the distribution of failures of a LEYP is a continuous extended Negative Binomial, that is,

$$\{N(t) - N(s) | N(b) - N(a) = j\} \sim$$
$$NB\left(\alpha^{-1} + j, \frac{\mu(b) - \mu(a) + 1}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right).$$

One important feature of a failure prediction model is to allow distinguishing the probability of failure in different pipes, using their different attributes. Thus the intensity function can be based on pipe covariates. Le Gat (2009) suggests that:

$$\lambda(t) = \delta t^{\delta-1} e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}}. \quad (14)$$

With:

$$\Lambda(t) = t^{\delta} e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}}, \quad (15)$$

$$\mu(t) = e^{\alpha t^{\delta}} e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}}. \quad (16)$$

5.1 Estimation of parameters

The LEYP parameters were estimated through the maximum likelihood method. Le Gat (2009) builds the likelihood function and presents the log-likelihood function for the overall system. However, due to some computational problems when maximising the likelihood function, a simplified version is suggested (Equation 17).

$$\ln L(\alpha, \delta, \boldsymbol{\beta} | \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{n}, \boldsymbol{a}, \boldsymbol{b}) =$$
$$\sum_{i=1}^{m} \Bigg(n_i \ln \alpha + \sum_{k=0}^{n_i - 1} \ln\left(\alpha^{-1} + k\right)$$
$$- \left(\alpha^{-1} + n_i\right) h(\alpha, \delta, \boldsymbol{\beta} | \boldsymbol{x}, \boldsymbol{n}, a_i, b_i) + n_i \ln \delta$$
$$+ n_i \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + (\delta - 1) \sum_{j=1}^{n_i} \ln t_{ij} + \sum_{j=1}^{n_i} \alpha t_{ij}^{\delta} e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}} \Bigg), \quad (17)$$

where:
m is the number of pipes;
$\boldsymbol{n} = [n_1 \ ... \ n_m]$, with n_i representing the number of failures in pipe i;
$\boldsymbol{a} = [a_1 \ ... \ a_m]$, with a_i representing the age of pipe i at the beginning of observations;
$\boldsymbol{b} = [b_1 \ ... \ b_m]$, with b_i representing the age of pipe i at the end of observations;
$\boldsymbol{X} = [\boldsymbol{x}_1 \ ... \ \boldsymbol{x}_m]$, with $\boldsymbol{x}_i$ representing the covariate vector of pipe i;
$\boldsymbol{T} = [\boldsymbol{t}_1 \ ... \ \boldsymbol{t}_m]$ with t_{ij} representing the age of pipe i at the j^{th} failure;
and

$$h(\alpha, \delta, \boldsymbol{\beta} | \boldsymbol{x}, \boldsymbol{n}, a_i, b_i) =$$
$$\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}} + \ln\left(1 - e^{\alpha e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}} \left({a_i}^{\delta} - {b_i}^{\delta}\right)} + e^{-\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}}\right) \quad (18)$$

With the parameters estimated, the probabilities of failure can be computed. For instance, the probability of some pipe to fail during $[s, t]$, knowing its failures during $[a, b]$, is given by Equation (19).

$$P\{N(t) - N(s) > 0 | N(b) - N(a) = j\} =$$
$$1 - P\{N(t) - N(s) = 0 | N(b) - N(a) = j\} =$$
$$1 - \left(\frac{\mu(b) - \mu(a) + 1}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right)^{\alpha^{-1} + j}. \quad (19)$$

The expected value can also be computed in order to compare with the results obtained with the WALM and the Poisson process:

$$E[N(t) - N(s) | N(b) - N(a) = j] =$$
$$\left(\alpha^{-1} + j\right) \frac{\mu(t) - \mu(s)}{\mu(b) - \mu(a) + 1}. \quad (20)$$

5.2 Test of significance of the estimated parameters

A statistical hypothesis test was conducted to establish the significance of each estimated parameter of LEYP. The test was based on the likelihood ratio test, using the null hypothesis of each parameter: $\alpha = 0$; $\delta = 1$; $\beta_j = 0, \forall_{j=0,..,p}$.

For every parameter, except α, the log-likelihood function of the null hypothesis was obtained directly from Equation 17, replacing the value of the parameter by the null hypothesis. To test the significance of α the null hypothesis leads to a Non-homogeneous Poisson process (NHPP), with intensity $\lambda(t)$, and a different likelihood function.

The log-likelihood function for the NHPP, using the failure records of all pipes, is presented in Equation (21).

$$\ln L(\alpha, \delta, \boldsymbol{\beta} | \mathbf{T}, \boldsymbol{X}, \mathbf{n}, \mathbf{a}, \mathbf{b}) =$$
$$\sum_{i=1}^{m} \left(\Lambda(a_i) - \Lambda(b_i) + \sum_{j=1}^{n_i} \ln \lambda(t_{ij}) \right). \quad (21)$$

Once defined the likelihood function for the null hypothesis for each estimated parameter, the likelihood ratio, LR, can be computed.

$$LR = \ln \frac{\sup\{L(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta_0\}}{\sup\{L(\boldsymbol{\theta}) | \boldsymbol{\theta} \in \Theta\}}. \quad (22)$$

LR represents the ratio between the values of the maximum likelihood of the null model and the complete model. Wilks proved that $-2 \ln LR$ is asymptotically distributed to a χ_k^2, where k is the difference between the dimensionality of Θ and Θ_0. In this case, the p-value, for each parameter, is obtained as $P\{\chi_1^2 > -2 \ln LR\}$.

6. DATA

The failure data used in this work were provided by SMAS O&A. They were presented in two data tables: the pipe inventory table and the maintenance records table. The pipe inventory table consists of a collection of all pipes identified by a unique code IPID. Each pipe is characterised by several attributes, such as pipe material, diameter, length and installation year.

The maintenance records table consists of all rehabilitation actions operated in the water network system, from 01-01-2001 to 31-03-2011. Each rehabilitation action is identified by a unique code (WOID) and is uniquely associated to the pipe identifier in which the rehabilitation action occurred (IPID). In addition to these attributes, each rehabilitation action is characterised by the failure type that caused the need of rehabilitation and its date and duration.

The only failures considered in this study were pipe breaks.

These two tables were combined using the IPID columns, in order to associate attributes, such as number of failures of each pipe and age of pipe at each failure.

6.1 Water supply network

Failure data of water network systems can present several issues. In the particular case of the data provided by SMAS

O&A, some of the issues are the missing data (empty attribute records) and the anomalous data (inconsistencies between the pipe inventory and the maintenance records). Despite the reduction of the overall data set size, the option, taken in this work, was to remove incomplete and anomalous records (i.e. pipe and failure records that do not have basic information or are inconsistent). Other approaches such as, imputation and survival models with missing data could be considered, but this would bring new complications in data analysis.

The available data is characterised as follows:

- Number of pipes: 11,472.
- Total length: 367km.
- Number of failures: 1,921.
- Number of pipes with no failures: 10,329 (90%).
- Total length of pipes with no failures: 287km (78%).

To characterise the water supply system used in this work, Figures 2 and 3 show the distribution of pipes according to material and installation decade. As can be seen in Figure 2,

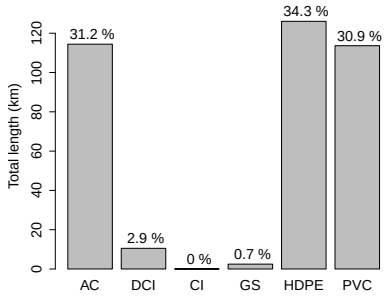

Figure 2: Total length frequencies by pipe material.

the more common pipe materials in this system are the Asbestos Cement (AC), High-Density PolyEthylene (HDPE) and PolyVinyl Chloride (PVC). These three materials together represent more than 95% of the entire water supply network. Ductile Cast Iron (DCI) still has some significance, whereas Galvanised Steel (GS) is of little relevance, representing only 2.5km, and Cast Iron (CI) is completely negligible, since it only represents 0.035km. Therefore, when failure data are divided into different material categories, CI and GS pipes are disregarded.

To assess the age of the water network, the decade of installation was considered rather than the year, because the older pipes have only the record of installation decade. Figure 3 shows the plot of the total length per installation decade. From this plot, it can be seen that the network is relatively recent. The less representative decades are the two oldest ones. Decade 2010 is a special case, as it represents only one year of installation. So, the ageing of pipes may not be very noticeable.

6.2 Pipe variables

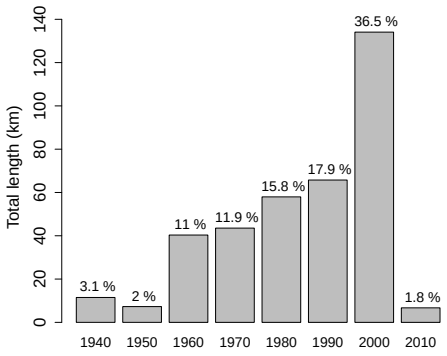

Figure 3: Total length frequencies by installation decade.

The available variables were pipe material, diameter, length, installation year and number of previous failures.

The single-variate Poisson process differentiates pipes using their variables as grouping criteria. Therefore there should be considered a limited number of variables to not produce short categories. The grouping criteria considered in this study were pipe material, three categories of diameter and three categories of length. The installation year was disregarded, because of the high correlation between this variable and the pipe material. In fact, some material categories, such as HDPE, would all belong to the same age category. Furthermore, the effect of the pipe age in the failure rate, in most pipe material categories, was not significant.

Another important variable that was disregarded in this failure model is the number of previous failures. To understand how this variable acts on the future failure rate, it was necessary to divide the training data into to two sets. The first one to count the number of previous failures and divide the pipe data set into categories of previous failures. The second one, to estimate the failure rate for each category of previous failures. As the time period of the failure history was already short, dividing it did not seem a good solution, and so, this variable was disregarded.

The length and the diameter of pipes were used as grouping criteria, because the data analysis showed that these variables had a direct effect on the failure rate.

The Weibull accelerated lifetime model fitted in this work used the pipe material as grouping criteria. The logarithm of length, the diameter, the previous failures indicator variable and the pipe age at the last recorded failure entered the regression model as covariates.

The linear extended Yule process used the pipe material as grouping criteria and considered the logarithm of length and the pipe diameter as covariates. The pipe age and the number of previous failures are already taken into account during the model construction, so they were not used as covariates.

7. COMPARISON OF MODEL PREDICTIONS

To assess the quality of the predictions obtained using the three failure prediction models, the failure data were divided in two data sets: the training and the test sets. Failure distributions were estimated using the training set, whereas failure predictions were carried out using the test set. Finally, failure predictions were compared with the observations of the test set, in order to assess the quality of the predictions. Two methods of choosing the training and test samples were considered: the temporal division and the random division methods.

In the temporal division method the training set is composed of failures occurring before 01-01-2007 in all pipes installed before this date. The test sample comprises the failures that happened between 01-01-2007 and 31-03-2011 in the same pipes of the training sample. This method assesses the ability to predict future failures in a water network with an organised failure history.

However, good failure data history may not be available, due to recent collection and organisation of the failure data, or due to data inconsistencies. In that case, failure predictions should be based on failure data provided by similar water supply systems. The random division method is important to assess the ability of a model, based on a training set, to predict failures in a different pipe data set, with no failure history.

The training sample of the random division method consists of a random 50% sample of all pipes and all failures occurring in those pipes. The test sample is composed by the other 50% pipes and all associated failures. Both training and test time windows go from 01-01-2001 to 31-03-2011.

7.1 Prediction of failures based on a temporal divided sample

In this section the prediction results of the three models are compared when using the temporal division method.

To evaluate the accuracy of each prediction models, the failure predictions are compared with the observed failures in the following way: all pipes of the test sample are sorted according to their predicted number of failures (for each model); after sorting, 0.1-quantiles of the predictions are built; for each prediction quantile the mean of the observed failures is compared with the mean of the predicted failures. The comparison results are presented in Figure 4, where the points represent the observed failures in each predicted quantile and the lines represent the predicted failures. In all prediction models, the mean number of observed failures showed a clear tendency to increase with the quantiles of predicted failures. This indicates that, in all models, the pipes with more predicted failures are, indeed, the ones with more observed failures. The prediction model that seemed more accurate was the WALM, where there is no significant difference between observed and predicted failures in each quantile. LEYP, on the other hand, showed a clear tendency to overestimate the number of failures in the last quantiles, i.e. in pipes more prone to fail.

Failure prediction models can be very useful to prioritise the pipes according to their likelihood of failure. To evaluate this ability, the number of failures that can be avoided, renewing a defined percentage of the water supply system, is compared for each model. If the failure prediction model is good then a large percentage of failures occurring after 2006 can be avoided rehabilitating a small percentage of the water network, after prioritising pipes according to their predicted failure rate. Table 1 presents the results obtained using the Poisson process, the WALM using the pipe age at last failure approach and LEYP. The past individual failure rate was also computed for each pipe, dividing the number of failures occurring before 2006 over its length, and used as a prioritisation criterion.

Table 1: Percentage of avoided failures by prediction model.

Models	Rehabilitated length				
	0.5%	1%	5%	10%	20%
Poisson process	1.7%	3.0%	13.2%	27.2%	49.1%
WALM	4.3%	6.0%	22.3%	35.1%	49.4%
LEYP	5.3%	6.7%	22.2%	32.3%	49.4%
Individual failure rate	4.1%	7.0%	21.3%	31.3%	42.0%

In Table 1 it can be seen that LEYP and WALM results are very close to each other, detecting quite efficiently pipes with higher probability of failing. The Poisson process, on the other hand, lacks the capacity of detecting the pipes that are more likely to fail. Clearly, the small percentage of priority pipes detected by the Poisson process did not suffer as much failures as the ones detected by the two other models. This fact can be explained due to the lack of ability of this model to differentiate pipes. Since it is purely based on categories, every pipe in the same category has exactly the same predicted failure rate. It is important to note that the simple method of prioritising pipes according to their past individual failure rates (Individual failure rate) allowed avoiding a considerable amount of future failures, rehabilitating a small percentage of the water supply network. This shows that pipes more likely to fail are those with higher past failure rates. Therefore, even with few available attributes, water utilities can prioritise effectively rehabilitation actions in a small percentage of the water network, using a well organised failure data. This method of prioritising should only be used to prioritise a portion under 15% of pipes, since 85% of pipes have no failure history, presenting a past individual failure rate of zero.

To assess the accuracy of the predicted number of failures of each model, the absolute error of these predictions was calculated, using Equation (23).

$$\text{Abs. Error} = \sum_{i=1}^{m} |o_i - \hat{e}_i|, \tag{23}$$

where m is the number of pipes; o_i is the observed future failures in pipe i; and $\hat{e}_i$ is the predicted number of future failures in pipe i.

Table 2 presents the absolute error of each model using the prediction results in each pipe material category.

In this table, the LEYP and the Poisson process errors are very similar. WALM is clearly the more accurate model of

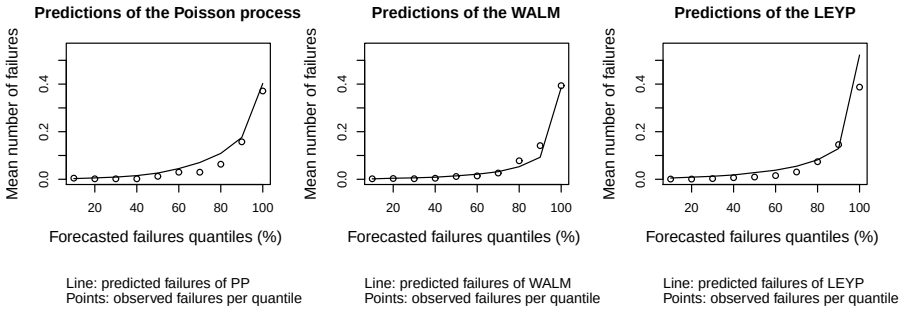

Figure 4: Predictions per forecasted quantiles in temporal divided sample

Table 2: Absolute error of the predicted number of failures per pipe material.

Models	HDPE	PVC	AC	DCI	All pipes
Poisson process	204.6	326.8	785.5	10.6	1327.4
WALM	120.9	259.6	687.6	7.1	1075.3
LEYP	169.8	330.0	782.6	10.1	1292.6

the three, as the absolute error is significantly smaller in each pipe material category.

Based on these comparisons, WALM showed to produce more accurate results. LEYP shared the same ability to detect efficiently pipes with higher likelihood of failing, but it presented a clear tendency to overestimate the number of failures. The Poisson process is limited in prioritising pipes, but general predictions were not very biased. Thus it can be used as a simple model to predict the overall failures, but it would be a poor method to prioritise rehabilitation actions. If a water supply system network has a complete (and long) failure history of pipes, then the past individual failure rate seems to be an efficient prioritising criterion.

7.2 Predictions of failures based on a random sample

In this section the training and test samples were built using the random division method. The training sample is composed by a random 50% sample of all pipes and all associated failures, whereas the test sample is composed by the other pipes and their associated failures. Therefore the training and test windows go now from 01-01-2001 to 31-03-2011.

In the first term of comparison, the test sample was divided according to the forecasted failures quantiles, for each model. The number of observed failures was compared to the predicted failures in each quantile, as conducted in the previous section; results are presented in Figure 5. Once more, WALM was apparently the most accurate model, with the prediction line standing very close to the observed failures. Furthermore, the mean number of observed failures in the last forecasted quantile was higher than in the other models. Comparing Figures 4 and 5, it can be seen that LEYP improved the accuracy of failure predictions. The overestimation that LEYP presented in Figure 4 is no longer noticeable when the test sample consists of pipes with no failure history.

The next term of comparison consisted in estimating the number of failures that could be avoided by renewing some portion of the water supply network. Table 3 shows the number of avoided failures when a small percentage of the test sample is renewed, for each prediction model.

Table 3: Avoided failures in random sample of pipes.

Models	Rehabilitated length				
	0.5%	1%	5%	10%	20%
Poisson process	0.9%	3.0%	13.9%	23.0%	48.4%
WALM	1.0%	2.3%	16.4%	30.8%	51.3%
LEYP	0.5%	2.7%	17.0%	30.3%	51.3%

The results presented in Table 3 show there were no significant differences between the percentages of avoided failures using the three models. Comparing Tables 1 and 3, it can be seen that there is a significant decrease of the avoided failures of LEYP and WALM when prioritising pipes with no failure history. This fact indicates that the main factor of prioritisation of LEYP and WALM were the number of previous failures. Nevertheless, the Poisson results were still slightly worse than the ones presented by the two other models.

The absolute error of the predictions of each model, in each pipe material category, is presented in Table 4.

Table 4: Absolute error of the predicted number of failure per pipe material

Models	HDPE	PVC	AC	DCI	All pipes
Poisson process	242.2	362.7	756.7	17.3	1378.8
WALM	198.9	359.1	753.4	13.2	1324.7
LEYP	205.9	358.7	754.9	13.9	1333.4

In this table, it can be seen that the absolute error values were very close for all prediction models, nevertheless WALM is the one that appeared to present more accurate predictions.

The model results, detecting pipes with higher likelihood of failing, when using a random sample of pipes seemed to be

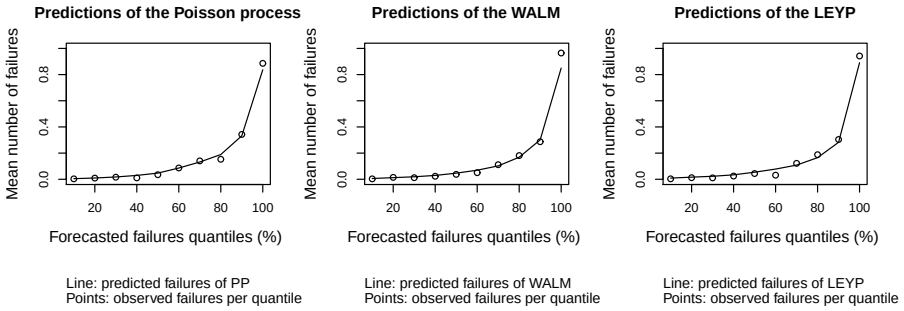

Figure 5: Predictions per forecasted quantiles in a random sample.

worse than using the temporal divided sample (see Tables 1 and 3). However, while the Poisson process results maintain almost the same in both tables, there is a significant downgrade in LEYP and WALM results. The application of the single-variate Poisson process did not take into account the number of previous failures variable, which is one of the most important explanatory variables. Therefore, when the pipe failure history is unknown, performances of LEYP and WALM get more similar to that of the Poisson process.

Despite the importance of the previous failures variable, the general predicted number of failures when no pipe failure history is available was very accurate. In fact, the total number of predicted failures in all models were close to the observed failures: there were 958 observed failures; the Poisson process, the WALM and the LEYP predicted 960, 912 and 946 failures, respectively.

The overestimation in LEYP results, when using the temporal divided sample, is no longer noticeable when predicting pipes with no failure history (i.e. using the random sample). The linear increment of the LEYP intensity (and expected number of failures) with the number of past failures may not be realistic and may explain the mentioned issue.

8. CONCLUSIONS

The three models considered in this study presented good prediction results. WALM appeared to be the best of the three models, because it combined accurate predictions with a good ability of detecting pipes more prone to fail. On the other hand, the Poisson process was the one that presented the worst general performance. Furthermore, WALM and LEYP present an important advantage over the Poisson process: the use of covariates allows a better understanding of the different effect of the pipe variables and does not require the division of the failure data, which could be a problem when dealing with small failure data.

The Poisson process was implemented to study the behaviour of a simple and intuitive prediction model, easy to understand and to implement. The only parameter to be estimated is the failure rates for each defined category, presenting an analytical and intuitive expression (Equation 2), not requiring the use of numerical algorithms. However, its prediction results are slightly worse than the results obtained by the other two models.

The fact that the Poisson process is defined by categories rather than covariates, implies that every pipe in the same category have the same failure rate. This fact leads to a difficulty in differentiating pipes and selecting the pipes that are more likely to fail. That is why the number of avoided failures obtained using this model, is not as high as when using other models (Table 1). The individual past failure rate could be used to prioritise pipes within the same pipe category. Nevertheless, when studying pipes with no failure history, the difference between the Poisson process and the two other methods decreases significantly.

Accelerated lifetime models are different from the other studied models, for they fit the time between failures, rather than the number of failures. In this study, the distribution of the time between failures was chosen to be a Weibull distribution, because it combines several features: it presents an intuitive hazard function (allowing to easily understand the covariates effect); the hazard function is not constant over time, allowing to better fit the failure data; it presents a simple survival function, which is important to obtain the likelihood function (Equation 6) and to generate random times during the Monte Carlo simulations (Equation 9).

Weibull distributions present the disadvantage of not being analytically convoluted. So, it is not possible to find the general number of failures distribution. Thereby, the expected number of failures is obtained using Monte Carlo simulations, which can be a time-consuming procedure.

WALM was the model that presented the best results in all comparisons, combining a great capacity of detecting pipes with higher likelihood of failing, which can be translated by the percentage of avoided failures (Table 1) and accurate predictions (Figure 4 and Table 2).

The linear extended Yule process, proposed by Le Gat (2009), is characterised by the linear relationship between the intensity function and the number of past events. The linear increment of the intensity function with the number of previous failures is probably the reason why LEYP presented a clear tendency to overestimate the number of failures, when the failure history of each pipe is known (Figure 4), but does not overestimate failures in pipes with no failure history (Figure 5). Therefore a non-homogeneous birth process where the intensity function does not depend linearly of the number of past events could be studied. For instance, a

limited function (continuous convergent function or a finite valued function) could be a good solution. However, considering other functions of the number of previous events can increase the complexity of the prediction model.

Despite the overestimation, LEYP presented a really good performance when detecting pipes more prone to fail (Table 1).

As might be expected, it was confirmed that the past failures variable was extremely important when predicting future failures. Pipes that have failed before, present a much higher probability of failing in the future. This may be explained by the fact that, in general, a pipe becomes more fragile after a repair than before the failure happens. Another possible reason is that there are unknown attributes that influence the failure rate, such as environmental, traffic or operating pressure conditions. And so, the higher failure values could be explained by other characteristics associated to the installation site rather than the fragility of those pipes. This is what makes predictions in pipes with no recorded history more difficult and the reason why it is so important to build a complete and trustworthy failure database.

All models require an organised information system with a complete inventory with all pipes well characterised. Even if there is not an extensive pipe database with a lot of variables, good predictions can be done, as proved by the application of the Poisson process, using only three variables. It is not necessary long maintenance records, LEYP and specially WALM offer very good predictions, only using six years of rehabilitation records. What is strictly necessary is to have a complete and up-to-date pipe inventory of all pipes and a reliable rehabilitation database properly linked to the pipe inventory. The information system should be periodically reviewed in order to detect and correct possible inconsistencies.

Acknowledgments

I acknowledge SMAS Oeiras e Amadora by having provided the failure data, with the complete pipe inventory and maintenance records, that was essential to this study.

References

Alegre, H., Covas, D., Coelho, S. T., Almeida, M. C. and Cardoso, M. A. (2011, Sep.). An integrated approach for infrastructure asset management of urban water systems. Mülheim An Der Ruhr, German. IWA 4th LESAM.

Clark, R. M., Stafford, C. L. and Goodrich, J. A. (1982). Water distribution systems: a spatial and cost evaluation. *Journal of the Water Resources Planning and Management Division, ASCE* **108**(WR3), 243–256.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistics Society* **34**(B), 187–220.

Gustafson, J. M. and Clancy, D. V. (1999). Modelling the occurrence of breaks in cast iron water mains using methods of survival analysis. *Proceedings of the AWWA Annual Conference*.

Herz, R. K. (1996). Ageing processes and rehabilitation needs of drinking water distribution networks. *Journal of Water Supply: Research and Technology - AQUA* **45**(5), 221–231.

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

Le Gat, Y. (2009). *Une extension du processus de Yule pour la modélisation stochastique des événements récurrents. Application aux défaillances de canalisations d'eau sous pression.* Ph. D. thesis, Cemagref Bordeaux, Paristech.

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

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reed, W. (2011). A flexible parametric survival model which allows a bathtub-shaped hazard rate function. *Journal of Applied Statistics* **38**(8), 1665–1680.

Ross, S. (2006). *Introduction to Probability Models, Ninth Edition*. Orlando, FL, USA: Academic Press, Inc.

Shamir, U. and Howard, C. (1979). Analytic approach to scheduling pipe replacement. *Journal of American Water Works Association* **71**(5), 248–258.

Silva, G. O., Ortega, E. and Cordeiro, G. M. (2009). A log-extended weibull regression model. *Computational Statistics & Data Analysis* **53**(12), 4482–4489.