Is the service life of water distribution pipelines linked to their failure rate?
Yves Le Gat, Ingo Kropp and Matthew Poulton

ABSTRACT
This paper aims to enable the relevant use of water main service lifetime and failure data to build a medium or long term infrastructure management plan. Firstly, how to estimate the service lifetime distribution of water mains using observations of decommissioning times which are possibly left-truncated and predominantly right-censored, is shown. Three methods are presented: a non-parametric method another based on the parametric Weibull distribution, and a third based on the parametric Herz distribution. An application with actual data related to grey cast iron water mains of two large French and German water distribution networks illustrates the implementation of the theoretical methods. The paper then investigates the link between failure rate and pipe renewal, and discusses the use of observation-based service time survival functions for infrastructure asset management.

Key words | left-truncation, right-censoring, survival data analysis, water main failure rate, water main service lifetime

INTRODUCTION
Detectable leaks, breaks, reduction of hydraulic capacity and degradation of water quality are the main consequences of the ageing of water mains. The rate of occurrence of repairs due to leaks and breaks is a relevant indicator of the performance of water mains, and its reduction constitutes a sound objective for building a long term Infrastructure Asset Management (IAM) plan, as stated by Saegrov (2005). Methodologies to build IAM plans, as presented by Alegre (2008) and Ugarelli et al. (2010), are most often based on the concept of service life survival functions; this is particularly the case for the methodology presented by Herz (1996) and Herz & Baur (2005), used within the framework of the CARE-W FP5 project (see Saegrov 2005; Vanreenterghem-Raven 2006), and based on a cohort survival model and the so-called Herz distribution function, initially presented by Herz (1995). The main difficulty of such a methodology pertains to the need for an accurate service life survival function as an input to cohort survival computations. As building an IAM plan involves comparing various rehabilitation strategies, it is essential to have, as a reference, service lifetime survival functions based on actual data, as well as their link with the failure rate. Good quality estimates of survival curves of drinking water network segments are therefore a pivotal condition for performing relevant long term budget simulations of pipe replacement.

In the next section, this paper states the concepts of left-truncated and right-censored observations, and the concept of survival functions. Three statistical methods for estimating a survival function from observed service lifetimes are then presented in the following three sections and illustrated with two actual datasets related to French and German grey cast iron (CI) water mains: Turnbull's non-parametric estimate, Weibull parametric estimate, and Herz parametric estimate. Based on the comparison of both actual CI datasets the effect on the survival curve of the preferential decommissioning of the most prone to failure water mains is then investigated. The involvement of this effect for improving IAM strategies is finally discussed in the concluding section.
The observation of a random variable (such as the service lifetime) on an incomplete population is said to be ‘left-truncated’; a formal presentation of this concept can be found in Aalen et al. (2008). Informally speaking, estimating the proportion of pipes decommissioned at a given age consists of dividing the number of pipes decommissioned at this age by the size of the initial population, which has to be estimated in the presence of left-truncation. Another important problem, arising when attempting to estimate the distribution of the service lifetime of the water mains, is due to the limited observation window that only allows the observation of the actual decommissioning of a low proportion of the mains; for most of the observed pipes, the actual value of the service lifetime remains unknown, and is to be

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning (unit when relevant)</th>
<th>Range</th>
<th>Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$, $t$</td>
<td>Service time random variable $T$, with possible value $t$ (years)</td>
<td>$[0, +\infty)$</td>
<td>(2)</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of observed segments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n$</td>
<td>Number of groups of segments with same $a_i$, $b_i$, and $c_i$ values</td>
<td></td>
<td>(5)</td>
</tr>
<tr>
<td>$[a_i, b_i]$</td>
<td>Age interval in which the segment group $i \in { 1, \ldots, n }$ is observed</td>
<td>$[0, +\infty]$</td>
<td>(3, 9)</td>
</tr>
<tr>
<td>$a_i, b_i$</td>
<td>$a_i = \min_{i \in {1, \ldots, n }} a_i$, $b_i = \max_{i \in {1, \ldots, n }} b_i$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_i$</td>
<td>Censoring indicator variable $-c_i = 0$ if the segments of group $i$ were decommissioned at age $b_i$, $c_i = 1$ otherwise</td>
<td>$[0, 1]$</td>
<td>(3, 9)</td>
</tr>
<tr>
<td>$e_i$</td>
<td>Number of segments in group $i$</td>
<td></td>
<td>(5, 9)</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of distinct $b_i$ values, provided $c_i = 0$</td>
<td></td>
<td>(3, 4, 5)</td>
</tr>
<tr>
<td>$S(t)$</td>
<td>Survival function, i.e. probability of $T &gt; t$</td>
<td>$[0, 1]$</td>
<td></td>
</tr>
<tr>
<td>$S(t \mid a)$</td>
<td>Conditional survival function given $T \geq a$</td>
<td>$[0, 1]$</td>
<td></td>
</tr>
<tr>
<td>$I()$</td>
<td>Indicator function which takes the value 1 when its argument is true, 0 otherwise</td>
<td>$[0, 1]$</td>
<td>(3, 11)</td>
</tr>
<tr>
<td>$S_N(t \mid a)$</td>
<td>Non-parametric estimate of $S(t \mid a)$</td>
<td>$[0, 1]$</td>
<td>(2, 11)</td>
</tr>
<tr>
<td>$S_\theta(t)$</td>
<td>Parametric estimate of $S(t)$ with parameter $\theta$</td>
<td>$[0, 1]$</td>
<td>(6, 10, 11)</td>
</tr>
<tr>
<td>$S_\theta(t \mid a)$</td>
<td>Parametric estimate of $S(t \mid a)$ with parameter $\theta$</td>
<td>$[0, 1]$</td>
<td>(7, 8, 9)</td>
</tr>
<tr>
<td>$f_\theta(t \mid a)$</td>
<td>Conditional probability density with parameter $\theta$ given $T \geq a$</td>
<td>$[0, +\infty)$</td>
<td>(8, 9)</td>
</tr>
<tr>
<td>$L(\theta)$</td>
<td>Likelihood function of the parameter $\theta$</td>
<td>$[0, +\infty)$</td>
<td>(9)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Scale parameter of the Weibull distribution (years)</td>
<td>$[0, +\infty)$</td>
<td>(6)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Shape parameter of the Weibull distribution</td>
<td>$[1, +\infty)$</td>
<td>(6)</td>
</tr>
<tr>
<td>$\eta$</td>
<td>‘Aging factor’ of the Herz distribution</td>
<td>$[0, +\infty)$</td>
<td>(10)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>‘Failure factor’ of the Herz distribution (years$^{-1}$)</td>
<td>$[0, +\infty)$</td>
<td>(10)</td>
</tr>
<tr>
<td>$\tau$</td>
<td>‘Resistance time’ (shift) of the Herz distribution (years)</td>
<td>$[0, +\infty)$</td>
<td>(10)</td>
</tr>
<tr>
<td>$\Delta_+$</td>
<td>Area between $S_\theta(t)$ and $S_N(t \mid a)$ $S_\theta(a)$ for $t$ values for which $S_\theta(t) \geq S_N(t \mid a)$ $S_\theta(a)$</td>
<td>$[0, +\infty)$</td>
<td>(11)</td>
</tr>
<tr>
<td>$\Delta_-$</td>
<td>Area between $S_N(t \mid a)$ $S_\theta(a)$ and $S_\theta(t)$ for $t$ values for which $S_N(t \mid a)$ $S_\theta(a) \geq S_\theta(t)$</td>
<td>$[0, +\infty)$</td>
<td>(11)</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>Sum of $\Delta_+$ and $\Delta_-$</td>
<td>$[0, +\infty)$</td>
<td>(11)</td>
</tr>
</tbody>
</table>
greater than the age reached at the end of the observation window. The observation of the service lifetime is then said to be ‘right-censored’, as presented by Kalbfleisch & Prentice (1980) and Aalen et al. (2008).

Figure 1 illustrates the concepts of left-truncation and right-censoring. The random variable T stands for the service lifetime, and is observed on a water main installed at age \( t = 0 \) and observed between ages \( a \) and \( b \):

- Any lifetime value \( T < a \) is unobservable, i.e. left-truncated.
- The value \( T = b \) is exactly observed when the actual decommissioning of the segment stops the observation.
- Any value \( T > b \) is right-censored.

**Survival function**

It is a usual practice to describe the distribution of a random lifetime variable by the so-called ‘survival function’, formally defined as the probability to survive beyond a given age:

\[
S(t) = \Pr(T > t)
\]  
(1)

The related survival curve has a starting value \( S(0) = 1 \), is non-increasing, and tends to 0 when \( t \) tends to infinity. The knowledge of survival functions for all categories of segments that make up a water network allows the utility to perform a long term simulation of the budget needs in terms of pipe replacements. The cohort survival model (Herz & Baur 2005) is a good example of the use of survival curves for IAM simulations. It is clear from this study that the quality of the estimates of the survival curves is a crucial condition for obtaining relevant IAM simulations. Three main statistical methods for estimating survival functions from observations of service lifetimes will be examined in the remainder of this paper:

- Turnbull’s non-parametric estimation.
- Parametric modelling based on Weibull distribution.
- Hybrid estimation consisting of fitting Herz survival function to the non-parametric estimate.

It will be shown, in particular, how these methods can be used to correctly process left-truncated observations.

**Turnbull’s non-parametric estimate**

A non-parametric estimate of the survival curve is a decreasing step function that jumps down at each observed lifetime value, while remaining constant (horizontal) between two successive lifetime values. Roughly speaking, the non-parametric method consists of estimating the heights of the jumps. A more formal presentation of the computation procedure necessitates stating of the specific notational conventions.

**Notations**

The random service lifetime variable, denoted \( T \) is observed on a set of \( N \) network segments, split into \( n \) groups. Each group consist of \( e_i \) segments (with \( N = \sum_{i=1}^{n} e_i \)) observed within age interval \([a_i, b_i]\); the random indicator variable \( C \) takes the value \( c_i = 0 \) if the segments of the group \( i \) were actually decommissioned at age \( b_i \), otherwise the value \( c_i = 1 \) if the segments of the group \( i \) have not yet been decommissioned when their observation stopped at age \( b_i \). The first case \( c_i = 0 \) means exact observation \( T = b_i \), whereas the second case \( c_i = 1 \) means right-censored observation \( T > b_i \). Over the set of observed \( b_i \) values, \( m \) are not censored; these values are sorted in increasing order to build up the set \( \{b_j, j = 1, \ldots, m\} \). Let \( a = \min_{i=1, \ldots, n} a_i \) and \( b = \max_{i=1, \ldots, n} b_i \). It is assumed without loss of generality that \( b > t_m \).

**Non-parametric survival function**

The non-parametric survival function \( S_N(t|a) \) is an empirical estimate of the conditional probability \( S(t|a) = \Pr(T > t|T \geq a) \) as no information is available about service lifetimes less than \( a \). No information is available beyond age \( b \) either. As illustrated by Figure 2, \( S(t) \) is then defined over the age interval \([a, b]\) by the vector of jumps \( s = (s_1, s_2, \ldots, s_{m+1}) \) with
Estimation of the non-parametric survival function

The estimation procedure of the vector \( s \) relies on the pivotal work of Turnbull (1976). This method formally consists of calculating both \( n \times (m + 1) \)-matrices of terms:

\[
\begin{align*}
\alpha_{ij} &= c_i I(t_j > b_i) + (1 - c_i) I(t_j = b_i) \\
\beta_{ij} &= I(t_j \geq a_i)
\end{align*}
\]

then both \( n \times (m + 1) \)-matrices of terms:

\[
\begin{align*}
\mu_{ij}(s) &= \alpha_{ij} s_j / \sum_{k \in \{1, \ldots, m+1\}} \alpha_{ik} s_k \\
v_{ij}(s) &= (1 - \beta_{ij}) s_j / \sum_{k \in \{1, \ldots, m+1\}} \beta_{ik} s_k
\end{align*}
\]

and finally the \( (m + 1) \)-vector:

\[
\pi_j = \sum_{i \in \{1, \ldots, n\}} e_i \left( \mu_{ij}(s) + v_{ij}(s) \right) / \sum_{k \in \{1, \ldots, m+1\}} e_k \left( \mu_{ik}(s) + v_{ik}(s) \right)
\]

Starting with an initial estimate \( s^{(0)} \), such that \( s_j^{(0)} = 1/(m + 1) \) for any \( j \), then computing \( s_j^{(1)} = \pi_j(s^{(0)}) \) for any \( j \), and iterating \( r \) times until \( s^{(r)} \approx s^{(r-1)} \), it is proven by Turnbull (1976) that a self-consistent estimate of \( s \) is obtained.

**Weibull parametric model**

Estimating \( S(t) \) with a parametric model requires assuming a theoretical distribution for \( T \). The first one that has been considered in this study is the Weibull distribution that has proven to be practically relevant in many reliability studies for a wide variety of technical systems (see Kalbfleisch & Prentice 1980).

The analytical form of the Weibull survival function depends on two parameters, a scale parameter \( \lambda > 0 \) and a shape parameter \( \delta \geq 1 \), gathered in the parameter vector \( \theta = (\lambda, \delta) \):

\[
S_\theta(t) = \exp(- (t/\lambda)^\delta)
\]

To correctly handle left-truncation and right-censoring the following conditional survival and probability density functions are to be considered:

\[
S_\theta(t | a) = \Pr(T > t | T \geq a) = S_\theta(t) / S_\theta(a) \quad (7)
\]

\[
f_\theta(t | a) = -dS_\theta(t | a) / dt \quad (8)
\]

The estimation of \( \theta \) relies upon the maximum likelihood theory, consisting of finding the parameter value that maximises the likelihood function \( L(\theta) \), i.e. the joint probability of the observed service lifetimes:

\[
L(\theta) = \prod_{i \in \{1, \ldots, n\}} f_\theta(b_i | a_i)^{e_i(1 - c_i)} S_\theta(b_i | a_i)^{e_i c_i} \quad (9)
\]

**Hybrid estimate using Herz distribution**

In practice, the implementation of the parametric method with actual service lifetime data may sometimes lead to lifetime underestimation, as illustrated by the next section. A solution to avoid such a bias with the parametric method consists of estimating the non-parametric survival first, and then fitting a parametric model to the non-parametric curve by least square regression. The parametric model
could be the Weibull one as discussed in the previous section, but the use of the Herz distribution is preferred, as it is the one developed within the cohort survival methodology. The Herz distribution is comprehensively presented by Herz (1997). The Herz survival function with parameter \( \theta = (\eta \gamma \tau) \) is defined by:

\[
S_\theta(t) = 1, \text{ whenever } 0 \leq t \leq \tau, \\
S_\theta(t) = (\eta + 1)/(\eta + \exp(\gamma(t - \tau))), \text{ whenever } t \geq \tau
\]  

(10)

Parameter \( \tau \) stands for a resistance time below which no pipe renewal is assumed to occur; this parameter cannot however be estimated because only \( S_\theta(t)/S_\theta(a) \) can be fitted to Turnbull’s \( S_N(t|a) \).

**RESULTS AND DISCUSSION**

**Non-parametric survival curve versus parametric models**

A way to graphically assess the discrepancy between the parametric and the empirical estimates of the survival curve is to consider the area contained between the graphs of \( S_\theta(t) \) and \( S_N(t|a)S_\theta(a) \) (rescaled empirical estimate). The discrepancy indicators \( \Delta_+ \) (part of the area where the parametric survival is greater than the empirical), \( \Delta_- \) (part of the area where the parametric survival is less than the empirical) and \( \Delta \) (total area) have been designed to that end; they are defined as:

\[
\Delta_+ = \int_{[a,b]} (S_\theta(t) - S_N(t|a)S_\theta(a))I(S_\theta(t) \geq S_N(t|a)S_\theta(a))dt \\
\Delta_- = \int_{[a,b]} (S_N(t|a)S_\theta(a) - S_\theta(t))I(S_N(t|a)S_\theta(a) \geq S_\theta(t))dt \\
\Delta = \Delta_+ + \Delta_-
\]  

(11)

Step curves in Figures 3 and 4 illustrate the Turnbull’s estimates obtained, with respectively, French CI data observed between ages 25 (last installations of 1971 observed in 1995) and 153 years (first installations of 1855 observed in 2007), and German CI data observed between ages 15 (last installations of 1970 observed in 1985) and 106 years (first installations of 1896 observed in 2002). French CI data consist
of \( N = 83,000 \) segments split into \( n = 1,022 \) groups according to their ages at the beginning and end of observation, and whether their decommissioning was observed or not; \( m = 110 \) distinct ages at decommissioning were observed. For German CI data, these values are \( N = 8,500 \) segments, \( n = 796 \) groups and \( m = 84 \) distinct ages at decommissioning.

Figures 3 and 4 also illustrate, for French and German CI segments, the rather close agreement between the Weibull survival curve \( S_\theta(t) \) and \( S_N(t|a)S_\theta(a) \). For French CI segments, the agreement seems to be satisfactory except for age values greater than 110 years, poorly represented in the dataset; \( \Delta_+ \) and \( \Delta_- \) were computed without ages greater than 133 years (in order to discard the last big step of the empirical survival), and show a loss of balance between both values. For German CI segments, the Weibull model seems to systematically underestimate the empirical survival (\( \Delta_+ = 0 \)). This systematic underestimation has been confirmed with random synthetic data following a known Weibull distribution; it may be due to the maximum likelihood estimation method itself, and is to be investigated in future research.

The non-parametric estimate has the great advantage that it faithfully represents the data. An important drawback however is that it cannot tell anything outside the ages \([a, b]\). That is a good reason for attempting to complete it with a parametric estimate that has the disadvantage of relying on a strong modelling hypothesis, but allows, on the other hand, the estimation of the proportion of segments decommissioned before age \( a \), and also predictions beyond age \( b \). It is reasonable to trust these predictions outside \([a, b]\) provided it can be ensured, at least graphically, that the parametric estimate is sufficiently close to the non-parametric one inside \([a, b]\).

The hybrid method using Herz distribution has also been implemented with both French and German CI datasets. In this case, a value \( \tau = 10 \) (years) has been assumed for the Herz resistance parameter. Other values within the likely range between 0 and 15 years (the smallest \( a \) value among both datasets) have been tried, and the goodness-of-fit seems to be insensitive to this parameter. Table 1 shows the Herz distribution parameters estimated for both CI datasets, as well as the main distribution quantiles. Figures 5 and 6 illustrate the shapes of the Herz survival curves, rather similar to the Weibull ones. The comparison with Figures 3 and 4 illustrates the bias correction, as \( \Delta_+ \) and \( \Delta_- \) values appear to be much better balanced.

The German CI segments show shorter service times compared with their French counterparts; this somewhat surprising difference, considering the similarity of material technologies and network operations, is further discussed in the next section.

**Preferential decommissioning of less reliable segments**

An important question that arises when examining historical decommissioning data is to guess whether decommissioning
has mainly been motivated by road maintenance decisions, or has also been at least partly driven by the repeated failures of water mains. It may be not possible to give a general answer, but when extensive datasets (83,000 French and 8,500 German CI segments) are available, it is highly informative to compare the distributions of the failure rates between segments still in service at the end of their observation period versus decommissioned segments. Figures 7 and 8 compare:

- the proportions of segments that did not fail within their observation period;
- the empirical distribution functions of non-zero failure rates.

Both Figures 7 and 8 clearly reveal that decommissioned pipes had much higher failure rates than pipes still in service. Water main replacements in big western European cities are indeed reputed to be mainly driven by road works, with little consideration of the pipe condition; the deviation between the failure rate distribution for still in service (grey curve) and decommissioned mains (dark curve) in Figures 7 and 8 is somewhat surprising. A partial explanation could perhaps be found by considering that the condition of road surfaces is frequently highly impacted by the repair works of water main failures. An important consequence of the
preferential decommissioning of the less reliable mains is the so-called selective survival phenomenon, which is referred to in the next section.

It is also highly informative to observe that the contrast between the failure rates of segments still in service versus decommissioned segments is much higher for German CI segments than for French ones. This suggests a much lower tolerance for repeated failures in the German utility compared with the French one, without prejudice to any significant difference in national practices. Moreover the average failure rate is 0.24 km\(^{-1}\) year\(^{-1}\) for French CI segments versus 0.20 for German ones. This explains well why German CI segments have a much shorter service time distribution as illustrated by the smaller quantiles in Table 1. This makes clear that assets like water mains cannot be characterised by any intrinsic predetermined longevity. A sound pipe renewal policy thus cannot consist of replacing pipes as soon they reach some predetermined lifetime: at a given age, the reliability of mains is indeed likely to greatly vary, according to their characteristics, their environment and their failure history (see e.g. Kropp et al. (2009)).

CONCLUSION AND PERSPECTIVES FOR IMPROVING IAM POLICIES

This paper has shown how to estimate a non-parametric survival function for water mains using commonly available left-truncated and right-censored information. This is a valuable basis that allows the estimation of a parametric survival function that can be extrapolated from the observed age interval, and consequently used for IAM simulations.

The study has also incidentally revealed a possible bias of the Weibull parametric maximum likelihood estimate, which raises a theoretical problem worth investigating in the future.

The indication that water main replacements are highly influenced by segment failure rates is a sound argument for proposing parametric survival curves such as those obtained with the Herz model as relevant inputs for IAM simulations. This approach could be further developed by modelling the link between the survival function of the service time and the distributions of the failure rates of still in use and decommissioned pipes. The concept of selective survival developed by Le Gat (2009) is a possible way; it is based on the assumption that observed segments, and especially the oldest ones, have survived until the observation period due to their relatively lower failure rate. This perspective is complementary to the use of a parametric model of survival curve derived from a failure model as the one proposed by Kropp et al. (2009) based on a linear extension of the Yule process.

REFERENCES


First received 8 August 2011; accepted in revised form 23 July 2012