Is the service life of water distribution pipelines linked to their failure rate?

Yves Le Gat[a], Ingo Kropp[b], Matthew Poulton[c]

[a] REBX, Cemagref-Bordeaux, 50 avenue de Verdun, 33612 Cestas Cedex (France)
Email address: yves.legat@cemagref.fr - URL: http://www.cemagref.fr
[b] BAUR+KROPP, Nieritzstraße 5, 01097 Dresden (Germany)
Email address: kropp@baur-kropp.de - URL: http://baur-kropp.de
[c] WTSim, Impasse Fauré, 33000 Bordeaux (France)
Email address: matthew.poulton@wtsim.com - URL: http://www.wtsim.com

Abstract

This paper aims at helping the relevant use water main service lifetime and failure data to build a medium or long term infrastructure management plan. It is first shown how to estimate the service lifetime distribution of water mains using observations of decommissioning times which are possibly left-truncated and predominantly right-censored. Three methods are presented, a non-parametric one, an other 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 asset management.

Keywords: Left-truncation, Right-censoring, Non-parametric estimation, Weibull distribution, Herz distribution, Water Main Service Time, Water Main Failure rate, Cohort Survival Model, Asset management.

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 asset management (AM) plan. Methodologies to build AM plans are most often based on the concept of service life survival function; this is particularly the case of KANEW, presented by Herz and Baur (2005), used within the framework of the CARE-W FP5 project (see Saegrov, 2005), and based on a cohort survival model and the so-called Herz distribution function (see Herz, 1995). The main difficulty of such a methodology pertains to the need of an accurate service life survival function as an input to cohort survival computations. As building an AM plan involves comparing various rehabilitation strategies, it is essential to have as a reference service life 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 replacements.

This paper firstly states the concepts of left-truncated and right-censored observations, and the concept of survival function in the next section. Three statistical methods for estimating a survival function from observed service lifetimes are then presented and illustrated with two actual datasets related to French and German grey cast iron (CI) water mains in the next three sections: 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 AM strategies is finally discussed in the concluding section.

LEFT-TRUNCATED AND RIGHT-CENSORED OBSERVATIONS

Many water mains were most often laid out a long time before water utilities began to keep up to date electronic archives of maintenance operations. Decommissioning data necessary to estimate the distribution of the service lifetime of the water mains are therefore only available relatively recently (generally 1985 or later), and the mains observed which were laid prior to this year can only be considered as the remnants of their cohort (defined as the set of pipes laid the same year). The observation of a random variable (such as the service lifetime) on an incomplete population is said to be “left-truncated”. 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 as well in 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 only sure to be strictly 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”.

Fig. 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 laid out 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.

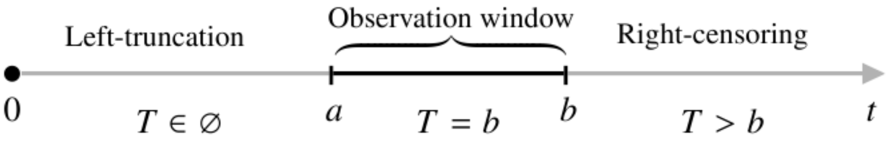

Figure 1: Concepts of left-truncation and right-censoring

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:

$$S(t) = \Pr\{T > t\}$$

The related survival curve has a starting value $S(0) = 1$, is non-increasing, and tends to 0 when *t* tends to infinity. Fig. 3 and 4 illustrate the appearance of such curves. 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 KANEW software is a good example of such use of survival curves for AM simulations. It is consequently clear that the quality of the estimates of the survival curves is a crucial condition for obtaining relevant AM simulations. We will then examine in the sequel of this paper three main statistical methods for estimating survival functions from observations of service lifetimes:

- Turnbull's non-parametric estimation,
- parametric modelling based on Weibull distribution,
- hybrid estimation consisting of fitting Herz survival function to the non-parametric estimate.

We will, in particular, see 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 consist of estimating the heights of the jumps. A more formal presentation of the computation procedure necessitates *ante omnia* to state specific notational conventions.

Notations

The random service lifetime variable, denoted T is observed on a set of N network segments, gathered in n groups. Each group consist of e_i segments (with $\sum_{i=1}^{n} e_i = N$) 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 be 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 $\{t_j, j = 1, \ldots, m\}$. Let $a = \min_{i=1}^{n}(a_i)$ and $b = \max_{i=1}^{n}(b_i)$. It is assumed without loss of generality that $b > t_m$.

Non-parametric survival function

The non-parametric survival function is an empirical estimate of the conditional probability $S(t \mid a) = \Pr\{T > t \mid T > a\}$ as no information is available about service lifetimes less than a. And no information as well is available beyond age b. As illustrated by fig. 2, $S(t)$ is then defined on the age interval $[a, b]$ by the vector of jumps $\boldsymbol{s} = (s_1 s_2 \ldots s_{m+1})$ with $\sum_{j=1}^{m+1} s_j = 1$:

$$\begin{cases} S(t) = 1, \text{ whenever } t \in [a, t_1[\\ S(t) = 1 - s_1, \text{ whenever } t \in [t_1, t_2[\\ \ldots \\ S(t) = 1 - \sum_{k=1}^{j} s_k, \text{ whenever } t \in [t_j, t_{j+1}[\\ \ldots \\ S(t) = s_{m+1}, \text{ whenever } t \in [t_m, b] \end{cases}$$

Estimation of the non-parametric survival function

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

$$\alpha_{ij} = c_i \mathrm{I}\left(t_j > b_i\right) + (1 - c_i)\mathrm{I}\left(t_j = b_i\right)$$
$$\beta_{ij} = \mathrm{I}\left(t_j \geqslant a_i\right),$$

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

$$\mu_{ij} = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m+1} \alpha_{ik} s_k}$$
$$\nu_{ij} = \frac{(1 - \beta_{ij}) s_j}{\sum_{k=1}^{m+1} \beta_{ik} s_k},$$

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

$$\pi_j(\boldsymbol{s}) = \frac{\sum_{i=1}^{n} e_i(\mu_{ij} + \nu_{ij})}{\sum_{k=1}^{m+1} \sum_{i=1}^{n} e_i(\mu_{ik} + \nu_{ik})}$$

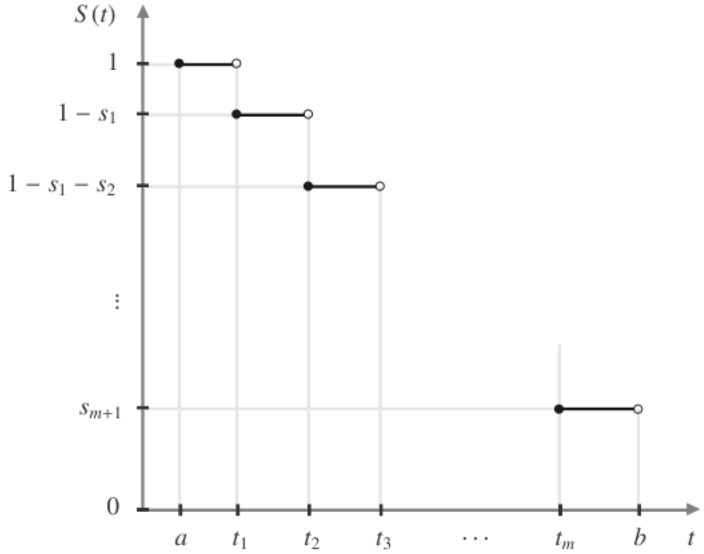

Figure 2: non-parametric Survival Function

Starting with an initial estimate $\boldsymbol{s}^{(\mathbf{0})}$, such that $s_j^{(0)} = 1/(m+1)$ for any j, computing $s_j^{(1)} = \pi_j(\boldsymbol{s}^{(\mathbf{0})})$ for any j, and iterating until $s_j^{(N)} \simeq s_j^{(N-1)}$, it is proven by Turnbull (1976) that a self-consistent estimate of $\boldsymbol{s}$ is obtained. Step curves in fig. 3 and 4 illustrate this kind of result 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).

WEIBULL PARAMETRIC MODEL

Estimating $S(t)$ with a parametric model requires assuming a theoretical distribution for T. The one that has been chosen 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.

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

$$S_{\boldsymbol{\theta}}(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\delta}\right)$$

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

$$S_{\boldsymbol{\theta}}(t \mid a) = \Pr\{T > t \mid T \geqslant a\} = S_{\boldsymbol{\theta}}(t)/S_{\boldsymbol{\theta}}(a)$$
$$f_{\boldsymbol{\theta}}(t \mid a) = -\mathrm{d}S_{\boldsymbol{\theta}}(t \mid a)/\mathrm{d}t$$

The estimation of $\boldsymbol{\theta}$ relies on the maximum likelihood theory, consisting of finding the parameter value that maximizes the likelihood function, *i.e.* the joint probability of the observed service lifetimes:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(b_i \mid a_i)^{e_i(1-c_i)} S_{\boldsymbol{\theta}}(b_i \mid a_i)^{e_i c_i}$$

Fig. 3 and 4 illustrate, for French and German CI segments, the rather close agreement between the Weibull survival curve $S_{\theta}(t)$ and the rescaled non-parametric estimate $S(t \mid a)S_{\theta}(a)$. For French CI segments, the agreement seems to be satisfactory except for age values greater than 110 years, poorly represented in the dataset; for German CI segments however, the Weibull model seems to systematically underestimate the empirical survival. 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.

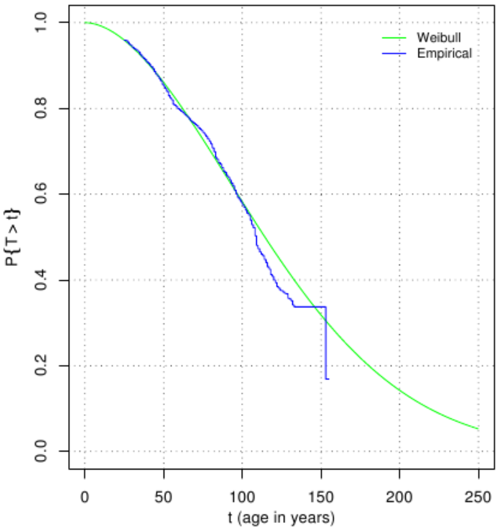

Figure 3: Weibull parametric and Turnbull's non-parametric estimates of the survival curve for French CI segments observed between 1995 and 2007

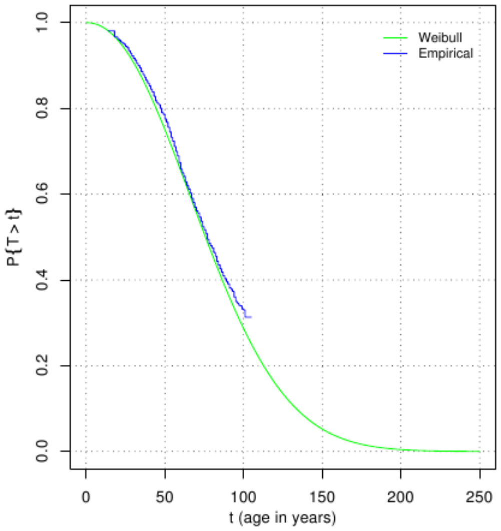

Figure 4: Weibull parametric and Turnbull's non-parametric estimates of the survival curve for German CI segments observed between 1985 and 2002

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 can, on the other hand, estimate the proportion of decommissioned segments before age a, and make 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 $[a, b]$.

HYBRID ESTIMATE USING HERZ DISTRIBUTION

A solution to avoid a possible underestimation bias with the parametric method consists of estimating the non-parametric survival first, and then in fitting a parametric model to the non-parametric curve by least square regression. The parametric model could be the Weibull one as in previous section, but we have preferred the use of Herz distribution, as it is the one developed within the KANEW methodology. The Herz distribution is comprehensively presented by Herz (1996). The Herz survival function is defined by:

$$S(t;\eta,\gamma,\tau) = 1, \text{ whenever } 0 \leqslant t \leqslant \tau$$

$$S(t;\eta,\gamma,\tau) = \frac{\eta + 1}{\eta + \exp(\gamma(t-\tau))}, \text{ whenever } t \geqslant \tau$$

Parameter τ stands for a resistance time below which no pipe renewal is assumed to occur; this parameter cannot however be estimated because only $S(t;\eta,\gamma,\tau)/S(a;\eta,\gamma,\tau)$ can be fitted to Turnbull's $S(t \mid a)$. In the case of our actual French and German CI data, we assume a value $\tau = 10$ (years). 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. Fig. 5 and 6 illustrate the shapes of the Herz survival curves, and especially the globally shorter service times of the German CI segments.

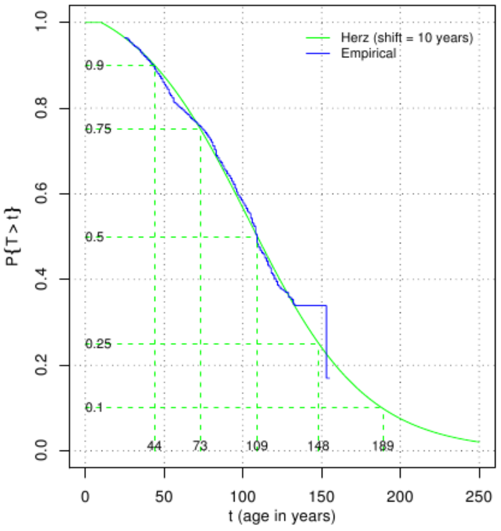

Figure 5: Herz parametric and Turnbull's non-parametric estimates of the survival curve for French CI segments observed between 1995 and 2007

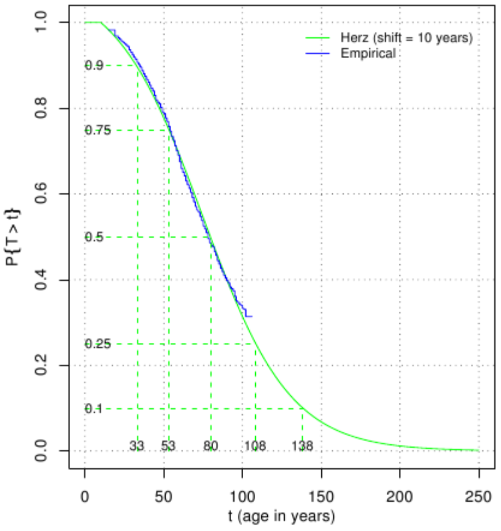

Figure 6: Herz parametric and Turnbull's non-parametric estimates of the survival curve for German CI segments observed between 1985 and 2002

Table 1: Herz distribution parameters and quantiles (years) for French and German CI datasets

Statistics	French CI data	German CI data
η	12.41	10.26
γ	0.027	0.036
τ	10	10
Q90%	44	33
Q75%	73	53
Q50%	109	80
Q25%	148	108
Q10%	189	138

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 is maybe 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 window versus segments decommissioned. Fig. 7 and 8 compare:

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

Both fig. 7 and 8 clearly reveal that decommissioned pipes had much higher individual failure rates than pipes still in service. As water main replacements are reputed to be mainly driven by road works in big western European cities, this result is a bit surprising. It is however less surprising if we consider that the condition of road surfaces is frequently highly impacted by the repair works of water main failures.

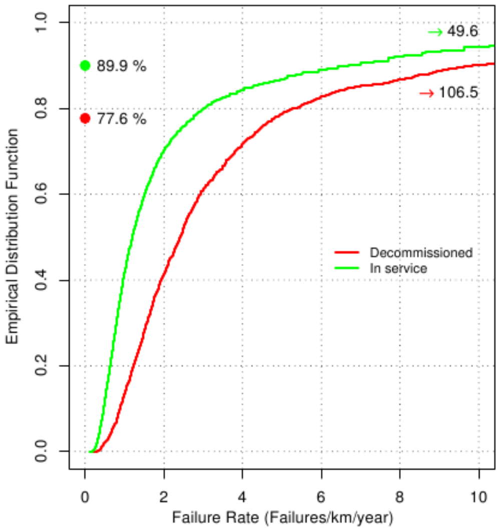

Figure 7: Comparison of failure rates for French CI segments decommissioned vs still in service - proportions of segments with no failure, empirical distribution functions for segments that failed at least once

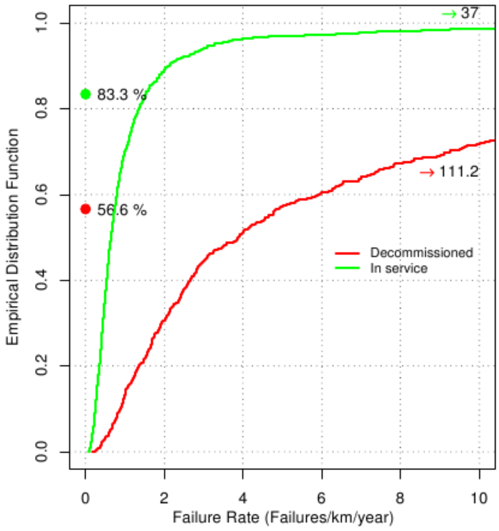

Figure 8: Comparison of failure rates for German CI segments decommissioned vs still in service - proportions of segments with no failure, empirical distribution functions for segments that failed at least once

It is also highly informative to observe that the contrast between the failure rates of still in service and 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 to the French one. 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 *a priori* characterized by any intrinsic predetermined longevity. A sound pipe renewal policy cannot consequently consist of replacing pipes as soon they reach some predetermined lifetime.

CONCLUSION AND PERSPECTIVES FOR IMPROVING AM 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 then the estimation of a parametric survival function that can be extrapolated out of the observed age interval, and consequently used for AM simulations.

The study has also incidentally revealed a possible bias of the Weibull parametric MLE, which raises a theoretical problem worth to be investigated in the future.

The evidence that water main replacements have been highly influenced by the segment failure rates is a sound argument for proposing parametric survival curves as those obtained with the Weibull model as relevant inputs for AM simulations using a software like KANEW. Simulations are particularly ensured to account for segment reliability, and hence to not only mimic pipe replacements based on the sole age of the assets. This approach could moreover 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 indeed on the assumption that observed segments, and especially the oldest ones, have survived until the observation window 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

Herz, R., 1995. Alterung und Erneuerung von Infrastukturbeständen - ein Kohortenüberlebensmodell. *Jahrbuch für Regional-Wissenschaft* 14-15, 5–29.

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

Herz, R., Baur, R., 2005. Erneuerungsstrategie und Prioritäten bei Rehabilitationsplanung von Rohrleitungsnetzen. *DVGW - Energie Wasser Praxis* 56 (5), 22–27.

Kropp, I., Gat, Y. L., Poulton, M., 2009. The application of the leyp failure forecast model at the strategic asset management planning level. In: *Proceedings of LESAM 2009*, Miami.

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, Engref - Paris.

Saegrov, S., 2005. *Care-w: Computer Aided Rehabilitation for Water Networks*. IWA Publishing.

Turnbull, B., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Serie B (Methodological)* 38 (3), 290–295.