Water Resour Manage (2011) 25:1229–1238
DOI 10.1007/s11269-010-9679-8

Water Network Assessment and Reliability Analysis by Use of Survival Analysis

Symeon E. Christodoulou

Received: 31 October 2009 / Accepted: 1 June 2010 /
Published online: 19 June 2010

Abstract A holistic and sustainable strategy for the management of urban water distribution networks should be composed of two equally important pillars: (1) efficient methods for monitoring, repairing or replacing aging infrastructure, and (2) effective tools for modelling the deterioration in the network and for proactively assessing the risk of failure of its components so as to devise preventive measures for avoiding such failures. The paper presents a framework for devising such a proactive risk-based integrity-monitoring strategy for the management of urban water distribution networks. The framework presented is based on a combination of artificial neural network, parametric and nonparametric survival analysis and it is utilized in the estimation of time-to-failure metrics for pipe networks.

Keywords Water distribution networks · Decision support system · Risk analysis · Survival analysis

1 Introduction

Each year, thousands of kilometres of pipes across the globe are upgraded or replaced in an attempt to maintain the uninterrupted transport of water. The goal is threefold: (1) to safeguard the health of urban populations, (2) to increase the reliability of the pipe networks and the service provided to consumers, and (3) to increase the sustainability and cost-efficiency of the operations and maintenance of such networks. In attaining this goal, water distribution agencies are required to develop and implement new methods for monitoring, repairing or replacing aging infrastructure, as well as modelling deteriorating infrastructure conditions and proactively devising strategies to keep the networks in operation. In essence,

S. E. Christodoulou (✉)
Department of Civil and Environmental Engineering, University of Cyprus,
P.O. Box 20537, Nicosia 1678, Cyprus
e-mail: schristo@ucy.ac.cy

water distribution agencies are faced with the increasingly more complex task of intelligently and efficiently assessing (or modelling) the condition of a pipe network while managing the network in ways that maximize its reliability and minimize its operational and management costs. The question that usually arises is whether an organization should repair or replace deteriorating water mains and, in either case, what should the sequence of any such repairs be as part of a long-term network rehabilitation strategy.

The work presented herein reports on the investigation of possible risk-of-failure parameters, the application of survival analysis techniques and their utilization by a multi-criteria decision support system (DSS) for the modelling of water distribution networks.

2 State of Knowledge

Research to-date has helped identify a number of potential risk factors that contribute to pipe breaks, ranging from time-invariant to time-dependent factors.

Most of related past studies have focused on identifying and developing the relationships between break rates and influential risk factors contributing to the risk of failure in the piping network. Examples of risk factors studied in the past are the age, diameter and material of a pipe, the corrosiveness of surrounding soil, the operating pressure and temperature in the piping network, possible external loads and the history of pipe breaks. Most such studies have shown a relationship between failure rates and the time-to-failure, with some studies suggesting methods to optimize a pipe's replacement time. For example, Shamir and Howard (1979) reported an exponential relationship and Clark et al. (1982) developed a linear multivariate equation to characterize the time from pipe installation to the first break and a multivariate exponential equation to determine the breakage rate after the first break. Andreou et al. (1987) suggested a probabilistic approach consisting of a proportional hazards model to predict failure at an early age, and a Poisson-type model for the later stages. They further asserted that data stratification (based on specific parameters) would increase the accuracy of the model. A non-homogeneous Poisson distribution model was later proposed by Goulter and Kazemi (1988) to predict the probability of subsequent breaks given that at least one break had already occurred. Finally, Kleiner and Rajani (1999) developed a framework to assess future rehabilitation needs using limited and incomplete data on pipe conditions.

Prasad et al. (2003) outlined a multi-objective genetic algorithm method and introduced a new reliability measure, called network resilience. The measure tried to provide surplus head above the minimum allowable head at nodes and reliable loops with practicable pipe diameters. Fattahi and Fayyaz (2009) considered integrated urban water management as a multi-objective problem and presented a mathematical model based on compromise programming for optimizing this multi-objective problem. Three objectives involving water distribution cost, leakage water and social satisfaction level were considered. Kanakoudis and Tolikas (2004) addressed the question 'what should system managers prefer? Systems that never fail or systems that remain safe during a failure?' by presenting a method that hierarchically analyses the possible preventive maintenance actions in a water system. The whole attempt was based on indices that assess the performance level of a system prior to any action being taken, in conjunction with a techno-economical analysis that takes into account the costs associated to the repair or replacement of failing system components. A study on New York City pipe inventories and models for the structural degradation of urban water distribution systems was also reported by Vanrenterghem-Raven et al. (2004), Aslani (2003) and Christodoulou et al. (2003). The knowledge gained by the New York City case study was furthered and reported upon by Christodoulou et al. (2007) in a developed framework for integrated GIS-based management, risk assessment and prioritization of water leakage actions. The study was also expanded to include neurofuzzy decision support systems (Christodoulou and Deligianni 2009). Frequently used methods for the risk analyses of water supply systems were also presented by Tuhovcak et al. (2006). The methods address the identification of qualitative and quantitative risks posed by the individual system components, the evaluation methods and the interpretation of results, with an emphasis on the Hazard Analysis and the Critical Control Points (HACCP) method. A similar approach was also utilized by Park (2008) who presented a method to assess and track changes in the hazard functions between water main breaks by using the proportional hazards model. The method provides a systematic framework for analyzing changes of the hazard functions as more breaks occur, and for identifying a critical point at which the hazard function changes to a different functional form. More recently, a general theory of vulnerability of water pipe networks (termed TVWPN) was outlined by Pinto et al. (2010). The theoretical concepts of TVWPN have a basis in the structural vulnerability theory and the fundamental contribution of this theory is to help design water pipe networks that are more robust against damage to the pipelines by use of an analysis of the network's form.

3 Nonparametric Survival Analysis

In most past cast studies the analysis of the risk of failure has focused on statistical analysis of the observed failure times and the application of parametric models to such data. Most notably, failure times were often fitted to Weibull-based parametric models and analyzed through data-stratification and the examination of the various risk factors one at a time. The underlying premise is that a pipe's failure time can be approximated by parametric equations relating it to the presumed risk-of-failure factors through mathematical equations. Of course, if such mathematical relationships can not be identified by use of known probability distributions then the level of difficulty in estimating the underlying multi-factored probability distribution function increases significantly.

A possible solution to the problem of devising time-related risk-of-failure relationships can be sought in a method termed "survival analysis". Survival analysis is a branch of statistics dealing with deterioration and failure over time and involves the modelling of the elapsed time between an initiating event and a terminal event (Cox 1972; Klein and Moeschberger 1997; Hintze 2006). In the case of piping networks such initiating events can be the installation of a pipe, a water-leak observation or the start of a pipe treatment, and cases of terminal events can be a relapse of a previous leak, a fix or a failure. The method is based on estimating the reliability of a system and its lifetime subject to multiple risk factors, and aims to provide answers on the population fraction that survives past an expected lifetime, the effect of the various risk factors on the system's lifetime and on the probability of survival, and the expected mean time to failure (Hosmer et al. 2008; Lee and Wang 2003; Hintze 2006). The data values used in the analysis are a mixture of complete and censored observations. In the former case a terminal event is thought to have occurred, while in the latter case a terminal event has not occurred. A terminal event is assumed to occur just once for each subject.

The basic equations used in survival analysis are listed below (Eqs. 1, 2, 3, 4 and 5), with T denoting the elapsed time until the occurrence of a specified event, $f(t)$ being the probability density function (denoting the probability that an event occurs at time t), and $F(t)$ being the cumulative distribution function (denoting the probability that an individual survives until time t). Similarly, a survival function, $S(t)$, denotes the probability that an individual survives beyond time t and the hazard rate, $h(T)$, denotes the probability that an individual at time T experiences the event in the next instant. A cumulative hazard function, $H(T)$, is the integral of $h(t)$ from 0 to T.

$$F(t) = \int_0^t f(x)dx \tag{1}$$

$$S(t) = \int_T^\infty f(x)dx = 1 - F(T) \tag{2}$$

$$S(t) = \exp\left[-\int_0^T h(x)dx\right] = \exp\left[-H(T)\right] \tag{3}$$

$$h(T) = f(T)\big/S(T) \tag{4}$$

$$H(T) = \int_0^T h(x)dx = -\ln\left[S(T)\right] \tag{5}$$

The survival function is usually the primary quantity of interest and it is estimated by use of kernels such as the Epanechnikov kernel (Epanechnikov 1969; Hintze 2006) and the nonparametric Kaplan–Meier estimator (Kaplan and Meier 1958) (Eq. 6), while the hazard rate (also known as the conditional failure rate in reliability) is a metric which may be used for identifying the appropriate probability distribution of a particular mechanism (Hintze 2006).

$$\hat{S}(t) = \begin{cases} 1 & \text{if } T_{\min} > T \\ \prod\limits_{A \le T_i \le T}\left[1 - \dfrac{d_i}{r_i}\right] & \text{if } T_{\min} \le T \end{cases} \tag{6}$$

The parameter d_i in Eq. 6 denotes the number of failures occurring in time T_i (D_i is the set of these failures); r_i denotes the number of the individuals that are at risk immediately before time T_i (R_i is the set of all individuals that are at risk immediately before time T_i); T_i is the termination time (failure event) for the ith event; and A is the minimum time considered below which failures are not considered (that is the case when data is left-truncated). A plot of the Kaplan–Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations is assumed to be constant. As already mentioned, an important advantage of the Kaplan–Meier estimator is that the method can take into account censored (i.e. missing) data, thus being able to approximate the underlying probability distribution function even in the presence of incomplete datasets (Klein and Moeschberger 1997). When no censoring occurs the Kaplan–Meier curve is equivalent to the empirical distribution.

It is often useful, though, to go beyond estimating the underlying probability distributions and simply be able to compare the hazard rates of two groups of similar attributes within the examined dataset. This is most often accomplished by employing the hazard ratio (*HR*), a metric that assesses the effect of a variable on the risk of an event (thus being an estimate of relative risk). If, for example, the hazard function of two variables is given by a regression model of the form

$$h(t, x, \beta) = h_0(t)\, r(x, \beta) \tag{7}$$

with $h_0(t)$ characterizing how the hazard function changes as a function of survival time, and $r(x,\beta)$ characterizing how the hazard changes as a function of the covariates of the two data sets, then the hazard ratio of the two data sets with covariate values of x_0 and x_1 is given by

$$HR(t, x_1, x_0) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} = \frac{h_0(t)\, r(x_1, \beta)}{h_0(t)\, r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)} \tag{8}$$

As Eq. 8 shows, the hazard ratio depends only on the function $r(x,\beta)$ and not the baseline hazard function, $h_0(t)$. This property allows the comparative examination of the risk factors and their effects on the survivability of the subject in examination (in this case the pipes in the network). Proportional hazards regression models include the Cox semi-parametric proportional hazards model, and the exponential, Gompertz and Weibull parametric models. In Cox's proportional hazard model (Cox 1972), $r(x,\beta) = \exp(x\beta)$, the hazard function is

$$h(t, x, \beta) = h_0(t)\, e^{x\beta} \tag{9}$$

and the hazard ratio is

$$HR(t, x_1, x_0) = e^{\beta(x_1 - x_0)} \tag{10}$$

Further discussion on the hazard ratio is given in the succeeding section, with a dataset on pipe breakage from the city of Limassol (Cyprus) as a backdrop.

4 Case Study

The water distribution network used as a case study is the network of the city of Limassol, Cyprus. The network is over 50 years of age, serves approximately 170,000 residents through approximately 64,000 consumer meters in an area of 70 km^2, consists of about 800 km of pipes and carries an annual volume of potable water of about 14 million cubic meter (about €7.0 million in value).

The dataset used in the study spans a 5-year period (2002–2007) and about 2,000 break incidents and its analysis considered only a subset of the risk factors examined in a previous study of a New York City dataset (Christodoulou et al. 2003), based on the applicability of these risk factors to the given locale and available dataset. The risk factors examined were the number of observed previous breaks (*NOPB*), the material (*MAT*) and diameter (*D*) of each pipe, and the traffic load in the vicinity of the pipe (*TRAF*). At first, a four-layer back-propagation artificial neural network (ANN) was employed (Christodoulou et al. 2007) for identifying the underlying data patterns and the interactions of these risk factors (treated as inputs to the ANN). The ANN was also used in identifying the possible contribution of each risk factor to a pipe's failure and to the pipe's estimated life-cycle (the *BreakOrNot* and *TimeToFailure* ANN outputs respectively). The ANN analysis indicated that the order of significance (in descending order) of the aforementioned risk factors was *NOPB*, *D*, *MAT* and *TRAF*, with relative importance values of 34.387, 29.430, 23.934 and 12.349 respectively. In addition to the aforementioned risk factors, data was also collected on the type of pipe in which water loss was observed (*PipeType*: water main or house connection), and the presumed reason for the pipe's failure (*IncidentType*: pipe deterioration, corrosion, interference by others, tree roots, connection hose, other).

Further to the ANN analysis, the incident data was analyzed by means of both Poisson regression (Table 1, Fig. 1) and Cox regression (Table 2) by which the *PresumedPipeAge* output was fitted to models of the basic variables *NOPB*, *MAT*, *PipeType*, *IncidentType* and *D*. For the purpose of data stratification, the factors *D* and *NOPB* were grouped in the subcategories *DiameterRange* and *NOPBRange* respectively. *DiameterRange* takes the values 'Small' for diameters of 0.00–25.4 mm, 'Medium' for 25.4–127.0 mm and 'Large' for values greater than 127.0 mm. *NOPBRange* takes the values 'Small' for 0–4 incidents, 'Medium' for 4–8 and 'Large' for more than eight incidents.

The basis for the Poisson model is the incidence rate (i.e. the expected number of events per time unit). Poisson regression further assumes that the response variable, Y, has a Poisson distribution, and that the logarithm of its expected value can be modelled by a linear combination of unknown parameters, p (Eq. 11). The

Table 1 Poisson regression analysis of risk factors *NOPB*, *MAT*, *PipeType*, *IncidentType* and *DiameterRange*

Variable	Coeff.	Error
Intercept	9.37716	0.00429
(MAT='DI')	0.05252	0.00744
(MAT='GI')	0.02445	0.00391
(MAT='MDPE (BLACK)')	0.02654	0.00389
(MAT='MDPE (BLUE)')	0.02876	0.00405
(DiameterRange='MEDIUM')	0.00723	0.00187
(DiameterRange='SMALL')	0.01223	0.00302
(IncidentType='CORROSION')	−0.01275	0.00175
(IncidentType='INTERF. BY OTHERS')	−0.03334	0.00186
(IncidentType='OTHER')	−0.00878	0.00235
(IncidentType='PIPE DETERIORATION')	−0.01628	0.00150
(IncidentType='TREE ROOTS')	−0.00738	0.00183
NOPB	0.00115	0.00003
(PipeType='WATER MAIN')	0.02870	0.00411
Dispersion Phi	21.1626	

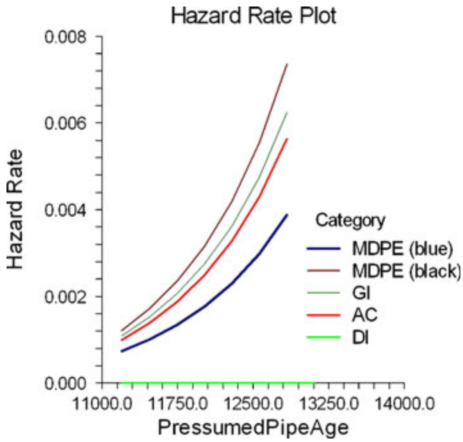


Fig. 1 Hazard rate vs. presumed pipe age (in days) for data stratification based on material type

regression coefficients β_0,β_1,...,β_p represent the effects of the covariates on the response variable. Similarly, the Cox model assumes that the response variable can be modelled by an exponential function (Eq. 12), with $\lambda_0(t)$ being the baseline hazard for a subject with covariates 0, β_i being the regression coefficients and X_i being the risk parameters.

$$\log(Y) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \tag{11}$$

$$\lambda_i(t) = \lambda_0(t) \exp\left[\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right] \tag{12}$$

Table 2 Cox regression analysis of risk factors *NOPB*, *MAT*, *PipeType*, *IncidentType* and *diameter range*

Variable	Regression coefficient (B_i)[a]	Standard error of B_i	Risk ratio[b]
B1 (MAT='DI')	0.779548	1.320206	2.1805
B2 (MAT='GI')	0.133945	0.676191	1.1433
B3 (MAT='MDPE (BLACK)')	0.214512	0.666312	1.2393
B4 (MAT='MDPE (BLUE)')	0.204553	0.688775	1.2270
B5 (DiameterRange='MEDIUM')	−0.040946	0.427816	0.9599
B6 (DiameterRange='SMALL')	−0.528034	0.859742	0.5898
B7 (IncidentType='CORROSION')	0.161740	0.412117	1.1756
B8 (IncidentType='INTERF. BY OTHERS')	0.242169	0.423415	1.2740
B9 (IncidentType='OTHER')	−0.241108	0.527373	0.7858
B10 (IncidentType='PIPE DETERIORATION')	0.112606	0.359604	1.1192
B11 (IncidentType='TREE ROOTS')	−0.057751	0.428092	0.9439
B12 NOPB	0.015908	0.009615	1.0160
B13 (NOPBRange='MEDIUM')	0.186366	0.221957	1.2049
B14 (NOPBRange='SMALL')	−0.198185	0.218228	0.8202
B15 (PipeType='WATER MAIN')	−0.391782	0.722911	0.6759

[a]This is the estimate of the regression coefficient, β_i. It is the amount that the log of the hazard rate changes when x_i is increased by one unit

[b]This is the value of $\exp(\beta_i)$. It is the ratio of two hazards whose only difference is that x_i is increased by one unit

As the residual errors from the first analysis indicate (Table 1) the Poisson regression model is a good fit. The Cox regression, though, with the added information on the factors' proportional hazards provides a more powerful analysis tool. Of particular interest is the 'risk ratio' metric (Table 2) showing the value of $\exp(\beta_i)$, which indicates the ratio of two hazards whose only difference is that x_i is increased by one unit.

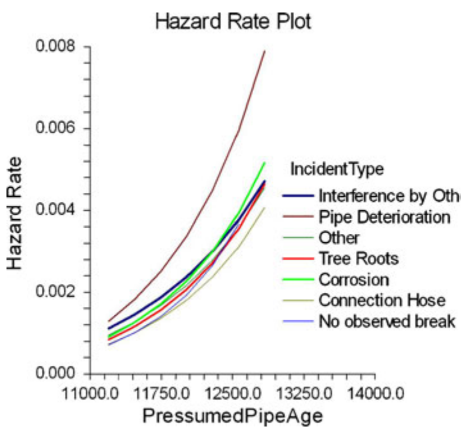
For example, in the dataset used and based on the analysis results (Table 2) if the number of observed previous breaks within the subgroup '*NOPBRange=medium*' is increased by one, then the hazard rate increases by 20.49%. The *NOPB* risk ratio shows a 1.6% increase in the hazard rate for each additional incident, but this covers the entire data range from 0 to 8+ incidents which is overweighed by the number of incidents in the range 0–3 and it is, thus, not representative of the whole data population.

Hazard rate plots are then developed for each different data group: material type (Fig. 1), incident type (Fig. 2), diameter range (Fig. 3).

The following can further be deduced from the plotted hazard rate graphs:

- The hazard rate for the black-colour medium density polyethylene (MDPE) pipes is higher than the rate for galvanized (GI), asbestos cement (AC) and MDPE-blue pipes. The data sample on ductile iron (DI) pipes is too small to produce statistically significant results and should therefore be ignored.
- The hazard rate related to pipe deterioration greatly outpaces the hazard rate of the other incident types (corrosion, interference by others, tree roots, connection hose, other). What starts as a relatively small deviation at the start of the observed failure time (at about 30 years) soon doubles (0.008 vs 0.004) at the upper end of the expected lifetime (at $t \sim 35$ years). And while the hazard rate for incidents attributed to tree roots, corrosion, connection hose or other reasons increases almost uniformly over time, the deterioration-related hazard rate accelerates over time (from 0.001 to 0.008). This is an indication that aging pipes should be replaced at smaller time-intervals (at about 11,000 days $\sim$30 years), otherwise the risk of failure increases substantially. Also of interest is the observation that the hazard rate related to tree roots accelerates in time and surpasses the rate of increase of corrosion-related, interference-related and

Fig. 2 Hazard rate vs. presumed pipe age (in days) for data stratification based on incident type



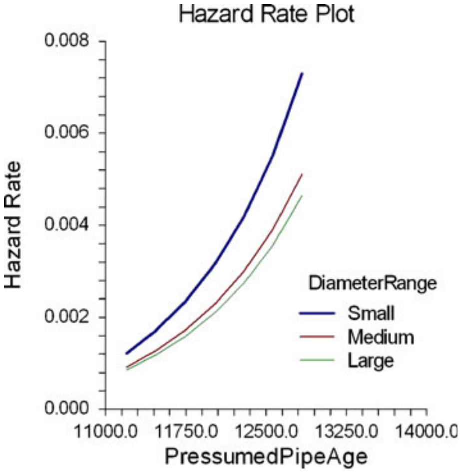


Fig. 3 Hazard rate vs. presumed pipe age (in days) for data stratification based on diameter range

connection-hose incidents. This is an indication that pipes in the vicinity of trees should be monitored more closely and/or replaced in smaller time intervals than other pipes.

- Medium and large diameter pipes have approximately the same hazard rate over time but small-diameter pipes have an increasing hazard rate. This is an indication that small-diameter pipes should be replaced at smaller time intervals than medium and large-diameter pipes. The observation may be attributed to the fact that the pipe network in study consists primarily of small-diameter pipes (neighbourhood and house connection pipes) that operate under relatively high pressures and therefore are subject to increased risk of failure.

5 Conclusions

The study identifies several risk factors (such as pipe deterioration, material type, pipe diameter, number of observed previous breaks) and their significance to failure, based on a 5-year dataset (about 2,000 water-loss incidents). A deeper insight to these factors' contribution to pipe failure is gained through survival analysis and the data stratifications employed in the analysis. The described methodology furnishes the managers of urban water distribution networks with a powerful decision support tool for devising 'repair or replace' strategies. Furthermore, the nonparametric nature of the Kaplan-Meier estimator does not necessitate a-priori knowledge of multi-factored probability distribution functions for the estimation of the pipes' probability of failure, thus simplifying the analysis. A limitation of the method is its reliance on fairly large datasets in terms of time-span, or on reasonably accurate estimation of the lifetimes of failed pipes.

Acknowledgements The work presented herein is based partially upon work supported by the Cyprus Research Promotion Foundation (under Grants No. AKGEN/1204/01 and PLYPH/0506/21).

References

Andreou SA, Marks DH, Clark RM (1987) New methodology for modelling break failure patterns in deteriorating water distribution systems: applications. Adv Water Resour 10:2–10.

Aslani P (2003) Hazard rate modeling and risk analysis of water mains. MSc. thesis, Polytechnic University, Brooklyn

Christodoulou S, Deligianni A (2009) A neurofuzzy decision framework for the management of water distribution networks. Water Resour Manag 24(1):139–156

Christodoulou S, Aslani P, Vanrenterghem A (2003) A risk analysis framework for evaluating structural degradation of water mains in urban settings, using neurofuzzy systems and statistical modeling techniques. In: Proc. world water & environmental resources congress 2003 and related symposia, Pennsylvania, USA

Christodoulou S, Charalambous B, Adamou A (2007) Managing the 'repair or replace' dilemma on water leakages. In: Proc. international water association's (IWA) third specialty conference on water loss reduction (Water loss 2007), Bucharest, Romania

Clark RM, Stafford CL, Goodrich JA (1982) Water distribution systems: a spatial and cost evaluation. J Water Resour Plan Manage Div 108:243–256

Cox DR (1972) Regression models and life-tables. J R Stat Soc B 34(2):187–220

Epanechnikov VA (1969) Nonparametric estimation of a multidimensional probability density. Theor Probab Applt 14:153–158

Fattahi P, Fayyaz S (2009) A compromise programming model to integrated urban water management. J Water Resour Plan Manage 24(6):1211–1227

Goulter IC, Kazemi A (1988) Spatial and temporal groupings of water main pipe breakage in Winnipeg. Can J Civ Eng 15:91–97

Hintze J (2006) NCSS, PASS and GESS. NCSS, Kaysville

Hosmer DW, Lemeshow S, May S (2008) Applied survival analysis: regression modeling of time to event data. Wiley, Hoboken

Kanakoudis VK, Tolikas DK (2004) Assessing the performance level of a water system. Water Air Soil Pollut: Focus 4(4):307–318

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Klein JP, Moeschberger ML (1997) Survival analysis techniques for censored and truncated data. Springer, New York

Kleiner Y, Rajani B (1999) Using limited data to assess future needs. J Am Water Works Assoc 91:47–61

Lee ET, Wang JW (2003) Statistical methods for survival data analysis. Wiley, Hoboken

Park S (2008) Identifying the hazard characteristics of pipes in water distribution systems by using the proportional hazards model: theory. KSCE J Civ Eng 8(6):663–668

Pinto J, Varum H, Bentes I, Agarwal J (2010) A Theory of Vulnerability of Water Pipe Network (TVWPN). Water Resour Manag. doi:10.1007/s11269-010-9655-3 (in print)

Prasad TD, Hong S-H, Park N (2003) Reliability based design of water distribution networks using multi-objective genetic algorithms. KSCE J Civ Eng 7(3):351–361

Shamir U, Howard C (1979) An analytical approach to scheduling pipe replacement. J Am Water Works Assoc 71:248–258

Tuhovcak L, Rucka J, Juhanak T (2006) Security of water supply systems: from source to tap. NATO science for peace and security series C: environmental security, vol 8. Springer, Netherlands, pp 169–182

Vanrenterghem-Raven A, Eisenbeis P, Juran I, Christodoulou S (2004) Statistical modeling of the structural degradation of an urban water distribution system: case study of New York City. In: Proc. world water & environmental resources congress 2003 and related symposia, Pennsylvania, USA