Reliability Engineering and System Safety 144 (2015) 311–318

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Modelling the failure risk for water supply networks with interval-censored data

B. García-Mora [a], A. Debón [b], C. Santamaría [a], A. Carrión [b]

[a] Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Spain
[b] Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain

ARTICLE INFO

Article history:
Received 24 October 2014
Received in revised form
29 July 2015
Accepted 5 August 2015
Available online 12 August 2015

Keywords:
Interval-censored data
Reliability analysis
Generalized non-linear model

ABSTRACT

In reliability, sometimes some failures are not observed at the exact moment of the occurrence. In that case it can be more convenient to approximate them by a time interval. In this study, we have used a generalized non-linear model developed for interval-censored data to treat the life time of a pipe from its time of installation until its failure. The aim of this analysis was to identify those network characteristics that may affect the risk of failure and we make an exhaustive validation of this analysis. The results indicated that certain characteristics of the network negatively affected the risk of failure of the pipe: an increase in the length and pressure of the pipes, a small diameter, some materials used in the manufacture of pipes and the traffic on the street where the pipes are located. Once the model has been correctly fitted to our data, we also provided simple tables that will allow companies to easily calculate the pipe's probability of failure in a future.

1. Introduction

Worldwide, water supply systems (WSS) face the problem of aging infrastructures and increasing maintenance costs. The profits of drinking water supply companies and service quality for citizens depend on the reliability of the pipes. The classical reactive approach (used by most companies) is to wait until there is a failure in the network and then repair it, which is obviously not the best way to handle this essential public service from the point of view of either quality or reliability while by contrast other proactive strategies based more on prevention are required. These require information, quantitative tools, and advanced reliability modelling to evaluate and predict risks of failures to assess current and future state of the network. The need for these proactive strategies is even greater in developing countries with stronger economic restrictions than advanced countries. Thus, companies with these proactive strategies would have a clearer framework to make decisions on the diagnosis and rehabilitation of the pipes for effective prevention of failures in the network. We analyze failure data registered in a water supply network in order to evaluate the probability of pipe failure. This study also assesses and identifies those factors that may affect the risk of failure in order to better plan breakdown service.

In reliability analysis, data refer to time from a well-defined *time origin* until the occurrence of some particular event or *end-point*. In this analysis, the variable of interest is the time T (in years) from the installation of the pipe (*time origin*) until its first failure (*end-point*). T is the time of the pipe until the first failure being as in the database is only recorded the only failure, no more. Possible subsequent repairs are not registered. Therefore, we consider the only registered failure as the end-point of the lifetime. In a standard analysis of the time until the occurrence of an event, failure times are known and observed exactly or right-censored. In this type of data the proportional hazard model [1] has been widely used. However, in some situations these failure times can occur in a given time interval as, for example, in survival analysis, where the event of interest, the relapse of a patient, occurs between two visits to the surgery (time interval). The data in this form are referred to as *grouped* or arbitrarily *interval-censored data*. In our case, we analyze failure data registered in a water supply network in order to evaluate the probability of pipe failure. For this, we have considered an *observation window* from the year 2000 until 2005 for the pipe failure times, a brief recorded pipe break history. This sampling scheme induces left-truncation into the data set (since failures before 2000 are not considered in the sample information) and right-censoring (for pipes that fail after 2005). Left-truncation is a common problem for water pipes' data sets. Some studies have dealt with this problem: [2] compare the risk associated with different statistical survival models applied to these same data sets of the present paper, under the assumption that left-truncation is a minimal problem and, more recently, it is used as an extended

E-mail addresses: magarmo5@imm.upv.es (B. García-Mora), andeau@eio.upv.es (A. Debón), crisanna@imm.upv.es (C. Santamaría), acarrion@eio.upv.es (A. Carrión).

http://dx.doi.org/10.1016/j.ress.2015.08.003
0951-8320/

version of the Nelson—Aalen estimator, modified to accommodate left-truncation as well as right-censoring [3]. Mailhot et al. [4] present a methodology to estimate calibration parameters of statistical models in municipalities with short recorded pipe break histories. Kleiner and Rajani [5] also deal with recorded pipe break histories. Their model is based on an exponential relationship. Finally [6] propose a formal statistical approach to extend the likelihood function of a pipe failure model by a replacement model because common data management practices mean that replaced pipes are often absent from available data sets leading to a survival selection bias, as pipes with frequent failures are more likely to be absent from the data.

In this study we used a method for modelling *interval-censored data* developed by Farrington [7] in the parametric model framework. This model assumes proportional hazards and it is based on a non-linear model for binary data. We use this method of easy implementation with a standard statistical package and interpretation analogous to the Cox model. Moreover, the same author develops a comprehensive account of diagnostic methods to use with proportional hazard models for interval-censored data which provides a validation of his own model [8]. Previously, we used this methodology with another database in the survival analysis framework [9], rounding off the analysis [10] carried out with the Cox model.

This study is organized as follows: firstly, we describe Farrington's model and the implementation to an *interval-censored* database. Secondly, we present our water supply network database and how the variable T is calculated for each pipe by an interval of time. Thirdly, we apply the model by identifying the main characteristic factors which can affect the failure risk, comparing the results with two other well-established models in reliability analysis: the Cox and the Generalized Linear Models. Next, we validate the Farrington model by means of standard diagnostic tools developed by that author. Finally, under the assumption that the model is correctly fitted to our set of interval-censored data, we provide simple risk tables that will allow companies to easily calculate the pipe's probability of no failure at ten, thirty and fifty years.

2. Farrington's model for interval-censored data

First of all, we consider one basic function in reliability analysis, which is the *reliability function*, $R(t)$. This function is the probability that the time to failure, T, is larger than or equal to t

$$R(t) = P(T \geq t) = 1 - F(t) \tag{1}$$

with $F(t)$ being the distribution function of the variable T.

Farrington's model supposes that the failure time for the ith pipe is observed in an interval $(a_i, b_i]$, that is, the failure has not occurred by time a_i but has occurred by time b_i, where the values a_i and b_i are different for each pipe. There are three types of interval-censored observations:

- If the event time for a pipe is *left-censored* at time b_i, so that the event is only known to have occurred some time before b_i, then $a_i = 0$.
- If the event time is *right-censored* at time a_i, so that the event is only known to have occurred after time a_i, the upper limit of the interval, b_i, is then effectively infinite.
- If the values of both a_i and b_i are observed for a pipe, the interval-censored observation is said to be *confined*.

The probability of failure occurring in the interval $(a_i, b_i]$ for the ith pipe is $R_i(a_i) - R_i(b_i)$. Thus, the likelihood function for the n pipes is

$$\prod_{i=1}^{n} (R_i(a_i) - R_i(b_i)) \tag{2}$$

Now let us suppose from these n pipes that l observations are *left-censored*, r observations are *right-censored* and c observations are *confined* with $n = l + r + c$, in such a way that the first l observations are left-censored, the next r ones are right-censored and the remaining c observations are *confined* $(0 < a_i < b_i < \infty)$.

Since $R_i(0) = 1$ and $R_i(\infty) = 0$, the contributions of a *left*- and *right-censored* observation to the likelihood function will be $1 - R_i(b_i)$ and $R_i(a_i)$, respectively. Consequently, the overall likelihood function (2) can be rewritten as follows:

$$\prod_{i=1}^{l} (1 - R_i(b_i)) \prod_{i=l+1}^{l+r} R_i(a_i) \prod_{i=l+r+1}^{n} (R_i(a_i) - R_i(b_i)) \tag{3}$$

where the last component $R_i(a_i) - R_i(b_i)$ of confined observations can be rewritten by $R_i(a_i)(1 - R_i(b_i)/R_i(a_i))$.

Farrington shows that this likelihood function (3) is equivalent to the following expression:

$$\prod_{i=1}^{n+c} p_i^{y_i} (1 - p_i)^{1 - y_i} \tag{4}$$

where $y_1, y_2, \ldots, y_{n+c}$ are a set of $n + c$ independent binary observations from a Bernoulli distribution with a response probability p_i, $i = 1, \ldots, n + c$. To see the relationship between the probabilities p_i in (4) and the values of the reliability function in (3), the following is considered:

- Each *left-censored* observation contributes a binary observation, with $y_i = 1$ and $p_i = 1 - R_i(b_i)$, $i = 1, \ldots, l$.
- Each *right-censored* observation contributes a binary observation with $y_i = 0$ and $p_i = 1 - R_i(a_i)$, $i = 1, \ldots, r$.
- Each *confined* observation contributes two binary observations to give the required component of $R_i(a_i)(1 - R_i(b_i)/R_i(a_i))$ in (3): one of these is $y_i = 0$ and $p_i = 1 - R_i(a_i)$, while the other is such that $y_{c+i} = 1$ and $p_{c+i} = 1 - R_i(b_i)/R_i(a_i)$, for $i = l + r + 1$, $l + r + 2, \ldots, n$.

Combining all these terms then leads to three new components of the likelihood in (4). Therefore, the final expression for this likelihood function is

$$\prod_{i=1}^{l} p_i \prod_{i=l+1}^{l+r} (1 - p_i) \prod_{i=l+r+1}^{n} (1 - p_i) p_{c+i} \tag{5}$$

Table 1 shows the intervals defined by Farrington for his model and values of the variable y_i for each type of interval-censored observation.

In a second step the expression for the reliability function $R_i(t)$ is constructed. Farrington assumes the proportional hazard model for the *reliability function* and so

$$R_i(t) = R_0(t) \exp(\beta' x_i) \tag{6}$$

with x_i being the vector of values of the p explanatory variables for the ith pipe, $i = 1, 2, \ldots, n$. The baseline reliability function $R_0(t)$ is modelled as a step function, where the steps occur at the k ordered censoring times, $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$, with $0 < t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ and the times $t_{(j)}$, $j = 1, 2, \ldots, k$, are a subset of the values of a_i and b_i, $i = 1, 2, \ldots, n$. The procedure for choosing the times $t_{(j)}$ is explained

Table 1
Definition of the A_i intervals in Farrington's model for *left–censored*, *right–censored* and *confined* observations.

Type of observation	Value of y_i	Interval A_i
Left-censored	1	$(0, b_i]$, $i = 1, 2, \ldots, l$
Right-censored	0	$(0, a_i]$, $i = l + 1,\ l + 2, \ldots, l + r$
Confined	0	$(0, a_i]$, $i = l + r + 1,\ l + r + 2, \ldots, n$
	1	$(a_{i-c}, b_{i-c}]$, $i = n + 1, n + 2, \ldots, n + c$