García Mora, MB.; Debón Aucejo, AM.; Santamaria Navarro, C.; Carrión García, A. (2015). Modelling the failure risk for water supply networks with interval-censored data. Reliability Engineering & System Safety. 144:311-318. doi:10.1016/j.ress.2015.08.003

Modeling and discrimination in failure risk models with interval–censored data. An application in water supply networks.

B. García-Mora* A. Debón † C. Santamaría * A. Carrión†

Problem: Since drinking water supply companies profits depend on pipe reliability it is important to be able to measure the risk of pipe failure with time accurately as its improvement could have important social and economic implications. Therefore, although awareness of the importance of predicting failure rates in reliability has existed in the literature for many years, the full power of advanced statistical modeling has only been used for engineering questions in recent times.

Approach: Using data from a real drinking water supply company in a medium–sized Spanish city, the network characteristics which affect the risk of failure and the models which best fit the data to predict service breaks were identified. As, in our data we do not know the exact count of times to failure of

*Instituto de Matemática Multidisciplinar. Universidad Politécnica de Valencia, Spain
†Dpto. de Estadística e Investigación Operativa. Universidad Politécnica de Valencia, Spain

each pipe we approximated the time until the failure of each pipe by means of an interval and we apply a model developed by Farrington for interval–censored data. This method is based on a non–linear model for binary data and uses standard statistical packages with interpretation analogous to Cox's. In order to check the consistency of Farrington's model we do an exhaustive validation of this method and we compare it with some well established models: Cox's model and the Generalized Linear model.

Results: This study shows that network characteristics affect the risk of pipe failure: an increase in the length and pressure, a small diameter, the material used and to make the pipes and a heavy traffic conditions of the street. So we propose a clear framework for decision support in the diagnosis and rehabilitation of water supply systems in that company. In order to compare the models, we have used the ROC curves and the Concordance Index to decide which models provide better discrimination in order to predict service breakdown: the Farrington model had the best discrimination and the Cox model the worst.

Key words: Interval–Censored Data, Reliability Analysis, Farrington Model, Concordance Index, ROC Curves.

Process Description

Worldwide, water supply systems (WSS) face the problem of aging infrastructures and increasing maintenance costs. Drinking water supply companies profits

and service quality for citizens depend on pipe reliability. The classic reactive approach used by most companies is obviously not the best way of managing this essential public service from the point of view of both quality and availability. Therefore, proactive strategies are required, however, these proactive approaches require models to evaluate risks, to predict the best measures and to forecast water supply network performance. In this way, it would be very important to be able to calculate failure probabilities of pipes over time as forecasting pipe failures has important economic and social implications. Moreover, the companies will have a clearer framework to make decisions in the diagnosis and rehabilitation of the pipes. Quantitative tools (statistical indicators, reliable databases, etc...) are required in the management of water supply systems in order to assess the current and future state of networks and so forecast the future deterioration of infrastructures. Nowadays, companies managing these networks try to establish models for evaluating the risk of failure in order to develop a proactive approach to the renewal process instead of using traditional reactive pipe substitution schemes. So the power of advanced statistical modeling in the field of reliability is needed although it has only been used for engineering questions in recent times.

The main objective of this paper is to compare and improve models of reliability data for evaluating the risk of pipe failure. We want to identify which main network characteristic factors affect the risk of failure and which models better fit data to evaluate the failure probability and predict service breakdown. Data

from the water supply network of a medium-sized city on the Spanish Mediterranean coast is used. In addition, we will outline the problems related to data collection and quality and the measures taken to "clean" the database of errors and inconsistencies.

In reliability analysis, data are related to time from a well-defined *time origin* until the occurrence of some particular event or *end-point*. In our study, the variable of interest is the time (in years) from the installation year of the pipe (*time origin*) to the deterioration of the pipe or failure time (*end-point*). This variable T is the time until the failure. Now, in standard time-to-event or survival analysis, occurrence times of the event of interest are observed exactly or are right-censored, however, in some situations the times of the events of interest may only be known to have occurred within an interval of time. For example, in clinical trials patients are often seen at pre-scheduled visits but the event of interest may occur between visits. As in our data, we do not know the exact number of all failure times of each pipe then we need to do an approximation of the failure time of each one by means of an interval. So each pipe may have a different time interval in which the failure has occurred, and so data are referred to as *grouped* or arbitrarily *interval-censored data*. We describe the database in detail below and the interval–censored data for our analysis.

On the other hand since Peto (1973) and Turnbull (1976) developed an estimator of the survival function in survival analysis for interval–censored data the literature on interval-censored data has grown. In order to evaluate the pos-

sible effect of several factors on the time T until the failure, parametric and semi–parametric models have mainly been used. In fact, many research analyses deal with interval–censored response data extending Cox's proportional hazard model (Cox, 1972). Therefore, Finkelstein (1986) proposed a method for fitting the proportional hazard model to interval–censored data. Lindsey and Ryan (1998) reviewed the use of parametric models for the analysis of interval–censored data. Huang and Wellner (1997) provided a rigorous theoretical account of these methods. This topic is dealt with in several general survival analysis books, e.g. Kalbfleisch and Prentice (2002), Lawless (2003) and more recently in Sun (2006).

From a practical point of view, an important problem for most of the available methods, as we have shown above, is the lack of standard packages of statistical software. In the parametric models framework, there is a method for modelling arbitrarily interval–censored data developed by Farrington (1996). It assumes proportional hazards and it is based on a non–linear model for binary data (see Collett et al., 2003, cap. 9). We use the Farrington model of easy implementation using standard statistical packages with interpretation analogous to Cox's. Moreover, the same author develops a comprehensive account of diagnostic methods to use with proportional hazard models for interval–censored data (Farrington, 2000) which provides a validation of the obtained model. We have used this methodology recently with another database in medicine (Santamaría et al., 2008) for survival analysis and we have improved the analysis carried out with Cox's model in García et al. (2005) and in Santamaría (2006).

On the other hand, as a model that places each pipe in the class to which it really belongs it could be said to have perfect discrimination. We also investigate the discrimination ability for all models in this study by means of the concordance C index introduced by Harrell et al. (1996) as a natural extension of the ROC curve area in survival analysis.

In the following sections, we present the water supply network database and we calculate the approximate failure times for each pipe by means of intervals. Next, we describe the Farrington model for interval–censored data and we apply it to analyze the main characteristic factors affecting the risk of failure. We also contrast and compare the Farrington model with other models in survival and reliability analysis, and the Cox model and the Generalized Linear model to study the consistency of the first one. Next, we validate the Farrington model by means of standard diagnostic tools developed by the same author for interval–censored data. Based on these findings, we analyze the ROC curve and the concordance C index for the three models. Finally, the most relevant conclusions are presented.

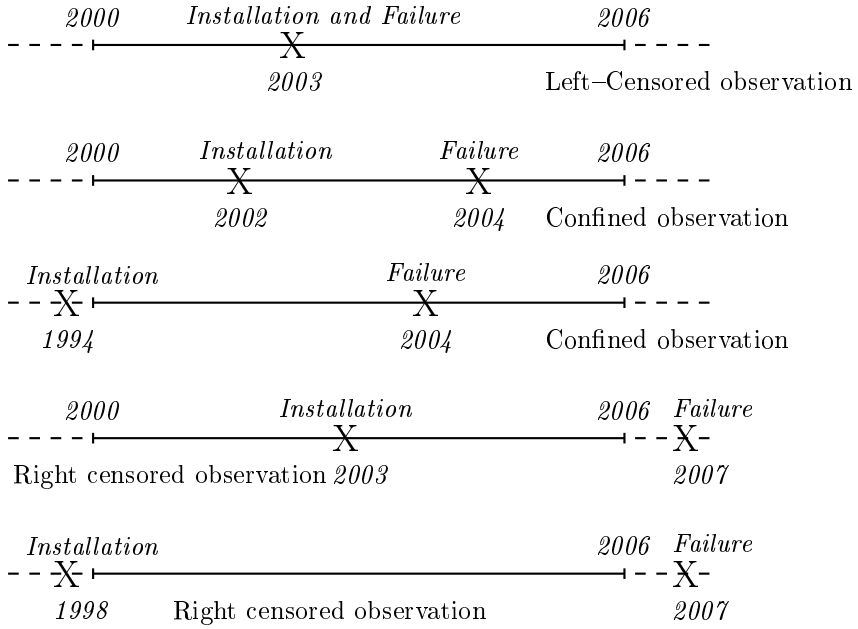
Data Collection

Interval–censored data

In the development of the research project we had access to data from the water supply company of a medium-sized Spanish city. The company gave us access to a database containing information on pipe sections making up the network.

The database includes 32387 entries corresponding to the sections. Among other variables, these entries contain section identification of the pipe, installation year, date of failure, general characteristics of the pipe (length, diameter and pressure), traffic conditions in the street (under sidewalk, normal traffic and heavy traffic) and four different materials that were used to make the pipes (ductile cast iron, gray cast iron, polyethylene and asbestos cement). *Traffic conditions* is classified by three levels and our reference is heavy traffic. *Material* is classified in four categories and the reference is asbestos cement.

There were some problems with the quality of the data. One of the major problems with the database was the information about the oldest pipe sections. Due to the lack of reliability of older data only those pipes installed after 1940 were considered. Failures have only been included in the database since 2000 (when the use of the GIS was established) and there was no possibility of recovering failure data from before that year. This means a very high censoring rate, up to 98%. Also, no consideration was given to the fact that a pipe section can fail more than once, because the database structure was not prepared to consider this fact. So these minor errors had to be corrected prior to using the database, frequently resulting in the loss of the corresponding failure entries.



Diagram

Since in the database the exact time of failure of each pipe is unknown we made an approximation of the time until the failure of each pipe by means of an interval. So each pipe may have a different time interval in which the failure occured and the database is referred to as grouped or arbitrarily interval–censored data. Therefore, for all pipes the failure times are only registered between 2000 and 2006, the installation year of the pipe can be before or within the interval [2000, 2006] and the date of failure can be before, within or after this interval. We defined the intervals A_i (Table 1) according to all possible cases in the database (see Diagram). For example, if the installation and the failure years for a pipe are the same we consider that the age of the pipe is between 0 and 1 year and

we approximate the age by means of the interval (0, 1], this being a *left–censored* observation. If the year of installation and the failure time are different but both are within the interval [2000, 2006] the observation will be *confined*. In this case we approximate the age of the pipe by means of an interval where the inferior limit will be the difference between the installation and failure years and the superior limit will be always one unit more. So we want to consider all months of the period from the beginning of the year of installation to the end of the year of failure. However if the installation year is before 2000 (the other case of confined observation) the age of the pipe will be approximated by an interval where the inferior limit goes from year 2000 to failure year and the superior limit goes from the year of installation to year of failure. We wanted to approximate the age of the pipe by means of that interval, as in this case we could exactly determine the inferior limit but not the superior limit; so we used that approximation as we did not known possible pipe failures before the year 2000. Finally, if the failure time is after the year 2006 the observation will be *right–censored* and we distinguish two possible cases: if the year installation is inside of the interval [2000, 2006] we approximate the age for that pipe by means of an interval from the year installation to 2006, and if the year installation is before the year 2000 we approximate the age in a similar way as the second case in confined observations.

Type of observation	year installation	year failure	Age	Interval A_i
Left-censored	2003	2003	0	$(0, 1]$
Confined	2002	2004	2	$(2, 3]$
Confined	1994	2004	10	$(4, 10]$
Right-censored	2003	2007	3	$(0, 3]$
Right-censored	1998	2007	9	$(0, 9]$

Table 1: Examples of intervals A_i.

Analysis and Interpretation

The Farrington model for Interval-Censored Data.

In reliability analysis there are two basics functions: the first one is the *reliability function* $R(t)$, the probability that the time until failure T is greater than or equal to t,

$$R(t) = P(T \geq t) = 1 - F(t),$$

with F the distribution function of T.

The second one is the *hazard function*, or hazard rate, $h(t)$. It is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

that is the probability of failure at time t, conditional on the pipe working until time t. It is assumed the proportional hazards assumption, tested in the

model checking process, that is, hazard ratio between different pipes is constant and independent of time.

Let us suppose the time until failure for the ith pipe is observed in the interval $(a_i, b_i]$. We have mentioned before that data in this form are referred to as interval-censored data. If this time is *left-censored* (the event occurred before the first observation of the pipe) at time b_i, then $a_i = 0$. If it is *right-censored* (the event occurred after the last observation) at time a_i, then $b_i = \infty$. We say the interval–censored recurrence time is *confined* when values a_i and b_i are observed for an pipe during the follow–up period. Let us suppose n pipes, where there are l left-censored, r right-censored, and c confined, so that $n = l + r + c$. Then the overall likelihood function for the n pipes can be written as

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

where $R_i(t)$ is the reliability function for the ith pipe.

Farrington shows that this likelihood is equivalent to

$$\prod_{i=1}^{n+c} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{2}$$

where $y_1, y_2, \ldots, y_{n+c}$ are $n + c$ independent binary observations from a Bernoulli distribution with response probability p_i, $i = 1, 2, \ldots, n + c$.

The left–censored observations contribute to the likelihood function as l binary observations with $y_i = 1$ and $p_i = 1 - R_i(b_i)$. The right–censored observations

contribute to the likelihood function as r binary observations with $y_i = 0$ and $p_i = 1 - R_i(a_i)$. Finally, each confined observation contribute to the likelihood function as two binary observations. The first one is defined as a binary observation with $y_i = 0$ and $p_i = 1 - R_i(a_i)$, while the second one is defined with $y_{c+i} = 1$ and $p_{c+i} = 1 - \dfrac{R_i(b_i)}{R_i(a_i)}$ (see Collett et al., 2003, p. 287 for more details).

In order to specify the model, the next step is to obtain the expression of the reliability function $R_i(t)$. For this purpose it is assumed that the hazards are proportional and the reliability function satisfies

$$R_i(t) = R_0(t)^{exp(\beta' x_i)} \tag{3}$$

with $R_0(t)$ the baseline reliability function and x_i the vector of values of p explanatory variables for the ith individual, $i = 1, 2, \ldots, n$. The baseline reliability function will be modelled as a step function, where the steps occur at the k ordered censoring times, $t_{(1)}, t_(2), \ldots, t_{(k)}$, with $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ (subset of the times at which observations are interval–censored). The procedure for choosing these times is explained later.

Let us define

$$\theta_j = \log \frac{R_0(t_{(j-1)})}{R_0(t_{(j)})} \tag{4}$$

where $t_{(0)} = 0$, so that $\theta_j \geq 0$, and at time $t_{(j)}$ we have

$$R_0(t_{(j)}) = e^{-\theta_j} R_0(t_{(j-1)}), \tag{5}$$

for $j = 1, 2, \ldots, k$.

The first step in the baseline reliability function occurs at $t_{(1)}$, so $R_0(t) = 1$ for $0 \leq t < t_{(1)}$. From $t_{(1)}$ and using (5) we have $R_0(t_{(1)}) = \exp(-\theta_1) R_0(t_{(0)})$ which implies that $R_0(t) = \exp(-\theta_1)$ for $t_{(1)} \leq t < t_{(2)}$. Similarly, from $t_{(2)}$ the reliability function is $\exp(-\theta_2) R_0(t_{(1)})$ which implies that $R_0(t) = \exp\{-(\theta_1 + \theta_2)\}$, $t_{(2)} \leq t < t_{(3)}$ and so on, up to $R_0(t) = \exp\{-(\theta_1 + \theta_2 + \cdots + \theta_k)\}$, $t \geq t_{(k)}$. Thus,

$$R_0(t) = \exp\left(-\sum_{r=1}^{j} \theta_r\right) \tag{6}$$

for $t_{(j)} \leq t < t_{(j+1)}$, and the baseline reliability function at any time t_i is given by

$$R_0(t_i) = \exp\left(-\sum_{j=1}^{k} \theta_j \, d_{ij}\right), \tag{7}$$

where $d_{ij} = 1$ if $t_{(j)} \leq t_i$ and $d_{ij} = 0$ if $t_{(j)} > t_i$ for $j = 1, 2, \ldots, k$.

Combining results from equations (3) and (7), we obtain the reliability function for the ith individual at times a_i, b_i.

From $R_i(a_i)$ and $R_i(b_i)$ the response probability p_i of (2) may be expressed in terms of unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$ and the unknown coefficients of the p explanatory variables in the model, $\beta_1, \beta_2, \ldots, \beta_p$. We obtain (see Collett et al., 2003, p. 289)

$$p_i = 1 - \exp\left\{-\exp\ (\beta' Z_i) \sum_{j=1}^{k} \theta_j \ d_{ij}\right\}, \tag{8}$$

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

Table 2: Definition of intervals A_i.

where $d_{ij} = 1$ if $t_{(j)}$ is in interval A_i (intervals A_i are as it is shown in Table 2) and $d_{ij} = 0$ in other cases for $j = 1, 2, \ldots, k$.

This leads to a non–linear model (*generalized non-linear model*) for a set of binary response variables, with values y_i and the corresponding probabilities p_i, given by equation (8) for $i = 1, 2, \ldots, n+c$. The model contains $k+p$ unknown parameters, $\theta_1, \theta_2, \ldots, \theta_k$ and $\beta_1, \beta_2, \ldots, \beta_p$. In the SAS software, the `proc nlmixed` code (SAS Institute Inc., 1999) is used to fit the model for the response probabilities in equation (8) by means of the definition of the distribution of the binary response variables y_i. In this way the estimated parameters $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ are obtained.

Once an appropriate model has been found, we may estimate the baseline reliability function from equation (6),

$$\hat{R}_0(t) = \exp\left(-\sum_{r=1}^{j} \hat{\theta}_r\right) \tag{9}$$

for $t_{(j)} \leq t < t_{(j+1)}$, $j = 1, 2, \ldots, k$ where $t_{(k+1)} = \infty$ and $\hat{\theta}_j$ is the estimated value of θ_j as mentioned above. The estimated reliability function for the ith pipe, $\hat{R}_i(t)$, is obtained by (3) substituting $R_0(t)$ for (9) and β for $\hat{\beta}$ (the vector of estimated coefficients of explanatory variables). The hazard can be obtained from the reliability function by means of well known relationships.

Let us now consider the selection of times C=$\{t_{(1)}, t_{(2)}, \ldots, t_{(k)}\}$. It would appear desirable that times $t_{(j)}$ were all different censoring times, which means shaving different values for a_i and b_i. This would introduce too many θ parameters into the non-lineal model and for this reason it is necessary to choose a subset of disposable times.

As mentioned above the procedure for choosing the partitioning $t_{(j)}$ times is carried out in the following way (see Collett et al., 2003, p. 291). Each interval used in this model of binary data, denoted earlier by A_i, must include at least one of the times $t_{(j)}$. If this is the case, at least one of the values of d_{ij} in equation (8) would be unity, and so $\sum_{j=1}^{k} \theta_j \, d_{ij}$ will be greater than zero. Let us suppose that interval A_i is $(u_i, v_i]$. We take $t_{(1)}$ to be the smallest of the values of v_i, $t_{(2)}$ the smallest v_i such that $u_i \geq t_{(1)}$. Again, we take $t_{(3)}$ to be the smallest v_i such that $u_i \geq t_{(2)}$, and so on, until $t_{(k)}$ as the smallest value of v_i such that $u_i \geq t_{(k-1)}$. In this way, we obtain the minimal set of times used in calculating the baseline reliability function, so that all the intervals A_i include at least one of these times $t_{(j)}$. The fitting model could be improved including a greater number of steps in the subset C=$\{t_{(1)}, t_{(2)}, \ldots, t_{(k)}\}$. This entails adding a new θ–parameter for

each additional time point. All these times are added one by one. So, each fitted model (one for each additional time) will lead to a value of the $-2\log\hat{L}$ statistic. So, the smaller the value of this statistic, the better the model.

Application to water supply networks

In the process of performing the reliability function of the model, a minimal set of ordered censoring times was chosen: C=$\{1,2,3,4,5,6,15\}$. With this minimal set C, in order to choose the variables, we use the statistic $-2\log\hat{L}$ for the model. In the fitting process we concluded that length, diameter, pressure, traffic conditions and the material are prognostic factors on the failure time of the pipes. Table 3 shows the estimated parameters for the hazard ratios of the prognostic factors from the generalized non–linear model. Each individual regression coefficient value can be interpreted in this way, for the explanatory variable length the increase in the risk of failure for an increase of 1 m in the section pipe is 0.3%. In the case of the diameter, the risk of failure decreases by approximately 1% for an increase of 1 mm while the risk increases by 2.11% for each increase of 1 pascal in pressure. On the other hand, pipes situated under sidewalks and normal traffic decrease the hazard rate by about 65% and 42% respectively compared to pipes under heavy traffic. Finally, pipes made with ductile cast iron were 0.23 times less likely to suffer failure than those made with asbestos cement. The rest of the materials gave results that were not significant.

In the second step we increased the number of censoring times of C one by

Variable	$\hat{\beta}$	Exp($\hat{\beta}$)	s.e.($\hat{\beta}$)	p-value.
length	0.0033	1.0033	0.000	$< .0001$
diameter	-0.0039	0.9961	0.000	$< .0001$
pressure	0.0209	1.0211	0.003	$< .0001$
under traf	-1.0079	0.3649	0.213	$< .0001$
normal traf	-0.5367	0.5846	0.218	0.0141
ductile	-1.4346	0.2382	0.094	$< .0001$
gray iron	-0.1124	0.8936	0.158	0.4770
polyethylene	-0.1016	0.9033	0.230	0.6586

Table 3: Estimated parameters for failure time. Generalized non–linear model.

one in the fitted model. The modelling procedure did not show any value which reduces $-2\log\hat{L}$ significatively. So the minimal set C provides the best fitted model so $A_i \bigcap C$ is non–empty and the expansion of the set C does not improve our fitted model.

In order to check the consistency of the Generalized non–linear model we contrasted the results obtained with some well established methods: the semi–parametric Cox regression model and the Generalized Linear Model (GLM). The Cox regression model, also called the Proportional Hazard Model, is designed to analyze the time lapse until an event occurs or time lapse between events. The covariables, one or more predictor variables, are used to predict the event. The Proportional Hazard Model Cox (1972) has been widely used in analizing survival data. The model specifies that the time until failure T, given the covariables vector x, has the hazard function

$$h(t,\, x) = h_0(t)e^{\beta' x}$$

where $h_0(t)$ is an unspecified baseline function and β is the regression coefficient vector.

Generalized Linear Models are an extension of linear models for non–normal distributions of the response variable and non–linear transformations. Therefore, we want to find a linear function

$$E(y|x) = m = \beta_0 + \sum_{i=1}^{p} \beta_i x_i,$$

where a constant variance for the variable Y is supposed. Then, GLM provides a method for estimating a function for the response variable mean as a linear combination of the set of predictive variables, that is,

$$l(E(y|x)) = l(m) = \eta(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

The function of the response mean, $l(m)$, is called a *link* function, and is considered to be the same as a linear function of the predictors, $\eta(x)$ which is called a linear predictor. Each component y_i of Y has a Binomial, Poisson or Gamma distribution. The GLM comprehensive reference is in McCullagh and Nelder (1989).

The results for both models are shown in Table 4. The interpretation of

the parameters in the Cox model is similar to the Farrington model. Length and pressure increase the risk of failure by 0.4% and 2.3% respectively for an increase of one unit in each of these measurements, while the diameter decreases the risk of failure by approximately 1% for each increase of 1 mm. Pipes under sidewalks and normal traffic decrease the hazard rate by about 52% and 42% respectively compared to heavy traffic. Finally, pipes made with ductile cast iron and polyethylene were 1.393 and 5.629 times more likelier to suffer failure compared to pipes made of asbestos cement.

The GLM allows us to analyze the data on the assumption that the number of pipe failures follows a Poisson distribution. The model is formally a Poisson Generalized Linear Model with a logarithmic link function. These models have been used before by Boxall et al. (2007) and more recently by Debón et al. (2010). We can conclude that the mean failure increases slightly for an increase of one unit in length and pressure of the pipes and decreases for an increase of one unit in diameter. On the other hand, pipes under sidewalks and normal traffic decrease the mean number of failures respectively compared to pipes under heavy traffic. Finally, ductile cast iron decreases the mean number of failures while gray iron increases it with respect to asbestos cement.

The analysis of the three models described above shows that pipes which were more prone to failure had the following characteristics: long lengths and large diameters, high pressure and installed under a heavy traffic. As regards the material, the three models do not provide the same results: pipes made with

	Cox model				*Generalized Linear model*		
Variable	$\hat{\beta}$	Exp($\hat{\beta}$)	s.e.($\hat{\beta}$)	p-value	$\hat{\beta}$	s.e.($\hat{\beta}$)	p-value
length	0.004	1.004	0.000	0.000	0.003	0.0004	0.000
diameter	-0.003	0.997	0.001	0.003	-0.003	0.0009	0.000
pressure	0.023	1.023	0.005	0.000	0.027	0.0045	0.000
under traf	-0.723	0.485	0.264	0.006	-1.156	0.2639	0.000
normal traf	-0.548	0.578	0.272	0.044	-0.787	0.2718	0.004
ductile	0.331	1.393	0.146	0.023	-1.795	0.1805	0.000
gray iron	-0.132	0.876	0.184	0.473	0.310	0.1827	0.089
polyethylene	1.728	5.629	0.266	0.000	0.168	0.2808	0.550

Table 4: Estimated parameters for failure time. Cox and Generalized Linear models.

ductile cast iron in the Farrington and GLM models and, pipes made ductile cast iron and polyethylene in the Cox model. We can observe that the values and signs of the coefficients are similar in the three models, despite small differences in the coefficients corresponding to materials. The results of the Cox and GLM are similar to as the Farrington method. This fact gives support to the Farrington model as a suitable model for dealing with the approximated age defined above by means of interval–censored data.

Validation of the model

In order to determinate if the Farrington model was fitted correctly to our set of interval–censored survival data we employed residuals derived from those for right–censored data. Farrington (2000) shows that many standard diagnostic tools of survival analysis have counterparts for interval–censored data. Specifically, it develops interval–censored versions of residuals of Cox and Snell (1968),

Lagakos (1980) and *deviance residuals* (Therneau et al., 1990). We used these results to carry out the validation of our results. All results obtained in this section were calculated with S–Plus software (Venables and Ripley, 2000).

In the first step, we checked the assumption that times until the failure of the pipes are independent of the observation process and in the second step, we highlight those pipes whose times until failure are not well fitted by the model (possible outliers).

As usual we have assumed that the observation process that generates the interval censoring is independent of time until failure and covariates. Therefore it is useful to examine plots of the distribution of interval lengths by observation number and by all covariates on the fitted model. In Figure 1 we can see that the plots do not reveal any systematic differences in the observation process from covariates or prognostic factors.

The model–checking procedure to detect possible *outliers* is based on *martingale and deviance residuals* developed by Farrington (2000) for interval–censored data. Given a sample of interval–censored observations $(a_1, b_1], \ldots, (a_n, b_n]$ and estimated reliability functions $\hat{R}_i$, the martingale residuals are given by the following expression

$$\hat{r}_i^M = \frac{\hat{R}_i(a_i)log(\hat{R}_i(a_i)) - \hat{R}_i(b_i)log(\hat{R}_i(b_i))}{\hat{R}_i(a_i) - \hat{R}_i(b_i)} \tag{10}$$

for $i = 1, 2, \ldots n$.

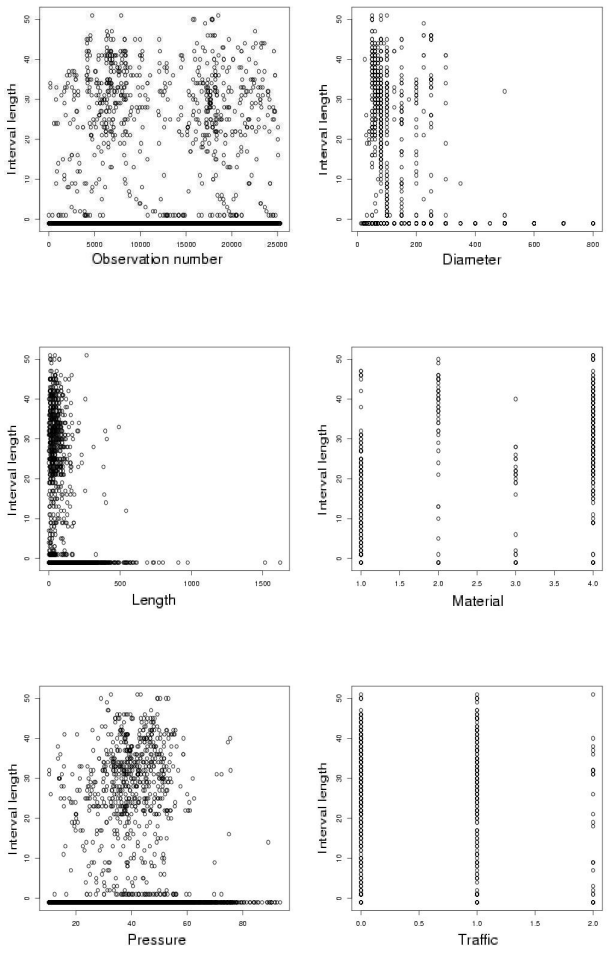


Figure 1: Distribution of interval length by observation diameter, length, material, pressure and traffic covariates.

In large samples, *martingale residuals* have zero mean under the correct model. In fact, we can see that only one pipe (pipe number 765) has a longer time until failure than the rest, however it has some bad prognostic factors (Figure 2). As the *martingale residuals* take values in the interval $(-\infty, 1]$, we have plotted *deviance residuals* (a transformation of the martingale ones), which are more symmetrically distributed about zero and so the plots based on these residuals

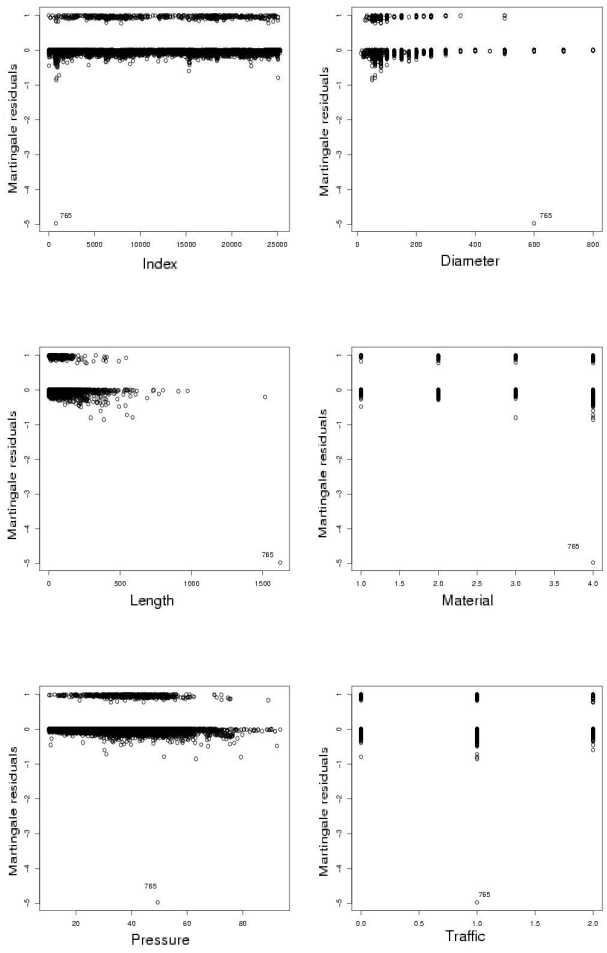


Figure 2: Martingale residuals by observation diameter, length, material, pressure and traffic covariates.

are easier to interpret. *Deviance residuals* are given by the expression

$$r_i^D = \text{sgn}(r_i^M)\left[2log\left\{\frac{R_0(a_i)^{\eta_i} - R_0(b_i)^{\eta_i}}{R_0(a_i)^{exp(\beta' Z_i)} - R_0(b_i)^{exp(\beta' Z_i)}}\right\}\right]^{1/2} \tag{11}$$

where

$$\eta_i = \frac{log\{\Lambda_0(b_i)\} - log\{\Lambda_0(a_i)\}}{\Lambda_0(b_i) - \Lambda_0(a_i)}, \tag{12}$$

$\eta_i = 0$ if $b_i = \infty$, $\eta_i = \infty$ if $a_i = 0$, and Λ_0 is the cumulative hazard function defined by $\Lambda_0 = 0$ if $t = 0$ and, $\Lambda_0 = \theta_1 + \cdots + \theta_j$ if $t = t_j$, $j = 1, \ldots, k$.

Figure 3 shows *deviance* and *martingale residuals* plotted against the logarithm of interval length (or -1 for right–censored observations). The deviance transformation again only suggests one apparent outlier slightly separated from the bulk of the data, pipe number 765.

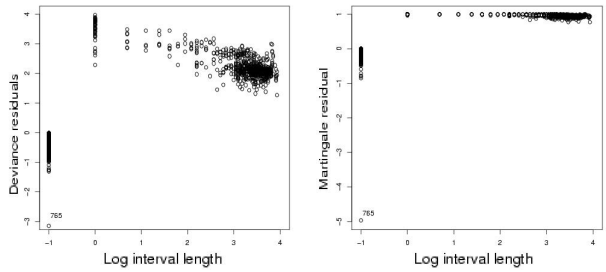


Figure 3: Log-cumulative hazard functions for explanatory variables against logarithm of selected times.

Discrimination ability of the model

In order to compare the Farrington model above with the Cox and the generalized linear models we assesed the discrimination ability of each one of them. Discrimination quantifies the ability of the model to correctly classify pipes into one of two categories (failures and non-failures). The model that places each pipe in the category to which it truly belongs can be said to have perfect discrimination. Initially a measurement for the discrimination was suggested by Harrell et al. (1996) in dichotomous outcomes whose development was motivated by the extension of

the concept of the receiver operating characteristic (ROC) curve which quantifies discrimination in logistic regression. However, measuring discrimination in reliability analysis is more difficult than logistic regression. Therefore, Pencina and D'Agostino (2004) developed a measure for good discrimination in the context of survival analysis: the *C index*. We show how, in reliability analysis, the C index allows us to discriminate between two models in our analysis.

The *C index* considers that pipes with longer predicted failure time actually survive longer without experiencing any failure. we have the *actual survival time* X_i and the *predicted survival* T_i given by the model. Harrell et al. (1996) points out that predicted probabilities of survival until any fixed time point, Y_i, can be used instead of the predicted survival times T_i. Moreover, this interchange between probabilities can be used in the most common models in reliability analysis such as proportional hazards and accelerated failure time models. Next we order the pipes from the smallest actual survival time to the highest actual survival time. We form all posible pairs of pipes and then we compare them. We have the following four posible situations for each pair of pipes:

1. $X_i > X_j$ and $Y_i > Y_j \Rightarrow$ *concordant pair.*

2. $X_i < X_j$ and $Y_i < Y_j \Rightarrow$ *concordant pair.*

3. $X_i > X_j$ and $Y_i < Y_j \Rightarrow$ *discordant pair.*

4. $X_i < X_j$ and $Y_i > Y_j \Rightarrow$ *discordant pair.*

It is assumed that distributions of survival times and predicted probabilities of reliability analysis are continuous. This avoids any ties in Y's, but not in X's. Not all pairs are either concordant or discordant. Thus in constructing the C index, only those pairs of pipes in which at least one had a failure are used. This results in either failure versus failure or failure versus non-failure comparisons. Such pairs are called *usable*. On the other hand, if two pipes did not develop any failures by the end of the study we can not compare them in terms of the predictions. In this case, pairs formed by such pipes are called *unusable*. For each pair of pipes it is defined

$$c_{ij} = \begin{cases} 1 \ if \ X_i < X_j \ and \ Y_i < Y_j \ or \ X_i > X_j \ and \ Y_i > Y_j \\ \\ 0 \ otherwise \end{cases} \tag{13}$$

In order to construct the *C index* we assume N pipes such that $N = k + n$ where k is the number of failures and n is the number of non–failures. Let c_h denote the number of pipes that are concordant with the hth pipe in the data, $h = 1, 2, \ldots, N$. With the definition (13) of c_{ij} we can write

$$c_h = \sum_{h \neq j} c_{hj}$$

Let d_h be the corresponding number of discordant pairs. Then the C index is the proportion of all *usable pairs* in which the predictions and outcomes are concordant

$$C = \frac{p_c}{p_c + p_d}$$

with

$$p_c = \frac{1}{N(N-1)} \sum_h c_h$$

and

$$p_d = \frac{1}{N(N-1)} \sum_h d_h$$

where p_c denotes the number of pipes that are concordant with the h-th pipe in the sample, $h = 1, \ldots, N$ and p_d are defined in the same way.

The estimated value of the C index was found to be 0.83 for the Farrington model and 0.70 for the Cox model. We concluded that the agreement between the predicted and observed outcomes are closer in the Farrington model. Also for the GLM model we calculated the ROC curve area following the methodology used by Debón et al. (2010) which was 0.82. Since the *C index* is an natural extension discrimination method to the ROC method we were able to compare the three models. We did not use the ROC curve for the Farrington and Cox model because the data are censored and so the discrimination would not be good. It can be seen that the Cox model produces the worst fit.

Conclusions

This paper is motivated by the research of time until failure of pipes and by the prognostic factors associated to them. As we did not know some failure times of pipes before year 2000 we approximated the age of the pipe's sections by means of interval–censored data and therefore we decided to employ a generalized non–linear model developed by Farrington. We think that this option is very interesting because its implementation is easy using standard statistical software and the interpretation of results is analogous to Cox's model. This author also provides residuals derived from those for right–censored data and as interval–censored data are more awkward to examine than right–censored data, these residuals at least provide a easy way of doing so.

In the model checking, we have presented index plots of interval lengths to check if the observation process that generates the interval censoring was independent of time until failure and the covariates: index plots of martingale and deviance residuals to identify outliers and the log–cumulative hazard functions for each explanatory variable against the logarithms of selected times (those used in constructing the baseline reliability function). From these plots we can see that the Farrington model is well fitted to our data. Only one pipe is highlighted because it presented some bad prognostic factors and its time until failure was long. From the deviance residual plot we can conclude that it was not actually an outlier observation.

The study sought to provide insight into the impact of different variables on the risk of failure in water supply networks. All variables resulted significant in the three models. We conclude the main network characteristics affect the risk of failure: an increase in the length and pressure, a small diameter, the material used and to make the pipes and a heavy traffic conditions of the street. So we propose a clear framework for decision support in the rehabilitation of water supply systems for that company.

We compared the Farrington model with the Cox and Generalized Linear models. Although the ROC (receiver operating characteristic) analysis method is widely used to compare two competing diagnostic tests, we only have calculated it for the GLM model because the ROC analysis don't show a correct discrimination with censored data. So we calculated the *C index* (motivated by the extension of the concept of the ROC) for the Farrington and Cox models. The *C index* is an natural extension discrimination method to the ROC curve allowing comparison of the three models. In that comparison we concluded that the Cox model produces the worst fit. Moreover, this methodology allowed us to establish a threshold at which a pipe can be considered high-risk, which lead the company to make decisions about the renewal of the network.

References

Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J. and Unwin, D. M. (2007). Estimation of burst rates in water distribution. *Water Mangement*, pages 83–88.

Collett, D. (2003). *Modelling Survival Data in Medical Research* 2^{th} *ed.* Chapman & Hall/CR, Boca Raton, Florida.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, pages 187–220.

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society*, Series B, pages 248–275.

Debón, A., Carrión, A., Cabrera, E. and Solano, H. (2010). Comparing Failure Risk Models in Water Supply Networks using ROC Curves. *Reliability Engineering and System Safety*, pages 43–48.

Farrington, C. (1996). Interval censored survival data: A generalized linear modelling approach. *Statistics in Medicine*, pages 283–292.

Farrington, C. (2000). Residuals for proportional hazards models with interval–censored survival data. *Biometrics*, pages 473–482.

Finkelstein, D.M. (1986). A proportional hazards model for interval–censored failure time data. *Biometrics*, pages 845–854.

García, B., Rubio, G., Santamaría, C., Pontones, J., Vera, C. and Jiménez, J. (2005). A predictive mathematical model in the recurrence of bladder cancer. *Mathematical and Computer Modelling*, pages 621–634.

Hölmang, S., Hedelin, H., Anderström, C. and Johansson, S. (1995). The relationship among multiple recurrences, progression and prognosis of patients with stage ta and t1 transitional cell cancer of the bladder followed for at least 20 years. *Journal of Urology*, pages 1823–1827.

Harrell F.E. et al. (1996). Tutorial in biostatistics: multivariate prognostics models: issues in developing models evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, pages 361–387.

Huang, J. and Wellner, J.A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seatle Symposium in Biostatistics: Survival Analysis*, Lin DY, Fleming TR (eds). Springer, New York, pages 123–169.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*, 2^{th} ed. Wiley-Intercience, Hoboken, New Jersey.

Lagakos, S.W. (1958). The grafical evaluation of explanatory variables in proportional hazards regression. *Biometrika*, pages 93–98.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2^{th} ed. Wiley–Interscience, Hoboken, New Jersey.

Lindsey, J.C. and Ryan, L.M. (1998). Tutorial in biostatistics: Methods for interval–censored data. *Statistics in Medicine*, pages 219–238.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.*, London: Chapman and Hall.

Peto, R. (1973). Experimental survival curves for interval censored data. *Applied Statistics*, pages 86–91.

Pencina, M. J. and D'Agostino R. B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population values and confidence interval estimation. *Statistics in Medicine*, pages 2109–2123.

Santamaría, C. (2006). *Modelización matemática de los factores de riesgo en el carcinoma vesical superficial. Nomogramas de predicción de recaída para el seguimiento individualizado de los pacientes.* PhD thesis, Universidad Politécnica de Valencia. Departamento de Matemática Aplicada.

Santamaría, C., García–Mora, B., Rubio, G. and Pontones, J. (2008). Modelling the recurrence of bladder cancer. *Acta Aplicanda Mathematicae*, pages 91–105.

SAS Institute Inc. *User's Guide. Version 8.*, Cary, NC: SAS Institute Inc.

Venables, W. N. and Ripley, B. D. (2000). *S Programming.*, Springer, New York.

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data.* Springer, New York.

Therneau, T.M., Grambsch, P.M. and Fleming, T.R. (1990). Martingale–based residuals for survival models. *Biometrika*, pages 147–160.

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