INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Stochastic models for prediction of pipe failures in water supply systems

André Damião da Costa Martins

Dissertação para obtenção do Grau de Mestre em
Matemática e Aplicações

Júri

Presidente:	Prof. Doutor António Manuel Pacheco Pires
Orientador:	Prof. Doutora Maria da Conceição Esperança Amado
Co-orientador:	Doutor João Paulo Correia Leitão
Vogal:	Prof. Doutora Ana Maria Nobre Vilhena Nunes Pires de Melo Parente
Vogal:	Doutora Maria do Céu de Sousa Teixeira de Almeida

Outubro de 2011

Abstract

The failure prediction process plays an important role in infrastructure asset management of urban water systems. This process aims at assessing the future behaviour of a urban water network. However, failure prediction in urban water systems is a complex process, since the available failure data often present a short failure history and incomplete records.

In the study presented in this thesis, three different failure prediction models, the single-variate Poisson process, the Weibull accelerated lifetime model and the linear extended Yule process, were implemented and explored in order to identify robust and simple models that combine good failure prediction results using short data history. The three models were applied to a Portuguese urban water supply system.

The Weibull accelerated lifetime model presented the best results throughout the comparison of the three models, presenting accurate predictions and a good ability to detect pipes with high likelihood of failure. Nevertheless, the expected number of failures can only be obtained using a Monte Carlo simulation, which can be a time-consuming procedure.

The linear extended Yule process could also effectively detect pipes more prone to fail. However, it presented a clear tendency to overestimate the number of future failures.

The single-variate Poisson process is a very simple stochastic process, that is easy to understand and to apply. Due to its simplicity, this prediction process lacks the ability of differentiating effectively the water network pipes, which leads to lower quality individual failure predictions.

It is noteworthy that no significant difference between the three models results was found when predicting failures in pipes with no failure history.

Key words: Failure prediction; Infrastructure asset management; LEYP; Poisson process; Water supply networks; Weibull regression.

Resumo

O processo de previsão de falhas tem um papel extremamente importante na gestão patrimonial de infra-estruturas em sistemas urbanos de água. Este processo tem como objectivo avaliar o comportamento futuro dos sistemas urbanos de água. No entanto, a previsão de falhas em sistemas urbanos de água é um processo complexo, uma vez que os dados de falhas disponíveis apresentam frequentemente um historial de falhas curto e registos incompletos.

No estudo efectuado nesta tese, três modelos de previsão de falhas, o processo de Poisson univariado, o modelo de regressão Weibull e a extensão linear do processo de Yule (LEYP), foram estudados com o objectivo de identificar modelos simples e robustos que permitam boas previsões de falhas com base num historial de falhas curto. Os três modelos foram ajustados aos dados de falha provenientes de um sistema de abastecimento de água português.

O modelo de regressão Weibull foi o modelo que apresentou melhores resultados durante a comparação dos três modelos, apresentando previsões exactas e uma boa capacidade em identificar as condutas com maior tendência para falhar. Contudo, o valor esperado do número de falhas tem de ser obtido através de simulações de Monte Carlo, o que pode ser um processo com alguma complexidade temporal.

O LEYP consegue, igualmente, identificar eficazmente as condutas com maior tendência para falhar. No entanto, apresenta uma clara tendência para sobrestimar o número de falhas.

O processo de Poisson univariado é um processo estocástico muito simples, que é fácil de compreender e de aplicar. Devido à sua simplicidade, este processo de previsão de falhas não tem a capacidade de diferenciar eficazmente as várias condutas do sistema urbano de água, o que leva a piores resultados no que toca a previsões de falha em cada conduta.

É importante notar que a diferença entre os resultados dos três modelos não foi muito significativa quando aplicados a condutas sem historial de falhas.

Palavras Chave: Gestão patrimonial de infra-estruturas; LEYP; Previsão de falhas; Processo de Poisson; Regressão Weibull; Sistemas de abastecimento de água.

Acknowledgements

I owe my deepest gratitude to my supervisors, Conceição Amado and João Paulo Leitão, for all the guidance throughout the development of this thesis.

I want to thank LNEC, and specially Sérgio Coelho, for welcoming me in the urban water division, providing all the conditions to develop this work, including the financial support through the AWARE-P and TRUST projects.

I acknowledge SMAS Oeiras e Amadora by having provided the failure data, with the complete pipe inventory and maintenance records, that was essential to the work presented in this thesis.

Finally, I want to thank my family for their support, in good and in bad times.

Index

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminaries

Managing urban water systems is not a simple activity, and, due to several factors, such as climate change, economic restrictions, ageing of the systems, increasing customer demands, it is becoming even more complex. Currently, for environmental and economic reasons, water utilities are increasingly concerned on minimising water losses and in meeting customer demands. Traditional infrastructure asset management methodologies, such as the *like-for-like* strategy, are not adequate for urban water system long-term planning. To help urban water utilities facing these new challenges, new methodologies are being developed, such as the AWARE-P methodology (Alegre *et al.*, 2011).

Failure prediction plays a major role during planning and decision support processes, whether in evaluating different solutions for the identified urban water system problems, whether in evaluating the current system performance under several scenarios. Nevertheless, urban water utilities are only recently becoming aware of the importance of keeping an organised, updated and complete inventory and failure database. Failure data history is very limited in the majority of the urban water systems. For this reason, the task of predicting failures is more difficult than expected. A great challenge is to find a failure prediction model that can produce good quality results, even in the cases of lack of failure data.

The aim of this work is to identify robust and simple models that combine: good predictions for short failure records; robustness when applied to different pipe samples; and simplicity. The studied failure prediction models were: the single-variate Poisson process; the Weibull accelerated lifetime model; and the linear extended Yule process. These models were applied to a Portuguese urban water supply failure data provided by the Serviços Municipalizados de Água e Saneamento de Oeiras e Amadora (SMAS O&A). The prediction results were compared and discussed, regarding the prediction accuracy and the ability of each model to identify the pipes with higher likelihood of failing; some improvements were also suggested.

The work presented in this thesis has been developed within the AWARE-P project framework, allowing the author working with several water utilities during this study. This fact contributed to a better awareness of the main concerns of Portuguese water utilities in infrastructure asset management. Working with these water utilities highlighted the importance of developing flexible and simple failure prediction models, that can produce accurate predictions with limited failure records.

All statistical analysis in this work was conducted with the use of the statistical software R (R Development Core Team, 2011).

1.2 Brief review of failure prediction models

The first statistical failure prediction models applied to urban water systems were developed from the early 1980s. The first models were deterministic, in which the decision variable (number of failures or time to next failure) is obtained directly from a function of explanatory variables. One of the first models was developed by Shamir and Howard (1979) that related the number of failures per unit length per year with the exponent of the age of a pipe. In this model no covariates were used, pipes were divided into homogeneous groups according to some of their attributes, such as diameter and material. Other time exponential models were proposed, in which more covariates were considered, e.g. Clark *et al.* (1982).

A different group of deterministic models describes the time to the first failure as a linear combination of several pipe attributes. The main disadvantage of these models is the lack of capability of dealing with censored data.

Deterministic models, whose behaviour is predicted by mathematical functions, can give an estimated value of one or more known variables. Nevertheless, to model failures in water systems, stochastic models might be more suitable, since they take into account the random nature of these failures.

There are two different types of stochastic models: single-variate and multivariate models.

Herz (1996) proposed a new single-variate survival analysis. In his work, Herz defines a new lifetime probability distribution, with simple density and survival functions, to describe the lifetime of a pipe. Pipes were divided into several homogeneous groups in which the parameters of the Herz distribution were estimated. Other models that use the attributes of pipes as grouping criteria, rather than covariates, are the Bayesian diagnostic model proposed by Kulkarni *et al.* (1986) and the semi-Markov chain analysis suggested by Gustafson and

Clancy (1999).

The Poisson process, implemented in this work, is a single-variate model with a constant failure rate for each pipe group. In this work, it is assumed that the rate of the process is proportional to the pipe length. As such, the probability distribution associated with the number of pipes during a time period t is Poisson($\lambda l_i t$), where l_i defines the length of pipe i and λ is assumed constant in each pipe group. This model is described in detail in Chapter 3.

Single-variate models have one main issue: failure data need to be divided into several homogeneous groups, assuming that the failure rate is the same in each group. However, a large number of groups would lead to a shortness of records in each group, which could consequently lead to biased predictions.

Multivariate models, which use covariates, allow to differentiate the pipe failure distributions without splitting failure data. In addition, these models allow a better understanding of how the different pipe attributes influence the occurrence of failures.

Cox (1972) developed the Proportional Hazard Model (PHM), where the hazard rate is given as a function of the time and a vector of covariates:

$$h(t|\boldsymbol{x}) = h_0(t)e^{\boldsymbol{x}^\intercal\boldsymbol{\beta}}, \tag{1.1}$$

where t is the elapsed time from the last failure, $\boldsymbol{x}$ is the covariates vector, $\boldsymbol{\beta}$ is a coefficients vector and $h_0(t)$ is a baseline function that can be estimated.

The proportional hazard model assumes that each covariate acts multiplicatively in the hazard function (or the covariates are assumed to act additively on $\ln h(t|\boldsymbol{x})$). Cox (1972) suggests a method for predicting coefficients $\boldsymbol{\beta}$ not regarding the baseline function h_0; this method is based on the partial likelihood estimation. The baseline function h_0 can, then, be approximated in order to fit the empirical hazard function. Jeffrey (1985) was the first to apply this method to a water network failure data. In his work, after $\boldsymbol{\beta}$ has been estimated, h_0 was approximated to a second degree polynomial, in order to translate the bathtub effect in the hazard function.

An accelerated lifetime model was applied in water network research by Lei (1997). This model defines the logarithm of the time to next failure as the linear combination of the covariates vector and a random error variable. A particular case, is the Weibull accelerated lifetime model, applied in Le Gat and Eisenbeis (2000), when the times between failures are given by a Weibull distribution. When the proportional hazard model uses an hazard base function with the Weibull power law form (i.e. $h_0(t) = \delta t^{\delta-1}$) these two models are analogous. That is the reason Eisenbeis and Le Gat call it the Weibull proportional hazard model. This model is detailed in Chapter 4.

New lifetime distributions, with higher complexity, have been developed recently in order to approximate the hazard function to a bathtub curve, such as: the log-extended Weibull distribution (Silva *et al.*, 2009); and the lognormal-power function distribution (Reed, 2011). In this work, the empirical hazard function was estimated for the available failure data, using function *muhaz* of the statistical software R, which is based on a method described in Gefeller and Dette (1992). The empirical hazard function, presented in Figure 1.1, does not seem to have a bathtub curve shape. This fact is probably due to the short period of failure records used in this work. In fact, the two-parameter Weibull distribution can already present a hazard rate that approximates the empirical hazard rate in Figure 1.1.

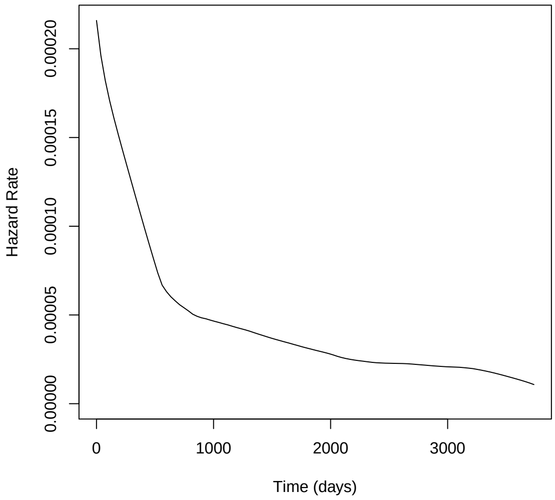

Figure 1.1: Hazard rate estimate.

Another recent developed approach, suggested by Le Gat (2009), is the linear extended Yule process. This is a counting process where the rate of the process is given by a linear function of the number of past events. A particular case of this model, with the process rate depending on the pipe age and other covariates, is studied in Chapter 5.

Multivariate models present the disadvantage of fixing *a priori* how the covariates act on the failures distribution. In WALM and LEYP it is assumed that the Weibull scale parameter and the process rate, respectively, are proportional to the exponent of a linear combination of the covariates vector.

A more complete description of failure prediction models can be found in Kleiner and Rajani (2001).

1.3 Data issues in prediction of failures

The collection of urban water systems failure data is, in general, a relatively uncommon process. Recently it is becoming a new concern, as water utilities start realising the importance of keeping accurate information related to their systems. This is a recent concern and, as such, difficulties in collecting information arise. Available failure data used in the development of failure models present several weaknesses that are exposed in this section. Some refer only to the particular case of the data set used in this work, but others can be applied to most of failure data sets.

Since the components (e.g. pipes, sewers and manholes) of a urban water network failure analysis have, in average, a very long lifetime, it is not easy to have recorded the complete history of all these components. Failure data of urban water systems are typically left-truncated and right-censored.

Failure data are traditionally left-truncated, because there is a number of failures, that occurred in the components before the beginning of the recorded history, that is unknown. Any component installed before the beginning of the recorded observations may have, or not, failed in the past. In addition the failure data set used in this study comprises only operating pipes. This means that there is a considerable amount of pipes that were already decommissioned that is completely unknown; this fact can significantly bias the failure predictions.

Another typical characteristic of failure data is that they are often right censored. Whenever records show the elapsed time without failing of some pipe, but does not show the exact time of failure, this elapsed time is a right-censored information: only the time the pipe survived is known, but the exact time of occurrence is unknown. This is an important factor that needs to be taken into account when failure prediction models are applied.

The oldest failure records in the failure data set used in this study date back to 2001. That is, the failure observation period is of 10 years only. Although the size of the failure data set is relatively large, about 11,500 pipes and 1,900 failures, with such an observation period it will be difficult to assess the deterioration process. The tight observation window causes a great amount of pipes with no recorded failures; more than 90% of pipes have never failed during the 10 years of observation. This makes difficult to fit the lifetime distribution and the distribution of the number of failures.

Another issue that needs to be taken into account is the scarcity of the collected component characteristics. The majority of failure models developed to predict pipe failures in urban water systems use the following variables: pipe diameter, material, installation year and length. Besides these, they often make use of soil and traffic characterisation. Some models even use environmental characteristics, such as temperature and precipitation, and operating pressure. Nevertheless, the available data, used in this work, present only five variables: diameter, material, installation year, length and roughness. More available variables could help to differentiate pipes, helping to derive the failures distribution according to the different pipe characteristics.

The inconsistencies that can be found when analysing the failure data are another frequent problem. Failure data are generally presented in two data tables, the pipe inventory and the maintenance records table. It is from the combination of these two tables that several inconsistencies can be found, such as pipes that have associated failures before their installation year. These inconsistencies are usually caused by careless update of the tables and also due to the lack of a verification process that should be conducted after the update procedure.

These data issues are an important part of the failure data analysis. It is important to develop and apply models that are robust to these problems and can still give good predictions.

1.4 Thesis outline

This thesis is organised in six chapters, as follows. Chapter 1 introduces the concepts and the motivation of the work presented in the thesis. The task of predicting failures in water supply systems is introduced and discussed briefly in Section 1.1; a summarised review of failure prediction models is presented in Section 1.2 , in order to put this work in context; in Section 1.3, data limitations in predicting failures are presented.

Chapter 2 describes the failure data used throughout this study. Explaining how the different variables are related among them and how they influence the failure rate.

In Chapters 3, 4 and 5 the failure prediction models were fitted. Each of these chapters presents a brief theoretically description of the model, the application of the model to the failure data and issues and improvements proposed in by the author.

In Chapter 6 is conducted a thorough comparison between the three failure prediction models.

Finally, Chapter 7 summarises the main findings of this thesis, discussing the advantages and disadvantages of each model.

Chapter 2

Exploratory data analysis

In this chapter, the results of the failure data analysis, from a urban water supply network, are presented. As mentioned in Chapter 1, failure predictions depend on the quality of the available data. It is of utmost importance to explore the failure data in order to develop good prediction models and to analyse objectively the failure prediction results.

First, a basic description of the data source is done, after that, a correlation matrix is computed to try to understand the relationship between the different attributes of the pipe network. Finally, an analysis is conducted to understand how failure rate is influenced by each of these attributes.

2.1 Failure data structure

The failure data used in this work were provided by SMAS O&A. They were presented in two data tables: the pipe inventory table and the maintenance records table. The pipe inventory table consists of a collection of all pipes identified by a unique code IPID. Each pipe is characterised by several attributes, such as pipe material, diameter, length and installation year.

The maintenance records table consists of all rehabilitation actions operated in the water network system. Each rehabilitation action is identified by a unique code (WOID) and is uniquely associated to the pipe identifier in which the rehabilitation action occurred (IPID). In addition to these attributes, each rehabilitation action is characterised by the failure type that caused the need of rehabilitation and its date and duration.

The only failures considered in this study were pipe breaks.

These two tables were combined using the IPID columns, in order to associate attributes, such as number of failures of each pipe and age of pipe at each failure.

2.2 Water supply network

As mentioned in Chapter 1, failure data present several issues. In the particular case of the data provided by SMAS O&A, some of the issues are the missing data (empty attribute records) and the anomalous data (inconsistencies between the pipe inventory and the rehabilitation records). Despite the reduction of the overall data set size, the option, taken in this work, was to remove incomplete and anomalous records (i.e. pipe and failure records that do not have basic information or are inconsistent). Other approaches such as, imputation and survival models with missing data could be considered, but this would bring new complications in data analysis.

The available data is characterised as follows:

- Number of pipes: 11,472.
- Total length: 367km.
- Number of rehabilitation actions: 1,921.
- Number of pipes with no failures: 10,329 (90%).
- Total length of pipes with no failures: 287km (78%).

To characterise the water supply system used in this work, the study of the basic graphical and numerical summaries of data pipes according to each main variable was carried out. Results for the pipe material are presented in Figure 2.1.

As can be seen in Figure 2.1, the more common pipe materials in this system are the Asbestos Cement (AC), High-Density PolyEthylene (HDPE) and PolyVinyl Chloride (PVC). These three materials together represent more than 95% of the entire water supply network. Ductile Cast Iron (DCI) still has some significance, whereas Galvanised Steel (GS) is of little relevance, representing only 2.5km, and Cast Iron (CI) is completely negligible, since it only represents 0.035km. Therefore, when failure data are divided into different material categories, CI pipes should be discarded since it has no significance in this case; GS pipes may or may not be discarded. Another option would be to include these pipe materials in other material categories, if they share similar characteristics.

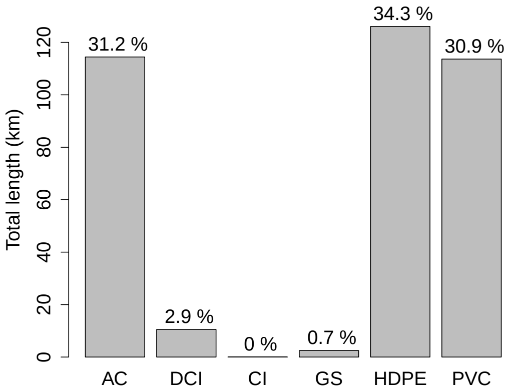

Figure 2.1: Total length frequencies by pipe material.

To assess the age of the water network, the decade of installation was considered rather than the year, because the older pipes have only the record of installation decade. Only pipes installed after 1980 have the exact installation year information. Figure 2.2 shows the plot of the total length per installation decade.

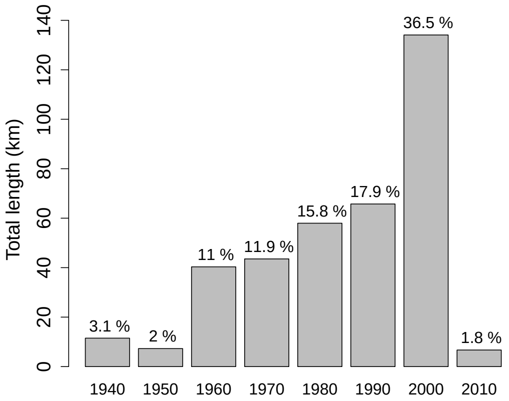

Figure 2.2: Total length frequencies by installation decade.

From this plot, it can be seen that the network is relatively recent. The less representative decades are the two oldest ones. Decade 2010 is a special case, as it represents only one year of installation. Therefore if the failure data set is divided by installation decade, the oldest decades can probably be aggregated in one class and decade 2010 can be included in decade 2000.

The general summary of the frequencies of the three remaining basic variables is presented in Table 2.1.

Table 2.1: Summary of diameter, length and number of failures variables.

Variables	Min	1st Quartile	Median	3rd Quartile	Max	Mean	St.Dev.
Diameter (mm)	20	90	110	125	630	120	62.3
Length (m)	0.1	2.4	10.5	44.5	904.8	32.0	51.9
Number of failures	0	0	0	0	16	0.17	0.66

One fact that can be drawn from Table 2.1 is that the length variable has a great variability, with range values between 0.1m and 905m. Since most models study each pipe individually, all pipes contribute equally to the model, so the results of the fitted models (Chapters 3, 4 and 5) could be improved if the available data had more uniform pipe lengths.

2.3 Statistical relationships between variables

In this section, the possible relationship between variables is investigated. Some conclusions drawn in this section will support the choice of the covariates to be used in the implemented multivariate models.

Pipe material and roughness are completely related, as can be seen in Table 2.2. Therefore it is redundant to use both variables. Since it is the different pipe material characteristics that are believed to influence the failure rate, it was decided to use pipe material as a covariate rather than roughness. However, the fact that roughness is a continuous variable makes this attribute useful during some calculations, such as the estimation of the Pearson correlation coefficients.

In Figure 2.3 the roughness variable is used with the purpose to translate the relationship between other variables and pipe material. The high correlation coefficient between roughness and installation decade indicates a significant dependence between the pipe material and the installation decade variables. Figure 2.4 clearly shows that the three main pipe materials are associated with a specific installation period. The most recent pipe material is HDPE, having been installed mainly during the 2000 decade. Polyvinyl chloride pipes (PVC) present a large range of installation years focused in the 1990 decade. Asbestos cement (AC) pipes are the oldest pipes of the water supply system used in this study.

The high dependence between pipe material and the installation decade is a good reason to use pipe material as grouping criteria and not as a covariate. When using two dependent

Table 2.2: Frequency table between roughness and material.

	Roughness					
Material	**0**	**90**	**110**	**120**	**135**	**145**
AC	1	0	0	3082	0	2
DCI	0	284	50	0	0	0
CI	0	0	4	0	2	0
GS	0	80	2	0	0	0
HDPE	4	1	0	0	6	3979
PVC	0	0	0	0	3975	0

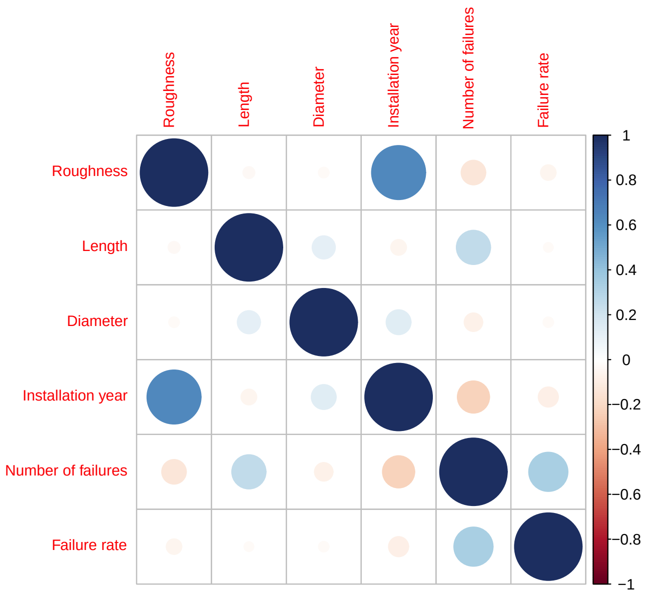

Figure 2.3: Correlation matrix plot.

variables as covariates it becomes difficult to understand the effect that each variable has in the failure distribution. Whereas if pipe material is used as grouping criteria, it is possible to understand the effect of the age of pipes in failure rate, for each pipe material group.

From Figure 2.3 it appears that there are no other significantly dependent variables. Number of failures and failure rate variables were introduced in the correlation matrix to understand how other pipe attributes may influence them. Failure rate is obtained for each pipe as the number of failures over its length. Number of failures seems to be influenced by every

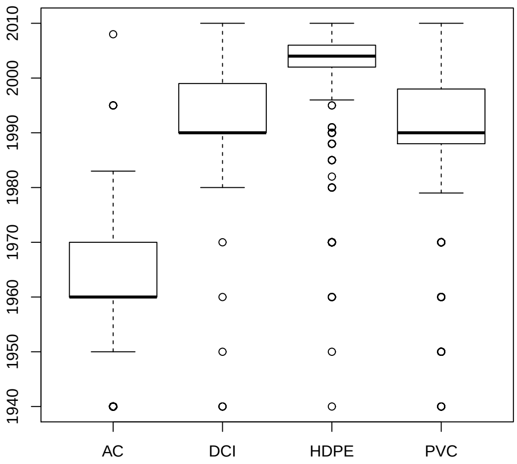

Figure 2.4: Boxplot of installation year per material.

other variable. Besides failure rate, length is the variable that influences most the number of failures. A high correlation coefficient between these two variables is expected, as common sense indicates that these two variables are strongly related. From the correlation matrix of Figure 2.3 it is difficult to indicate the pipe attributes that can influence the most the failure rate.

2.4 Effect of variables in failure rate

In this section, the variable effect on failure rate is analysed. For categorical variables, the failure rate is estimated for each category, assuming it is constant. For continuous variables, they are categorised first and then the failure rate is similarly estimated in each category, as the categorical variables.

Assuming a Poisson process in each category, the failure rate is the maximum likelihood estimate of the Poisson process rate (Equation 3.2, for further details on Poisson process, see Chapter 3). The estimated failure rate, $\hat{\lambda}_k$, in some category C_k, is given by the number of all failures in pipes of category C_k over the sum of the product of each pipe length by the pipe observation time (Equation 2.1).

$$\hat{\lambda}_k = \frac{\sum_{i \in C_k} n_i}{\sum_{i \in C_k} t_i l_i}, \tag{2.1}$$

where n_i is the number of failures, l_i is the length and t_i the observation period of pipe i.

In this thesis, throughout the text it will be used "failure rate" instead of "estimated failure rate", hopping the reader will understand the context within it is used.

2.4.1 Pipe Material

Material is a categorical variable, therefore it is already divided in different categories, it is only needed to compute the failure rate for each material. The given results are presented in Table 2.3.

Table 2.3: Failure rate per material.

Material	Failure rate	Total Length
	(No. of failures/km/year)	(km)
AC	1.041	114.4
DCI	0.139	10.5
CI	3.100	0.035
GS	0.354	2.5
HDPE	0.253	126.1
PVC	0.396	113.7

Failure rate for CI pipes is not significant, CI pipes represent less than 0.01% of the entire water supply network. Therefore this material shall not be considered when pipe material is taken into as a covariate or as grouping criteria. The pipe material with a smaller failure rate is DCI, which, once more, is not as representative as the three main materials: AC, HDPE and PVC. Asbestos cement is, in this case, the material that presents the highest failure rate; it is also the oldest material that has been installed in this water supply network. On the contrary, HDPE is the material with smaller failure rate of the most common pipe materials of this water supply network; it is also the newer material that has been installed. As discussed above, there is a strong correlation between pipe material and installation decade variables. Thereby, it is needed to determine if this effect on failure rates is due to the pipe material or the pipe installation decade (i.e. pipe age).

2.4.2 Pipe installation decade

To assess how the pipe age influences the failure rate, pipes were categorised according to its installation decade, computing the failure rate for each pipe installation category.

Table 2.4: Failure rate per Installation decade.

Decade	Failure rate (No. of failures/km/year)	Total Length (km)
1940	1.35	11.5
1950	1.37	7.3
1960	1.24	40.3
1970	0.93	43.5
1980	0.48	58.0
1990	0.33	65.8
2000	0.23	134.1
2010	0.12	6.7

Table 2.4 shows a tendency for a decreasing failure rate with the installation decade. This clearly indicates that the pipe ageing has a direct effect on failure rate. To understand if this effect is due to pipe ageing or pipe material, the evolution of failure rate as a function of installation decade is plotted in Figure 2.5, for each pipe material.

When dividing the failure data into several categories there is the risk of obtaining very small categories, which are not representative. Plotting the failure rate in those categories could lead to misinformation, thus categories with less than 1km total length were not considered in Figure 2.5. For HDPE material, since more than 96% of HDPE pipes only were installed after 2000, the evolution of failure rate could only be plotted when computed for each installation year, rather than considering the installation decade.

From Figure 2.5, it seems that there is a decreasing failure rate with the installation decade (and year); this is more clear in AC material, which is represented by a wider installation period. For PVC and HDPE this tendency is not visible. For HDPE material, since most pipes are new (less than 10 years), the ageing is probably not yet noticeable. Polyvinyl chloride material is represented in four decades, so the ageing factor would be expected. One possible explanation could be that plastic materials deteriorate by mechanisms other than the affecting AC material.

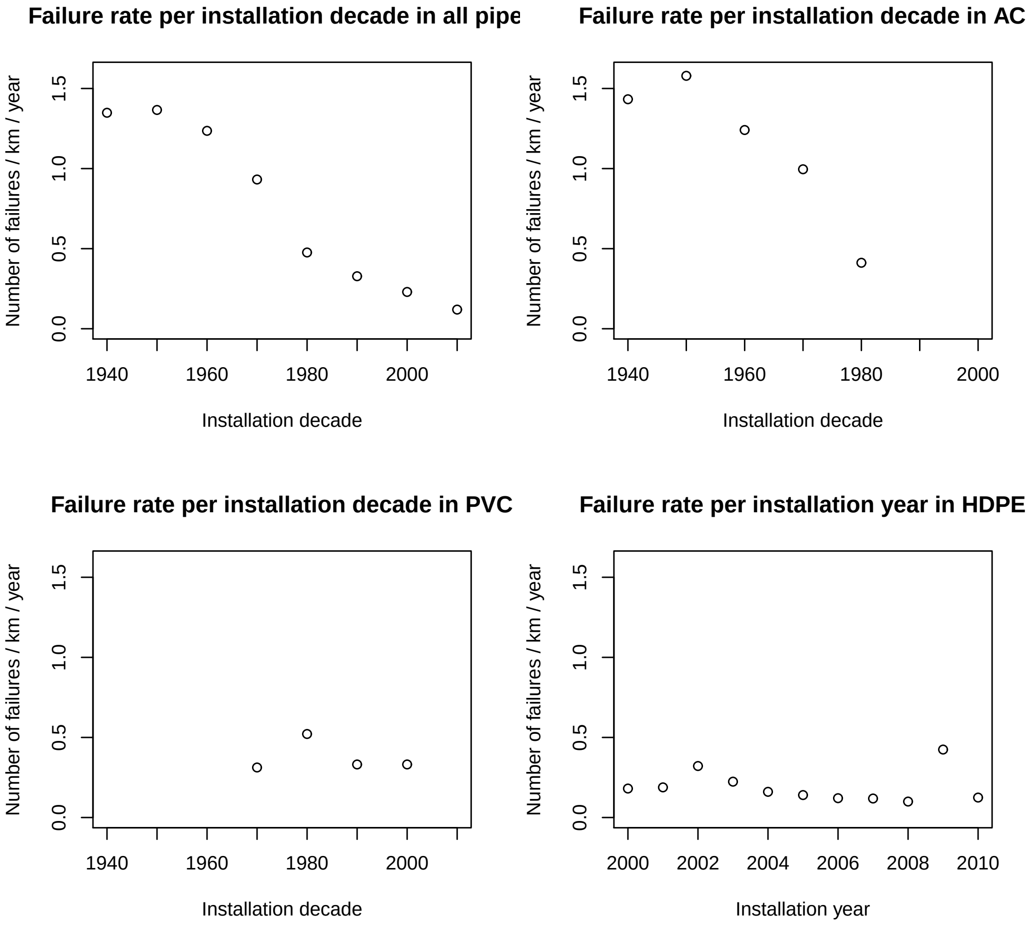

Figure 2.5: Failure rate per installation decade and years.

2.4.3 Pipe diameter

To evaluate the failure rate for different pipe diameters, it was necessary to split the failure data into different diameter classes. The daiameter variable was divided into seven classes of equal frequency. Failure rate was calculated for each class; results are presented in Table 2.5. The effect of diameter on failure rate is less visible than the effect of pipe installation decade (i.e. pipe age). Nevertheless, it appears, from Table 2.5, that failure rate is lower for larger pipe diameters.

Table 2.5: Failure rate per Diameter.

Diameter	Failure rate	Total Length
(mm)	(No. of failures/km/year)	(km)
[20,63]	1.14	46.4
(63,80]	1.05	33.1
(80,100]	0.69	60.0
[110]	0.23	104.0
(110,150]	0.70	32.9
(150,160]	0.36	36.9
(160,630]	0.23	53.9

2.4.4 Length

Empirical knowledge states that, when other variables are fixed, the number of failures is proportional to the length of pipes and failure rate is the coefficient of proportionality. That is, failure rate is independent of pipe length. The failure data was divided in seven classes such that the sum of the lengths of every pipe belonging to class C_k is the same for $k = 1, .., 7$.

From Table 2.6 it would appear that when pipes are longer failure rate tends to slightly decrease. This dependence is more noticeable in longer pipes classes, i.e. the failure rate in pipes that measure more than 123.6m is significantly smaller. The failure rate in medium-sized pipes is almost constant. If the failure data is to be divided into several categories with homogeneous failure rates, Table 2.6 suggests that the pipes could be categorised in three classes: small-sized pipes; medium-sized pipes; and long pipes.

Table 2.6: Failure rate per Length.

Length (m)	Failure rate (No. of failures/km/year)	Total length (km)
[0.1,28.9]	0.88	52.7
(28.9,50]	0.63	52.4
(50,69.2]	0.59	52.6
(69.2,91.6]	0.61	52.4
(91.6,123.6]	0.57	53.4
(123.6,192.3]	0.44	52.5
(192.3,904.8]	0.31	52.4

2.4.5 Previous failures

To study how previous failures influence the future failure rate, the failure data had to be divided in two sets. The first set has the failure records until the end of 2006 and the second has only the records of the failures that have occurred afterwards in the same pipes. Having divided the failure data, it is possible to compute the number of previous failures and the past individual failure rate (failures before 2007) for each pipe and group all pipes according to it. Then the mean future failure rate (failures from 2007 to 2011) can be computed for each group.

Results presented in Tables 2.7 and 2.8 suggest that there is a clear relationship between the occurrence of previous failures and the occurrence of future failures. This fact is even more noteworthy when comparing the past individual failure rates with the mean future failure rates (Table 2.8). The higher the past individual failure rate is, higher is the tendency to fail in the future. This evolution shows a considerable jump between the class of no failing pipes and the subsequent class. In fact, the failure rate in pipes without previous failures is 0.35, whereas failure rate in pipes with one or more failures is 1.19.

After this analysis it can be stated, for this specific case, that all variables have a direct effect on the failure rate. As such, when possible, these variables should be taken into account in the models considered in this study. The main attributes that influence the failure rate are suggested to be material, previous failures (previous failure rate) and age of pipes. However, since pipe age and pipe material are strongly dependent variables, the other studied pipe attributes can be as important as one of these two variables.

Table 2.7: Failure rate associated with number of previous failures.

Number of previous failures	Failure rate (No. of failures/km/year)	Total length (km)
0	0.35	281.8
1	0.89	34.0
2	1.31	13.3
3	2.48	4.1
4	0.70	2.7
5 to 7	3.02	2.6

Table 2.8: Failure rate associated with previous failures rate.

Previous failure rate (No. of failures/km/year)	Failure rate (No. of failures/km/year)	Total length (km)
0	0.35	281.8
(0,1.23]	0.66	11.4
(1.23,2.014]	0.81	11.4
(2.013,2.92]	0.98	11.4
(2.92,4.825]	1.28	11.4
(4.825,556]	2.24	11.2

Chapter 3

Single-variate model: Poisson process

The Poisson process is the only single-variate failure prediction model considered and implemented in this thesis. The water supply network pipes were divided into several groups, each associated with a specific failure probability distribution. Analysis conducted in Chapter 2 on how different variables influence failure rate was used further to define the grouping criteria.

In the first section of this chapter, the Poisson process is explained in detail. In the last section it is applied to the failure data. Further, in Chapter 6, a thorough comparison between this and the other implemented models is conducted.

3.1 Definition of the Poisson process

According to Ross (2006) a Poisson process is a counting process $\{N(t), t \geq 0\}$ with rate γ satisfying the following conditions:

$\forall \gamma \in \mathbb{R}^+$ and $\forall s, t, u, v \in \mathbb{R}^+$, such that $s < t < u < v$,

1. $N(0) = 0$.
2. Independent increments, i.e. $N(t) - N(s) \perp\!\!\!\perp N(v) - N(u)$.
3. $N(t) \sim Poisson(\gamma t)$.

A property of the Poisson process that can be derived from conditions 2 and 3 is the stationary increments, i.e. $N(t+s) - N(t) \sim N(u+s) - N(u) \sim N(s)$. Another property is that the expected number of events is proportional to the observation time, where γ is the coefficient of proportionality and defines the intensity of the process.

When analysing failure data in urban water systems, it is assumed that the number of events (i.e. failures) is also proportional to the length of pipes. If a pipe is twice as long, then the expected number of failures is twice as high. The rate of the Poisson process in some pipe i is $\gamma_i = \lambda l_i$, where l_i is the length of the pipe. Therefore, the failure rate per km in the overall system is represented usually by λ (number of failures / km / year), where λ is the proportional coefficient between the rate γ_i of the counting process $N_i(t)$ and the length of the respective pipe l_i.

To define the distribution of the number of failures in each pipe, it is necessary first to estimate λ, which can be obtained using the maximum likelihood method.

Considering a failure data $\mathbf{n} = \{n_i\}_{i=1,..,m}$, with n_i = number of observed failures in pipe i during the observation time t_i, the likelihood function (Equation 3.1) can be written using the Poisson probability function.

$$L(\lambda|\mathbf{n},\mathbf{t},\mathbf{l}) = \prod_{i=1}^{m} \frac{e^{\lambda l_i t_i}(\lambda l_i t_i)^{n_i}}{n_i!}, \tag{3.1}$$

where $\mathbf{t} = \{t_i\}_{i=1,..,m}$ and $\mathbf{l} = \{l_i\}_{i=1,..,m}$.

The solution of the likelihood maximisation problem (applying the logarithm to Equation 3.1 and maximising the resulting function) is:

$$\hat{\lambda} = \arg\max_{\lambda\in\Theta} L(\lambda|\mathbf{n},\mathbf{t},\mathbf{l}) = \frac{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} t_i l_i}. \tag{3.2}$$

If λ is estimated using the entire data set, then the failure rate will be the same for all pipes, no matter their properties. Nevertheless, the data set can be divided based on the pipe characteristics, such as material and diameter, creating different categories. To create these categories a preliminary analysis of data needs to be performed in order to define categories that gather pipes with similar failure rate. Once the pipe data is categorised, the failure rate λ_k can be estimated for each category C_k using Equation 3.2.

After the λ_k is estimated for each category C_k, probabilities of failures can be calculated for each pipe. The maximum likelihood estimator of the probability of a pipe i, belonging to C_k, to suffer n failures during a time period t is presented in Equation 3.3 (by the invariance property of maximum likelihood estimators).

$$P(\widehat{N_i(t) = n}) = \frac{e^{\hat{\lambda}_k l_i t}\left(\hat{\lambda}_k l_i t\right)^{n}}{n!}. \tag{3.3}$$

The estimator of the expected number of failures of pipe i, during the time period t, is presented in Equation 3.4. This expected value is used to predict the number of failures for each pipe during the Poisson process validation.

$$\widehat{E[N_i(t)]} = \hat{\lambda}_k l_i t. \tag{3.4}$$

3.2 Fitting of the Poisson single-variate model

To assess the quality of the predictions obtained using the Poisson process, the failure data were divided in two data sets. Failure distributions were estimated using the training set. Failure predictions were carried out using the test set. Finally, failure predictions were compared with the observations of the test set, in order to assess the quality of predictions using the Poisson process. Two methods of choosing the training and test samples were considered: temporal division and random division methods.

In the temporal division method the training set is composed of failures occurring before 01-01-2007 in all pipes installed before this date. The test sample comprises the failures that happened between 01-01-2007 and 31-03-2011 in the same pipes of the training sample. This method assess the ability to predict future failures in a water network with an organised failure history.

The training sample of the random division method consists of a random 50% sample of all pipes and all failures occurring in those pipes. Whereas the test sample is composed by the other 50% pipes and all associated failures. Both training and test time windows go from 01-01-2001 to 31-03-2011.

This method is important to evaluate the ability of a model, based on a training sample, to predict failures in a different data set. When failure history of a water supply network is too short (too few failures) to build predictions, a different data set from another similar water network can be used as a training sample. In order to have good prediction results, both water supply networks need to be very similar, sharing the same pipe characteristics.

The analysis of how the pipe variables influence the failure rate, conducted in Chapter 2, is taken into account to choose the variables to be used as grouping criteria.

The main variables to influence the failure rate were pipe material, pipe age and the previous failures variables. However, pipe age and pipe material are strongly dependent variables, the influence of the pipe ageing in each pipe material group was not clear; only in AC pipes the effect is noticeable (Figure 2.5).

Table 3.1: Continuous variables categories.

Variables	Class 1	Class 2	Class 3
Length (m)	$[0.1, 28.7]$	$[28.8, 188.9]$	$[189.6, 904.8]$
Installation years	$[1940, 1969]$	$[1970, 1979]$	$[1980, 2006]$
Diameter (mm)	$[20, 80]$	$(80, 100] \cup (110, 160]$	$\{110\} \cup [160, 630]$

The previous failures variables (number of previous failures and past individual failure rate) are very difficult to use as grouping criteria. Previous failures variables, unlike the other pipe attributes, depend on the failure history of the pipes. Therefore the use of this variable requires the splitting of the training sample in two sets: one to build the failure history of the pipes; and another to estimate how this failure history influences the failure rate, the same way it was done in Chapter 2. This method is rather complex and presents the major problem of splitting an already small failure data.

Thereby, the Poisson process is applied using pipe material, length and diameter variables. As seen in the previous Chapter 2, GS and CI do not present a significant sample, thus these two materials will be removed from analysis and the study will be based only in AC, DCI, HDPE and PVC pipes.

Over splitting the training data can lead to non-significant categories, which can lead to biased predictions. Therefore, it was decided to build only three different categories for each continuous variable. These categories were defined using the Ward's method of clustering.

Using the temporal division method of selecting the training and test samples, the continuous variable categories are defined as presented in Table 3.1.

Once defined the different categories, λ_k can be estimated for each category, using Equation 3.2. The resulting failure rate for each material, diameter and length categories are presented in Table 3.2.

It can be seen that the failure rates tend to decrease when the length or diameter classes increase. As expected, AC material presents the higher failure rate among the pipe material categories considered, whereas HDPE appears to be the one with the lower failure rates.

Splitting the failure data in categories has the problem of creating small-size categories. This may lead to non-significant failure rates. The total length of pipes in categories with a failure rate of 0 is, typically, very small, thus these failure rates are not significant. The NA value in Table 3.2 means that there are no pipes in that category, so no failure rate could be estimated.

Table 3.2: Failure rate per category.

Asbestos cement			
	Length 1	Length 2	Length 3
Diameter 1	1.70	1.37	1.12
Diameter 2	2.05	0.89	0.55
Diameter 3	0.40	0.33	0.26

Ductile cast iron			
	Length 1	Length 2	Length 3
Diameter 1	0	0	NA
Diameter 2	0.21	0.38	0
Diameter 3	0.93	0	0

HDPE			
	Length 1	Length 2	Length 3
Diameter 1	0.45	1.00	0
Diameter 2	0.42	0.52	0.22
Diameter 3	0.62	0.23	0.14

PVC			
	Length 1	Length 2	Length 3
Diameter 1	1.53	0.62	0.19
Diameter 2	0.98	0.45	0.35
Diameter 3	0.75	0.29	0.28

After estimating λ_k for each category, the distribution of the number of failures is fitted for each pipe. Then, the expected number of failures occurring in the test sample can be estimated using Equation 3.4. These predictions can be compared with the observed failures, in the following way: all pipes are sorted according to its expected number of failures (predicted failures); after sorting, 0.1-quantiles of the predictions are built; for each prediction quantile the mean of the actual observed failures is compared with the mean of the predicted failures.

This method, used in Le Gat and Eisenbeis (2000), allows to understand the ability to detect pipes with higher likelihood of failing. Figure 3.1 presents these results, in which dots represent the average of the number of observed failures in each 0.1-quantile and the red line represents the average of the predicted number of failures in each 0.1-quantile.

In Figure 3.1, the mean number of observed failures show a clear tendency to increase with the quantiles of predicted failures. The observed failures of the last quantile really stands out from the lower predicted failures quantiles. From these results, it appears that the Poisson process can detect the pipes more prone to fail. However, there is a clear overestimation in the last four quantiles.

This overestimation may have to do with the distinct failure rates of the training and the test samples. There were 0.59 failures/km/year in the training sample, whereas there were only 0.49 failures/km/year in the test sample. Since there is such a significant difference between both failure rates, it is only natural that the Poisson process (which assumes the same behaviour in both samples) tends to overestimate the number of failures.

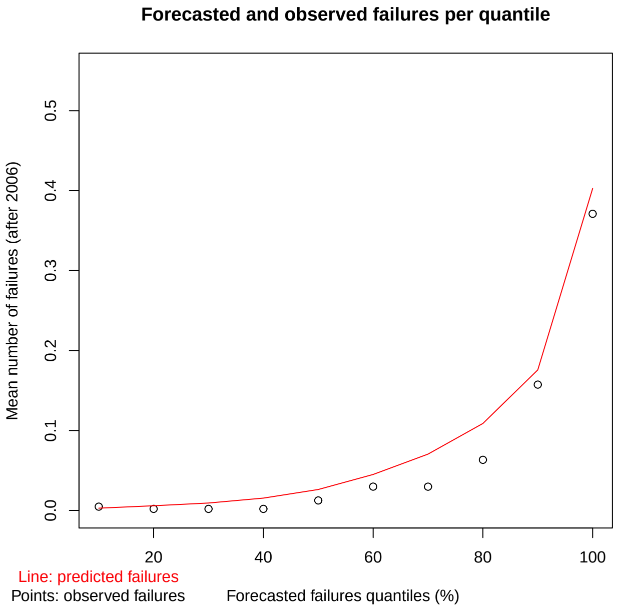

Figure 3.1: Poisson process predictions.

Another test was carried out using the random division method to build the training and test samples. The same categories used in the previous method (Table 3.2) were chosen as grouping criteria. A plot comparing the number of observed failures and the predicted number of failures for each prediction quantile (comparison method suggested by Le Gat and Eisenbeis (2000)) is presented in Figure3.2.

In Figure 3.2 it seems that, in addition to detect the pipes that are more likely to fail, the Poisson process presents an accurate prediction in almost every quantile. The reason that the Poisson process presents better results using this method(Figure3.2) rather than the time division method (Figure 3.1) is that in the random division method the training and test samples show similar behaviours. The failure rate in the test sample is of 0.53 failures/km/year and the failure rate in the training sample is of 0.49 failures/km/year.

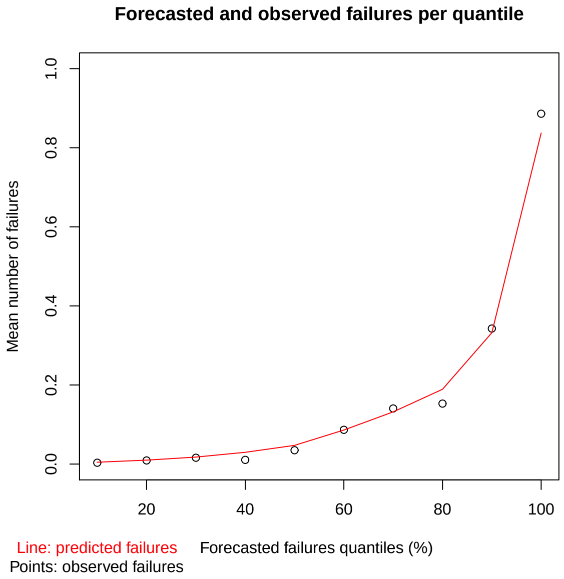

Figure 3.2: Poisson process predictions in random sample.

Chapter 4

Weibull accelerated lifetime model

The second failure prediction model to be fit is an accelerated lifetime model. This failure prediction model greatly differs from the Poisson process, as it models the time between failures rather than the number of failures. This multivariate model is described in detail in the next section. A failure prediction tool suggested by Le Gat and Eisenbeis (2000) is described, presenting some possible issues and suggesting some improvements. Finally, the Weibull accelerated lifetime model (WALM) is applied to the failure data studied in Chapter 2 and some prediction plots are presented in the last section.

4.1 Definition of the Weibull accelerated lifetime model

Accelerated lifetime models relate the logarithm of the time to failure with a linear combination of p covariates, $\boldsymbol{x} = [1\ x_1\ x_2\ ...\ x_p]$, and an error term, Z.

$$\ln T = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} + \sigma Z, \tag{4.1}$$

where $\boldsymbol{\beta} = [\beta_0\ \beta_1\ ...\ \beta_p]$ are unknown regression parameters and σ is a scale parameter.

Equation 4.1 shows that the distribution of the random variable Z defines the distribution family of T. In particular, if Z follows the standard Gumbel distribution, then T will be a Weibull random variable, as proved from Equation 4.2 to Equation 4.4.

$$P\{Z > z\} = e^{-e^{-z}}. \tag{4.2}$$

$$P\{T > t\} = P\{\ln T > \ln t\} = P\{Z > \frac{\ln t - \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\}. \tag{4.3}$$

$$P\{T > t\} = e^{-e^{\frac{\ln t - \boldsymbol{x}^\intercal \boldsymbol{\beta}}{\sigma}}} = e^{-\left(\frac{t}{e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}}}\right)^{\frac{1}{\sigma}}}. \tag{4.4}$$

From Equation 4.4, it can be seen that the random variable T follows a Weibull distribution, i.e. $T \sim \text{Weib}(\sigma^{-1}, e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}})$.

The Weibull distribution has the great advantage of having simple survival and hazard functions. This allows to easily understand the direct effect of the covariates in the hazard function. Moreover, this model is equivalent to a proportional hazard model as proved in Cox and Oakes (1984), since the covariates act multiplicatively in the hazard function (Equation 4.5).

$$h(t) = \frac{1}{\sigma} t^{\frac{1}{\sigma}} e^{-\frac{\boldsymbol{x}^\intercal \boldsymbol{\beta}}{\sigma}}. \tag{4.5}$$

4.2 Estimation of parameters

To fit the model (Equation 4.1), parameters σ and $\boldsymbol{\beta}$ should be estimated. These parameters were obtained maximising the appropriate likelihood function.

In this model, the sample was composed by the time between recorded failures t_i and the explanatory covariates $\boldsymbol{x}_i$, for each individual i. Nevertheless, when conducting survival analysis, the decision variable, time between failures, is, in many cases, right censored. In all pipes, the time that they survived without failing until the end of the observation period is a right censored time. It is called a right censored time, because it was not ended by an observed failure, hence it is only known that the time between failure is greater than the censored time.

Although the time between failures is not observed, discarding these censored times would lead to a biased survival results. Therefore these right censored times should enter in the likelihood function. However, instead of entering with the Weibull density function, right censored times will enter with the Weibull survival function, since the only available information is that the due pipe survived that right censored time.

In this study it is assumed that water supply networks are repairable systems, one pipe may present several times between failures, whether they are observed or censored. Each of these times will enter in the model independently. Whether they occurred in the same pipe or not, they only depend on the covariates of the respective pipe. Hence, using a sample of observed times, each associated with p covariates, $\{(t_i, \boldsymbol{x}_i)\}_{i=1,\ldots,n}$, and a sample of censored times, each associated with p covariates, $\{(c_j, \boldsymbol{y}_j)\}_{j=1,\ldots,m}$, the likelihood function can be expressed using the Weibull density and survival functions (Equation 4.6).

$$L(\sigma, \boldsymbol{\beta} | \mathbf{t}, \mathbf{c}, \boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{n} f(t_i | \sigma, \boldsymbol{\beta}, \boldsymbol{x}_i) \prod_{j=1}^{m} S(c_i | \sigma, \boldsymbol{\beta}, \boldsymbol{y}_j). \tag{4.6}$$

$$L(\sigma, \boldsymbol{\beta} | \mathbf{t}, \mathbf{c}, \boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{n} \frac{1}{\sigma e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} \left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} \right)^{\frac{1}{\sigma} - 1} e^{-\left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} \right)^{\frac{1}{\sigma}}} \prod_{j=1}^{m} e^{-\left(\frac{c_j}{e^{\boldsymbol{y}_j^\intercal \boldsymbol{\beta}}} \right)^{\frac{1}{\sigma}}}, \tag{4.7}$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}$ and $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}$ are the matrices of covariates.

Applying the logarithm, Equation 4.8 is obtained.

$$l(\sigma, \boldsymbol{\beta} | \mathbf{t}, \mathbf{c}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{n} \ln \frac{1}{\sigma e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} + \left(\frac{1}{\sigma} - 1 \right) (\ln t_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}) - \left(\frac{t_i}{e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}} \right)^{\frac{1}{\sigma}} - \sum_{j=1}^{m} \left(\frac{c_j}{e^{\boldsymbol{y}_j^\intercal \boldsymbol{\beta}}} \right)^{\frac{1}{\sigma}}. \tag{4.8}$$

Unlike the Poisson process, the maximum likelihood estimators can not be analytically expressed. Therefore they must be obtained through numerical maximisation.

In this work, the estimation of parameters was done using function *survreg* of the survival package of the statistical software R.

The distribution of the times between failures is estimated for each pipe, and so the probability of surviving during some time can also be estimated, according to the survival function expressed in Equation 4.4.

4.3 Prediction of the number of failures

It is very important to estimate the expected number of failures in a repairable system, which allows calculating the expected total cost of repairs in the repairable system during some time period. One of the biggest drawbacks of the Weibull distributions is that their convolution can not be analytically obtained. Thus, the distribution of the number of failures during some time can not be derived.

In order to predict the number of failures, Le Gat and Eisenbeis (2000) presented an algorithm based on Monte Carlo simulations, described in this section.

The concept behind Le Gat and Eisenbeis (2000) algorithm is to generate a large number of simulations in each pipe and, consequently, determine the mean number of failures obtained in all simulations for each pipe.

To build simulations over a pipe i, with covariates $\boldsymbol{x}_i$, it is necessary to generate times between failures. As seen before, the survival function for this pipe is given by Equation 4.4, that can be rewritten as Equation 4.9, where $e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}$ is replaced by η.

$$S(t) = e^{-\left(\frac{t}{\eta}\right)^{\frac{1}{\sigma}}}. \tag{4.9}$$

Solving the survival function $S(t)$ as a function of t, the expression to generate random times is obtained (Equation 4.10).

$$t = -\eta \left(\ln S\right)^{\sigma}. \tag{4.10}$$

The Monte Carlo simulations in pipe i will be built as follows. Successive times between failures are generated until their sum overlaps the prediction time window. Subsequently, the number of generated times is recorded, ignoring the last one, since it falls outside the prediction window. This experiment is repeated 1,000 times and finally the predicted number of failures will be the mean of all 1,000 simulated numbers of failures. This procedure is repeated for all pipes, obtaining a number of predicted (expected) failures for each one.

4.3.1 Improvements of the WALM prediction process

In this section, it is suggested an improvement of the failure prediction process presented by Le Gat and Eisenbeis (2000). Figure 4.1 shows the time line of a pipe from the beginning of the failure history until the end of the prediction window.

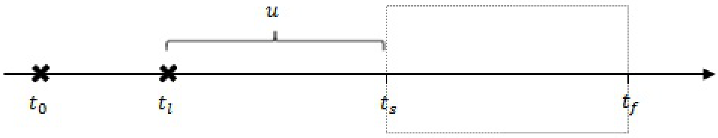

Figure 4.1: Time line and failure instants.

Let:

- t_0 is the instant where failure history begins for some pipe i;
- t_l is the time of the last recorded failure occurring in pipe i;
- t_s and t_f represent the start and finish instants of the prediction window, respectively;

- u is the time between t_l and t_s.

The approach presented in Le Gat and Eisenbeis (2000) ignores the elapsed time u between the last recorded failure and the beginning of the prediction window. Therefore, the first generated time to failure is counted from t_s, using Equation 4.10. However, since the times between failures follow a Weibull distribution, the failure counting process does not have the stationary increments property, that is $T|T > u$ and $T + u$ are not equally distributed. The improvement consists in generating a time to failure from t_l using the distribution of $T|T > u$ as described in Equations 4.11 and 4.12.

$$S(t,u) = P\{T > t|T > u\} = \frac{e^{-\left(\frac{t}{\eta}\right)^{\frac{1}{\sigma}}}}{e^{-\left(\frac{u}{\eta}\right)^{\frac{1}{\sigma}}}} = e^{-\left(t^{\frac{1}{\sigma}} - u^{\frac{1}{\sigma}}\right)\eta^{-\frac{1}{\sigma}}}. \tag{4.11}$$

$$t = \left(-\eta^{\frac{1}{\sigma}} \ln S + u^{\frac{1}{\sigma}}\right)^{\sigma}. \tag{4.12}$$

Equations 4.11 and 4.12 are used only to generate the first predicted time of failure, since u will be 0 when generating the following times to failure. When a pipe has no recorded failures it is assumed that the time of last failure is equal to the beginning of the pipe history, that is $t_l = t_0$ and $u = t_s - t_0$.

4.3.2 Dynamic variables and prediction of failures

The number of failures prediction process can become more complex if some of the covariates are dynamic, i.e. they can change during the process. One example of this type of covariate is the number of previous failures, which is considered one of the most important variables. This covariate should thus be updated whenever a new time to failure is generated. Therefore, $\boldsymbol{x}_i$ is not constant during the whole simulation, it is updated in every iteration of the failures prediction process.

Issues using the covariate number of previous failures

When applying the WALM regression, one of the most important covariates that was found was the number of previous failures. However, a new problem arose during the failure prediction process, related to this covariate. The Weibull accelerated lifetime model assumes that the covariates influence exponentially the hazard function. The expected time to failure of some pipe is given by Equation 4.13.

$$\Gamma(1+\sigma)e^{\boldsymbol{x}^{\intercal}\boldsymbol{\beta}} = \Gamma(1+\sigma)e^{\boldsymbol{x}^{*\intercal}\boldsymbol{\beta}^{*}}e^{x_{nopf}\beta_{nopf}}, \tag{4.13}$$

where x_{nopf} represents the number of previous failures covariate; β_{nopf} is the coefficient associated to x_{nopf}; $\boldsymbol{x}^{*}$ is the covariates vector without x_{nopf}; and $\boldsymbol{\beta}^{*}$ is the coefficients vector without β_{nopf}.

When applying the Monte Carlo simulation to predict the number of failures of a pipe, in each iteration the expected value of the next time to failure drops exponentially, since x_{nopf} increases (Equation 4.14).

$$\frac{E\left[T|x_{nopf}+1\right]}{E\left[T|x_{nopf}\right]} = \frac{\Gamma(1+\sigma)e^{\boldsymbol{x}^{*\intercal}\boldsymbol{\beta}^{*}}e^{(x_{nopf}+1)\beta_{nopf}}}{\Gamma(1+\sigma)e^{\boldsymbol{x}^{*\intercal}\boldsymbol{\beta}^{*}}e^{x_{nopf}\beta_{nopf}}} = e^{\beta_{nopf}}. \tag{4.14}$$

Therefore, the expected sum of times to failure is given by Equation 4.15.

$$\begin{aligned} E\left[\textstyle\sum_{k=0}^{\infty} T_k\right] &= \sum_{k=0}^{\infty} E\left[T|x_{nopf}=k\right] = \sum_{k=0}^{\infty} E\left[T|x_{nopf}=0\right]\left(e^{\beta_{nopf}}\right)^{k} \\ &= E\left[T|x_{nopf}=0\right]\sum_{k=0}^{\infty}\left(e^{\beta_{nopf}}\right)^{k}. \end{aligned} \tag{4.15}$$

Since the failure rate should increase with the number of previous failures, it is expected that $\beta_{nopb} < 0$, hence $e^{\beta_{nopb}} < 1$. And so, the expected sum of the times between failures is a sum of a geometric progression of ratio $r = e^{\beta_{nopf}} < 1$, which means that the series is convergent to $\frac{1}{1-r}$. In this case, if $E\left[T|x_{nopf}=0\right]$ and β_{nopf} are sufficiently small, there is no guaranty that the sum of the simulated times between failures overlaps the prediction window. So, the Monte Carlo simulation can enter a never ending cycle.

Issues using pipe age covariate

Other covariates which are continuously changing with t may increase the method complexity; an example of this covariate is pipe age.

Although the pipe age variable is one of the most important explanatory variables, its introduction in the model increases the complexity of the failure prediction process. Unlike the number of previous failures covariate, which only needs to be incremented after each failure time is generated, the age of the pipe variable is continuously increasing with T. So, T can not be generated with a fixed distribution, since this distribution will continuously change throughout the duration of T.

One way to avoid this issue is to use a fixed covariate that translates the age of the pipe variable, e.g. the installation year or decade. However, as good and simple this solution may seem, the use of a fixed covariate to translate a time dependent covariate may not be realistic. By using the installation year as a covariate, the predicted number of failures from 2006 to 2011 will be the same as the predicted number of failures from 2020 to 2025.

Pipe age at last failure approach. The pipe age at last failure approach considers the age of the pipe at the last recorded failure as a covariate. This way, the covariate only needs to be updated at every iteration of the failures prediction process, as the number of previous failures covariate. One disadvantage of this approach is the fact that the pipe age covariate is updated only with the occurrence of a failure. This means that the age of the pipe will not act on the failure distribution as long as the pipe does not fail.

New approaches to use dynamic variables

In order to deal with the issues regarding the number of previous failures and the pipe age covariates, three new approaches are suggested:

Logarithm of the number of previous failures. This approach is to apply the logarithm to the number of previous failures. Instead of considering x_{nopf} as covariate, it is considered $\ln(1 + x_{nopf})$. With this new covariate, in each iteration of the failure prediction process the expected time to the next failure no longer drops exponentially (Equation 4.16).

$$\frac{E\left[T|x_{nopf}+1\right]}{E\left[T|x_{nopf}\right]} = \frac{e^{\beta_{nopf}\ln\left(x_{nopf}+1+1\right)}}{e^{\beta_{nopf}\ln\left(x_{nopf}+1\right)}} = \left(\frac{x_{nopf}+2}{x_{nopf}+1}\right)^{\beta_{nopf}}. \tag{4.16}$$

The expected time to failure is expressed in Equation 4.17.

$$E\left[T|x_{nopf}=k\right] = E\left[T|x_{nopf}=0\right]\left(\prod_{i=1}^{k}\frac{i+2}{i+1}\right)^{\beta_{nopf}} = E\left[T|x_{nopf}=0\right]\left(\frac{k+2}{2}\right)^{\beta_{nopf}}. \tag{4.17}$$

Assuming that $\beta_{nopf} < 0$, then the sum of expected times between failures is presented in Equation 4.18.

$$\sum_{k=0}^{\infty} E\left[T|x_{nopf}=k\right] = E\left[T|x_{nopf}=0\right]\sum_{k=0}^{\infty}\left(\frac{2}{k+2}\right)^{|\beta_{nopf}|}. \tag{4.18}$$

Equation 4.18 is a divergent series if $|\beta_{nopf}| \leq 1$. So only if $\beta_{nopf} \geq -1$ it can be guaranteed that the Monte Carlo simulation will halt.

Finite covariate instead of discrete number of previous failures. A simpler solution of the number of previous failures covariate issue, is to use a finite valued covariate. For instance, a binary covariate pf, where $pf = 1$ if the pipe has failed in the past and $pf = 0$ otherwise. The binary covariate may not give the same information that the discrete number of previous failures, nevertheless, it guarantees that the Monte Carlo simulations halt, whatever the estimated $\boldsymbol{\beta}_{nopf}$.

Age classes approach. A new approach is presented in order to deal with the pipe age variable. The aim of this approach is to allow the ageing effect on the failure distribution, independently of the occurrence of previous failures.

The age class approach categorises the pipe age variable into different classes, taking into account how this variable influences the failure rate. No many classes should be considered, in order to keep the model's simplicity, e.g. three classes. Subsequently, for each pipe i, the prediction window is divided in three subwindows, such that pipe i will belong to the same age class in each subwindow. The time elapsed from the previous failure to the instant where pipe i changes from age class k to class $k+1$ is denoted by s_k.

The approach assumes that each age class presents a different distribution of the time between failures. The time between failures in age class k is represented by T_k. In order to estimate the distribution of T_k, the distribution parameters are obtained using the maximum likelihood estimation (Equation 4.8). However, in this approach, the failure times entering the log-likelihood function will only be the times elapsed in age class k; this means that if some pipe starts the observation in age class k but only fails in class $k+1$, then the time elapsed in class k will enter in the regression of T_k as a right censored time and the time elapsed in class $k+1$ will enter in the regression of T_{k+1} as an observed time.

The failure prediction process will behave as before in each class. However, when a pipe is in class k and the next generated time t (using the distribution of T_k) exceeds s_k then t is discarded and a new time will be generated using the distribution of T_{k+1} from s_k.

This approach presents a major disadvantage: when dividing the failure data into different classes it becomes difficult to understand the age effect on the failure rate. Recent pipe materials are only represented in the first age classes. This approach makes impossible to predict the number of failures in these pipes when they belong to older classes. For example, it may be impossible to predict failures in HDPE pipes for the decade 2020, because HDPE pipes during that time will already belong to older age classes. However, if the pipe age variable could enter the model as a numeric covariate, then, even if the behaviour of HDPE is unknown at the age of 20 years, a prediction could be computed based on the ageing of HDPE in the early stages.

4.4 Fitting of the Weibull accelerated lifetime model

In order to assess the quality of predictions of the Weibull accelerated lifetime model, this model was applied to the available failure data. As done in Chapter 3, the distribution parameters are estimated using the training sample and the failure predictions are conducted in the test sample.

The two failure data division methods used in Chapter 3 were used to define the training and test samples: the temporal division and the random division methods.

The covariates used in all WALM predictions were: ln length, diameter and previous failures (binary variable). The logarithm is applied to the length variable, because it is believed that the hazard rate should be proportional to the length of pipe. Entering with ln length the hazard rate will be: $h(t) = h^*(t) \ (x_{length})^{-\beta_{length}}$. So, if β_{length} is close to -1 the hazard rate will be proportional to the length of pipe.

The pipe material and the pipe age variables entered the model differently according to the WALM approaches. Two different approaches were tested: the pipe age at last failure approach and the age class approach. In the pipe age at last failure approach, the pipe material was used as grouping criteria and the pipe age at the last failure was used as a numeric covariate. In the age class approach the pipe age was used as grouping criteria, as explained in the according section, whereas pipe material entered the model as a covariate.

The first tests were carried out using the temporal division method. So the training sample is composed of the failures occurring before 01-01-2007 and the test sample is composed of the failures occurring after this date.

Table 4.1 shows the estimated parameters using the pipe age at last failure approach, for each pipe material. In Table 4.1, S.E. represents the standard error of each estimated parameter. The p-value, for each parameter, is obtained using the Wald test, in which the null hypothesis is zero. The scale parameter, σ, is constrained to being positive and its proposed value is one, thus the Wald test is carried out using $\ln \sigma$ with null hypothesis zero. In order to evaluate the model significance, a likelihood ratio test is conducted, where LR represents the ratio of the maximum likelihood value using only the parameters β_0 and σ and the maximum likelihood value using all regression parameters.

The most significant covariates seem to be the previous failures indicator and the logarithm of length, whereas the less significant variable is pipe diameter. To be noticed that the coefficient values associated with ln length are usually close to -1, thus the hazard function is almost proportional to the length of pipes. Coefficients of DCI pipes have higher p-values and the likelihood ratio test presents a higher p-value than other materials. This is probably due to the fact that DCI pipes compose a small failure data. The fitness of the model seems higher in AC pipes, where all variables are statistically significant, probably because this material is the one that presented more failures.

Table 4.1: Regression estimates and p-values of Wald tests

	HDPE				**PVC**			
Parameters	Estimate	S.E.	$^{\hat{\theta}}/_{\text{S.E.}}$	p-value	Estimate	S.E.	$^{\hat{\theta}}/_{\text{S.E.}}$	p-value
β_0	17.808	1.004	17.73	<0.0001	15.187	0.574	26.44	<0.0001
Age at last failure (β_{age})	-0.046	0.021	-2.23	0.026	-0.026	0.009	-2.92	0.004
Diameter (β_{diam})	3.1e-03	0.003	0.99	0.322	0.001	0.003	0.27	0.785
ln Length ($\beta_{lnlength}$)	-1.225	0.167	-7.33	<0.0001	-1.002	0.088	-11.33	<0.0001
Previous failures (β_{pf})	-4.052	0.465	-8.72	<0.0001	-1.714	0.222	-7.71	<0.0001
$\ln \sigma$	0.688	0.081	8.53	<0.0001	0.446	0.054	8.29	<0.0001
σ	1.99				1.56			
$-2\ln(LR)$	210				404			
p-value	0.0001				<0.0001			

	AC				**DCI**			
Parameters	Estimate	S.E.	$^{\hat{\theta}}/_{\text{S.E.}}$	p-value	Estimate	S.E.	$^{\hat{\theta}}/_{\text{S.E.}}$	p-value
β_0	13.337	0.368	36.28	<0.0001	15.557	2.839	5.48	<0.0001
Age at last failure (β_{age})	-0.020	0.005	-4.07	<0.0001	0.082	0.097	0.85	0.394
Diameter (β_{diam})	0.007	0.001	5.79	<0.0001	0.001	0.005	0.19	0.846
ln Length ($\beta_{lnlength}$)	-0.871	0.060	-14.53	<0.0001	-0.768	0.436	-1.76	0.078
Previous failures (β_{pf})	-1.533	0.131	-11.71	<0.0001	-4.836	1.687	-2.87	0.004
$\ln \sigma$	0.441	0.035	12.73	<0.0001	0.614	0.279	2.20	0.028
σ	1.55				1.85			
$-2\ln(LR)$	759				15.7			
p-value	<0.0001				0.0035			

The comparison method, used in Le Gat and Eisenbeis (2000) and in Chapter 3, to compare predictions with the observed records is used; results for each pipe material are presented in Figure 4.2. Predictions are very accurate for all materials except for DCI pipes. In DCI pipes the model can not detect effectively pipes with higher likelihood of failing, this is probably due to the short failure records of this pipe material.

In Figure 4.3 all pipe materials are assembled and predictions are grouped. Comparing both Figures 4.3 and 3.1, WALM seems to be a more accurate model than the Poisson process,

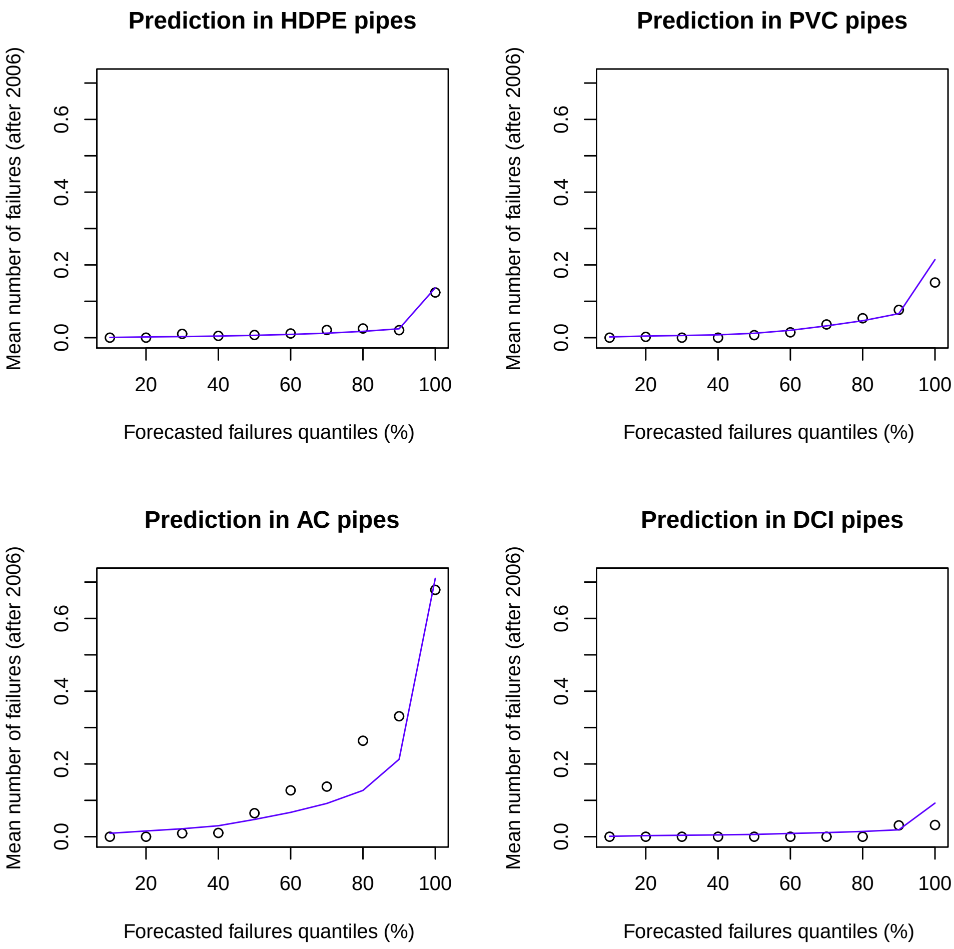

Figure 4.2: WALM predictions in each material category.

nevertheless a more thorough comparison of all models is conducted in Chapter 6.

The age class approach was also tested using the same training and test samples. The prediction results are plotted in Figure 4.4. Results are very similar, although there is a small overestimation in pipes of the last quantile. Since this approach is significantly more complex and results do not seem better, the age classes approach will be dropped in further analysis of the WALM.

The third test was conducted using the random division method of the failure data, that is: the training sample is composed of a 50% random sample of the water supply system pipes; and the test sample is composed of the other 50% of the water supply system. In this random division method only the pipe age at last failure approach was tested. The estimated coefficients were very similar to the ones presented in Table 4.1 and thus they are not presented. The plot of the mean observed failures for each prediction 0.1-quantile is

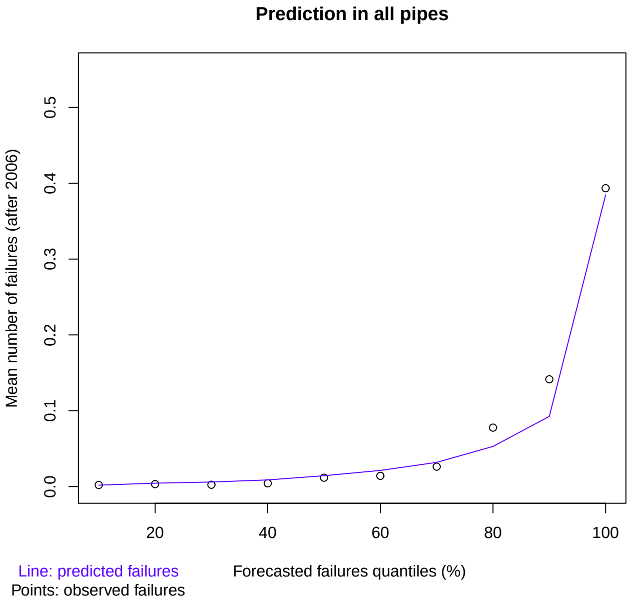

Figure 4.3: WALM predictions in all pipes.

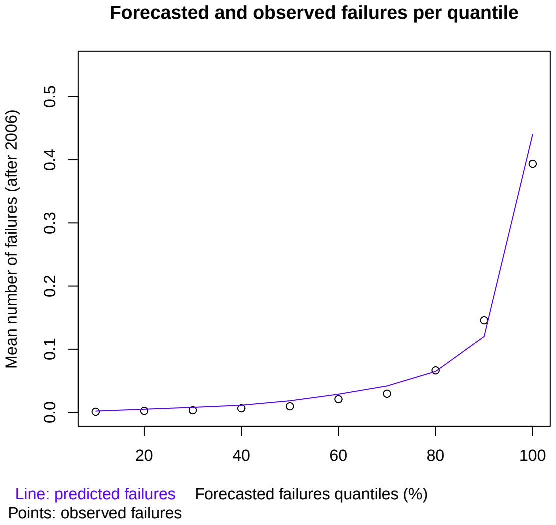

Figure 4.4: WALM predictions using age classes.

presented in Figure 4.5. The results presented in Figure 4.5 are very accurate, with a slight underestimation in the last prediction quantile.

In all WALM prediction plots (Figures 4.3, 4.4 and 4.5) the mean number of observed failures of the last quantile always stands out from the previous quantiles. This indicates that this

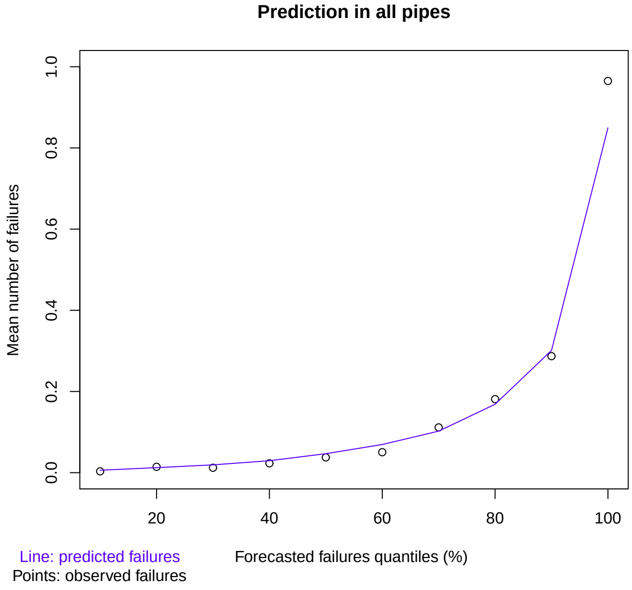

Figure 4.5: WALM predictions in random sample of pipes.

failure prediction model detects efficiently pipes with higher risk of failing.

Chapter 5

Linear extended Yule process

The third model fitted in this study is the linear extended Yule process presented in Le Gat (2009). This is a multivariate counting process, where the process rate is a linear function of the number of past failures, depending on the age of pipes and influenced by covariates. The logarithm of the likelihood function as presented in Le Gat (2009) is described and some simplifications are suggested. Results of the fitting of the linear extended Yule process are plotted in the last section.

5.1 Definition of the linear extended Yule process

The Linear Extended Yule Process (LEYP) is based on the pure birth process. Let $N(t)$ be the number of individuals in a Yule process with a single initial progenitor and birth intensity $\lambda > 0$. The pure birth process or classical Yule or Yule-Furry process is defined by:

$\forall_{t\in\mathbb{R}^+}\forall_{n\in\mathbb{N}}\ ,\ \lambda \in \mathbb{R}^+$

$$\begin{cases} N(0) = 1 \\ P\{N(t+dt) - N(t) = 1 | N(t) = n\} = n\lambda dt \end{cases} \tag{5.1}$$

This process presents several interesting properties:

(i) it has the Markov property, i.e. the distribution of future individuals depends only on the number of individuals at the present time;

(ii) there are at most one occurrence at some time instant;

(iii) the distribution of the number of individuals follow a Geometric distribution.

The last property (iii) can be derived from the solution of the following differential equations:

$$dP\{N(t)=n\}=(-n\lambda P\{N(t)=n\}+(n-1)\lambda P\{N(t)=n-1\})\,dt. \tag{5.2}$$

Using the initial condition $P\{N(0)=1\}=1$, the solution of Equation 5.2 is given by Equation 5.3.

$$P\{N(t)=n\}=e^{-\lambda t}(1-e^{-\lambda t})^{n-1}. \tag{5.3}$$

For a general n_0, the counting process with initial condition $N(0)=n_0$ can be thought as the sum of n_0 simple processes. Thus, the distribution of the number of individuals will be the sum of n_0 independent Geometric distributions, which is equivalent to a Negative Binomial $(n_0, e^{-\lambda t})$, with probability function presented in Equation 5.4.

$$P\{N(t)=n\}=\binom{n-1}{n-n_0}e^{-\lambda n_0 t}(1-e^{-\lambda t})^{n-n_0},\quad n=n_0,n_0+1,... \tag{5.4}$$

In order to adapt this pure birth process to failure data, several changes were applied by Le Gat (2009):

- $N(t)$ represents the number of failures occurring in a pipe, in the time interval $[0,t]$.
- The process starts at 0, since the number of failures right after the installation of the pipe is 0, i.e. $N(0)=0$.
- The rate of the process is free to vary with time, since it is common knowledge that time influences failure rate in water systems, that is, $\lambda=\lambda(t)$.
- The intensity of the counting process is no longer forced to be proportional to the number of previous failures.

With these three new changes, is defined the Non-Homogeneous Birth Process (NHBP), with rate $\lambda(t)\geq 0$,

$$\forall_{t\in\mathbb{R}^+}\forall_{n\in\mathbb{N}}\ ,\ \alpha_n\in\mathbb{R}^+\ ,\ \lambda(t)\in\mathbb{R}^+$$

$$\begin{cases} N(0)=0 \\ P\{N(t+dt)-N(t)=1|N(t)=n\}=\alpha_n\lambda(t)dt \end{cases} \tag{5.5}$$

The probability function of the counting process is the solution of the differential Equation 5.6, using $P_n(s) = P\{N(t+s) - N(t) = n|N(t) = j\}$.

$$dP_n(s) = (1 - \alpha_{j+n}\lambda(t+s))P_n(s)ds + \alpha_{j+n-1}\lambda(t)P_{n-1}(t+s)ds; \quad P(N(0) = 0) = 1. \tag{5.6}$$

Solving Equation 5.6 and by induction on n, the probability function of $N(t+s) - N(t)|N(t) = j$ is presented in Le Gat (2009) and is given by Equation 5.7.

$$P\{N(t+s) - N(t) = n|N(t) = j\} = \left(\prod_{k=0}^{n-1} \alpha_{j+k}\right) \sum_{k=0}^{n} \frac{e^{-\alpha_{j+k}(\Lambda(t+s)-\Lambda(t))}}{\prod_{l=0,l\neq k}^{n}(\alpha_{j+l} - \alpha_{j+k})}. \tag{5.7}$$

The LEYP is a particular case of the NHBP, in which the function $\alpha_j = \alpha(j) = (1 + \alpha j)$, with $\alpha \in \mathbb{R}^+$.

It is proven by Le Gat (2009) that the probability function of the LEYP is described by Equation 5.8.

$$P\{N(t) - N(s) = n|N(b) - N(a) = j\} =$$
$$= \frac{\Gamma(\alpha^{-1} + j + n)}{\Gamma(\alpha^{-1} + j)n!}\left(\frac{\mu(b) - \mu(a) + 1}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right)^{\alpha^{-1}+j}\left(\frac{\mu(t) - \mu(s)}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right)^{n}, \tag{5.8}$$

where $\mu(t) = e^{\alpha\Lambda(t)}$ and $\Lambda(t) = \int_0^t \lambda(u)du$.

Equation 5.8 implies that the distribution of the number of failures of a LEYP is a continuous extended Negative Binomial, that is,

$$\{N(t) - N(s)|N(b) - N(a) = j\} \sim NB\left(\alpha^{-1} + j, \frac{\mu(b) - \mu(a) + 1}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right).$$

One important feature of a failure prediction model is to allow distinguishing the probability of failure in different pipes, using their different attributes. Thus the intensity function can be based on pipe covariates. Le Gat (2009) suggests an intensity function built upon the Weibull power law and the Cox factor,

$$\lambda(t) = \delta t^{\delta-1} e^{\boldsymbol{x}^\intercal \boldsymbol{\beta}}. \tag{5.9}$$

For the intensity function described in Equation 5.9,

$$\Lambda(t) = t^{\delta} e^{\boldsymbol{x}^{\intercal}\boldsymbol{\beta}}, \tag{5.10}$$

$$\mu(t) = e^{\alpha t^{\delta} e^{\boldsymbol{x}^{\intercal}\boldsymbol{\beta}}}. \tag{5.11}$$

5.2 Estimation of parameters

The LEYP parameters were estimated through the maximum likelihood method. Le Gat (2009) builds the likelihood function and presents the log-likelihood function for the overall system, Equation 5.12.

$$\begin{aligned}\ln L(\alpha, \delta, \boldsymbol{\beta} | \boldsymbol{T}, \boldsymbol{X}, \mathbf{n}, \mathbf{a}, \mathbf{b}) = &\sum_{i=1}^{m} \Bigg(n_i \ln \alpha + \ln \Gamma\left(\alpha^{-1} + n_i\right) - \ln \Gamma\left(\alpha^{-1}\right) \\ &- \left(\alpha^{-1} + n_i\right) \ln\left(\mu(b_i) - \mu(a_i) + 1\right) + \sum_{j=1}^{n_i} \ln \lambda(t_{ij}) + \alpha \Lambda(t_{ij}) \Bigg),\end{aligned} \tag{5.12}$$

where:
m is the number of pipes;
$\mathbf{n} = [n_1 \ldots n_m]$, with n_i representing the number of failures in pipe i;
$\mathbf{a} = [a_1 \ldots a_m]$, with a_i representing the age of pipe i at the beginning of observations;
$\mathbf{b} = [b_1 \ldots b_m]$, with b_i representing the age of pipe i at the end of observations;
$\boldsymbol{X} = [\boldsymbol{x}_1 \ldots \boldsymbol{x}_m]$, with $\boldsymbol{x}_i$ representing the covariate vector of pipe i;
$\mathbf{T} = [\mathbf{t}_1 \ldots \mathbf{t}_m]$ with t_{ij} representing the age of pipe i at the j^{th} failure.

The estimated parameters were obtained through the numerical maximisation of Equation 5.12. However, the log-likelihood function presents some computational problems. In Equation 5.11, the function $\mu(t)$ can easily reach very high values; for common values of b_i and a_i, $\mu(b_i)$ and $\mu(a_i)$ are considered infinity, due to exceeding the machine precision, and thus $\ln(\mu(b_i) - \mu(a_i) + 1)$ can not be computed.

In order to solve this problem a simplification of the log-likelihood function is suggested. Let $h(a_i, b_i, \boldsymbol{x}_i) = \ln(\mu(b_i) - \mu(a_i) + 1)$. $h(a_i, b_i, \boldsymbol{x}_i)$ can be simplified as follows:

$$\begin{aligned}
h(a_i, b_i, \boldsymbol{x}_i) &= \ln\left(\mu(b_i) - \mu(a_i) + 1\right) \\
&= \ln\left(e^{\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}} - e^{\alpha {a_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}} + 1\right) \\
&= \ln\left(e^{\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}}\left(1 - e^{\alpha {a_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}} - \alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}} + e^{-\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}}\right)\right) \\
&= \alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}} + \ln\left(1 - e^{\alpha e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}\left(a_i^{\delta} - b_i^{\delta}\right)} + e^{-\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}}\right). \qquad (5.13)
\end{aligned}$$

Equation 5.13 is an important simplification: when b_i assumes high values, $e^{\alpha e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}\left(a_i^{\delta} - b_i^{\delta}\right)}$ and $e^{-\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}}$ will be close to zero and, thus, $h(a_i, b_i, \boldsymbol{x}_i)$ is computed as $\alpha {b_i}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}}$.

Therefore the final expression of the log-likelihood function to be maximised is given by Equation 5.14.

$$\begin{aligned}
\ln L(\alpha, \delta, \boldsymbol{\beta} | \boldsymbol{T}, \boldsymbol{X}, \mathbf{n}, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{m} \Bigg(& n_i \ln \alpha + g(\alpha, n_i) - \left(\alpha^{-1} + n_i\right) h(a_i, b_i) \\
& + n_i \ln \delta + n_i \boldsymbol{x}_i^{\intercal}\boldsymbol{\beta} + (\delta - 1) \sum_{j=1}^{n_i} \ln t_{ij} + \sum_{j=1}^{n_i} \alpha t_{ij}^{\delta} e^{\boldsymbol{x}_i^{\intercal}\boldsymbol{\beta}} \Bigg), \qquad (5.14)
\end{aligned}$$

where $h(a_i, b_i, \boldsymbol{x}_i)$ is given by Equation 5.13 and $g(\alpha, n_i)$ is given by Equation 5.15.

$$g(\alpha, n_i) = \ln \Gamma\left(\alpha^{-1} + n_i\right) - \ln \Gamma\left(\alpha^{-1}\right) = \sum_{k=0}^{n_i - 1} \ln\left(\alpha^{-1} + k\right). \qquad (5.15)$$

With the parameters estimated, the probabilities of failure can be computed. For instance, the probability of some pipe to fail during $[s, t]$, knowing its failures during $[a, b]$, is given by Equation 5.16.

$$\begin{aligned}
P\{N(t) - N(s) > 0 | N(b) - N(a) = j\} &= 1 - P\{N(t) - N(s) = 0 | N(b) - N(a) = j\} \\
&= 1 - \left(\frac{\mu(b) - \mu(a) + 1}{\mu(t) - \mu(s) + \mu(b) - \mu(a) + 1}\right)^{\alpha^{-1} + j}. \qquad (5.16)
\end{aligned}$$

The expected value can also be computed in order to compare with the results obtained with the WALM and the Poisson process:

$$E[N(t)-N(s)|N(b)-N(a)=j]=\left(\alpha^{-1}+j\right)\frac{\mu(t)-\mu(s)}{\mu(b)-\mu(a)+1}. \tag{5.17}$$

5.2.1 Test of significance of the estimated parameters

A statistical hypothesis test was conducted to establish the significance of each estimated parameter of LEYP. The test was based on the likelihood ratio test, using the null hypothesis of each parameter: $\alpha = 0$; $\delta = 1$; $\beta_j = 0, \forall_{j=0,..,p}$.

For every parameter, except α, the log-likelihood function of the null hypothesis was obtained directly from Equation 5.14, replacing the value of the parameter by the null hypothesis. To test the significance of α the null hypothesis leads to a Non-homogeneous Poisson process (NHPP), with intensity $\lambda(t)$, and a different likelihood function.

The likelihood function of the NHPP for each pipe is:

$$\begin{aligned}&L(\alpha,\delta,\boldsymbol{\beta}|\mathbf{t},\boldsymbol{x},n,a,b)=\\&\prod_{t\in[a,b]}\left(P\left\{N(t+dt)-N(t)=1\right\}\right)^{\Delta N(t)}\left(1-P\left\{N(t+dt)-N(t)=1\right\}\right)^{1-\Delta N(t)},\end{aligned} \tag{5.18}$$

where:
$\Delta N(t) = N(t) - N(t^-)$;
$P\left\{N(t+dt)-N(t)=1\right\} = \lambda(t)dt$;
$\prod_{t\in[a,b]}$ defines the product integral in $[a,b]$.

$$L(\alpha,\delta,\boldsymbol{\beta}|\mathbf{t},\boldsymbol{x},n,a,b)=\prod_{j=1}^{n}\lambda(t_j)\prod_{j=0}^{n}\prod_{t\in[t_j,t_{j+1}]}(1-\lambda(t)dt), \tag{5.19}$$

with $t_0 = a$ and $t_{n+1} = b$.

Using the product integral property $\prod_{t\in[a,b]}(1-dA(t)) = e^{-\int_a^b dA(t)}$, Equation 5.20 is obtained.

$$\begin{aligned}L(\alpha,\delta,\boldsymbol{\beta}|\mathbf{t},\boldsymbol{x},n,a,b)=&\prod_{j=1}^{n}\lambda(t_j)\prod_{j=0}^{n}e^{-\int_{t_j}^{t_{j+1}}\lambda(t)dt}\\=&\left(\prod_{j=1}^{n}\lambda(t_j)\right)e^{-\int_a^b\lambda(t)dt}\\=&e^{\Lambda(a)-\Lambda(b)}\prod_{j=1}^{n}\lambda(t_j).\end{aligned} \tag{5.20}$$

The log-likelihood function for the NHPP, using the failure records of all pipes, is presented in Equation 5.21.

$$\ln L(\alpha, \delta, \boldsymbol{\beta}|\mathbf{T}, \boldsymbol{X}, \mathbf{n}, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{m} \left(\Lambda(a_i) - \Lambda(b_i) + \sum_{j=1}^{n_i} \ln \lambda(t_{ij}) \right). \tag{5.21}$$

Once defined the likelihood function for the null hypothesis for each estimated parameter, the likelihood ratio, LR, can be computed.

$$LR = \ln \frac{\sup\{L(\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta_0\}}{\sup\{L(\boldsymbol{\theta})|\boldsymbol{\theta} \in \Theta\}}. \tag{5.22}$$

LR represents the ratio between the values of the maximum likelihood of the null model and the complete model. Wilks proved that $-2\ln LR$ is asymptotically distributed to a χ^2_k, where k is the difference between the dimensionality of Θ and Θ_0. In this case, the p-value, for each parameter, is obtained as $P\{\chi^2_1 > -2\ln LR\}$.

5.3 Fitting of the linear extended Yule process

The first test was carried out defining the training and test data sets as a temporal division of the failure data. The training set includes failures before 2007 and the test set includes all failures from 01-01-2007 to 31-03-2011. Parameters of the LEYP were estimated using a quasi-Newton algorithm to maximise Equation 5.14, using the training data set. As for predictions, they were built using Equation 5.17, where:

a is the age of pipe at 01-01-2001, or 0 if installed after 2001;
b is the age of pipe at 01-01-2007;
s is the age of pipe at 01-01-2007 ($s = b$);
t is the age of pipe at 31-03-2011;
j is the number of observed failures between 01-01-2001 and 01-01-2007.

Pipe material was used as grouping criteria and the logarithm of the length and the pipe diameter enter the model as covariates. The age of pipes and the number of previous failures are already taken into account during the model construction, so there is no need to use them as model covariates. The estimated model coefficients are presented in Tables 5.1 and 5.2.

In Table 5.1, parameters were estimated using the likelihood function restricted to positive values of α and δ. Restrictions were included in the likelihood function using the variables transformations: $\alpha = e^{la}$ and $\delta = e^{ld}$. The la and ld parameters are estimated with no restrictions using the maximum likelihood method, and then, $\hat{\alpha}$ and $\hat{\delta}$ are obtained applying the exponential to $\hat{la}$ and $\hat{ld}$. Since Le Gat (2009) restricts $\delta > 1$, the maximum likelihood estimation using this restriction is presented in Table 5.2. In this table, the restriction in δ was imposed, using the variable transformation $\delta = e^{ld} + 1$.

Table 5.1: Maximum likelihood estimates of LEYP coefficients by material category, restricted to $\alpha > 0$ and $\delta > 0$.

	HDPE			**PVC**		
Parameters	Estimate	$-2\ln(LR)$	p-value	Estimate	$-2\ln(LR)$	p-value
α	6.122	103.54	<0.0001	2.527	89.83	<0.0001
δ	0.633	37.34	<0.0001	0.737	14.10	0.0002
β_0	-5.606	400.86	<0.0001	-5.243	472.44	<0.0001
Diameter (β_{diam})	-0.003	4.31	0.0379	-0.001	0.14	0.7043
ln Length ($\beta_{lnlength}$)	0.587	91.94	<0.0001	0.482	234.20	<0.0001
without covariates		92.44	<0.0001		234.11	<0.0001
	AC			**DCI**		
Parameters	Estimate	$-2\ln(LR)$	p-value	Estimate	$-2\ln(LR)$	p-value
α	1.989	242.94	<0.0001	12.371	7.16	0.0075
δ	0.668	13.23	0.0003	0.522	2.47	0.1159
β_0	-3.700	79.20	<0.0001	-4.677	35.93	<0.0001
Diameter (β_{diam})	-0.003	45.40	<0.0001	-0.001	0.02	0.9002
ln Length ($\beta_{lnlength}$)	0.304	313.77	<0.0001	0.250	3.28	0.0702
without covariates		344.79	<0.0001		3.70	0.0545

As presented in Table 5.1, the estimates associated with ln length are far from one, which means that the intensity of the process will not be proportional to the length of pipes. Nevertheless, this variable is still one of the most important variables of the model, as can be seen by the p-value in all pipe material categories. On the contrary, pipe diameter seems to be the less significant variable.

Another interesting fact is that the estimate $\hat{\delta}$ is smaller than 1. This indicates that the intensity process is decreasing with the elapsed time. However, using the log-likelihood function of the NHPP, fixing $\alpha = 0$ (Equation 5.21), the estimated $\hat{\delta}$ is greater than one in all materials except HDPE. It seems, although the general network is ageing, the intensity of the process is very high after the occurrence of a failure and tends to decrease until the next failure occurs. This is explained by the fact that a pipe is very likely to fail soon after the occurrence of a failure and this likelihood tends to decrease with time; the empirical hazard function presented in Figure 1.1 translates this fact.

Table 5.2 presents the estimated parameters and their likelihood ratio test, restricting $\delta > 1$,

Table 5.2: Maximum likelihood estimates of LEYP coefficients by material category, restricted to $\alpha > 0$ and $\delta > 1$.

	HDPE			**PVC**		
Parameters	Estimate	$-2\ln(LR)$	p-value	Estimate	$-2\ln(LR)$	p-value
α	3.635	74.79	<0.0001	1.904	76.05	<0.0001
δ	1.000	0.00	1.0000	1.000	0.02	0.8875
β_0	-6.478	767.50	<0.0001	-6.191	795.80	<0.0001
Diameter (β_{diam})	-0.000	0.01	0.9372	-0.000	-0.08	1
ln Length ($\beta_{lnlength}$)	0.568	89.15	<0.0001	0.513	234.63	<0.0001
without covariates		89.59	<0.0001		234.58	<0.0001
	AC			**DCI**		
Parameters	Estimate	$-2\ln(LR)$	p-value	Estimate	$-2\ln(LR)$	p-value
α	1.789	265.91	<0.0001	7.938	5.04	0.0248
δ	1.000	0.00	1.000	1.000	0.00	1.000
β_0	-5.182	1433.74	<0.0001	-6.122	64.55	<0.0001
Diameter (β_{diam})	-0.003	37.83	<0.0001	-0.001	0.19	0.6593
ln Length ($\beta_{lnlength}$)	0.337	318.01	<0.0001	0.275	2.97	0.0849
without covariates		341.74	<0.0001		3.88	0.0488

as imposed in Le Gat (2009). In this table, it can be seen that $\hat{\delta}$ is always one and that the associated p-value is very high. This indicates that, for this failure data set, the δ of LEYP is less than one, which is explained by the decreasing empirical hazard function presented in Figure 1.1. Therefore, predictions of LEYP are conducted using the estimates in Table 5.1, where δ is only restricted to being positive, since that $\delta > 1$ is not a necessary condition to derive the LEYP distribution.

Having the parameters estimated using the training set, the number of failures occurring during the test window were predicted for each pipe. These predictions were compared with the observed values of the test sample, as explained in Chapters 3 and 4. Figure 5.1 shows the mean observed failures in each 0.1-quantile of predicted failures, for each pipe material. It can be seen that there is a general overestimation of predicted failures of the last quantile in each pipe material. Failures in DCI, once more, are poorly forecasted, reinforcing the point that it is important to have an extended failure data to build accurate failure prediction models.

Figure 5.2 groups the predictions in all pipes. In spite of the overestimation of the number of failures in the last quantile, it seems that LEYP still detects effectively pipes that are more prone of fail. The mean number of observed failures in the last quantile is as high as in WALM, Figure 4.3, but it is higher than the one in the Poisson process, Figure 3.1.

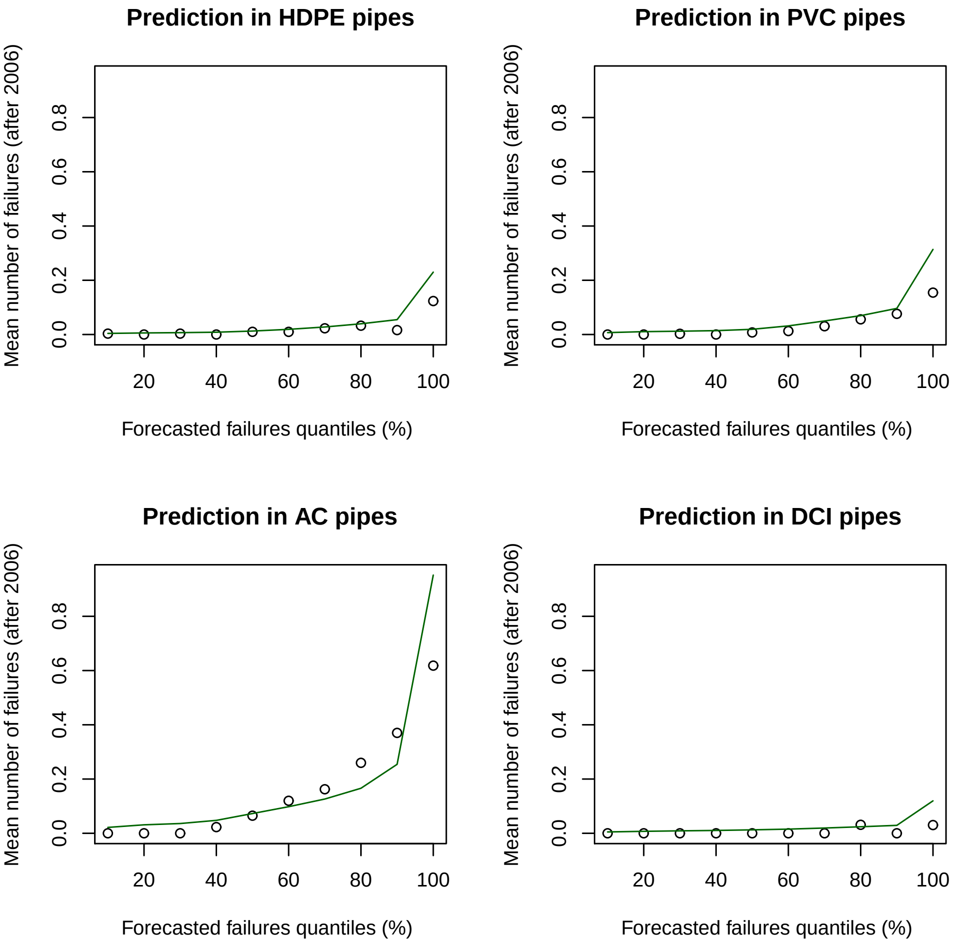

Figure 5.1: LEYP predictions in each material category.

A second test was carried out using the random division method to define the training and test samples. Parameters of LEYP are estimated using the maximum likelihood method restricted to $\alpha > 0$ and $\delta > 0$, using a random sample of 50% of the water network pipes; Figure 5.3 presents failure predictions in the other 50% sample of the water network, using Equation 5.17, where:

a is 0;
b is 0;
j is 0;
s represents the age of pipe at 01-01-2001, or 0 if installed after 2001;
t represents the age of pipe at 31-03-2011.

As can be seen in Figure 5.3, predictions are much closer to the observed values than in the previous prediction plots. The mean number of observed failures in the last prediction

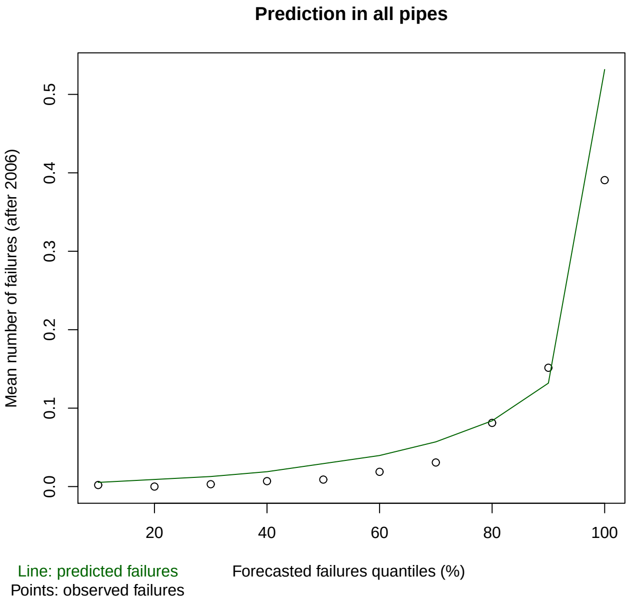

Figure 5.2: LEYP predictions in all pipes.

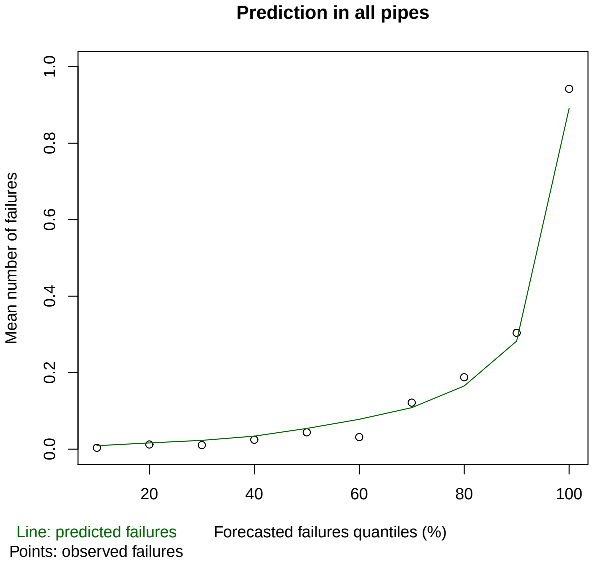

Figure 5.3: LEYP predictions in random sample.

quantile stands out from the previous quantiles (twice as high), which translates a good capacity of detecting the 10% pipes that are more prone to fail.

Chapter 6

Comparison of model predictions

After analysing the failure data in Chapter 2 and applying the three models in Chapters 3, 4 and 5, the obtained results are compared in this chapter. Several prediction plots are presented to better visualise the differences between the studied models. The comparisons conducted and presented in this chapter are: number of failures that can be avoided using the different models; Receiver Operating Characteristic (ROC) curves comparison; the mean of the predicted number of failures for each observed failures category; and the absolute error of the predicted number of failures.

6.1 Prediction of failures based on a temporal divided sample

In this section, the prediction results of the three models are compared when predicting future failures of a water network, using the available pipe history. The training and test sets are built according to the temporal division method. That is, the training set is the failure data before 01-01-2007, whereas the test set is composed of all failures occurring after 01-01-2007 in those pipes.

First, the number of failures that can be avoided renewing a defined percentage of the water supply system is compared. The aim of this comparison method is to evaluate the ability the models have to prioritise pipes according to their likelihood of failing. If the failure prediction model is good then a large percentage of failures occurring after 2006 can be avoided rehabilitating a small percentage of the water network. Figure 6.1 illustrates this term of comparison.

The pipes are sorted according to the predicted failure rates (for each model) and the cumu-

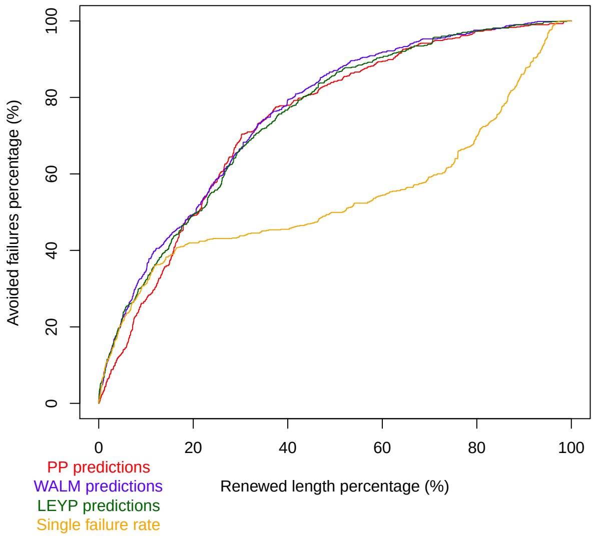

Figure 6.1: Avoided failures per length of rehabilitated pipes.

lative observed failures are plotted against the cumulative pipe length. The model results plotted in Figure 6.1 were obtained using the Poisson process (red line), the WALM using the pipe age at last failure approach (blue line) and LEYP (green line). Individual failure rates were also computed for each pipe, dividing the number of failures occurring before 2006 over its length, and used as a prioritisation criteria (orange line).

In Figure 6.1 it can be seen that LEYP and WALM results are very close to each other. The Poisson process, in spite of having a general good prediction, lacks the capacity of detecting the pipes at risk (more likely to fail). Clearly the priority pipes detected by the Poisson process did not suffer as much failures as the ones detected by the two other models. This fact can be explained due to the lack of ability of this model to differentiate pipes. Since it is purely based on categories, every pipe in the same category has exactly the same predicted failure rate.

On the other hand, the individual failure rate approach can effectively detect the pipes more likely to fail until 15% of the water supply system is renewed; after this value it is a rather poor method. The reason is that only 15% of pipes have recorded failures before 01-01-2007, so all other pipes have an individual failure rate of zero.

WALM and LEYP both seemed to detect quite efficiently pipes with higher probability of failing. Nevertheless WALM curve appears to avoid more failures for most of values of the renewal percentage, specially when renewing a large network percentage. To illustrate better these conclusions, Table 6.1 shows the percentage of avoided failures when renewing a small portion of the water supply system network.

Table 6.1: Percentage of avoided failures by prediction models.

Models	Rehabilitated length		
	0.5%	1%	5%
Poisson process	1.7%	3.0%	13.2%
WALM	4.1%	6.0%	22.3%
LEYP	5.3%	6.7%	22.2%
Single failure rate	4.1%	7.0%	21.3%

From Table 6.1 it can be seen that failures avoided using LEYP were very close to those using WALM. It is important to note that the simple method of prioritising pipes according to their past individual failure rates allowed avoiding a considerable amount of future failures, rehabilitating a small percentage of the water supply network. This shows that pipes more likely to fail are those with higher past failure rates. Therefore, even with few available attributes, water utilities can prioritise effectively rehabilitation actions using a well organised failure data.

A receiver operating characteristic (ROC) graph was built for the three model results. ROC graphs are visualisation techniques that allow to compare different classification models performance. ROC curves are based on confusion matrices, using the sensitivity and specificity of a classifier. Although ROC curves were used mostly in medicine and psychology, it has been now introduced in many other fields and it has been found useful in decision problems. Débon *et al.* (2010) apply these performance assessment method to a water supply network in order to compare several prediction models. Problems that can be assessed with ROC graphs have a decision binary variable (in this case if pipe will fail or not) and a classification model that produces a continuous response (in this case the probability of failing) that will classify objects in both prediction classes, according to different discrimination thresholds. For each threshold the classifier presents a true positive rate and a false positive rate. Table 6.2 shows a confusion matrix in order to explain what are true and false, positive and negative observations.

The true positive rate, TPR, of a classifier is the proportion of the actual positive cases that were correctly identified, and is calculated using Equation 6.1.

$$TPR = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}. \tag{6.1}$$

The false positive rate, FPR, is the proportion of the actual negative cases that were incor-

Table 6.2: Confusion matrix.

		Actual class	
		P′	N′
Predicted class	P	True positives	False positives
	N	False negatives	True negatives

rectly classified as positive, and is calculated using Equation 6.2.

$$FPR = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}. \tag{6.2}$$

A ROC graph consists simply in plotting these two rates for different discrimination thresholds. As the threshold decreases both rates will increase. Nevertheless, for a good prediction model the true positive rate will increase much faster than the false positive rate, meaning that the ROC curve will be way above the identity line $f(x) = x$.

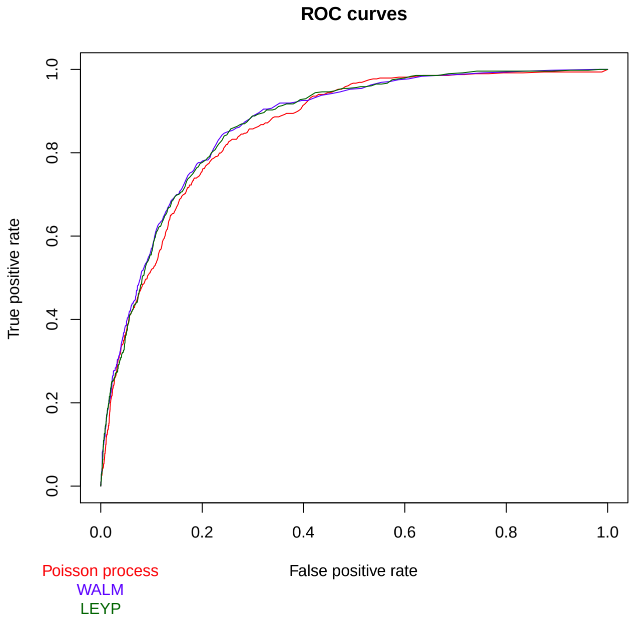

Figure 6.2: ROC curves failures after 2006.

Figure 6.2 presents the ROC curve of each model, using the probability of failing for each pipe as the continuous response variable.

In this comparison LEYP and WALM results were also very alike, standing out from the

Poisson process results, which presented a ROC curve slightly under LEYP's and WALM's curves for most of the considered thresholds.

Another comparison was conducted, dividing the test sample in five classes according to the number of observed failures in each pipe. The mean of the predicted number of failures for each observed class is plotted in Figure 6.3.

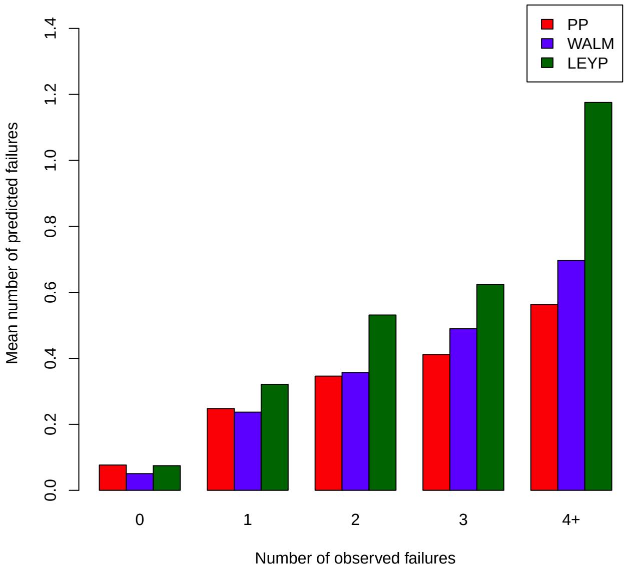

Figure 6.3: Mean predicted failures per observed failures categories.

LEYP appeared to be the most accurate model, because the obtained predictions, in pipes that failed more, were higher than predictions obtained using other models. However, it overestimates failures in pipes with no observed failures. Although it appears that this overestimation is very insignificant, the first category of pipes with 0 failures is 297km long, while the total length of all other categories is only 39km; therefore this overestimation is not negligible.

The Poisson process showed the same overestimation in pipes with no observed failures and a significant underestimation in pipes with more observed failures, which indicated that this model was less accurate than the other models.

The absolute error of the predicted number of failures can be calculated, using Equation 6.3.

$$\text{Abs. Error} = \sum_{i=1}^{n} |o_i - \hat{e}_i|, \tag{6.3}$$

where n is the number of pipes; o_i is the observed future failures in pipe i; and $\hat{e}_i$ is the predicted number of future failures in pipe i.

Table 6.3 presents the absolute error calculated using the model results in each pipe material category.

Table 6.3: Absolute error of the predicted number of failure per pipe material.

Models	HDPE	PVC	AC	DCI	All pipes
Poisson process	204.6	326.8	785.5	10.6	1327.4
WALM	120.9	259.6	687.6	7.1	1075.3
LEYP	169.8	330.0	782.6	10.1	1292.6

In this table, the LEYP and the Poisson process errors are very similar. WALM is clearly the more accurate model of the three, as the absolute error is smaller in each pipe material category.

In most of comparisons LEYP and WALM seemed to present similar results, leaving the Poisson process slightly behind. However, LEYP had the tendency to overestimate failures, while WALM seemed to produce more accurate predictions. In fact, there are 703 observed failures after 01-01-2007. Failures predicted by the Poisson process, WALM and LEYP were 899, 642 and 938, respectively.

Based on these comparisons, WALM showed to produce more realistic results, but LEYP seemed also to determine very efficiently pipes with higher likelihood of failing. The Poisson process is limited in prioritising pipes, but general predictions were not very biased. Thus it can be used as a very simple model to predict the overall failures, but it would be a poor method to prioritise rehabilitation actions. If a water supply system network has a complete (and long) failure history of pipes, then the past failure rate seems to be an efficient prioritising criteria.

6.2 Predictions of failures based on a random sample

Good failure data history may not be available, due to recent collection and organisation of the failure data, or due to data inconsistencies; therefore the data set is not trustworthy. In that case, failure predictions should be based on failure data provided by similar water supply systems. So, it is important that the failure prediction models present good results, not only predicting failures in pipes with failure history, but also predicting failures in other sets of pipes with no recorded failure history.

In this section the training and test samples were built using the random division method. The training sample is composed by a random 50% sample of all pipes and all associated failures, whereas the test sample is composed by the other pipes and their associated failures. Therefore the training and test windows go now from 01-01-2001 to 31-03-2011.

The first comparison consisted in estimating the number of failures that could be avoided by renewing some portion of the water supply network. Figure 6.4 shows the results visually, Table 6.4 shows the resulting number of avoided failures when a small percentage of the water supply network is renewed.

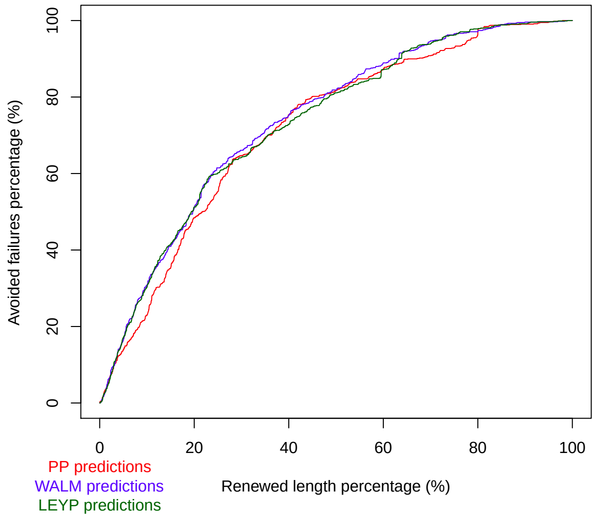

Figure 6.4: Avoided failures in random sample of pipes.

Table 6.4: Avoided failures in random sample of pipes.

Models	Rehabilitated length		
	0.5%	1%	5%
Poisson process	0.9%	3.0%	13.9%
WALM	1.0%	2.3%	16.4%
LEYP	0.5%	2.7%	17.0%

All three models seemed to present similar results. Although in Figure 6.4 WALM and LEYP results appeared slightly better, when a small portion of the water supply system network is rehabilitated the three models are very similar (Table 6.4). In fact, this is not due to an improvement of the predictions obtained using the Poisson process, it is rather caused by a decrease of the percentage of avoided failures using LEYP and WALM, as can be seen by comparing Tables 6.1 and 6.4.

To study these models as classifiers, a new ROC curve was built for each model; the plot is presented in Figure 6.5.

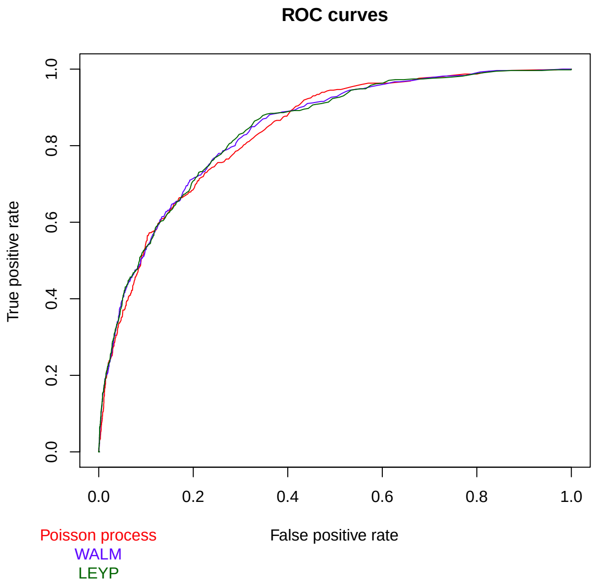

Figure 6.5: ROC curves in random sample.

In Figure 6.5 all three models presented very similar results, being difficult to select the best model to use as a classifier. Comparing Figures 6.5 and 6.2, it can be seen that the WALM and the LEYP results became much closer to the Poisson process results.

Another comparison performed was plotting the mean of the predicted number of failures for each category of observed failures.

In Figure 6.6, LEYP stands out from the other two models in the pipe categories with more observed failures. Nevertheless, in the categories with few observed failures (0, 1 and 2), the predictions obtained using the three models were very similar. The Poisson process seemed to be the model with the worst results, since it is the one that showed higher overestimation in the non-failing pipes category, while being the one that underestimated more in the last pipe categories.

The absolute error of the predictions of each model, in each pipe material category, is presented in Table 6.5.

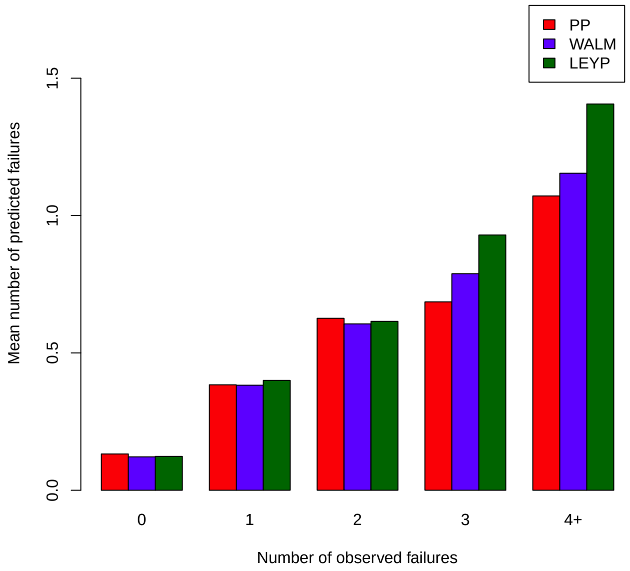

Figure 6.6: Mean predicted failures per observed failures categories in random sample.

Table 6.5: Absolute error of the predicted number of failure per pipe material

Models	HDPE	PVC	AC	DCI	All pipes
Poisson process	242.2	362.7	756.7	17.3	1378.8
WALM	198.9	359.1	753.4	13.2	1324.7
LEYP	205.9	358.7	754.9	13.9	1333.4

In this table it can be seen that the absolute error values were very close for all prediction models. WALM is the one that appeared to present more accurate predictions. Comparing both Tables 6.3 and 6.5, the absolute error of predictions in pipes with no failure history seems higher. However, despite the test sample is smaller using the random division method, the time period of the test sample is about twice as long. In fact, 958 observed failures occurred in the test sample when using the random division method, whereas only 703 observed failures occurred in the test sample of the temporal division method.

The model results, detecting pipes with higher likelihood of failing, when using a random sample of pipes seemed to be worse than using the temporal divided sample (see Tables 6.1 and 6.4). However, while the Poisson process results maintain almost the same in both tables, there is a significant downgrade in LEYP and WALM results. The study of the significance of each covariate in the WALM (Chapter 4) and the analysis of the variables effect in the failure rate (Chapter 2) show that one of the most important pipe variables is the previous failures variable. The application of the single-variate Poisson process did not take into account this variable. Therefore, when the pipes failure history is unknown, performances of LEYP and WALM get more similar to that of the Poisson process.

Despite the importance of the previous failures variable in determining the future failures, as discussed earlier, the general predicted number of failures when no pipe failure history is available was very accurate. In fact, the total number of predicted failures in all models were close to the observed failures: there were 958 observed failures; the Poisson process, the WALM and the LEYP predicted 960, 912 and 946 failures, respectively.

The overestimation in LEYP results, when using the temporal divided sample, is no longer noticeable when predicting pipes with no failure history (i.e. using the random sample). The linear increment of the LEYP intensity (and expected number of failures) with the number of past failures may not be realistic and may explain the mentioned issue.

Chapter 7

Conclusions

This chapter summarises the main findings of this study, describing the advantages and disadvantages of each model and pointing out the importance of having a complete and reliable failure data.

The three models considered in this study presented good prediction results. WALM appeared to be the best of the three models, because it combined accurate predictions with a good ability of detecting pipes more prone to fail. On the other hand, the Poisson process was the one that presented the worst general performance.

The Poisson process was implemented to study the behaviour of a simple and intuitive prediction model. This model is easy to understand and easy to implement. The only parameters to be estimated, in the case of the Poisson process, are the failure rates for each defined category, presenting an analytical and intuitive expression (Equation 3.2). So, it does not require numerical algorithms to estimate the failure distribution. Finally it allows the user to easily define the categories to build the model. However, its results are slightly worse than the results obtained by the other two models.

The fact that the Poisson process is defined by categories rather than covariates, implies that every pipe in the same category have the same failure rate. This fact leads to a difficulty in differentiating pipes and selecting the pipes that are more likely to fail. That is why the number of avoided failures obtained using this model, is not as high as when using other models (Figure 6.1). In order to better differentiate pipes, there should be defined an increased number of categories. However, a high number of categories would cause an over splitting of the failure data leading to non-significant failure rates, which could lead to completely biased predictions. The individual past failure rate could be used to prioritise pipes within the same pipe category.

When studying pipes with no failure history, the difference between the Poisson process and the two other methods decreases. This supports that the previous failures variable is of utmost importance when predicting future failures. The previous failures variable could enter the Poisson process as grouping criteria. However, failure data would have to suffer an additional time division: one to build the pipe categories associated with the number of past recorded failures and another to calculate the observed failure rate in each category. This is a complex method and requires an extensive failure data set.

Accelerated lifetime models are significantly different from the other studied models. These models fit the time between failures, rather than the number of failures. In this study, the distribution of the time between failures was chosen to be a Weibull distribution, because it combines several features:

(*i*) Weibull presents an intuitive hazard function, which can be important to understand the different covariates effect in the lifetime distribution;

(*ii*) the Weibull hazard function is not constant over time, allowing to better fit the failure data;

(*iii*) it also presents a simple survival function, which is important to obtain the likelihood function (Equation 4.7) and to generate random times during the Monte Carlo simulations (Equation 4.10).

The fact that WALM fits the time between failures rather than being a counting process, can produce a different knowledge about pipes and their failures. Different estimated quantiles of this distribution could give useful information about pipes, such as the service life of pipe materials. However, quantiles of the estimated distribution are far from realistic. The 50% quantile of an average pipe of Asbestos Cement is about 100 years and in other pipes can go over 500 years. This is probably due to the short time records of the failure data used in this study. The short period of observations does not allow efficiently understanding the deterioration process of pipes. With a more extended failure data it would be expected to obtain more realistic results.

Weibull distributions present the disadvantage of not being analytically convoluted. So, it is not possible to find the general number of failures distribution. Thereby, the expected number of failures needs a Monte Carlo simulation process. Nevertheless, this simulation process can give very good predictions with a high number of experiments. During the WALM analysis the number of failures prediction process was carried out using 1,000 experiments, for each pipe. This number of experiments seemed to be adequate, since more tests were conducted using more experiments, but predictions were very similar.

Despite the fact that the predicted number of failures are not directly given by a random variable distribution, these predictions were very accurate. Actually, this was the model that presented the best results in all comparisons, among the results of the three models. The WALM combines a great capacity of detecting pipes with higher likelihood of failing, which can be translated by the percentage of avoided failures using this model (Table 6.1) and accurate predictions (Figure 4.3 and Table 6.3).

Prioritising the pipes more likely to fail among those with no failure history is clearly a more difficult task, specially since the previous failures variable is one of the most significant covariates in the WALM regression (Table 4.1). However, WALM results obtained using a random sample of pipes were still very accurate (Figure 4.5 and Table 6.5) and the number of avoided failures was similar to the number of avoided failures using other models (Table 6.4).

The linear extended Yule process was first applied to water supply network failure data by Le Gat (2009). This model is born from a pure birth process (also called Yule process), where the rate of the process increases linearly with the number of events. A linear extension of this model, maintains this proportion between the process rate and the number of previous events. Le Gat (2009) supports the use of this model based on the fact that an increasing number of previous failures leads to a higher future failure rate. Nevertheless it is not clear that the relation between past number of failures and future failure rate is linear. Therefore a non-homogeneous birth process using other functions of the number of previous failures, α_n, (see Equation 5.5) to describe the intensity of the process could be studied. For instance, a limited function (continuous convergent function or a finite valued function) could be a good solution. However, considering other α_n functions can increase the complexity of the prediction model.

The nature of LEYP is probably the reason why this model presents a clear tendency to overestimate the number of future failures (Figure 5.2). Nevertheless, LEYP presented a really good performance when detecting pipes more prone to fail (Figures 6.1 and 6.3). When applied to pipes with no recorded failure history, LEYP does not make the same overestimation; predictions are significantly more accurate, as can be seen in Figure 5.3.

The predictions obtained using the three models can be considered accurate. WALM and LEYP present important advantages over the Poisson process. The use of covariates allows a better understanding of the different effect of the pipe variables and does not require the division of the failure data, which could be a problem when dealing with small failure data. The Poisson process has the advantage of being a simple method, which makes it easy to understand and to apply.

As might be expected, it was confirmed that the past failures variable was extremely important when predicting future failures. Pipes that have failed before, present a much higher probability of failing in the future. This may be explained by the fact that a pipe repair is not the same as a pipe replacement; in general, a pipe becomes more fragile after a repair than before the failure happens. Another possible reason is that there are other unknown attributes that influence the failure rate, such as environmental, traffic or operating pressure conditions. And so, the higher failure values could be explained by other characteristics associated to the installation site rather than the fragility of those pipes. This is what makes predictions in pipes with no recorded history more difficult. This is the reason why it is so important to build a complete and trustworthy failure database.

All models require an organised information system with a complete inventory with all pipes well characterised. Even if there is not an extensive pipe database with a lot of variables, good predictions can be done, as proved by the application of the Poisson process, using only three variables. It is not necessary long maintenance records, LEYP and specially WALM offer very good predictions, only using six years of rehabilitation records. What is strictly necessary is to have a complete and up-to-date pipe inventory of all pipes and a reliable rehabilitation database properly linked to the pipe inventory. The information system should be periodically reviewed in order to detect and correct possible inconsistencies.

References

Alegre, H., Covas, D., Coelho, S. T., Almeida, M. C. and Cardoso, M. A. (2011, Sep.). An integrated approach for infrastructure asset management of urban water systems. Mülheim An Der Ruhr, German. IWA 4th LESAM.

Clark, R. M., Stafford, C. L. and Goodrich, J. A. (1982). Water distribution systems: a spatial and cost evaluation. *Journal of the Water Resources Planning and Management Division, ASCE* **108**(WR3), 243–256.

Cox, D. and Oakes, D. (1984). *Analysis of survival data.* London: Chapman & Hall.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistics Society* **34**(B), 187–220.

Débon, A., Carrión, A., Cabrera, E. and Solano, H. (2010). Comparing risk of failure models in water supply networks using roc curves. *Reliability Engineering and System Safety* **95**(1), 43–48.

Gefeller, O. and Dette, H. (1992). Nearest neighbour kernel estimation of the hazard function from censored data. *Journal of Statistical Computation and Simulation* **43**, 93–101.

Gustafson, J. M. and Clancy, D. V. (1999). Modelling the occurrence of breaks in cast iron water mains using methods of survival analysis. *Proceedings of the AWWA Annual Conference*.

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

Jeffrey, L. A. (1985). *Predicting urban water distribution maintenance strategies: a case study of New Haven, Connecticut.* Massachusetts Institute of Technology.

Kleiner, Y. and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* **3**(3), 131 – 150.

Kulkarni, R., Golabi, K. and Chuang, J. (1986). Analytical techniques for selection of repair-or-replace options for cast-iron gas piping systems - phase i. Technical report, Chicago.

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, Cemagref Bordeaux, Paristech.

Le Gat, Y. and Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water* **2**(3), 173 – 181.

Lei, J. (1997). *Statistical approach for describing lifetimes of water mains: case Trondheim municipality.* Trondheim, Norway: SINTEF, Civil and Environmental Engineering, Water and Waste Water.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reed, W. (2011). A flexible parametric survival model which allows a bathtub-shaped hazard rate function. *Journal of Applied Statistics* **38**(8), 1665–1680.

Ross, S. (2006). *Introduction to Probability Models, Ninth Edition.* Orlando, FL, USA: Academic Press, Inc.

Shamir, U. and Howard, C. (1979). Analytic approach to scheduling pipe replacement. *Journal of American Water Works Association* **71**(5), 248–258.

Silva, G. O., Ortega, E. and Cordeiro, G. M. (2009). A log-extended weibull regression model. *Computational Statistics & Data Analysis* **53**(12), 4482–4489.