STE - A Survival Tree Ensemble: Application to the Prediction of Pipe Failures in Water Supply Systems

Inês Oliveira
ines.m.r.oliveira@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

December 2019

Abstract

Drinking water is provided to the majority of urban population through a complex water distribution system, which is constantly being affected by economic constraints, climate changes, infrastructure deterioration, and even increasing urbanisation and customer demand. Hence, water utilities are challenged with an increasingly demanding task, and they are being forced to venture in long-term strategies which direct their efforts in planning and prioritisation of rehabilitation actions. Despite the common reactive approach in which water utilities base their efforts, the truth is they are recently moving towards a more proactive approach and using failure prediction strategies in infrastructure asset management. The goal of this project is to apply some Machine Learning techniques to build prediction models and then to evaluate their performance when compared to existing approaches such as the one implemented in the Baseform software.

Keywords: Infrastructure Asset Management, Water Distribution Systems, Failure Prediction, Machine Learning, Survival Analysis.

1. Introduction

Managing a water distribution system is a challenging task, making water utilities more concerned and focused on applying proactive approaches which facilitate the process of prioritising critical zones and pipes of the system. Traditional condition assessment techniques are often not the most adequate not only because they are time and resource consuming tasks but also because of their inability in planning in the long-term. The cost associated with regularly checking all pipes makes this procedure virtually impossible, aggravated by the fact that the majority of pipes are buried. As a result, most water utilities end up using more reactive approaches despite often being unsustainable. The cost and necessity to guarantee service quality lead most water utilities to move towards a more proactive approach and scheduled preventive maintenance.

Failure analysis has, therefore, a fundamental role in Infrastructure Asset Management (IAM). Failure data has some particular specificities that distinguish it from other types of data. Such characteristics, combined with the fact that one of the main interests in these problems is to study the time to next failure, make survival analysis emerge as a very appealing technique. Taking advantage of the latest developments of the field, a machine learning approach under a survival analysis perspective is going to be experimented. The goal of this work can be summarised as developing machine learning methodologies for survival analysis, which are simultaneously able to predict failures and to help plan and prioritise rehabilitation actions.

2. Pipe Failure Prediction

Only recently, water utilities have become aware of the importance of keeping an organised system with updated and complete records on both the network inventory and corresponding failure reports. For the majority of water utilities, failure history is very limited when compared to the age of the network itself, and usually contains inaccurate, estimated and missing data, which largely contribute for a more challenging failure analysis. In practice, only some variables are actually needed when collecting the data. Information generally comes from two different sources: one is the asset inventory usually from a Geographic Information System (GIS), and the other one is a failure record usually from a work order system. Regarding the asset data, each pipe has to be characterised by a unique code, installation date, material, diameter and length. As for the work orders, each record has to be associated with the ID of the intervened pipe and the corresponding date of failure. Moreover,

an association between both sources of information can be done resulting in a better characterisation of each work order and in the possibility to obtain for each pipe additional variables, such as the number of previous failures, distance between failures and even its age at the time of the event. As it was mentioned, several problems in the data contribute to the challenging task of pipe failure prediction. Besides inaccurate and missing records, common problems involve failure dates prior to the installation of the corresponding pipes, intervened assets not present in the inventory or even not registered in the work order record. Other problems concern the recent and short history of the maintenance actions undertaken in the network and the unavailability of which assets are still active or not. If such problems are not corrected or if a careful verification is not done to the available data, then this will certainly affect the posterior built models.

2.1. Data Specificities

Other issues can difficult the analysis of failures in water distribution systems, and they are some characteristics which distinguish the pipe failure data from other common data in the field of survival analysis. The first distinction is the fact that events are recurrent, i.e. pipes have the possibility to fail more than once. Generally, the propensity to fail increases with the pipe ageing and with the more maintenance actions which are executed. This kind of actions have a limited power when it comes to extend pipeline life, so their eventual replacement will be inevitable in the future. Therefore, in order to preserve the nature of failure data, it is fundamental to account for the recurrence of failures throughout their lifetime. Such characteristic is usually introduced in the form of a time-varying covariate. The number of previous failures is transformed into a variable dependent on time, which is updated with each new event. The second distinction is the fact that failure data is usually considered to be Left-Truncated and Right-Censored (hereafter denoted by LTRC). It is considered to be left-truncated since it is generally impossible to determine when and how many failures occurred (if in fact, any occurred) before the beginning of the recorded history. For all pipes installed before the start of the observation period, their past history is unknown. The right-censored notion occurs when pipes are followed up during a certain period, and they have not failed in this time, that is, the elapsed time without failing is known but not the exact instance of the failure. The event of interest not happening up to the censoring time is the only thing available. In this context, such censoring might happen if for example a pipe is abandoned without registering any failures during the observation period, or if their survival time is larger than the end of the study (whether or not they have failed before).

2.2. Literature Review

Various approaches have already been proposed in an attempt to predict water pipe failures and to create prediction models, leading to significant developments in the field of IAM. In fact, failure prediction models have been applied to water systems since late 70s, where models were mainly deterministic, that is, the response variable (such as number of failures or time to next failure) was directly obtained from a function of the covariates. However, in the context of pipe failure prediction, it may be beneficial to take into account the random nature which characterises pipe failures. Thus, stochastic models (single-variate and multivariate) arise as a much more suitable approach to model failures in water systems. One of the most well-known models in the context of multivariate stochastic models is the Proportional Hazard Model (PHM), proposed by [4], where the hazard rate can be obtained as a function of time and covariates. This model and its extensions are often used with survival data due to its capability of dealing with censored information. Currently, an approach suggested by [10] and later improved by [12], is one of the best in obtaining failure rates and probability estimations for each pipe along time. It is called Linear Extended Yule Process (LEYP), and it is essentially a multivariate counting process in which the process rate depends on the age, previous failures and the available covariates. Presently, this is one of the models implemented in Baseform software [1].

Studies involving time-to-event data have been increasingly using such methods due to their capability of obtaining simple interpretations of the effects of each covariate of the data. However, these methods also present some weaknesses, and the main one is related to the fact that a specific relationship between the covariates and the response has to be forced. More flexible alternatives to these (semi)-parametric approaches are, therefore, needed. Tree-based methods are popular non-parametric substitutes. In contrast with the (semi) parametric models, these methods offer the flexibility to automatically detect interactions without having to define them beforehand. Even though these methodologies were initially developed for categorical or continuous outcomes data, from the mid-1980s to the mid-1990s, the interest in using them for censored survival data rapidly boosted new approaches that took the failure data specificities into account. A review on these survival tree methodologies can be found in [2]. Once the basic survival tree methodology was established, research moved in different directions. Over the last two decades, much of the methodological

work has concentrated in introducing time-varying covariates, increasing the predictive power of a single tree by means of ensemble methods and extending the basic survival setup (which involves only right-censoring) to left-truncation.

3. Survival Trees

During the last years, machine learning techniques have been very successful in numerous fields, mainly due to their capacity in modelling non-linear relationships and due to the quality of the built models. Therefore, it does not come as a surprise the fact that there are already machine learning approaches specifically developed to survival analysis. However, one of the main challenges when dealing with survival data is the difficulty to work with censored information. Wang et al. [13] reviewed the existing literature on survival analysis and its recent developments from a machine learning perspective.

The general formulation of a problem in survival analysis can be presented for a given instance i, represented by $(T_i, \delta_i, \mathbf{X}_i)$:

- T_i corresponds to the observed event time or the censored time (in case there is censoring);
- $\delta_i \in \{0, 1\}$, i.e., $\delta_i = 1$ for uncensored observations (and T_i is the event time) and $\delta_i = 0$ for censored observations (and T_i is the censored time);
- $\mathbf{X}_i \in \mathbb{R}^{1\times p}$ is the covariates vector.

where the goal is to estimate the time-to-event T_j for a new observation j with features $\mathbf{X}_j$. Notice that to adjust this formulation to a LTRC problem, it is enough to represent each instance as $(L_i, R_i, \delta_i, \mathbf{X}_i)$, where (L_i, R_i) are the left-truncation time and the observed/censored time, respectively.

In such context, the main challenge of machine learning methods when handling survival data is to appropriately deal with censored data and time estimation of the model. Models such as survival trees stand out precisely for their capability in dealing with censored data maintaining the tree structure. The majority of tree methodologies are able to handle the most basic setup for survival outcomes, which corresponds to right-censored data with time-independent covariates. Fu and Simonoff [6] proposed an extension of two popular tree methods to LTRC data, as well as adding the capability to deal with time-varying covariate data. Their strategies are implemented in the `R` package `LTRCtrees` [5]. The first approach, developed by [7], implements a survival tree, embedded into a condition inference framework, using the log-rank score as the splitting method, whereas the second approach was proposed by [11] and it concerns the construction of relative risk trees.

Through data reformulation, these algorithms can be much more versatile than other existing approaches, simply because they can fit survival data with time-varying covariates. The key idea is to take the classical formulation of the problem and change the representation of instances. Instead of the classical approach (*time*, *event*), the idea is to have a triplet of the form (*start*, *stop*, *event*) where *start* corresponds to the left-truncation time and *stop* is the right-censored time. In this work, *start* will correspond to the age of the pipe at the beginning of the study (and it will be zero if it has not been installed before the start of the study), and *stop* will correspond to the age of the pipe at the time of the failure, or to its age at the end of the observation period (in case it has never failed before).

3.1. Conditional Inference Trees

Recursive binary partitioning is a popular tool nowadays. One of the most popular implementations includes the CART (Classification and Regression Trees) algorithm proposed by [3]. Typically, most of these approaches perform an exhaustive search over all possible splits maximising an information measure of node impurity in order to select the variable to which corresponds the best split. However, two fundamental problems arise with such approaches. One of them is overfitting, and the other one is the existence of bias in the selection procedure of covariates with many possible splits.

The overfitting problem can be solved through pruning, which is basically a tree size reduction technique used in the building process and which removes sections of the tree that can represent noisy data. The bias comes from maximising the splitting criterion over all possible partitions simultaneously and has been identified as a problem over the years.

In [7], a tree method embedded in a condition inference framework and based on the statistical properties of the variables is proposed, allowing these trees to deal with both the overfitting and the variable selection problems. The main idea is that the conditional distribution of the statistics measuring the association between responses and covariates is the basis for an unbiased selection among covariates measured at different scales. Moreover, defining a stopping criteria based on the non-rejection of a hypothesis of independence between the response variable and each covariate, [7] proved that it is possible to obtain models whose predictive performance is equivalent to the performance of optimally pruned trees. These trees are called conditional inference trees.

The algorithm can be briefly outlined as testing the null hypothesis of independence between the

response variable and each covariate. If the null hypothesis cannot be rejected, stop the algorithm. Otherwise, the covariate with the strongest association to the response variable **Y** is selected. Then, the best binary split for this variable is determined and the data is divided accordingly. These steps are repeated until the null hypothesis of independence cannot be rejected.

The presented independence tests are in fact fundamental to understand this procedure. The null hypothesis is tested using a linear test statistic of the form:

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = vec\left(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(\mathbf{Y}_i, (\mathbf{Y}_1, ..., \mathbf{Y}_n))^T\right) \in \mathbb{R}^{p_j q}$$

where $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is a non-random transformation of the covariate X_j, $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ is the influence function, which may depend on the responses $(\mathbf{Y}_1, ..., \mathbf{Y}_n)$ in a permutation symmetric way, and the vec operator transforms the $p_j \times q$ matrix into a $p_j q$ column vector by column-wise combination.

For a univariate numeric response **Y**, the most natural influence function is the identity ($h(\mathbf{Y}_i) = \mathbf{Y}_i$). However, when it comes to handling censored information, other functions can be chosen, such as the log-rank score. For right-censored information, subjects can be represented in the form of a triple $(T_i, \delta_i, \mathbf{X}_i)$, $i = 1, ..., n$, whilst the response variable is $\mathbf{Y}_i = (T_i, \delta_i)$. Thus, the influence function for such bivariate response is the log-rank score, defined as:

$$h(\mathbf{Y}_i) = U_i = \delta_i - \sum_{j=1}^{r_j(t)} \frac{\delta_j}{n - r_j(t) + 1}$$

where $r_j(t) = \sum_{i=1}^{n} I(T_i \leq T_j)$ is the number of observations which had the event, or were censored before or at time T_j. The main goal of this score is to assign a univariate value U_i to this bivariate response $\mathbf{Y}_i = (T_i, \delta_i)$ in order to allow the algorithm to proceed the same way as in the univariate numeric response case. Hence, extending this methodology to LTRC data can be done by building a log-rank score which transforms the response triple into a scalar.

Having into account that $\hat{S}(\infty) = 0$, the authors in [6] show that the log-rank score for an LTRC observation is given by:

$$U_i = \begin{cases} 1 + \log \hat{S}(R_i) - \log \hat{S}(L_i), & \text{if } \delta_i = 1 \\ \log \hat{S}(R_i) - \log \hat{S}(L_i), & \text{if } \delta_i = 0 \end{cases}$$

where $\hat{S}$ is the non-parametric maximum likelihood estimator of the survival function.

Each terminal node of this tree displays the fitted Kaplan-Meier curve for the observations which belong to that node and the median survival time, which will be used as the predicted response.

Henceforth, the extended LTRC tree, using this framework, is going to be referred to as LTRCIT (LTRC tree based on Conditional Inference Trees).

3.2. Relative Risk Trees

In [11], the authors developed a method for obtaining tree-structure relative risk estimates for censored survival data, based on the assumption of proportional hazard. Despite adopting most aspects of the widely used CART algorithm, the proposed model uses the one-step full likelihood strategy, which has the advantage of being based on the popular PHM.

For a tree B, the full likelihood of the learning sample can be expressed as:

$$L = \prod_{h \in \tilde{B}} \prod_{i \in S_h} \lambda_h(t_i)^{\delta_i} \exp^{-\Lambda_h(t_i)} \quad (1)$$

where $\tilde{B}$ is the set of terminal nodes, S_h is the set of labels $\{i : x_i \in \mathcal{X}_h\}$ for observations belonging to the region $\mathcal{X}_h$ to which corresponds node h, $\lambda_h(t)$ is the hazard function, and $\Lambda_h(t)$ is the cumulative hazard function for node h. If the PHM $\lambda_h(t) = \theta_h \lambda_0(t)$ is assumed to be true, where θ_h is the non-negative relative risk of node h and $\lambda_0(t)$ is the baseline hazard, then replacing this term in Equation 1 gives:

$$L = \prod_{h \in \tilde{B}} \prod_{i \in S_h} (\lambda_0(t_i)\theta_h)^{\delta_i} \exp^{-\Lambda_0(t_i)\theta_h}$$

where $\Lambda_0(t)$ is the baseline cumulative hazard function. If this value is known, then the maximum likelihood estimates of $\{\theta_h : h \in \tilde{B}\}$ are simply:

$$\tilde{\theta}_h = \frac{\sum_{i \in S_h} \delta_i}{\sum_{i \in S_h} \Lambda_0(t_i)}.$$

However, in practice, the cumulative hazard is not known and therefore has to be estimated. One alternative, proposed by [11], is to use the Nelson-Aalen estimator for $\Lambda_0(t)$. Each terminal node of this tree displays the relative risk on that node, although it is also possible to obtain the Kaplan-Meier curves on each terminal node, obtaining this way a comparable response with LTRCIT. This relative risk is simply given by $\tilde{\theta}_h$, and it is relative to the estimated baseline hazard.

A key feature in the proposed algorithm is the fact that a relative risk tree can be grown using the CART methodology. In fact, a tree is grown using recursive partitioning methods, which split the data and the covariate space into regions that maximises the reduction of the node deviance residual. The splitting procedure continues until a large

tree is obtained. Lastly, just like is done in CART, a pruning algorithm is applied in order to find the optimally pruned subtrees.

The adaptation to LTRC data is executed by transforming the logarithm of the full likelihood function presented in Equation 1. Extending the log-likelihood for right-censored data to data of the form $(L_i, R_i, \delta_i, \mathbf{X}_i)$, $i = 1, ..., n$ can be easily done using:

$$\log L = \sum_{i=1}^{n} [\delta_i \log \lambda(R_i) - (\Lambda(R_i) - \Lambda(L_i))]$$
$$= \sum_{i=1}^{n} \left[\delta_i \log \lambda(R_i) - \int_{L_i}^{R_i} \lambda(z) dz \right]$$

Fu and Simonoff [6] show that is enough to do such transformation to extend the survival tree of [11] to this special data. This extended survival tree is going to be referred to as LTRCART (LTRC tree based on CART).

4. Survival Tree Ensemble

The basic idea of ensemble learning methods is to build prediction models by combining the strengths of a collection of simpler base learners, also called weak learners. Even though the presented survival trees provide valuable predictions and have proved to appropriately deal with the specificities of the failure data, the truth is that trees can be very unstable. In fact, small perturbations in the training data can produce significant changes in their predictive function. Ensembles of survival trees emerge as an attractive approach to handle such problems.

Published attempts to use ensemble methods in survival settings is limited due to the difficulties to deal with censored information. To our knowledge, there is no publicly ensemble methodology of survival trees adapted to LTRC data and time-varying variables. Hence, in this section a new technique of ensemble of survival trees is proposed, based on the presented survival trees (LTRCIT and LTRCART), which is capable of dealing with those difficulties. Such method will be called STE (Survival Tree Ensemble).

The algorithm for this methodology can be written as:

1. Define training and test sets according to temporal division.
2. Choose randomly 80% of the pipes from the training data (and corresponding instances), and randomly select `mtry` variables from the available variables to split the data.
3. Grow a tree for the sampled data set, either using a LTRCIT tree or a LTRCART tree and, using the grown tree, predict for the test data.
4. Repeat steps 2 and 3 until `ntree` trees are built.
5. For each pipe in the test set, compute an ensemble response using the available approaches.

The desired response obtained by each tree method is the time (in days) until the next failure. However, obtained responses may be infinite and, for that reason, a special care has to be taken when formulating a procedure to combine the responses obtained from the trees used in the ensemble. Several options were experimented, but only two reveal themselves better, either intuitively and in terms of the predictive performance results.

First Quartile

An approach that immediately stands out is the minimum predictive response. However, there is the possibility that this value does not appropriately characterise the history of the corresponding pipes. It could simply be a lucky strike and considering such value as the final predicted value could compromise the final accuracy of the model. Hence, an alternative approach could be using the first quartile, since instead of considering exactly the minimum value of the responses, it considers the middle value between the smallest response and the median.

Reduced Mean

Let $\{v_{i,1}, ..., v_{i,\texttt{ntree}}\}$ be the obtained responses for pipe i by each one of the `ntree` trees and $\hat{y}_i$ be the final and combined predicted response for pipe i. The reduced mean also arises as an alternative to considering the classical mean. The later is very unstable, especially when infinite values are allowed. This reformulated mean can be easily defined as ignoring all the infinite values of the predicted sample for a certain pipe, and calculating the mean with the remaining values.

$$\hat{y}_i = \frac{\sum_{j=1}^{\texttt{ntree}} v_{i,j} I(v_{i,j} \neq \text{Inf})}{\sum_{j=1}^{\texttt{ntree}} I(v_{i,j} \neq \text{Inf})}$$

where $I(v_{i,j} \neq \text{Inf}) = 1$ if $v_{i,j} \neq \text{Inf}$ and $I(v_{i,j} \neq \text{Inf}) = 0$ if $v_{i,j} = \text{Inf}$. In fact, for a certain pipe, averaging over the non-infinitive values can give an idea of how trees, which are capable of obtaining a numerical response, actually behave, without their expected value being skewed by the infinite values.

5. Variable Importance

One important question that often arises when doing such analysis is which variables are in fact the most relevant. Several variable selection methods are already available and a very popular one, in the context of random forest, is commonly known nowadays as VIMP (Variable Importance). It was initially defined in CART by [3]. The original definition refers that VIMP corresponds to the increase

(or decrease) in the prediction error when a variable is permuted while predicting on test data. Such permutation of variables is called as noising it up.

It was already shown in the literature that VIMP can be a very effective approach in various fields. However, several limitations restrict the ability to develop a more general methodology based on such approach. Its dependency on the prediction error used and its arduous difficulty to define it theoretically lead to the search of alternative measures. It is in this context that maximal subtrees and minimal depth appear.

Intuitively, the closer the variables are to the root, the stronger is their effect on the prediction accuracy. Similarly, variables which split farther down the tree, have less impact. Using this idea as a basis, it is possible to define a structure such that the importance of variables is obtained through their positioning in the tree. This idea was formalised by [8] and it originated the concept of maximal subtrees.

Definition 5.1 (Maximal Subtree)**.** *For each variable v, T_v is a v-subtree of tree T if the root node of T_v is split using v. Then, T_v is a maximal v-subtree if T_v is not a subtree of a larger v-subtree.*

As pointed by [9], maximal subtrees permit to peer inside what can be called as the ”black-box” of VIMP. In fact, according to the authors, there are strong reasons suggesting that maximal subtrees can be used in addition to, or even instead of VIMP. The proposed approach assumes that variables which have a great impact on the prediction are the ones that most frequently split the nodes closest to the root, since it is near the root where trees partition large samples of the data. This is where the concept of minimal depth arises, which simply corresponds to the distance between the root node and the closest maximal v-subtree. That is, given the depth of the root node of all maximal v-subtrees, minimal depth is the minimum of those values.

For a certain variable, minimal depth measures how an observation travels down the tree until encountering the first split on that variable, allowing to have an idea of its predictiveness. Thus, the smaller the minimal depth, the greater is its impact on the prediction, and consequently, the larger is its relative importance. When it comes to ensemble techniques, the minimal depth approach measures which risk factors are the most relevant, by averaging, for each variable, the depth of their first split over all trees used in the ensemble.

6. Performance Metrics

Due to the presence of censored information, traditional evaluation metrics are not the most suitable for measuring the performance of survival models. Survival analysis requires more specialised measures which can accommodate its specificities. The main form of comparison between the built models and LEYP is based on a score obtained through the Baseform Failure Prediction App [1].

Currently, LEYP is one of the best approaches in the field of pipe failure analysis. It was originally proposed by [10], and it is a multivariate counting process capable of obtaining a failure rate and a probability estimation for each individual pipe along time. Baseform implementation of LEYP is based on the definition of [12], and allows an arbitrary number of covariates besides the most usual ones (diameter and length), as well as, the inclusion of categorical variables.

A performance measure capable of detecting which pipes have higher failure risks is, then, possible to define. To do that, a predictive performance curve is built to measure how actual failures concentrate on pipes with higher expected failure rates.

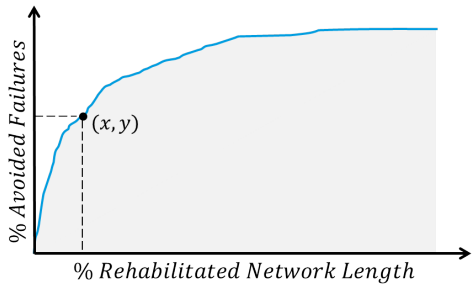


Figure 1: Predictive performance curve, where a certain point (x, y) indicates that rehabilitating $x\%$ of the water distribution system would avoid $y\%$ of all observed failures.

It is this kind of curve which allows to define the Baseform Score (hereafter Score), and despite its simple designation it is a major model assessment measure capable of transforming the cumulative gain curve into a very informative value. The Score corresponds to the area under the predictive curve, that is, to the area of the grey region in Figure 1.

Models' ability to prioritise pipes according to their likelihood of failure can, in fact, be accomplished by comparing the number of failures that might be avoided by rehabilitating a certain percentage of the water network. A good failure prediction model is obtained when rehabilitating only a small percentage of the system leads to a large percentage of avoided failures. Hence, such performance metric provides water entities with the possibility to compare and measure the success of implemented prioritisation policies. First percentiles are therefore the most important ones, since when planning those policies only a small percentage of the entire network is considered, due to limited budgets.

In the software, there are several prioritisation

strategies available:

- **Baseform prediction:** pipes are sorted according to the estimated failure rate by LEYP;
- **Random selection:** pipes are randomly ranked;
- **Pipes with most failures:** pipes are sorted according to the number of failures each pipe had in the previous years;
- **Older pipes:** pipes are arranged by age, i.e. pipes with the earliest installation date are prioritised.

Besides these strategies, it is still possible to include extra numerical variables that allow the assortment (increasing or decreasing) of pipes. This will be particularly useful to compare these four approaches with the implemented models in this work. Therefore, this will be the privileged performance measure since it allows the best comparison between the experimented models and LEYP.

Other possible measures could be used. The Brier Score or the Concordance Index are equally common metrics to evaluate the performance of survival analysis problems.

7. Exploratory Data Analysis

Before tackling any kind of failure prediction results and in order to build good prediction models, it is fundamental to understand what type of data is on the stand and to comprehend its specificities by doing a good exploratory data analysis. All data used was kindly provided by Baseform [1], as well as the expert knowledge about water distribution systems essential to this work. Notice that failures in this sense correspond to the registered work orders during the observation period.

7.1. Dataset A

The first dataset was provided by a large Portuguese utility, and is characterised as follows:

- **Number of pipes:** 12439
- **Total length:** 383.14km
- **Number of work orders:** 15366
- **Number of pipes with no failures:** 6186 (49.7%)
- **Total length of pipes with no failures:** 282.22km (73.7%)
- **Observation period:** 2000/01/08 to 2019/01/24

Some of the available variables for the asset data (besides the mandatory ones) include the number of connections (a water service connection corresponds to the service pipe that goes from the public water main to the service line which enters the customer's property), distance to the nearest railways (which for this area includes a train and a metro line), type of road where the pipe is located, soil's type, and finally soil pH range and type.

7.2. Dataset B

The second dataset to be analysed was also provided by Baseform [1]. Immediate distinctions between the two datasets involve the dimension. The corresponding network for dataset B is much larger than the one for dataset A, whereas the observation period is much smaller. Its characterisation is given by:

- **Number of pipes:** 49688
- **Total length:** 3331.02km
- **Number of work orders:** 3213
- **Number of pipes with no failures:** 47383 (95.4%)
- **Total length of pipes with no failures:** 2912.86km (87.4%)
- **Observation period:** 2013/03/25 to 2018/12/26

Some of the available variables include again the number of connections and distance to railways (and type of railways), as well as other attributes such as depth of the assets and more geological variables (lithology and permeability). In contrast with dataset A, there is an indication of which pipes were already removed, abandoned or decommissioned. This is an information very well appreciated, since it allows to evaluate the impact of having the history of abandoned pipes also available. Two possible situations emerge to explore, and they are: (i) remove the abandoned pipes, and treat the data as in dataset A; or (ii) keep the abandoned pipes and extend the explored methods for this special case.

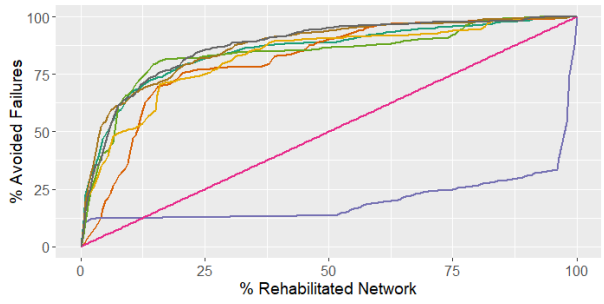
8. Results and Discussion

This section will now present some results and findings made during this work. The presented failure prediction models were fitted to the available failure data, and then evaluated in terms of the quality of their predictions. To apply the proposed models, the dataset was divided into training and test samples according to a temporal division. The training sample will be composed of all the failures occurred before a time cut point in all pipes installed before this date and, analogously, the test sample will be composed of all failures which occurred after that point in all pipes considered in the training data. To assess the quality of the obtained predictions for each method, the Baseform Score and the cumulative gain curve, as presented in 6, will be used. This method has the ability not only to prioritise pipes according to their likelihood of failure, but also it gives the best means of comparison between the experimented models and the implemented model (LEYP) at Baseform. The results will be presented for the datasets discussed in Section 7.

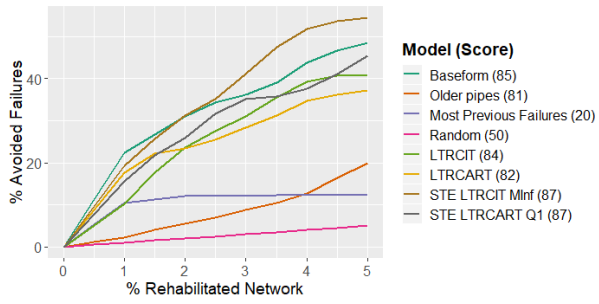
8.1. Dataset A

Several models were experimented and playing with different combinations of variables and methods allowed to obtain the best approach for each technique presented above. Therefore, a LTRCIT and LTRCART (whose output is transformed in the Kaplan-Meier curve) model with the same variables as LEYP - material as a grouping attribute, diameter, length, age and number of previous failures - and the reduced mean for the LTRCIT and the first quartile for LTRCART ensembles (with 100 trees each) will be compared with the ranking strategies offered by Baseform Failure Prediction App [1] (Figure 2).

Immediately looking at Figure 2a, the approach which prioritises pipes with the most previous failures stands out for its behaviour. The smaller Score corroborates the fact that this approach is worse than sorting pipes randomly. The remaining strategies (also excluding the random one) seem to have a similar behaviour from a more ample perspective. Examining the first percentiles (Figure 2b), one can see that the older pipes strategy is one of the worst, while the most previous failures strategy behaves similarly to the LTRCIT model for the first percentile.



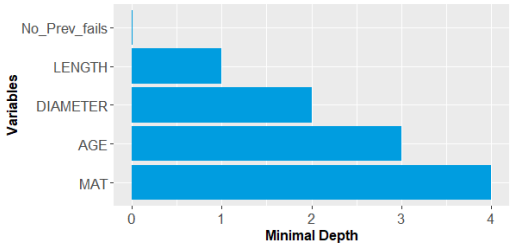
(a) All percentiles.



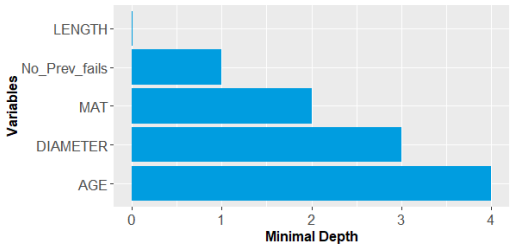
(b) Zoom 5%.

Figure 2: Comparison between all the best approaches (dataset A).

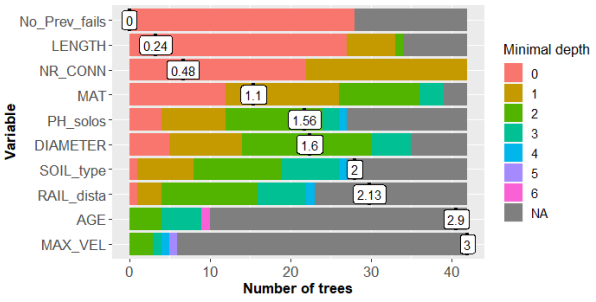
Analysing the minimal depth results obtained by each model presented above, one obtains Figure 3. Recall that the smaller the minimal depth for a certain variable, the larger is their importance for the analysis. Notice that for the ensemble methods (Figures 3c and 3d), the average minimal depth for all variables is represented. That is, the values associated with each variable correspond to the mean of its minimal depths over the trees, and the colour scale indicates the number of trees that variable splits at a certain depth. Also the x-axis ranges from zero to the maximum number of trees in which any variable is used for splitting.



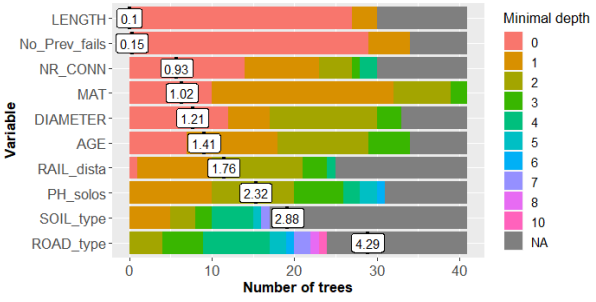
(a) LTRCIT.



(b) LTRCART.



(c) LTRCIT ensemble.



(d) LTRCART ensemble.

Figure 3: Minimal Depth for tree-based models for dataset A.

Several conclusions can be drawn from the analysis of minimal depth, and from Figure 3a, number of previous failures and length seem to be the most important variables, since they present the smallest minimal depth among all variables used. The same conclusions are drawn from Figure 3b, despite length overcomes the number of previous failures. This suggests that length and number of previous failures are in fact the most relevant to

explain the occurrence of failures in water distributions systems. As for the ensemble methodology, besides the mentioned variables, the number of connections arises curiously as a relevant variable as well. As for Figure 3d, which corresponds to the LTRCART ensemble, the material arises as an important variable, whereas the number of connections appears more down in the ranking.

To better examine how each experimented model behaves in the first percentiles Table 1 displays the percentage of avoided failures for fixed percentages of rehabilitated length:

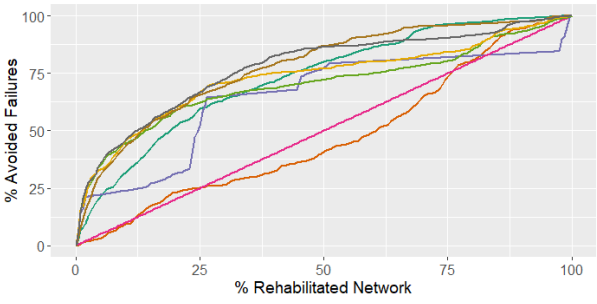
Table 1: Percentage of avoided failures per percentage of rehabilitated length.

	Rehabilitated Length (%)			
Models	0.5	1	2	5
Baseform	**11.2%**	**22.5%**	30.9%	48.5%
LTRCIT	5.2%	10.3%	23.5%	40.9%
LTRCART	8.8%	17.5%	23.4%	37.2%
STE LTRCIT	9.7%	19.4%	**31.2%**	**54.5%**
STE LTRCART	7.8%	15.5%	25.8%	45.4%

From Table 1, one sees that despite not achieving the best result for the first two percentiles, above that the ensemble methods start to yield very good predictions.

8.2. Dataset B

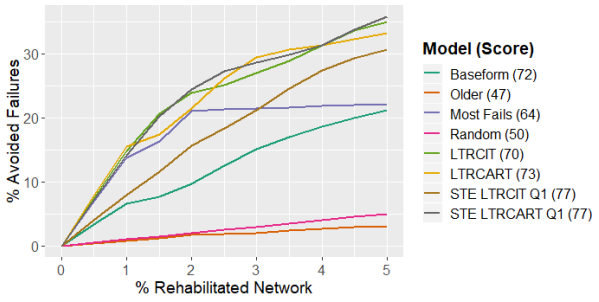
Results for dataset B can be obtained in a similar way as for dataset A. For that reason, only the best found models will be again displayed. Besides the ranking strategies offered by Baseform software, it will be compared a LTRCIT and LTRCART models using all available covariates, and an ensemble for each type of tree methodology (with 20 trees and using the first quartile strategy to combine results). The results will be compared in a dichotomous perspective, corresponding to two analyses: $A1$ refers to the dataset where deactivated pipes were removed and $A2$ consists of the dataset where deactivated pipes are still considered.



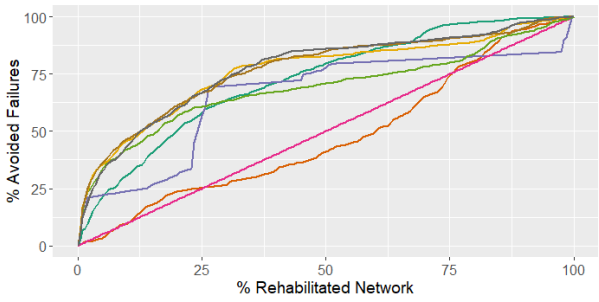
(a) A1: All percentiles.

Figure 4: Comparison between all the best approaches (dataset B).

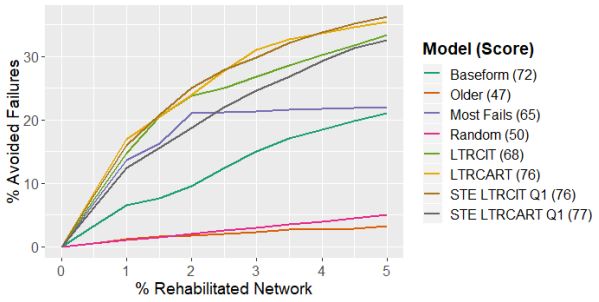
At first sight, both analyses seem to produce similar results. The strategy where pipes are ranked according to age immediately stands out for its predictive performance very similar to the random one. The poor estimation of the age for the assets in this network may explain why such behaviour is observed. The strategy where pipes are ordered according to the number of failures also produces some mixed results. Even though for the first percentiles its predictive ability is very close to other approaches, its behaviour almost resembles a step function (something similar is actually verified for dataset A). Regarding the remaining models, from a global perspective all of them seem to overtake LEYP in the first half of the rehabilitated length, whereas this later starts to gain ground, especially in analysis $A2$ (Figure 4c). The first percentiles allow to take a peek on how these methods behave in the most crucial percentiles. In Figure 4b, models such as the LTRCIT, the LTRCART and the ensemble of relative risk based-trees have very similar performance curves, while the ensemble of LTRCIT is a little behind and it is not capable of avoiding that many failures as the referred models above. Despite that, the global Scores are much better for the ensemble methodologies than



(b) A1:Zoom of 5%.



(c) A2: All percentiles.



(d) A2: Zoom of 5%.

Figure 4: Comparison between all the best approaches (dataset B) - cont.

for the individual trees and even for the Baseform model. On the contrary, in Figure 4d, all the models achieve similar results, while the ensemble of LTRCART is the one a little behind. However, its global Score is the largest for all the displayed models.

Minimal depth results are very similar to both analyses. Just like it happened for dataset A, the most relevant variables to the study seem to be length and number of previous failures. Also the number of connections is a strong candidate, maybe because of its strong correlation with length. Moreover, age usually appears in the bottom half of the ranking, which can be a consequence of its problematic estimation.

Table 2: Percentage of avoided failures per percentage of rehabilitated length (dataset B).

A1: Removing abandoned pipes

	Rehabilitated Length (%)			
Models	0.5	1	2	5
Baseform	3.3%	6.6%	9.6%	21.2%
LTRCIT	**7.4%**	14.8%	23.9%	34.9%
LTRCART	7.7%	**15.5%**	21.5%	33.2%
STE LTRCIT	4.0%	8.0%	15.6%	30.7%
STE LTRCART	7.1%	14.1%	**24.5%**	**35.8%**

A2: Keeping abandoned pipes

	Rehabilitated Length (%)			
Models	0.5	1	2	5
Baseform	3.3%	6.6%	9.6%	21.0%
LTRCIT	7.4%	14.7%	23.8%	33.4%
LTRCART	**8.4%**	**16.9%**	23.9%	35.5%
STE LTRCIT	8.0%	16.0%	**25.0%**	**36.3%**
STE LTRCART	6.2%	12.4%	18.7%	32.5%

The conclusions taken from Table 2 only corroborate the observations already made from Figure 4. In fact, and in contrast to what happened for dataset A, the proposed models have a largest predictive ability than LEYP and for the crucial percentiles these models are able to avoid more failures than the already implemented approach.

Comparing both analyses, it seems that their results yield very similar outcomes. Nevertheless, for the majority of the results, it was noticed that when considering all pipes (including the abandoned ones) the models were not capable of prioritising pipes in the final percentages of the ranking strategies plots, giving some improvement margin to LEYP. This is, indeed, surprising since one would expect that having such information would benefit the adjusted models.

9. Conclusions

The importance of failure prediction has been established throughout this work, regarding either planning and prioritisation processes. The goal is to predict the time until next failure and the given data includes information on the assets and on the maintenance records undertaken in the network. Survival analysis is the field of statistics which offers the tools to tackle time-to-event data, and combined with the simplicity and developments in machine learning turn this into a powerful technique to apply to failure prediction.

Failure data has some specificities which distinguish it from more common types of data, such as being left-truncated and right-censored and having a time-varying covariate (number of previous failures). Such specificities call for more specific models as well, which can accommodate them. It is in this perspective that survival trees arise. An ensemble approach was also tested, as a way to combat the potential instability caused by trees and to improve their predictive performance.

Even though the proposed models were not capable of overtaking LEYP for all percentiles, in general they were able to yield very good performances being able to avoid a larger number of failures than the Baseform model for a large part of the rehabilitated length.

References

[1] Baseform. BF Software, Lda. `http://baseform.com`, 2019. [Online; accessed July-2019].

[2] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

[3] L. Breiman, J. Friedman, C. Stone, and R. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[4] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[5] W. Fu and J. Simonoff. *LTRCtrees: Survival Trees to Fit Left-Truncated and Right-Censored and Interval-Censored Survival Data*, 2018. R package version 1.1.0.

[6] W. Fu and J. S. Simonoff. Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. *Biostatistics*, 18(2):352–369, 2017.

[7] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Stat*, 15(3):651–674, 2006.

[8] H. Ishwaran and U. B. Kogalur. Random survival forests for R. *Rnews*, 7:25–31, 2007.

[9] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

[10] Y. Le Gat. *Une extension du processus de Yule pour la modélisation stochastique des événements récurrents. Application aux défaillances de canalisations d'eau sous pression*. PhD thesis, Cemagref Bordeaux, Paristech, 2009.

[11] M. Leblanc and J. Crowley. Relative risk trees for censored survival data. *Biometrics*, 48(2):411–25, 1992.

[12] A. Martins. Stochastic Models for Prediction of Pipe Failures in Water Supply Systems. Master's thesis, Instituto Superior Técnico, University of Lisbon, 2011.

[13] P. Wang, Y. Li, and C. K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6):110:1–110:36, 2019.