UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Gestionale
A.a. 2021/2022
Sessione di Laurea Marzo 2022

Tesi di Laurea Magistrale

A new machine learning approach to support asset management in water distribution networks

Relatori:
Cagliano Anna Corinna
Perez Magrane Ramon

Candidato:
Brucia Paolo Gioele

Acknowledgements

With the writing of this thesis, I conclude a wonderful journey that started more than 5 years ago when, maybe unaware of my crazy choice, I decided to include the word “engineering” deeply in my life, which drove me through three different countries and 4 different cities. This chapter has finally come to its end.

I am profoundly grateful to my two supervisors, Prof. Cagliano Anna Corinna and Prof. Perez Magrane Ramon, that successfully took me to the end of this work, always showing enthusiasm and positivity, even when I felt this feeling disappearing, because of tiredness and discouragement. They have been believing in me and in this work and this awareness gave me strength and determination to hold on. No other words rather than a big “thank you both”.

I want to thank my siblings, not the blood ones (not their turn yet): the crew of *Il Posto&Co.*, all my friends that have given meaning to my life, throughout all the unforgettable moments we have lived together, until this moment. No other way to reward you to be here rather than with all my love. Who more, who less, but all of them have done their part to shape what, so far, is the best part of me. Not the best… the best is yet to come.

I cannot forget all my colleagues in these years of university and all the people from all around the world I met that have given a different reading key of life and of the world. And my “mentors” and friends of my adventure in Barcelona, becoming my “little Spanish family”: it is also thanks to them that now I am, safe and sound, at the end of this Master.

A particular thanks go to a special person, lifejacket during the weakest moment of my life and partner during my happiest moments. It is thanks to this person that I further matured the belief that I can be a “winning horse”, as she uses to say: grazie Ale.

Last but not least, my wonderful family, indestructible pillar of my whole life, the balance of choice and light in my darkness, always supporting me in my run toward the attempt of “taking the flight”: people willing to give up their self-interest for other's happiness. I truly hope I made them as happy and proud as they did to me conveying their love to me.

This thesis not only represents the end of an intense and full learning path, but it also signs the starting point of an unknown and challenging journey, made of failures and successes, but of which the straight path has still to be discovered.

Abstract (English)

One of the main causes of the widespread problem of freshwater scarcity lies in unfruitful maintenance of distribution infrastructure, leading to failures with consequent waste of precious resources. It is estimated that more than 25% of the annual loss of water is due to poor conditions of the distribution networks and, in a scenario of continuously increasing demand for water, effects of such inefficiency might be even more dramatic, beyond the merely economic aspect. However, with the rise of data analysis, the awareness of the power of predictive technologies and machine learning techniques, the opportunity to make use of these tools to support decision making has become more than a hope.

With this study, the author attempts to address the problem of usage of historical data of pipes and their failures in the Spanish city of Manresa to deduce conclusions on how to conduct maintenance interventions. After conducting an explorative study on how pipes intrinsic factors may have reflections on breakages, machine learning algorithms (Logistic Regression and Random Forest have been chosen in this thesis) are used to predict pipe failures over time. Lately, results from predictions will be used to take out conclusions from two different assessment models. The first method, given the structure of cost of a general distribution company, tries to establish the optimal ratio between sensitivity and sensibility of a predictive model to return the best economic benefit from the predictive maintenance. The second approach wants to assess how the uptime of the service level can be improved whether relying on prediction to replace pipes, given a certain agreed investment budget.

In an old industry such as water distribution, difficulties come up not only during the development of predictive models but also during the reconstruction of the data on which training and testing models, since they can suffer from inconsistencies. Indeed, data gathering has not unique and standardized methodologies and time and people take-over have changed procedures during the data collection, making the whole work harder.

Table of contents

List of Tables

List of figures

List of equations

List of optimization systems

1 Introduction

The first chapter of this study aims to build the structure of the whole thesis, outline the background behind the issue of freshwater resources and the reason why management of water distribution network has become an important topic in terms of sustainability but also economically talking. Many are the challenges affecting the distribution of freshwater, starting from scarcity of pure sources because of climate change, until the continuously increasing demand of water due to inexorable growth of Mondial population.

Among all the efforts to deal with the critical situation, effects can be immediate and surely goal-oriented thanks to improvement in the management of distribution networks, which, because of low investments for maintenance, suffer from serious problems of leaks and breakages.

Over time, studies have tried to identify deteriorating paths and causes of bursts to prevent waste of water. The rise of the word of data analysis has given hope even to the topic in the matter, with the application of new tools attempting to give a solution to tough issues.

1.1 Background

European water demand has been rising over the last 50 years, mainly due to an always increasing population. This scenario has been responsible for a reduction of around 24% of renewable freshwater resources (such as rivers, groundwater or lakes) per capita in the Old Continent [1]. An increase in drought phenomenon in Europe, especially in countries such as Italy and Spain, climate change and global warming threaten the already scarce availability of water, worsening an already worrying and risky situation.

According to a report released by United Nations World Development Report, by 2050, 6 billion people will suffer from real water scarcity, keeping the current usage trends, and this prediction could even be an underestimation (WWAP, 2018). The three main driving causes are increases in water demand, reduction in water sources and an always higher level of pollution of freshwaters. Actually, the world would be able to face demand increases, but only with massive changes in the way water is used, managed and shared. Regulation of population and economic increase rate, as well policies and rules to reduce pollution of water sources are urgent measures that need to be undertaken to preliminarily face the water shortage threat.

It is estimated that between 25% and 50% of all annual globally distributed water is lost due to inefficient water distribution networks and poor condition of infrastructure [2].

Going beyond mere business aspects and the consequences of this inefficiency, there is an additional aspect that, over the last decades, has been catching the attention and interest of all authorities around the world. We are talking about the impact that water waste can

have on the environment and the sustainability of the distribution process. Data on water availability and its effects over the year are dramatically discouraging, with almost half of the world's wetlands disappeared since 1900 and damaging ecosystems [3].
Infrastructure deterioration, inaccurate water pressure management and limited budget for maintenance are some of the causes leading to low performances of water distribution networks. Among them, leakages, due to pipe breaks, can be charged to be responsible for above 80% of all the lost water [2].
What is a break? "A break is a rupture of the line causing a cessation of service" and the reasons for breaks are grouped into four classes (Clark R. M., 1982):

1. Quality and age of pipe itself.
2. The environment where the pipe is located.
3. Quality of workmanship in laying the pipe.
4. Service conditions, such as pressure.

The phenomenon of water main burst has become a very frequent problem in pipes management and some study says that the frequency of breaks has gone up by over 27% over the last six years. (Folkman, 2018). Increases in pipe breaks are problematic either because they increase repair costs, interrupt services provided to customers and also potentially impact water quality.

In one of his articles, Raziyeh Farmani, Associate Professor of water engineering at University of Exeter and Chair of Intermittent Water Supply Specialist Group, defines a failure as "a cumulative effect of various pipe-intrinsic (such as material, diameter, and age), operational (such as corrosion, pressure, external stresses) and environmental factors (such as temperature, rainfall, soil conditions) acting on mains" (Farmani, Kakoudakisb, Behzadianc, & Butlerd, 2017). Additionally, environmental and intrinsic factors can either be static or dynamic, while operational factors belong only to the dynamic group (Farmani, Kakoudakisb, Behzadianc, & Butlerd, 2017).
Basically, there is a widespread hypothesis assumption that pipes sharing the same intrinsic properties are expected to have the same breakage pattern. Actually, even pipes absolutely equal can react differently to external dynamic factors, making the previous assumption unrealistic. However, it is unreasonable to carry predictive studies on the breakage behaviour of every single pipe because not enough data could be gathered for each tube object of study. Statistical analysis with such a low-sampled methodology would result to be not significant and powerful (Kleiner & Rajani, 2012).

In a total water supply system, the distribution network represents a big share of the total expenditure, up to 80% and its management and maintenance is crucial for optimal functionality. Indeed, as water mains deteriorate into the network, the breakage rate increase, the reliability of the service and hydraulic capacity decrease and the quality of water will be negatively impacted. Therefore, asset management, including replacement and repair strategies, is crucial for providing optimal service to consumers and for reaching good cost-effective decision-making, since capitals are always scarce and limited (Kleiner & Rajani, 2001).

Improvement in data collection and the introduction of the concept of *data mining* for a better understanding and use of data have represented an important step forward in the world of water distribution network management. One of the principal ways to monitor water flow conditions within a distribution network is the pressure sensor system, enabling it to operate safely. Thanks to the use of pressure sensors alongside pipes and all information gathered to know in real-time events occurring at a depth of meters. In fact, while pipe bursts can be easily identified by civils if the water reaches the ground, detecting leakages, but also burst occurring in not visible or easily reachable places, with the aid of pressure sensors can be extremely easier and quicker (Qi, et al., 2018).

Identification of leaks and bursts, the monitoring of pipes evolution and the collection of huge amounts of data over time have enabled the application of machine learning technology and artificial intelligence to predict future events and anomalies. However, already developed models present many forms of inaccuracies, mainly coming from the fact that these models are built relying on relatively little data availability, for lack of historical awareness of collecting data, which leads to incomplete knowledge.

1.2 Origins of this work

Aigües Manresa is the municipal company managing the water supply cycle for around 20 towns, mainly belonging to the *"Bages comarca",* in the center of Catalunya.

As known, Spain suffers from dramatic water supply losses, as almost a quarter of all distributed water does not reach households, as it goes lost because of breakages and leaks for around the 60% of the cases [4]. Besides the real water loss, part of the water is not recorded due to *"faked losses",* due to mistakes that occurred during collection and handling of data. The main cause is surely the lack of investments, either of distribution system, data detection technologies or data utilization [5].

As almost the totality of company of any industry, also *Aigües Manresa* has long started the gathering of data through sensors and other technologies, to monitor daily water behaviour alongside buried pipes. Over years, the company has installed alongside their pipes and upstream tanks sensors to detect water pressure and tanks' water level and outflow, and these tools represent the main sources of data for the company, besides the self-reading of water meters by consumers. They have strengthened their data analysis department to take advantage of all the data that have been collected over years to make predictions. Outcomes coming from the department would represent support for decision-making processes and solid help for developing better maintenance strategies with mainly three gains. First of all, starting from the most considerable aspect, reducing losses due to breaks represents an important step forward more solid sustainability of water distribution process, reducing environmental impacts. Secondly, as some places suffer from shortage of hydro-sources during the year, the pumping capacity of electromechanical equipment is not enough for ensuring a supply service appropriate to quality standards. Last but not least, data understanding and usage may be of use to economic results.

However, sorting, arranging, combining and using a large amount of data could be challenging work that requires time and resources. Also, the department of data analysis has complained about asynchronism between data they collect, which makes all the jobs more complex and long. For this reason, thanks to the intermediation of Professor Peréz Magrané Ramon, the company got in touch with the author of this thesis proposing the application of *machine learning* knowledge for making predictions about future water main breakages and, eventually, attempting to assess how ML can improve company's financial performances.

Although machine learning is widespread and with wide application in any field, according to what was said from the data analysis department, in *Aigües Manresa* it has not passed the innovation department yet and therefore the company has not exploited the power of these technologies yet. Therefore, the use of machine learning methods for predicting future

pipe failures to support the decision-making process would represent a concrete innovation, bringing also new insight to the company management.

2 Literature review on water main management

Watermain management has always been object of interest for engineers. Losses in distribution network efficiency and their impact on economic performances have pushed people to carry out studies attempting to give a contribution to the current knowledge and *state of the art* over years.

The present chapter is a summary of some of the main literature that was analyzed to outline the current state of the art for water main management. The reader will have a general but concise idea about how the topic of this thesis has been undertaken over the last decades and how it evolved thanks to research.

The methodology of researching these sources comes from selecting papers, articles, reports found on databases such as Scopus [6], ScienceDirect [7], Google Scholar [8]. These platforms have been chosen for their wide gamma of resources, both in terms of topics covered but also for the time of publication. In some cases, reports, conference minutes or presentation has been found directly on specific authorities' websites. Although it was established to refer basically to sources of the last 10-15 years, some articles cited in the review are dated back to the early twenties and even before. In fact, although a massive effort of researchers to improve technologies in water main management, some subsequently methodologies still rely on such dated back studies.

Watermain management is a very wide topic, including basically three main sub-topics:

- Maintenance strategies.
- Predictive analysis.
- Economic assessment.

Usually, authors, in their studies, focus on only one of these three aspects, which turns out to be the main theme of the research, with only a few references to the other two aspects. Only recently, it was increasingly common to have papers jointly dealing with all three aspects of water mains management.

In this literature review, we are going to outline, for each of these subtopics, how knowledge has been evolved over time and how we have reached the current state of the art.

2.1 Maintenance strategies

Various decision-making strategies have been developed to deal with water pipe deterioration, to find the optimal, in terms of reliability and costs, sequence of reparation and replacement. All strategies can be grouped into two main families: reactive strategies and proactive strategies.

Reactive strategies are still the most used worldwide and they refer to those maintenance interventions that take place only once a break occurs (Canadian Water Network, 2018). Reactive strategies do not attempt to predict *ex-ante* future water main conditions; therefore, these strategies are not recommended to make long-term maintenance plans. Usually, a pipe is replaced after it has experienced a certain number of breaks, followed by a corresponding number of repairing interventions. Generally, reactive strategies are considered more money-consuming and therefore less cost-efficient. In fact, without a proper study regarding the expected life of a pipe, a company risks allocating capital and time to repair a break that can be avoided because predicted. In other cases, the repair was not necessary because the broken pipe had already reached its useful life and replacement would be the best option.

Proactive strategies work trying to predict and estimate future conditions of water pipes to develop well-structured long-term maintenance plans, with the right allocation of capital. Whether for reactive strategies, the resources and capital are used *ex-post* for repairing or replacing, in proactive strategies a strong effort must be done before using or developing predictive models to identify the optimal time for replacing a pipe. In fact, if a pipe is replaced before its optimal time, a company does not exploit fully the efficient service life of the main, using now resources that can be postponed. On the other hand, the utility company can wrongly keep on repairing and spending money in maintaining in service a pipe that has overcome its useful and cost-efficient life. However, the use of models for making predictions does not ensure optimal identification of the correct replacement time, because these models present often inaccuracy (Snider & McBean, 2021). One of the main sources of inaccuracy comes from directly the fact that these models are built relying on relatively little data availability, for lack of historical awareness of collecting data, which leads to incomplete knowledge. Regarding this last point, it has not to be forgotten as these data come from pipes buried at a depth of several meters. Therefore, mechanisms leading to pipe failures, besides being complex, are not fully explained and understood (Kleiner & Rajani, 2001).

However, although the results of predictive models could not be perfectly accurate, they help utilities to estimate the optimal time for a replacement. As long as a new pipe ownership cost, the sum of replacement, operational and maintenance cost (Boulos, 2017), exceeds the existing pipe ownership cost (operational and maintenance), it is economically

inconvenient to renovate the water main. Once keeping in life an existing pipe, continuously repairing breaks, occurring due to deterioration, becomes too expensive, going above costs of replacing and maintaining a new pipe, that is the moment when water distribution company must substitute the pipe to allocate efficiently capitals.

Effects of pipe failures do not regard only the economic performance and efficiency of water utilities. In fact, besides maintenance cost, consequences involve not only economic aspects, but also operational, environmental and social. Economic consequences concern factors of asset utility and society in monetary terms, such as loss in revenues of direct and indirect business, repairing cost. Operational effects refer to the loss of operational availability of infrastructure assets and surrounding (e.g., loss of production, loss of hydraulic functionality, etc.). Effect on habitats, water bodies, service areas, archaeological sites etc. are part of those called “environmental consequences” (Mazumder R. , Salman, Li, & Yu, 2021). Finally, the social consequences concern all the impacts on public life deriving from inconveniences to public life due to inefficiencies, traffic slowdowns, etc. (Salman & Salem, 2011).

2.2 Prediction field

Over years, researchers and scientists have studied pipe breaks, especially from the 80s (Kettlere & Goulter, 1985; Kazei & Goultier, 1988; O'Day, 1985). With the improvement of data quantity and source, more and more articles have been published in the most important journals for urban infrastructure management. Intending to increase economic performance and asset management, the two main undertaken problems have been causal inference and prediction (Konstantinou & Stoianov, 2020). Going through the first problem, researchers attempt to identify factors and mechanisms responsible for pipe failures and for accelerating deterioration phenomena. The outcome of these studies is the set of variables that will be later included in predictive models, that, on the other side, allows to reach a specific result such as pipe break, a hazard function, predicted by using a specific set of factors (Konstantinou & Stoianov, 2020).

Predictive models can be grouped, depending on input data and procedures, into physical, statistical and machine learning models.

Physical models predict the breaking of tubes by evaluating the load to which the tube is subjected and the ability of the tube to handle the load efficiently. Some other of the variables used in these models are corrosion, stress acting on pipes and residual strength and remaining pipe thickness. As soon as stresses of load acting on pipes’ surfaces exceed the remaining strength of the pipe, a break is expected to occur. The main limitation of physical models is the high requirement of accurate data, with infield measurements, that

are often difficult to obtain for time and financial constraints (Marlow, Davis, Beale, Burn, & Urquhart, 2010).

More common and with easier applications are statistical methods, that use available data regarding breaks history and pipe intrinsic information for predicting future failures.

The early developed statistical methods are those called deterministic, which typically use two or three parameters (e.g., pipe length, age, breakage history) to predict specific break rate or time-to-next-break prediction. An important application point for these models is that they are best applied to a group of water mains that are homogenous in respect to the parameters influencing the breakage patterns (e.g., diameter, age of the pipe, pipe type, number of repairs etc.) (Kleiner & Rajani, 2001). Among some of the deterministic models that have been the backbone of this approach, it is worth remembering the time-exponential models, e.g. (Shamir & Howard, 1979; Clark, Stafford, & Goodrich, 1982), time-linear regression models, e.g. (Kettler & Goulter, 1985; Jacobs & Karney, 1994). Actually, many weaknesses come up from the deterministic models such as low values of R^2 in some cases, a sign that the model is not enough powerful to predict occurrences but only to see which variables accelerate more degradation of pipes. Or more the strict requirement to over-partitioning pipes in homogenous groups, e.g., using ANOVA, to apply the model.

The second macro-group of models takes the name of *probabilistic multivariate* models because they consider many covariates that influence breakage patterns. The output is usually a *hazard function*, representing the probability to have a break at each unit of time during the pipe's life (Cox D. R., 1972; Andreou, Marks, & Clark, 1987). Unlike deterministic models that usually are relatively simple, the mathematical framework of probabilistic multivariate models is much more complex and able to handle many variables (Kleiner & Rajani, 2001). On the other hand, the inclusion of any factor for the prediction of failure probability models reduces the need to divide water networks into homogeneous groups. Indeed, these models are also more suitable for pipe individual analysis rather than families of homogeneous pipes, making these approaches *ad-hoc* for single replacement strategies planning. However, the most relevant strength of some multi-variate models is the ability to take into consideration also right-censored data, which instead are usually discarded from samples for inability to include them into models. "Right-censorships occur when the event of interest (pipe break) has not occurred within the study period. Right-censorship occurs within pipe break datasets in two forms: (1) the pipe is removed before the break is recorded, or (2) the pipe break has not yet occurred (pipe is still in service but the break has not yet occurred). In either instance, the event of interest (break) has not been recorded for the pipe and a right-censored event is said to occur at the last observed time for the pipe (i.e., the date of pipe removal or the latest date of pipe breaks records)" (Snider & McBean, 2021).

Eventually, the last group, *probabilistic single-variate*, gather those models using probabilistic processes to predict probabilities of pipe life expectancy, probability of break and probabilistic analysis of break clustering phenomenon (Kleiner & Rajani, 2001). These models, e.g. (Goulter, Davidson, & Jacobs, 1993; Kleiner & Rajani, 2012; Clark, J., Thurnau., R., & S., 2010), have the feature to be enough versatile even though they are not appropriate for medium and long-term plans, but only for the short term.

Recently, machine-learning technologies have increased their use and popularity in water distribution network management. Machine-learning algorithms, with a data-driven approach, can strongly identify relationships between several input factors, probably responsible for pipes' deterioration, and breaks (Snider & McBean, 2021). Generally, they seem to be more accurate and easier to calibrate, making them nowadays the most spread and used methods for studying pipes' life cycle.

Machine-learning models result to be really different from each other, with several approaches going from clustering through *k-mean cluster* technology pipes (Farmani, Kakoudakisb, Behzadianc, & Butlerd, 2017), or use of non-additive models such as decision tree, random forest and gradient boosting machine (Chen, Beekman, & Guikema, 2017).

With the parallel improvement of data collection by water utilities and advanced technology in machine learning, predictive models have significantly increased their accuracy and efficacy. Furthermore, last studies have also dealt with the problem of handling right-censored data, overcoming the removal of these data that, introducing bias, causes prediction of early pipe breakage. Survival machine learning is a relatively new field that tries to include right-censored data for developing predictive models. By combining survival analysis techniques and machine learning algorithms, models such as Random Survival Forest (Snider & McBean, 2021) have opened significant new horizons for water main management.

2.3 Economic goal

Most of the studies about predicting pipe life expectancies were born to contribute to the improvement of economic performances of water utilities. Therefore, besides the prediction model itself, they often include some assessment of the economic convenience of maintenance strategies (Kibum, et al., 2019; Snider & McBean, 2021; Kleiner, Nafi, & Rajani, 2010; Li, Ma, Sun, & Mathew, 2011).

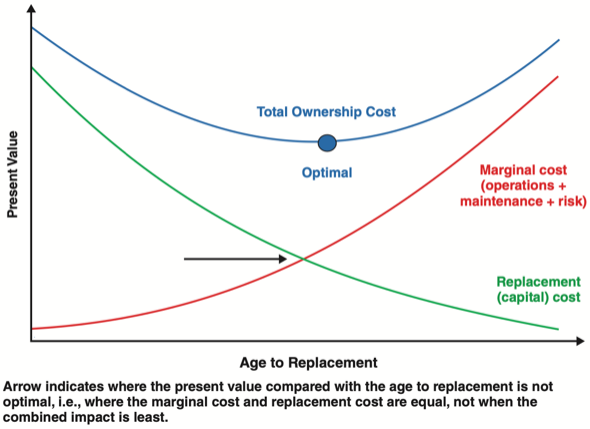
Generally, although slight differences between authors' ways of allocating some costs, economic approaches are aligned. Indeed, water main management presents mainly two kinds of cost, namely marginal cost and replacement cost and the financial of asset is based on the comparison of these two costs (Boulos, 2017).

Marginal cost includes all the expected risk costs of keeping an existing pipe (e.g., the consequences of an upcoming break, such as water loss, other direct damage such as adjacent infrastructure, road damages etc.), the accelerating cost of maintenance of the pipe (more the pipe fails, faster breaks occur and higher expenses for fixing) and costs due to declines of service level (e.g., social costs such as pollution, time loss, loss of business, disruption etc.). The curve of marginal cost rises with increasing rates as all cost components increase dramatically over time.

All the costs involved in replacing a pipe account for replacement costs, including mobilization components (e.g., costs for setting up the job site, signage, discovery and marking of adjacent infrastructure) and variable components (e.g., material, new pipes) (Kleiner, Nafi, & Rajani, 2010). The value is discounted and the present value decrease as pipe renewal is deferred. Therefore, the replacement cost curve decreases as time increases (Boulos, 2017).

The sum of these two mentioned curves represents the curve of the total cost a company must bear for replacing an existing pipe. The economic optimal time to decide not to repair the main but to substitute it is uniquely determined with the time when the minim of the curve occurs. If the replacement takes place before the optimal time, the company does not exploit all the useful life of a pipe. On the other hand, substituting a pipe beyond its optimal time means wasting capital in sustaining, maintaining and repairing a pipe that has overcome its economic useful life (Figure 1).

Figure 1 Optimal timing of pipe replacement



Source 1 Boulos Paul F., Optimal time of pipe replacement, 2017, Journal AWWA, 109 (1), 45

Further studies have been carried out including additional factors in the economic assessment of replacing. For example, Kleiner, Rajani and Nafi (2010) included in one of their publications the concept of economies of scale and the convenience of replacing also some adjacent water main once the excavation is already done for replacing a pipe. To do so, they defined the cost of replacement as the sum of a fixed and variable contribution. The first fixed addend includes costs such as for setting up the job, signage, marking of adjacent infrastructure, while the variable cost depends on the length-unit cost, pipe material, diameter etc. They described two types of economies of scale: quantity discount, which applies to the variable component of pipe cost and contiguity discount, which applies to the mobilization (fixed) component. Quantity discount occurs when pipe material installed exceeds a certain quantity lower bound, from which pipe unitary cost start decreasing.

Contiguity discount is defined as follows: “if pipe *j* is contiguous to pipe *i* (both share the same node) and both are replaced in a given year *t* they are assumed to be part of the same replacement project and therefore only one mobilization component is levied. Therefore, if *k* contiguous pipes are replaced in a given year, their total replacement cost will comprise the sum of all their unit costs plus one mobilization charge (i.e., *k-1* mobilization charges were saved compared to the cost of replacing *k* non-contiguous pipes)” (Kleiner, Nafi, & Rajani, 2010).

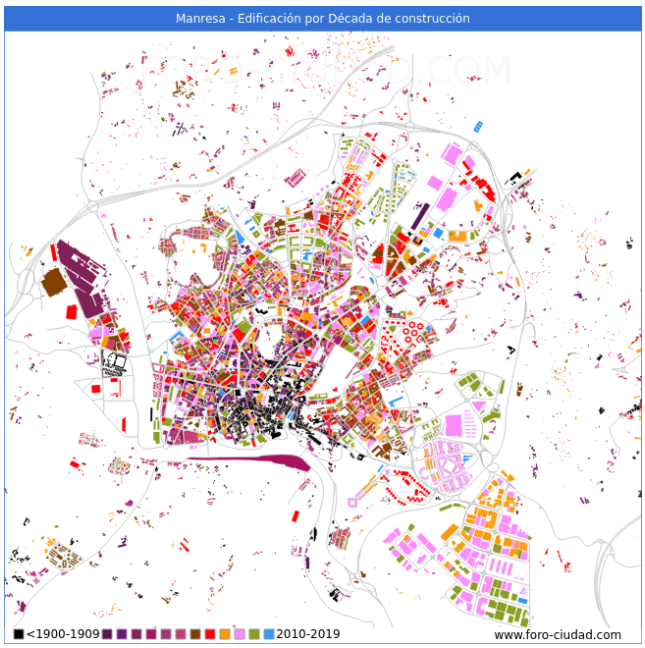
2.4 The goal of the thesis as a bridge for research gaps

After reviewing the literature selected for this thesis, we have concluded that the industry of water main management cannot be classified as an industry with “disruptive technologies” coming over previous predictive methods and maintenance strategies. Improvements and innovation arise slowly, and, in many cases, they concern only little modifications of already existing methods. For this reason, most of the publications are updated literature reviews where usually the authors also include a specific case of study, with application of selected predictive models, economic analysis and maintenance approach.

Many cases of study have Canada as the place of interest (Snider & McBean, 2021; Wang, Zayed, & Moselhi, 2009; Kleiner, Nafi, & Rajani, 2010), but also the USA (Chen, Beekman, & Guikema, 2017; Mazumder R. K., Salman, Li, & Yu, 2021) or Korea (Shin, Joo, & Koo, 2016; Kibum, et al., 2019).

However, cases of study regarding Mediterranean locations and more generally, European sites, are rare with no recent studies. Therefore, since environmental conditions are among those most affecting corrosion, burst and breakages, an application to a location such as Manresa, in Catalunya, may represent an original case of application of existing knowledge. The town has a Mediterranean subhumid climate with a continental tendency climate, with cold winters, with 1-2 months with 0°-5° averaged temperature, and hot, moderately dry summers, while the rainiest seasons are spring and autumn [9]. Moreover, the city has experienced an important urbanistic expansion from the 60s, due to the increase of population, with almost the 60% of the total built surface dated from the 60s to the first decade of the 21st century [10]. This background has implied a relevant development of the entire distribution network to fulfil water demand in the newly expanded area of the city.

Figure 2 Manresa - building by decade of construction



Source 2 [10]

In addition, generally, each author undertakes their own path for the application of predictive models and maintenance investment analysis. Over time, the literature has not intrinsically outlined common guidelines for future study, letting anyone interested in the topic follow different roads. Furthermore, especially for unexperienced company, leveraging Machine Learning power to predict future distribution network conditions has been obstructed by the complexity of the subject matter.

The goal of this work is to provide readers and future interested a well-structured approach the study of water main failures, streamlining the methodology for the application of ML algorithms and to develop economic assessment model to address both real business needs and theoretical knowledge.

This thesis will be carried out starting from the company's needs, discussed with one employee of the data analysis department of *Aigues Manresa*, and we will try to give an external consideration about how to face and solve their problems using new theoretical, but also practical, methods.

3 Data exploration

The following chapters will immerse the reader into the world of water distribution networks through data analysis techniques. The goal is to make use of information provided by the company “Aigüe Manresa” to identify potential behaviours of pipes as a response to changes of variables such as diameter, age or pressure. Charts and other data visualization tools will be used for this purpose. By doing so, the author could already delineate particular behaviours that can be meaningful also in the later stage of prediction.

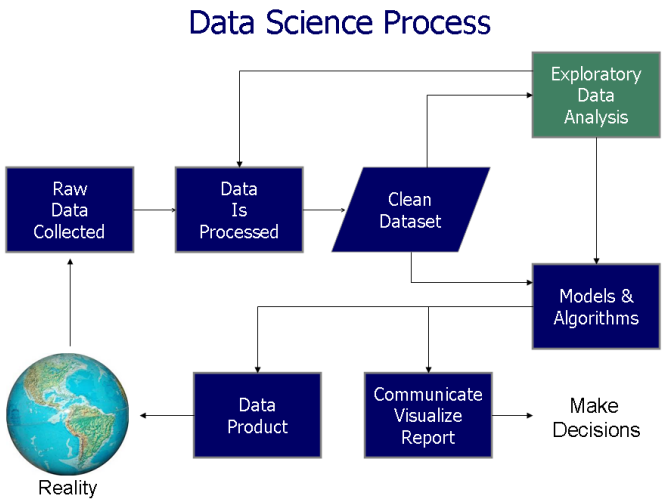
3.1 Procedure structure

Data analysis is a structured process, made of defined and not-officially standardized steps, to convert raw and apparently meaningless raw data into useful information to support decision-making (Judd & McCleland, 1989). It can be seen as the pairing of people who develop technology that can learn from data with people who have data and who have problems to solve.

Although the discipline has various approaches, the main phases on the entire analysis are quite defined and they are iterative, as the result of a later phase may result in additional work in earlier stages (Schutt & O'Neil, 2013).

In order of being performed, the phases of a proper data analysis are the *data requirement, data collection, data cleaning, analysis of data* and eventually *interpretation of results* (Erdelyi, 2021).

Figure 3 Data science flowchart



Source 3 Schutt, R. & O'Neil, C., Doing Data Science, O'Reilly Media, 2013

3.2 Data requirement and data collection

3.2.1 Which data are needed?

As previously explained in *chapter 1*, pipe-intrinsic, operational and environmental factors are those mostly used in pipes' failure management. A proper amount of data would allow either to perform analysis for visualization of insight and for applying ML to predict future failures. Therefore, the author asked the company to provide as much data as they were able to collect and give, in the matter of pipes and their failures over years.

3.2.2 Source of data

To accomplish the goal of this study, *Aigues Manresa* has provided the author with a considerable amount of data regarding pipes and failures in the distribution network. Unfortunately, *Aigues Manresa* has never collected data regarding environmental factors such as soil corrosivity or average traffic load, so only intrinsic and limited operational variables have been collected. Moreover, the importance of collecting data to support decision-making processes has been deeply understood only in the early 2000s, data regarding failures and other intrinsic factors affecting and characterizing pipes are available only for about the last 15 years of operation. However, the sample is considered enough big to be used for the purpose of this thesis.
Data have been provided as.*xlsx* extension files, containing three different datasets, respectively *pipe_all*, *pipe leak* and *arc_minsector.*

. *Pipe_all*

The dataset *pipe_all* (henceforth *pipes*) contains detailed information about all the pipes within the whole distribution network between 2005 and 2019. Going in detail into which information we have for each element of the dataset, the following are the *category of information,* also known as *variables*, of the dataset, in their original configuration:

1. *Arc_id:* pipe's ID.
2. *Builtdate:* pipe's installation date, as *day/month/year.*
3. *Length:* pipe's longitude.
4. *Matcat_id:* material a pipe is made of.
5. *Pnom:* average pressure detected inside a certain pipe.
6. *Dnom:* nominal diameter of a pipe.

As it happens frequently when handling a large amount of data over years, some information can miss or go lost for unexpected circumstances or lack of information. For example, for some pipes the entry on *builtdate* column is 01/01/1900, corresponding to an unknown date.

Therefore, during the cleaning data stage of the study, these values must be properly handled.

Regarding the way on how to measure the nominal diameter of the pipe, for some pipes, such as those made of polyethylene, it is an exterior measurement (including also the section of the pipe). Iron pipes, as well as those in fiber cement, have their diameter measured from the inside. However, the operative method used for measuring the diameter cannot be deducted from the dataset and this represents a clear example of hidden information that cannot be spread objectively through data, but only verbally.

. *Arc_minsector*

The dataset *arc_minsector* (henceforth *sectors*) introduces two new concepts never encountered through this study so far, “minimum sector” and “node”.

As explained by a correspondent of *Aigues Manresa*, a “sector” is the whole group of pipes that will be affected negatively by a burst of one of the pipes belonging to the sector itself. Obviously, companies used to gather pipes into *minimum* sectors because one of their goals is to impact as few pipes as possible given a certain burst, reducing a wider sector into a *minimum sector* identifying the minimum number of pipes likely involved in a failure. In case of a breakage, the only pipe shut down for placing the maintenance would be those belonging to the same *minimum sector* of the broken one.

The other new term is “node”, identifying a point of connection of two or more pipes. Each pipe will so have two nodes, one for each extremity.

Once clarified the meaning of these new concepts, the core of the speech can move toward the analysis of the dataset, which presents seven variables:

1. *Arc_id:* ID of a pipe.
2. *Minsector:* minimum sector the pipe belongs to.
3. *Node_1:* connection node 1 of a pipe.
4. *Node_2:* connection node 2 of a pipe.
5. *Arccat_id:* a string containing information about material and nominal diameter of a pipe.
6. *Custom_length:* Longitudinal length of a pipe.
7. *Builtdate:* date of installation of a pipe.

. *Pipeleak*

This last dataset (henceforth *leaks*) lays the foundation for this study, as it contains information regarding failures that occurred from 2005 to 2020. It is not exactly known whether the starting date of this sample coincides with the campaign of data detection from

Aigues Manresa, but a period of 16 years might constitute enough big sample to give meaning and solidity to this study.

The v*ariables* of this dataset are eight, respectively:

1. *Arc_id:* ID of the pipe affected by a failure.
2. *Builtdate:* installation date of the leaking pipe, as *day/month/year.*
3. *Data:* date when the leakage has been firstly detected.
4. *Length:* longitude length of the broken pipe.
5. *Matcad_id:* material of the broken pipe.
6. *Pnom:* nominal pressure of the broken pipe.
7. *Dnom:* nominal diameter of the broken pipe.
8. *Minsector:* "minimal sector" the broken pipe belongs to.

Even for this dataset are valid the same instructions regarding "builtdate" and nominal diameter: date corresponding to January 1st, 1900, are "not known" entries, while for the detection of pipe thickness of different material, the same rules must be applied.

3.3 Data cleaning

3.3.1 Software setting

The third step, so-called "data cleaning", is dramatically vital for the accuracy and quality of the analysis and also extremely impacting on the pace of progress. It consists in amending, fixing and removing incorrect, corrupted, superfluous data as well as possible inconsistencies. In fact, a predictive model, or more generally any outcome may result to be unreliable whether based on not-cleaned data.

The language selected to accomplish the goal of this study is *R 4.0.2 GUI 1.72 Catalina build (7847)*, many of the packages on CRAN, containing all additional packages useful to exploit R resources, while the statistical software *Rstudio*. Among all the libraries into CRAN, R has some packages constituting the backbone of the software itself, which are its graphical libraries allowing the coder to display graphs and make them interactable with the user. In addition, R offers several advanced data analysis options such as forecasting model development, machine learning algorithms, etc. [11]. All these features make R a valid and suitable tool for the purpose of this work at the most of its possibilities and expectancies.

3.3.2 Data cleaning

Among the activities conducted during the data cleaning, "not-available" entries and management of variables format are worth to be mentioned. Actually, right before moving on to checking data, an analyst should be sure that the upload of data from the data source onto the used software has taken place successfully. This basically means checking the proper *reading function* to use, dimensions of the uploaded dataset and heads of variables.

In fact, it could happen that during the upload of a dataset, some rows and/or columns go lost for mere problems of reading.

```
leaks <- read_excel("leaks.xlsx")
pipes <- read_excel("pipes.xlsx")
sectors <- read_excel("sectors.xlsx")
```

. *Aligning formats*

The first typical step of data cleaning is the variables' format check, as in many cases they do not present the right format, suitable for being manipulated easily and giving easily all the information they have. It is for example the common case of dates, that usually are uploaded as integers that count the number of days from a certain day (usually January 1st 1900 or December 31st 1899) up to the date they represent. Therefore, this integer must be converted in a data format. Also, in case data are missing in the original dataset, with empty entries, these cells would be seen as containing zero and in the calculation of the date, R would assign them the origin date from which the calculation starts. For this reason, all the rows dated to December 31st, 1899, would get their date substituted with NAs.
The same procedure is carried out for "NULL" entries to uniform the nomenclature to *NA* to indicate missing values.

. *Discarding duplications*

After checking whether the dataset' dimensions on R match with those from the *.xlsx* files, sign of a successful upload, and transform variables into proper formats, an important step of data cleaning consists in checking and removing potential redundancies in the dataset. Indeed, duplicate observation happens really often when collecting data, especially when they come from partners, clients, other or multiple departments. De-duplication is one of the largest areas to be considered in this process [12].
For each dataset, it is needed to identify which, among all the variables, is a *primary key* that distinguishes a univocal row from a duplication. In some cases, not a single variable can alone determine the uniqueness of an observation and only the combination of two or more values can succeed in the purpose.
Starting with the analysis of *pipes* dataset, *arc_id* is identified as *primary key*, as it will be an error to find two rows with the same value in the columns *arc_id*, because they would give the same information regarding a single item of the network. It is found out that there are 5,393 duplications, destined to be removed from the dataset: they represent a threat for a good study and for reliable outcomes. Out of 18,022 observations from the original dataset, the final, at least for now, the dataset will be reduced by almost 30%. Just to have

an idea on how much important the data cleaning process is as data never are in a “ready-to-be-used” status.

Same *primary key*, *arc_id*, is used for the same purpose for *sectors.* The output, in this case, is definitely happier since no duplication is detected, for the luck of the analyst. Indeed, reducing a database’s size is always a loss, as a bigger sample means more data to validate, but also to destroy, results. Reliability goes up as the data sample increases in dimensions.

Eventually, for the dataset reporting information about failures in the distribution network, *leaks,* the analyst had to choose during the detection of duplication, two *primary keys.* In fact, a single pipe might experience multiple failures in its life cycle, therefore two or more rows with identical values in the column *arc_id* are allowed. However, we assume that two failures cannot simultaneously occur on the same pipe. Therefore, a row presenting *arc_id* and *date* entries identical to a previous one will be seen as redundancy, so it is destined to be eliminated from the dataset.

By asking Rstudio to display the number of duplications with the formula *duplicated(),* we get aware of a discouraging output: out of 3,007 observations of failures, the backbone of our study, 1,776 are not unique rows: almost 60% of the dataset is made of duplication.

Actually, the *duplicated()* formula considers redundancies also those rows with identical *arc_id* entry and NAs as *date,* but it can also be likely scenario that the latter value just misses for detection issues. Therefore, even though two failures for the same pipe happened in different moments but without recording the breakage time, *duplicated()* would see the two events as duplication. As there are 337 NAs in the column *data*, the worst scenario we could run into is that all 337 observations were actually failures that occurred to pipes at different moments, and by using the *duplicated()* formula, useful information is discarded from the dataset. In the best scenario, all 337 leaks are really duplications, and the formula is efficiently cleaning the dataset up. An analyst would opt for having a less numerous dataset but surely clean rather than keeping rows that might bring bias into the study.

However, as the phenomenon is of a relevant magnitude that may lead to rising doubts regarding the reliability of the data source, it is has been decided that before moving on with the following step of data analysis, to report all the issues to *Aigues Manresa*, with the purpose to understand a reason of the found inconsistencies and maybe, find a solution.

. *Arc_id cross-check*

As we know, *pipes* dataset represents the "born-list" of all items in the distribution network and, in theory, should be the most reliable and accurate source of information we have. We expect that all events recorded in the *leaks* dataset and all the information contained in *sectors* are about pipes registered into *pipes* dataset.

By defining a formula that shows the elements of a vector not into another vector, it is possible displaying how many are the element in *leaks$arc_id* and *sector$arc_id* are not into pipes$arc_id column.

```
'%!in%' <- function(x,y)!('%in%'(x,y))
filter(leaks, leaks$arc_id%!in%pipes$arc_id)%>%nrow()

filter(sectors, arc_id%!in%pipes$arc_id)%>%nrow()
```

The discovery in the aftermath of this process is not negligible: 120 leaks regard pipes not registered into *pipes* and even 1,048 rows from the *sectors* dataset. This is a dramatic discovery for the solidity of the data and their source because we expected that everything happening as a failure, involve a pipe registered in the system and all sectors include pipes the company know everything about. This inconsistency is even more worrying than the one about duplications: a deeper analysis and a cross-check directly from the source of data origin is strictly required.

. *Length consistency between **pipes** and **sectors** dataset*

Already warned by the company of this discrepancy of pipes' length between *pipes* and *sectors* dataset, in this subchapter the focus will be on trying to make uniform values of this variable in the two data frames.

The analysis will be carried out according to the following procedure:

1. Create a temporary data frame including *arc_id* and the corresponding value of length from *pipes* and *sectors* dataset, respectively *length* and *custom_length.*
2. Omitting *arc_id* where one of the two values, either in *length* or *custom_length* is *NA* because for those pipes we only have one value of length that will be taken anyway.
3. Check whether there are still pipes with different values of length after the first cleaning about *NAs*. If not, the inconsistency was due to the value of length not registered in one data set or the other and the length analysis ends. If it does, we should keep on investigating, moving to the following step.
4. The following step is to understand the magnitude of the *delta* between length values, creating a new column in the temporary data frame, *delta_length.*
5. The first check on the delta regards numbers rounding. As lengths value are rounded at the second decimal digit, *arc_di* with *length_delta* equal to $\pm$ 0.01 will not

be considered different as the delta is only due to approximation. Therefore, those rows are taken away from the data frame.

6. Once step 5 is performed, by checking the dimension of the resulting dataset, the number of residual pipes with divergent lengths between *pipes* and *sectors* data frame will be displayed.
7. Depending upon step 6 results, deciding how to handle the situation.

After removing a consistent number of rows presenting *NAs* (step 3), 453 pipes have different length values. However, around 300 of these differences is due to rounding, but still, 152 pipes are recoded into *pipes* and *sectors* dataset with different longitudinal dimension.

To decide how to proceed, it has been considered appropriate to study what is the distribution of delta values for the interested pipes.

```
summary(consistency.test$length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.21    2.22    9.98   34.08   46.96  440.00
```

Already 25% of delta values diverge for more than 2 meters and half of them for almost 10 meters, and moving up to the following quartiles, values incredibly worsen. It is not a matter of differences of centimeters, rather entire meters of pipes in excess or lack: another important inconsistency to be reported to the company.

3.4 Data validation and check

Given the high number of inconsistencies found on the three datasets provided by *Aigues Manresa,* the author has believed that the work could not be kept on board without a cross-check from the data source. Therefore, all the previous conclusions from the data cleaning have been reported to the responsible department of the company.

Indeed, all mismatches highlighted from the analysis have been confirmed by the company's correspondent, who has gladly embraced these starting points to run a new extrapolation of the dataset from their database. In fact, as the installation of their new Geographical Information System (GIS) took place only in March 2018, at the time when data have been extrapolated, late 2019, the GIS was still on a debug phase that led to duplications, mismatches and inconsistencies in the data.

Once becoming aware of their data *Aigues Manresa* have also relied and worked on, a new data extrapolation has been performed, at this time based on a solid and more reliable GIS, even able to record and store additional information. Indeed, the new dataset created and used for the continuation of this work not only is more solid for very high rates of consistency

throughout all the datasets but also *all_pipes* dataset has been improved with two valuable variables:

1. *End_date:* date at which a pipe has been replaced.
2. *State:* categorical variable assuming value *0* whether a pipe has been replaced, *1* whether still in service, *2* whether a pipe is out of service.

Unlikely the first versions of the dataset, the new ones are presented with *.csv* extension, an aspect not particularly impacting by how *NA* data are expressed. Whether for *.xlsx* files, the origin counting date was 31/12/1899, in the new *.csv* file the count starts on 01/01/1900, that would be our *NAs* for dates. Another peculiarity of *.csv* files is that *NAs* can as *NULL, the* reason was needed to check for those columns with *NULL* entries and replace those cells with *NAs.*

Even though new datasets are reduced in terms of dimension (new *leaks* dataset is less than half of the previous given one) and present more missing values than first versions, data are definitely solid and reliable, as:

- Duplications of rows with same *arc_id* within *pipes* dataset are *0.*
- Duplications of failure detections are only 4 rows out of 1173 rows of *leaks* dataset.
- No pipes whose there are failures recorded are not registered in *pipes* dataset (before, 120 were the failures of unknown pipes).
- No pipes whose the belonging to a certain sector is unknown (before, for more than 1000 pipes the sector of belonging was not registered)
- No pipes whose sector belonging was known is not included in *pipes* dataset (before, we knew information about 762 sectors of pipes not in *pipes dataset)*
- No pipes with different longitudinal lengths between *pipes* and *sectors* dataset.

4 Data analysis

The goal of this chapter is to go through all the information available to get insight regarding pipes and try to understand whether or not there is a straightforward relationship between some pipe features, such as the material or the diameter, and a break. Firstly, some variable will be analyzed *stand-alone* to better understand the network and only on a second stage, they will be related to failure events.

4.1 Frequency of pipes by age

Once obtained cleaned data frames, to start with the extrapolation of meaningful information, seeing when pipes have been installed can give a general understanding of how pipes are distributed over year and whether the distribution matches Manresa's main expansion period.

To carry the analysis out, a new data frame only containing pipes with no *NAs* within the *builtdate* column was created and named *pipes.complete.* Basing upon the variable *builtdate* in *the new* dataset, a new variable *age* has been created as the difference between pipe year of installation and the *enddate,* which for this study has been set on the 1st September 2021. The function `eeptools::age_cal` has been used as follows:

```
pipes.complete <- mutate(pipes.complete, age=age_calc(builtdate, enddate
= as.Date("2021-09-01") , units = "years"))
```

By checking the range of the new variable *age*, the youngest pipe has only 0.098 years (indeed installed only on July the 27th 2021) while the oldest is more than 77 years old, dating back to 1944. The quartile distribution is the following:

```
quantile(pipes.complete$age)
```

```
##          0%         25%         50%         75%        100%
##  0.09863014 15.04109589 23.66849315 44.67121791 77.66849315
```

For plotting the distribution of age, bins of 10 years have been created because it has been believed that analyzing pipes' age per decade would be more insightful rather than per single year.

Figure 4 shows that among the entire water distribution network, there are especially two age bins with the highest frequency: pipes aged between 10 and 20 years and 40 and 50 years, therefore respectively built in the decade 1970-1980 and 2000-2009. The result is perfectly aligned with Manresa's main expansion periods, where there was a boost in surface construction as Figure 5 shows. 1970-1980 and 2000-2009 were the two decades where most of the city surface has been built and consequently, pipes have been installed to enlarge the water distribution network.

Figure 4 Histogram: distribution of age of pipe

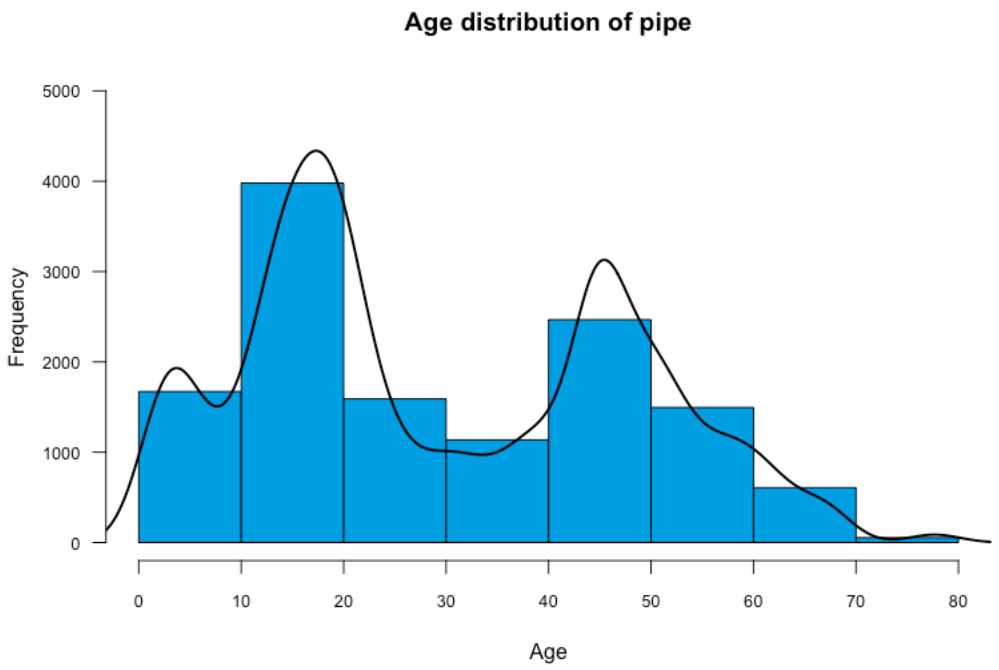


Figure 5 Manresa data: built surface over decades

Superficie Construida			
Decada	% Total	Top Provincial	Top Nacional*
<1900	7.39%	<- 2º ->	<- 13º ->
1900-1909	2.26%	<- 15º ->	<- 202º ->
1910-1919	1.77%	<- 2º ->	<- 16º ->
1920-1929	3.51%	<- 3º ->	<- 23º ->
1930-1939	1.72%	<- 6º ->	<- 73º ->
1940-1949	2.67%	<- 5º ->	<- 68º ->
1950-1959	5.89%	<- 5º ->	<- 49º ->
1960-1969	11.16%	<- 10º ->	<- 64º ->
1970-1979	16.27%	<- 14º ->	<- 103º ->
1980-1989	10.53%	<- 7º ->	<- 140º ->
1990-1999	15.40%	<- 11º ->	<- 101º ->
2000-2009	18.81%	<- 9º ->	<- 147º ->
2010-2019	2.65%	<- 15º ->	<- 180º ->

Source 4 [10]

4.2 Pipe material

Surely among the most relevant variable of a pipe, materials used for distribution network have been changed over the years, as technology and discoveries have progressed.
By grouping pipes by the same material, a bar chart displaying the absolute frequency of usage of a single material has been created to visually identify widely used materials.

Figure 6 Bar chart: absolute frequency of pipes by material

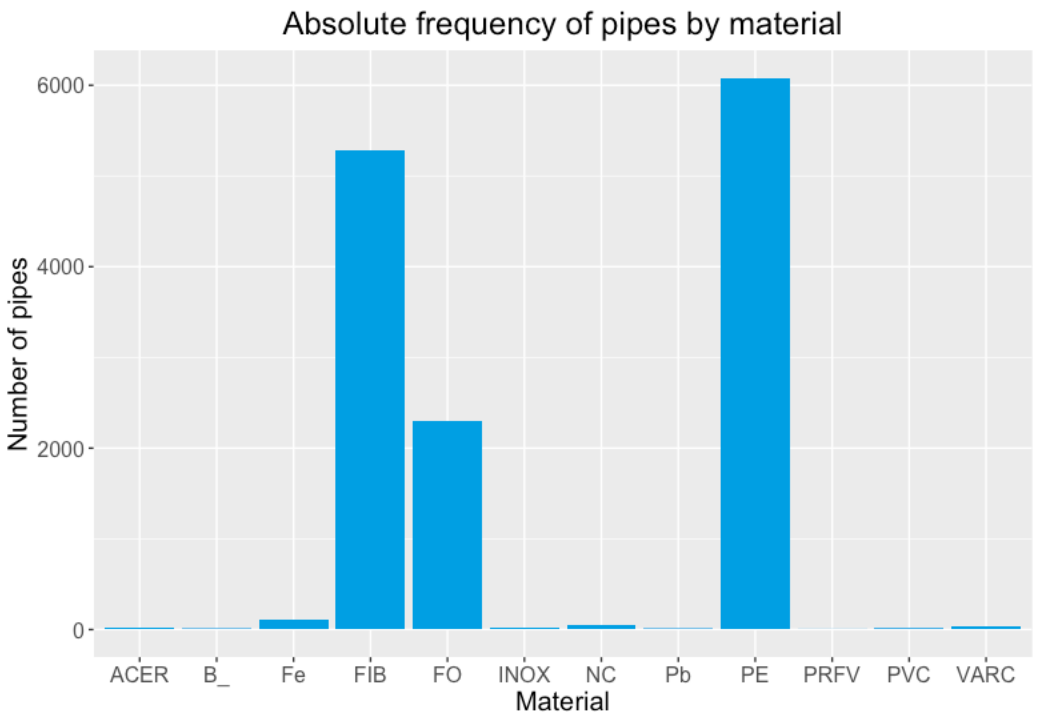
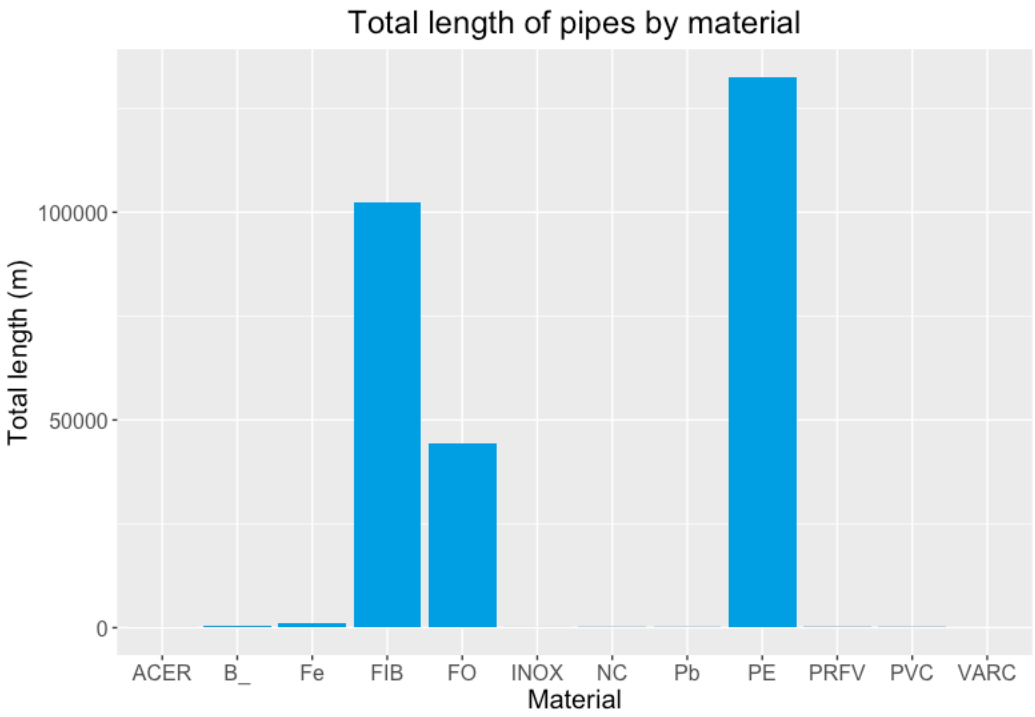


Figure 6 gives a piece of important information for the continuation of the study: the almost totality of distribution network in Manresa is made of “FIB” (Fibrocemento), “FO” (fundición dúctil) and “PE” (Polietileno) standing respectively for *fiber cement, ductile cast iron*, *polyethylene*.
Besides a light relevant usage of “Fe” (iron), the use of other materials can be considered negligible.
However, could be possible that the absolute number of pipes made of a certain material is low, but the total length of those tubes might represent an important share in the entire network. This consideration has led to the construction of the following graph, where not merely the frequency of usage is displayed, but the total length of pipes per material.
However, no relevant changes come out from Figure 7, as the same three materials, *FIB, FE* and *PE* remain the highest usage also in terms of meters extension.
Since the impact of all other materials is really irrelevant, the study can proceed only focusing the attention on pipes where the variable *matcat_id* assumes the values “FIB”,

"FE" and "PE". All the rows where this condition is not respected will be taken away from the dataset, filtering the information by what really matters. The risk of keeping pipes made of those "low-used material" is that in a future predictive stage, predictive models could get distorted as bias can be generated by those classes of materials.

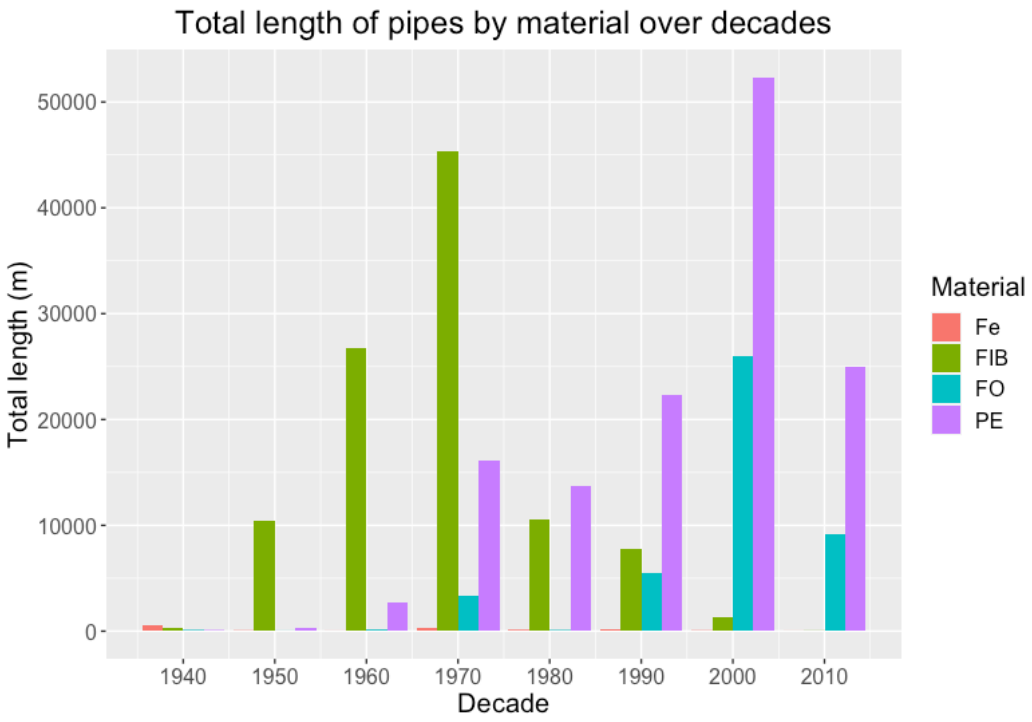
Figure 7 Bar chart: total length of pipe by material



4.2.1 Material per decade

Once having understood the major used materials within the network, the focus moves on how materials have been used over years, with the attempt of identifying possible trends.

Figure 8 Bar chart: Total length of pipes by material over decades

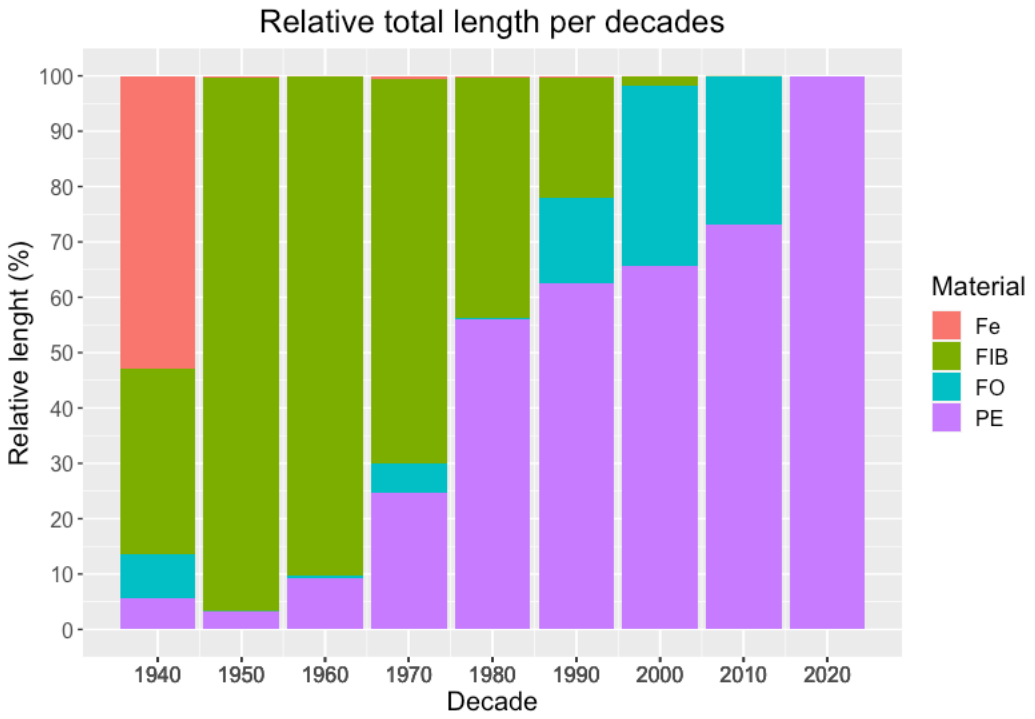


It is straightforward as “FIB” that in the past was largely the most used material, sometimes almost the unique one, has been abandoned from the 70s and replaced by a more and more used “PE” and, over the last 3 decades, by “FO”. The cause behind the stop of fiber cement as the main material for pipes construction is due to the presence of asbestos, that once exposed to weather and erosion elements, can be a source of airborne toxic fibers, threatening human health. The “Orden de 7 de diciembre de 2001 por la que se modifica el anexo I del Real Decreto 1406/1989, de 10 de noviembre, por el que se imponen limitaciones a la comercialización y al uso de ciertas sustancias y preparados peligrosos”[1] had officially prohibited the presence of asbestos within fibers of fiber cements material.

To fully understand how materials have been used relatively within a specific decade, the following graph displays out of the total 100% of length installed in each decade, which is the share per each material.

[1] Orden de 7 de diciembre de 2001, por la que se modifica el anexo I del Real Decreto 1406/1989, de 10 de noviembre.

Figure 9 100% bar chart: relative total length of pipes per decade

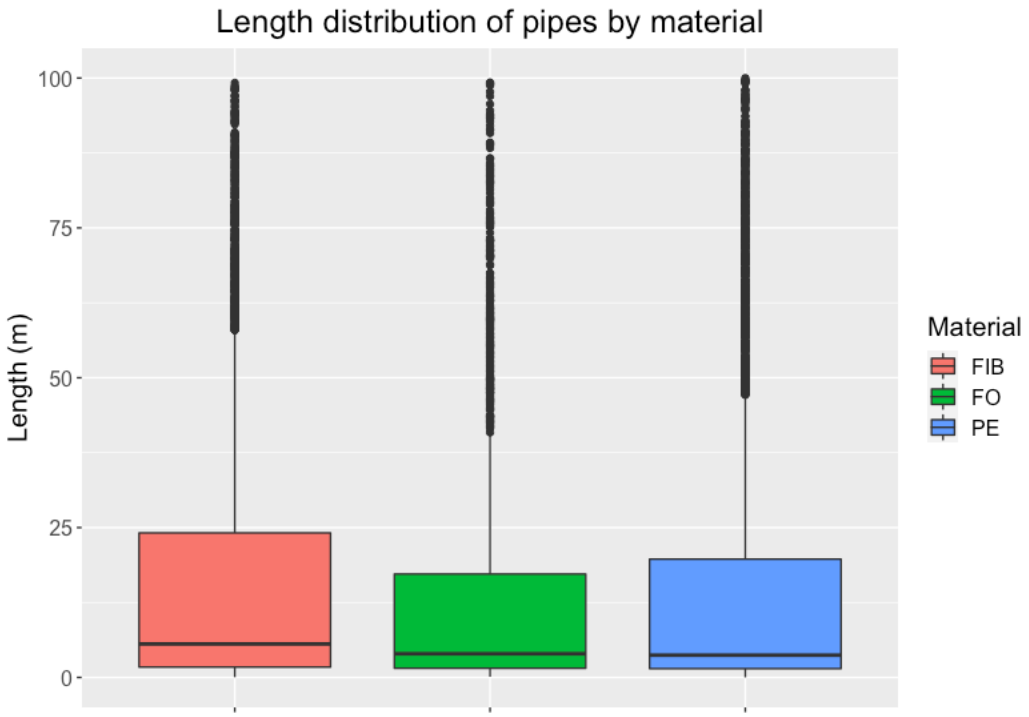


In the late 60s, regulations around the world against asbestos risks have been revamped and, in the early 70s, deeply strengthened [13], trigging strong reflection on fiber cement usage also in Spain. The main percentage reduction in length of pipes in "FIB" occurred during the three decades from 1970 and 2000, for the benefit of "PE" that had been increasingly replaced "FIB" in the distribution network. However, whether from the 1950s the share of length lost by "FIB" was almost completely taken by "PE", from the 90s, also pipes made of "FO" started to be largely used.

4.2.2 Material and average pipes length

In this sub-session of relationship with used material, the goal is to investigate possible preferred material to be used depending on the length of the pipe. Maybe some materials are better suited for longest mains, for a matter of monetary affordability (lower price per unit of length) or physical characteristics. By grouping data from *pipes* dataset by the variable *matcat_id,* a *boxplot* could say which are length distribution within each class of material and where the median is located. Median has been chosen as the main indicator rather than average as some outlier could negatively affect the mean per each material. In addition, a *boxplot* may also tell more than a bar graph as it also shows the distribution and the dispersion of values around the median (1st and 3rd quartile values).

Figure 10 Boxplot: length distribution of pipes by material



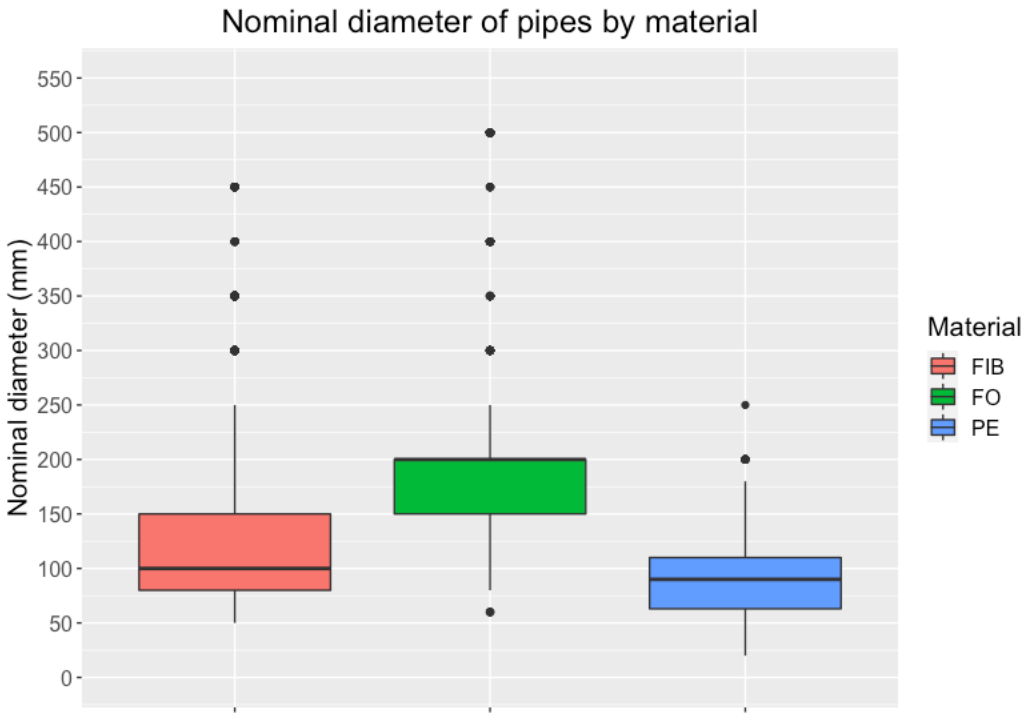
However, Figure 10 shows that pipes apparently measure homogeneously between materials, with very similar widespread distribution: there is not any material preferred basing upon the length of a pipe. To be meticulous, although medians are almost perfectly aligned between them, there is a slight difference among lengths of pipes made of *FO* and other material, as it seems as they are less widespread, and its 3rd quartile value (upper side of the green box) is slightly lower than for the other two boxes.

4.2.3 Material and nominal diameter

Once found that pipe length is not a determining variable for the choice of material, the attention falls onto the other main dimension of a pipe, that is the nominal diameter, variable *dnom* in *pipes* dataset.

The same approach has been adapted for this analysis, that is plotting a *boxplot* able to display simultaneously median and quartile distribution of values the variable *dnom* assume within material groups.

Figure 11 Boxplot: nominal diameter of pipes by material



Unlikely longitudinal length, pipe material is chosen depending on the nominal diameter of the pipes. Indeed, from Figure 11 it is possible to detect a clear upward trend for "FO" diameter dimension, compared to the other two materials. In addition, "PE" box is narrower than "FIB" box, than in overall results to be the most widespread material in terms of diameter. Once brainstormed with the correspondent from *Aigües Manresa,* the reason behind this particular behaviour has been identified and it is merely based on economic assessments. Indeed, for smaller diameters, *polyethylene* (PE) is much cheaper than *ductile cast* (FE). However, this economic convenience disappears once diameter dimension rises, and since *ductile iron* is stronger than polyethylene, and so a better material at an equal price, it is preferred for biggest mains. Regarding *fiber cement,* the large area of the box is because when in the past it was almost the only one used material, both big and small dieter pipes were built with FIB, that therefore presents a wider distribution of nominal diameter values.

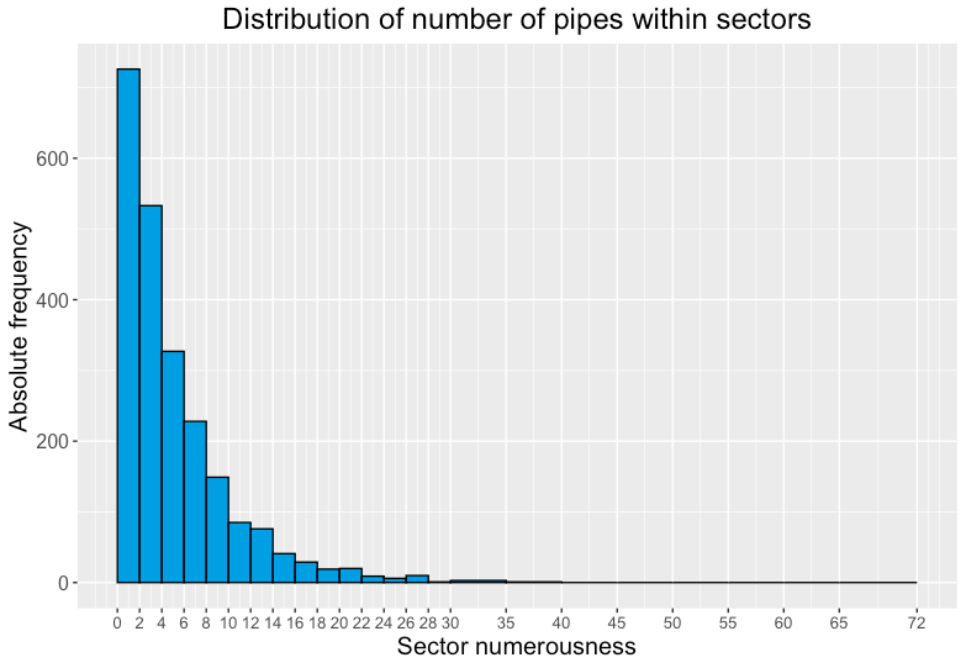
4.3 Sector

As already defined, a "sector" is the whole group of pipes that will be affected negatively by a burst of one of the pipes belonging to the sector itself. What could be an object of interest about sectors is to understand how many pipes are contained into sectors and which frequency a certain numerousness within a sector is repeated in the network. Indeed, if on one side a utopic scenario would be to have sectors made of only one pipe to reduce

minimally the spread of failures' impacts, on the other side, such scenario would lead to an extreme granular network and too complex to build and manage (due to very high duplication of resources). In a hypothetical graph, the "complexity curve" would decrease as the numerousness of sectors increases. On the other hand, the spread of failure's impact would increase as the numerousness of sectors increases. The optimal numerousness would be given by the interception of the curve where there is the perfect compromise between network complexity and risk.

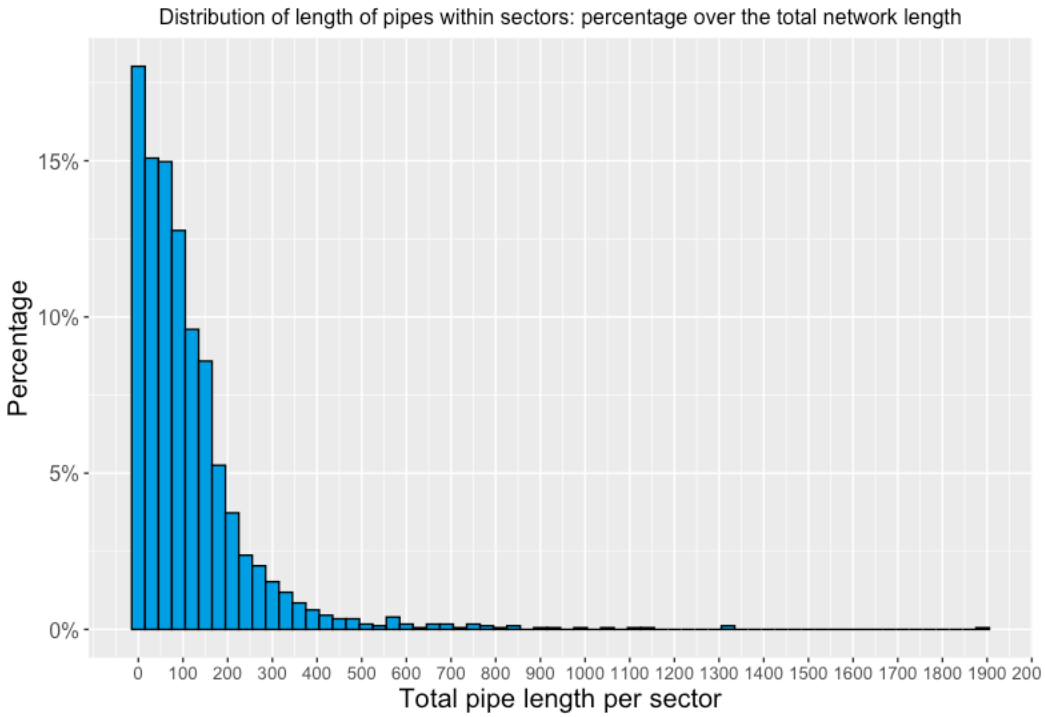
As previously done for the distribution of age of pipes, the methodology for showing how pipes are distributed within sectors is to calculate the absolute frequency of a certain numerousness of pipes in a sector (Figure 12).

Figure 12 Histogram: distribution of pipes within sectors



However, instead of calculating merely the frequency of the number of pipes per sector, once again, it is widely believed that length is a more robust indicator rather than frequency. In addition, in the following histogram (Figure 13), the length per sector will be displayed in terms of relative percentage over the total length of pipes.

Figure 13 Histogram: distribution of length of pipes within sectors: percentage over the total network length



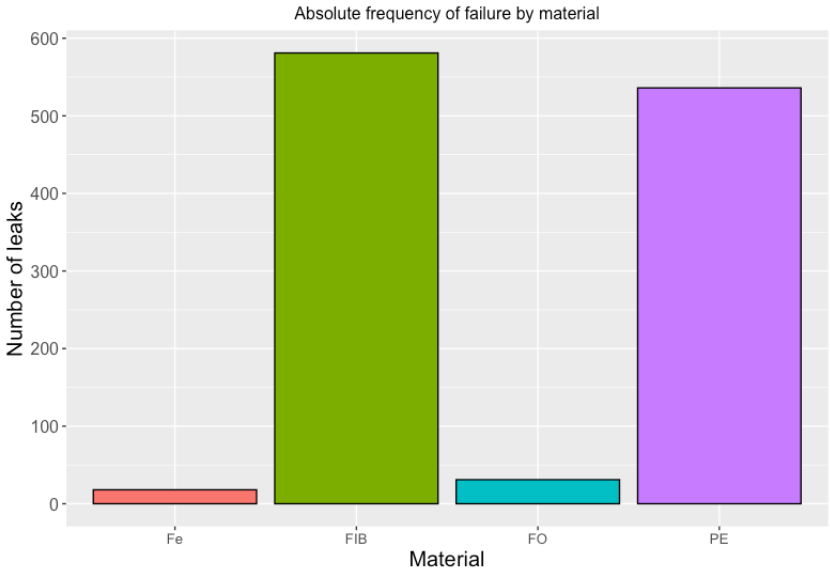
As visible from the histogram in Figure 13, the majority of sectors have less than 200 meters of pipes within them. In particular, in 75% of the cases, a failure involves less than 150 meters of the distribution network.

4.4 Leaks

4.4.1 Leaks per material

Once having explored interesting correlation among variables of *pipes* and *sectors* dataset, the focus of the study moves on the core dataset of this study, that is *leaks*.
The first preliminary visualization will regard the absolute number of failures per group of material within the dataset. By doing so, the reader would have a general idea which are pipe materials mostly populating the dataset of failures.

Figure 14 Bar chart: absolute frequency of leaks by material



Keeping in mind that “FIB”, together with “Fe”, pipes are the oldest in the network, while “FO” pipes are the newest, the bar chart Figure 14 shows as, in absolute terms, pipes in *fiber cement* mainly fail, followed by those in polyethylene. On the other hand, very few are the breaks of pipe made of “FO” recorded in the company’s database. However, to get a deep insight on failure rate by material, absolute failure figures should be compared to two values:

- Total number of pipes of each material group
- Total length of pipes of each material group.

In the first case, the ratio would give a relative break frequency, showing the percentage of broken pipes per each material, while the second one will compare rates of failure per meter of materials.

Figure 15 apparently shows that the relative frequency of failure has radical differences between materials, since “FO” has a very percentage of breaks compared to the other two materials, being almost *1/3* of the other percentages.

Although the gap in failures could seem very high, once comparing the number of failures throughout the network with the corresponding total extension of pipes, figures come out more aligned as shown in Figure 16. Indeed, the number of failures per unit of length rewards one more time pipes in *fiber cast* on a privileged position, that, as also confirmed by the company’s expectation, results to be the more resilient and resistant material, with a failure rate per meter just above 0.01.

However, results have to be interpreted carefully, otherwise risking falling into wrong conclusions. Indeed, pipes in *ductile cast* (FO) have started to be installed in the network mainly only from the 90s, as Figure 9 clearly shows. Therefore, on the date of data

gathering, approximately September 2021, their age is on average lower than pipes in *fiber cement* and *polyethylene*, as Table 1 clearly displays. So, by that time, the failure of pipes in *ductile cast* are expected to be fewer than for other materials.

Figure 15 Bar chart: relative frequency of leaks per material

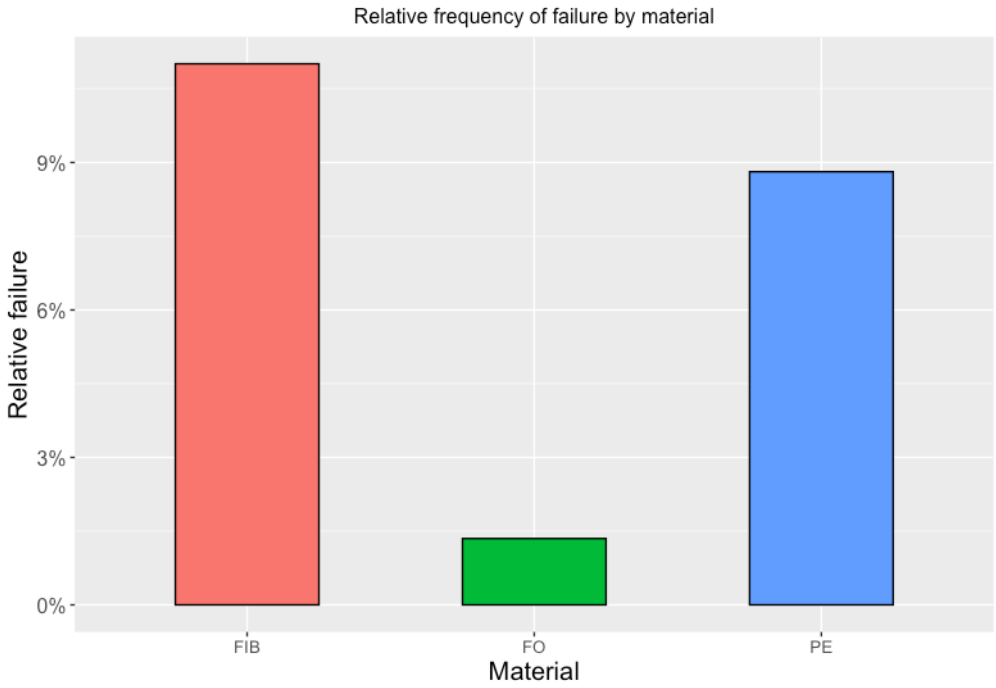
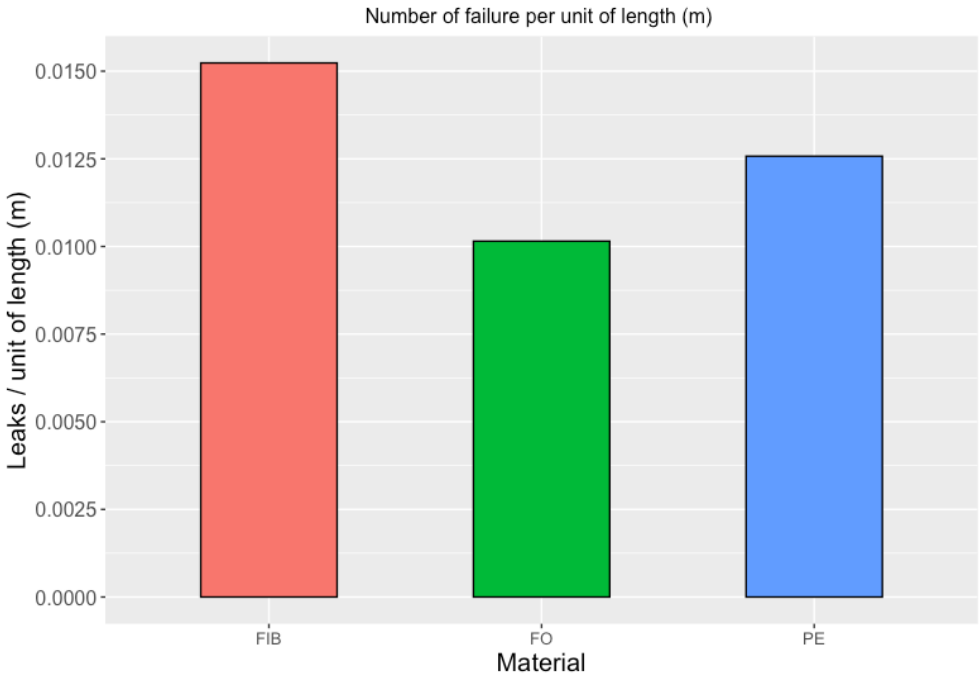


Figure 16 Bar chart: number of failures per unit of length (m)



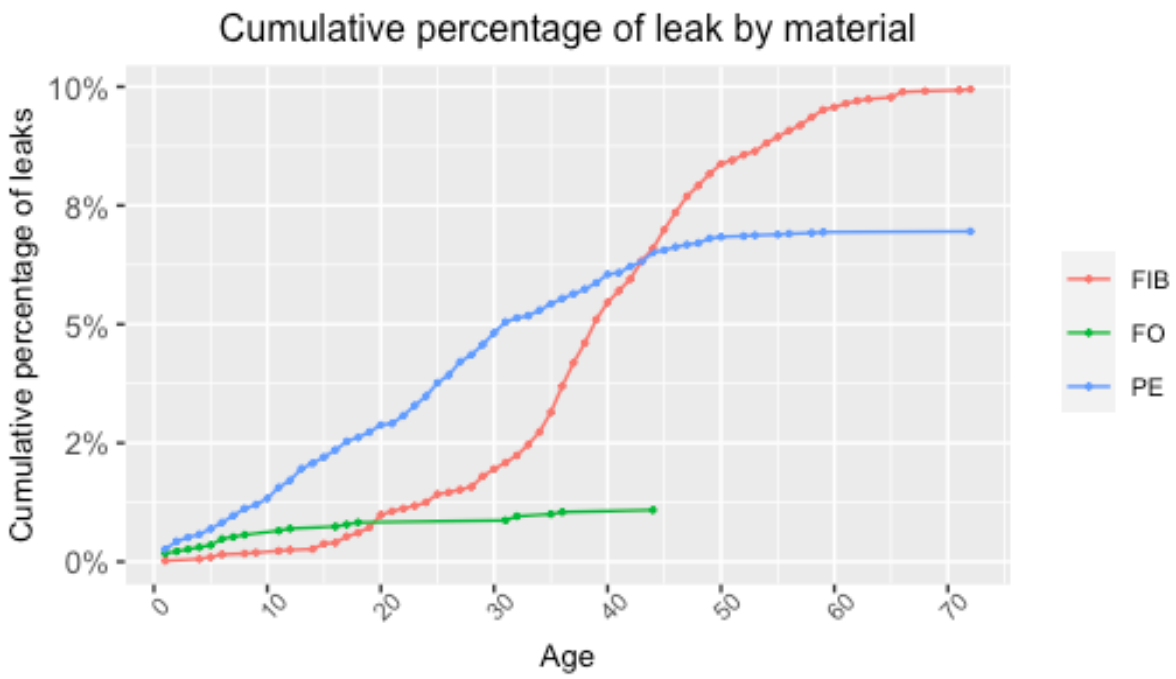
Material	Average age (year)
FIB	46.54
FO	16.22
PE	19.51

Table 1 Average age per material (year)

The proper analysis therefore should be carried on by comparing breaks rates of pipe with same age at the observation day.

After having calculated the age at which each pipe in the *leaks* dataset has experienced a failure, with the help of the formula `dplyr::group_by`, leaks have been grouped first by pipe age at the occurrence and then by material: the goal is to detect particular behaviour in failure trends. The goal is to understand how results in Figure 15 are obtained over years. Figure 17 shows the (cumulative) rate of failure per year. Basically, it answers the question "out of the total number of pipes of a certain material, how many got broken by an age of *x* years old?". Curves do not sum up to 100%, because they do not show the path towards the total number of failures, rather they end to a percentage that is equal to the ratio between total.leaks.per.material / total.pipe.per.material (values in Figure 15).

Figure 17 Cumulative percentage of leaks by material



Depending on the material, trend lines adopt really different shapes.

The slope of the cumulative percentage of leaks of pipes in *polyethylene* is quite constant over age, with the percentage of leaks increasing constantly as the age increases, up to 45 years old when the curve flattens. Indeed, whether in the first part of their life, "PE" pipes

"get old" quickly, once turning 45 years old, the deterioration process takes an opposite trend, up to end in a flat curve in the latest years. Differently, "FIB" seems to be more resistant in the first years of their life but falls in a dramatic deterioration from year 30 to 50, going even above the curve of "PE" once around 45 years old.

Eventually, "*FO*" pipes apparently have a very low and smooth cumulative percentage rate of failure, going up to 0.01 extremely slowly. However, this trend line is not enough reliable and comparable for two reasons:

- Pipes in ductile cast iron have been installed only starting from 1990; all the pipes aged more than 25 years represent exceptions.
- The sample size is less than half of the other two material categories.

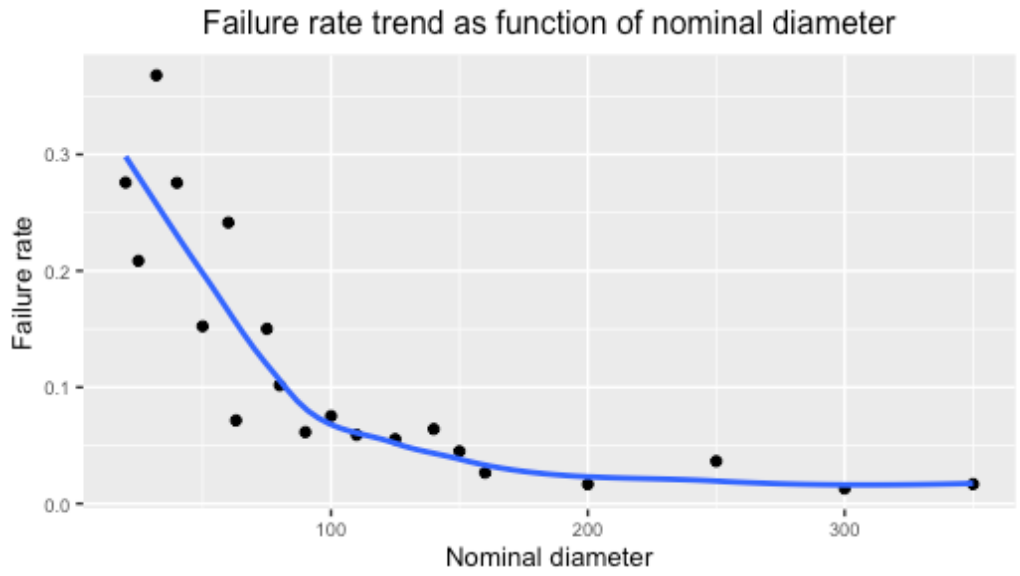
4.4.2 Leaks and nominal diameter

An important discovery regarding failures is explained in this sub-chapter, where the author has related pipes failure rate and pipes nominal diameter.

Figure 18 clearly shows as there is a strong relationship between pipe size and the probability of having a break. In particular, the trend of failure rate is downward as diameter increases. The failure rate is calculated as the ratio between the number of pipes with a certain diameter and the total number of failures for that diameter size.

The result reflects what the experience of employees of *Aigüe Manresa* predicted, since over years the trend was already noticed, biggest pipes used to break less than thinner ones.

Figure 18 Scatter plot and trendline of failure rate as a function of nominal diameter



4.5 Conclusions from data exploration

Even though the complexity of the data, it was possible catching some possible relationships between pipes characteristics and failures, such as the link between material and breaks or the downward of failures given the nominal diameter. However, unlike it was expected at the beginning of this study, the study of impacts of many variables has been omitted because apparently with no meaningful direct relationship with breaks. It is the case of the minimum, maximum and average pressure detected within pipes, as well as the nominal pressure and length. This does not mean those variables do not influence the probability of failing, but only that the effect cannot be seen by plotting a two dimensions graph. Or also, standing alone, variables such as pressure does not show any relationship, but if associated with other variables (e.g., length), within each group of length, pressure presents a noticeable relationship. However, during the next analysis about the prediction model, it would be possible to understand whether a variable actually has an impact on failures or not, but iteratively including and taking out the variable from the model and seeing how the goodness changes. If after taking out a variable the model is weaker, it means that the left-out variable does have a relationship with failures probability, even though was not possible to show it visually.

Eventually, how matching variables to outline potential trends in pipes failure is a topic left out for future studies.

5 Prediction

The content of Chapter 0 has given the reader a meaningful understanding of how pipes react to different stimuli and variables and which of them may be more impacting on the survival of the network. The probability to have a break is not only a matter of randomness, rather the generation of a regression model may attempt to explain some of the variability leading to a failure and to predict future breaks based on pipes variables.

5.1 Machine learning as a powerful tool

In chapter 2, the author reported some of the most famous work from the past in which the topic of prediction for water pipes failures has been undertaken. Most of them are based on traditional statistical and probabilistic methodologies, using parameters to estimate *time-to-next-break* or a *hazard function.*

However, the author of this study wants to follow a more recent trend in the management of assets with stochastic occurrences: Machine Learning (ML). For this purpose, *R* and *Rstudio* are very valuable tools, both for creating predictive models and analyzing the goodness of results.

Predicting in ML means generating output using an algorithm, chosen among many for its adaptability to the particular scope of the work. The algorithm will be trained and tested based on historical data for which complete information is reported, and once reached a certain desired level of goodness, the predictive model will be ready to be applied in future scenarios where the output variable is unknown to forecast the likelihood of that particular outcome [14].

Nowadays, there is a group of widespread machine learning algorithms, such powerful to be able to solve mainly any data problems (Ray, 2017). Following, the most common have been listed:

- Linear regression
- Logistic regression
- Decision tree
- Naïve Bayes
- K-means
- Random forest
- Gradient boosting algorithm

In cases such as the one interesting this study, where the goal is to predict when and with which probability a pipe will “die” due to a break, there is another specific branch of statistic, known as “survival analysis”, handling the matter. Although “survival analysis” lies the

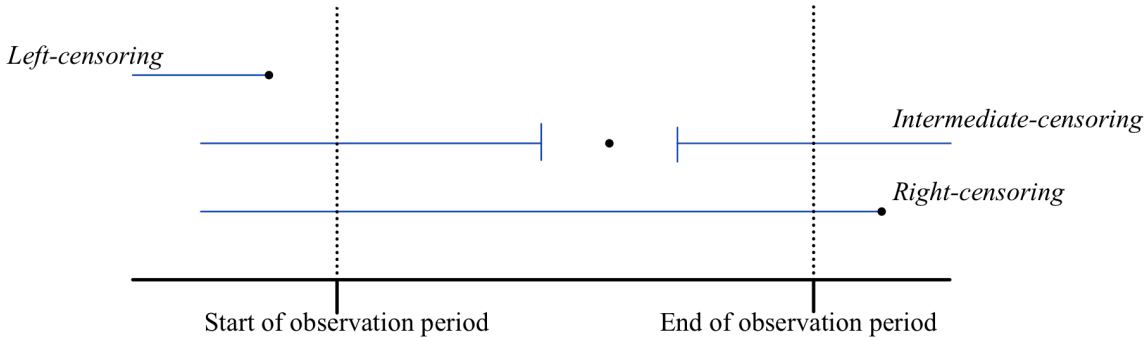
foundation on traditional statistical studies, that led to the development of non-parametrical (e.g. (Kaplan & Meier, 1958), (Nelson, 1972)), semi-parametric (e.g. (Cox R. D., 1972)) and parametrical methods (e.g. (Lee & Wang, 2003)), the latest frontier of survival analysis involves ML technics, with new developed algorithms such as survival trees, artificial neural networks and random survival forest. The main strength of these is to be able to handle successfully the issue with censoring and truncation that highly often "damage" datasets.

5.2 Censorship and truncation

"Samples obtained with data collection and /or observation is restricted over some portions of the sample space are, depending on the nature of the restriction, designated as either truncated or censored" (Cohen, 2020).

Censoring occurs when it is not known the exact time-to-event for an included observation and depending on when the event takes place, and the sample may be classified as left, interval or right-censored. When it is known that the time-to-event on an element of the sample is *less* than some value, the inclusion of this observation lead to left-censoring. The case of right-censorship has an opposite scenario, with time-to-event *greater* than some value. The last case includes both right and left censoring, with the time-to-event of an element of the sample *between* two specific values.

Figure 19 Type of censoring



The dataset used in this study suffers from censorship, regards the population of *pipes* dataset. Indeed, the inclusion of pipes experiencing leaks before 2005 is a straightforward example of left censorship, while all the pipes included in *pipes* set with no breaks recorded by 2020 represent instances of right-censoring observation. Moreover, the absence of intermediate-censorship cannot be absolutely stated as it is not exactly known whether, during the detection period, there was some time frame when leakages were not detected due for instance to technical issues. Since the author is not fully aware of this possible

scenario, it is assumed that during the time of observation, 2005 – 2020, no cases of intermediate censorship are present.

As long as the set includes pipes without recorded breaks because happening before 2005 or after 2020, the data set will be subjected to the censoring problem.

The *truncation* phenomenon is caused by the nature of virtue of time-to-event of some observations. Left-truncation happens when occurrences, with very short survival time, evade sampling, while right-truncation when the value is too big to be measured.

In this study, truncation regards the population of *leaks*, because all the leaks before 2005 were not included in the study, although pipes did have breaks before that temporal lower-bound: *leaks* dataset suffers from left-truncation.

Although the two problems represent an important obstacle in the resolution of such a problem analysis, the bias from right censoring can be resolved or mitigated by the use of a survival model because a survival model incorporates the right-censoring issue by its mathematical definition. Snider and McBean (2020) reconfirmed the advantages of a survival model by comparing it with machine learning algorithms that do not incorporate right-censored data. They concluded that removing censored events from the machine learning model results in predicting earlier pipe breaks than occur (Hao Xu & Sunil, 2021).

However, survival analysis would solve the problem of *pipes* dataset, but not the truncation affecting *leaks*, the application of these advanced methods would not be enough to deal with all the issues involved in this study. Therefore, with the approval of the department of *Aigues Manresa*, it has been decided to approach the study with traditional machine learning methodologies.

5.3 Choose of predictive methods

As the purpose of this study is to assess how ML methodology can improve the company's performance in asset management in Manresa and given the high complexity behind the concept of truncation and censoring, it has been agreed that traditional ML techniques might already give valuable insights. Later studies could then focus on the inclusion of problems such as censoring and truncation to go deeply into the problem.

5.4 Logistic regression as preliminary ML method

To perform the first prediction and see early results, logistic regression has been elected as the preliminary ML method.

Logistic regression is a learning classification algorithm, used to predict the probability of occurrence of the dependent variable, with a dichotomous nature: it can only assume values "0" and "1", respectively in presence of a failure or of a non-failure. By taking as input a list of variables *X,* logistic regression computes the probability that the output variable *Y* is 1*,* mathematically:

$$P(Y = 1) = f(X)$$

Equation 1 Logistic regression

Before creating the predictive model, training and testing it on the dataset, it is needed to re-modulate the given information to create a structure suitable to the object. In detail, given *pipes, sectors* and *leaks* dataset, and keeping in mind that *leaks* reports information about failures from 2005 to 2020, the final framework of the *break.history* data frame would be the following:

- For each pipes in *pipes* dataset, *break.history* will have a number of rows equal to 15 whether the installation year of the pipe is antecedent data collection beginning (2005) or equal to the difference, in number of years, between data collection end (2020) and installation year.

$$\#\boldsymbol{row}(\boldsymbol{i}) = \begin{cases} \mathbf{15}, & \boldsymbol{built.date} \leq \mathbf{2005} \ (1) \\ \mathbf{2020} - \boldsymbol{year.installation}(\boldsymbol{i}) + \mathbf{1}, & \boldsymbol{built.date} > \mathbf{2005} \ (2) \end{cases}$$

Equation 2 Calculation of number of rows in break.history dataset

- Column will be those already known variables included in all the three original datasets, plus two new columns:
  - *Year*, standing for "observation year", going from 2005 to 2020 in case (1) in Equation 2, or from *installation.year(i)* to 2020 in case (2).
  - *Failure,* a binary value assuming value *0* or *1* whether for the pipe *i* a failure occurred in the "observation year" *j*.

To generate *break.history,* that from now on will represent our information source for building predictive models, it was necessary to code a *for-if* cycle.

As explained, the cycle will calculate whether a failure occurred for each year within the range 2005-2019 in case a pipe was installed before 2005 or within the range *installation.year*-2019 in case of installation after the beginning of data collection.

In the end, considering all the pipes in the *pipes* dataset, a *break.history* dataset has been created with 195.660 rows. After another round of data cleaning to provide the new dataset with a proper format, and to delete critical NAs that would represent a threat to the effectiveness of predictive models, the path towards the creation of the predictive model may proceed.

The following table shows the frequency of values *0* and *1* within *break.history*.

VALUE	FREQUENCY
0	69532
1	237

Table 2 Frequency of observation with failures (1) and non-failures (0)

Based on these figures in Table 2, the baseline, indicating the percentage of occurrence of the event "failure" is calculated as:

$$baseline = \frac{f(1)}{\mathrm{f}(0) + \mathrm{f}(1)} = \frac{237}{69532 + 237} = 0.0034 = 0{,}34\%$$

Equation 3 Baseline equation

Only 0.34% of our dataset has a failure

Once calculated the baseline, the dataset *break.history* has been split into a *training* and *test* subset, both maintaining the same ratio of *1s* and *0*s in the *failures* column: the baseline value stays constant also in the new two subsets. In this case, subsets have been created with a proportion of 75% of observation in *training* and 25% in *test*.

It is time to create the logistic prediction model (`failure.log`) and to train it on the *training* subset. At this first attempt, all variables will be included as possible regressors, even though the significance of coefficients must be tested once the model is generated. However, for obvious reasons of collinearity, only one variable between age and decade will be held in the model. Indeed, they would express a linear relationship in the regression model.

The used formula for the generation of the logistic model is `stats::glm`, where it will be said to use the logistic approach by indicating *binomial* in the input "family".

```
failure.log = glm(failure ~ year + length + matcat_id + pnom + dnom + de
cade + max_pressure + min_pressure + avg_pressure, data = training, fami
ly = binomial)
summary(failure.log)

##
## Call:
## glm(formula = failure ~ year + length + matcat_id + pnom + dnom +
##     decade + max_pressure + min_pressure + avg_pressure, family = bin
omial,
```

```
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9741  -0.0856  -0.0665  -0.0448   4.2015
##
## Coefficients:
##                 Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)  145.7184358  34.5671444   4.216            0.0000249 ***
## year          -0.0578083   0.0170536  -3.390             0.000699 ***
## length         0.0074094   0.0007336  10.101 < 0.0000000000000002 ***
## matcat_idFO    2.7720457   1.1570127   2.396             0.016581 *
## matcat_idPE    0.4728648   0.3907540   1.210             0.226227
## pnom          -0.1272491   0.0382823  -3.324             0.000887 ***
## dnom          -0.0076448   0.0022821  -3.350             0.000808 ***
## decade        -0.0165595   0.0058483  -2.832             0.004633 **
## max_pressure  -0.0390213   0.0278516  -1.401             0.161201
## min_pressure  -0.0324052   0.0152625  -2.123             0.033738 *
## avg_pressure   0.0731541   0.0375840   1.946             0.051605 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2378.7  on 52326  degrees of freedom
## Residual deviance: 2131.2  on 52316  degrees of freedom
## AIC: 2153.2
##
## Number of Fisher Scoring iterations: 9
```

The first step after creating the predictive model is to go through regression coefficients and evaluate their significance. by displaying the summary of the built model, Rstudio shows coefficients significance with support of visual indicator, as following:

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a. significant level of 5%, we can keep all the regressors marked from 3 asterisks `'***'` up to those marked with a dot`'.'`. For all the other regressors, our training dataset does not provide enough evidence to reject the null hypothesis. For these variables, there is a zero correlation with the *failure* values. It is apparently the case of *max_pressure* and the binary *matcat_idFO.* Indeed, as *matcat_id* is a categorical variable assuming only 3 possible values (PE, FO and FIB) the algorithm generates 2 binary variables, *matcat_idPE* and *matcat_idFO,* assuming value “1” whether the pipe is made of the specific material. A row with both binaries “0” would be made of FIB. Regarding the significance, it seems as being made of polyethylene is determining for a failure, whereas FO does not. The variable m*ax_pressure* has also a p-value not enough low to reject the null hypothesis and a non-zero correlation with *failure.*

However, before taking out not significant regressors, we would like to see the results of this first model. By applying `failure.log` to the training dataset, results are the following:

```
predictTrain <- predict(failure.log, type = "response")
tapply(predictTrain, training$failure, mean)
          0           1
0.003374193 0.011456117
```

	Probability
0	0.0033711
1	0.0123585

Table 3 Probability of being predicted as a 0/1 when a failure occurred (training subset)

Table 3 must be read as following:

- If an observation did not experience a failure, the probability that it is predicted as a failure is 0.3%.
- If a pipe had in a certain year a failure, the logistic predictive model would assign a "1" with a probability of 1.2%.

At a first sight, the results may appear really weak and discouraging, but by thinking about the goal of this study, which is to understand whether and how ML can improve the company's economic performances, conclusions need to be taken out only at the end. Moreover, the 1 unit of magnitude difference between the two probabilities is already a sign of the efficiency of the model.

The next step would be to convert probabilities into predictions, by setting a threshold: if the probability of observation to be predicted positive is below the threshold, *failure* would assume value 0, otherwise, a leak would be predicted.

For the choice of the right threshold, marking the edge to assign *0* or *1* to each observation, the *receiver operating characteristic* (ROC) curve is useful. The curve helps decide the threshold by comparing sensitivity and specificity for different thresholds. It is possible to reduce the threshold value as long as the increase in *true positive rate* causes a less proportional increase in *false positive rate*. Once the side-effect of *false positive rate* increases more than how much the *true positive rate* rises, it is not favourable to reduce the threshold anymore. Geometrically, as long as the straight-line tangent to the curve has a slop higher than 45°, the threshold can be reduced.

Figure 20 ROC curve, first logistic model

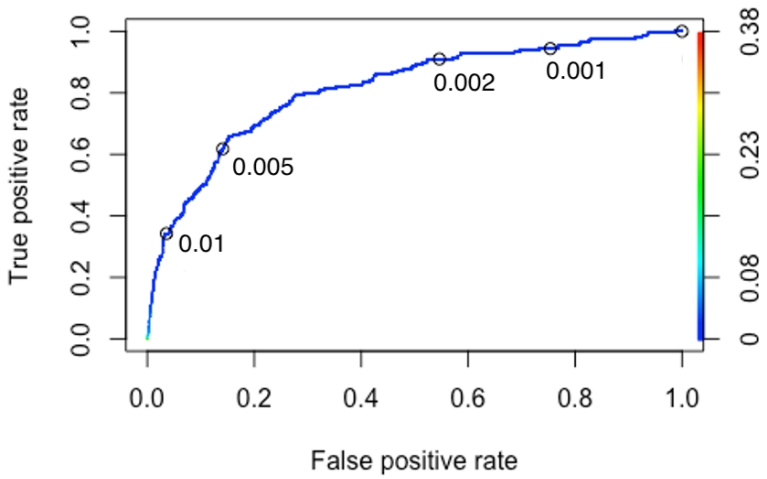


Figure 20 plot the ROC curve for the first logistic model and some cutoffs, from which 0.005 emerges to be the threshold for an optimal compromise of sensitivity and sensibility. Considering a threshold of 0.005, the prediction table is the following:

	FALSE	**TRUE**
0	44,699	7,450
1	77	101

Table 4 Aggregated output of prediction

- Rows (*0* and *1*): indicates actual values of observations.
- Columns (*FALSE* ad *TRUE*): output of prediction.

Out of a total of 52,327 observations in *training* subset, the preliminary logistic model has obtained the following results:

- Out of 52,149 observations without failures, 7,450 observations have been classified as TRUE, so as they experienced a failure (*false positive*).
- Out of 178 observations with failures, 77 have been predicted as without failures (*false negative*).

To efficiently summarize Table 4 figures and predictive performances, recalling some performance indicators would be useful for a deep understanding.

$$Specificity = \frac{True\ negative}{True\ negative\ +\ False\ positive}$$

Equation 4 Specificity equation

The meaning of *specificity* (Equation 4) is to show the capacity of the predictive model to predict negative values (without a leak) over the total actual observation without failures.

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative}$$

Equation 5 Sensitivity equation

Sensitivity (Equation 5) explains the portion of actual pipes with failures that our model has been able to predict as such.

$$Accuracy = \frac{True\ Negative + True\ Positive}{Total\ \#\ of\ observations}$$

Equation 6 Accuracy equation

The last indicator is the *accuracy* (Equation 6) shows to which extent the predictive model has been able to predict in aggregate correctly the observation.
In our case, the results are the following.

Indicator	Value
Specificity	0.8571401
Sensitivity	0.5674157
Accuracy	0.8561546

Table 5 Performance from training the first logistic prediction

For being the first predictive model, results are more than satisfying, although the first feeling after obtaining figures in Table 4 seemed discouraging. Furthermore, recalling the basic goal of this study, which is to plan maintenance supported by prediction, figures from Table 5 could be interpreted with a different reading key, providing additional relieving information.
Indeed, the total number of predicted failures (sum of figures in column *TRUE*) is 7,551, summing up for only 14.4% of the observations. The number of real failures that would be discovered during these 7,551 maintenance interventions is 101, accounting for 56,7% of actual total network failures. Considering a measure *"pipe x year"*, by working only on the 14.4% of the total *pipe x year,* the company would fix the 56% of the total actual failures in the distribution network over the same time frame.

5.4.1 Testing the model

Once training the model and having found such a positive outcome, it is time to test `failure.log` and compare results from training and testing. Basically, the same methodologies adopted for training will be followed for the testing stage, heading to computing the same indicators of performance as before.

```
predictTest <- predict(failure.log, type = "response", test)
tapply(predictTest, test$failure, mean)
```

	Probability
0	0.0033711
1	0.0123585

Table 6 Probability of being predicted as a 0/1 when a failure occurred (test subset)

Probabilities to predict a pipe with at least a failure in a certain year as "without failures" or "with failures" are perfectly aligned with probabilities in Table 3.

Indicator	Value.test
Specificity	0.8699304
Sensitivity	0.5423729
Accuracy	0.8688224

Table 7 Performance from testing the first logistic prediction

Eventually, the last step for validating the model is to compare results with those from the training.

Indicator	Training	Test
Specificity	0.8689716	0.8699304
Sensitivity	0.5280899	0.5423729
Accuracy	0.8678120	0.8688224

Table 8 Comparison between training and test indicators, first logistic model

Although the training subset is only 25% of the total sample dataset, prediction indicators are absolutely aligned with those from the training. The logistic model is solid and enough powerful for being applied in a company real situation for supporting decision making, in particular maintenance policies.

5.5 An amendment to the first logistic model

As already mentioned right before starting building the predictive model, once generated *break.history* dataset, a new round of data cleaned has been necessary, especially for dealing with *NAs*. Indeed, by default, most of the regression models in R work complete information and missing values can be problematic. Therefore, once deciding which variables were candidates as regressors, *break.history* was adjusted in order to erase any inconvenient *NA* occurring in those variables. In particular, dataset dimension dropped from more than 190 thousand to only around 67 thousand, mainly because of *pnom* variable,

presenting around 113 thousand missing values. This deduction of rows has dramatically reduced the sample size, and in this subchapter, the objective is to evaluate the effect of taking out *pnom* as a regressor and keeping a larger dataset. The idea is to give up a variable with a relevant significance (it is evaluated as a "three asterisks" regressor of the 1st logistic model) but gaining a larger sample where training and testing the model.

In addition, by displaying which variables has a coefficient of collinearity of variables higher than 0.7, related variables will be also taken out from the model: once again, the goal is to build a more simple but pure model.

In the following code and R output, *TRUEs* mark pairs of variable with collinearity higher than 0.7, which is a reasonable threshold for excluding regressors.

```
abs(cor(select(break.history, c(4, 6, 7,10, 11, 12, 13, 14)))) >0.7
```

```
##          length  dnom state   age max_pressure min_pressure avg_pressu
re
## length     TRUE FALSE FALSE FALSE        FALSE        FALSE     FALSE
## dnom      FALSE  TRUE FALSE FALSE        FALSE        FALSE     FALSE
## state     FALSE FALSE  TRUE FALSE        FALSE        FALSE     FALSE
## age       FALSE FALSE FALSE  TRUE        FALSE        FALSE     FALSE
## max_pressure  FALSE FALSE FALSE FALSE    TRUE         TRUE      TRUE
## min_pressure  FALSE FALSE FALSE FALSE    TRUE         TRUE      TRUE
## avg_pressure  FALSE FALSE FALSE FALSE    TRUE         TRUE      TRUE
## failure  FALSE FALSE FALSE FALSE        FALSE        FALSE     FALSE
##              failure
## length         FALSE
## dnom           FALSE
## state          FALSE
## age            FALSE
## max_pressure   FALSE
## min_pressure   FALSE
## avg_pressure   FALSE
## failure         TRUE
```

In particular, as could be expected, *max_pressure*, *min_pressure* and *avg_pressure* have a linear relationship with each other. Therefore, by attempting of keeping one of them in the model and taking out all the others and assessing the effect on significance and goodness of the model, we have ended up to the conclusion to only keep *avg_pressure* included as *a* regressor.

Following, the result from summarizing logistic model indicators:

```
failure.log = glm(failure ~ year + length + matcat_id + dnom + decade +
avg_pressure, data = training, family = binomial)
summary(failure.log)
```

```
##
## Call:
## glm(formula = failure ~ year + length + matcat_id + dnom + decade +
##     avg_pressure, family = binomial, data = training)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.5193  -0.1124  -0.0868  -0.0644   3.9108
##
## Coefficients:
##                 Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)  181.0567981  18.2527749   9.919 < 0.0000000000000002 ***
## year          -0.0723028   0.0087113  -8.300 < 0.0000000000000002 ***
## length         0.0081544   0.0003599  22.659 < 0.0000000000000002 ***
## matcat_idFO   -0.8620553   0.2499758  -3.449             0.000564 ***
## matcat_idPE    0.0939436   0.1089888   0.862             0.388712
## dnom          -0.0053877   0.0009063  -5.945      0.0000000027648 ***
## decade        -0.0207721   0.0031416  -6.612      0.0000000000379 ***
## avg_pressure   0.0313787   0.0075489   4.157      0.0000322831276 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8452.3  on 124504  degrees of freedom
## Residual deviance: 7676.5  on 124496  degrees of freedom
## AIC: 7694.5
```

Once again, the binary *matcat_idPE* does not pass the *non-zero* correlation with *failures.* Let's check results from this tentative and let's try to give a comparative insight with what developed previously.

	Probability
0	0.0054117
1	0.0160062

Table 9 Probability of being predicted as a 0/1 when a failure occurred (training subset), 2nd model

Also in this case, the chosen threshold was 0.005, since below this value

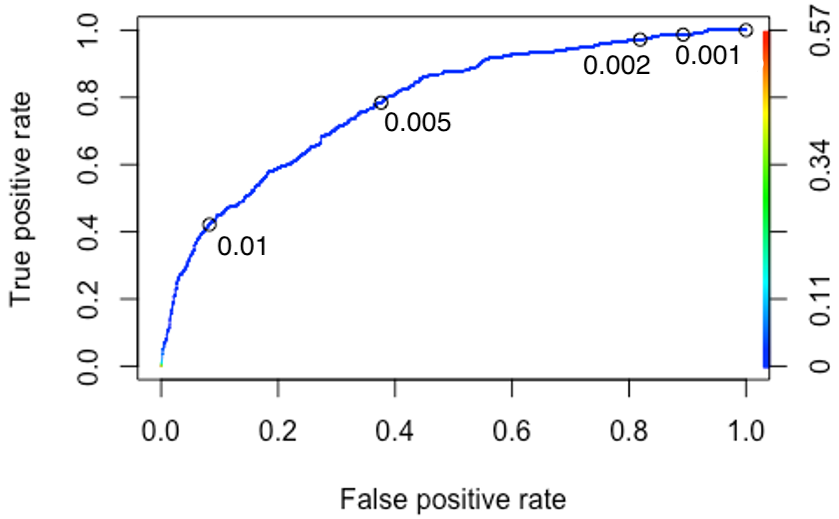


Figure 21 ROC curve, second logistic model

	FALSE	TRUE
0	77,415	46,409
1	155	526

Table 10 Aggregated output of 2nd prediction, training

Given figures from Table 10, it can be first said that by acting, in a certain time frame, on the 37% of the total *network x years,* 77% of the total failures of the corresponding time frame will be found and maintenance actions are taken.

Indicator	Value
Specificity	0.6252019
Sensitivity	0.7723935
Accuracy	0.6260070

Table 11 Performance from training the 2nd logistic prediction

Before, drawing some conclusions, let's extract specificity, sensitivity and accuracy from the testing process to validate the progress of the algorithm's training

Indicator	Training	Test
Specificity	0.6252019	0.6242186
Sensitivity	0.7723935	0.7929515
Accuracy	0.6260070	0.6221296

Table 12 Comparison between training and test indicators, 2nd prediction

Once all needed indicators are obtained, it is time to make conclusions and compare results.

Indicator	Training 1st	Test 1st	Training 2nd	Test 2nd
Specificity	0.8689716	0.8699304	0.6252019	0.6242186
Sensitivity	0.5280899	0.5423729	0.7723935	0.7929515
Accuracy	0.8678120	0.8688224	0.6260070	0.6221296

Table 13 Comparison first and second logistic prediction indicators

Going from the first logistic model to the second one generated taking out *pnom* and *min_pressure* and *max_pressure,* the model has lost power in specificity (more false-positive detected) but it has considerably gained sensitivity, being able to better detect failures. Therefore, in a true economical business scenario, a company should first estimate costs of intervention and costs of neglecting a failure. In case it is more worth to "waste" resources in acting on false-positive pipes but detecting more actual leaks, the second logistic model should be taken into consideration to support maintenance strategies. In the other case, where the cost of placing a maintenance action is too high compared with the

cost of leaking, it is better to adopt the first model, where fewer interventions will be vain because of false-positive occurrence.

5.6 Random forest

Once ultimate the generation of predictive models based on logistic regression, the second machine learning method applied in this study is *random forest*. It is a classification algorithm consisting of many decisions trees. By setting the number of trees in the *forest* and the minimum number of elements within each *leaf,* each tree releases a class prediction and the class with the most votes in the *forest* will be set as the final prediction. The power of random forest relies mainly on the ability of trees to protect each other from their individual error. Indeed, for giving a wrong output, the 50% + 1 of trees must go in the wrong direction, headed by error. (Yiu, 2021).

The only important step in data treatment required by *random forest* is to change the output format, applying the formula `base::as.factor`, transforming failure into a factor: it has not to be forgotten as random forest is a classification algorithm.

The R formula generating the predictive model is `randomForest::randomForest` the parameters to be specified are the number of trees in the forest (`ntree =`) and the minimum dimension of an ending node (`nodesize =`).

The number of trees usually rises as much as the sample size increases; therefore, the number of trees will be set to 2800 and the minimum number of elements inside each leaf to 10. Regarding this last point, a small chapter has to be open. Indeed, setting a too large *minbucket* (technical name) the model would be too simple, while if too small can bring overfitting in the model. In the latter case, the built model fits perfectly well the training set but once applied to another sample (test subset), it does not guarantee to do equally. However, in our case, it is necessary to set a relatively low *minbucket* as, otherwise, the algorithm would be too weak in predicting. The main cause is due to the data structure, where pipes with failure in a certain year are only 0.5%, therefore the model must go into deep detail to flush failures out. Luckily, there is some developed statistical technique able to solve this potential issue, such as cross-validation

The base used dataset will be the version of *break.history* where rows containing *NAs* values in the column *pnom* are not taken out, since this variable will not be included in the model as a regressor, following the same path as done for the logistic regressions.

```
training$failure <- as.factor(training$failure)
test$failure <- as.factor(test$failure)
failure.RNDFor <- randomForest(failure ~ year + length + matcat_id +
dnom + age + decade + avg_pressure, data = training, ntree = 2800,
nodesize = 10)
```

Once built the model, it is time to apply it to the training subset to assess its predictive ability.

```
predict.forest.training <- predict(failure.RNDFor, training)
```

Table 14 and Table 15 displays the output of training the random forest model:

	FALSE	TRUE
0	123,824	1
1	653	28

Table 14 Aggregate output random forest, training

Indicator	Value
Specificity	0.9999919
Sensitivity	0.0411160
Accuracy	0.9947472

Table 15 Performance from training random forest model

Results are divergent. Indeed, on one side, the model is absolutely able not to fall in false-positive prediction, with no case of failure predicted when they occurred, specificity equal to 1, as shown in Table 15. However, random forest-based model has a really weak sensitivity, as only 28 out of 681 actual failures are detected, while the remaining 653 failures are classified as *FALSE* (false negative). However, going through the testing phase is required, before starting amending the model to reach better goodness.

	FALSE	TRUE
0	41,273	1
1	217	10

Table 16 Aggregate output random forest, test

Indicator	Value
Specificity	0.9999757
Sensitivity	0.0440528
Accuracy	0.9946989

Table 17 Performance from testing random forest model

Testing the model has led to validating outcomes since all the performance indicators are aligned with those from training, which mean that no overfitting has "infected" the random forest. However, tools such as cross-validation can help to better choose a proper *minbucket,* heading to potential improvement of the model.

5.6.1 K-fold cross-validation

A widely used technique in data analysis to increase the goodness of ML models, especially when struggling with classification algorithms such as *random forest*, is cross-validation. Given the logic behind the statistical method, cross-validation is suited for choosing optimal parameters such as *minbucket* in *random forest.*

Basically, the algorithm is based on splitting the entire sample into *k* equally sized subsets (here the origin of *k-fold*), and, for each of the *k* group, doing the following general procedure:

- Take the group as a test dataset.
- Take the remaining *k-1* groups for training the model
- Fit the model on the training set and evaluate it on the test set
- Retain the evaluation score and discard the model

(Brownlee, 2021)

This aforementioned procedure is iteratively repeated *k* times and results are eventually summarized to show for example how accuracy moves up and down depending on a minbucket-size parameter. Each observation of the mother-set must be assigned to only one to the *k* group: in other words, each observation is used once in the testing process and *k-1* times for testing the model.

In this study, the author has opted for 10-fold cross-validation has been adopted, as a common use is to have a test group accounting for around the 10% of the original set. Required packages and functions are `caret::trainControl`, `caret::train` . In particular, the cross-validation is recalled checking how accuracy changes in the function of the "complexity parameter" (*cp*) parameter. The concept behind *cp* is the same as *minbucket*, involving the minimum number of observation final nodes of trees. Unlikely *minbuket*, assuming values that increase as the minimum number of elements in a *leaf* increase, *cp* works in the other way around: higher is the value, lower is the numerousness of a node.

By comparing the relationship between *accuracy ~ cp*, the optimal found *cp* value to maximize model accuracy is equal to 0.01: a new random forest model is therefore created, setting a complexity parameter at 0.01.

Results from the training section (Table 18) are literally amazing: with no prediction of *false-positive*, the algorithm has been able to predict correctly almost the 25% of the real failure in the network. However, by applying the model to the test dataset, although an improvement (Table 19) compared to the first attempt with random forest model, the author recognizes an obvious problem of overfitting. Indeed, sensitivity from the testing step is almost 1/3 of the incredible value from the training (Table 20), sing that likely, the model has been tailored too much to the data of the training sample. In fact, on that subset, its

predictive power is amazing, but once applied to a different dataset, the model loses part of its goodness.

	FALSE	TRUE
0	123,824	0
1	511	170

Table 18 Aggregate output from random forest after cross-validation, training

	FALSE	TRUE
0	41,269	5
1	209	18

Table 19 Aggregate output from random forest after cross-validation, testing

Indicator	Training	Test
Specificity	1.0000000	0.9998789
Sensitivity	0.2496329	0.0792951
Accuracy	0.9958957	0.9948435

Table 20 Comparison of performance from training and testing random forest model after cross-validation

Therefore, given the issue where the study fell into, it is needed to take out and keep in variables and rebuilt the regression structure of the model.

After an iterative procedure of "playing" with regressors, a potential structure that overcomes overfitting issue without losing predictive power has been found. By only omitting "*year*" from the list of regressors, the overfitting issue is solved, although the number of correct *true-positive* predictions drops.

Table 21, Table 22 and Table 23 show results from training and testing the new model, where besides a slight decrease in sensitivity, but negligible, nothing deserves to be commented because warning. Rather, almost maintaining the same specificity, sensitivity rose from a previous value of 0.033 to 0.11 during the training and to 0.08 during the testing! This is an incredible outcome demonstrating the extraordinary power of *k-fold cross-validation.*

	FALSE	TRUE
0	123,819	5
1	598	83

Table 21 Aggregate output from the 2nd random forest after cross-validation, training

	FALSE	TRUE
0	41,265	9
1	208	19

Table 22 Aggregate output from the 2nd random forest after cross-validation, testing

Indicator	Training	Test
Specificity	0.9999596	0.9997577
Sensitivity	0.1189427	0.0837004
Accuracy	0.9951408	0.9949158

Table 23 Comparison of performance from training and testing the 2nd random forest model after cross-validation

5.7 Simulating a more critical scenario

The main challenge of work such as the one this study attempts to address is the scarcity and not high accuracy of data, often affected by the phenomena of truncation and censoring. In other case, data collection procedures are not standardized, and the takeover of persons charged for gathering data lead to non-homogeneity of information.

Although results of previously validated data have been validated, passing successfully test phase, they are built on information about around 14,000 pipes that have experienced only 1,200 breaks from 2005 up to 2020. Therefore, it has been decided to emulate a more critical scenario by increasing the ratio of years with and without failure in *break.history.* perhaps, the model would not be appropriate to the case in analysis, but in a more quickly deteriorating distribution network (or sub-network).

Randomly, around 60 thousand rows with *0* as value of the variable “failure” are removed from the dataset, to raise the density of failures in the network over time. Whether Equation 3 show a baseline of 0.34% for the case of study, the percentage of failures over the sample size rises to 1.6%: this hypothetical network is 5 times more dramatic than the one in analysis.

Variables such as *pnom*, *min_pressure* and *max_pressure* are kept out from regressors not to infect the sample size by issues brought by *NAs*. *Training* and *test* dataset are created as always using respectively a split of 75% of total observations and 25%.

5.7.1 Logistic model

```
failure.log = glm(failure ~ year + length + matcat_id + dnom + decade +
avg_pressure, data = training, family = binomial)
summary(failure.log)

##
## Call:
## glm(formula = failure ~ year + length + matcat_id + dnom + decade +
##     avg_pressure, family = binomial, data = training)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1588  -0.1930  -0.1457  -0.1040   3.8991
##
## Coefficients:
##                  Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)  213.1788145  18.7474079  11.371 < 0.0000000000000002 ***
## year          -0.0879440   0.0089840  -9.789 < 0.0000000000000002 ***
## length         0.0098361   0.0004401  22.348 < 0.0000000000000002 ***
## matcat_idFO   -1.1812762   0.2821449  -4.187       0.000028295032 ***
## matcat_idPE    0.0169459   0.1114440   0.152              0.879142
## dnom          -0.0055953   0.0009165  -6.105       0.000000001029 ***
## decade        -0.0204928   0.0032144  -6.375       0.000000000183 ***
## avg_pressure   0.0100993   0.0029476   3.426              0.000612 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

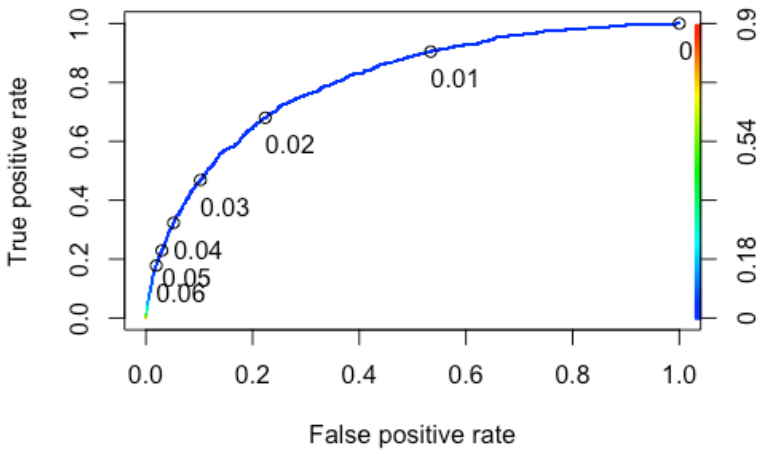


Figure 22 ROC curve, logistic model: more critical scenario

Looking at the ROC curve, a threshold giving an acceptable trade-off between true positive and false positive rate is right lower than 0.02.

Results of the prediction are shown in the following tables.

	FALSE	TRUE
0	32,039	9,113
1	222	459

Table 24 Aggregate output from logistic model in a more critical scenario, training

	FALSE	TRUE
0	10,818	2,899
1	76	151

Table 25 Aggregate output from logistic model in a more critical scenario, testing

Indicator	Training	Test
Specificity	0.7785527	0.7886564
Sensitivity	0.6740088	0.6651982
Accuracy	0.7768508	0.7866466

Table 26 Comparison of performance from training and testing logistic model in a more critical scenario

Given the higher density of failure, the algorithm can better outline a path towards failures of pipes and outcomes are by far above what previously found. Although techniques such as logistic and random forest attempt to define responsibilities in the deterioration process, there are too many factors affecting pipes life, some of them neither included in this study because of shortage of data (e.g., environmental and soil information). The difficulty is even hyperbolized since the topic involves events with rare probability, making the whole predictive process tougher. Once the occurrence of events becomes more frequent, the algorithm easier detects the relationship between the output variable and regressors, giving solidity to prediction and definitely better performances.

The following Table 27 compares performance from testing the 2nd logistic model built in this thesis and the last one coming from the critical scenario. In overall, the latter has higher accuracy, since in 78% of the cases the prediction match reality, against a 62% of the 2nd logistic model.

However, as already said, the aforementioned example cannot be applied to a general and well-working distribution network, but in a not steady scenario, such as in a very old network, it can find a proper application.

Indicator	2nd logistic	Critical scenario
Specificity	0.6252019	0.7886564
Sensitivity	0.7723935	0.6651982
Accuracy	0.6260070	0.7866466

Table 27 Comparison between performance of 2nd logistic and logistic in the critical scenario

6 Economic assessment

Chapter 5 has given a list of trained and tested ML models able to predict, with certain levels of accuracy, sensitivity and specificity whether a pipe will be affected at least by a failure in a certain year. These models can support the company's management in decision making for network maintenance, deciding when and how to perform interventions of reparation or replacement of pipes, towards a better allocation of resources, especially economic.

In section 2.3, the author already went through some of the approaches currently used for planning maintenance activities to optimize costs. However, in this specific case of study, those approaches do not find easy application, due to data structure and especially the output of the prediction. Indeed, the majority of the economic assessment methodologies rely on outputs of the type of "time-to-next-failure" or hazard-function saying the probability of failure over time.

For an economic evaluation of the impact of ML on maintenance strategies, the author proposes two different roads, one recalling a model already explained in the chapter of the literature review and an *ex-novo* approach.

The first of the two methodologies will be developed based upon the publication of Kleiner, Rajni and Nafi *"Planning renewal of water mains while considering deterioration, economies of scale and adjacent infrastructure"* and will have a more theoretical physiognomy, even though with very wide application. Indeed, any company, knowing their function of replacement cost, may be able to estimate saving due to the usage of a predictive model, by comparing costs in the absence and presence of predictive tools.

The second approach relies on a more realistic approach. Indeed, companies do not only place interventions based on the necessity to maintain the network, but they also need to take into consideration the maintenance budget allocated over a certain period. Therefore, with this second methodology, the aim is not to quantitively evaluate savings thanks to ML, rather to see how prediction can improve the efficacy of maintenance investment and assess how service level increases if compared to a baseline scenario without the support of predictive algorithms.

6.1 1st assessment model: theoretical methodology with general application

In particular, as Kleiner, Rajni and Nafi say in their publication (Kleiner, Nafi, & Rajani, 2010), there are savings a company could benefit from due to simultaneous maintenance interventions, due to mobilization of workers teams and quantity of purchased materials. The first factor comprises costs such as setting up the job site, signage and the gathering of all resources needed during a replacement intervention. Quantity discount depends on

pipe material, diameter, location and other circumstances (Kleiner, Nafi, & Rajani, 2010). Therefore, the cost of replacing a pipe of a certain length l_i *is:*

$$C^{rep}{}_i = M + C_i l_i$$

Equation 7 Replacement cost of a pipe

With Cr_i the replacement cost per unit of length.

The second term of the equation presents a fixed component *M* and a variable factor $C_i l_i$. The quantity discount applies to the variable component of pipe cost while the mobilization discount to the fixed component. Indeed, more are the replacement the company has planned to put in place at a given year, higher will be the discount on the unitary cost. The factor γ says what is the discount rate applied in case quantity discount may be applied so that the unit cost per length drops to $C_i \cdot (1 - \gamma)$.

Moreover, if the management opts for multiple replacements during the same period, all the actions composing the fixed voice of the cost *M* might be carried out only once, or at least not continuously replicated for every replacement. In the best scenario, for replacing *k* pipes, a company might save fixed costs amounting to $M \cdot (k-1)$, as if the fixed cost was supported only once for all the replacements. In a more realistic scenario, the fixed cost would be discounted by a certain factor λ, with $0 < \lambda < 1$, depending on what percentage of the amount $k \cdot M$ can be saved, reducing the fixed cost for replacing *k* pipes from $k \cdot M$ to $k \cdot M \cdot (1- \lambda)$.

It is straightforward as both factors λ and γ get into action as if talking about replacement as well as for preventive maintenance: they do not come from the nature of the intervention, rather from the concept of planning and scheduling the maintenance. Indeed, the worst company can do is to deal with leaks as they were stand-alone and to wait for their occurrence to react and remedy the damage: under this non-strategy, all the economic advantages of planning maintenance come less.

The last important aspect of pipe management is timing, especially when maintenance relies on prediction. If a pipe is expected to stop its useful life at a certain year, depending on which moment of the year the intervention is put in place, avoiding that the failure already has occurred leads to avoiding the cost of a failure, accounting for four main different factors (Kleiner, Nafi, & Rajani, 2010):

1. C_i^{dir}, cost for expected direct damage (e.g., to adjacent infrastructure, basement flooding, road damage)
2. C_i^{indir}, cost of indirect damage (e.g., accelerated deterioration of roads, sewers, etc.)
3. C_i^{wat}, cost of lost water due to the leaks
4. C_i^{soc}, the social cost (e.g., disruption, time loss, pollution, loss of business etc.)

However, in this study, only direct costs and cost of lost water are taken into account, due to difficulties to gather data explaining social and indirect costs.

Therefore, given a certain period of *x* years and its related prediction outcomes as the following:

	FALSE	TRUE
0	TN	FP
1	FN	TP

Table 28 General prediction outcome: TN True Negative, FN False Negative, FP False Positive, TP True Positive

Maintenance costs, depending on the adopted strategy, are the following:

$$(1). Total\ cost_{No\ Prediction} = (FN + TP) \times C^{rep}$$

Equation 8 Total cost in absence of prediction

$$(2).\ Total\ Cost_{Prediction} = FN \times C^{rep} + (FP + TP) \times (C^{rep} - \lambda M - \gamma C_i) - TP \times C^{wat} - FP \times \rho \times C^{wat}$$

Equation 9 Total cost using prediction

Equation 8 regards the scenario in absence of prediction, with the total cost only equal to the total number of actual leaks (*false negative* FN + *true positive* TP) times the unitary cost of replacement.

If the company decides to use ML for predicting future failures, the function of cost follows Equation 9. Following, an explanation of each term of the sum:

- $FN \cdot C^{rep}$ is the cost due to the not total ability of a predictive model to detect properly all the failures; it is the cost to repair pipes that are *false negative*. The unitary cost for these elements will be the same as the prediction was not performed, without any quantity or mobilization discount.
- $(FP + TP) \cdot (C^{rep} - \lambda M - \gamma C)$: The model suggests maintaining pipes expected to get broken, either they actually will fail (*true positive* TP) or not (*false positive* FP). For these pipes, all the economies of scale mentioned by Klainer, Rajani and Nafi will get into action, therefore the unitary cost will be reduced by the quantity and mobilization discount factors γ and λ. Therefore, the final discounted unitary cost, considering *C* the unitary cost of a pipe of average length (Crl_{avg}), is:

$$C_{dis}^{rep} = (FP + TP) \times [M(1 - \lambda) + C(1 - \gamma)] = (FP + TP) \times (M + C - \lambda M - \gamma C) = (FP + TP) \times (C^{rep} - \lambda M - \gamma C)$$

Equation 10 Cost or reparation in presence of economies of scale

- $TP \cdot C^{wat}$: if the company was not able to predict failures in time, bursts will happen, and loss of water would represent a cost. However, as the company may hypothetically avoid that some breakages (*TP*) occur, the cost would be a cost avoided, that can be seen as a saving.
- $FP \cdot \rho \cdot C^{wat}$: the same logic behind the previous term, for saving the loss of water for those failures the predict will avoid, stands behind the last element of the equation. Indeed, prediction models will also lead to intervention on pipes that would not need maintenance (*false positive*), and their replacement would save the company money for possible future breaks, even though discounted by a factor ρ.

All the cost and discount parameters in Equation 9, C^{rep} , C^{wat} , γ, λ and ρ , are subject to the serval aspects that may change from a company to another and from a geographical area to another one. Indeed, depending on how much is the cost of replacing a pip and how much can be saved by preventively replacing a pipe, making use of economies of scale and the cost of lost water, a certain predictive model may be better than another one. Therefore, to minimize the total cost of Equation 9, a company should first assess all its cost and savings parameters and based on these values, understand for which ratio of *specificity/sensitivity* the cost equation touches its minimum value.

For reading easiness, all the cost parameters of Equation 9 are recalled as following:

- $C^{rep} = \alpha$
- $C^{rep} - \lambda M - \gamma C_i = \beta$
- $C^{wat} = \theta$
- $\rho \cdot C^{wat} = \delta$

The cost optimization problem is reduced to solve the following system of equation and to minimize the total cost function:

$$\begin{cases} TN + FP + FN + TP = N \\ Sp = \dfrac{TN}{TN + FP} \\ Se = \dfrac{TP}{FP + TP} \\ TotalCost = \alpha \cdot FN + \beta \cdot (FP + TP) + \theta \cdot TP + \delta \cdot FP \end{cases}$$

Optimization 1 Original system of equations

With *N* standing for “sample size”, that would be the size of the network over year.

The next mathematical steps are the following:

1. Expressing the equation of total cost in the function of only one of the 4 instances of prediction (TN, FN, FP, TP); in our case, calculous are done by expressing all the four equations in the function of *FN*.
2. Calculating the derivative of the total cost with respect to TP.
3. Imposing the derivative equal to 0.
4. Express *sensitivity* as a function of *specificity* and all the cost parameters.

Step 1 of the previous list is developed in the following equations, where second members of all the equations are expressed only in function of *FN, Se, Sp* and sample size *N:*

$$\begin{cases} FP = N \cdot (1 - Sp) - FN \cdot \dfrac{1 - Sp}{1 - Se} \\ TN = N \cdot Sp - FN \cdot \dfrac{Sp}{1 - Se} \\ TP = FN \cdot \dfrac{Se}{1 - Se} \\ TotalCost(FN) = \alpha \cdot FN + \beta \cdot \left[N \cdot (1 - Sp) + \dfrac{Sp + Se - 1}{1 - Se} \cdot FN \right] + \theta \cdot \dfrac{Se}{1 - Se} \cdot FN + \delta \cdot \left[N \cdot (1 - Sp) + \dfrac{Sp - 1}{1 - Se} \cdot FN \right] \end{cases}$$

Optimization 2 System of equation expressed in function of FN

Given the function *TotalCost(FN),* step 2 is carried out as follows:

$$\frac{dTotalCost}{dFN} = [\alpha \cdot (1 - Se) + \beta \cdot (Sp + Se - 1) + \theta \cdot Se + \delta \cdot (Sp - 1)$$

Equation 11 First derivative of the "TotalCost" function

By imposing Equation 11 equal to *0*, the relationship between *sensitivity* and *specificity* for which the function of the total cost of replacement can be found.

$$\frac{dTotalCost}{dTN} = 0$$

Equation 12 Imposition of the first derivative of "TotalCost" function equal to 0

Translated into:

$$Se = \frac{\alpha - \beta - \delta}{\alpha - \beta - \theta} + \frac{\beta + \delta}{\alpha - \beta - \theta} \cdot Sp$$

Equation 13 Relationship between sensitivity and specificity to minimize "TotalCost" function

After re-substituting parameters α, β, θ and δ with the proper values of cost:

$$Se = \frac{\rho \cdot C^w - \lambda \cdot M - \gamma \cdot C_i}{\lambda \cdot M + \gamma \cdot C_i + C^w} - \frac{(1-\lambda) \cdot M + (1+\gamma) \cdot C_i + \rho \cdot C^w}{\lambda \cdot M + \gamma \cdot C_i + C^w} \cdot Sp$$

Equation 14 Optimal relationship of Sensitivity and Specificity in the function of a company cost parameters

Equation 14 says for which relationship *Sensitivity* and *Specificity* the *TotalCost* function is minimized, given all the cost and discount parameters characterizing the replacement framework of costs of a company.

Depending on their function of cost, a company can establish whether strong to detect failures even though falling into many *false positives* (more sensitivity than specificity) is more favourable than an accurate model in not wrongly predicting "healthy" pipes as failures.

Therefore, given this relationship between sensitivity and specificity, the most proper prediction model and parameters can be elected, either a logistic with a certain threshold, a random forest with a specific *minbucket* value, etc.

Once chosen the adapt ML model and the corresponding parameter and run the prediction algorithm, by comparing Equation 8 with Equation 9, the economic impact of ML can be measured.

6.2 2nd assessment model

The second proposed way of measuring the impacts of prediction on pipes management relies on a different approach. Indeed, the aim is not to compare costs with and without prediction to assess the savings, rather to evaluate, by keeping constant the maintenance investment, how the service level may change thanks to the usage of a predictive tool.

A general water distribution company, such as *Aigües Manresa*, allocates a specific budget for maintenance over a certain period and plans which pipes will be replaced, based mainly on managers' experience. The number of interventions depends on the allocated budget and on the cost of putting in place maintenance, including costs of workers, raw material, worksite set up and the usage of diggers.

Aigües Manresa has provided a set of data about costs of intervention on different pipes, from which it is possible retrieving an average unit cost for replacing a pipe. The author has been told that the 2 main factors affecting the cost of intervention are material and length. To get the final average cost, the dataset has been first filtered to only keep pipes made of "FO", "PE" and "FIB". Then, since no info regarding the length of pipes maintained was included in the dataset, the author has assumed that the distribution of pipes length is the same as in the analyzed network. Eventually, the mean of costs has been calculated, accounting for 450€. Therefore, the number of interventions that can be put in place is given by the ratio between the allocated budget and the unit cost of a replacement.

How does the impact of prediction on service level is measured? In the following two sections, the framework of the assessment will be explained, for both cases with and without the support of prediction.

Anyway, for both scenarios, the evaluation will be carried out for the period from 2015 to 2020. To evaluate the potential impacts of carrying out predictions, a comparison between predicted output and the actual value is needed. Therefore, the prediction has been done on data from the last 6 years and then real and computed output values have been compared.

Assuming a level of investment *I* and that the number of replaced pipes is homogeneously spread over the 6 years, the number of replacements per year is determined as following:

$$R = No.replacement/year = \frac{I(€)}{450(€/replacement)} \cdot \frac{1}{6}$$

Equation 15 Number of pipes replaced per year

Two important assumptions are made in this evaluation, and they will be kept for all the scenarios analyzed:

1. A failure occurring in a certain year would not take place if the pipe is already replaced in the years before the year of the break.
2. A pipe cannot be replaced more than once in the 6-years horizon. Therefore, after each year, the *R* replaced pipes are taken out from the "candidate" list for the following year. The assumption is not unrealistic, since less than 1 failure out of 10 occurs within 6 years from the installation year. In addition, 50% of the leaks in the first 6 years of the life of a pipe take place within the first 1.4 years, likely due to human mistakes. Therefore, assuming no errors during the installation process, the assumption is proper enough not to distort reality.

6.2.1 Replacement without support of prediction

In the case the company does not embrace the opportunity of prediction, it is preliminarily assumed that pipes are replaced randomly, by picking *R* pipes from the network each year. This operativity is far from reality as it never happens that the choice of which pipes to replace is left totally to the chance. In predictive maintenance, usually, managers decide based upon their experience, but this would not be possible to replicate and to model in a simulating study.

In the year 2015, *R* random pipes will be replaced from all the pipes present this year in the network. This intervention would avoid all the breaks from 2015 to 2020 happening to those *R* pipes (from now, *R* is the vector including the replaced pipes). Therefore, the efficacy of

the maintenance is calculated as the sum of failures between 2015 and 2020 occurring to *R* divided the number of pipes in the network in 2015.

In the year 2016, *R* pipes have been already replaced in 2015. Other *R'* random pipes will be replaced, but from the picking process, the first *R* cannot be selected. The efficiency of this second maintenance is calculated as the number of failures occurring from 2016 to 2020 to *R'* divided by the total number of pipes in the distribution network in 2016.

In the following years until 2020, the procedure is replicated but deducting during the picking process all the pipes already replaced in the prior years.

Ending this iterative calculation, the avoided failures will be detected. It is important now to recall the concept of "sectors", defined as the group of pipes that would be affected (shut down) by a failure in any of the elements belonging to it. Therefore, it is possible, knowing the leaks avoided, how many pipes have been saved from being shut down and the corresponding percentage of the network. Once again, the goal is to reduce at the least the downtime level of the network since a black-out in the water distribution may lead to economic and social effects, besides customers' inconveniences.

Picking pipes to be replaced given the date of birth

Although the just explained methodology is a good representation of a total random election, it goes away from the reality of operating, since a company would not only rely on randomness for replacing pipes. Therefore, in the second assumed scenario without the support of the predictive model, the selection is carried out with the purpose to replace given the age of pipes, starting from the oldest ones. This choice is based on the idea that the probability to get broken increases as age goes by. Therefore, each year, the *R* oldest the oldest pipes are replaced, keeping the same two assumptions of not replacing items more than once in 5 years and imagining a reality where a new pipe cannot fail again by 2020.

6.2.2 Replacement supported by predictive models

If the company decides to rely on prediction to decide which pipes to replace, the only difference, but substantial, with the case without prediction support is that the vector *R* will not be created either randomly or based on age. In fact, the vector will include the *R* pipes with the highest predicted probability to have a break in the 5 years. Therefore, the support of the model acts right in the step of picking pipes that is based on the predicted probabilities.

For the calculation of the probability of failure, any kind of ML predictive model can be used, as long as the output of the model is a probability. In this thesis, the author will use the RandomForest predictive model, improved by *cross-validation*, rather than the logistic.

6.2.3 Results

Scenario 1

In the first simulated scenario, the investment level *I* is set at €540,000, a level capable to allow the replacement of 1200 pipes over 6 years, with *R=200* according to Equation 15, which means interventions on almost 2% of the distribution network each year.

Each pipe that is preventively saved by maintenance avoids that the sector to which the leak belongs is involved in the failure, decreasing distribution service level. Since each sector includes a specific number of pipes, a sector shutdown means stopping the distribution alongside that portion of the network belonging to that sector.

The goal is to understand which portion of the network will be safeguarded from disruption of the distribution service, obtained in the following way:

- The number of *sectors involved* represents the list of sectors with pipes prevented by leaks. In case a sector would have more than one pipe with failures, it will be counted as many as the number of predicted failures because the distribution for the sector will be interrupted multiple times (assuming not a timing overlap of failures).
- The multiplication of the number of sectors with failures by the number of pipes into each sector says how many pipes will be prevented to be shut down because of sectors disruption. This value would represent the “saved network”.
- The percentage of the “saved network” is given by the ratio between the number of pipes saved in each year and the total number of pipes in the network in the same year.

	USING PREDICTION		
YEAR	*No. sectors involved*	*No. of pipes saved*	***% NETWORK***
2015	29	268	2.504
2016	25	225	2.072
2017	24	220	1.986
2018	29	268	2.325
2019	19	181	1.59
2020	8	94	0.815
	134	**1256**	**1.882**

Table 29 Result from service level assessment (I = €540,000)

	WITH PREDICTION	WITHOUT PREDICTION	
YEAR	% Network	% Random selection	% Age selection
2015	2.504	0.028	0.14
2016	2.072	0.037	0.064
2017	1.986	0.09	0.135
2018	2.325	0.026	0.062
2019	1.59	0.035	0.13
2020	0.815	0.139	0.061
AVG	**1.882**	**0.059**	**0.099**

Table 30 Comparison of different scenarios of service level assessment (I = €540,000)

Table 29 clearly shows how prediction may positively impact the level of service of the distribution. Over the years, 134 breaks that would occur in 6 years, might be prevented and the corresponding sectors not involved in a shutdown, accounting for more than 1250 pipes not involved in a failure occurrence. Even though the efficacy goes down over years, on average 1.882% of the network is saved each year between 2015 and 2020. This percentage is by far above what can be reached by adopting the other two methodologies of selecting pipes to replace, both random picking and based on age, as figures in Table 30 show. In particular, it is more than 30 times higher than the random-selection scenario and almost 20 times higher than the selection by age. As expected, replacing the oldest pipes is a method leading to a higher efficacy (0.099% of the network saved on average) than randomly substituting 200 pipes (only 0.056% on average).

If a company was able to retrieve a cost per each percentage of shut down network τ, it might also be understood how much a replacement strategy, driven by predictive models, can save extra cost due to unexpected events such as leaks.

Changing investment level – scenario 2

As said at the beginning of Section 6.2, results from the 2nd assessment method show the impact of predictive model *ceteris-paribus*, which is under the same investment budget.

By acting on the 2% of the network each year (200 pipes over the total network), the average network prevented to fall into a shut-down thanks to prediction is 30 times more than a random selection and around 20 times more than a replacement strategy by age.

The goal of this section is to see what happens when changing the allocated budget. Here, a budget equal to *I* = €1,350,000, that has been set such to allow the replacement of 3000 pipes in 6 years, that means 4.6% of the network each year.

		USING PREDICTION	
YEAR	*No. sectors involved*	*No. of pipes saved*	*% Network*
2015	56	522	4.877
2016	44	417	3.84
2017	40	389	3.511
2018	30	289	2.539
2019	26	258	2.238
2020	11	120	1.041
TOTAL	**207**	**1995**	**3.007**

Table 31 Result from service level assessment (I = €1,350,000)

	WITH PREDICTION	***WITHOUT PREDICTION***	
Year	*% Network*	*% Random selection*	*% Age selection*
2015	4.877	0.168	0.234
2016	3.84	0.212	0.111
2017	3.511	0.09	0.226
2018	2.539	0.079	0.105
2019	2.238	0.104	0.217
2020	1.041	0.156	0.104
AVG	**3.007**	**0.135**	**0.166**

Table 32 Comparison of different scenarios of service level assessment (I = €1,350,000)

207 are the sector not disrupted because in failures that without preventive maintenance based on prediction would have happened, involving 1995 pipes. On average, 3.007% of the network would be saved from disruption each year.

The level of investment is 2.5 higher than the first scenario when it was set at €540.000. However, the effects of this strong financial increase had not the same impact on replacement performance. Table 33 shows that the percentage of network "saved" by the predictive model is only 1.6 times higher than in the first case study. Therefore, it seems the positive impacts of being supported by a predictive model reduce their effect the more the investment budget goes up. Let's analyze the results from Table 33:

- With *I=€540,000,* each €1,000 of investment prevents 2.3 pipes to stop serve because of a shutdown.
- Increasing the budget by 150% leads to an increase of 59% of the network saved with respect to the first investment, with almost 1.5 pipes prevented to be closed for each €1,000 of extra budget.
- Eventually, with a further increase of budget by 100%, from 1,350,000 to 2,700,000, the increase in network saved accounts for 25% compared to the second investment, with only less than a pipe saved for each €1,000 of investment.

INVESTED BUDGET	*Δ budget*	*Pipes saved*	*Δ Pipes*	*Pipes saved / €1,000 Invested*
540,000		1,256		*2.326*
1,350,000	150%	1,995	*59%*	*1.478*
2,700,000	100%	2,488	*25%*	*0.921*

Table 33 Comparison of performances of different investment levels

The optimal value of investment to take advantage of the power of predictive model can be found by studying the curve % *Network saved – invested budget* and to identify the maximum of the curve *I*.* However, if the company was aware of the cost τ, previously defined as the cost of shutting down 1% of the network, the economic optimal value of investment *I*** may change, since it is given by the intersection of the curve of invested budget and money saved per each percent of network saved. Indeed, the capital a company allocates for pipe replacement can even go beyond I*, as long as the money saved by preventing failures is higher than the invested budget.

7 Summary of results

7.1 Budget summary

As deeply detailed in the attached document “Budget”, performing such work cannot be considered negligible from an economic standpoint. The following table shows the total cost estimated to carry out such a study, totally due to pay people in charge for the work. In fact, personnel expenses would be the only cost to bear, since the absence of drafting any physical design. Moreover, software used for the study do not need license.

Type of cost	Total time	Total cost (€)
Personnel expenses	816 h	8,160.00

Table 34 Cost assessment

However, given the nature of the job, most of the job needs to be done once and stays valid for a quite long period. Indeed, the initial part of understanding the current state of the art is a step that does not need to be done continuously, since as said at the beginning of this thesis, changes do not occur frequently. Similarly, the generation of the predictive models is something that is done once reused over time (anytime a company decides to plan preventive maintenance), as well as for the study of the relationship between variables. In case, the data analysis department can worry about doing periodical training and testing to validate models. Eventually, once given the structure of the economic assessment methods, companies only need to insert data and retrieve results, but the whole study behind the model is already performed and the cost covered. The only procedure that necessarily needs to be done any time new data are collected is the data cleaning and validation, accounting for approximately the 10% of the entire cost (840€).

7.2 Conclusions

Data represents an incredibly strong source to take wise and convenient decisions, especially when the ability to observe and study the phenomenon can be impeded by physical barriers, such as being buried under the soil. Trying to predict events such as breaks may be an optimal strategy to adopt preventive maintenance, and new ML techniques are a viable solution to undertake the problem.

For example, by exploring data, it was possible to establish as ductil cast (*FE*) material has a lower break rate per unit of length than fiber cement and polyethylene pipes. Therefore, management should concentrate resources in monitoring pipes of these last two materials, because of higher possibility of breaking. Moreover, as Figure 17 shows, before turning 40 years old, the probability of detecting a failure is higher for pipes in polyethylene than for fiber cement. The trend reverses after the breakeven age when fiber cement resistance

drops. One last important insight about which physical factor affects durability is the nominal diameter, with the failure rate continuously decreasing as the diameter rises (Figure 18).

Even though some data issues, i.e., truncation and censorship have not been addressed in this thesis, *logistic* and *RandomForest* predictive models have been able to return important insights in terms of operativity. Indeed, even though sometimes sensitivity and sensibility had not enthusiastic values from a data analysis standpoint, a holistic view and the meaning behind those numbers radically change the interpretation of results. E.g., combining *random forest* and *k-fold cross-validation,* by only placing 88 interventions, in a network of more than 13,000 items, in 6 years, around 12% of the total failures over the same period can be avoided. The logistic regression model is stronger in detecting properly failures, with sensitivity up to 0.76, enabling the prevention of more than 90% of the failure. However, this would be very consuming and far from a real operative scenario due to the enormous consumption of resources due to high *false-positive* prediction. The choice of the proper predictive model is a matter of compromise between the risk of falling into false alarms during the research of leaks and the necessity to provide the best service possible and reduce consumption. This object lies in the scope of the first economic assessment method developed in this thesis, where, based on the function of the cost of a company, the most suitable relationship between sensitivity and sensibility can be determined to optimize the total cost of maintenance.

The second methodology gets close to the real process of maintenance planning of a company such as *Aigües Manresa*, which first agrees upon the budget to invest in maintenance in a certain period and then put into action interventions. We have seen as relying on a predictive model, such as random forest, leads to higher efficiency than random selection of pipes or replacing by “seniority”, confirming one more time what a powerful tool ML can be also in maintaining a good service level also in water distribution. Water distribution companies, such as *Aigües Manresa*, can undertake preventive maintenance strategies, using tools and approaches explained in this study to give positive contributions to their performances. Without never forgetting the “green” impact ML can have on the environment, since preventing failure not only avoid capital waste but also waste of water, that as said in the introduction of this report, more and more becomes a scarce resource.

To conclude, the approach with which this thesis has been carried out can be followed for further studies. Reducing complex *time-to-next-event* problems to a scenario with simply binary *0-1* output, by remodulating the mother-dataset, is a banal but efficient trick able to easily drive to meaningful conclusions, especially when ML knowledge is limited. This simplification, without banalization, of the problem is an important outcome this job was able to reach and that may represent an accessible starting point for future interests. However, future inclusions of more advanced ML techniques such as *random survival forest,* allowing

to overcome left-truncation and censorship, would enable a better reconstruction of pipes history, to build more powerful predictive models. Also, data imputation techniques may give further support to stronger results, always with the aim of building a more meaningful dataset on which training and testing predictive models. Eventually, the first point of interest to deepen and improve this work stands on the ability to provide to ML algorithms a better dataset, for consistency, solidity and size, on which training and testing. The more predictions are accurate and with both high sensitivity and sensibility, the more a company can increase economic, but also social and environmental, performances.

8 References

2030 Water Research Group. (2009). *Charting our water future: Economic frameworks to inform decision-making.* Washington DC: 2030 WRG.

Andreou, S. A., Marks, D. H., & Clark, R. M. (1987). A new methodology for modelling break failure patterns in deteriorating water distribution systems: Theory. *Advance in Water Resources, 10*, 2-20.

Boulos, P. F. (2017, January). Optimal Scheduling of Pipe Replacement. *Journal American Water Works Association, 109*(1), 42-46.

Brownlee, J. (2021, December). *A Gentle Introduction to k-fol cross-validation.* Retrieved from Machine learning mastery: https://machinelearningmastery.com/k-fold-cross-validation/

Canadian Water Network. (2018). *Balance the books: Financial sustainability for Canadian water systems.* Waterloo: Canada: Canadian Water Network.

Chen, T. Y.-J., Beekman, J. A., & Guikema, S. D. (2017). Drinking Water Distribution Systems Asset Management: Statistical Modelling of Pipe Breaks. *Pipeline 2017.* Phoenix, Arizon.

Clark, R. M. (1982, October). Water distribution systems: a spatial and cost evaluation. *Water resources planning and management*, pp. 243 - 256.

Clark, R. M., Stafford, C. L., & Goodrich, J. A. (1982). Water distribution systems: A spatial and cost evaluation. *Journal of Water Resources Planning and Management Division, ASCE, 108 (3)*, 243-256.

Clark, R., J., C., ThurnauR., R., K., & S., P. (2010). Condition assessment modelling for distribution systems using shared frailty analysis. *American Water Works Association J. 102 (7)*, 81-91.

Cohen, A. C. (2020). *Truncated and censored samples - Theory and applications.* New York: CRC Press.

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

Cox, R. D. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, 34*(2), 187-220.

Erdelyi, L. (2021, October). *The Five Stages of Data Analysis .* Retrieved from LightHouseLab: https://www.lighthouselabs.ca/en/blog/the-five-stages-of-data-analysis

Farmani, R., Kakoudakisb, K., Behzadianc, K., & Butlerd, D. (2017). Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. *Procedia Engineering 186*, 117 - 126.

Folkman, S. (2018). Water main break rates in the USA and Canada: A comprehensive study. *Mechanical and Aerospace Engineering Faculty Publications Paper 174*.

Goulter, I. C., Davidson, J., & Jacobs, P. (1993). Predicting water-main breakage rate. *Journal of Water Resources Planning and Management, ASCE, 119(4)*, 419-436.

Hao Xu, S. M., & Sunil, K. S. (2021, August). Modeling Pipe Break Data Using Survival Analysis with Machine Learning Imputation Methods. *Journal of Performance of Constructed Facilities, 35*(5).

Jacobs, P., & Karney, B. (1994). GIS development with application to cast iron water main breakage rate. *2nd international conference on water pipeline systems.* Edinburgh, Scotland.

Judd, C., & McCleland, G. (1989). In C. a. Judd, *Data analysis: a model-comparison approach.* New York: Harcourt Brace and Jovanovich.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Kazei, A., & Goultier, I. C. (1988, February). Spatial and temporal groupings of water main pipe breakage in Winnipeg. *Canadian Journal of Civil Engineering, 15*(1), 91-97.

Kettler, A. J., & Goulter, I. C. (1985). An analysis of pipe breakage in urban water distribution networks. *Canadian Journal of Civil Engineering, 12*, 286-293.

Kettlere, A. K., & Goulter, I. C. (1985). An analysis of pipe breakage in urban water distribution networks. *Canadian Journal of Civil Engineering, 12*(2), 286-293.

Kibum, K., Jeewon, S., Jinseok, H., Taehyeon, K., Jaehag, K., & Jayong, K. (2019). Economic-based approach for predicting optimal water pipe renewal period based on risk and failure rate. *Environmental Engineering Research, 24*(1), 63-73.

Kleiner, Y., & Rajani, B. (2001, May 11). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water (3)*, 131 - 150.

Kleiner, Y., & Rajani, B. (2012). Comparison of four models to rank failure likelihood of individual pipes. *J. Hydroinformatics 14 (3)*, 659-681.

Kleiner, Y., Nafi, A., & Rajani, B. (2010). Planning renewal of water mains while considering deterioration, economies of scale and adjacent infrastructure. *Water Science & Technology: Water Supply - WSTWS, 910.6*, 897-906.

Konstantinou, C., & Stoianov, I. (2020). A comparative study of statistical and machine learning methods to infer causes of pipe breaks in water supply networks. *Urban Water Journal*, 534-548.

Lee, E., & Wang, J. (2003). *Statistical methods for survival data analysis.* Wiley-Interscience.

Li, F., Ma, L., Sun, Y., & Mathew, J. (2011). Group Maintenance Scheduling: A Case Study for a Pipeline Network. *Engineering Asset Management*, 163-177.

Marlow, D., Davis, P., Beale, D., Burn, S., & Urquhart, A. (2010). *Remaining Asset Life: A State of the Art Review.*

Mazumder, R. K., Salman, A. M., Li, Y., & Yu, X. (2021, May). Asset Management Decision Support Model for Water Distribution Systems: Impact of Water Pipe Failure on Road and Water Networks. *Journal of Water Resources Planning and Management, 147*(5).

Mazumder, R., Salman, A. M., Li, Y., & Yu, X. (2021). Asset Management Decision Support Model for Water Distribution Systems: Impact of Water Pipe Failure on Road and Water Networks. *Journal of Water Resources Planning and Management, 147 (5)*.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14*(4), 945-966.

O'Day, D. K. (1985). Water utility main break patterns. *Proceedings - Distribution System Symposium*, 195-295.

Qi, Z., Zheng, F., Guo, D., Maier, H. R., Zhang, T., Yu, T., & Shao, Y. (2018, July). Better Understanding of the Capacity of Pressure Sensor Systems to Detect Pipe Burst within Water Distribution Networks. *Journal of Water Resources Planning and Management, 144*(7).

Ray, S. (2017, September 9). *Commonly used Machine Learning Algorithms (with Python and R Codes).* Retrieved 12 2021, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/ (accessed November 2021)

Salman, B., & Salem, O. (2011). Risk assessment of wastewater collection lines using failure models and criticality ratings. *Journal of Pipeline Systems Engineering and Practice, 3 (3)*, 68-76.

Schutt, R., & O'Neil, C. (2013). In R. Schutt, & C. O'Neil, *Doing Data Science.* O'Reilly.

Shamir, U., & Howard, C. D. (1979). An analytic approach to scheduling pipe replacement. *Journal of AWWA, 71(5)*, 248-258.

Shin, H., Joo, C., & Koo, J. (2016). Optimal Rehabilitation Model for Water Pipeline Systems with Genetic Algorithm. *Procedia Engineering, 154*, 384-390.

Snider, B., & McBean, E. A. (2021, September). Combining Machine Learning and Survival Statistics to Predict Remaining Service Life of Watermains. *Journal of Infrastructure Systems, Vol. 27*.

Wang, Y., Zayed, T., & Moselhi, O. (2009). Prediction Models for Annual Break Rates of Water Mains. *Journal of Performance of Constructed Facilities, 23*(1), 47-54.

WWAP, (. N. (2018). *The United Nations World Water Development Report 2018: Nature-Based Solutions for Water.* UNESCO, Paris.

Yiu, T. (2021, December). *Understanding Random Forest .* Retrieved from Towards data science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed December 2021)

9 Electronic support

Baptiste Auguie (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Hadley Wickham and Jennifer Bryan (2019). *readxl: Read Excel Files*. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Ishwaran H. and Kogalur U.B. (2021). *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC).* R package version 2.11.0.

Jared E. Knowles (2020). *eeptools: Convenience Functions for Education Data*. R package version 1.2.4. https://CRAN.R-project.org/package=eeptools

Jarek Tuszynski (2021). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc*. R package version 1.18.1. https://CRAN.R-project.org/package=caTools

Kuhn Max (2020). *caret: Classification and Regression Training*. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Liaw A. and Wiener M. (2002). *Classification and Regression by randomForest*. R News 2(3), 18--22.

Matt Dowle and Arun Srinivasan (2020). *data.table: Extension of `data.frame`*. R package version 1.13.6. https://CRAN.R-project.org/package=data.table

Meyer David, Dimitriadou Evgenia, Hornik Kurt, Weingessel Andreas and Leisch Friedrich (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-4. https://CRAN.R-project.org/package=e1071

Milborrow Stephen (2020). *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. R package version 3.0.9. https://CRAN.R-project.org/package=rpart.plot

Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2021). *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.6.1. https://CRAN.R-project.org/package=naniar

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio 1.3.1093, 2009-2020 RStudio, PBC.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). *ROCR: visualizing classifier performance in R*. _Bioinformatics_, *21*(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

Stefan McKinnon Edwards (2020). *lemon: Freshing Up your 'ggplot2' Plots*. R package version 0.4.5. https://CRAN.R-project.org/package=lemon

Terry Therneau and Beth Atkinson (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Yihui Xie (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30.

Wickham et al., (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham H.. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

Yihui Xie (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30.

10 Website

[1] https://www.eea.europa.eu/signals/signals-2018-content-list/articles/water-use-in-europe-2014

[2] https://www.interreg-central.eu/Content.Node/Digital-Learning-Resources/Water-loss.html

[3] https://www.worldwildlife.org/threats/water-scarcity

[4] https://www.euroweeklynews.com/2019/01/11/water-waste-spain-loses-the-third-highest-amount-of-water-in-europe/

[5] https://www.iagua.es/noticias/locken/vuelven-aumentar-perdidas-agua-espana

[6] https://www.scopus.com/home.uri

[7] https://www.sciencedirect.com

[8] https://scholar.google.com

[9] https://es.meteosolana.net/estacion/0149X

[10] https://www.foro-ciudad.com/barcelona/manresa/habitantes.html#EvolucionGrafico

[11] https://data-flair.training/blogs/using-r-for-data-science/

[12] https://www.tableau.com/learn/articles/what-is-data-cleaning

[13] https://www.mesotheliomahelp.org/asbestos/history/

[14] https://www.datarobot.com/wiki/prediction/