Predicting water pipe failures
a neural Hawkes process approach

Verheugd, J.T.

Award date:
2020

Department of Industrial Engineering and Innovation Sciences - Information Systems group

Predicting water pipe failures: A Neural Hawkes Process approach

Performed at Vitens, Zwolle, The Netherlands

Jeroen Verheugd (1004842)

In partial fulfillment of the requirements for the degree of
Master of science in Operations Management and Logistics

Supervisors:

Dr. Y. (Yingqian) Zhang, Technical University of Eindhoven
Dr. L. (Laura) Genga, Technical University of Eindhoven
S. (Sjoerd) Boersma, Vitens
R. (Ron) Bergers, Deloitte Netherlands

1st version

Eindhoven, January 2020

Abstract

The water distribution network of Vitens is slowly becoming of age and starts to show a yearly increased rate of failure. In this work, the failure intensity function of water pipes is learned by applying the deep learning technique recurrent neural networks to point process modeling. Compared to current solutions proposed when modeling water pipe failures, our model is able to predict the time to next failure on an individual water pipe level. The learned failure intensity function is used to identify value points in the deterioration process of water pipes that represent their economical end-of-life.

Our method is believed to be expressive enough to learn the underlying dynamics of water pipe failures without a predefined parametric function that is often seen in related academic papers. The failure intensity function is learned based on two components: (1) the base failure rate that is determined by the unique profile attributes; (2) and the effect of past failures on this base failure rate. Our Neural Hawkes Process model is able to predict the timestamp of the first failure of a water pipe unit with a mean absolute error of 8.8 years. When a prediction is made for a consecutive failure on a water pipe, the mean absolute error is 1.6 years.

According to the economical end-of-life definition of Vitens, our model has identified for a total of 326 groups of water pipes it is more profitable to replace the water pipe than to perform a repairment.

Keywords - water pipes, predictive maintenance, point process model, Hawkes process, intensity function, recurrent neural networks

Executive summary

Research motivation

The asset management department of Vitens has the responsibility to monitor the quality of the water distribution network. To prevent failures within this system and guarantee top quality drinking water, there must be continuously invested in new technologies and infrastructure. The water distribution network of Vitens consists of 49.600 km of pipelines and is slowly becoming of age and starts to show a yearly increased rate of failure of 1.7% per year. Unexpected water pipe failures have a negative impact on the customer service level that is measured in the amount of time a client of Vitens has no access to quality drinking water. It is the responsibility of the asset management department of Vitens to prevent water pipe failures and act on 'early warnings' or so-called value points in the deterioration process of individual water pipes. These value points represent the economical end-of-life. It is not beneficial to keep water pipes operational beyond the economical end-of-life due to the fact of increased maintenance costs in comparison to the replacement costs.

The current predictive maintenance capabilities of Vitens consist of a condition assessment tool that is able to score each individual water pipe on a scale of 'healthiness'. This model, called the Pipe Replacement Potential, combines a probability of failure and its consequence of failure to be able to prioritize the replacement of water pipes in their distribution network. The probability of failure is determined by calculating Kaplan-Meier curves for homogeneous groups of water pipes. The new ambitions of Vitens are defined to enhance their current predictive maintenance capabilities. The motivation for this thesis project is to investigate the potential of machine learning techniques to be able to predict failures on an individual water pipe level. Subsequently, this predictive model can be used to identify which water pipes have reached their economical end-of-life. This research motivation has led to the formulation of the following research question:

How can a machine learning approach determine the economical end-of-life of water pipes in a Dutch water distribution system?

Methodology

The literature study performed has identified the current state-of-the-art on modelling an intensity function. An event sequence, like failure events of water pipes, carries important clues about the underlying failure dynamics of water pipes. One mathematical approach to model such behaviors is called point process and is also known as the Hawkes process. A neural embodiment of this classical Hawkes model is used to model the failure intensity function with recurrent neural networks and long short-term memory cells called the Neural Hawkes Process model. The failure intensity function involves two components: the base failure rate determined by the water pipe profile attributes and the effect of past failures on this base failure rate.

Two major data preparation steps have been performed on the distribution network dataset and the historical failure records. First, individual water pipes are grouped into asset units based on their shared material type, diameter, age and their connectivity. The creation of asset units is fundamentally different than homogeneous groups due to the fact of a required physical connection. Second, a geospatial algorithm has been developed to allocate failure events to these created asset

units. In total 16182 failure events have occurred on 10203 unique asset units, which are used for training the Neural Hawkes Process model. Failure events are transformed into failure event sequences per asset unit. These failure event sequences consist of tuples with the type of failure and time of occurrence denoted in years.

The Neural Hawkes Process model is expressive enough to be able to learn the underlying dynamics of the deterioration process of water pipes to predict the timestamp of the next failure. This implies that no prior domain knowledge is needed to develop the Neural Hawkes Process model, which is seen as a big advantage for this model approach. The failure event types are included in the event sequence tuples to allow the model to learn exciting or inhibiting effects of past failures on future failures.

Results

The Neural Hawkes Process model has been tested and evaluated in two situations. In the first situation, the Neural Hawkes Process model is trained on all instances of water pipes. The accuracy of the Neural Hawkes Process model of predicting the timestamp of the first failure event on asset units has a mean absolute error of 8.8 years. The base intensity rate captured in the profile vector and the influence of the failure event sequence has resulted in the best performance, compared to the individual performance of the profile vector and failure event sequence. The effect of the failure event sequence component is limited due to the fact that 73.4% of the asset units have only experienced one failure during the failure observation period. In the second situation, the Neural Hawkes Process model is trained on all instances of water pipes that have at least experienced two failures. The effect of the failure event sequence component is more leveraged in this situation compared to that of situation 1. The ability of the Neural Hawkes Process model in this situation is to predict the time to next failure when an asset unit has already experienced one failure. In this case, the Neural Hawkes Process model has an accuracy of a mean absolute error of 1.6 years. Regarding exciting and inhibiting effects of failure event types, incorporating the failure event type in the failure event sequences, have increased the accuracy of the Neural Hawkes Process model in both situations 1 and 2, respectively with an improvement of 0.2 years and 0.14 years.

The Neural Hawkes Process model has led to the identification of 326 asset units that have reached their economical end-of-life. These model insights are presented to the asset management department of Vitens and can be used in their maintenance strategy.

Conclusion and recommendations

The Neural Hawkes Process model is able to predict the economical end-of-life and identity value points in the deterioration process of the asset units. The Neural Hawkes Process model is expressive enough to learn the intensity function without any prior domain knowledge. When compared to the current predictive capabilities of Vitens, the predictive model presented in this report is able to make accurate predictions on an asset unit level instead of a homogeneous group; and is able to predict the time to next failure for these individual asset units. This exceeds and contributes to the current predictive maintenance capabilities at Vitens of scoring individual water pipes on a relative scale of structural condition.

This work has shown that the event-sequential model approach in combination with the profile attributes exceeds the current modelling techniques in the field of predicting water pipe failures. Further research should focus on incorporating time-varying profile attributes, i.e. climate related variables, in the Neural Hawkes Process model by replacing the static profile vector with a time-series RNN to allow for a dynamic base failure rate of asset units.

To further advance the predictive maintenance capabilities of Vitens, the following is recommended: (1) the process of water pipe failure registration must be aligned with the predictive maintenance ambitions, (2) the data quality of the distribution network should be improved to enhance future predictive maintenance projects, (3) the developed data transformation steps should be made readily available on a centrally accessible location to form the basis for future data science projects.

Preface

This report is the final result of my graduation research, as a final project within the master program Operations Management and Logistics at the Eindhoven University of Technology. This research is conducted at the Dutch water distribution company Vitens, Zwolle, and co-supervised by the Analytics & Cognitive department of Deloitte, Amsterdam.

This report concludes on my academic career of the past seven years as a student and marks the beginning of my professional career as an industrial engineer. Before the findings of this report are presented, I would like to take the opportunity to express my appreciation to the people that have contributed to this achievement.

First of all, I would like to thank Yingqian Zhang for supervising me during this master's project. Thank you for trusting me in arranging and formulating my master thesis topic. You have provided me with the right questions at the right time that gave me direction to finish this project successfully. Secondly, I would like to thank my second university supervisor Laura Genga for her feedback and suggestions that have contributed to this final version.

I would also like to thank my company supervisor at Vitens, Sjoerd Boersma, for the amount of time he has invested in my thesis project, which eventually became our shared predictive maintenance project. Your expertise in both the domain knowledge in the water domain as in the field of data science has taught me a lot and has contributed to the quality of this report. I would also like to thank my Deloitte supervisors, Ron Bergers and Edwin Wanner, for their industry knowledge and guidance during my intern period of the last 8 months.

Furthermore, I would like to thank the support of my parents, sister and brother during the last 7 years. Although most of the time you barely understood the things that I have shared with you, the support and admiration made me feel confident to pursue this master.

At last, I want to thank all my friends that have supported me during my master period. Academic or personal related, you were always there when I needed to talk or needed advice. A special thanks to Bram Nick for all the great memories we have made during our master period.

Jeroen Verheugd

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter is an introduction to the problem and motivation of developing predictive maintenance capabilities in the water distribution network of Vitens. There will be elaborated on the current predictive capabilities of their Asset Management department and the challenges that Vitens face. This has led to a new set of ambitions that will contribute to their predictive maintenance capabilities. From these new ambitions, the main research questions and sub-questions have been formulated. Thereafter, the project scope is defined and the outline of this report is discussed.

1.1 Research motivation

Vitens is a Dutch water distribution company that provides water to 5.7 million households in the provinces Utrecht, Gelderland, Flevoland, Overijssel, and Friesland. The distribution network of Vitens consists of 49.600 km of pipe lines and has distributed 354 million m^2 of produced water in 2017. Their core business is defined as producing, distributing and supplying top quality drinking water to their customers and the associated service delivery Vitens (2017).

The asset management department of Vitens has the responsibility to monitor the quality of the water distribution network. To prevent failures within this system and guarantee top quality drinking water, there must be continuously invested in new technologies and infrastructure. The size of the distribution network has slowly expanded since the 19th century. This has led to the use of four main types of material that are present within the system in a variety of different sizes. These material types are cast iron, asbestos cement, polyvinylchloride (PVC), and polyetheen (PE). In general, pipe lines of the type cast iron and asbestos cement were mainly installed during the last century. Nowadays these materials are no longer used for the installation of new pipe lines and are completely replaced by PVC or PE due to better performance and lower installation and maintenance costs. Each material type used for pipe lines has different material characteristics that lead to a different estimated life expectancy. This in combination with different pipe sizes and environmental effects near the asset location, makes the distribution system prone to water pipe breaks and leaks (St. Clair & Sinha, 2014).

A key performance indicator (KPI) of Vitens is measured by the amount of time a person has no access to drinking water from their distribution network. The occurrence of water pipe breaks and leaks temporally cut-off the water supply to customers or lead to undrinkable water. The duration of these failures directly negatively affect the performance KPI of Vitens and the customer service level. Therefore, a strategy to control and prevent these events must be maintained.

As previously mentioned, the failure of a water pipe cannot be solely explained with the aging of a water pipe. In addition, several breaks or leaks per year on the same asset unit does not necessarily mean that this specific asset has reached its technical end-of-life. This can be best explained by 'The Bathtub Curve' principle shown in Figure 1.1. As illustrated in this figure, the behaviour of failures can be best described with three different phases, namely infant

mortality, normal life and end-of-life wear-out. Premature failure can occur due to incorrect storage and/or installation of water pipes. This probability of failure rapidly decreases after a successful installation. The premature failure phase is followed by the normal life of the asset with a low and constant failure rate. By the time the asset reaches the end of his technical lifetime, the rate of failure will increase.

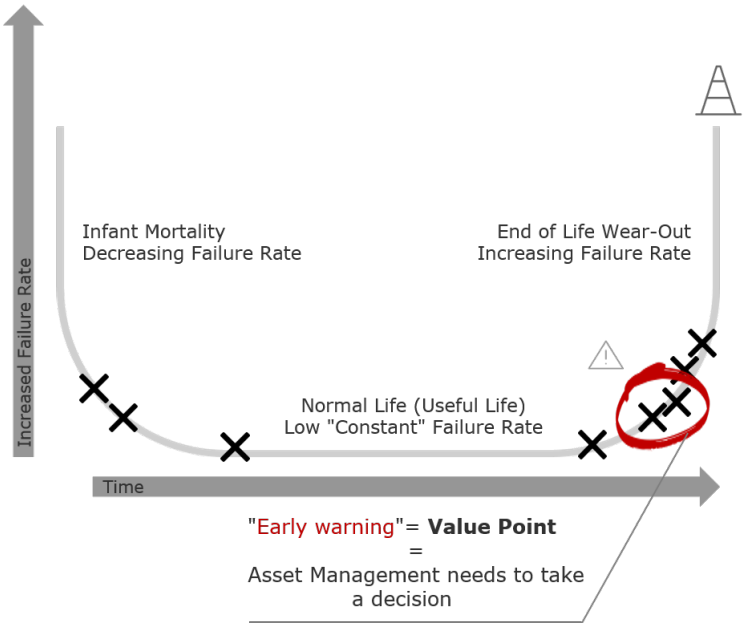


Figure 1.1: Bathtub curve principle

Since the impact on customer service is important to Vitens, it is the responsibility of Asset Management to detect water pipe failures and act on 'early warnings' or so called value points, indicated with a red circle in Figure 1.1. The Asset Management department has stated that these value points are an indicator that a water pipe line has reached his economical end-of-life, meaning that it is economically more profitable to perform a replacement than a repairment. To determine the economical end-of-life, not only the probability of failure is important, but also the cost of replacement and the execution of the maintenance activities are relevant. The financial factor and the consideration of maintenance costs is simplified in this report, Asset Management has defined the economical end-of-life as follows:

"When a consecutive water pipe failure is predicted within two years since the last occurred failure, within a range of 400m of the last failure, the water pipe has reached its economical end-of-life. "

The ability to identify these values points will prevent Asset Management in making unnecessary premature replacements of water pipes that have not yet reached their economical end-of-life. More importantly, predicting the economical end-of-life of water pipes will reduce the impact of unforeseen water pipe failures for their customers. This will contribute to the customer service level and KPI of Vitens.

1.2 Current predictive capabilities at Vitens

The Asset Management department of Vitens currently has a predictive model in operation that is able to asses the condition of their water distribution network. This condition assessment tool is called the Pipe Replacement Potential (PRP) model (Figure 1.2)'and is a decision support model

that consists of two factors, probability of failure and the consequence of failure. The PRP model uses three distinct data sources to predict the condition of a water pipe. A schematic overview is given in Figure 1.2.

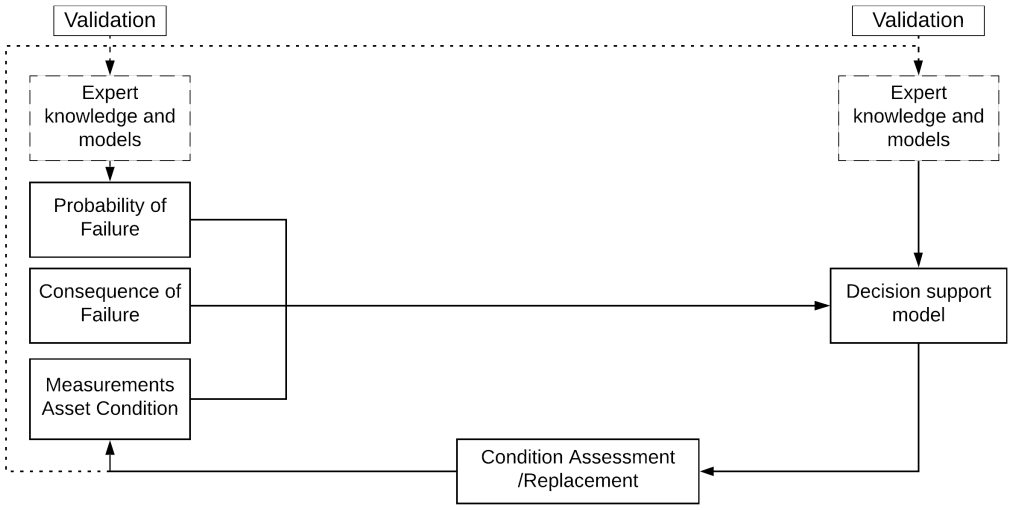


Figure 1.2: Pipe Replacement Potential

First, the **Probability of Failure (PoF)** is determined by making Kaplan-Meier curves for homogeneous groups of water pipes. Kaplan-Meier curves requires a lot of water pipe failures in order to be created, which is insufficiently available when making these deterioration curves on an individual water pipe level. The water pipes in the distribution network, from here on also referred as assets, are grouped into homogeneous groups based on their characteristics age, diameter and material type. It is important to note that the individual water pipes in these homogeneous groups can be scattered throughout the whole distribution system of Vitens. The Kaplan-Meier curves, which is a common type of survival analysis, are made for each homogeneous group based on the respective failures occurred on water pipes in that group. The Kaplan-Meier curves are used to determine the probability of failure past a certain point of age. This probability of failure is used for all individual water pipes within a homogeneous group. However, the probability of failure of the homogeneous group is the averaged probability of failure of all individual water pipes. The creation of homogeneous groups resolves the issue of the limited amount of failures for making Kaplan-Meier curves, but does not allow to accurately determine the probability of failure on an individual water pipe level.

Secondly, some assets have a higher customer service contribution to the overall distribution system than others, e.g. a water pipe failure in a rural area affects more households than one in a remote farm field. A **Consequence of Failure (CoF)** is introduced to bring a nuance in the PoF analysis. To do so, different areas in the distribution network are scored on several factors such as water usage in m^3, priority of clients, and a category of urbanisation.

These two factors are combined into a prediction of the water pipe's condition - that combined forms the decision support model. The outcome of this model is a condition score of all current water pipes within their water distribution network. This score is given on a relevance scale from 0 to 100 percent, whereby the best performing water pipes are given a score of 0 and the worst performing water pipes a score of 100. The asset management department of Vitens uses this outcome as a data-driven decision support model. However, the asset management department does not solely make decisions based on this model. There is a lack of validity whether the predictions are in line with the actual state of the water pipe. Therefore, they combine expert knowledge and the condition scores of the PRP-model to make a budget and planning for the

maintenance activities in their distribution network. The expert knowledge is learned from field inspections measurements that are evaluated with physical deterioration models for each material type. Based on the expert knowledge, critical parts of the distribution network are identified and these water pipes will be planned for replacement. The insights in whether the results of the inspections are in line with the predicted results of the PRP-model are limited due to a lack of reliable feedback from field inspection measurements. However, to some level, these inspections are used to validate the PRP-model in combination with the expert knowledge present at Asset Management. As has been previously mentioned, the results of the PRP model are not decisive and are currently only consulted to assist in the expert's decision making process. For example, premature replacements of water pipes is performed whenever an opportunity arises if a third party contractor, e.g. an electricity company, offers to share the costs to break ground for performing maintenance on the below infrastructure.

1.3 Contribution to the scientific literature

This master thesis project will investigate the possibilities of machine learning techniques to model water pipe failures. The contributions of this research to the scientific literature is listed as follows:

- Successfully modelled the failure intensity function of water pipes that consists of a base intensity rate and the long-term effects of historical failure events, respectively determined by the unique profile attributes and failure event sequences. This combination leads to a better predictive performance of the time to next failure than a failure intensity function learned from only the profile attributes or historical failure events.
- An innovative approach to incorporate the effect of water hammer to the profile attributes of water pipes by including the distance to the closest water production location. It has been proven that water pipes closer to a production location experience the largest effects of water hammer.
- In the water application domain, this work is the first attempt to determine the economical end-of-life in the life cycle of individual water pipes.

1.4 Research questions

The objective of this master thesis research is to explore how and which machine learning techniques can be applied to model water pipe failures on an individual water pipe level. This predictive model can be used to predict the next failure event and determine when an asset has reached its economical end-of-life. The accuracy of this predictive model will directly contribute to a more efficient planning of replacements of assets while reducing the probability of unexpected failures.

The combination of these ambitions has led to the following main research question:

How can a machine learning approach determine the economical end-of-life of water pipes in a Dutch water distribution system?

Answering this research question will be structured by answering the following sub-questions:

1. What are the existing methods for modelling water pipe failures and what are their advantages or shortfalls?
2. What technical, environmental, and operational variables are fundamental predictors for modelling water pipe failures?
3. Based on the literature, which machine learning method is promising to model the underlying dynamics of water pipe failures?

4. How can these fundamental variables be combined into a dataset that is prepared for time sequential analysis?
5. What is the performance of the machine learning method when being tested on the historical water pipe failure records of Vitens?

1.5 Scope

Vitens goal for this thesis project is to develop a predictive model that is able to predict the economical end-of-life of water pipes. When compared to the current predictive maintenance capabilities, the novel desired model must be able to make accurate predictions on an individual water pipe level. Due to a limited available time for this master thesis project, a project scope is defined:

- The Consequence of Failure factor included in the Pipe Replacement Potential model is not incorporated in the development of the new predictive model
- The actual maintenance costs of water pipe failures is left out of scope when the economical end-of-life of water pipes is determined
- Validating the results of the new predictive model with actual field inspections is left out of scope.
- The implementation of the new predictive model in the day-to-day business of Vitens is not included in this thesis project.

1.6 Outline

The remainder of this report is structured as follows. In Chapter 2, there will be elaborated on the state-of-the-art of modelling water pipe failures and the suggested solution approaches. In addition, the literature will be consulted to determine important predictors on the life time expectancy of water pipe assets, e.g. the environmental effects on different material types.

Thereafter, in Chapter 3 there will be elaborated on the distribution network dataset and the historical failure records. The data insights that lie within these two datasets will be extracted and discussed, respectively.

This is followed by Chapter 4, wherein these two datasets are prepared for time sequential analysis. Two major data transformation steps are the creation of asset units and a geospatial analysis to connect water pipe failures to these asset units.

In Chapter 5, the method for modelling water pipes failures is introduced. A recurrent neural network approach is used to model this behaviours with a classical survival analysis technique, called Point Process models or the Hawkes process. Subsequently, the results of this predictive model are discussed in Chapter 6. The model insights gained are discussed in a implementation section in Chapter 7.

In Chapter 8, a conclusion and discussion on the results of modelling water pipe failures in the water distribution network of Vitens is given. In addition, the limitations of this research will be addressed and a section is devoted to provide recommendations to Vitens related to their predictive maintenance capabilities.

Chapter 2

Literature review

The majority of the current distribution system of Vitens has been installed shortly after World War II, wherein this period only minimal maintenance was performed on water pipe lines. Now, more and more of the installed water pipe lines are approaching the end of their expected life time, and has starting failing at an increasing rate. This literature review has been performed in order to answer the first three sub questions of this master thesis project.

This literature review is structured as follows. In the first section there will be elaborated on the existing methods of model development and their advantages or shortfalls. Subsequently, a profound exploration into the understanding of the deterioration process of the water pipes and its predictors has been performed. Furthermore, the state-of-the-art of modelling water pipe failures is investigated. Finally, a discussion on the position of this research in the literature will be given.

2.1 Method of model development

The physical mechanisms that contribute to the failure of water pipes are complex and not well understood, and normally there is little data available on the breakage modes due to the difficulty of inspection and a lack of historical data due to a low rate of failure (Rajani & Kleiner, 2001). This makes it difficult to develop a universal, reliable predictive water pipe failure model for usage in all regions. This has led to different methods of model development, which can be divided into physical, statistical, probabilistic and advanced mathematical based approaches.

2.1.1 Physical models

In the work of Rajani & Kleiner (2001) a comprehensive overview of physical models, also referred as deterministic models, is provided that has attempted to model the structural performance of water mains. The residual structural capacity of water mains is affected by material deterioration due to environmental and operational conditions. In these models extensive efforts have been made to model the physical processes of the deterioration and failure of buried pipes through analysis of the loads the pipe is subjected to as well as the capacity of the pipe to resist the loads (Wilson et al., 2017). There are two different approaches in which a physical model may be developed, namely empirical and mechanistic. The empirical approach relates failure rates to the attributes of the asset. This approach is only applied to cohorts of pipes with certain similar characteristics. The mechanistic approach, on the other hand, predicts the service lifetimes of individual assets. Most of the existing prediction models have been developed through regression analysis, combined mechanistic-empirical analysis and expert knowledge (St. Clair & Sinha, 2012).

The problem with physical models is that the applicability of each individual model is restricted to a specific location due to the variation of environmental effects in other regions of the water distribution system. Moreover, there are difficulties in implementing physical models for a large distribution system, as many of the models rely on data from inspections. The inspection of large-

diameter water mains can be technically difficult and expensive, problems which are exacerbated by the lack of mature inspection techniques. Individual inspection techniques are also limited in their scope, in terms of the applicability to different pipe materials, the information gained and their overall performance (Wilson et al., 2017).

2.1.2 Statistical models

Many of the models that predict water main failure are developed through the analysis of historical failure data, as an understanding of the physical mechanisms that lead to the failure is not required (Wilson et al., 2017). In contrast to physical models, statistical models usually require fewer resources and can capture hidden statistical failure patterns caused by different physical reason (Li & Wang, 2018). Statistical modeling is commonly used to predict the lifetime or failure time of infrastructure. There are many approaches to develop a statistical model for predicting water pipe failures. One of the characteristics they all have in common is that they are based on long-term observed field data and processed through regression analysis (St. Clair & Sinha, 2012).

Since statistical models heavily rely on long-term observation, their applicability is limited when considering newer assets or assets with insufficient historical data (St. Clair & Sinha, 2012). These approaches are applied to homogeneous groups of pipe infrastructure and require a large number of long-term observed field data. This amount of data is in many cases limited available. This is in line with Asnaashari et al. (2013), wherein they state that most of the times records of the pipes installed, including data on pipe length, diameter, pipe material and date of installation is present. However, records of pipe breakage longer than a decade are rarely available. There exist several methods, such as the use of probabilistic analysis that can reduce the data requirement, but at the price of a more complex mathematical framework. As well, since these models are often based on historical failure data from all times in the service life of the assets, some are capable of modelling the entire life cycle (Wilson et al., 2017).

A disadvantage of statistical models is that the models developed are only as accurate as the quality of the input data used to develop the model. Extrapolation of data points is not always possible if a non-parametric statistical model is used. This could be problematic for modelling failures of water pipes older than 100 years due to the fact that data beyond that point of age is limited (St. Clair & Sinha, 2012). A second disadvantage is that the models are specific to the region from which the data is obtained to develop the models, and are thus not as globally applicable as physically-based models. The third disadvantage is that often only a limited number of variables are analysed in developing the models, due to lack of data, which can lead to inaccuracy (Wilson et al., 2017). In the work of St. Clair & Sinha (2012) they state that in many cases the regression-based approach is not suitable for modeling the actual deterioration process of pipe infrastructure since the sampling data used in the regression analysis suffers from various limitations, such as pipe structure, loading and environmental variables.

Probabilistic modelling analyses the probability or relative frequency of an event occurring. The likelihood of these occurrences is practical to describe the failure of an asset. More specifically, the probability of an event to occur is denoted by a 1; while an event that cannot occur is denoted by a 0. These types of models use condition related data of water pipes and their physical characteristics in modelling the probability of failure. Moreover, these models are based on extensive data sets and are commonly utilised in pavement, bridge and other infrastructure network management concerning repair, rehabilitation and replacement priority (St. Clair & Sinha, 2012).

2.1.3 Advanced models

Based on the literature, among the more advanced models that are used to model the deterioration or failure of pipe infrastructure are machine learning methods, such as Artificial Neural Networks (ANN). ANN provides a useful tool for modeling asset deterioration or failure due to its non-linearity, adaptivity and learning capabilities (Haykin, 1994). Due to the increased levels of skill and training required to develop these complex networks and its ”black-box” approach, it is currently limited used in the water utility industry. (St. Clair & Sinha, 2012).

2.2 Understanding of water pipe failures

All water distribution systems consist of different material types used for their assets. The performance of these material types over the long-term are affected by many parameters. An extensive historical background for each material group can be found in Appendix A. In this section the different parameters that effect the deterioration process of water pipes will be discussed. Subsequently, the literature is examined to determine which parameters are used in case studies wherein is attempted to model water pipe failures.

2.2.1 Predictors of water pipe failures

In order to develop a predictive model that models water pipe failures it is necessary to have a complete understanding of the failure modes and mechanism of buried pipe infrastructure systems. Many parameters affect water pipe infrastructure systems. Logically, physical factors such as pipe diameter, age, and the used material type have an effect on the deterioration process of a water pipe. This, combined with operational factors and environmental effects, will lead to failures such as cracks, fractures, and holes, which in their turn lead to the leakage of water. Also, the effects of one parameter may amplify the effects of other parameters.

Some early research in this field has proven that pipe breaks are not directly related to the age of pipes and other environmental factors such as the water temperature and soil temperature also influence pipe breaks (O'Day, 1982)(Kawakita, 1986). Moreover, the shallower the depth of pipe installation, the larger the fluctuation of the soil temperature due to the susceptibility to the atmospheric temperature (Ahn et al., 2005).

In the work of Puust et al. (2010) they state that leakage in distribution systems can be caused by a number of different factors. Some examples include, bad pip connections, internal or external pipe corrosion or mechanical damage caused by excessive pipe load (e.g. by traffic on the road above or by a third party working in the system). In addition, they state that other common factors also influence leakages such as ground movement, high system pressure, pipe age, winter temperature, and other ground conditions.

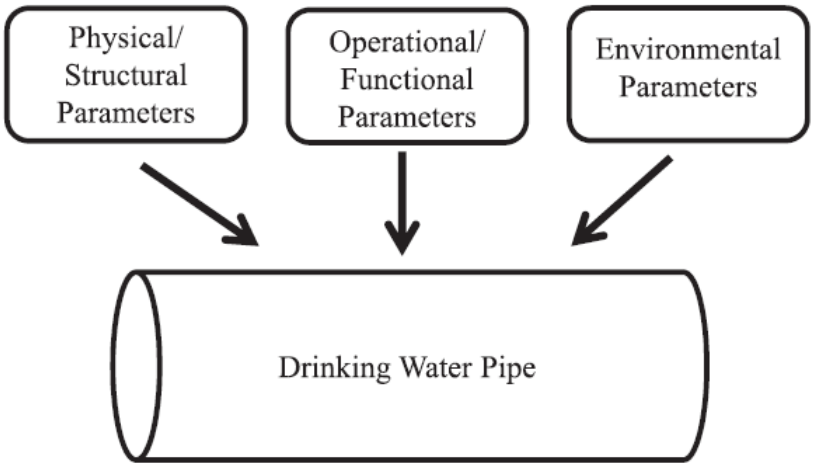


Figure 2.1: Standard data structure for modeling water pipe failures (adapted from (St. Clair & Sinha, 2014))

In the paper Tabesh & Delavar (2003) is described that combining hydraulic modelling software and Geographical Information System (GIS)[1] data are promising sources of information to predict

[1] A method to store and control data variables based on a GPS location.

water pipe failures. Of which the latter is generally freely available from municipalities or national governments.

St. Clair & Sinha (2014) provided a standard data structure for predicting the remaining physical life of water pipes. They classify all water pipe parameters into five categories of which two categories are out of scope for the Vitens case study, namely financial parameters and third party damages. As shown in Figure 2.1, this has resulted in a limited version of their standard data structure.

Subsequently, St. Clair & Sinha (2014) identify the parameters per classification category needed to include according to their proposed standard data structure for modelling water pipe failures. This overview is presented in Table 2.1.

Parameter Classification	**Parameters**
Physical/structural	Node identification number; node length; material type; nominal diameter; age; depth; location; and design life; join type; wall thickness; installation
Operational/functional	Operational pressure; pipe renewal record; pipe failure record; water quality violations; operation and maintenance practices; consumption and service type pressure limitations; water temperature; water flow velocity; pressure surges; pipe inspection record; leakage amount; pressure complaints; discolored water; and leakage allowance
Environmental	Land cover; soil type; water source; dead load; and live load; climate/ temperature; soil corrosivity; groundwater table; frost penetration

Table 2.1: Parameters of standard data structure for modelling water pipe failures (St. Clair & Sinha, 2014)

2.2.2 Analysis of case studies

As previously mentioned in Chapter 2.2, a water distribution system consist of several material types. The performance of these material types over the long-term are affected by many parameters. As a result of these external factors in relation to a particular material type - cracks, fractures, and holes may occur which in turn lead to a pipe failure. A total of 21 case studies presented in the literature will be examined to investigate which parameters are commonly used to model water pipe failures. All of the case studies are performed at small to medium sized cities with a total water distribution network length between 300 and 5500 kilometers. An overview is given in Table 2.2.

Several studies in North America and Europe over the last decade have documented the influence of pipe structure and soil parameters on water pipe failures. Based on the overview presented in Table 2.2, the significance of the diameter, length, age, material, and failure history along with soil type for predicting failure rate has been widely established. Only in the papers Karimian (2015) and Shirzad et al. (2014) has the buried depth of water pipes been included as a physical parameter. However both papers state that there is no significant effect of the installation depth of water pipes on its failure rate. The effect of tree root intrusion on nearby water pipe lines has been captured in the model of Tran et al. (2006) and is classified as a land cover parameter. This variable is calculated by counting the number of surrounding trees, but based on their findings they are unable to conclude if this effects the structural deterioration of water pipe lines. Information on pipe renewal records like in the paper of Winkler et al. (2018) is barely used, because most water distribution companies do not record failures of water pipe lines that have been replaced and this, unfortunately, implies that a lot of valuable historical data is lost.

The water pressure within a water distribution system is an important indicator of water pipe failures. However, this operational parameter is difficult to measure. Moreover, there is a wide variation in approaches to capture the effect of water pressure on water pipe lines. In the paper of

Kumar et al. (2018) they categorized each water pipe in a water pressure zone due to the elevation differences of the city they used for their case study. A common approach is determining the nominal water pressure per water pipe with hydraulic models, as is done in the papers Winkler et al. (2018), Snider et al. (2018), Karimian (2015), Shirzad et al. (2014), and Jafar et al. (2010). A more extensive effort to capture the effect of water pressure is made in the paper op Martínez-Codina et al. (2016). Besides the nominal water pressure, they were able to include the maximum and minimum water pressure per water pipe and created additional features as 'pressure range'. Based on their results, the most influential indicator among all analysed is the pressure range indicator. This could imply that in general the effect of water pressure on failures is moderate but high pressure peaks has a larger influence on water pipe deterioration. Another water pressure related phenomenon is water hammer, which can be best explained by a a sudden water pressure surge or wave caused when water in motion is forced to stop or changes direction. To the author's knowledge, no paper has researched the effect of water hammer on water pipe failures.

	Physical/Structural							**Operational**			**Environmental**			
Paramaters / Case study	Material type	Nominal diameter	Pipe age	Depth laid	Pipe length	Location	Wall thickness	Operational Pressure	Pipe renewal record	Pipe failure record	Land cover	Soil type	Live load	Climate/Temperature
Verheugd (2020)	✓	✓	✓		✓			✓	✓	✓	✓	✓	✓	
Kumar et al. (2018)	✓	✓	✓		✓	✓		✓		✓		✓	✓	
Winkler et al. (2018)	✓	✓	✓		✓			✓	✓	✓				
Liang et al. (2018)			✓		✓					✓				
Snider et al. (2018)	✓	✓	✓		✓			✓		✓		✓		
Zhang et al. (2018)	✓	✓	✓		✓					✓			✓	✓
Li & Wang (2018)	✓	✓	✓		✓					✓				
Farmani et al. (2017)	✓	✓	✓		✓					✓				✓
Kakoudakis et al. (2017)	✓	✓	✓		✓					✓		✓		
Martínez-Codina et al. (2016)	✓	✓	✓		✓			✓		✓				
Karimian (2015)	✓	✓	✓	✓	✓		✓	✓		✓				
Kimutai et al. (2015)	✓	✓			✓					✓				
Kutyłowska (2015)	✓	✓	✓		✓					✓				
Nishiyama & Filion (2014)	✓	✓	✓		✓	✓				✓		✓		
Shirzad et al. (2014)	✓	✓	✓	✓	✓			✓		✓				
Asnaashari et al. (2013)	✓	✓	✓		✓					✓		✓		
Jafar et al. (2010)	✓	✓	✓		✓	✓	✓	✓		✓		✓		
Berardi et al. (2008)		✓			✓					✓				
Rogers & Grigg (2008)	✓	✓	✓		✓					✓		✓		
Tran et al. (2006)	✓	✓	✓		✓	✓				✓	✓	✓		✓
Pelletier et al. (2003)	✓	✓	✓		✓					✓				
Røstum (2000)		✓	✓		✓					✓		✓		

Table 2.2: Overview of parameters used in case studies for modelling water pipe failures

2.3 Predictive abilities of existing methods

All predictive models for modelling water pipe failures that have been reviewed in the case study analysis can be categorized as a probabilistic or machine learning approach. However, each predictive model has a different predictive ability. These predictive abilities can be classified in the following categories presented below:

1. **Condition assessment**: Within the water distribution practice, companies often utilise a condition rating index (CRI) that evaluates the existing water main condition by rating the pipe on a scale from excellent to an inferior quality condition (St. Clair & Sinha, 2012).

2. **Survival analysis**: By performing survival analysis, the time to the next failure or the probability of failure within a specific time range can be predicted.

3. **Recurrent event analysis**: With a strong foundation in biomedical research, recurrent event analysis focuses on modeling the number of occurrences of failures over time rather than the length of time prior to the first failure.

2.3.1 Condition assessment

One of the greatest challenges facing municipal engineers is the condition rating of buried infrastructure assets, particularly water mains. This is because water mains are typically underground, operated under pressure, and usually inaccessible. Condition rating is a mandatory process to establish and employ management strategies for any asset (Al-Barqawi & Zayed, 2006). The purpose of the condition rating is to objectively rate or scale the current condition of buried pipes. These condition ratings will be used in a maintenance replacement strategy wherein the worst conditioned water pipe lines are given priority. In the work of Winkler et al. (2018) they use a Gradient Boosted Decision Tree (GBDT) technique to predict the network state in 5 and 10 years which are subsequently used for the creation of a tactical rehabilitation plan where the model is employed. In another work of Zhang et al. (2018) a Point Process model is used to predict a risk score per water pipe line for three cases: the risk score for leak failure over the next year; the risk score burst over the next year; and the 'overall' risk score regardless of the failure types. Their Point Process model is classified as probabilistic and is trained with observed historical failure records and structural characteristics of water pipe lines. In an Australian case study for storm water pipes a probabilistic neural network (PNN) approach has been used to categorize the condition of storm water pipes. Based on structural, operational and environmental parameters each water pipe is categorized in three separate levels - good; fair; and poor, in need of further inspection (Tran et al., 2006).

To the author's knowledge there is no standard condition assessment approach for water pipe lines provided in the literature and. Each approach and its scoring metric has been developed in cooperation with the water distribution company.

2.3.2 Survival analysis

Survival analysis (biostatistics), failure-time analysis (engineering), event-history analysis (sociology) or duration analysis (econometrics) provide statistical techniques to estimate the expected time until the occurrence of an event (Donev & Hoffmann, 2019). Schober & Vetter (2018) state that survival analysis, or more generally, time-to-event analysis, refers to a set of methods for analyzing the length of time until the occurrence of a well-defined end point of interest.

Survival analysis can be categorized as proportional hazards model. The proportional hazard model is a general failure model that is used to calculate the hazard function, $h(x)$ of an asset (Wilson et al., 2017). The survival function describes the probability of surviving past a specified time point, or more generally, the probability that the event of interest has not yet occurred by this time point. A hazard rate (or failure rate) is the rate of occurrence of the event during a given time interval (Schober & Vetter, 2018).

In relation to pipe failure modelling, the hazard function can be used to calculate three different parameters, depending on the approach used:

1. The time elapsed between installation and decommissioning (service life).
2. Time elapsed between two consecutive pipe failures.
3. The development of the consecutive failures over the pipe lifetime.

The first and second parameter is applied to the service lifetime or inter-failure time random variables (Wilson et al., 2017). If T denotes the random duration (either service lifetime or inter-failure time), then the hazard function $h(t)$ is defined as a function of time t:

$$h(t) = \lim_{\Delta t \to 0+} \frac{P\{T \in [t, t + \Delta t | T \geq t\}}{\Delta t} \tag{2.1}$$

The numerator of this expression is the conditional probability that the event will occur in the interval $[t, t + \Delta t]$ given that it has not occurred before, and the denominator is the width of the interval. Dividing one by the other we obtain a rate of event occurrence per unit of time (Rodríquez, 2020).

When analyzing survival data, there are three common classes of methods that are distinguished, namely non-parametric; semi-parametric; parametric methods.

Non-parametric Kaplan-Meier method

The non-parametric Kaplan–Meier is based on conditional probability and makes no assumptions about the form of the lifetime distribution. The Kaplan-Meier method can only handle right-censored data and cannot accommodate left- and interval-censored data. In case of water pipe failures, observations of water pipes are terminated before the event occurs, so the actual time-to-event, if it were to occur, is longer than the observation time, but it is unknown by how much. Out-of-sample prediction or estimation of survival probability beyond the last failure time is also not possible. As a further disadvantage, the model can only account for time-constant categorical explanatory variables (e.g. climatic zone freeze/non-freeze) by stratifying the data and separate analysis (Donev & Hoffmann, 2019).

Semi-parametric Cox method

Another popular method is the semi-parametric Cox proportional hazard (PH) model allowing for simultaneous control of covariates without specifying the baseline hazard function. Continuous time-independent variables and interaction terms (product of variables) may be used as explanatory variables (Donev & Hoffmann, 2019). A limitation of the Cox semi-parametric proportional hazard model is that the baseline hazard function cannot be extrapolated outside the range of observed age values. If such an extrapolation is needed, a parametric PHM must be used (Wilson et al., 2017)

Parametric method

In parametric survival analysis the survival and the hazard are fully specified smooth functions being convenient for subsequent optimisation and simulation. Parametric models account for left-, interval- and right censoring providing even out-of-sample estimates beyond the last failure time. The survival time may be modelled as a function of explanatory variables based on underlying proportional hazards.

A parametric survival analysis may be used as an alternative or better in addition to regression analysis for predicting time-to-events. Referring to predicting water pipe failure, the process of crack initiation as well as the occurrence of some random distresses can be described properly

with survival and not regression analysis. However, duration models may be employed also for an estimation of the whole service life (Schober & Vetter, 2018). Survival analysis is a commonly used approach for predicting water pipe breaks within a certain time period and is normally applied to homogeneous groups of water pipes based on pipe age, material type and soil type characteristics. For example, in the paper Kumar et al. (2018) they used the machine learning technique Gradient Boosted Decision Trees to predict which water pipes are most likely to experience a failure within the next three years. More case studies have attempted to model the failure rate of water pipes in Kakoudakis et al. (2017), Karimian (2015), Kimutai et al. (2015), Kutyłowska (2015), Asnaashari et al. (2013), Jafar et al. (2010), Rogers & Grigg (2008), Pelletier et al. (2003) and Røstum (2000).

Recurrent event analysis

Event sequence is becoming increasingly available in a variety of applications such as e-commerce transactions, social network activities, conflicts, and also equipment failures such as water pipe failures. The event data can carry rich information not only about the event attribute but also the timestamp indicating when the event occurs (Xiao et al., 2017). This analysis is classified as recurrent event analysis and is the third parameter that can be calculated by the hazard function. This parameter pertains to the technique of point process and the random variable is $N(t)$, the cumulated number of failures between pipeline installation and age t (Wilson et al., 2017). The hazard function, also referred to as the conditional intensity function, is then defined :

$$h(t) = \lim_{\Delta t \to 0+} \frac{P\{N([t, t+\Delta t])|\mathcal{H}_{t-})\}}{\Delta t} \tag{2.2}$$

The past in a point process is captured by the concept of the history of the process. If we consider the time t, then the history $\mathcal{H}_{t-}$ is the knowledge of times of all events up to but not including time t; $\mathcal{H}_t$ also includes the information whether there is an event at time t. In this equation $N(A)$ denotes the number of points falling in an interval A with the assumption that no failures coincide in time. This conditional intensity function gives the mean number of failure events in a region conditional on the past (Rasmussen, 2018).

The baseline hazard function can be interpreted as a time-dependent ageing component while the covariates represent the risk factors acting to increase or decrease the hazard function (Kleiner & Rajani, 2001). This technique is particularly interesting for those that want to predict events that can occur more than once or in a series of events in which each event has its own failure time (Schober & Vetter, 2018). Furthermore, recurrent event models are also very usable in situations wherein a failure event is determined by one or more previous occurred failure events.

An overview of the earlier discussed case studies is given with their corresponding method used, the output type and category, presented in Table 2.3.

Work	Model classification	Output type	Category
Verheugd (2020)	Recurrent Neural Networks	Failure rate	Continuous
Kumar et al. (2018)	Gradient Boosting Decision Trees	Pipe failure	Binary
Winkler et al. (2018)	Gradient Boosting Decision Trees	Risk score	Continuous
Liang et al. (2018)	Recurrent Neural Networks	Failure rate	Continuous
Snider et al. (2018)	Gradient Boosting Decision Trees	Pipe failure	Binary
Zhang et al. (2018)	Point process	Risk score	Continuous
Farmani et al. (2017)	Evolutionary Polynomial Regression	Pipe failure	Binary
Kakoudakis et al. (2017)	Evolutionary Polynomial Regression	Failure rate	Continuous
Martínez-Codina et al. (2016)	Poisson process	Failure rate	Continuous
Karimian (2015)	Evolutionary Polynomial Regression	Failure rate	Continuous
Kimutai et al. (2015)	Weibull/Poisson/Cox	Failure rate	Continuous
Kutyłowska (2015)	Artificial Neural Networks	Failure rate	Continuous
Nishiyama & Filion (2014)	Artificial Neural Networks	Pipe failure	Binary
Shirzad et al. (2014)	Artificial Neural Networks	Pipe failure	Binary
Asnaashari et al. (2013)	Artificial Neural Networks	Failure rate	Continuous
Jafar et al. (2010)	Artificial Neural Networks	Failure rate	Continuous
Berardi et al. (2008)	Evolutionary Polynomial Regression	Failure rate	Continuous
Rogers & Grigg (2008)	Artificial Neural Networks	Risk score	Continuous
Tran et al. (2006)	Artificial Neural Networks	Pipe failure	Binary
Pelletier et al. (2003)	Weibull/Exponential	Failure rate	Continuous
Røstum (2000)	Poisson process	Failure rate	Continuous

Table 2.3: Overview of model classifications and predictive abilities applied to case studies

2.4 Modelling the intensity function

For preventative maintenance applications such as water supplies network, recent studies of Ertekin et al. (2015) and Yan et al. (2013) have suggested that: First, the failure intensity for an asset i.e. its hazard rate, normally stays at a relatively stable level. This can be regarded as a base component depending on its profile covariants. Second, the occurrence of a failure can often lead to an instantaneous rise of its vulnerability. The reason behind this is that when an water pipe failure occurs, it often becomes more fragile to failures due to the fundamental physical damage. This vulnerability gradually fades back to the baseline when the asset recovers by a certain means e.g. a performed maintenance activity. Third, different types of failures have different triggering effects to each other, e.g. a pipe burst will cause more damage to a leak failure (Xiao et al., 2017).

One expressive mathematical approach for modelling such failure events is point process. Since its first publication in Hawkes (2017) it has many variations relevant to our predictive maintenance case. Some of these variations have been modelled with recurrent neural networks to model the underlying behaviour in various application domains. In addition to the point process models that considers event sequence as input, also recurrent neural networks models have been used with a time-series as input.

2.4.1 Introduction to Recurrent Neural Networks

Recurrent neural network (RNN) is a subclass of ANNs designed for capturing information from sequence/time series data and are specifically designed for understanding long-term dependencies. In a feed forward network signals flow in only one direction from input to the output, one layer at a time. Unlike feed forward networks, a recurrent network can receive a sequence of values as input, and it can also produce a sequence of values as output. The ability to operate with sequences opens these networks to a wide variety of applications, such as predictive maintenance applications

(Abbasi et al., 2019). Illustrated in Figure 2.2, is a feed-forward neural network structure presented where additional edges, referred to as the recurrent edges, are added such that the outputs from the hidden units at the current time step are fed into them again as the future inputs at the next time step. In consequence, the same feed-forward neural network structure is replicated at each time step, and the recurrent edges connect the hidden units of the network replicates at adjacent time steps together along time, that is, the hidden units with recurrent edges not only receive the input from the current data sample but also from the hidden units in the last time step. This feedback mechanism creates an internal state of the network to memorise the influence of each past data sample (Liang et al., 2018).

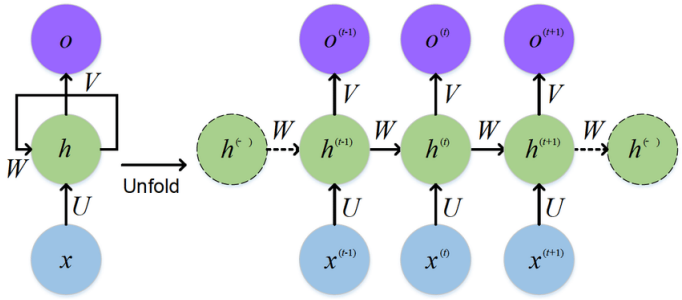


Figure 2.2: Recurrent Neural Network structure (Feng et al., 2017)

Within a RNN structure there can be made use of different types of hidden units, one of the more widely used is Long Short Term Memory (LSTM). LSTM networks are a special kind of RNN that are capable of learning long-term dependencies and was introduced by Hochreiter & Schmidhuber (1997). The LSTM cell is explicitly designed to remember information for long periods of time, avoiding the long-term dependency problem of normal RNN structures. As shown in Figure 2.2, a RNN has a form of a chain of repeating modules with a simple activation function, such as a tanh layer. This chain like structure is also the case for LSTM RNN, but the repeating module has a different structure as shown in Figure 2.3. Instead of one neural activation layer, a LSTM cell has four different layers interacting in a unique way.

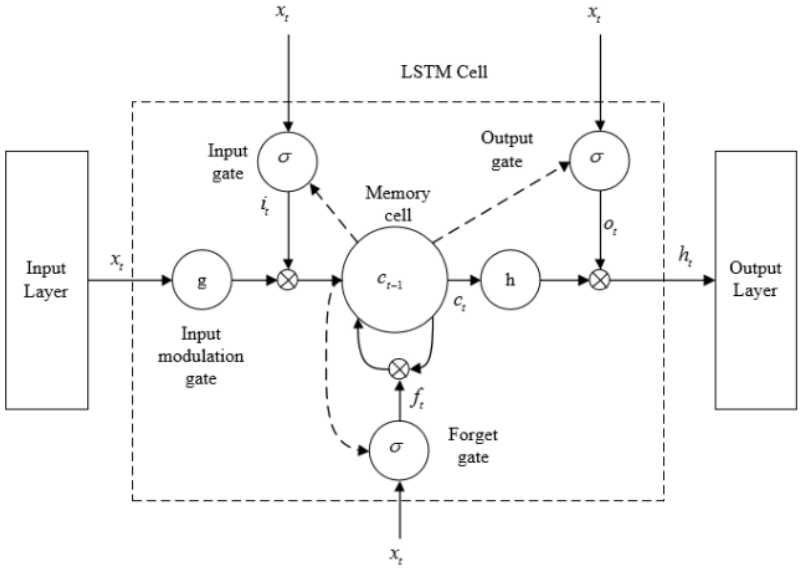


Figure 2.3: Structure of a LSTM cell (Fu et al., 2016)

The key to LSTM cell is its current state, that runs straight down the entire chain via the horizontal line running through the middle of the diagram. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. These gates give the possibility to update the current state of the LSTM with new information, which is composed by a sigmoid layer and a point wise multiplication operation. This sigmoid layer outputs numbers between zero and one, determining how much of each new type of information should be let through. An LSTM has three of these gates, subsequently discussed below.

Forget gate layer

The forget gate layer f_t is there to decide what information needs to be thrown away from the cell state. This decision is made based on the point wise multiplication of the previous LSTM unit state h_{t-1} and the current cell input vector x_t, multiplied by the weight of the forget layer W_f and its bias b_f. The sigmoid layer σ will transform this output to a number between 0 and 1 for each number in the cell state C_{t-1} (Eq. 2.3).

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{2.3}$$

Input gate layer

Now that has been determined what information can be forgotten, there must be decided which new information is going to be stored in the cell state. This is done in two consecutive steps. First, a sigmoid layer σ functions as the input gate layer i_t that decides which values will be updated (Eq. 2.4), based on its weight W_i, previous cell state h_{t-1}, current cell input x_t and its bias b_i. Next, a tanh layer creates a vector of new values, $\tilde{C}_t$, that could be added to the state (Eq. 2.5). This is done by the same point wise multiplication between the previous cell state h_{t-1}, the current cell input x_t and its corresponding weight $W_{\tilde{C}}$ and bias $b_{\tilde{C}}$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{2.4}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{2.5}$$

The old cell state C_{t-1} will be updated to the new cell state C_t. To forget what has been decided before, the old cell state is multiplied by the forget layer f_t. In addition, the old values are updated with $i_t * \tilde{C}_t$ (Eq. 2.6).

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.6}$$

Output gate layer

The last step is determining which information will be put through the output gate layer o_t. This output will be dependent on the cell state, $W_0 * [h_{t-1}, x_t] + b_o$, but first will be filtered by a sigmoid layer σ that decides which parts of the cell state will be transferred to the next cell (Eq. 2.7)

$$o_t = \sigma(W_0 * [h_{t-1}, x_t] + b_o) \tag{2.7}$$

Subsequently, the cell state will be put into a tanh layer to transform the values between -1 and 1. These transformed values will multiplied by the result of Eq (2.7) to only output the desired part of the cell state h_t (Eq. 2.8).

$$h_t = o_t * \tanh(C_t) \tag{2.8}$$

Since the first introduction of RNN in the work of Hochreiter & Schmidhuber (1997), many variants on this initial RNN have been introduced. Without going into details, some of the most notables variants are that of introducing peepholes in the memory cells that connect the cell state with its gates (Gers & Schmidhuber, 2000) and combining the input and forget gate into one Gated Recurrent Unit (Cho et al., 2014).

2.4.2 Point process model

Some events in the world are correlated. A single event, or a pattern of events, may contribute to cause or prevent future events. The ability to discover correlations among events is crucial to accurately predict the future of a sequence given its past, i.e., which events are likely to happen next and when they will happen (Mei & Eisner, 2017). An event sequence, like failure events of water pipe lines, carry important clues about the underlying dynamics, and has lent the event data fundamentally different from the time-series whereby series is indexed with fixed and equal time interval. These events are often associated with continuous time stamps and additional information such as event type or profile features (Xiao et al., 2017).

Hawkes process

As earlier mentioned, previous occurred failures can have a temporal effect on the failure rate λ_k of an asset. One approach to model this underlying behaviour is by means of the Hawkes process (Hawkes, 2017). A Hawkess process assumes that past events at time t can temporarily raise the probability of future events, assuming that such excitation is positive, additive over the past events h, and exponentially decaying with time (Mei & Eisner, 2017). This effect is also known as self-exciting multivariate point process (Eq. 2.9):

$$\lambda_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h)) \tag{2.9}$$

where $\mu_k \geq 0$ is the base intensity of an asset of event type k, $\alpha_{j,k} \geq 0$ is the degree of to which an event of type j initially excites type k, and $\delta_{j,k} > 0$ is the decay rate of that excitation. So when an event occurs, the failure rate will be temporarily elevated, but will gradually decay toward its base rate μ.

Self-Modulating Hawkes process

There are some limitations to the original Hawkes process. For example, some events like inspections and repair activities will decrease the vulnerability of water pipe lines and therefore its failure rate. This is a self-regulating or inhibition property (Ertekin et al., 2015). Moreover, the base line intensity μ does not capture the inherent inertia of some events, which are unlikely until their cumulative excitation by past events crosses some threshold (Mei & Eisner, 2017).

In the work of Mei & Eisner (2017) they propose a model to relax the positivity constraints on $\alpha_{j,k}$ and μ_k, which allows inhibition ($\alpha_{j,k} < 0$) and inertia ($\mu_k < 0$). To prevent that the total activation equation 2.9 becomes negative, they pass it through a non-linear softplus function $f(x)$ to obtain a positive intensity function:

$$\lambda_k(t) = f_k(\tilde{\lambda}_k(t)) \tag{2.10a}$$

$$\tilde{\lambda}_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h)) \tag{2.10b}$$

As t increases between events, the intensity $\lambda_k(t)$ may increase and decrease, but eventually approaches the base rate f_{μ_k+0}, as the influence of each previous event still decays toward 0 at a rate $\delta_{j,k} > 0$.

Profile specific base intensity

In a self-modulating Hawkes process the same base intensity failure rate is shared by different types of water pipes. However this model ignores profile covariants of the water pipe, such as material type and diameter, which can be rather informative to model since the instance level covariants are widely used in classification and regression models for failure prediction (Zhang et al., 2018).

In the work of Zhang et al. (2018) they propose a profile specific self-exciting Hawkes process to model the past failure events in combination with a profile specific base intensity μ_d^s. This profile specific base intensity is implemented in the Hawkes (Eq. 2.9) and incorporates the inherent profile attributes associated with a water pipe, irrespective how the past failure occurred:

$$\lambda_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h)) \tag{2.11a}$$

$$\mu_d^s = \beta_d / \Big(1 + \exp(-\theta_d^T x^s)\Big) \tag{2.11b}$$

This model (Eq. 2.11) incorporates the K profile covariates where $x^s = [1, x_1^s, x_2^s, ..., x_K^s]^T$ and $\theta_d = [1, \theta_{d0}, \theta_{d1}, ..., \theta_{dK}]^T$ is the encoding coefficients for x^s. They also include a scalar parameter β_d governing the weight between base term and the exciting term. It is important to mention that in this profile specific model each water pipe has its own base intensity, but the base intensity is constant over time. To prevent their model of overfitting, they model the base intensity with a relatively small number of parameters.

The predictive ability of their model is the one-year forward failure prediction, as shown in Figure 2.4. Past failures are placed in an even set since the installation date and the end of the failure observe period and the prediction window is set to a one-year prediction period. They associate an m event set $c_s^m{}_{s=1}$ for every water pipe s. Each event set consists of n_s events and every event has a time stamp t_i^s and a failure type d_i^s, therefore $c_s = (t_i^s, d_i^s)_{i=1}^{n_s}$. The observation period is dependent on the event set of the specific water pipe. The starting point is installation date t_s^a, and the ending date is the right-censored point t_s^b which is the current time-stamp.

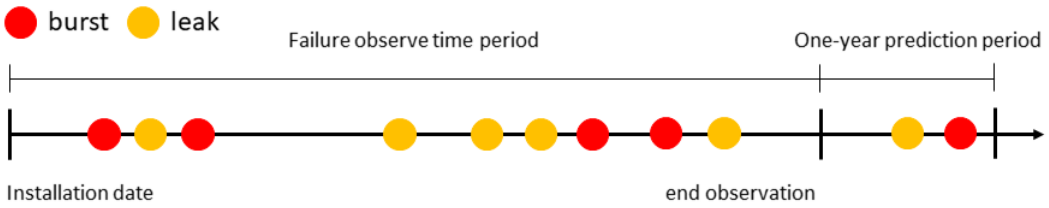


Figure 2.4: One-year forward failure prediction protocol of the profile specific mutual-exciting Hawkes model (adapted from Zhang et al. (2018)).

Neural Hawkes Process

The final model proposed in Mei & Eisner (2017) removes the restriction that the past events have independent, additive influence on $\tilde{\lambda}_k(t)$. They use a RNN LSTM to learn the complex dependence of the intensities on the number, order, type and timing of past events. They refer to this model as the Neural Hawkes Process.

This neural version of the Hawkes behaves differently when compared to the previous variations of the Hawkes model. First, the base rate in the RNN is not a constant μ_k but rather changes per occurred event. Second, the self-regulating and inhibiting influences on $\lambda_k(t)$ can decay at different rates based on different elements of the LSTM hidden unit $h(t)$. Third, due to a sigmoidal or tanh decay function instead of an exponential decay the hidden unit $h(t)$ will behave differently, e.g. the sigmoid function will allow for the possibility that some pair of events will not influence one and another. This can be best described by a schematic example in Figure 2.5 presented in their paper.

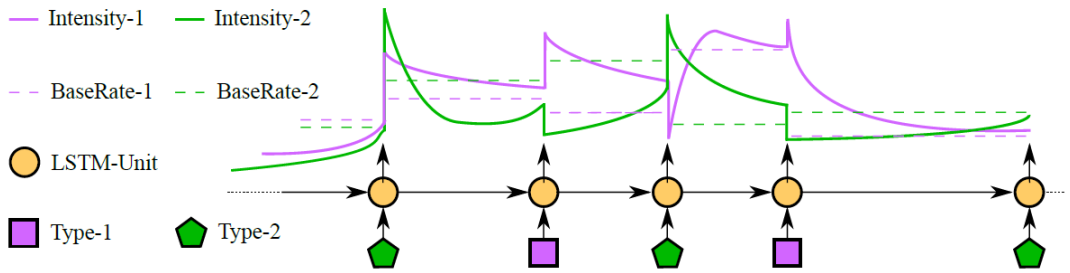


Figure 2.5: Event stream of Neural Hawkes Process (Mei & Eisner, 2017)

An example given in Mei & Eisner (2017), an LSTM reads the sequence of past events (polygons) to arrive at a hidden state (orange). That state determines the future "intensities" of the two types of events—that is, their time-varying instantaneous probabilities. The intensity functions are continuous parametric curves (solid lines) determined by the most recent LSTM state, with dashed lines showing the steady-state asymptotes that they would eventually approach. In this example, events of type 1 excite type 1 but inhibit type 2. Type 2 excites itself, and excites or inhibits type 1 according to whether the count of type 2 events so far is odd or even. Those are immediate effects, shown by the sudden jumps in intensity. The events also have longer-timescale effects, shown by the shifts in the asymptotic dashed lines.

2.4.3 Modelling the intensity function with RNN LSTM

The intensity functions of many point processes involve two components: the background and the effect by the history. In the work of Xiao et al. (2017) they model the background by a Recurrent Neural Network (RNN) with its units aligned with the time series index, while the history effect is modeled by another RNN whose units are aligned with asynchronous events to capture the long-range dynamics. They state that their approach of nonlinear mapping is complex and flexible enough to model various characters of real event data for its application utility.

Their model consists out of two RNNs and is illustrated in Figure 2.6. The top row is a time series that is more suitable to understand the synchronously and possible regularly updated profile features. The bottom row represents the event series which is able to catch the effects of events for the long-term period. The time steps of the RNN of the event sequence are arbitrary while the time steps of the RNN time series are aligned with the evenly spaced unit index of the time series data.

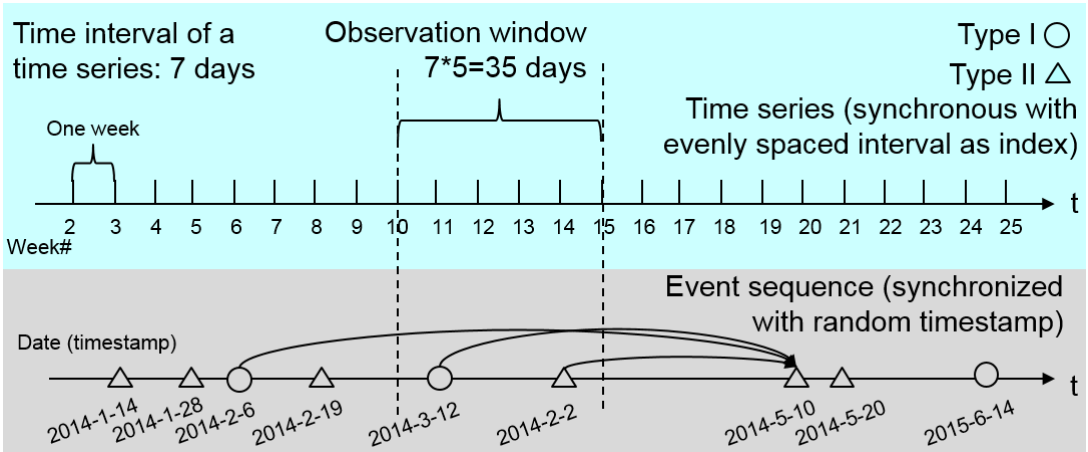


Figure 2.6: Modelling the background intensity and history events with two RNNs (Xiao et al., 2017)

To the best knowledge of Xiao et al. (2017), their work is the first to fully model the conditional intensity function with fused time series and event sequence RNNs. The inputs of the time series and event sequence will be transferred through an embedding mapping layer. Subsequently, the data transformed by the embedding layer is used to classify the main-type and subtype of the next event with a classification loss layer. In addition to this, the timestamp of the next event is predicted with a regression loss layer.

The performance of this model is tested on a predictive maintenance case of a large number of ATMs in North America. For the time series RNN, features such as cash inventory, age, model type and maintenance and error frequency are concatenated as the features of each time series point. The event series consist of the event type and the time interval between two events. They have benchmarked their results with other baseline methods based on several performance indicators, such as the precision indicator. The results of the intensity RNN of Xiao et al. (2017) and the other baseline models are presented in Table 2.4.

	model	main-type	subtype	Hierachical output main-type	Hierachical output subtype
precision	Time series RNN	0.673	0.554	0.582	0.590
	Event sequence RNN	0.671	0.570	0.623	0.614
	Intensity RNN	0.714	**0.620**	**0.642**	**0.664**
	Hawkes process	0.457	0.387	-	-
	Logistic prediction	**0.883**	0.385	-	-

Table 2.4: Precision of Intensity RNN in comparison to baseline methods (Xiao et al., 2017)

In case of the main-type prediction the logistic prediction outperforms all other baseline methods including the intensity RNN of (Xiao et al., 2017). However for the subtype prediction and event timestamp prediction, their method significantly outperforms the other baseline models.

2.5 Position of this research in literature

In this literature review there has been elaborated on the current state-of-the-art on modelling water pipe failures. This literature discusses three main topics: the existing methods for modelling water pipe failures; predictors of water pipe failures; and promising machine learning methods to model water pipe failures. First of all, this literature has shown that recent advancements in the embodiment of a Hawkes Point Process into a RNN structure to model the conditional intensity function, seem promising for predictive maintenance cases. Although many machine learning approaches exist to model water pipe failures, a Neural Hawkes Process is a completely new approach to predict future water pipe failures based on the conditional failure history.

First, this report has shown that the objective of predictive models can differ per water distribution company and this subsequently leads to a preference for a predictive ability. A condition assessment enables a water distribution company to prioritize certain assets in their water pipe replacement program, but lacks the ability to predict when a failure event will occur. The predictive ability of survival analysis is able to determine the remaining useful lifetime of an asset, the time to next failure and the growth rate of failures over the total life time of assets. However, recent works that predict the timing of water pipe failures are applied to homogeneous groups of water pipes or approached as a classification problem for bounded time periods. In the problem considered in this report, there will be attempted to predict the unbounded timestamp of the next failure on an individual water pipe level.

Subsequently, a case study analysis has been performed to analyze which type of physical, operational, and environmental attributes are used to predict these water pipe failures. This case study analysis has shown that already a wide variation of attributes have been tried. However, the effect of water hammer in water distribution systems is yet undiscovered and will be investigated in this research.

In section 2.4, the technique of RNN and its advantages are discussed as opposed to other machine learning techniques. The main advantage of RNN and the use of LSTM cells is its performance in capturing the long-term dependencies of past failures on the failure rate of physical assets. Furthermore, there is elaborated on the literature of modelling the intensity function, i.e. failure rate, wherein RNN is combined with the theory of Point Process, first proposed in Hawkes (2017) as the Hawkes process. The work of Zhang et al. (2018) is the first work that models water pipe failures with a Hawkes process. However, their model requires some level of prior domain knowledge in the form of a parametric failure rate function and does not allow for a time-varying base intensity rate. Both these restrictions are removed when the intensity function is modelled with a RNN, as has been done in the work of Xiao et al. (2017) in another predictive maintenance domain. These characteristics of their Neural Hawkes Process model is seen as a big advantage to the current solutions proposed in the literature. The Neural Hawkes Process model is able to learn the background failure intensity based on the temporal and static water pipe profile attributes with a time series RNN. The long-term effects of historical water pipe failures, such as exciting and inhibiting effects, are captured with an event sequence RNN. In addition, the Neural Hawkes Process model allows to model these background intensities and long-term effects of past failures per individual instances of water pipes. Depending on the availability of time-varying water pipe attributes in our problem, the time-varying base intensity component in the model version of Xiao et al. (2017) could be replaced by a constant base intensity component. We believe that this state-of-the-art Neural Hawkes Process model approach applied in the water utility domain is innovative and contributes to the literature on modelling water pipe failures.

Chapter 3

Extracting data insights

In order to model the intensity function of water pipe failures at Vitens there are two sources of data that are relevant, namely data on the distribution network assets, i.e. water pipes, and the records of the historical maintenance activities that have been performed on these assets. This chapter is devoted to the process of gathering, exploring and extracting data insights from these two datasets. This chapter is for both the distribution network dataset and historical maintenance activities structured in the following manner. First, the gathering of the respective dataset used for modelling the intensity function is discussed. Second, based on the expert knowledge of the asset management department of Vitens, the domain knowledge used to determine outliers in the dataset is introduced. Subsequently, important insights that lie within both datasets are extracted and discussed. Finally, a summary of the data quality is presented and evaluated.

3.1 Dataset: Water distribution network

The current distribution network of Vitens consists of more than 1.2 million unique water pipe IDs. Only a selection of water profile attributes within this dataset are relevant for our predictive maintenance problem. In the following subsections, data insights found within these attributes will be discussed and checked for their feasibility.

3.1.1 Data collection of the distribution network dataset

First of all, the asset management department of Vitens has provided a data file in the form of a shapefile[1] of all water pipes that were in operation at the time of June 2019. However, water pipe renewal records are not present in this most recent overview of operational assets. The failure events that have occurred on these replaced water pipes are important data on the deterioration process of water pipes. Two older archived shapefiles of the distribution network, respectively 2011 and 2017, have been used to determine which unique water pipe IDs are present in 2011 but not in 2017, and which are present in 2017 and not in 2019. These expired water pipes have been added to the current distribution network shapefile. The results of the data gathering for the distribution network are **1212205** operational assets in 2019; **6784** expired assets between 2017-2019 and **66885** expired assets between 2011-2017.

3.1.2 Data quality

In cooperation with the asset management department of Vitens, the data quality of the distribution network dataset is validated with the domain knowledge of Vitens, which is shown in Table 3.1. Since Vitens has undergone several mergers with different water companies over the years, the data set of the distribution network consists of impurities due to data migrations. All

[1]The shapefile format is a geospatial vector data format for geographic information system (GIS) software.

non-valid categories and minimum and maximum feasible values are determined for respectively the categorical and continuous attributes. It is important to note that only water pipes with a diameter between 35 - 800 mm are in scope. Furthermore, installation dates between 1857 and 2019 and water pipe lengths between 1 m and 16955.06 m[2] are considered feasible.

Attribute	Characteristic	Values
Material	Non-feasible categories	Not Select, —, Unknown, **
External diameter	Min. feasible value	35
	Max. feasible value	800
Internal diameter	Min. feasible value	35
	Max. feasible value	800
Nominal diameter	Min. feasible value	35
	Max. feasible value	800
Status	Non-feasible categories	Unknown
Installation date	Min. feasible value	1857
	Max. feasible value	2019
Function	Non-feasible categories	Not Select, Other
Length	Min. feasible value	1 m
	Max. feasible value	16955.06 m
Geometry	Non-feasible values	All non-numerical values

Table 3.1: Domain knowledge for water distribution network dataset

Now that the domain knowledge for the water distribution network dataset is determined, the following subsections will discuss the data insights found in the water distribution network dataset.

3.1.3 Material

The material types present in the distribution network of Vitens can be reduced to seven main categories, as shown in Figure 3.1. The majority of the distribution network consists of PVC, PE, cast iron (GIJ) and AC, while the remaining material types are slowly phasing out or are limitedly used.[3] The material types PVC and PE are currently the preferred material type for their durability and resistance to environmental corrosion.

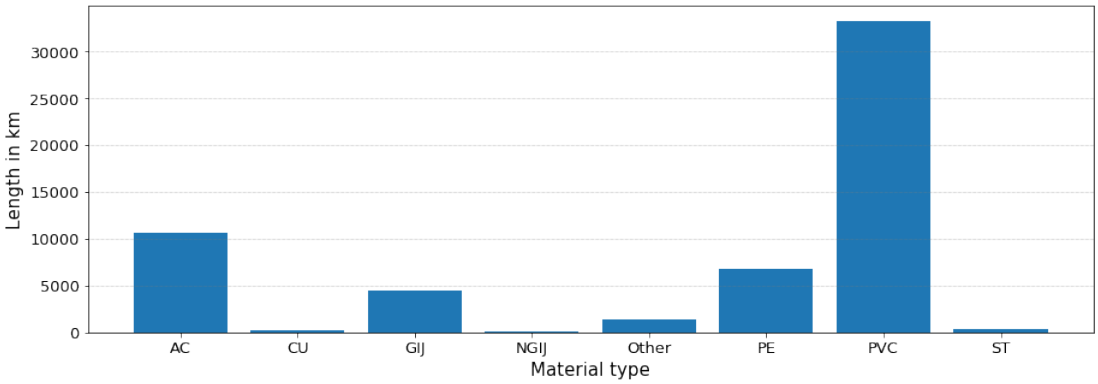


Figure 3.1: Total network length per material type

[2]Highest value has been validated as it concerns a transportation pipe to the west Frisian Islands.
[3]More information on the material types can be found in Appendix A.

After World War II, the use of the material AC became really popular for its low production costs and has been widely used until further use was forbidden around 1990-1995. Unfortunately, the material type AC turned out to be very sensitive to the effect of leaching caused by flowing water. The oldest water pipes present in the water distribution network are made of the material type GIJ and are prone to metal corrosion due to environmental effects.

3.1.4 Water pipe diameter

For all water pipes, three measurements of the structural diameter are included in water distribution dataset. First, the external diameter is the diagonal length measured from the outer edges of a water pipe. Second, the internal diameter is the diagonal length measured from the inner edges of a water pipe. At last, the nominal diameter is the diagonal length measured from the point that is exactly between the external and internal edge towards the same point at the opposite side of the water pipe. The network length of per nominal diameter used in the water distribution network is shown in km in Figure 3.2.

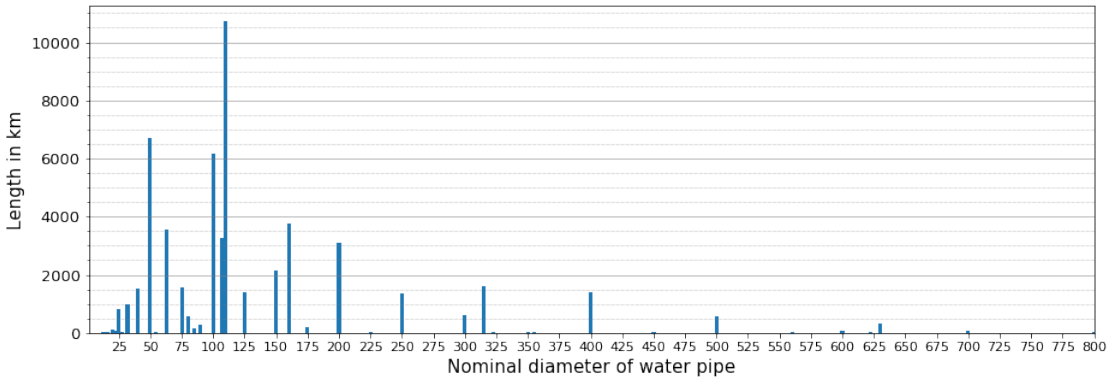


Figure 3.2: Total network length per nominal diameter

It can be concluded that the majority of water pipe diameters fall within a range of 50mm and 200mm with a peak around 100mm. According to Asset Management, the peak of these diameter sizes belong the finely-meshed distribution network, which is the largest part of the total distribution network. All diameter sizes below 35mm function are small connection pipes and are based on the domain knowledge left out of scope in this research. Furthermore, the largest diameters are used for water pipes that have a water transportation function. Their large sizes allow for a high volume and rapid velocity of water.

A missing value analysis per diameter type is presented in Table 3.2. In general, the nominal diameter is of the highest quality compared to the internal and external diameter. Furthermore, the material types Other and ST have a far lower data quality compared to the other categories.

	External	**Internal**	**Nominal**
AC	10.2	0.2	0.2
GIJ	9.9	0.2	0.2
NGIJ	9.8	0.2	0.2
Other	80.6	95.8	78.9
PE	0.2	3	0.1
PVC	0.3	1.1	0.3
ST	8.8	18.2	8.9

Table 3.2: Missing values per diameter measured in total percentage %

3.1.5 Status

The status variable indicates whether the water pipe is still in operation or already has been replaced. The exploratory analysis on this variable has determined that 94.7% of all water pipes are still in operation and the remaining 5,3% have been replaced. Eventually, no failure event predictions will be made for water pipes that are out of operation. However, the deterioration patterns of these replaced water pipes are valuable to predict failure events of water pipes still in operation.

3.1.6 Installation date

The year of installation for each unique water pipe represents its age, which is an important attribute to determine the life cycle phase on a water pipe. The amount of kilometers of water pipe installed between 1857 and 2019 is shown in Figure 3.3. In compliance with statements made in the literature review and Asset Management, the majority of the distribution network of Vitens has been installed after World War II. One of the assumptions of Asset Management is that the water pipes installed shortly after World War II will require replacement in the near future, causing a sudden increase in required maintenance. Since the last 20 years, a stable amount of km water pipes has been replaced. This could be explained by an expansion of the distribution network or the replacement of older water pipes for newer ones. A more detailed figure per material type can be found in Appendix B. The most import data insights from these detailed figures is that the material AC and the use of PE and PVC is the preferred material type for the future distribution network.

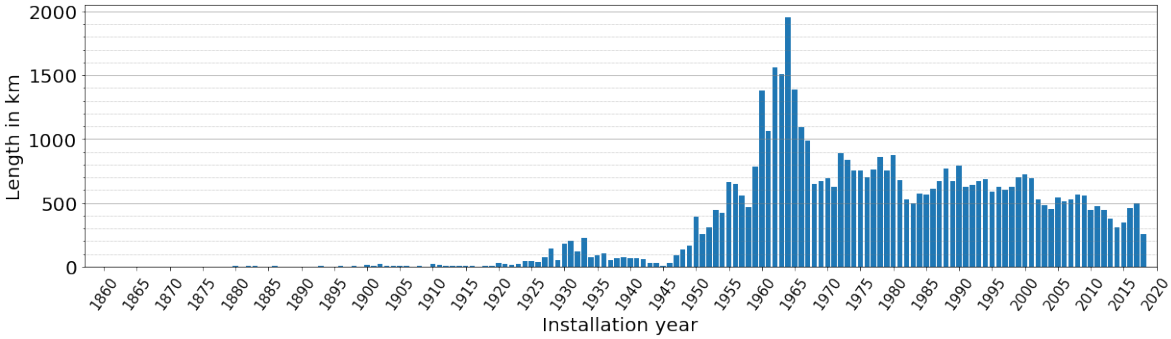


Figure 3.3: Total network length per installation year

3.1.7 Function

All water pipes in the distribution network of Vitens have a different operational function. Transportation pipes transport large amounts of water over a long distance, distribution pipes function as a finely-meshed network within cities and, connection pipes enable water flow from the distribution pipes to households. The remaining and more specific functions are categorized as 'Other'. The amount of network length per function type is presented in Figure 3.4. It can be concluded that most water pipes are part of the finely-meshed distribution network. Although there are more connection pipes than transportation pipes, the total network length of transportation pipes is larger because their function is to transport water over long distances.

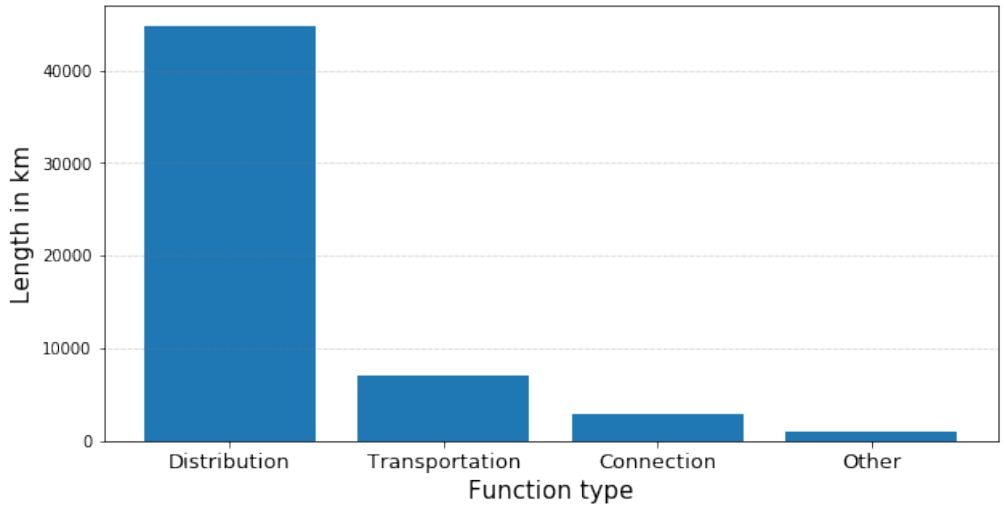


Figure 3.4: Total network length per function type

3.1.8 Length of water pipe

This attribute represents the length of the water pipe of which a histogram is shown in Figure 3.5. Based on this figure, it can be concluded that most water pipes have a length between 0 to 5 meters. An important statement of Asset Management is that an installed water pipe can be split into several unique water pipe IDs. They explain, that some water pipes IDs share the same attributes, are physically connected and installed on the same date. In theory, these water pipes IDs can be considered as one unique water pipe and therefore share the same deterioration pattern. To find these cases and combine the water pipe ID's together into a single ID that represents the original water pipe installed, is an important data transformation step to improve the predictive accuracy of the Neural Hawkes Process model.

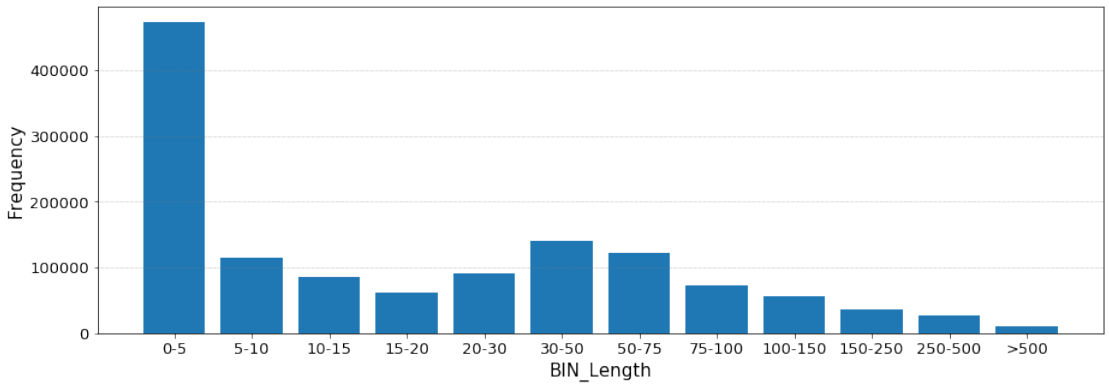


Figure 3.5: Binned histogram of the distribution of the length of water pipes

3.1.9 Geometry

Each unique water pipe ID is linked to a geographical location, denoted as geometry. This geometry attribute is used to gain insights in the physical location and connections within the distribution network of Vitens. The geographical coordination system used is the Dutch 'Rijksdriehoekscoördinaten' and is not similar to the standard GPS coordinates. The coordinate reference systems is coded as EPSG:28992.

3.2 Dataset: Failure records of water pipes

Historical failure records give insights in the life cycle of water pipes in the distribution system of Vitens. These failure records have been put together based on historical maintenance work orders since the year 2005. Unfortunately, all work orders before that year have not been processed into failure events. Within this historical failure data set, only a select amount of variables are relevant. In the following subsections these variables will be introduced, explored and checked for their feasibility with the domain knowledge at Vitens.

3.2.1 Data quality

In the period of 2005 till 2019 there are 35347 data points in the historical failure record dataset. As well for the historical failure records, the domain knowledge of Asset Management is used to check the feasibility of these data points. The domain knowledge is shown in Table 3.3.

Attribute	Characteristic	Values
Cause of failure	Non-feasible categories	Third parties, no failure, Other
Failure year	Min. feasible value	2005
	Max. feasible value	2019
Type of object	Non-feasible categories	Fire hydrant, Service faucet, Other
Material	Non-feasible categories	Unknown, not assigned, Other
Diameter	Min. feasible value	35
	Max. feasible value	800
X & Y	Non-feasible values	All non-numerical values

Table 3.3: Domain knowledge for historical failure records dataset

3.2.2 Cause of failure

From the total amount of water pipe failures recorded, the majority of failures are related to the ageing process of the water pipes, fractures due to soil subsidence and corrosion of the material type, as is shown in Figure 3.6.

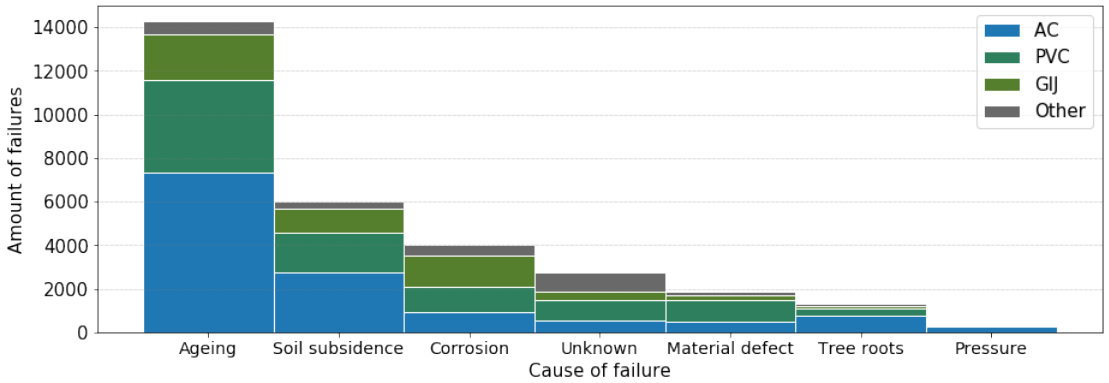


Figure 3.6: The total amount of failure events per type of cause

Although the largest part of the distribution network consists of the material type PVC, most failure events caused by ageing and soil subsidence have occurred on the material type AC. In the last 50 years the material type PVC is more frequently used, while the water pipes of material type

AC are installed 20 to 30 years before that period. In general, it can be concluded that currently water pipes made of AC are becoming of age more often than those made of PVC. Based on the data insights gained in the corrosion category, it seems that the material type GIJ is relatively most prone to this effect. Most failures caused by material defects are seen for PVC because it is the preferred choice of material type. This logically leads to a percentage of water pipes installed that are of bad quality. At last, the lowest number of failures are caused by water pressure but have all occurred on the material type AC. This leaves the assumption that the structural design of AC water pipes is vulnerable to high water pressures.

3.2.3 Type of object of failure

An overview of the amount of failures attributed to a specific object is presented in Figure 3.7. Most failures have occurred on the water pipe itself and its connecting structures. Based on the domain knowledge, the failures occurred on service cranes and the category Other are out of scope. A small amount of failures have occurred on the valves that are used to stop the water flow in a specific part of the distribution network.

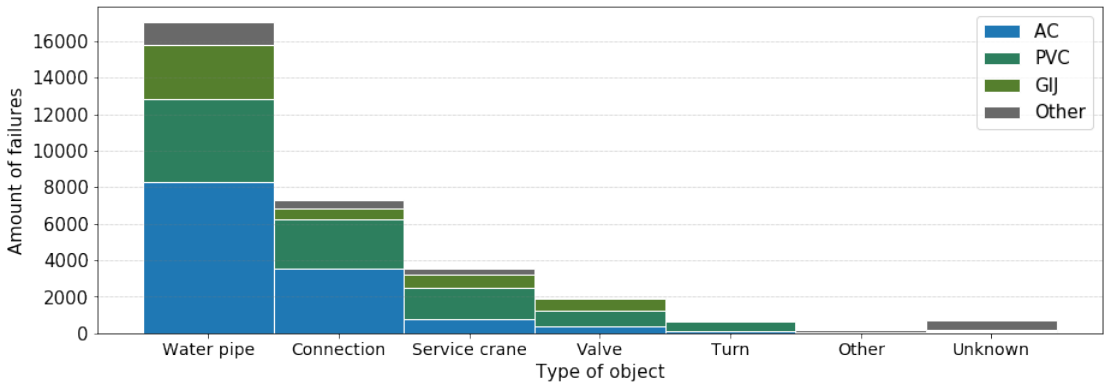


Figure 3.7: The total amount of failure events per type of object

3.2.4 Year of failure

As can be seen from the distribution of failure events per year in Figure 3.8, only a small number of failure events have been recorded in the period between 2005 and 2009. The first event recorded dates from November 2005 until August 2019.

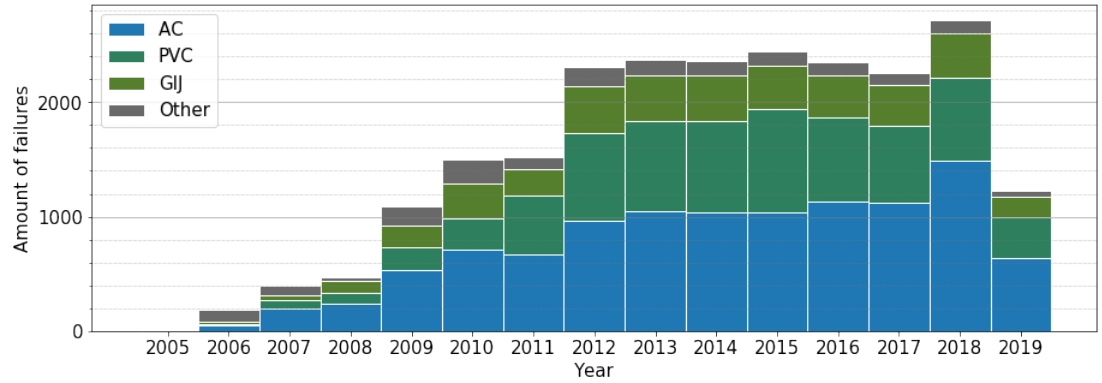


Figure 3.8: The amount of failure events per year

The failure events that are out of scope or have occurred on out of scope objects are already excluded from this overview. Not all failure events are recorded with an accurate timestamp, namely 15.77% of all data points contain a missing year of occurrence. Although Asset Management has stated that on average 2000 failure events occur per year, the data becomes of quality in the period 2009 to 2010. The increased number of failures in the year 2018 can be explained by the extremely dry summer of that year, that caused the ground water to that has led to soil subsidence which resulted in more water pipe bursts.

3.2.5 Material

The exploratory analysis on the material type of the failure events is shown in Figure 3.9. The data insights gained from this figure support the earlier found data insights, the vast majority of failures have occurred on the material types AC, PVC and GIJ. However, in accordance with Asset Management, it can be be concluded that there is a lack of sufficient failure events for the remaining material types to truly understand their deterioration pattern.

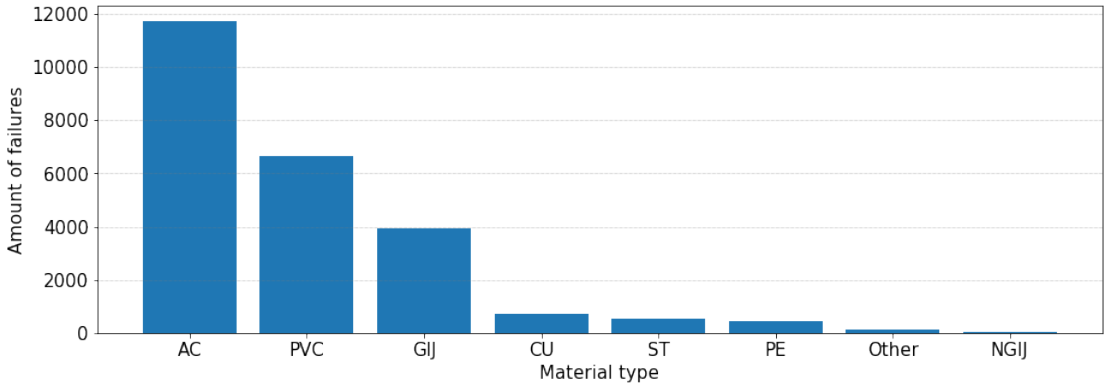


Figure 3.9: The amount of failure events per material type

3.2.6 Diameter

The distribution of failure events occurred per diameter category is presented in Figure 3.10. The domain knowledge states that only the failure events occurred on water pipes with a diameter between 35mm and 800mm are in scope. The majority of failures have occurred on water pipes with a diameter between 100-150 mm but this is also the most frequent diameter size of water pipes installed in the distribution network of Vitens.

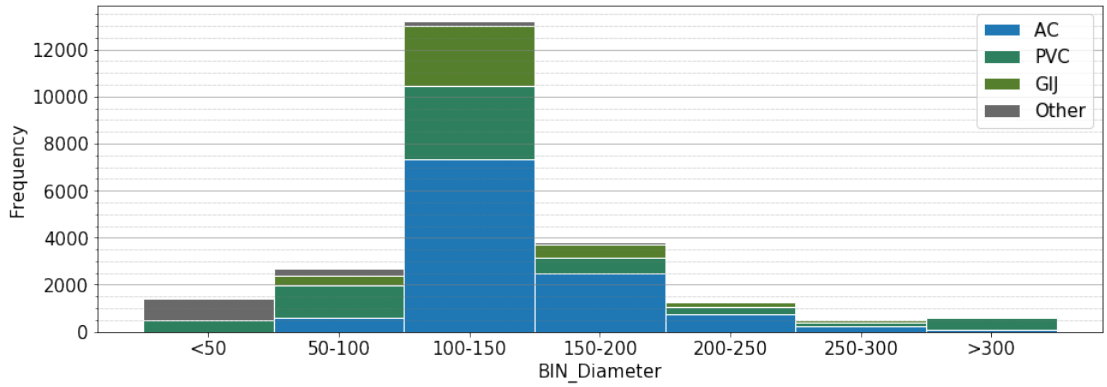


Figure 3.10: Histogram of distribution of failure events per diameter category

The diameter type of the water pipe on which a failure has occurred, differs per material type. Asset Management has stated that the internal diameter is used for the material types AC and GIJ, while the nominal diameter is used for PVC and the remaining material types.

3.2.7 X & Y coordinates

Whenever a water pipe failure has occurred, the maintenance performed is logged with GPS coordinates, with the coordinate reference system EPSG:4326. However, Asset Management explained that the actual location of the failure is not always the recorded X & Y coordinate. In case of old maintenance records, sometimes the street address of clients that reported an initial warning of a water pipe break is used as the location. This negatively affects the data quality of the actual location of the failure event.

3.3 Summary on data quality

These two datasets discussed in this chapter will form the basis for modelling the failure intensity rate of water pipes. Therefore, the data quality of these datasets have been checked by use of the domain knowledge given by Asset Management. An overview of the data quality of both datasets is given in Table 3.4. This has resulted in the fact that 67.2% of the water pipe IDs in the distribution network dataset fall within scope to include in the modelling phase. The data reduction has led to 864030 unique water pipe IDs and 19579 failures that are feasible.

The majority of failures have occurred on water pipes of the material types AC, PVC and GIJ. There is a lack of sufficient failure events of the other material types to truly understand their failure pattern. Together with Asset Management, it has been decided to exclude these material types for further analysis. Another data insight gained is that for the material types AC and GIJ the internal diameter, instead of the nominal diameter, is used for logging failure events. Furthermore, for a large amount of water pipes there is no information available on the date of installation, which is important to use to determine the phase of life cycle of the water pipes. Another important data insight gained is that the historical water pipe failures are not yet linked to unique water pipe IDs, which is needed for the Neural Hawkes Process model to include past failures on an individual water pipe level.

	Attribute	**Missing values in %**	**Data quality in %**
Distribution network	Material	1.8	96.6
	External diameter	2.0	84.9
	Internal diameter	1.9	86.1
	Nominal diameter	0	84.9
	Status	0	99.9
	Installation date	13.2	83.8
	Function	0	97.5
	Length	0	100
	Geometry	0	100
Historical failure records	Cause of failure	4.0	87.0
	Failure year	17.8	82.2
	Type object	1.7	78.9
	Material	1.8	98.2
	Diameter	1.1	94.3
	X & Y	0	100

Table 3.4: Summary on data quality

Chapter 4

Enrichment of the datasets

In Chapter 3 the datasets of the distribution network and historical failure records have been cleaned from noisy and non-feasible data points to improve the data quality. This chapter will discuss the data transformation steps taken to be able to model the failure intensity function with the Neural Hawkes Process model. Furthermore, the water profile attributes will be enriched with operational and environmental attributes from internal datasets of Vitens and external required geographical charts.

4.1 Retrieving installation year

In predictive maintenance cases it is important to know the age of each individual asset. If the installation date of individual water pipe is unknown, it is difficult to compare the life cycle phases between individual water pipes. The data insights gained in Chapter 3 showed that for 13.2% of all unique water pipe IDs, the installation data is unknown or not feasible. Therefore, a dataset published by the Dutch government that includes all houses in the Netherlands along with its construction year (NLextract, 2019), is used to retrieve the installation date for these groups of water pipes. The assumption made at Vitens is that most water pipes are installed during the same construction period of the nearby buildings.

The results of this data retrieval analysis is shown in Table 4.1. Based on the selections made in Chapter 3, only 96137 water pipe IDs that are in scope have an unknown installation year. Three iterative steps are performed to retrieve the installation year, namely nearby households within a buffer of 20m, 40m and 60m. From the construction years of the houses within the buffer, the median value of this selection is used as the installation year for that specific asset. In total 92.1% of missing values have been retrieved and imputed in the distribution network dataset.

Iteration	Retrieved installation years	Retrieved in %
20 m	68366	70.8%
40 m	15438	16.0%
60 m	5108	5.3%

Table 4.1: Results of the retrieval of the installation year of water pipes

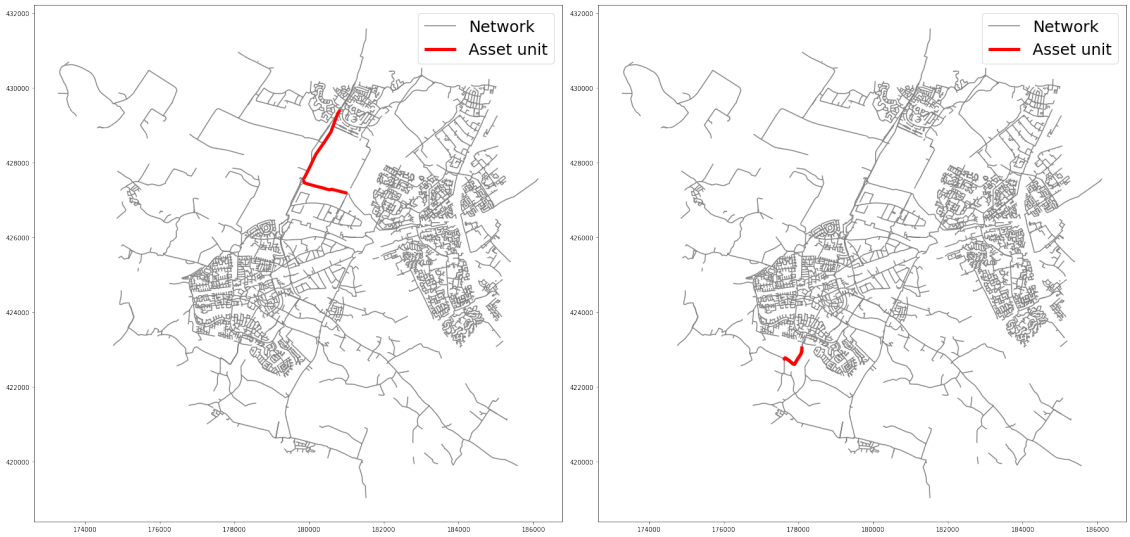
4.2 The creation of asset units

In the previous chapter, Asset Management has stated that some water pipe IDs share the same profile attributes, are physically connected and have been installed on the same date. For an unknown reason these asset units have been split into several unique water pipe IDs. Asset Management argues that in theory these water pipes have the same age, experience the same local

environmental effects and therefore have a similar deterioration process. Chapter 3 has also shown that the location registration of failure events is not always accurate enough to precisely allocate a failure event to unique water pipe ID. Every failure record holds information on the material type and diameter size and can be used to determine the water pipe that belongs to that failure. When the location of a failure event lies near three complete identical water pipes (material type, diameter size and installation year) it cannot be concluded to which water pipe the failure must be allocated. However, since these water pipes share the same deterioration pattern, this failure can be allocated to this group of water pipes. This section describes the approach of finding water pipes that share the same attributes. These groups are from now on referred to as asset units. It is important to note that the creation of asset units is fundamentally different than homogeneous groups of water pipes. Although the homogeneous groups of water pipes discussed in the case study analysis have been created based on the material type, diameter size and installation date, for the creation of asset units the individual water pipes must be physically connected. In this manner, the individual water pipe IDs in the asset unit share the same local operational and environmental effects. Compared to homogeneous groups, water pipes within these groups can be scattered all over the distribution network and experience different environment and operational effects. Based on the domain knowledge of Asset Management the unique water pipe IDs will be grouped based on the following characteristics:

- The asset unit will share the same material type;
- The asset unit will share the same internal or nominal diameter;
- The asset unit has been installed in the same year;
- All water pipes in the asset unit must be connected within a buffer of 16 meter.

In order to compensate for the partial replacement of pipelines within an asset unit, a buffer of 16 meters is used. The buffer of 16 meter is used because Asset Management states this is the standardized tube length. The pseudo code of the algorithm of the creation of asset units is provided in Appendix C.



(a) Asset unit of material type AC and length 3396m (b) Asset unit of material type PVC and length 1730m

Figure 4.1: Visualization of created asset units

Two visualization examples of the created asset units are presented in Figure 4.1. In these two visualizations a small part of the water distribution network is presented, wherein the grey lines

represent the individual water pipe IDs. The red line presents the created asset units, respectively of material type AC and length 3396m and material type PVC and length 1730m. A summary of this data transformation is presented in Table 4.2. This data transformation step resulted in a reduction of around 70% of individual water pipe IDs.

Material	# of water pipes	# of asset units	Data reduction in %
AC	148282	34906	76.5
PVC	703189	222632	68,3
GIJ	95002	26998	71.6

Table 4.2: Summary of created asset units

4.3 Linking failures to asset units

Chapter 3 revealed that a total of 19579 failure events are in scope to modelling the failure intensity function with the Neural Hawkes Process model. Currently, failure events are not linked to specific asset units and are only logged with the corresponding material type, diameter and a geographical location of occurrence of the water pipe. As has been previously discussed, the method of recording the GPS coordinates of the location of the water pipe failure has not always been so accurate.

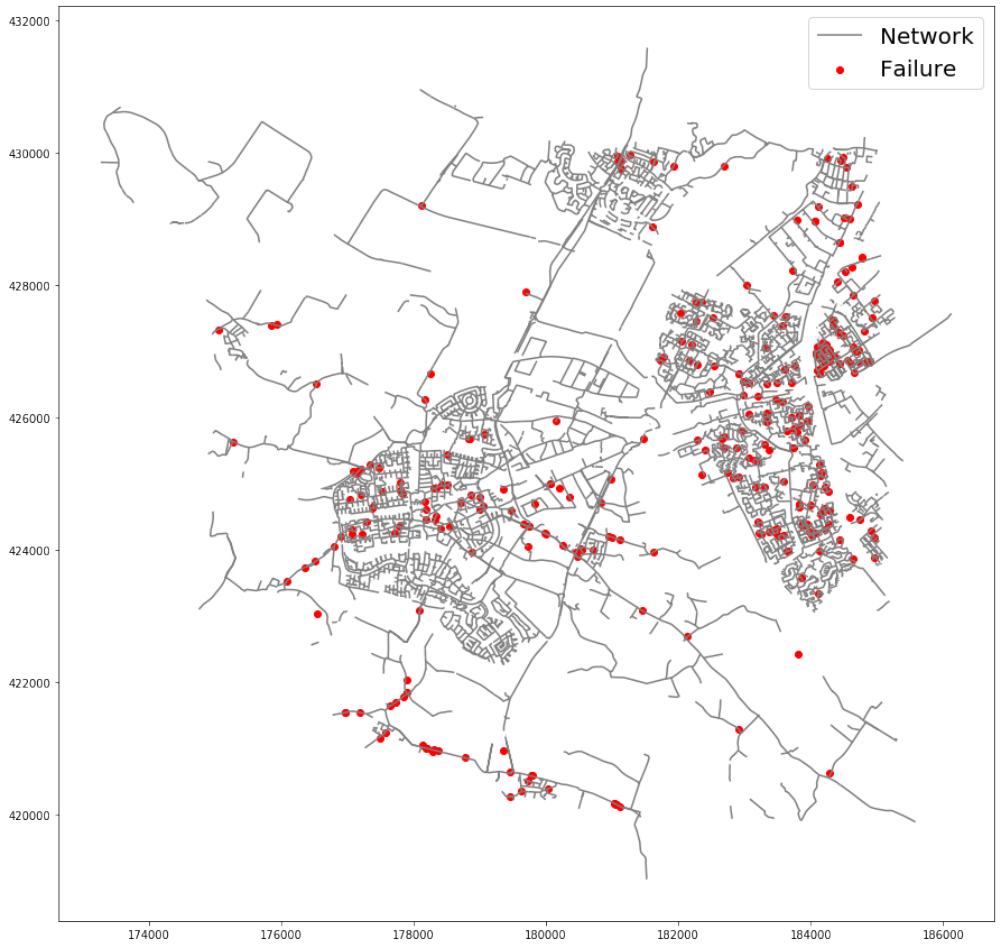


Figure 4.2: Location of failure events

In Figure 4.2, the same part of the distribution network as in Section 4.2 is visualized with the occurred failures of that region. The grey lines represent the distribution network that is made of asset units and the red dots visualize the location of the failure events as included in the dataset. First of all, although most failure events seem to be located on a grey line, i.e. asset unit, for some red dots it is left unclear to which asset unit they need to be allocated. Especially in the areas were the density of water pipes is high, a small inaccuracy in the GPS coordinates can result in a false allocation.

To be able to allocate failure events to asset units in an accurate way, an geospatial algorithm has been developed with GeoPandas[1]. First, the GPS coordinates of the failure events are converted to the 'Rijksdriehoekscoördinaten' reference system of the geometry column in the distribution network dataset. Subsequently, based on the characteristics of the failure event, i.e. material type and diameter, the nearest asset unit of the same characteristics will be allocated to that failure. The original GPS location of the failure will be changed to the location of the found asset unit. As a measure of accuracy, the amount of distance moved is calculated. In accordance with Vitens, all failures within 100 meters moved are assumed reliable. The pseudo code of this geospatial algorithm is provided in Appendix D.

This data transformation step is crucial to link failure IDs to asset units and is the first step in creating a sequence of historical failure events per asset unit. This event set of past failures is one of the components that is used in the Neural Hawkes Process model. The amount of failures allocated to asset units per material type and the percentage of failures moved within a specified range is presented in Table 4.3. It can be concluded that in total 16182 failures are in scope for further analysis.

Material	**# of failures**	**<200 m**	**<100 m**	**<50 m**	**<10m**	**<1m**
AC	10199	0.91	0.88	0.85	0.64	0.55
PVC	5369	0.86	0.83	0.78	0.59	0.49
GIJ	3183	0.88	0.86	0.84	0.65	0.51

Table 4.3: Accuracy of allocated failures to asset units per material type

4.4 Feature engineering

The base intensity rate of an asset unit is determined by its profile attributes. The literature study has shown that all attributes can be categorized as structural, operational and environmental variables. The operational and environmental characteristics for each asset unit vary per location. The case study analysis, summarized in Figure 2.2, has identified the most promising predictors for water pipe failures. This has led to four additional data attributes that are added to the profile of every asset unit.

4.4.1 Water hammer

Vitens has a total of 196 production locations where water is extracted from below the ground and pumped into the distribution network. The effect of water hammer is explained by a sudden surge of water pressure that can cause a water pipe to break. In addition, there are in total 26 water accelerators to ensure a sufficient velocity of flow of water. These production locations and accelerators can cause a sudden surge of water pressure, which could cause a water pipe to break. With the absence of water pressure sensors installed in the distribution network, it is difficult to measure the maximum operational pressure experienced per asset unit. The coordinates of the production locations and accelerators are used to calculate the celestial latitude between the geometrical centroid of the asset unit and the nearest production location or accelerator. The

[1]GeoPandas is a geospatial analysis package for the programming language Python

assumption is made that the asset units closest to these locations will experience the largest effect of water hammer.

A visualisation heatmap of the distance in km is shown in Figure 4.3, wherein each asset unit is presented with a color. The red colored asset units are closest to these production locations while the yellow colored asset units are the furthest away from these production locations. The production locations and water accelerators are visualised as black dots.

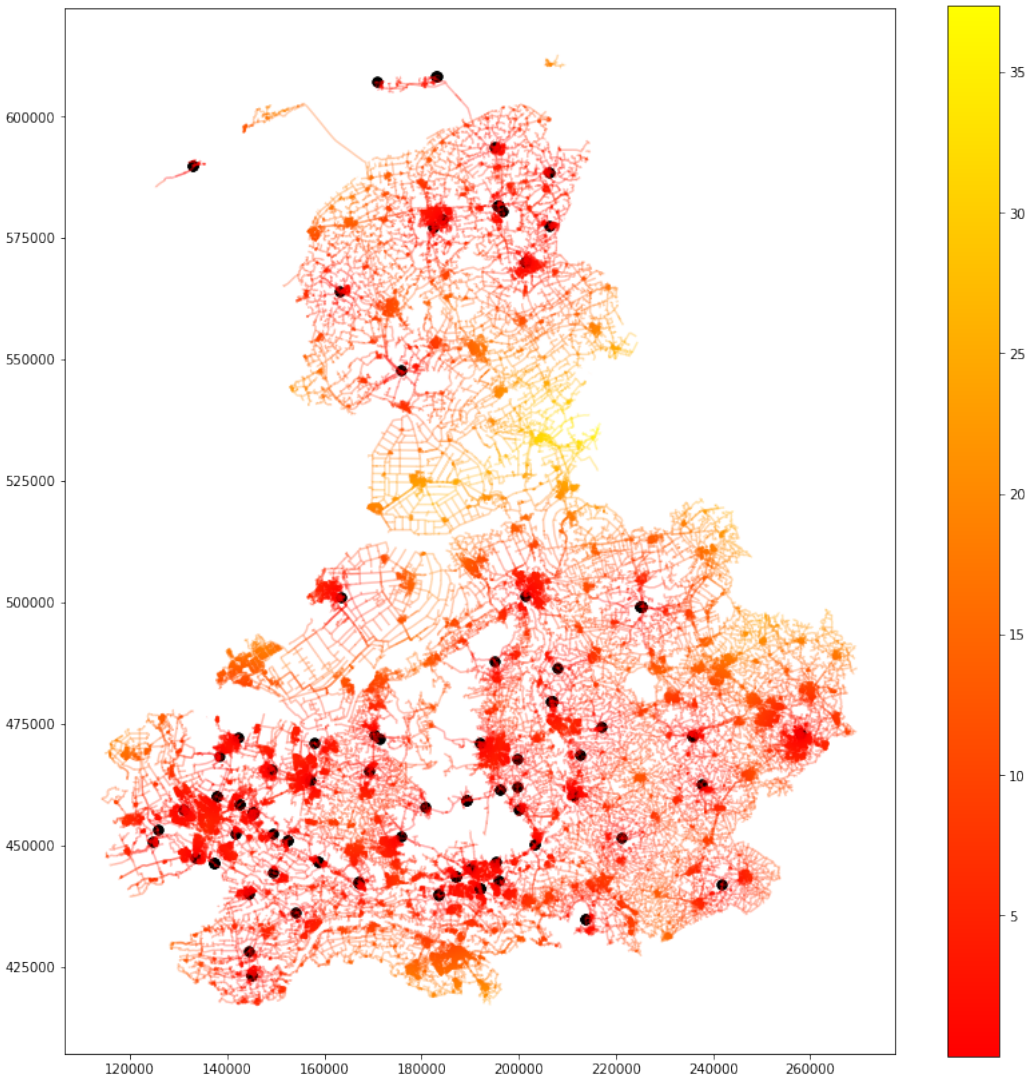


Figure 4.3: Distance (km) to nearest production location or accelerator

4.4.2 Appendages

Asset management provided a shape file with in total 2.2 million appendages, i.e. household connections, that are connected to distribution pipes. The literature study and expertise within Vitens has stated that the number of appendages per water pipe is related to a higher risk of leaks and breaks. The number of appendages per 100m length of the water pipe is added as a profile attribute. In this way the effect the number of appendages is independent of the length of the water pipe. The results of this data preparation step is shown in Table 4.4.

# appendages	# asset units	Percentage of total %
0	133232	46.8
1 - 5	72866	25.6
6 - 10	22073	7.8
10 - 20	18852	6.6
>20	16163	5.7

Table 4.4: Overview of number of appendages per asset unit

4.4.3 Ground soil

The water distribution network covers five provinces in the Netherlands. This implies that environmental effects can vary per region. One of these environmental effects is the soil category. Each type of soil has a different effect on a specific material type, making it an important feature to include in the asset unit profile. A geographical chart of the different types of soil in the Netherlands was available within Vitens. Each asset unit will be assigned to a soil category or a combination of several categories. The results of this data enrichment step is shown in Table 4.5.

Soil type	# of asset units	Percentage of total %
Buildings, etc.	159498	44.3
Sand	94402	26.2
Sandy clay	44680	12.4
Clay	28035	7.8
Combination	22966	6.4
Other	10082	2.8

Table 4.5: Overview of asset units per soil type

4.4.4 Vegetation

The land cover above an asset unit can be categorized into four classes, namely high green, low green, rural and non green. The amount of vegetation near the location of a specific asset unit represents the risk of tree roots that can damage the structure of the water pipes.

It is possible that an asset unit can cross through numerous regions with a different type of vegetation. Therefore, a priority heuristic will be used to maintain the highest class experienced per asset unit. The distribution of the vegetation classes is presented in Table 4.6 with the highest class as the first entry and respectively descending.

Vegetation	# of asset unit	Percentage of total %
High green	37834	12.8
Low green	4232	0.02
Rural	196344	66.8
Non green	55475	18.8

Table 4.6: Overview of asset units per vegetation class

4.5 Conclusion on data enrichment

To conclude on this chapter, several data transformation steps and the enrichment of the asset unit profile resulted in an improved quality of the datasets to be used for modelling water pipe failures. First of all, 92.1% of the missing installation years of water pipes have been retrieved through a land registry made available by the Dutch government. This was an important step to the problem of this thesis, because in order to determine the life cycle phase of a water pipe, the installation year is needed to compare the deterioration process of water pipes of different ages. Subsequently, asset units have been created to group individual water pipes into groups based on a similar material type, diameter, installation year and that they are physically connected. Asset Management has stated that the individual water pipe ID's in an asset unit have an identical profile and therefore have a similar deterioration process. The creation of asset units is fundamentally different than the use of homogeneous groups due to the fact that individual water pipe ID's in the asset unit must be physically connected. In total 946.473 unique water pipe ID's have been grouped into 284.536 asset units.

Subsequently, the inaccuracy of the location coordinates of failure events is resolved with a developed geospatial algorithm in GeoPandas. This geospatial algorithm has been used to geographically allocate failure events to the nearest asset units, given the same material and diameter characteristics. The distance between the GPS location of the failure to the nearest asset unit is used as a measure of accuracy. In agreement with Asset Management, all failures that have been allocated within 100 meters from the original location are assumed feasible. This resulted in the fact that 16182 failures will be used for modelling the failure intensity function of water pipes.

Furthermore, the profile of the asset unit is enriched with four operational and environment variables attributes. When compared to the attributes included in the case study analysis shown in Table 2.2, the profile of the asset units include the most promising predictors as stated in the literature. The effect of water hammer on water pipes, the number of appendages per 100 meter, the ground soil category and land vegetation above an asset unit is added the the profile. The effect of water hammer, measured by the distance to the closest water production location or water accelerator, is a new and innovative feature that has not been used for modelling water pipe failures in the literature and at Vitens.

Chapter 5

Failure intensity function of water pipes

This chapter will discuss how the enriched profile of the asset units and the allocated failures to these asset units are transformed in order to model the failure intensity function of water pipes. First, a motivation is given to use an adapted Neural Hawkes Process model to learn the failure intensity function of water pipes. The Neural Hawkes Process model is able to learn the base failure intensity rate of profile attributes of the asset units. The long-term effects of past failures on this base failure intensity rate are captured in a failure event sequence. Subsequently, the predictive label for training the model is introduced and a feature importance analysis is performed. At last, the characteristics of the Neural Hawkes Process model are presented.

5.1 Motivation for Neural Hawkes Process model

The literature study performed in this report has resulted in the state-of-the-art of modelling an intensity function, namely the embodiment of a Hawkes Point Process model into a recurrent neural network structure. In the paper of Xiao et al. (2017), the intensity function is modelled by combining time-varying background features and the sequence of past failure events, and is called the Neural Hawkes Process model. This model is able to capture the base intensity failure rate from the background features with a time-series RNN and the long-term effects of a sequence of past event with an event sequence RNN. Based on the combined outcome of both RNNs, the timestamp of the next failure event is predicted.

An adapted version of this Neural Hawkes Process model will be used for modelling the failure intensity function of water pipes to predict the next failure event on asset units. The enriched profile attributes of the asset units do not vary over time, e.g. the structural characteristics, type of soil and distance to the closest pump are all static features. Therefore, the background failure intensity RNN in the work of Xiao et al. (2017) is replaced by a static vector that incorporates the profile features during training of the model. The structure of the RNN event sequence is unchanged and the output will be merged with the static profile vector. This combined vector will be processed through several hidden layers before a timestamp prediction is made for the next failure event. The author believes that the combination of the base failure intensity rate determined by the profile features and the influence of past failures on that base intensity rate on an asset unit level, is innovative in the water distribution application domain.

The advantages of this adapted Neural Hawkes Process model structure is that the underlying behaviour of water pipe failures is learned by training the model weights end-to-end. No prior domain knowledge, i.e. no parametric behaviour function, is needed to incorporate the effect of past failures on this failure intensity function.

5.2 Base failure intensity rate

In the Neural Hawkes Process model the base intensity rate will be learned from the profile attributes of the asset units. The profile attributes of the asset units consist of structural, operational and environmental characteristics that determine their unique deterioration process. Four of these attributes have been added to the asset unit profile during feature engineering in Chapter 4. An overview of these attributes is shown in Table 5.1.

Attribute	**Data type**
Material (3)	Binary
External diameter	Numerical
Internal diameter	Numerical
Nominal diameter	Numerical
Installation date	Numerical
Function (3)	Binary
Length GIS	Numerical
Appendages per 100m	Numerical
Distance to closest pump	Numerical
Ground soil (6)	Binary
Vegetation (4)	Binary

Table 5.1: Attributes of asset unit profile

The profile attributes of the asset unit are all static and do not vary over time. In the original Neural Hawkes Process model of Xiao et al. (2017), the base intensity failure rate is modelled with time-varying features and the use of a RNN. Due to the availability of only static profile attributes of the asset units, the base intensity failure rate will be incorporated with a static vector in the Neural Hawkes Process model.

The Neural Hawkes Process model is not able to handle categorical data input, i.e. material type, the function of the asset unit, ground soil and vegetation. The categorical attributes in the static profile vector will be one-hot encoded to transform these attributes into binary columns per attribute value. The categorical profile attributes are transformed to 3 material, 3 function, 6 soil categories and 4 vegetation features. In addition, the numerical features will be min-max normalized according to Eq. 5.1.

$$x' = \frac{x_i - min(x)}{max(x) - min(x)} \tag{5.1}$$

For all values in a numerical feature x_i, its value will be transformed that it falls between the range 0 and 1, respectively the lower and upper bound. The lower and upper bound are respectively determined by the minimum and maximum value of the numerical feature.

5.3 Long-term effects of past failure events

Some events in the world are correlated, whereas a single or a sequence of events may help to cause or prevent future events. The same effect is assumed of the effect of past failure events on the probability of future failures of asset units. The long-term effects of past failure events per asset unit will be learned from a failure event sequence per asset unit. Two attributes of the historical failure records dataset are used for creating a failure event sequence per asset unit, namely the timestamp of failure and the type of failure. These two attributes will be used as the input for the event sequence RNN in the Neural Hawkes Process model.

5.3.1 Failure event sequence

The distribution of failure events will be constructed in tuples with a type of failure and time of occurrence $(k_1, t_1), (k_2, t_2), ...,$ where each $k_i \in \{1, 2, ..., K\}$ is an event type and $0 \leq t_1 \leq t_2 ...$ are times of occurrence, denoted in the amount of years since installation. The idea behind constructing the type of event and time of occurrence is that the effect of past events on the base failure intensity rate differs per failure type. This was explained in more detail in Chapter 2. Past events may now either excite or inhibit future events. They do so by sequentially updating the state of the event sequence RNN LSTM.

Boundary conditions on failure event sequence

All event sequences of the asset units will start with a special beginning-of-stream (BOS) event (k_0, t_0), where k_0 is a special event type and t_0 is set to 0. This special event type expands the RNN LSTM's input dimensionality by one and will represent the installation date of the asset unit, i.e. the birth event. The initial configuration determines the hidden state h_t and the intensity function $\lambda_k(t)$ over $t \in [0, t_1]$. Another reason for implementing BOS is that data before 0 are simply missing, e.g. the observation of the asset unit starts at an arbitrary point in time. In both kinds of settings, the initial configuration just after reading BOS event characterizes the model's belief about the unknown state of the true system just after time 0, as it waits for event 1.

Arbitrary length of the failure event sequence

The event sequence RNN LSTM in the Neural Hawkes Process model only functions with predefined input shapes. In case of failure events, not all asset units have experienced the same amount of failures during the observation period. This implies that the failure event sequences can have an arbitrary length of failures which cannot be directly used as input for the event sequence RNN LSTM. In order to tackle this problem, the event set will be left padded with zeros in order to obtain the required input size. For example, the desired failure event sequence must consist of the last 4 failures that have occurred on the asset unit. The total length of this failure event sequence will be these 4 failures and the initial BOS event. During the observation period of failures used in this report, some asset units will have experienced 4 or more failures. However, most asset units will have experienced 0 or a few failure events. To transform the size of all failure event sequences to the required input shape, the technique padding is used. An example of both a padded and unpadded event set of event size 5 is shown in Eq. 5.2, wherein the BOS event is denoted as $k = 1$. In the first example, the failure event sequence of the asset units consists of a BOS event and a failure event of type 3 with a time of occurrence of 47 years since its installation year. In the second example, the asset unit has experienced 4 failures of failure types 5, 2, 4 and 3. Respectively, these failure events occurred after 37, 47, 52 and 55 years since its installation year.

$$\Big((0,0),(0,0),(0,0),(1,0),(3,47)\Big) \tag{5.2a}$$

$$\Big((1,0),(5,37),(2,47),(4,52),(3,55)\Big) \tag{5.2b}$$

Finally, the tuples in the failure event sequences must be transformed in order to enable the event sequence RNN in the Neural Hawkes Process model to be able to read these event sequences. First, the event type will be one-hot encoded to binary column values. Secondly, the times of occurrences must be min-max normalized according to Eq. 5.1.

5.3.2 Characteristics of the created event sequences

In Chapter 4, asset units have been created and the water pipe failures in scope have been geographically allocated. In total 16182 feasible failures have occurred on 10338 asset units, resulting in an equal amount of failure event sequences of arbitrary length. Unfortunately, the installation year of 135 asset units in scope is unknown and could not be retrieved. These asset units are not included during further analysis. A histogram distribution of the number of failures per asset unit is shown in Figure 5.1. Most failure event sequences only consist of one failure event with a maximum of 20 failures occurred on one asset unit. All asset units that not have experienced a failure within the observation period of the historical failure records are excluded from this visualization.

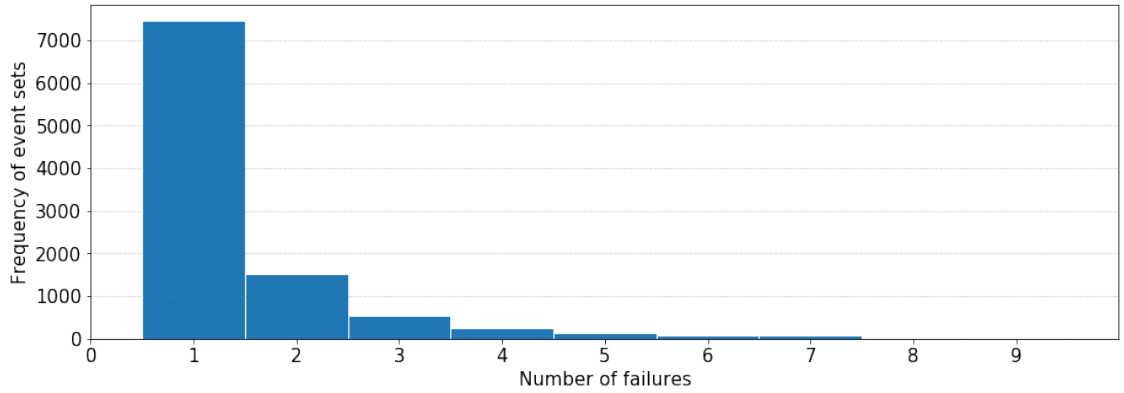


Figure 5.1: Histogram of the number of failures per event sequence

5.4 Feature importance

The profile vector of an asset unit and its failure event sequence have been introduced in the previous sections. The Neural Hawkes Process model will use these two components to predict per asset unit the occurrence of the next failure event. In this section, the predictive label used for training the Neural Hawkes Process Model will be introduced. Next, the correlations of the profile features of the asset units to this predictive label are tested. At last, a conclusion on this feature importance analysis is given.

5.4.1 Predictive label

The Neural Hawkes Process model is used to model the failure intensity function based on the profile features of the asset units and the long-term effects of historical failures event sequence. In Chapter 1 is explained that the ambition of Vitens is to use a machine learning technique to determine when an asset unit has reached its economical end-of-life. The economical end-of-life of an asset unit implies that when a failure occurs, it is economically more beneficial to replace the asset unit than to repair it. The asset management department of Vitens has defined the economical end-of-life as follows:

"When a consecutive water pipe failure is predicted within two years since the last occurred failure, within a range of 400m of the last failure, the water pipe has reached its economical end-of-life. "

Based on this definition, the Neural Hawkes Process model will predict the occurrence, i.e. timestamp, of the next failure on an asset unit level. The units of this predictive label is denoted in the number of years since the year of installation. The predictions of the Neural Hawkes Process model will be used to determine the time difference between the predicted timestamp and the last failure occurred on that asset unit. Based on the calculated time difference, it can be

concluded that for failures predicted in less than two years, the asset units will have reached its economical end-of-life if the failures occurred within 400 meters of each other.

Now that the predictive label has been introduced, this raises the problem that most asset units have not experienced a failure within our failure observation period. In total 274.201 asset units will not have a predictive label when chosen for this predictive label and model approach. It can be argued that these asset units are not relevant to include in further analysis based on the definition of the economical end-of-life. Only assets units that have experienced at least one failure can theoretically reach their economical end-of-life, because its needs a point of reference to determine if a next failure will occur within two years. Therefore, the asset units that have not experienced a failure event during the observation period are excluded from further analysis.

The distribution of the predictive labels of the 10203 asset units, used for training the Neural Hawkes Process Model, is shown in Figure 5.2. All labels presented in this figure denote the time of occurrence since the year of installation of the asset unit.

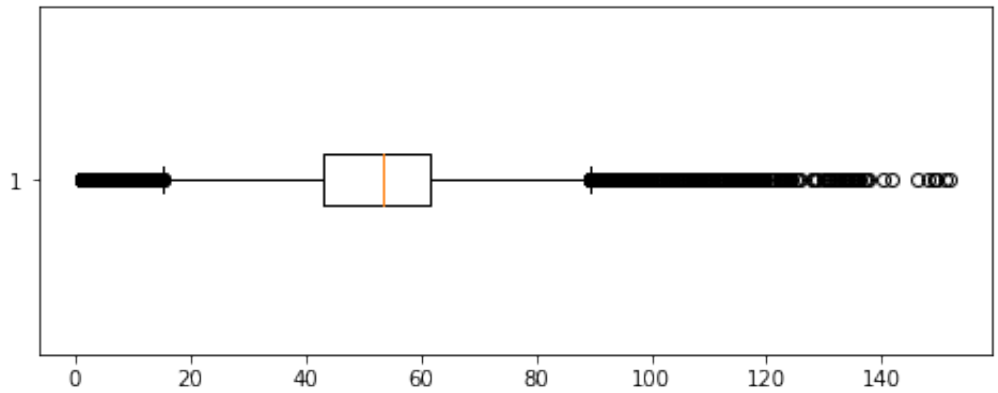


Figure 5.2: Box plot of the predictive label

From the box plot it can be concluded that 50% of the predictive labels, i.e failure events, fall between 18 to 90 years since the installation of the asset unit. The data points below the 25% whisker are explained due to the effect of premature failure of an asset unit. It is also important to include the data points above the 75% whisker of the box plot, because these data points represent the failure events on the oldest asset units present in the distribution system.

5.4.2 Correlation analysis

A final step before modelling the failure intensity function of asset units with the Neural Hawkes Process model is testing the correlation between the profile features and the continuous predictive label. The correlation between the event sequences and the predictive label can only be tested after training the Neural Hawkes Process model. This will be addressed during model evaluation. There are two main reasons for performing a feature importance analysis;

1. Reducing the total training time due to the exponential increase with the number of features;

2. Predictive models experience increased risk of over-fitting with a higher number of features.

Therefore, a feature importance analysis is performed to determine any correlation between the profile attributes and the predictive label, i.e. the timestamp of next failure event. There exist several appropriate techniques to compute the correlation between categorical and continuous variables. However, there exist no single technique to compare the correlation of categorical and continuous variables for a continuous label. Moreover, some techniques only test for linear correlation while there may also be a non-linear correlation.

First, a Pearson correlation matrix has been made to determine any linear relationships between the continuous variables and the continuous prediction label. Subsequently, a Spearman correlation matrix is made to determine any non-linear correlation between the continuous variables and the continuous predictive label. Based on both the Pearson and Spearman correlation matrix, it can be concluded that the feature of the year of installation of the asset unit, is almost perfectly correlated with the predictive label. The observation period of the water pipe failures ranges from 2005 until 2019. This logically implies that all predictive labels fall within the same time period. When the installation date of an asset unit is incorporated in the profile vector, it is assumed that undesirably knowledge is transferred to the prediction and will result in a Neural Hawkes Process model that is biased. Since the age of the asset unit is also incorporated in its event sequence, the installation year of the asset unit is removed from the profile vector during training of the Neural Hawkes Process model. Furthermore, it can be concluded that there exists a strong linear correlation between the external, internal and nominal diameter feature. The Spearman correlation matrix shows a stronger non-linear correlation between the prediction label and the features of appendages, water hammer, water pipe length and diameter. Both correlation matrices are included in Appendix E.

In order to test whether there is significant correlation between the categorical attributes and the continuous label, a point biserial analysis has been performed in Table 5.2. The point biserial analysis is a variation on the Pearson's correlation coefficient and is performed on the artificially binarized categorical attributes. In this table, the correlation coefficient and the probability is respectively shown for each categorical value. Based on the results of this analysis, four categorical values do not show a significant correlation with the prediction label. The effect of excluding these variables during training of the Neural Hawkes Model will be discussed during the results section. In addition, the importance of the failure event sequence will also be addressed during the results section.

		Correlation coefficient	**p value**
Function	Connection pipe	-0.00	0.46
	Distribution pipe	0.04	0.00
	Transportation pipe	-0.04	0.00
Material	AC	0.09	0.00
	GIJ	0.50	0.00
	PVC	-0.50	0.00
Soil type	Buildings, etc	0.10	0.00
	Combination	0.01	0.36
	Clay	-0.09	0.00
	Sand	0.05	0.00
	Sandy clay	-0.02	0.03
	Other	-0.02	0.03
Vegetation	High green	0.02	0.08
	Low green	0.04	0,00
	Rural	0.00	0.84
	Non green	-0.05	0.00

Table 5.2: Point biserial correlation analysis between categorical features and predictive label

5.5 Neural Hawkes Process model

The design of the Neural Hawkes Process model consists of two components, a static vector that incorporates the profile features of the asset units and a failure event sequence. First, the architecture of the model and its initial model parameters are discussed. Next, the activation behaviour of the Neural Hawkes Process model is discussed. To learn the model weights that represent the failure intensity function of the asset units, the loss function for the Neural Hawkes Process model is introduced. At last, the validation methods to evaluate the performance of the Neural Hawkes Process model is motivated.

5.5.1 Architecture

The architecture of the Neural Hawkes Process model is visualised in Figure 5.3. The architecture consists of two input layers, an input layer for the static profile vector and a three-dimensional input layer for the event sequence, i.e samples x event steps x features. The input of the event sequence layer is connected to the event sequence RNN with LSTM cells. The static profile features and the output of the event sequence RNN LSTM are merged into one vector and processed through two Dense layers. Finally, a timestamp prediction is made with a Dense output layer.

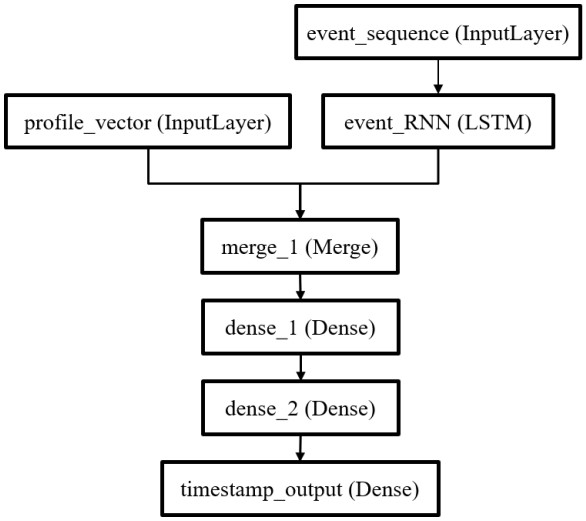


Figure 5.3: Architecture of the Neural Hawkes Process model

During training, the following initial parameters are used as baseline as proposed in the paper of Xiao et al. (2017):

- The state size of the event sequence RNN LSTM is set to 32;
- The event set contains of 6 time steps, namely a BOS event and the last 5 failure events occurred on the asset unit;
- Two hidden Dense layers, after merging the static profile vector and the event sequence RNN LSTM, of state size 64 with ReLu activation;
- A Dense output layer of state size 1 with a softplus activation layer;
- The Adam optimizer is used for learning the model weights during 60 epochs.

5.5.2 Activation behaviour

A different event type can cause an inhibiting or inertial effect on other event types. In order to allow the event sequence RNN LSTM to incorporate these effects, its activation needs to be passed through a non-linear activation function. A common non-linear activation function is the ReLu function $f(x) = max(x, 0)$, wherein x is the input to a neuron. However, it returns a 0 for negative x while the failure intensity rate must be positive at all times due to the probability that the failure cannot occur immediately after the previous failure. In the work of Mei & Eisner (2017), a softplus function $f(x) = s\ log(1 + exp(x/s))$ is used to keep the failure intensity strictly positive and approaches ReLu when x is far from 0. A disadvantage of this activation function is that x is expressed in units of time and the curvature strongly determines the effect of inhibition or inertia (Mei & Eisner, 2017). Therefore, a scaling parameter s could be used to control the curvature of $f(x)$. The effect of this parameter s is shown in Figure 5.4. However, investigating the performance improvement of optimizing this scaling parameter is left out of scope during the results section.

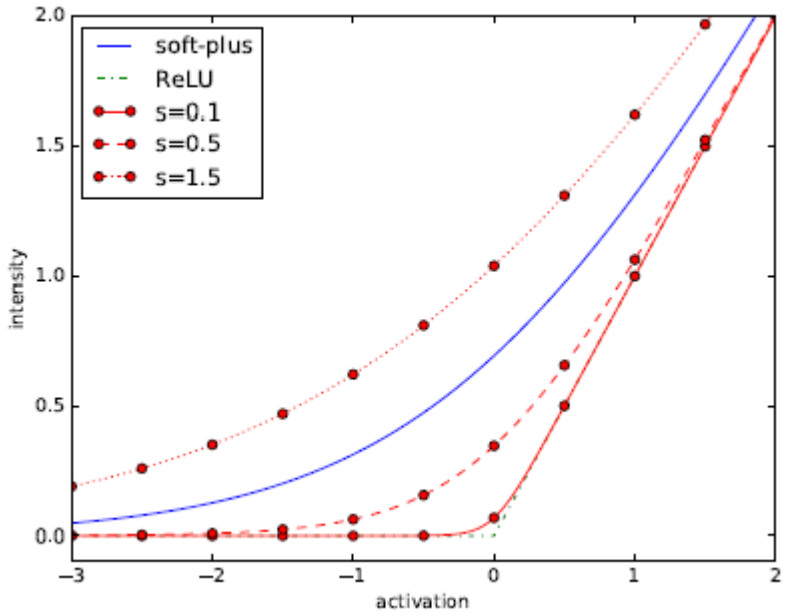


Figure 5.4: Activation behaviour comparison between ReLu and softplus (Mei & Eisner, 2017)

5.5.3 Loss function

The Neural Hawkes Process model is used to determine the economical end-of-life of asset units by predicting the timestamp of the next failure event. The timestamp is denoted in the number of years since the installation date of the asset unit. There is chosen to use the loss function of the policy Mean Squared Error (MSE) and not the Mean Absolute Error (MAE). This is motivated by the behaviour of MSE that larger deviations from the true label are more severely punished.

The formula of the Mean Squared Error is presented in Eq. 5.3, wherein n represents the size of data set, y the predicted value denoted in years since installation of the asset unit and $\tilde{y}$ the true value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{5.3}$$

5.5.4 Validation

The performance of the model Neural Hawkes Process Model will be validated based on the accuracy of predicting the timestamp of the next failure event. Three versions of the model will be evaluated to understand the predictive performance of the individual components of the Neural Hawkes Process model. These three versions are introduced subsequently:

- **Model_profile:** Profile attributes are modelled to predict the next failure event
- **Model_event:** Failure sequence events are modelled to predict the next failure event
- **Model_combined:** Both profile and failure sequence events are modelled to predict the next failure event

The first model version of the Neural Hawkes Process model, Model_profile, predicts the timestamp of the next failure event based on the profile features of the asset unit. The Model_event is the second version of Neural Hawkes Process model, that predicts the timestamp of the next failure event based on the event sequences. The Model_combined is the actual Neural Hawkes Process model architecture as shown in Figure 5.3. To determine the predictive performance of the individual components of the Neural Hawkes Process model, the performance of the Model_profile and Model_event are compared to that of Model_combined.

During the failure observation period used in this analysis, most asset units have experienced only one failure. The only failure event that has occurred on these asset units is used as the predictive label during training, which implies that the event sequence only consists of the BOS event. Therefore, it is assumed that the predictive performance of the failure event sequence in the Model_sequence and Model_combined is limited. To validate this assumption, the three versions of the Neural Hawkes Process model will be trained on two different training sets. The first training set includes all the asset units and the second training set includes all asset units that have experienced at least two failures. In this way the event sequence will have at least one failure during training of the model. Furthermore, the effect on the timestamp predictions when the failure event type is included in the tuples of the event sequences, is evaluated during the results section.

During training of all model versions, the profile vectors and their corresponding failure event sequence will be split into a training subset and a test subset, respectively 75% and 25% of the total dataset. A K-fold cross validation will be performed which cuts the training data in 10 subsets that are equally big. During every iteration, one subset is held out for testing and calculating the prediction error. The average prediction error of the 10 iterations is taken as the estimated prediction error. In addition to the loss function, i.e. Mean Squared Error, the Mean Absolute Error is calculated as an extra measurement of comparison between models.

Chapter 6

Results on failure event prediction

In the previous chapter, the three different versions of the Neural Hawkes Process model have been introduced. First, the feature importance analysis has determined the linear and non-linear correlation between the profile features of the asset unit and the predictive label. An initial performance comparison will be made by training the Model_combined with a full and a reduced profile vector. Subsequently, the performance of the Model_combined with different model parameters will be analyzed. After the final model architecture has been chosen, the three model versions of the Neural Hawkes Process model are trained and tested in two scenarios. Finally, a discussion on the obtained results is given. Performing these analysis has been performed on a Processor Intel(R) Core(TM) i7 CPU @ 2.80GHz with 16GB of RAM.

6.1 Architecture and parameter tuning

First, the insights gained from the feature importance analysis will be used to improve the predictive performance of the model. Next, several state sizes of the event_sequence component in the Neural Hawkes Process model are tested and evaluated. Subsequently, the learning behaviour of the Model_combined with different activation functions is visualized.

6.1.1 Static profile vector

The feature importance analysis has indicated a degree of correlation between the profile features of the asset unit and the predictive label, i.e. timestamp of the next failure event. First of all, the profile feature that represented the installation year of the asset unit is removed from the profile vector. When this feature is included in the profile vector, information is leaked from the model input to the predictive label which causes the model to be biased. If the installation year is included, the model will learn to predict all failures between the observation period of failures, which started in 2005 and ended in 2019. Since it is desired that the Neural Hawkes Process model predicts failures in the future, this feature is removed to increase the validity of the model.

Secondly, the nominal, internal and external diameter showed strong correlation with each other. Furthermore, some of the artificially binarized categorical features did not show a significant correlation to the predictive label. These are the features 'Function - Connection pipe', 'Soil type - Combination', 'Vegetation - High green' and 'Vegetation - Rural'.

The performance of the Model_profile that is trained with three different versions of the profile vector are evaluated in Table 6.1. Respectively, an original profile vector (1) that includes the installation year, a complete profile vector (2) that excludes the installation year and a reduced profile vector (3) that excludes the non-significant categorical features and only has one of the diameters, namely the nominal diameter.

	Metric	Original profile	Complete profile	Reduced profile
Train	MSE	2.99	14.67	14.88
	MAE	2.47	10.22	10.34
Validation	MSE	3.12	15.01	14.85
	MAE	2.56	10.49	10.39

Table 6.1: Performance comparison of different profile vectors

It can be concluded that the original profile vector, that includes the installation year, clearly outperforms the other profile vectors. However, earlier has been described that this can be explained that knowledge on the predictive label is leaked to the model's input, resulting in a biased model. When the complete and reduced profile vectors are compared, it can be concluded that the actual difference in performance is small, but the reduced profile vector has a higher accuracy. In addition, the reduced profile vector shows less over-fitting on the training set than in case of the complete profile vector. The reduced profile vector is used for further analysis.

6.1.2 Event sequence RNN LSTM

The failure event sequences that consist of tuples of the failure event types and the corresponding time of occurrences, are processed through a RNN LSTM. In the Model_combined, i.e Neural Hawkes Process model, the output of the RNN LSTM is merged with the static profile vector before a timestamp prediction of the next failure event is made. To evaluate this event sequence component, the performance of the Model_event with different state sizes of the RNN LSTM is presented in Table 6.2.

	Metric	State size 32	State size 16	State size 10
Train	MSE	17.48	17.65	17.61
	MAE	11.22	11.33	11.35
Validation	MSE	17.77	17.34	17.45
	MAE	11.39	11.08	11.26

Table 6.2: Performance comparison of different RNN LSTM state sizes

First of all, it can be concluded that the difference in state sizes did not result in a large performance improvement. However, the accuracy is best when the event sequences are processed with a RNN LSTM with state size 16. The predictive accuracy of the Model_event with state size 16 predicts the next failure event with a MSE of 17.34 years and a MAE of 11.08 years on the validation dataset. The state size parameter of 16 for the RNN LSTM is used for further analysis.

6.1.3 Learning behaviour of the activation functions

In Chapter 5, it has been described that the deterioration process of water pipes can be best modelled with a softplus activation function. In the work of Mei & Eisner (2017), the softplus activation function allows for inhibiting and inertial effects of failure events. This effect has been analyzed in Figure 6.1, wherein the learning behaviour of the Model_combined with different activation functions is visualized. In this model version, the Neural Hawkes Process model consists of three layers with a specified activation function, namely the two hidden dense layers and the final dense prediction layer. When referred to the model variation 'ReLu/Softplus', the two hidden layers consists of a ReLu activation function while the prediction layer uses a softplus activation function.

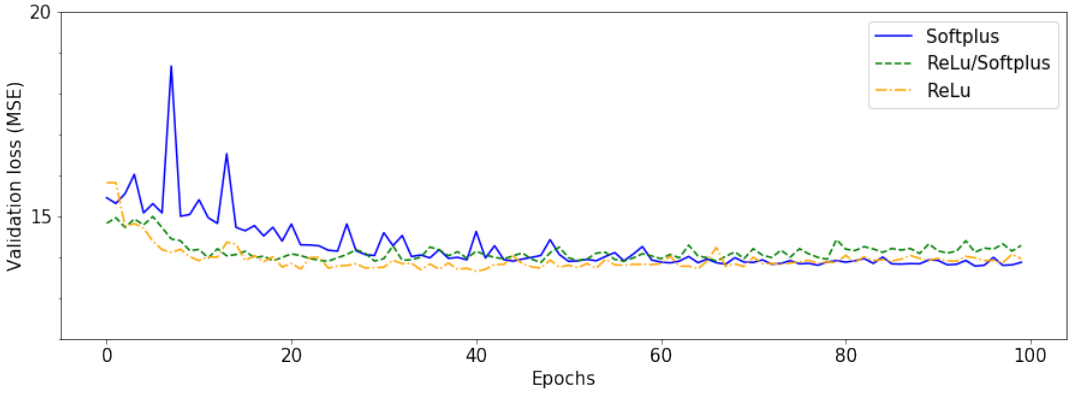


Figure 6.1: Learning behaviour of the activation functions

When compared, the softplus activation function learns slower but eventually arrives at the same threshold as the ReLu and the combination of ReLu and Softplus. Around 100 epochs, the validation loss of the MSE starts to stabilize and has reached its minimum loss. The performance results of these three model variations are presented in Table 6.3.

	Metric	**Softplus**	**ReLu/Softplus**	**ReLU**
Train	MSE	13.55	12.93	13.16
	MAE	8.61	8.24	8.37
Validation	MSE	13.79	13.77	13.69
	MAE	8.85	8.87	8.82

Table 6.3: Performance comparison of different activation functions

Overall, it can be concluded that there is no clear improvement in performance when a different activation function is used. On average, the MSE per timestamp prediction of the next failure event is 13 years and has a mean error of 8.8 years. Furthermore, it seems that the activation functions ReLu and the combination ReLu/Softplus are slightly over-fitting on the training dataset. In case of the softplus activation function, the accuracy on the validation dataset shows signs of over-fitting. Taken this into consideration and that the softplus function is used in the work of Mei & Eisner (2017), the activation function softplus is used for further analysis.

6.2 Performance results

In the previous section we have learned the model parameters that give the best model performance. In this section we will analyse and evaluate the performance of the Model_combined and its individual components, i.e. Model_profile and Model_event. These model variations of the Neural Hawkes Process model are learned in two different situations. First, the model weights are learned based on a training dataset that included all asset units in scope. In the second situation, the training dataset is reduced so that it only includes asset units that have experienced at least two failures. It is believed that when the event sequence of the asset unit has at least one failure event besides the BOS event, the performance of the model improves. At last, the long-term effects of the failure event types are tested and evaluated.

6.2.1 Performance of the model in situation 1

In situation 1, the performance of the Neural Hawkes Process model is tested when the model is trained on all instances of asset units. In Figure 5.1, the distribution of the number of failure events in the event sequences was presented. From this figure, it has been concluded that most event sequences only consist of one actual failure event. This results in the fact that for most training samples, the only failure in the event sequence is used as predictive label. For this reason, it is assumed that the predictive performance of the event_sequence component in the Model_combined is limited in this situation.

If the event sequence of the asset unit only consists of the BOS event, the Neural Hawkes Process model learns the failure intensity function of an asset unit that is unconditioned on the historic failures of that asset unit. As illustrated in the Bathtub curve principle in Figure 1.1, the probability of failures increases when a water pipe has reached its end of life wear-out phase. It is most likely that the first failures of the asset units will occur during this phase. However, environmental and operational effects can cause the asset unit to fail during its normal life time. For this reason, the many structural, environmental and operational features of the asset unit profile will result in the fact of a high variation in the prediction of the timestamp of the first failure event. The performance of the model versions of the Neural Hawkes Process model is shown in Table 6.4.

	Metric	**Model_profile**	**Model_event**	**Model_combined**
Train	MSE	15.03	17.61	13.55
	MAE	10.48	11.29	8.67
Test	MSE	15.18	17.63	13.99
	MAE	10.49	11.34	8.79

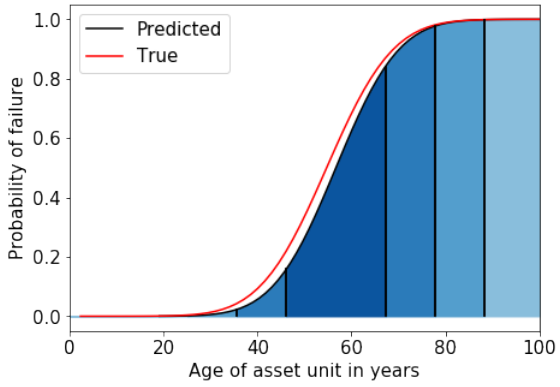
Table 6.4: Performance of the Neural Hawkes Process model in situation 1

First of all, it can be concluded that the Model_combined, i.e. the Neural Hawkes Process model, clearly outperforms its individual components Model_profile and Model_event. The MSE of the Model_combined is 14 years with an average absolute error of 8.8 years from the actual timestamp of the true label. The performance of the individual components of the Neural Hawkes Process model, respectively the Model_profile and Model_event, can be evaluated from the results in Table 6.4. In case of the Model_profile, the timestamp of the next failure is only predicted based on the base failure intensity rate determined by the profile features of the asset unit. The MSE of the Model_profile is 15.2 years on the training set and has a corresponding mean absolute error of 10.5 years from the true label. The performance of the Model_event represents the predictive ability of the long-term effects of past failures on the failure intensity function. The accuracy of the the Model_event is the lowest compared to the other two versions, which has a MSE of 17.6 years and an absolute error of 11.3 year from the true label. The low performance of the Model_event can be explained due to the fact that most failure event sequences for training only consist of a BOS event.

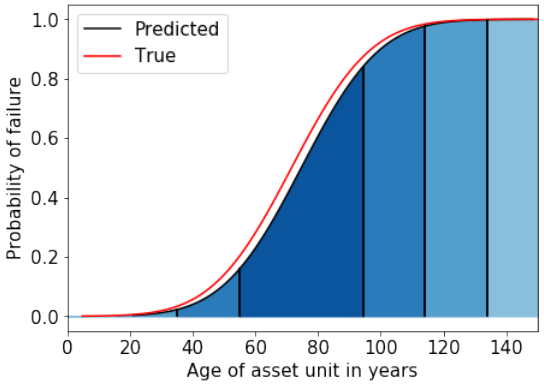
With the performance of the three model versions discussed, the overall conclusion of the performance of the Neural Hawkes Process model in situation 1 can be drawn. First, the combination of the base failure intensity rate and the long-term effects of past failures to learn the failure intensity function of water pipes, i.e. asset units, has led to the highest accuracy of predicting the timestamp of the next failure event. Secondly, the performance improvement of the Model_combined compared to Model_profile indicates that the Neural Hawkes Process model is able to distinguish whether it is predicting a first failure event or a consecutive failure. As explained in Chapter 2, Model_profile learns the failure intensity function based on Equation 2.1 while Model_combined learns a failure intensity function that is conditioned on the history of past failures of that specific asset unit (Equation 2.2).

The failure intensity function learned with the Model_combined is used to draw the probability of failure per material type, visualized in Figure 6.2. In these figures, the colors represent the confidence intervals σ, 2σ, etc., respectively from dark to light blue. The red line represents the actual distribution as observed and the black line represent the distribution predicted with the Neural Hawkes Process model. The Neural Hawkes Process model is able to distinguish the distributions that differs per material type. For example, on average an asset unit of material type GIJ experience its first failure at the age of 80 years. When compared to the material types AC and PVC, the average age is 59 and 43 years. It can be concluded that the probability of failure predicted by the Neural Hawkes Process model is similar to the actual distribution of failures of the observation period.

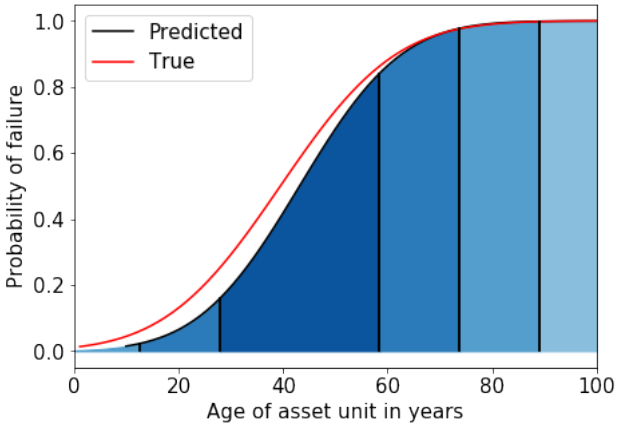
It is important to note that these probabilities of failure could be biased due to the fact of the small time range of the observation period. Nowadays, water pipes of the material type GIJ are no longer installed in the distribution network. The currently operational asset units of material GIJ have been installed since 1917, while PVC and AC became popular around 1960. Due to the fact that the observation period of failures ranges from 2005 - 2019, there is relatively more failure events available of the normal life phase of the material types AC and PVC. In case of the material type GIJ, most failure events occur during the asset units end of life wear-out.



(a) Material type AC



(b) Material type GIJ



(c) Material type PVC

Figure 6.2: Distribution of probability of failure in situation 1

6.2.2 Performance of the model in situation 2

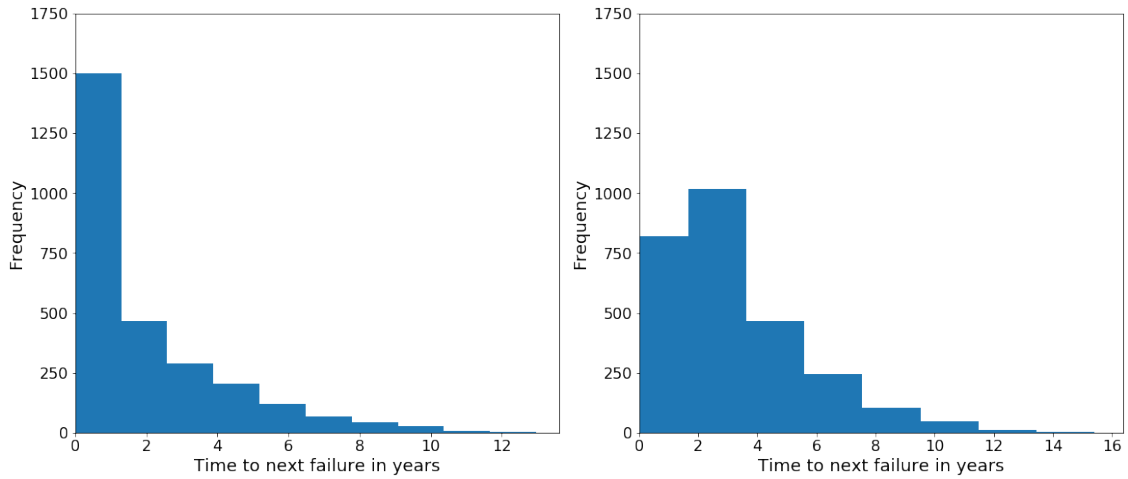
In situation 2, the performance of the Neural Hawkes Process model is tested when trained on all instances of asset units that have experienced at least two failure events. The motivation of testing the Neural Hawkes Process model in this situation, is that at least one failure event is included in the event sequence of the training dataset. It is assumed that the performance of Model_event increased when compared to the first situation due to the fact that most event sequences only consisted of the BOS event. Contrary to the first situation, there can be said that the failure intensity function of all asset units is now conditioned on past failures of the asset unit, which is incorporated in the event sequence. In total 10203 asset units have experienced at least one failure, of which only 2773 asset units are in scope in situation 2. The performance of the Neural Hawkes Process model in situation 2 is presented in Table 6.5.

	Metric	Model_profile	Model_event	Model_combined
Train	MSE	11.78	2.23	2.15
	MAE	8.12	1.79	1.59
Test	MSE	12.55	2.34	2.17
	MAE	8.60	1.78	1.64

Table 6.5: Performance of the Neural Hawkes Process model in situation 2

The performance of all three versions of the Neural Hawkes Process model represent its ability to accurately predict the time to next failure that is conditioned on the previous occurred failures. The performance of the Model_combined has improved significantly to a MSE of 2.17 years and an absolute error of 1.64 years from the true label. Compared to situation 1, the MSE and MAE were 14 and 8.8 years, respectively.

The performance of the Model_event in situation 2 is almost similar to that of the combined model. It can be concluded that when the failure intensity function is conditioned on the past failures, i.e. at least one failure event is included in the event sequence, the Neural Hawkes Process model is able to make accurate predictions. The performance of the Model_profile is slightly better than its performance in situation 1, but since the number of asset units in scope in situation 1 is less than in situation 2, the variance of the predictive label is lower. Furthermore, the Model_profile is still unable to distinguish whether to predict the first failure or a consecutive



(a) Time to next failure as observed

(b) Time to next failure as predicted

Figure 6.3: Predicted time to next failure of the Neural Hawkes Process model

failure on the asset unit, based on its performance of a MSE 12.6 years and an absolute error of 8.6 years from the true label.

The Neural Hawkes Process model in situation 2 is able to predict the timestamp of the next failure based on the profile features and the past failures incorporated in the event sequence. In Figure 6.3, the time to next failure in years as observed, i.e. true time to next failure, and the time to next failure predicted by the Neural Hawkes Process model is shown. It can be seen that both distributions are different from each other. The actual time to next failure, shown in Figure 6.3a, indicates that most consecutive failures occur within 2 years with a maximum of 12 years. Figure 6.3b shows that the Neural Hawkes Process model predicts that most consecutive failures will occur between 2 and 4 years since the last failure, with a maximum of 16 years.

6.2.3 Exciting and inhibiting effects of failure types

At last, the ability of the Neural Hawkes Process model to incorporate exciting and inhibiting effects of the failure event types on the failure intensity function, is researched. All failure events consist of a timestamp and a specific event type, in total seven failure types are in scope as described in Chapter 3. Based on the literature study and the work of Mei & Eisner (2017), modelling the failure intensity function with event types will enable the Neural Hawkes Process model to learn the exciting or inhibiting effects of past failures on the probability of future failures. To determine the existence of these effects, the event type is excluded from the event sequence and the performance of the Neural Hawkes Process model, i.e. Model_combined, is evaluated in both situation 1 and 2.

The effect of failure types on the performance of the Neural Hawkes Process model is presented in Table 6.6. Referring to situation 1, the effect of including the event types shows different results when the MSE and MAE are compared. First, the accuracy of the Model_combined, based on the MSE, has actually decreased when the failure event types are included. However, the mean absolute error of the timestamp prediction improves both on the train and test set. In case of situation 2, the performance accuracy has improved for both the MSE and MAE on the train and test set.

Overall, it can be concluded that the performance of the event sequence is more leveraged in situation 2, wherein including the event types results in a better MSE and MAE score on both the train and test set. In this situation, the performance improvement, when the exciting and inhibiting effects of failure events are incorporated, of modelling the failure intensity function is clearly seen. In situation 2, when the failure intensity function is conditioned on past failures, the performance on the model on the test set is 2.17 and 1.64 years, respectively the accuracy of the MSE and MAE. In case of situation 1, most event sequences have no actual failures during training of the model. However, still the average error of the timestamp prediction is improved when the failure events are included, respectively 8.79 and 8.96 years.

		Situation 1		**Situation 2**	
		Including	**Excluding**	**Including**	**Excluding**
Train	MSE	13.55	13.76	2.15	2.17
	MAE	8.67	8.89	1.59	1.69
Test	MSE	13.99	13.82	2.17	2.26
	MAE	8.79	8.96	1.64	1.78

Table 6.6: Performance comparison on exciting and inhibiting effects

6.3 Discussion

First, the model structure of jointly combining the profile vector with an event sequence of past failures, has proven to be effective for predicting the timestamp of next failure event. The performance of the Neural Hawkes Process model in situation 1 shows that the Model_combined is able to distinguish the first failure and a consecutive failure of an asset unit. The results of situation 2 has proven that when the number of failures in the event sequence increases, the predictive performance of the Model_event and therefore also the Model_combined increases and outperforms the Model_profile. The activation function softplus showed the best results in the performance of the Neural Hawkes Process model, which is in compliance with the work of Xiao et al. (2017). The exciting and inhibiting effects of past failures, specifically the event types, are seen in both situation 1 and situation 2.

In general, the applicability of the Neural Hawkes Process model to predict future water pipe failures depends on the quality and number of failures in the observation data available at Vitens. The failure intensity function of water pipes has been learned from historical failure records that started to be of quality from around 2010. Since the Neural Hawkes Process model makes no parametric assumptions on the underlying behaviour of water pipe failures, there can be argued whether the model can learn the failure intensity function based on small observation period when compared to the total life cycle of an asset unit. This limited observation period makes the dataset left censored, meaning that failures on assets units may have occurred before the start of the observation period.

Finally, the matter of extrapolation of the failure event timestamps must be addressed. Failures recorded during the observation period have occurred on asset units with different ages. The number of failure patterns observed during infant mortality, normal life and end of life wear-out determine if the Neural Hawkes Process model is able to accurately predict failure events. The failure intensity function learned by the model, only reflects the data the model has seen during training, which limits the extrapolating abilities. A predefined parametric hazard function could be used to extrapolate outside of the learned training domain.

Chapter 7

Implementation

This last chapter is devoted to implement the model insights gained into the maintenance replacement program of Vitens. The concept of early warning points has been described in Chapter 1. These value points are an indicator that a water pipe, i.e. also referred to as asset unit, has reached its economical end-of-life, meaning that it is economically more profitable to perform a replacement than a repairment. The developed and trained failure prediction model will improve the predictive maintenance capabilities of Vitens to be able to more accurately predict when these certain value points occur. First, the definition of the economical end-of-life of a water pipe is briefly reintroduced. Next, the predictive ability of the Neural Hawkes Process model is used to identify value points in the deterioration cycle of asset units. At last, an advice is given to implement these model insights into the maintenance strategy of Vitens.

7.1 Economical end-of-life of water pipes

An asset unit that has entered his end-of-life wear-out will be prone to an increased risk of failure. Before an asset unit will have reached the end of its technical life, several failures could have already occurred on this specific asset unit. When the economical end-of-life must be determined, not only the probability of failure is considered, but also the cost of replacement and the execution of the maintenance activity are relevant. Since these maintenance costs have not been considered during this thesis project, the asset management department of Vitens has defined the economical end-of-life as follows:

" When a next failure event is predicted two years after the last occurred failure, within a range of 400m water pipe length, the asset unit has reached its economical end-of-life. "

The repairment costs of two consecutive failures on a specific asset unit per 400m, within a time period of 2 years, exceeds the costs of installing a new water pipe. This is called a value point in the life cycle of an asset unit, which functions as an early warning point for the asset management department of Vitens. The failure intensity model is trained on the training set in situation 2 in order to learn the dependency of the failure event-sequence, which it was unable to learn in situation 1.

7.2 Value points

To be able to detect a value point, an previous failure must have been recorded on that asset unit during the observation period, which is used as a reference point for the two year interval. Therefore, the Model_combined in situation 2 is used to learn the failure intensity function to predict the timestamp of the next failure. Only the predictions for the assets units with a maximum length of 400m and with a last failure occurred since 2018 are included in this prediction.

The timestamp predictions are used to determine the time to next failure for the asset units in scope, presented in Figure 7.1.

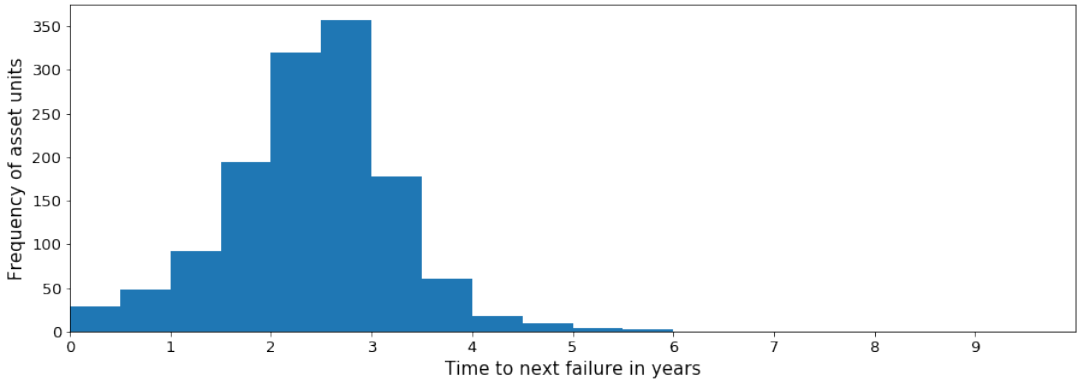


Figure 7.1: Distribution of predicted time to next failure

Based on the economical end-of-life definition, in total 326 assets units have been predicted to experience a next failure within 2 years. In total 204 asset units of material type AC, 121 asset units of material type PVC and 37 asset units of material type GIJ.

7.3 Maintenance strategy

To keep asset units operational beyond their economical end-of-life is not beneficial due to the fact of increased maintenance cost in comparison to the replacement cost. The insights gained from the Neural Hawkes Process model into the deterioration process of asset units can be used to prioritize water pipe replacements in the water distribution network of Vitens. The identification of value points in the life cycle of these asset units will enable Vitens to reduce maintenance costs by preventing unnecessary early replacements of water pipes. The Neural Hawkes Process model is used to predict the time to next failure for the asset units that already have failed in the past. The asset units IDs and their corresponding individual water pipe IDs of the 326 value points identified, will be distributed to the asset management department of Vitens to incorporate these insights in their maintenance planning.

It is recommended to monitor the 326 identified asset units during the next two years in order to evaluate the accuracy of predictions before taking any maintenance related actions based on these insights. The Neural Hawkes Process model is an additional tool to the current predictive maintenance capabilities at Vitens. The ability of the Neural Hawkes Process model to perform a timestamp prediction on an asset unit level makes this predictive model unique. However, when compared to the current Pipe Replacement Potential model within Vitens, it is not suitable to predict the growth rate of failures for larger homogeneous groups on the long-term. The availability of both prediction models will behave in a complementary manner and enhances the predictive maintenance capabilities of Vitens.

Furthermore, continuing the monitoring of failure events within the distribution network and accurately assigning these failures to the corresponding water pipe IDs is key for the predictive maintenance capabilities in general. Extending the observation period of failures and improving the data quality of these failure records will directly contribute to an increased performance of Neural Hawkes Process model

Chapter 8

Conclusion and Discussion

In this final chapter there will be concluded on the master thesis project performed at the water distribution company Vitens. In the first section, the sub questions and eventually the main research question will be answered. Subsequently, there will be elaborated on the limitations and advised on future research regarding this research project. To end this chapter, a set of recommendations are given to Vitens based on the results of this research.

8.1 Conclusion

In this work, we have shown that the recent embodiment of a classical survival analysis, the Hawkes process, into a neural network structure is able to model the failure intensity function of water pipes. We have proven that the failure event sequence of asset units incorporate valuable information on the deterioration life cycle as described in the Bathtub Curve principle. We applied an adapted version of the model proposed in the work of Xiao et al. (2017), called the Neural Hawkes Process model. The base failure intensity rate is learned based on a static profile of structural, environmental and operational water pipe attributes. The long-term dependencies of past failures on the failure intensity function is captured in a failure event sequence and feed into a RNN LSTM. The failure intensity model is considered expressive enough to learn the underlying behaviour of water pipes and therefore requires no model assumptions. This non-parametric model structure is considered a strong advantage and increases the applicability in other predictive maintenance cases.

In the first part of this thesis, we have performed an extensive literature research on the current state-of-art on modelling water pipe failures in order to answer the first three sub questions of this thesis. A case study analysis has been performed to determine the most promising water pipe failure predictors. In our predictive model we have included most of these important predictors and introduced an innovative way to incorporate the effect of water hammer. The literature research has shown that apart from the time-sequential effects of failure events, the type of failure events can have an exciting or inhibiting effect on the failure intensity function.

Two major data transformation steps were needed in order to model the failure intensity function of water pipes. First, individual water pipes that share the same characteristics and are physically connected, are grouped into asset units due to the fact that they share a similar deterioration life cycle. Subsequently, an geospatial algorithm has been developed to assign failure events to these created asset units in order to create a failure event history on an asset unit level. For every asset unit, a failure event sequence is created that consists of an arbitrary length of tuples. These tuples contain the failure event type and a corresponding timestamp of occurrence, denoted in the number of years since installation year.

The performance of the Neural Hawkes Process model is evaluated in two situations, respectively when the model is trained on all asset units and on those asset units that have experiences at least two failures. In the first situation the performance of the model represents the ability to

predict the first failure event on the asset units. On average, the Neural Hawkes Process model is able to predict the timestamp of the next failure event with an error of 8.8 years. The predictive performance is compared to its individual model components, i.e. Model_combined and Model_event. Based on this comparison, it can be concluded that the Neural Hawkes Process model is able to distinguish when to predict a first failure event or a consecutive failure. This proves that combining the base intensity rate captured in the profile vector with the failure event sequences leads to a better predictive performance.

In the second situation, only the asset units that have experienced at least two failures are used to learn the failure intensity function of asset units. It has been concluded that the predictive performance of the Neural Hawkes Process model increases with the size of the failure event sequence. The Neural Hawkes Process model is able to predict the time to next failure with an average error of 1.64 years. Furthermore, the Neural Hawkes Process model is able to learn the inhibiting or exciting effects of failure event types on the failure intensity function. The performance of including these failure event types have resulted to a more accurate prediction of the next timestamp in both situation 1 and 2.

A definition of the economical end-of-life of an asset unit has been formulated by the asset management department of Vitens. The maintenance costs exceed the cost of replacement when a consecutive failure occurs within a two year time period, also referred to as a value point. The Neural Hawkes Process model has been used to identify 326 value points for asset units that will experience a consecutive failure within two years.

Now that the sub questions of this research have been answered, the main research question of this thesis project can be answered:

How can a machine learning approach determine the economical end-of-life of water pipes in a Dutch water distribution system?

The Neural Hawkes Process model is able to predict the economical end-of-life and identify value points in the deterioration process of the asset units. The Neural Hawkes Process model is expressive enough to learn the intensity function without any prior domain knowledge. When compared to the current predictive capabilities of Vitens, the Neural Hawkes Process model is able to make accurate predictions on an individual water pipe level instead of a homogeneous group; and is able to predict the time to next failure for these individual water pipes. This exceeds the current predictive capabilities at Vitens of scoring individual water pipes on a relative scale of structural condition.

8.2 Discussion

This work has served as a first attempt to assess the applicability of modelling the failure intensity function of water pipes with a neural version of the Hawkes process. In this section there will be elaborated on the limitations of this research. To address the next steps to be taken in modelling the failure intensity function of water pipes, some directions for future research will be motivated.

8.2.1 Limitations

The first limitation of this research is the small observation period of failures that is used to train the Neural Hawkes Process model. The life time of an average water pipe varies per material pipe, but can range between 50-150 years. Although failures have occurred on all types of water pipes of different ages, the Neural Hawkes Process model could be biased because it can only learn from the data points in the training dataset. Extending the observation period would enrich the training dataset to improve the accuracy of the Neural Hawkes Process model.

A second limitation of this research is the availability of the installation dates of the assets units within the distribution network. A significant part of the failures have been linked to the asset units that contain missing data on their year of installation. For predictive maintenance cases and survival analysis in general, the age of the asset units is required to be able to compare

the deterioration of assets units in different life cycle phases. Although in total 92% of the data points have been retrieved based on a publicly available dataset of the Dutch government, this does not guarantee the highest data quality for training.

At last, the effect of water hammer is incorporated in the profile vector of the asset units by calculating the celestial distance between an asset unit and the nearest production location. A more accurate approximation of this effect is made when the actual distance through the network is calculated.

8.2.2 Future research

The original failure intensity model, as described in the work of Xiao et al. (2017), is able to learn a time-varying base intensity rate. In our model, the base failure intensity rate of asset units is learned from a static profile vector. A possibility for future research is to incorporate time-varying model features, i.e. temperature or climate, in a time-series RNN model to increase the expressiveness of our model and allow for a dynamic base failure rate.

We have assumed that our failure intensity model is expressive enough to learn the underlying failure behaviour of water pipe failures. However, implementing model assumptions in our current model structure could allow for extrapolation of failure timestamp predictions.

Our failure intensity model is able to predict the next failure event. Further research should focus on extending our model structure to be able to predict multiple consecutive failures on an asset unit level. At last, the maintenance costs of failure events have been left out of scope for predicting the economical end-of-life of asset units. Incorporating the financial aspect of water pipe failures would improve the impact of our model for the maintenance strategy of Vitens.

8.2.3 Recommendations to Vitens

This work has shown that the Neural Hawkes Process model is a promising technique to model the failure intensity function of water pipes. The development of our model compliments the current predictive maintenance capabilities of Vitens. Not only is our model able to predict the next failure event, it is expressive enough to make predictions on a detailed asset unit level. In Chapter 7, our Neural Hawkes Process model has identified 326 asset units that have reached their economical end-of-life. However, several other general recommendations can be made regarding to the predictive maintenance capabilities at Vitens:

1. The current process of failure event registration needs to be changed in order to enhance the predictive maintenance capabilities of Vitens. Standardization of maintenance work orders and directly linking failure events to individual water pipe IDs will contribute to the data quality for predictive modelling.

2. The data quality of the distribution network and the historical failure records could be improved and maintained at a more easily available platform in order to enhance the productivity of data scientist at Vitens.

3. Data transformation steps, such as the creation of asset units and connecting failures to these asset units, should be readily available for other predictive maintenance projects within Vitens. In this manner, less time is spend on data preparation and future predictive models will be developed faster.

References

Abbasi, T., Lim, K. H. & San Yam, K. (2019). Predictive maintenance of oil and gas equipment using recurrent neural network. In *Iop conference series: Materials science and engineering* (Vol. 495, p. 012067).

Ahn, J., Lee, S., Lee, G. & Koo, J. (2005). Predicting water pipe breaks using neural network. *Water Science and Technology: Water Supply*, *5*(3-4), 159–172.

Al-Barqawi, H. & Zayed, T. (2006). Condition rating model for underground infrastructure sustainable water mains. *Journal of Performance of Constructed Facilities*, *20*(2), 126–135.

Asnaashari, A., McBean, E. A., Gharabaghi, B. & Tutt, D. (2013). Forecasting watermain failure using artificial neural network modelling. *Canadian Water Resources Journal*, *38*(1), 24–33.

Berardi, L., Giustolisi, O., Kapelan, Z. & Savic, D. (2008). Development of pipe deterioration models for water distribution systems using epr. *Journal of Hydroinformatics*, *10*(2), 113–126.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Donev, V. & Hoffmann, M. (2019). Condition prediction and estimation of service life in the presence of data censoring and dependent competing risks. *International Journal of Pavement Engineering*, *20*(3), 313–331.

Ertekin, Ş., Rudin, C., McCormick, T. H. et al. (2015). Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, *9*(1), 122–144.

Farmani, R., Kakoudakis, K., Behzadian Moghadam, K. & Butler, D. (2017). Pipe failure prediction in water distribution systems considering static and dynamic factors. *Procedia Engineering*, *186*, 117–126.

Feng, W., Guan, N., Li, Y., Zhang, X. & Luo, Z. (2017, 05). Audio visual speech recognition with multimodal recurrent neural networks. In (p. 681-688). doi: 10.1109/IJCNN.2017.7965918

Fu, R., Zhang, Z. & Li, L. (2016). Using lstm and gru neural network methods for traffic flow prediction. In *2016 31st youth academic annual conference of chinese association of automation (yac)* (pp. 324–328).

Gers, F. A. & Schmidhuber, J. (2000). Recurrent nets that time and count. In *Proceedings of the ieee-inns-enns international joint conference on neural networks. ijcnn 2000. neural computing: New challenges and perspectives for the new millennium* (Vol. 3, pp. 189–194).

Hawkes, A. G. (2017). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, *33*(3), 438–443.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jafar, R., Shahrour, I. & Juran, I. (2010). Application of artificial neural networks (ann) to model the failure of urban water mains. *Mathematical and Computer Modelling*, *51*(9-10), 1170–1180.

Kakoudakis, K., Behzadian, K., Farmani, R. & Butler, D. (2017). Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with k-means clustering. *Urban Water Journal*, *14*(7), 737–742.

Karimian, S. F. (2015). *Failure rate prediction models of water distribution networks* (Unpublished doctoral dissertation). Concordia University.

Kawakita, K. (1986). The relation between the water temperature and the number of bursts in water mains. *J. JWWA*, *55*(5), 14–24.

Kimutai, E., Betrie, G., Brander, R., Sadiq, R. & Tesfamariam, S. (2015). Comparison of statistical models for predicting pipe failures: Illustrative example with the city of calgary water main failure. *Journal of Pipeline Systems Engineering and Practice*, *6*(4), 04015005.

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

Kumar, A., Rizvi, S. A. A., Brooks, B., Vanderveld, R. A., Wilson, K. H., Kenney, C., ... others (2018). Using machine learning to assess the risk of and prevent water main breaks. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 472–480).

Kutyłowska, M. (2015). Neural network approach for failure rate prediction. *Engineering Failure Analysis*, *47*, 41–48.

Li, Z. & Wang, Y. (2018). Domain knowledge in predictive maintenance for water pipe failures. In *Human and machine learning* (pp. 437–457). Springer.

Liang, B., Li, Z., Wang, Y. & Chen, F. (2018). Long-term rnn: Predicting hazard function for proactive maintenance of water mains. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 1687–1690).

Martínez-Codina, Á., Castillo, M., González-Zeas, D. & Garrote, L. (2016). Pressure as a predictor of occurrence of pipe breaks in water distribution networks. *Urban Water Journal*, *13*(7), 676–686.

Mei, H. & Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in neural information processing systems* (pp. 6754–6764).

Nishiyama, M. & Filion, Y. (2014). Forecasting breaks in cast iron water mains in the city of kingston with an artificial neural network model. *Canadian Journal of Civil Engineering*, *41*(10), 918–923.

NLextract, P. (2019). *Basisregistratie adressen en gebouwen (bag) (uigebreid).* Retrieved 14-12-2019, from `https://nlextract.nl/downloads/`

O'Day, D. K. (1982). Organizing and analyzing leak and break data for making main replacement decisions. *Journal-American Water Works Association*, *74*(11), 588–594.

Pelletier, G., Mailhot, A. & Villeneuve, J.-P. (2003). Modeling water pipe breaks—three case studies. *Journal of water resources planning and management*, *129*(2), 115–123.

Puust, R., Kapelan, Z., Savic, D. & Koppel, T. (2010). A review of methods for leakage management in pipe networks. *Urban Water Journal*, *7*(1), 25–45.

Rajani, B. & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: physically based models. *Urban water*, *3*(3), 151–164.

Rasmussen, J. G. (2018). Lecture notes: Temporal point processes and the conditional intensity function. *arXiv preprint arXiv:1806.00221*.

Reed, C., Robinson, A. & Smart, D. (2007). *Potential techniques for the assessment of joints in water distribution pipelines.* American Water Works Research Foundation.

Rodríquez, G. (2020). *The hazard and survival functions.* Retrieved 18-01-2020, from `https://data.princeton.edu/wws509/notes/c7s1`

Rogers, P. D. & Grigg, N. S. (2008). Failure assessment model to prioritize pipe replacement in water utility asset management. In *Water distribution systems analysis symposium 2006* (pp. 1–17).

Røstum, J. (2000). Statistical modelling of pipe failures in water networks.

Schober, P. & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesthesia and analgesia*, *127*(3), 792.

Shirzad, A., Tabesh, M. & Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE Journal of Civil Engineering*, *18*(4), 941–948.

Snider, B., McBean, E. A. et al. (2018). Improving time to failure predictions for water distribution systems using extreme gradient boosting algorithm. In *Wdsa/ccwi joint conference proceedings* (Vol. 1).

St. Clair, A. M. & Sinha, S. (2012). State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! *Urban Water Journal*, *9*(2), 85–112.

St. Clair, A. M. & Sinha, S. (2014). Development of a standard data structure for predicting the remaining physical life and consequence of failure of water pipes. *Journal of Performance of Constructed Facilities*, *28*(1), 191–203.

Tabesh, M. & Delavar, M. (2003). Application of integrated gis and hydraulic models for unaccounted for water studies in water distribution systems. In *Maksimovic, butler and memon (eds) proceedings of the international conference on advances in water supply management* (pp. 129–135).

Tran, D., Ng, A., Perera, B., Burn, S. & Davis, P. (2006). Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. *Urban Water Journal*, *3*(3), 175–184.

Verheugd, J. T. (2020). *Predicting water pipe failures: A neural hawkes process approach.*

Vitens. (2017). *Operational facts vitens.* Retrieved 03-04-2019, from `https://www.vitens.com/organisatie/kengetallen`

Wilson, D., Filion, Y. & Moore, I. (2017). State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal*, *14*(2), 173–184.

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W. & Tscheikner-Gratl, F. (2018). Pipe failure modelling for water distribution networks using boosted decision trees. *Structure and Infrastructure Engineering*, *14*(10), 1402–1411.

Xiao, S., Yan, J., Yang, X., Zha, H. & Chu, S. M. (2017). Modeling the intensity function of point process via recurrent neural networks. In *Thirty-first aaai conference on artificial intelligence.*

Yan, J., Wang, Y., Zhou, K., Huang, J., Tian, C., Zha, H. & Dong, W. (2013). Towards effective prioritizing water pipe replacement and rehabilitation. In *Twenty-third international joint conference on artificial intelligence.*

Zhang, C., Wu, H., Bie, R., Mehmood, R. & Kos, A. (2018). Dynamic modeling of failure events in preventative pipe maintenance. *IEEE Access*, *6*, 12539–12550.

Appendix

A Water pipe materials

As mentioned in the introduction, the water distribution system of Vitens is mainly constructed of four material types. Each pipe material is characteristically designed for certain applications and provides its own advantages. An overview of these water pipe materials is given subsequently.

Cast-Iron Water Pipe

Around the 1900s, gray cast-iron (CI) pipes were introduced to the market and were produced by casting molten iron in vertical sand molds. The disadvantage of CI water pipes was the lack of a uniform thickness because of the misalignment of the central core mold. Between the 1930s and 1960s a new type of CI pipes with better performance, called spun gray iron pipes, were widely used during this period. Today CI is no longer used as preferred material type; however, many water companies still have a very high percentage of this material type in their distribution systems.

The main advantage of this material type was its low cost. The disadvantages of CI are that it is structurally weak and subject to internal corrosion, which on its turn can lead to water quality problems. Moreover, age is not a good indicator of its structural condition because during that period different technologies were used for producing CI pipes that has resulted in different wall thicknesses (Reed et al., 2007).

Polyethylene Pipes

In the 1950s, polyethylene (PE) pipes were introduced and more modern material became available from 1990 to 1990. One of these more modern material types was medium-density PE, which was widely used as water pipe material during that period. PE pipes are primarily used for locations where the ground is subjected to movement and subsidence. Some of the main advantages of using PE pipes are their corrosion resistance; flexibility; ability to absorb impact loads; vibration, and ground movement; and that they are lightweight. The disadvantage of using PE pipes is that they are susceptible to permeation or degradation by certain organic contaminants; they are dependent on stable support from the ground to resist deformation; and they are susceptible to fatigue failure (St. Clair & Sinha, 2014).

Polyvinyl Chloride Water Pipes

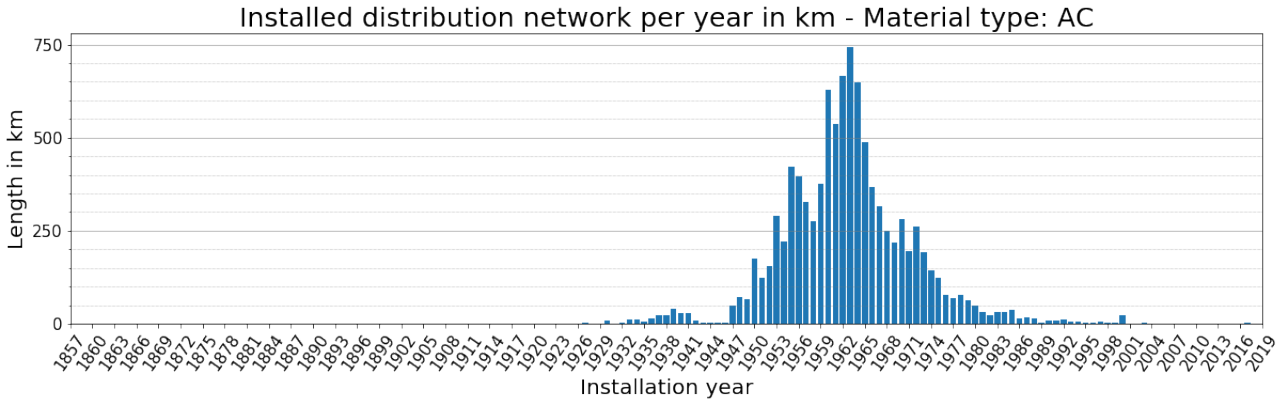
In the 1950s water distributions companies started to use PVC pipes. The application of PVC is primarily used in most low-stress environments.The advantages of using PVC pipes are their corrosion resistance, high resistance to chemical attack, and the fact that they are lightweight. The disadvantages of using PVC pipes include the following: they are susceptible to permeation or degradation by certain organic contaminants; they are susceptible to point loading and impact damage; certain grades are susceptible to UV degradation; they run the risk of fracture in contaminated land; they are dependent on stable support from the ground to resist deformation; they are susceptible to fatigue failure and buckling; they require care in handling; and they run the

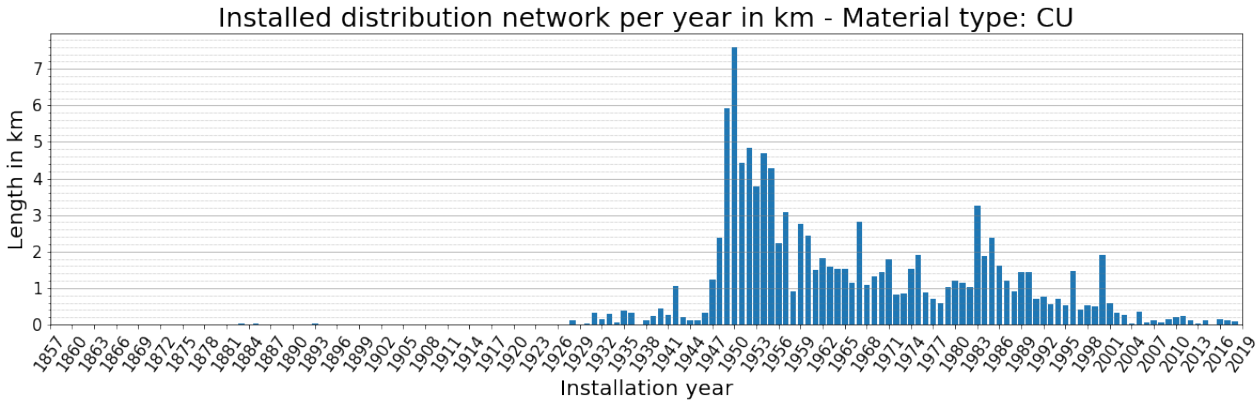
risk of floatation. Caution should be used where there are contaminated land, surge conditions, and ground subject to movement and subsidence (Reed et al. (2007), as cited in St. Clair & Sinha (2014)).

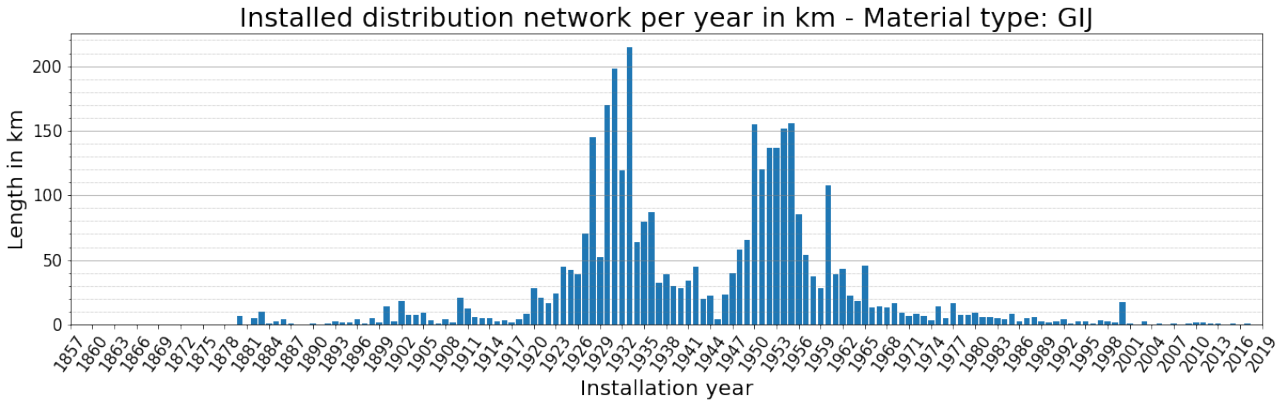
Asbestos Cement Pipes

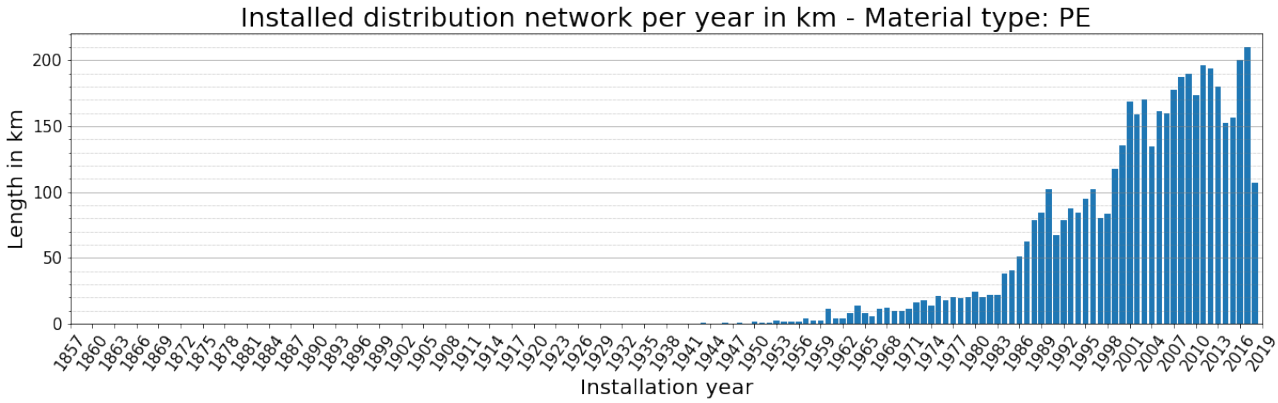
(AC) pipes were first introduced in the 1930s and were an alternative to CI pipe from 1950 to 1960. The production of AC pipes consisted of cement substance containing around 2% by weight of white asbestos fibers and portland cement. AC pipes were primarily used in heavy trafficked areas and contaminated land (organic) and where the ground was subject to minor movement. The advantages of using AC pipes were their strength and rigidity, ability to withstand fluctuating pressure and surges, and their corrosion resistance to most soils and waters. The disadvantages of using AC pipes include the following: they are susceptible to impact and accidental damage and they are vulnerable to chemical attack by certain soils and waters (Reed et al. (2007), as cited in St. Clair & Sinha (2014)).

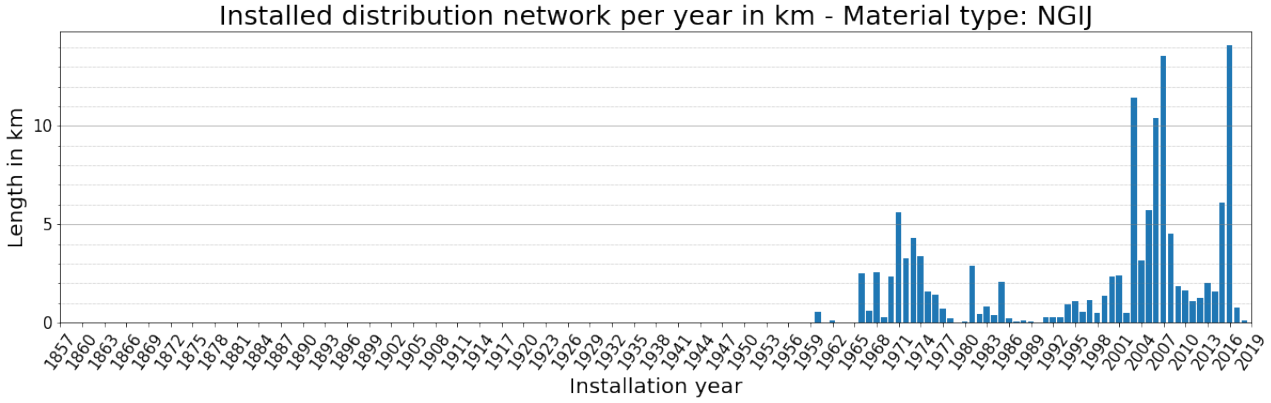
B Distribution network installed per year per material type

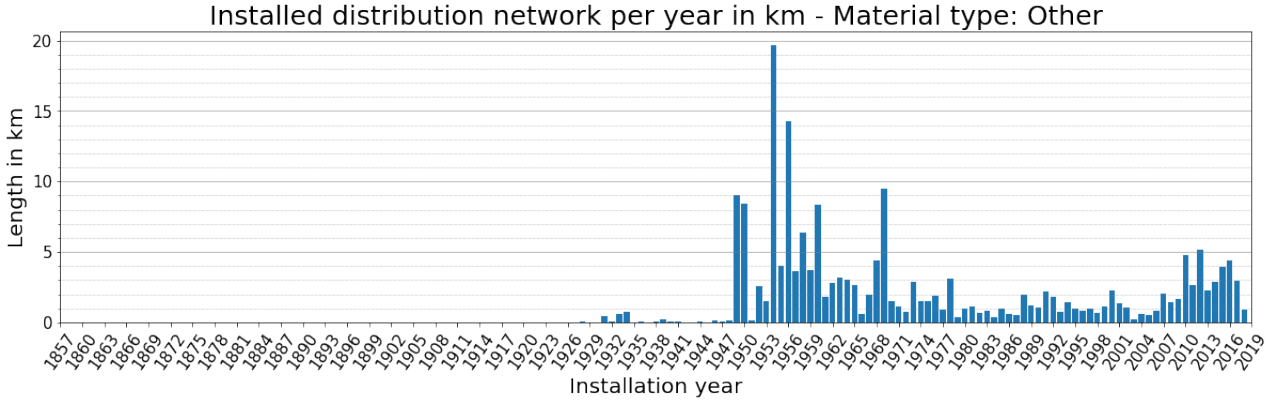


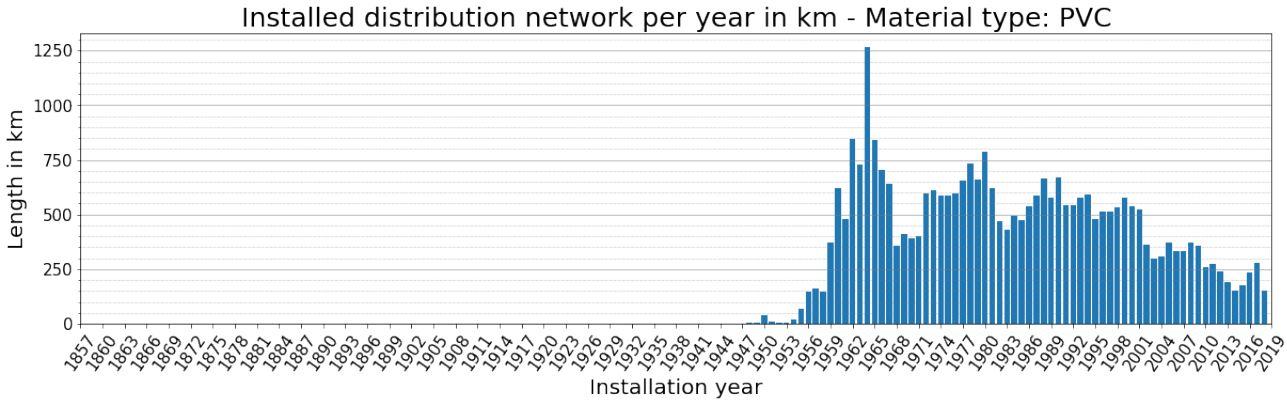


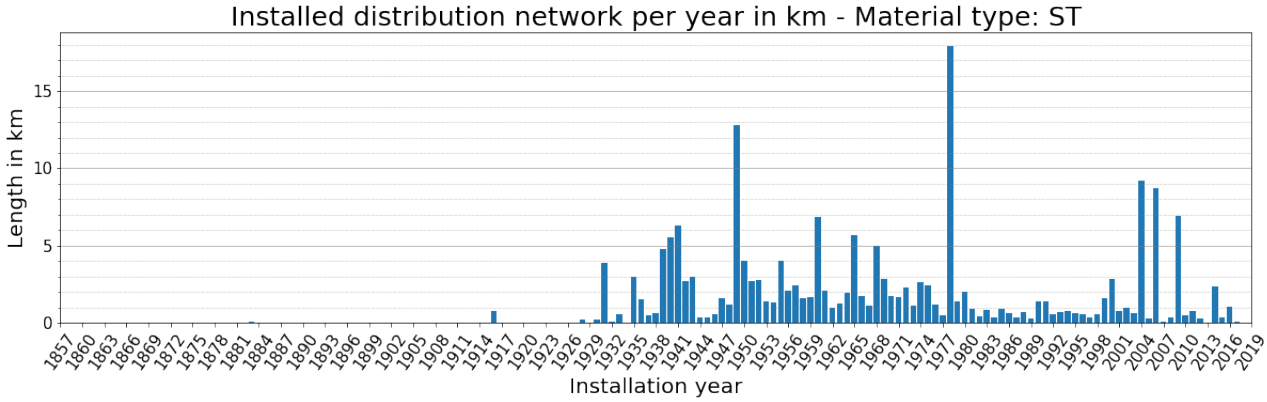



Installed distribution network per year in km - Material type: PE
Length in km
Installation year


Installed distribution network per year in km - Material type: NGIJ
Length in km
Installation year


Installed distribution network per year in km - Material type: Other
Length in km
Installation year





C Pseudo code of creation of asset units

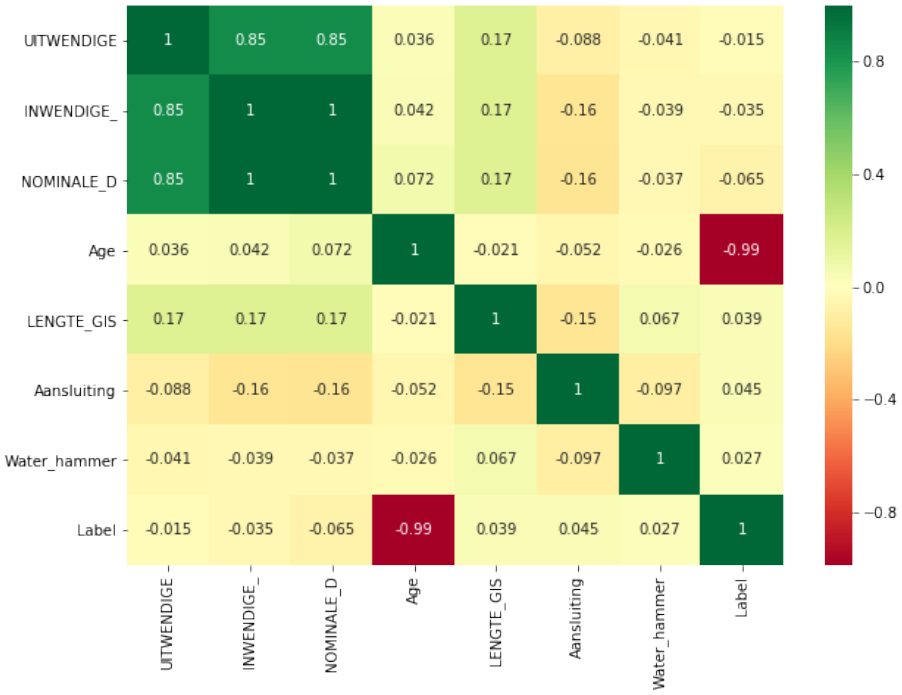
```
Input: Cleaned and validated distribution network
Output: Distribution network grouped in asset units

Initialize dictionary{asset_unit_ID : [asset_IDs]}
Create asset unit dataframe

Create a buffer around the geometry shape of individual water pipes

For each material type:
    net_m = network[network['MATERIAAL'] == materiaal]

    if materiaal in ['AC', 'GIJ', 'NGIJ']
        col_diameter = 'Internal diameter'
    else:
        col_diameter = 'Nominale diameter'

    diameters = net_m[col_diameter].unique()

    for diam in diameters:
        net_d = net_m[net_m['col_diameter'] == diam]
        installation_age = net_d['Age'].unique()

        for year in installation_age:

            tuples = geopandas.sjoin(net_d, net_d, op='intersects')
            Perform algorithm to determine which tuple connections belong to
                the same asset unit
```

```
            Store connected water pipe IDs in an array and give unique asset
                unit ID
            Add result to dictionary

For each key in dictionary:

    1. Merge water pipe IDs into asset unit and merge geometry lines together
    2. Concatenate result to asset unit dataframe
```

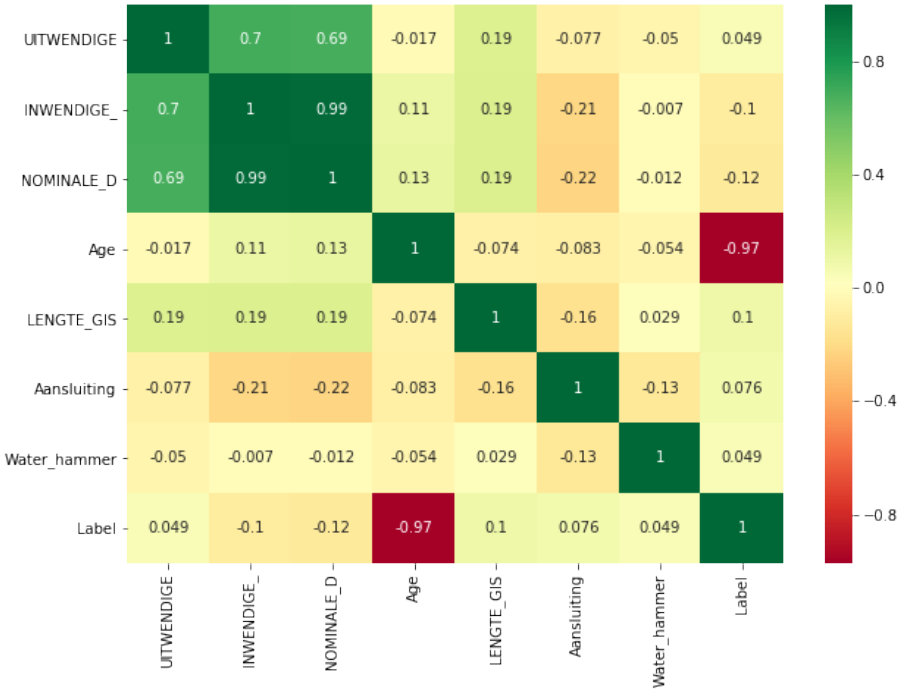
D Pseudo code of linking failures to asset units

```
Input: Created asset units and historical failure data set
Output: Failure event ID linked to asset unit ID

Transform GPS coordinates from EPSG:4326 to EPSG:28992

For each material type:
    net_m = network[network['MATERIAAL'] == materiaal]

    if materiaal in ['AC', 'GIJ', 'NGIJ']
        col_diameter = 'Internal diameter'
    else:
        col_diameter = 'Nominale diameter'

    diameters = net_m[col_diameter].unique()

    for diam in diameters:
        line = network[(network['MATERIAAL'] == mat) & (network[col_diameter]
            == 'diam)]
        point = failures[(failures['Materiaal'] == mat) & (failures['Diameter']
             == 'diam)]
        line = line.unary_union

        geometry_new = []

        for i point.index:
            result = line.interpolate(line.project(<location of failure of
                point.[i]>)
            geometry_new.append(result)

       distance = []

        for i in point.index:
            Calculate distance between old and new point

        df = geopandas.sjoin(network, <new locations of failures>, op='
            intersects', how='left')

        All failure IDs have been linked to asset unit ID in dataframe df
```

E Pearson and Spearman correlation matrix



Pearson correlation matrix on continuous variables



Spearman correlation matrix on continuous variables