Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

Joakim Skjelde

On the Transferability and Comparability of Sewer Deterioration Models

A Case Study on Norwegian Sewer Data

Master's thesis in Civil- and Environmental Engineering
Supervisor: Franz Tscheikner-Gratl, Marius Møller Rokstad, Bardia Roghani, Shamsuddin Daulat
June 2023

NTNU
Norwegian University of
Science and Technology

Joakim Skjelde

On the Transferability and Comparability of Sewer Deterioration Models

A Case Study on Norwegian Sewer Data

Master's thesis in Civil- and Environmental Engineering
Supervisor: Franz Tscheikner-Gratl, Marius Møller Rokstad, Bardia Roghani, Shamsuddin Daulat
June 2023

Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

NTNU
Norwegian University of Science and Technology

Acknowledgements

This thesis is the final work of my M.Sc. degree in Water and Wastewater Engineering at the Norwegian University of Science and technology, Trondheim. This thesis represents the culmination of two years of hard work, long study days, countless cups of coffee, and a wealth of new experiences and friendships. First and foremost, I would like to thank my supervisors:

- Associate Professor Franz Tscheikner-Gratl: For your in depth knowledge in urban water systems, asset management and deterioration modelling, and your eye for details.
- Associate Professor Marius Møller Rokstad: For your knowledge in sewer deterioration modelling, statistical modelling and machine learning methods, and for helping me with GompitZ when I was lost in calibration.
- Postdoc Bardia Roghani: For your knowledge in asset management, detailed reviews in the writing process and eyes for details regarding all the parts of my project.
- PhD candidate Shamsuddin Daulat: For your knowledge regarding this specific project and data used in my thesis, your knowledge regarding the Random Survival Forest, and always being helpful when I have been stuck with my modelling. Additionally, I appreciate the opportunity to contribute to your work, as it has been a valuable learning experience.

To each of you, I extend my deepest gratitude for your involvement, interest, and assistance throughout my project thesis last autumn and the completion of this thesis. Our discussions have helped me develop critical thinking skills and extended my knowledge regarding scientific methods, urban water asset management and deterioration modelling.

I would like to express my appreciation to Thomas Fugledalen at Sweco for providing me with his Python code for running the GompitZ model. His contribution greatly facilitated the implementation of the model in my work. Furthermore, I am thankful to all those who have contributed to this research by providing data for modeling and offering insights into the methods employed.

Additionally, a special note of gratitude goes out to all my fellow students whom I have had the pleasure of working together with the last two years. Thanks for all our discussion, both professional and non-professional, for all our social gatherings, and the support we have given one another, especially during the final weeks of completing our theses. A special thank you goes to Christian Skagsoset, whom I have had the pleasure of sharing reading room with the last year, for all our discussions regarding Python, our thesis topics and everyday stuff in general.

Lastly, I want to express my deepest gratitude to my girlfriend, Elly, for her support these years, for telling me that I need to take breaks from studying, and always believing in me. I would also like to thank my family and friends, particularly my parents and sisters, for their constant support in both my academic and personal life, without which none of this would have been possible. Additionally, I am grateful to my nieces, Emma and Ida, whose presence always brightens my mood, providing much-needed relief during the stressful periods of my studies.

Trondheim
2023-06-10

Abstract

Sewer systems serves as critical infrastructure for a municipality, and as the sewer pipes deteriorates over time, knowledge about their current and future condition is an important part for better planning of rehabilitation and inspection. In Norway, the estimated investments needed for the sewer system the next 20 years is approximately 114 billion NOK. Therefore, utilizing models to predict the condition of the sewer systems will play an important role in the planning of rehabilitation and replacement. The use of machine learning for sewer deterioration modelling has shown good results, but the models requires a significant amount of data, which small municipalities often lack.

This thesis aims to evaluate if a "Global" sewer deterioration model trained in data from several municipalities can be used to predict the sewer condition in another municipality. This is done using the survival model Random Survival Forest, and the classifier Support Vector Machine. Sewer data from five Norwegian municipalities are used, where four of them are used for training and one for testing, repeated for each municipality. The results from the global model is compared with a local model trained on data from the specific municipality. Further, as the Random Survival Forest has seldomly been used in sewer deterioration modelling, its output are compared on a network level against the GompitZ model, and on the pipe level against the Support Vector Machine. Furthermore, the feature importance of the different models are addressed and discussed.

Results from the study indicates that sewer deterioration models can be transferred between representative municipalities, and the performance scores for both models shows that they are significantly better than guessing in most cases. For the comparability, the Random Survival Forest achieved reasonable survival curves laying between the pessimistic and optimistic curves derived with the GompitZ model. Nevertheless, the curves for the transition probabilities between the good pipes differs significantly for pipes younger than 50 years due to deviation in the initial survival probability. Furthermore, using the Random Survival Forest for predicting good and bad pipes gave almost identical predictions as the Support Vector Machine, using a probability cutoff of 0.88. Lastly, the feature importance study indicates the pipe length as the most important variable, probably serving as a proxy for an unknown variable.

Keywords: Sewer Deterioration Modelling, Random Survival Forest, Transferability, Comparability, Feature Importance

Sammendrag

Avløpssystemer er kritisk infrastruktur for kommuner, og ettersom tilstand til avløpsrørene forverres over tid, er kunnskap om deres nåværende og fremtidige tilstand ett viktig ledd for bedre planlegging av rehabilitering og inspeksjon. I Norge er det estimert at omtrent 114 milliarder kroner må investeres i avløpssystemene de neste 20 årene. Bruk av modeller for å forutse tilstanden vil derfor spille en viktig rolle i planleggingen av rehabilitering og utskifting. Bruken av maskinlæring til slik modellering har vist gode resultater, men er avhengig av store mengder data, noe små kommuner ofte mangler.

Denne studien har som mål evaluere om en "global" tilstandsmodell som er trent på data fra flere kommuner kan brukes til å forutsi tilstanden i en annen kommune. Dette gjøres ved bruk av overlevelsesmodellen Random Survival Forest og klassifiseringsmodellen Support Vector Machine. Data fra fem norske kommuner er benyttet, der fire av dem brukes til trening av modellen, og den siste brukes til testing, som gjentas for hver kommune. Resultatene fra den globale modellen sammenlignes med en lokal modell som er trent på data fra den spesifikke kommunen. Videre, ettersom Random Survival Forest sjelden har blitt brukt i tilstandsmodellering av avløpsrør, blir resultatene sammenlignet på nettverksnivå mot modellen GompitZ og på rørnivå mot Support Vector Machine. Videre blir signifikansen til forklaringsvariablene i brukt i modellen diskutert.

Resultatene fra studien viser at en global modell trent på data fra representative kommuner kan brukes til å forutse tilstanden til avløpsrør i en annen kommune. Modellene er generelt sett betydelig bedre enn ren gjetting i de fleste tilfeller. Resultatene rundt sammenlignbarhet viser at overlevelseskurvene fra Random Survival Forest ligger rimelig bra plassert mellom de optimistiske og pessimistiske kurvene lagd av GompitZ. Likevel viste studien relativt store avvik i tilstandssannsynlighet for rør yngre enn 50 år, som resulterer i store avvik mellom kurvene i dette tidsrommet. Videre ga bruken av Random Survival Forest til å forutse om ett rør var i god eller dårlig tilstand gode resultater, nesten identisk med Support Vector Machine, gitt en sannsynlighetsterskel på 0.88. Til slutt viser studien at lengden på avløpsrøret er den viktigste variabelen, da den sannsynligvis opptrer som en stedfortreder for en ukjent variabel.

Nøkkelord: Tilstandsmodellering, Random Survival Forest, Overførbare Modeller, Modellsammenligning, Signifikante Variable

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC: Area Under Curve

CC: Condition Class

CCTV: Closed-Circuit Television

C-Index: Concordance Index

FNR: False Negative Rate

FPR: False Positive Rate

ROC: Receiver Operating Characteristic

RSF: Random Survival Forest

SVM: Support Vector Machines

TPR: True Positive Rate

TNR: True Negative Rate

1 Introduction

The access to clean water and functioning sanitation systems, e.g. sewer systems, is crucial for peoples health, and is one of the Sustainable Development Goals stated by the United Nations (2015). For the urban environment, the water infrastructures are necessary for a well functioning municipality, where systems for collecting and transporting sewer are an important asset (Hahn et al., 2002). In general, sewer systems convey two types of water, namely wastewater and stormwater (Butler et al., 2018). The former is water produced from community's use of freshwater, e.g. toilet flushing or washing, whereas the latter is water generated by precipitation. In the European Standard EN 752 (2017), four objectives of the sewer systems are defined. These are: (1) services to public health and safety, (2) occupational health and safety, (3) environmental protection, and (4) sustainable development. A well functioning sewer system is therefore necessary to achieve these objectives. In 2010, 2015 and 2021, the Norwegian association of consulting engineers (RIF) have published reports on the state of infrastructure in Norway, named "State of the nation" (RIF, 2021). The reports look into and highlights the values and needs of the infrastructure in Norway, among others including the state of the water- and wastewater systems. In the reports, a grading system from 1 to 5 is used to describe the current condition of the infrastructure, where grade 1 is a system that does not achieve their objectives given the demands and needs of today, and grade 5 is a system that fulfill their objectives and is capable of managing future changes. In the latest report from 2021, the municipal sewer systems are given grade 3 of 5 for whole Norway on average, meaning that the system is in an acceptable, but not good state. The system will therefore need an extensive amount of rehabilitation and investments to deliver its objectives. In the study of Bruaset et al. (2021), the investments needed to upgrade the Norwegian sewer systems from grade 3 to grade 4 were addressed. It is estimated that the investment need is approximately 320 billion NOK from year 2021 to 2040, divided between the systems owned by the municipalities (186 billion NOK) and private systems (134 billion NOK), including both sewer pipes and sewer treatment. For the municipalities, 114 billion NOK is estimated to be needed for renewal and upgrading of the sewer pipes. According to Statistics Norway (2023), approximately 86% of the Norwegian population in 2021 was connected to municipal owned sewer systems. A lag in investments and rehabilitation will therefore affect a significant amount of people. To be able to plan for future investments and schedule maintenance, implementing and utilizing an asset management methodology will play a key role. The need of extensive investments are not limited to Norwegian sewer systems, in fact several European countries are in the need for substantial amount of investments which in most cases aren't being met (Tscheikner-Gratl et al., 2019). Countries such as Austria, Germany, the Netherlands and France all face investment needs of millions and billions of euros, but in most of the cases, the rate of rehabilitation is not being met.

Ugarelli (2008) describes four different asset management strategies: (1) Operative-reactive, (2) Inspection-condition-based, (3) Proactive-preventive and (4) Predictive-advanced. Ugarelli et al. (2010) describes the duty of all municipalities should be to achieve the latter strategy, i.e. the predictive approach. Despite this, there is no best or worst approach, as they all have their specific role within the asset management methodologies. The reactive approach is simple as decisions often are made ad hoc or based on experience. The approach realizes the assets full lifetime, as it is only rehabilitated or replaced when it has failed. The downside of this is the unplanned, and often increased, costs of rehabilitation or replacement of the asset, and potential costs for the users. The other strategies differs from the reactive approach as they all include the actual condition of the asset in their methodology, known as condition assessment. Tscheikner-Gratl et al. (2019) describes condition assessment as a vital component in any asset management strategy, as a backbone to use risk-based approaches. The above condition based strategies are all useful within risk-based asset management, depending on the estimated risk of the asset (Ugarelli et al., 2010). For an asset with low risk, the simpler approach, e.g. the inspection, could be suitable, while assets with higher risk will need a more advanced model, e.g. a proactive or predictive approach.

Generally, most of the methodologies used for condition assessment are aiming to provide an overall grade of the sewer system (Kley et al., 2013). Defects in the sewer, usually recorded during inspections using Closed-Circuit Television (CCTV), are used as input for classifying the sewer condition. The use of CCTV is by Roghani et al. (2019) described as the industry standard for sewer inspections, mainly due to the low cost of use compared to other methods. Nevertheless, the use of CCTV has been criticized for being heavily affected by human factors (e.g. Dirksen et al., 2013). The probability of an inspector to not recognize a defect is significantly higher than the probability of recognising a defect that isn't there (Dirksen et al., 2013). Fugledalen et al. (2021) quantified and studied the effect of uncertainty in CCTV inspections from the city of Trondheim in Norway, concluding that the uncertainty has significant effect when the inspections are used for modelling tasks. Other methods for sewer inspections do exist, such as sonar- or laser technology (Butler et al., 2018) and image recognition on CCTV footage (Meijer et al., 2019), but the use of human inspectors and CCTV is still by far the most used method.

Rokstad and Ugarelli (2015) describes the goal of condition assessment of sewer pipes as addressing which condition the pipes are in, usually based on the defects registered during the CCTV inspections. Based on the findings after the inspection, each sewer pipe are given a condition class (CC), usually divided into 4 or 5 classes, depending on which standard are being used. In Norway, the standard given by Norsk Vann uses a 5 class approach, where a class 1 pipe are described as "good as new" and class 5 are described as "very bad" (Haugen, 2018). As the condition classification follows specific, standardised protocols, Rokstad and Ugarelli (2015) argue that the resulting CC from an inspection is in principle an objective term, which can be used as a response in a prediction model. On the other hand, in the study of Fugledalen et al. (2021), the uncertainties in the inspection data gave significant uncertainty in the output from the condition model that was used, showing the effect of the human subjectivity in CCTV analysis.

Condition assessment tools are important for utilities to gain knowledge about the state of deterioration of their infrastructure (Hawari et al., 2020). As utilities gain knowledge about the state of their sewer system by performing inspections, it is possible to use the sewer data obtained to construct models that say something about deterioration of the sewer pipes. These models are commonly known as deterioration models, where the idea is to find relationships between the factors that affects the deterioration process, and the condition of the pipe. In general, sewer pipes deteriorates with age, but pipes with different characteristics can experience significant variations in the deterioration process, depending on other factors, such as material, sewer type, diameter or soil condition (Hawari et al., 2017). Due to this, predicting the condition of sewers is considered to be a complex task, as there are multiple factors acting at once, that affects the deterioration process. Therefore, a lot of different models has been developed over the past years, with the goal of predicting the condition of sewer pipes (Tscheikner-Gratl et al., 2019).

Infrastructure deterioration models are usually classified into three groups (Yang, 2004): (1) physical models, (2) statistical models and (3) artificial intelligence-based models. Physical models are based on understanding the physical mechanisms that influence the pipe deterioration process (Ana & Bauwens, 2010). Analogous to water main breaks, the models could consider factors and mechanisms such as the structural properties of the pipe, internal and external loads, and chemical environment (Rajani & Kleiner, 2001). Still, the mechanisms leading to deterioration of pipes are often complex, and not fully understood, making it difficult to apply physical models. However, the basis for the statistical models is the relationships between the factors that affect the deterioration process of the sewer pipe, treating one or more of the factors as random variables (Rokstad & Ugarelli, 2015). The statistical models are usually divided into two subgroups, defined by Ana and Bauwens (2010) as: (1) pipe group models and (2) pipe level models. A pipe group model considers the whole network, or a section with similar properties, known as a cohort. A pipe level model on the other hand, takes the properties or features of individual pipes as as model input, and uses this to predict the deterioration of individual pipes. Several statistical models have been developed and used in sewer deterioration modelling over the years, such as the cohort survival model (e.g. Baur and Herz, 2002) and the Markov model (e.g. Dirksen and Clemens, 2008). The latter was used as the basis for the GompitZ model developed by Le Gat (2008), which has been used in various studies on sewer deterioration modelling (e.g. Rokstad and Ugarelli, 2015; Caradot et al., 2018; Fugledalen et al., 2021). The artificial intelligence-based models, i.e. machine learning models, differs from the statistical models, as they don't need any assumptions on the model structure, as they are purely information-driven (Tscheikner-Gratl et al., 2019), commonly known as data-driven (Hawari et al., 2020). The mathematical relationships between the factors driving the deterioration and the condition class of the sewer pipe are constructed by "learning" the deterioration behaviour of inspected pipes. The strength of these models compared to the others mentioned, is their capability to handle complex problems that are difficult to describe with statistical models (Ana & Bauwens, 2010). On the other hand, a disadvantage with machine learning-based models, is that they are known as "black boxes", meaning that the internal processes are somewhat unknown (Tu, 1996). The models also require a substantial amount of computational power and a high demand for data to be trained on. The latter has an substantial impact on the use of machine learning models in small municipalities, as they often lack inspection and condition data of their water- and wastewater systems (e.g. Jenkins et al., 2015; Kabir et al., 2020; Chen et al., 2022). Nevertheless, machine learning-based sewer deterioration modelling has been used in various studies over the recent years, from the elementary Decision Trees (e.g. Syachrani et al., 2013) to the more advanced Artificial Neural Networks (e.g. Sousa et al., 2014). Numerous studies have compared several of the available algorithms, such as Harvey and McBean (2014a) who compared Decision Trees and Support Vector Machines on sewer data from Guelph, Ontario, Canada, and Nguyen et al. (2022) who compared 17 different machine learning models on sewer data from the city of Ålesund, Norway. Also, different studies has compared statistical models with machine learning-based models (e.g. Rokstad and Ugarelli, 2015; Caradot et al., 2018; Laakso et al., 2019).

In general, machine learning-based sewer deterioration modelling has been based on predicting the condition classes of pipes, but a different approach that has not yet been used in any large extent, is machine learning-based survival analysis. This type of survival analysis has over the years been used in fields such as medicine (e.g. Senanayake et al., 2020; Kim et al., 2019), finance (e.g. Li et al., 2022) and even social science (e.g. Saadati, 2022). Also, these type of models have been used for predicting breaks on water distribution systems, showing good performance (e.g. Almheiri et al., 2021; Daulat et al., 2022), but they are seldomly used on sewer pipes. The use of survival models are shown to be useful for supporting estimation of investments and rehabilitation on a network level, especially when the time until an event occurring is of interest (Laakso et al., 2019). In this study, two machine learning models, the Random Survival Forest for survival analysis and the Support Vector Machine for classification, will be tested on sewer data from five Norwegian municipalities. Hence, the goal of this research is to address the possibility of transferring models trained on sewer data from several representative municipalities, and use them for predictions in municipalities who lack data to trained their own models. Furthermore, as machine learning-based survival analysis has seldomly been used for sewer deterioration modelling, the predictions from this model will be compared both on a network level against the statistical model GompitZ, and on a pipe level against a the Support Vector Machine. Additionally, the importance of the different explanatory variables is addressed for the machine learning models, to investigate which factors are contributing to the deterioration process in this study.

2 Method and materials

This section starts with description of survival analysis, before the survival analysis model and its performance metrics are described in detail. Then a description of the concept of machine learning-based classification, before the algorithm applied and its performance metrics are described, before the statistical model is described. Further, the process of hyperparameter tuning is presented, before the concepts used for model transferability and comparability is addressed. Lastly, the sewer data used for modelling are being described. The choice of models used in this thesis, is based on the work conducted by Skjelde (2022), where different machine learning algorithms used in deterioration modelling were reviewed, and the framework for comparability is described. Table 1 provides a summary of the chosen algorithms, and the criteria used when choosing the models.

Table 1: Summary of chosen algorithms (based on Skjelde (2022))

Model	Survival Analysis	Classification	Statistical
Algorithm	Random Survival Forest	Support Vector Machine	GompitZ
Model output	Cumulative Hazard Function estimated at each node of each tree	Optimal separating hyperplane between classes	Survival functions for each pipe cohort
Source	Algorithm by Ishwaran et al. (2008). Python implementation through Scikit-survival (Pölsterl, 2020)	Algorithm by Cortes and Vapnik (1995). Python implementation through Scikit-learn (Pedregosa et al., 2011)	Model by Le Gat (2008). Python implementation provided by Fugledalen et al. (2021) and calibration software by Le Gat (2011)
Implementation and proficiency	Promising results on water distribution pipes in Norway and Canada, and on sewer pipes in Finland	Promising results on sewer data from Australia, Portugal and Colombia, among others	Used on sewer data from two Norwegian municipalities, Oslo and Trondheim

2.1 Random Survival Forest

Survival analysis is by Kleinbaum and Klein (2012) defined as a cluster of statistical procedures for analyzing data where the outcome is the time until an event occurs. Another common name for survival analysis is therefore time-to-event analysis (Mills, 2010). The major difference between ordinary statistical methods, or machine learning methods, and survival analysis, is that survival models takes censoring of data into account. Klein and Moeschberger (2003) divides censoring of data into three categories: (1) Right censoring, (2) left censoring and (3) interval censoring. Right censoring is when all that is known is that the event hasn't occurred yet, left censoring is when all that is known is that the event has occurred before the start of the study, while interval censoring is when all that is known is that the event has occurred within a time interval. A visual description of the three cases is shown in Figure 1.

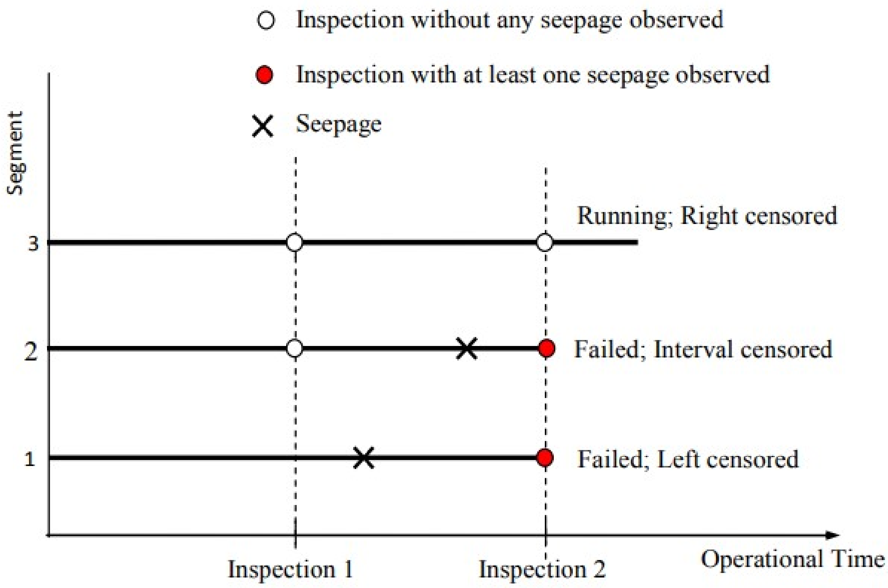

Figure 1: Example of censored data (Jiang et al., 2014)

There exists several survival models for estimating time-to-event, where two of the most popular ones are the Kaplan-Meier estimator (Kaplan & Meier, 1958) and the Cox proportional hazards model (Cox, 1972). The Kaplan-Meier formula is used to compute survival probabilities based on survival times for an event, and the number of events for each survival time (Kleinbaum & Klein, 2012). The general formula is (Klein & Moeschberger, 2003):

$$\hat{S}(t) = \prod_{t_i \leqslant t} (1 - \frac{d_i}{n_i}) \tag{1}$$

In the above equation (Equation 1), $\hat{S}(t)$ is the estimated survival function at time t, d_i is the number of events at time t_i and n_i is the number of individuals at risk at time t_i. The formula is non-parametric, meaning no parameters need to be specified, as it only uses a sample of right-censored data for creating the survival curve (Klein & Moeschberger, 2003). To get the cumulative hazard function, the Nelson-Aalen estimator can be used (Aalen, 1978; Nelson, 1972). This is similar to Kaplan-Meier, being non-parametric, on the form (Klein & Moeschberger, 2003):

$$\hat{H}(t) = \sum_{t_i \leqslant t} \frac{d_i}{n_i} \tag{2}$$

Equation 2 takes the same input as Equation 1, with the number of events, d_i, and individuals at risk, n_i, at time t_i. In general, the relationship between $\hat{S}(t)$ and $\hat{H}(t)$ is defined as (Kleinbaum & Klein, 2012):

$$\hat{S}(t) = e^{-\hat{H}(t)} \tag{3}$$

The Kaplan-Meier and Nelson-Aalen estimators do not use any explanatory variables except the time of an event, making it too simple for many survival analyses (Kleinbaum & Klein, 2012). To consider several explanatory variables at ones, the Cox proportional hazard model (Cox, 1972) is a commonly used method. The general formula is (Kleinbaum & Klein, 2012):

$$h(t, \mathbf{X}) = h_0(t) * e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_n)} \tag{4}$$

In Equation 4, the hazard at time t with a set of explanatory variables $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ is given as the product of a baseline hazard function, $h_0(t)$, and the exponential expression e with the linear sum over the n explanatory variables. As the equation gives the hazard function, the survival curves can be estimated, called adjusted survival curves, as they are adjusted to take the explanatory variables into account. The formula is on the form (Kleinbaum & Klein, 2012):

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_n)}} \tag{5}$$

Equation 5 uses a baseline survival function, $S_0(t)$, raised to a power equal to the exponential part of Equation 4.

Within the field of medicine, survival analysis are being used to a large extent, usually to predict the time of death for patients with different medical conditions, e.g. leukemia or heart transplantation (Kleinbaum & Klein, 2012). Different statistical methods for survival analysis have also been used to predict failure of water distribution pipes (e.g. Røstum, 2000; Park et al., 2008), and sewer pipes (e.g. Le Gat, 2008; Egger et al., 2013) over the years. A new approach has been to combine machine learning methods with the concepts of survival analysis, both in medicine (e.g. Kvamme et al., 2019; Moncada-Torres et al., 2021; Deepa and Gunavathi, 2022) and water distribution prediction (e.g. Almheiri et al., 2021; Daulat et al., 2022). Laakso et al. (2019) compared a machine learning-based survival analysis model with a statistical survival analysis model on the network level. Beyond this, machine learning-based survival analysis haven't been used for sewer deterioration modelling to any great extent.

The Random Survival Forest algorithm, developed by Ishwaran et al. (2008), is a combination between the Random Forest algorithm by Breiman (2001) and survival analysis. To understand Random Survival Forest, the concepts behind Random Forest will be explained before describing how survival analysis is implemented. As a Random Forest is an ensemble of Decision Trees voting for the most popular class (Breiman, 2001), a description of Decision Trees is given. The CART (classification and regression tree) algorithm developed by Breiman et al. (1984) is the most common algorithm used for tree-based classification (Géron, 2019). The Decision Tree algorithm predicts the target variables by implementing a set of prediction rules which are arranged in a tree-like structure (Syachrani et al., 2013), as seen in Figure 2.

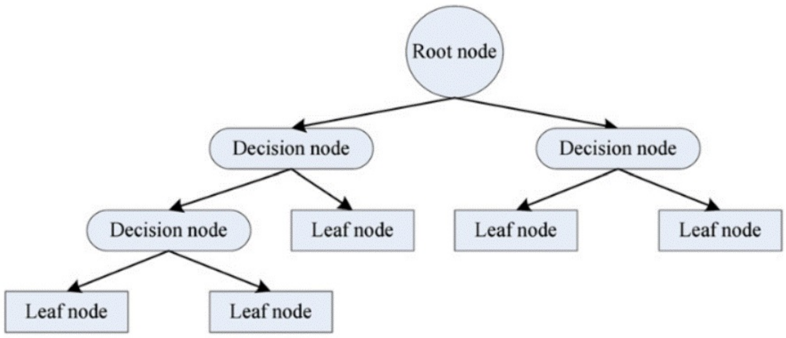

Figure 2: Decision Tree structure (Syachrani et al., 2013)

In the training process, the construction of prediction rules starts with a root node where all the observations are initially assigned (Syachrani et al., 2013). The process moves further by splitting the root into branches named decision nodes, based on the values of the provided predictor variables. Further, the number of observations on the higher node is distributed to the lower nodes. This splitting process will be recursively repeated for each of the branches until all the observations in a decision node have the same classification. The decision nodes where the splitting process have stopped, are known as leaves. The objective of the splitting is usually to minimize the so called Gini criterion (Breiman, 2002). The performance of the ability the split has to classify the output is evaluated. The Gini criterion has its maximum value if the output distribution in both nodes are similar, i.e. poor classification, while a value of zero is obtained if the outputs are separated perfectly between the two nodes, i.e. excellent classification (Caradot et al., 2018). The drawback of Decision Trees are that they are easily overfitting to the data they are trained on (Kotsiantis, 2013). Combining them into an ensemble, as described by Breiman (2001), is done to prevent overfitting, based on the Law of Large Numbers. The Random Forest algorithm is based on growing multiple decision trees, i.e. a forest, where the class with most votes at the end is the resulting prediction. By using these votes, it is possible to calculate the probability of belonging to a class (Rokstad & Ugarelli, 2015). The algorithm is random as it is using bagging (bootstrap aggregating), where each decision tree is trained on a random subset of the data (Breiman, 1996), and a randomly selected subset of explanatory variables are used at each node of the tree (Breiman, 2001). Using the theory of Random Forests for right-censored survival analysis, is proposed by Ishwaran et al. (2008) who describes the Random Survival Forest (RSF). The algorithm is similar to the Random Forest, but the trees grown are called survival trees, where the node splitting is based on maximizing the survival difference between the next nodes. The tree is grown to full size with the criteria that a leaf node has above $d_0 > 0$ unique deaths. At the end, a cumulative hazard function (CHF) is calculated for each tree, where the CHF for the ensemble (forest) is the average of all trees. The CHF is calculated using the Nelson-Aalen estimator given in Equation 2.

A commonly used parameter to evaluate survival models is the concordance-index (c-index) by Harrell et al. (1982). The index is a measure of the ability the model has of ranking event based on the highest risk of failure. A pair of samples are drawn, in this case two randomly chosen pipes, and if the model predicts a higher risk of failure for the pipe that fails first, the pair is concordant (Laakso et al., 2019). Also, the calculation is based on if the pairs are comparable or non-comparable (Harrell et al., 1982). A pair is comparable if both elements have experienced an event, i.e. they are uncensored, or if one of the elements are censored and the survival time of the censored case is greater than the uncensored. If the pair consist of two censored elements, or if one of them is censored its survival time is lower than the uncensored element, the pair is non-comparable. The C-index is calculated by the formula (Harrell et al., 1982; Schmid et al., 2016):

$$C = \frac{\sum_{i,j} I(T_i > T_j) \cdot I(\eta_i < \eta_j) \cdot \Delta_j}{\sum_{i,j} I(T_i > T_j) \cdot \Delta_j} \tag{6}$$

where i and j refers to pars of observations, T_i, T_j are survival times and η_i, η_j are predictions. The

Δ_j element discards pairs that are non-comparable as the smaller survival time is censored, giving $\Delta_j = 0$. In general, the formula can be summarized as the number of concordant pairs over the sum of concordant pairs and discordant pairs, meaning that a value of 1 denotes a perfect model, while a value of 0.5 denotes a random model.

2.2 Support Vector Machines

To predict the condition class of a sewer pipe, the machine learning algorithms used need to be able to learn from a set of input(e.g. pipe material, dimension, age) with a given output (condition class) (Harvey & McBean, 2014b). Within supervised learning, this is known as classification (Géron, 2019). Here, the training data used in the machine learning model includes the solution, called a label. A lot of classification algorithms exists, and several of them has been used for sewer deterioration modelling, such as Decision Trees (e.g. Harvey and McBean, 2014a), Random Forest (e.g. Rokstad and Ugarelli, 2015), Support Vector Machines (e.g. Hernández et al., 2021) and Neural Networks (e.g. Atambo et al., 2022).

Support Vector Machines (SVM) is an algorithm for classification based on the work of Cortes and Vapnik (1995), originally for two-class (binary) problems. It is also possible to use SVM for multiclass output, e.g. five class sewer condition assessment, by using a one versus one approach between the outputs (Pedregosa et al., 2011). The concept of the algorithm is to map the input vectors, e.g. pipe characteristics, on a high dimensional feature space, where a linear decision surface is constructed. The method uses so-called kernel functions to do this mapping (Shawe-Taylor & Cristianini, 2004). Some available kernel functions are linear, polynomial and radial basis function (RBF) (Pedregosa et al., 2011). The goal of SVM is to find the optimal hyperplane that maximizes the separation between the classes, who also generalizes well (Cortes & Vapnik, 1995). Optimally, each class should be on its own side of the separation margin, known as hard margin classification (Géron, 2019). The main issue is that this is only possible for linearly separable data, and it is sensitive to outliers. A more flexible model is needed to avoid this, where the separation distance is as large as possible, and violations of the margin, i.e. data points at the margin area or on the wrong side, is low. This is known as soft margin classification, and is controlled by the hyperparameter C in the model. A lower value of C gives a wider margin, but more violations. Of the available kernels, the RBF kernel is chosen as it a common and successfully applied kernel, especially for sewer deterioration modelling (e.g. Mashford et al., 2011; Hernández et al., 2021). The RBF kernel takes in addition to C, γ as hyperparameter, where it acts like a regularization parameter (Géron, 2019). A small γ value gives larger variance, while a higher value gives smaller variance.

The predictions from the SVM model, or any classification model, can be sorted in a confusion matrix, showing the amount of correct and incorrect predictions (Géron, 2019). An example of a confusion matrix for a binary classifier is shown in Table 2.

Table 2: Confusion Matrix for a Binary Classifier (Based on Harvey and McBean (2014b))

		Predicted Condition	
		Good	Bad
Actual Condition	Good	True Positive (TP)	False Negative (FN)
	Bad	False Positive (FP)	True Negative (TN)

The predictive performance of a classifier can be evaluated using the confusion matrix, or it can be addressed using the model accuracy, true positive rate (TPR), and true negative rate (TNR) (Harvey & McBean, 2014b). These measures are calculated by the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$TPR = Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

$$TNR = Specificity = \frac{TN}{TN + FP} \tag{9}$$

In addition to the above metrics, the false negative rate (FNR) and the false positive rate (FPR) can also be used for evaluating the model predictive performance (Harvey & McBean, 2014b):

$$FNR = 1 - Sensitivity = 1 - TPR \tag{10}$$

$$FPR = 1 - Specificity = 1 - TNR \tag{11}$$

Using accuracy solely as the performance measure of a model is unsuitable for imbalanced dataset (e.g. Géron, 2019, Harvey and McBean, 2014b), and other measures should be used in addition to better evaluate the model performance. A commonly used method is to plot the TPR as a function of the FPR, for different probability cutoffs. This is known as the receiver operating characteristic (ROC) curve (Géron, 2019), and is shown in Figure 3.

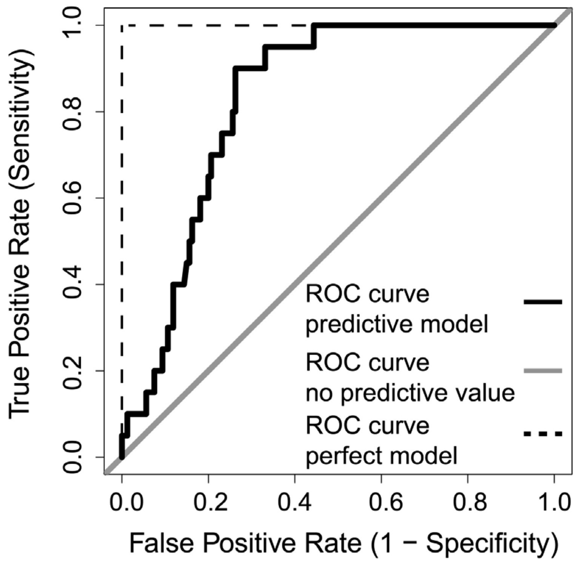

Figure 3: Example of ROC curve (Harvey & McBean, 2014b)

The straight line in Figure 3 denotes a random model with no predictive value, while the dotted line represent a perfect model (Harvey & McBean, 2014b). In general, the model that stays the furthest away from the random model line, is the better (Géron, 2019). One way to compare different classifiers with the ROC curve, is to calculate the area under the curve (AUC). An AUC equal to 1 denotes a perfect model, while an AUC equal to 0.5 denotes a random model. Further, (Hosmer & Lemeshow, 2000) states than an AUC above 0.7 is a good model, while above 0.8 it is a excellent model, but this could be dependent on the field of use. Caradot et al. (2018) proposed two performance metrics for sewer condition modelling, in a study where two sewer deterioration models were compared, the statistical GompitZ model and a Random Forest model. The study used three condition classes, good, medium and bad pipes. The two metrics are one on network level and one on pipe level. The network level metric is based on finding the absolute deviation between predicted and observed number of pipes in each of the condition classes. Furthermore, the same metrics are computed for a given age category of pipes. The metric formula can be written as follows (based on Hernández et al. (2021)):

$$K_{net} = \sqrt{\frac{\sum_{i=1}^{3} K_{\{DEV,i\}}^2 + \sum_{i=1}^{3} K_{\{OLD-DEV,i\}}^2}{6}} \tag{12}$$

The summary metric is defined as the root mean square error, which again is taken the square root of to give more weight to large errors. The pipe level metric is defined as maximizing the TPR, and minimizing FNR and FPR. The metric is on the form (based on Hernández et al. (2021)):

$$K_{pipe} = \sqrt{\frac{\sum_{i=1}^{3}(100 - K_{\{TPR,i\}})^2 + K_{\{FNR,1-2\}}^2 + K_{\{FNR,1-3\}}^2 + K_{\{FPR,3-1\}}^2}{6}} \tag{13}$$

Where the TPR factor is normalized, i.e. subtracted from 1, so that optimum value is 0.

2.3 GompitZ

The GompitZ model by Le Gat (2008) is a statistical sewer deterioration model based on the theory of non-homogeneous Markov chains, where the transition probabilities are derived from the Gompertz distribution. Using a non-homogeneous Markov Chain means that the probability of a pipe transitions from one condition class to another varies with time. Using the Gompertz distribution allows to include explanatory variables, both time-dependent and -independent. The time-dependent variables include the variables that can describe the deterioration rate of pipes over time, while the time-independent variables describes the initial deterioration state. The estimation of the parameters in the model is done by maximum likelihood estimation. The pipes are usually grouped into cohorts based on similar features, such as pipe diameter or materials, before it is calibrated (Caradot et al., 2018). During calibration, survival functions for each cohort are estimated. By using a finite number of condition classes, c, the deterioration can be described as the stochastic process $Y(t)\epsilon[1, ..., c]$, and the Gompertz survival function can be written as (Le Gat, 2008):

$$\begin{aligned} &\forall t\epsilon\mathbf{R}_+, \forall k\epsilon[1, ..., c-1]: \\ S_k(t) = P[Y(t) \leq k] = &exp[-exp(\alpha_k + t \cdot exp(\beta_1))] \\ &S_c(t) = 1 \\ &\text{assuming } \forall k\epsilon[1, ..., c-2]: \alpha_k > \alpha_{k+1} \end{aligned} \tag{14}$$

Where k ranges from the best to second worst condition class, c is the worst condition class, $S_k(t)$ is the survival function, α_k and β_1 are scalar parameters. By introducing explanatory variables, the function can be written as:

$$\begin{aligned} &\forall t\epsilon\mathbf{R}_+, \forall k\epsilon[1, ..., c-1]: \\ S_{ik}(t|u_i) = &exp[-exp(\alpha_k + \mathbf{Z}_{0i}^T\beta + t \cdot exp(\mathbf{Z}_{1i}^T\beta_{\mathbf{1}} + u_i))] \\ &S_{ic}(t|u_i) = 1 \\ &\text{assuming: } u_i \approx N(0, \sigma^2) \text{ and} \\ &\forall k\epsilon[1, ..., c-2]: \alpha_k > \alpha_{k+1} \end{aligned} \tag{15}$$

Here, $S_{ik}(t|u_i)$ is the survival probability of pipe i being in condition class k, at a given time t given the individual frailty factor (IFF) u_i of the pipe. The explanatory variables are accounted for in the vectors $\mathbf{Z}_{0i}$ and $\mathbf{Z}_{1i}$, where the first is affecting the initial deterioration state, and the second affects the time-dependent deterioration. Using the described survival function in Equation 15, it is possible to calculate the state probabilities p for each condition state at any time by the function:

$$\begin{aligned} p_1(t) &= S_1(t) \\ &\forall k\epsilon[2, ..., c-1]: \\ p_k(t) = P[Y(t) = k] &= P[Y(t) \leq k] - P[Y(t) \leq k-1] = S_k(t) - S_{k-1}(t) \\ p_c(t) = P[Y(t) = c] &= 1 - P[Y(t) \leq c-1] = 1 - S_{c-1}(t) \end{aligned} \tag{16}$$

The probability of staying in the current condition state, q_k, or jumping to the next condition state can be defined as:

$$\begin{aligned} q_k(t+1) &= P[Y(t+1) = k|Y(t) = k] \\ 1 - q_k(t+1) &= P[Y(t+1) = k+1|Y(t) = k] \end{aligned} \tag{17}$$

Further by defining the transition probability matrix $\mathbf{Q(t)}$ as:

$$\mathbf{Q}(\mathbf{t}) = \begin{pmatrix} q_1(t) & 1-q_1(t) & 0 & 0 & 0 \\ 0 & q_2(t) & 1-q_2(t) & 0 & 0 \\ 0 & 0 & q_3(t) & 1-q_3(t) & 0 \\ 0 & 0 & 0 & q_4(t) & 1-q_4(t) \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{18}$$

Introducing the state probability vector at time t as $\mathbf{p}(t) = (p_1(t) \cdots p_c(t))^T$, the Markov-Chain process can now be written as:

$$\mathbf{p}^T(t+s) = \mathbf{p}^T(t)\prod_{i=1}^{s}\mathbf{Q}(t+i) \tag{19}$$

The methods for calculating survival probability depends on if the pipe has been inspected or not. For an uninspected pipe the probability for belonging to a specific class is estimated directly from the calibrated survival curves. For an inspected pipe, the condition at year T is 100% certain, and the evolution further is simulated as a Markov Chain. The slopes of the survival curves are used to derive the transition matrix (Equation 18) (Le Gat, 2008). The transition probabilities are also restricted so that a pipe will not improve in condition, but will go to the next, "worse" condition class (Rokstad et al., 2014). This is due to the $exp(\beta_1)$ in Equation 14 always being positive (Le Gat, 2008). Pipes improving in condition class has shown to occur in machine learning models (e.g. Caradot et al., 2018), which is not a likely situation to happen in reality. Additionally, from the matrix in Equation 18 one can notice that the model assumes that the transition is happening one condition class at a time.

2.4 Hyperparameter Tuning

Tuning the parameters of a machine learning model plays a key role in increasing the model performance (Géron, 2019; Feurer and Hutter, 2019). The tuning process is based on adjusting the model parameters, and finding the parameter set that produces the best model results. Techniques such as Grid Search or Randomized Search is often used on machine learning models, but Grid Search is time consuming on large parameter sets, while Randomized Search are purely random and will be highly affected by luck (Géron, 2019). Other optimization techniques such as Bayesian optimization (Močkus, 1975) and Genetic Algorithms (Whitley, 1994) have been used to a great extent in optimizing hyperparameters in machine learning models (e.g. Snoek et al., 2012; Nikbakht et al., 2021). The Differential Evolution algorithm by Storn and Price (1997) is a Genetic Algorithm that has been used in earlier studies to optimize hyperparameters (e.g. Baioletti et al., 2020). Some of the characteristics of the algorithm are that it is metaheuristic, as it tries to improve the solution during the iterations, and it does work on optimization problems that are not differentiable, unlike classic optimization methods (Price et al., 2005). Hernández et al. (2021) used the Differential Evolution algorithm to optimize SVM based sewer deterioration models for the cities of Bogotá and Medellin in Colombia. The study implemented the metrics proposed by Caradot et al. (2018), i.e. Equation 12 and Equation 13. The findings of Hernández et al. (2021) indicate that the Differential Evolution algorithm is suitable for hyperparameter tuning of SVM models used for sewer deterioration modelling, and is therefore chosen to use in this study.

2.5 Transferability and comparability of models

In this study, transferability is defined as the process of training a deterioration model based on data from $n-1$ municipalities, where n is the total number of municipalities, and testing it on the left out municipality. This strategy is based on the leave-one-out (LOO) principle (e.g. Pedregosa et al., 2011), and is often used in validation of machine learning models. The transferability is tested for both the RSF- and SVM model, using the same strategy in terms of creating training data and testing data. To validate the predictions from the globally trained model, a local model trained on the left out data is used. This process is done for all available datasets. In total, 4 transition states will be considered for each local and global model, namely the transition between all condition class 1 and the rest, condition class 1, 2 and the rest, and so on. The transitions states are denoted as CC1/2, CC2/3, CC3/4 and CC4/5, but in theory there are no restrictions that a pipe only can move to the adjacent worse condition class, which is the case in the GompitZ model. Still, the transition states are denoted in the given way for simplicity.

Model comparability is here defined as the possibility of one type of deterioration model to reproduce the results from another deterioration model. As the RSF is the main model being used in this study, as it has seldomly been used in sewer deterioration modelling, the outputs will be compared with the other two models, both on a network level and on a pipe level. On the network level, a comparison between the RSF and GompitZ are made by comparing predicted survival curves from both models, to address the advantages and disadvantages between using a machine learning model or a statistical model. Further, on the pipe level, the RSF is compared with the predictions of the SVM model. The comparison is based on using a binary classification system, i.e. good and bad pipes. For the RSF model, this is tested by setting different probability cutoffs, i.e. below which survival probability should the pipe be classified as bad. The mentioned comparisons are conducted using the most representative municipality obtained from the transferability study, based on the best fit of C-index, survival curves and AUC-ROC score.

2.6 Sewer data

Sewer data from five Norwegian municipalities are used in this study. The municipalities are geographically dispersed, with varying climate conditions and environmental factors, such as annual temperature and precipitation. In total, 10180 CCTV inspections are considered, with a total length of approximately 318km, distributed between the municipalities. Figure 4 shows the distribution of pipe characteristics for all five municipalities combined.

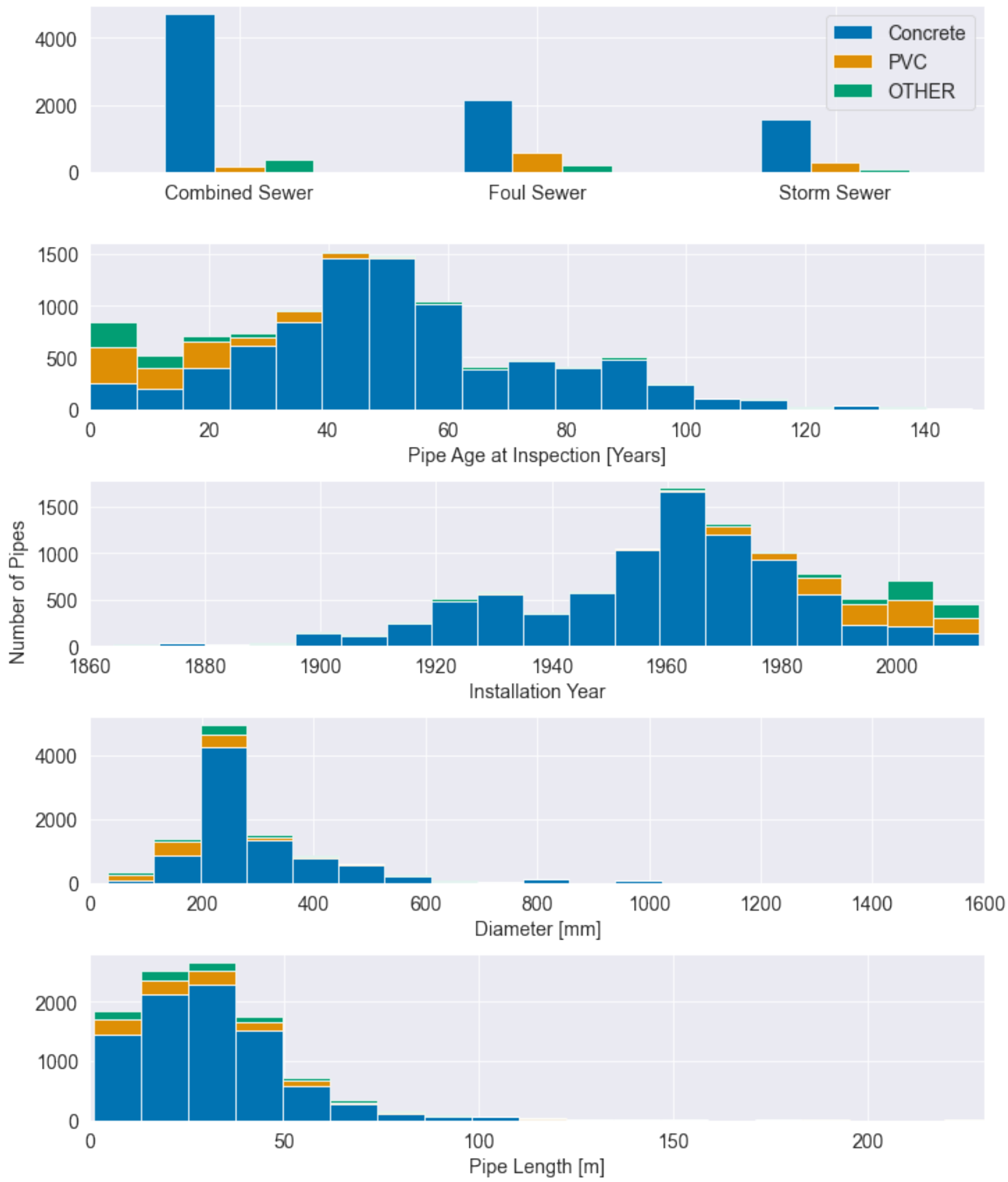

Figure 4: Pipe characteristics for the inspected pipes, with data from all municipalities combined

For the materials, concrete and polyvinyl chloride (PVC) were the only to be present in all municipalities. To be able to make global models, these where kept as they were, while the rest of materials were grouped in a combined cohort (OTHER), as seen in Figure 4.

In the given datasets, the damage score from each CCTV inspection are given, which need to be translated into a condition class. The thresholds between classes are based on the values in Table 3 and the resulting distribution of condition classes are shown in Figure 5.

Table 3: Thresholds between damage score and condition class (based on Haugen (2018))

Damage score	Condition class
1–10	1
11–40	2
41–100	3
101–150	4
151 $\leq$	5

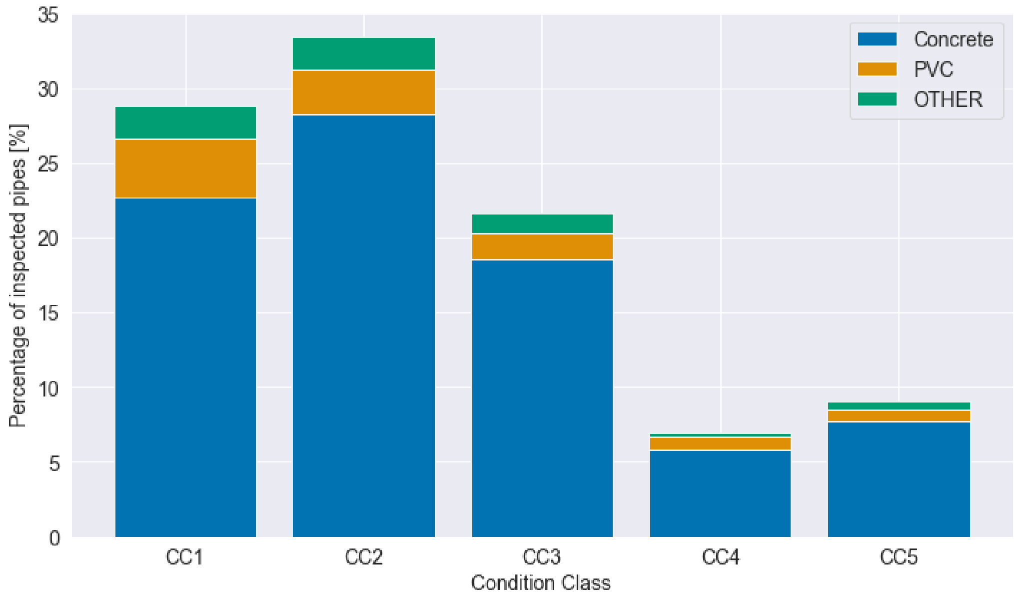

Figure 5: Distribution of condition class for the whole dataset given as percentage of inspected pipes

The distribution of condition classes for each municipality are shown in Table 4, where the total number of pipes for each municipality are given together with the distribution of condition classes as percentage of inspected pipes for each municipality.

Table 4: Distribution of Condition Class for each municipality

Municipality	Number of pipes	CC1 (%)	CC2 (%)	CC3 (%)	CC4 (%)	CC5 (%)
1	6435	27	35	23	6	9
2	263	72	26	2	0	0
3	586	23	33	22	8	14
4	1251	31	38	20	5	6
5	1645	28	26	21	13	12
Total	10180	29	33	22	7	9

The datasets had missing values for the explanatory variables, namely construction year, material and dimension. The construction year together with the time of inspection is used to calculate the age of the pipes. As pipe age in general is the main driver for sewer deterioration (Caradot et al., 2018), missing construction year or inspection year has major impact on the available data for modelling. Caradot et al. (2021) proposed methods to estimate the missing construction years of the sewers, one method using the median construction year of pipes in the same neighbourhood with similar pipe characteristics, and one method using a k-nearest neighbour approach using the five closest adjacent pipes with the same diameter and material. Tscheikner-Gratl et al. (2016) proposed methods for reconstruction missing construction year, material, dimension and failure records for water pipes, using a street section approach. The proposed methods uses information about the construction year of connecting infrastructure (e.g. buildings) and the connecting pipes.

A grading system is developed to find suitable candidates for filling missing data. Still, using information regarding year of installation or inspection from surrounding pipes for sewer could be problematic, as it is often the case that single pipes are rehabilitated or replaced individually from their upstream and downstream connections. Any pipes lacking information on the installation year or inspection year are therefore removed from the dataset. In the dataset, most pipes have a street code, which is used for filling missing material and dimension as pipes with similar street code are geographically close to each other. Over the length of a street section, the pipes will have approximately the same catchment size, where the pipe flow, and thus the dimension, will increase gradually downstream from the pipe furthest upstream. For a pipe with missing data where the material and dimension of the upstream and downstream pipes are known, the material and dimension will most probably be similar with either the upstream- or downstream pipe. The methodology for filling missing material and dimension is therefore to some extent based on the method proposed by Tscheikner-Gratl et al. (2016), with a street section based approach. The method used consist of three parts, and is similar for both dimension and material:

1. If the start and end node of the missing pipe is available, continue to step (2), if not skip to step (3).
2. Based on the start and end node of the missing pipe, if the upstream or downstream pipe are available, replace the missing value with the available value. If both upstream and downstream pipe are available, use the upstream value. This is a simplified method of the k-nearest neighbour approach described by Caradot et al. (2021).
3. If information on upstream or downstream pipes are missing, the mode of the current street section is used.
4. If information on street section is missing, the mode for the whole municipality is used.

The mode is used for imputing the values in step (3) and (4) above, as dimension and material are discrete values, and the mean will not be possible to use. The mode will be similar to the median for a fairly evenly distributed dataset, and thus this approach is similar to median imputation and mean imputation of missing data described by Kabir et al. (2020).

After the reconstruction of missing data, plausibility checks of the data were performed. As the construction year and time of inspection is used to calculate the age of the pipe, the inspection need to have happened after the construction of the pipe. The dataset operates with two different construction years for some pipes, namely old and new construction year. The new construction year is when the pipe has been replaced. In special cases, some pipes have been replaced, while the inspections have been performed on the old pipe, but the construction year of the old pipe is missing. These pipes are therefore removed from the dataset, as the age cannot be calculated. The dataset contained one pipe with a dimension of 4000 mm, which after studying the GIS-system, was found to be a highway culvert, coded with the wrong sewer type. This pipe was removed, and pipes over a certain threshold (in this case 2000 mm), where also removed. Still, the highway culvert was the only pipe over this threshold, so only one pipe where discarded.

3 Results and Discussion

3.1 Hyperparameter Tuning

The tuning of hyperparameters were done by using the earlier described Differential Evolution algorithm, by Storn and Price (1997). As the RSF models were more time consuming to train compared to the SVM models, and the need for 4 sub-models (i.e. CC1/2, CC2/3, CC3/4 and CC4/5) for each local- and global model, it was chosen not to tune the hyperparameters of these models. Example values for the model parameters from the Scikit-Survival module (Pölsterl, 2020) were used for all sub-models, listed in Table 5.

Table 5: Predefined parameter set for the Random Survival Forest

n-estimators	**min-samples-split**	**min-samples-leaf**	**max-features**
1000	10	15	4

Here, "n_estimators" denotes the number of trees in the forest, "min_samples_split" denotes the minimum of samples needed to split an internal node, "min_samples_leaf" denotes the minimum number of samples required at a leaf node, while "max_features" denotes the number of features (i.e. explanatory variables) to consider when looking for the best split (Pölsterl, 2020; Song et al., 2023). To assess which parameters to focus on in a potential tuning process, a sensitivity analysis for the four variables listed in Table 5 was conducted. The results are shown in Figure 6.

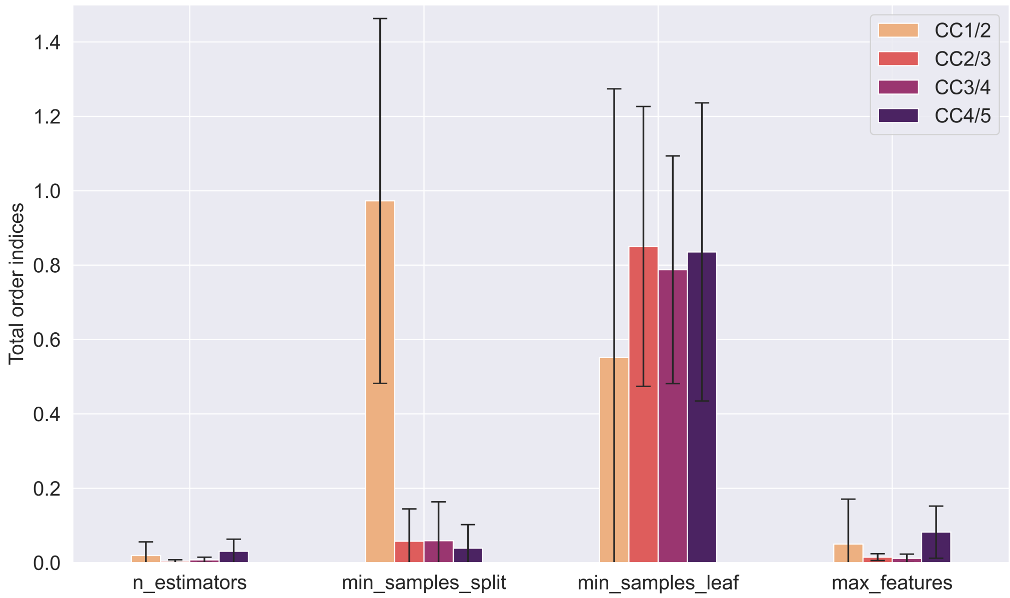

Figure 6: Sensitivity Analysis of the Global RSF model

For each parameter, the columns represent one of the sub-models, i.e. CC1/2, CC2/3, CC3/4 and CC4/5. The Sobol-indices (e.g. Sobol, 2001; Saltelli, 2002) were used as measures, and was computed using the SALib module (Herman & Usher, 2017) in Python. The results from the sensitivity analysis shows that the parameters "min-samples-split" and "min-samples-leaf" are the most sensitive for the CC1/2 sub-model, while the latter parameter are the most the sensitive for the other sub-models. Even though the sensitivity analysis is conducted using data from all municipalities combined, it is reasonable to believe that the two mentioned parameters are the ones to focus on in a possible calibration of the models. This will require further investigation before a potential tuning is conducted.

For the SVM models, no structured sensitivity analysis was conducted, mainly as SVMs are known to be very sensitive to changes in the parameters (Sadrfaridpour et al., 2019), and adjustments of the parameters during the setup of the models showed significant impact on the predictions. In total, 9 models were tuned, including 5 global and 4 local models. Only 4 local models were created and tuned, as the dataset from municipality 2 lack pipes in condition class 4 and 5, i.e. no bad pipes. Still, this dataset is used for training global model 1, 3, 4, and 5, and also global model 2 is used for predicting the condition classes in municipality 2. As the condition classes are imbalanced, which is the case for all municipalities (see Table 4), each condition class should be assigned asymmetric weights to penalize misclassification of the minority class (Caradot et al., 2018). In the study of Hernández et al. (2021), the weights for the condition classes were tuned together with the C and γ parameters. Another approach available in the Scikit-learn module is to balance the weights, such that the sum of all weights is equal to the total number of samples in the training data (Pedregosa et al., 2011). The latter approach was chosen in this study, both do to shorter computational time, and minor effects on the performance metrics when doing small tests with both balanced weights and adjusting them manually. The performance metric used for tuning was the K_pipe metric described by Caradot et al. (2018). The metric was expanded to cover the five class classification system, in total 25 K values were created and used in the numerator of Equation 13. These K values are distributed with 5 for TPR, 10 for FNR and 10 for FPR, covering all cells in a 5x5 confusion matrix. The FNR is computed as the number of pipes predicted in a worse class than observed over the total number in the predicted class, for all worse conditions than the one observed, while FPR is the number of pipes predicted better than observed over the total number in the observed class. The tuned parameters together with the K-pipe value are listed in Table 6.

Table 6: Parameter sets for the Support Vector Machine models

Model	**C**	γ	**K_pipe**
Local 1	162.03	2.11	0.38
Local 3	94.11	6.41	0.31
Local 4	0.50	0.41	0.33
Local 5	395.36	1.44	0.41
Global 1	18.50	0.08	0.39
Global 2	147.07	0.65	0.37
Global 3	397.52	0.29	0.38
Global 4	123.63	1.70	0.39
Global 5	8.46	1.30	0.40

As described earlier the optimal value for K_pipe is 0, equal to a 100% accuracy of the model. The values obtained by the tuned models are similar to the values Hernández et al. (2021) got in Bogotá and Medellin, with a K_pipe of 0.35 and 0.38 respectively. Also, the values obtained are similar to the ones by Caradot et al. (2018) on sewer data from Berlin. Here a Random Forest model got a K_pipe of 0.34, while the GompitZ model got a K_pipe of 0.51.

3.2 Model Transferability

Figure 7 shows the comparison between the C-indices obtained by the local- and global RSF model for each municipality, where the local model corresponds to a model trained and tested on data from the given municipality, and the global model is trained on the four other datasets, and tested on the given municipality. Each model contains four sub models, corresponding to the labels on the x-axis, namely the transition states between condition classes. The boxplot distributions are obtained by using different subsets for the training and testing of the models by setting the random state parameter to None, and sampling it for 15 times. For the local models, 80% of the data were used for training and 20% for testing, while for the global models, 80% of the data from the 4 combined datasets (leave out the one to predict) were used for training, and 80% of the data from the left out data were used for testing. As the data from municipality 2 only contained good pipes, i.e. CC1-3 (as seen in Table 4), no local model was made, but the global model was used to predict on the local data.

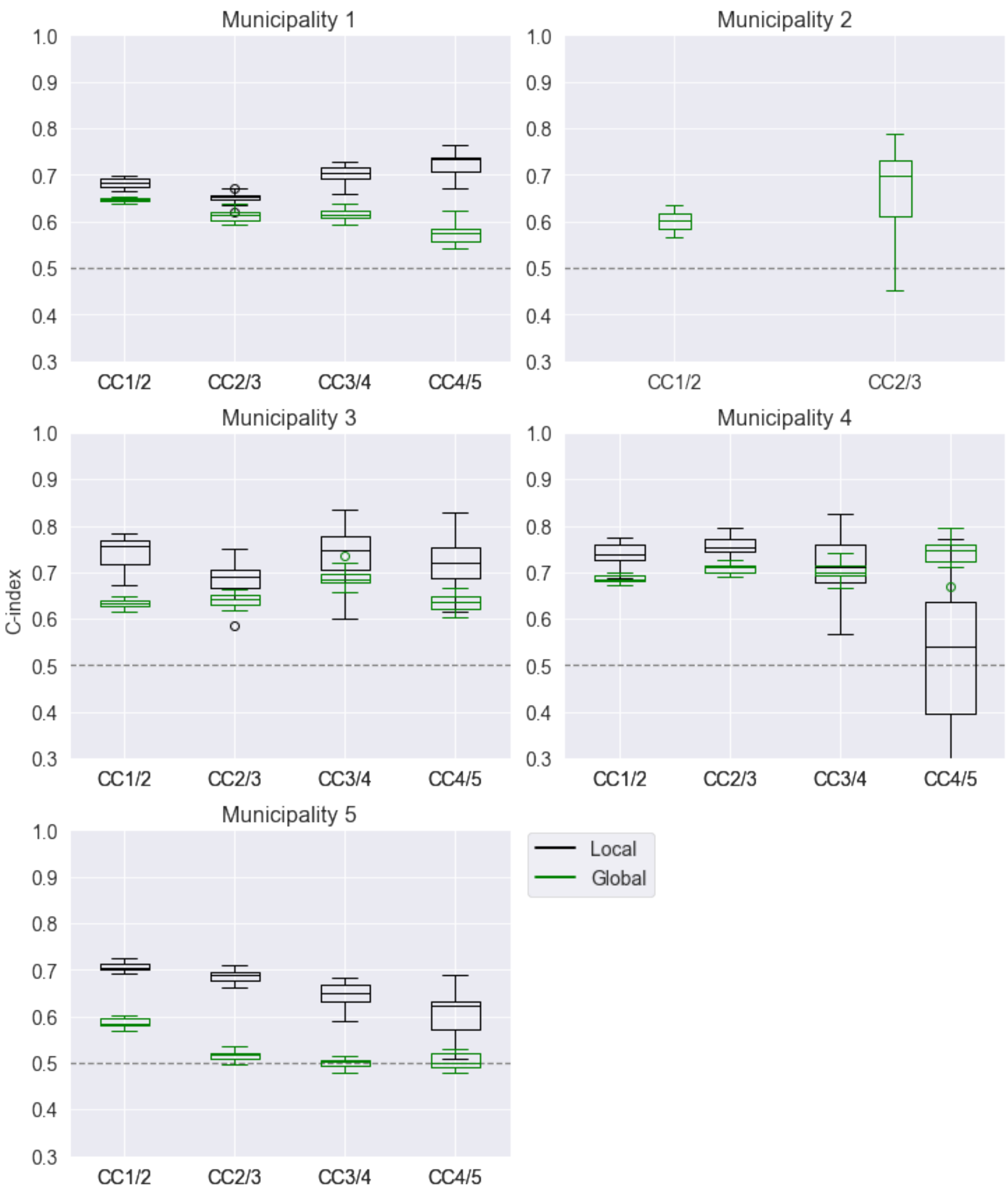

Figure 7: Sampled Concordance Index for all models. The gray dashed line marks the threshold above which the model is better than guessing.

In general, all local models perform quite good, with a C-index above 0.7 for the majority of the models. The biggest variability in C-index are the models for the bad pipe transitions, i.e. CC3/4 and CC4/5. This makes sense as the available data for the pipes who has experienced the transition to a higher class gets more sparse (according to Table 4). The trend of the global models is a lower predictive power than the local models, except for municipality 4, where the CC4/5 sub-model performs a significant amount better than the local model. Also, the local sub-model CC4/5 for municipality 4 shows a significantly higher spread in the sampled C-index, with a median value just above 0.5. The significant variation in the CC4/5 model for municipality 4 can be explained by studying the feature importance, which is done in Section 3.4.

Looking at the C-indices for municipality 5, the local model outperforms the global model to a higher degree than what is the case for the other municipalities. The local models has a step-wise decrease in the C-index when the condition class increase, which is not seen to the same extent in the other municipalities. The global model on the other hand performs reasonable for predicting pipes in CC1 or worse with a C-index around 0.6, but for the other sub-models, it does not perform better than a random model (C-index = 0.5). A reason behind this could be the difference in climate type and geography in the municipalities. Municipality 1-4 are more similar in terms of climate, especially regarding the mean annual temperature and geographical location, which again could affect typical deterioration factors, such as loads (due to increased frost depth), soil type or groundwater level. In the study by Laakso et al. (2019), the RSF model was compared to a Weibull regression model for a binary sewer condition case, giving a really good C-index for both of the models (above 0.8) on pipes younger than around 35 years old, with the RSF model being slightly better than the other. From the age of 35 years to about 55 years, the C-index varies significantly from 0.7 down to almost 0.5. The data used in their study only contained pipes with an age between 0 and 55 years, which could be an important reason for the in general higher C-index obtained. In Figure 8 the survival curves for each municipality, displaying both the curves from the local model and the global model for each transition state. The survival curves displays the probability of surviving as time goes by, in this case the probability of staying in a condition class or transferring to a worse condition class.

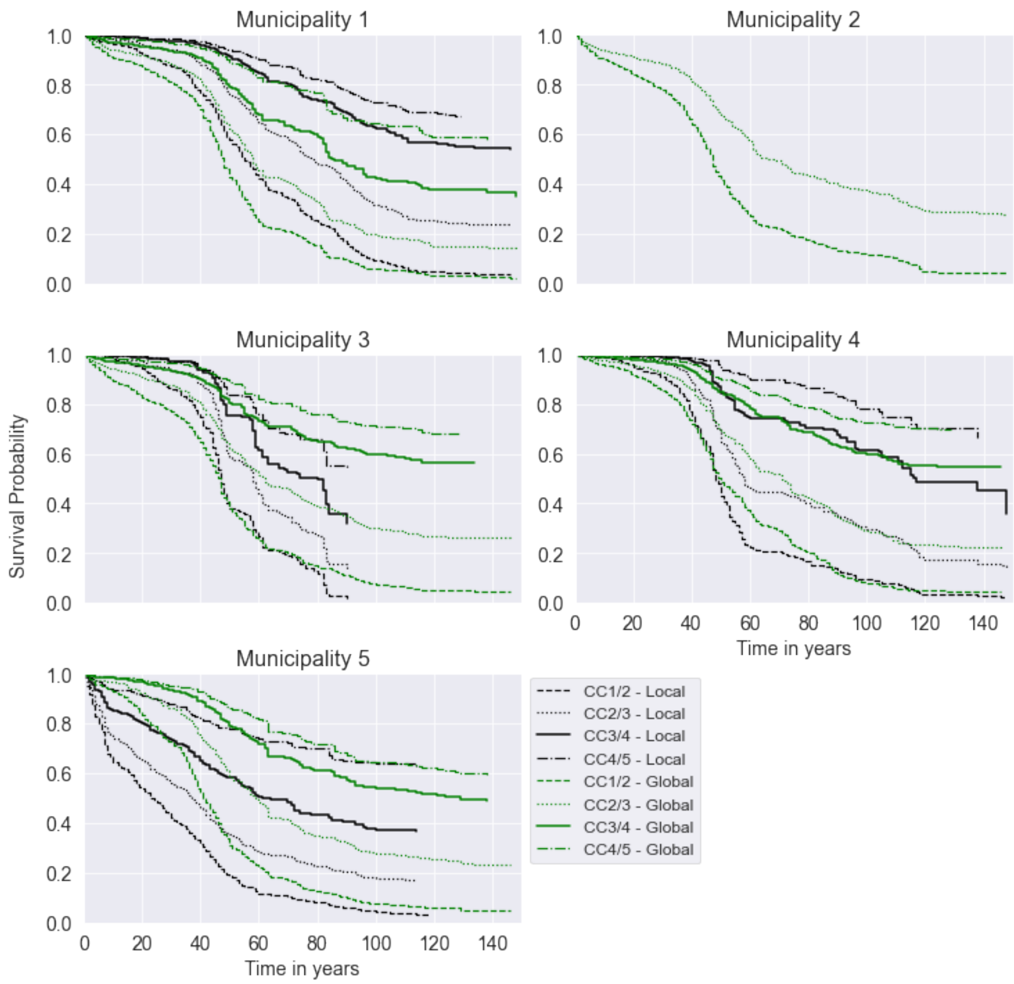

Figure 8: Comparison of survival curves generated from local model and global model. The transition between good and bad pipes, i.e. CC3/4. are displayed with a thicker line.

Looking into what impact the difference in C-indices in Figure 7 has on the survival curves will be of interest, to see if the global model is too optimistic or pessimistic in its predictions. The curves in Figure 8 is the averaged curves from the local- and global models. This is done by creating survival curves for each pipe and taking the averaged value for each time step. Normally, survival curves are divided into cohorts, typically based on materials (e.g. Rokstad and Ugarelli, 2015), but here the averaged curves for all pipes are used. One can argue that this does not necessarily impact the resulting curves to a large extent, as most of the pipes in the datasets are concrete pipes. Focusing on the transition between good and bad pipes, i.e. the CC3/4 curves, there are some interesting points to investigate for the different municipalities. As municipality 2 do not have a CC3/4 transition curve, it will not be discussed in the following. For municipality 1 the global model underestimates the survival in condition class 3 or better compared to the local model, almost up to 20% lower probability for pipes over 60 years. For municipality 3 and 5, the trend is a significantly more optimistic curve by the global model, also with a deviation of around 20% in general, especially for pipes older than 60 years. For municipality 4 on the other hand, both the curves from the global model and local model follows more or less the same path, which makes sense given the C-indices in Figure 7, where both models has more or less similar C-index for the CC3/4 sub-model. In general, the most reliable survival curve will probably be somewhere between the predictions of the local and global model. For all municipalities, the local model curves has a quite broad confidence interval for the 95% percentile (see Figure A.1), and the span between the most optimistic (upper bound) and pessimistic (lower bound) predictions are significant. Which curve to use for asset management decisions should preferably be based on the potential effects of asset failure, and dividing the curves further into smaller cohorts, e.g. material and diameter, will probably reduce the width of the confidence interval. One key point to address when using a global model for predicting local data, is the possibility to predict further than the timespan of the inspected pipes in the municipality. The RSF models are only able to make predictions up until the age of the oldest pipe in the dataset (Laakso et al., 2019), as it models survival using the Nelson-Aalen estimator (Equation 2). For example, the local model of municipality 3 is only able to predict the survival probability up until about 90 years, but the global model, which is trained on older pipes, are able to predict further. Laakso et al. (2019) showed that statistical models has their strength over the RSF in predicting beyond the last observation. One can argue therefore argue that using data from representative municipalities to train a machine learning model is beneficial for predictions beyond the ones a municipality is able to model itself. In general, the models studied are better than guessing, but the municipalities used should be representative in order to increase the predictive power.

When studying the transferability of the SVM models, a similar approach as the one for RSF were used. The local models were trained on 80% of the data for the specific municipality, and tested on the remaining 20%. The global models were trained on 80% of the data from 4 municipalities, and tested on 80% of the data on the left out municipality when computing the ROC-AUC scores. For the confusion matrices, the same 20% are used for testing from the given municipality. In general, the trend is that the models have difficulties in predicting correct condition classes out of the five available. The AUC value for a condition class is a similar measure as the C-index for the RSF models, as it measures the ranking ability of the model. Similar to the C-index, an AUC of 0.5 is just as good as guessing, while a value of 1 indicates a perfect model. The span of AUC scores for each model is significant, ranging from just above 0.40 to 0.87 (see Figure B.1). One method to quantify the overall performance of the model, is using the micro-averaged ROC curve, with its corresponding AUC score (Pedregosa et al., 2011). The TPR and FPR are calculated as the sum or TPs and sum FPs over all classes, over the sum of TP and FN, and FP and TN respectively. The resulting curves are shown in Figure 9.

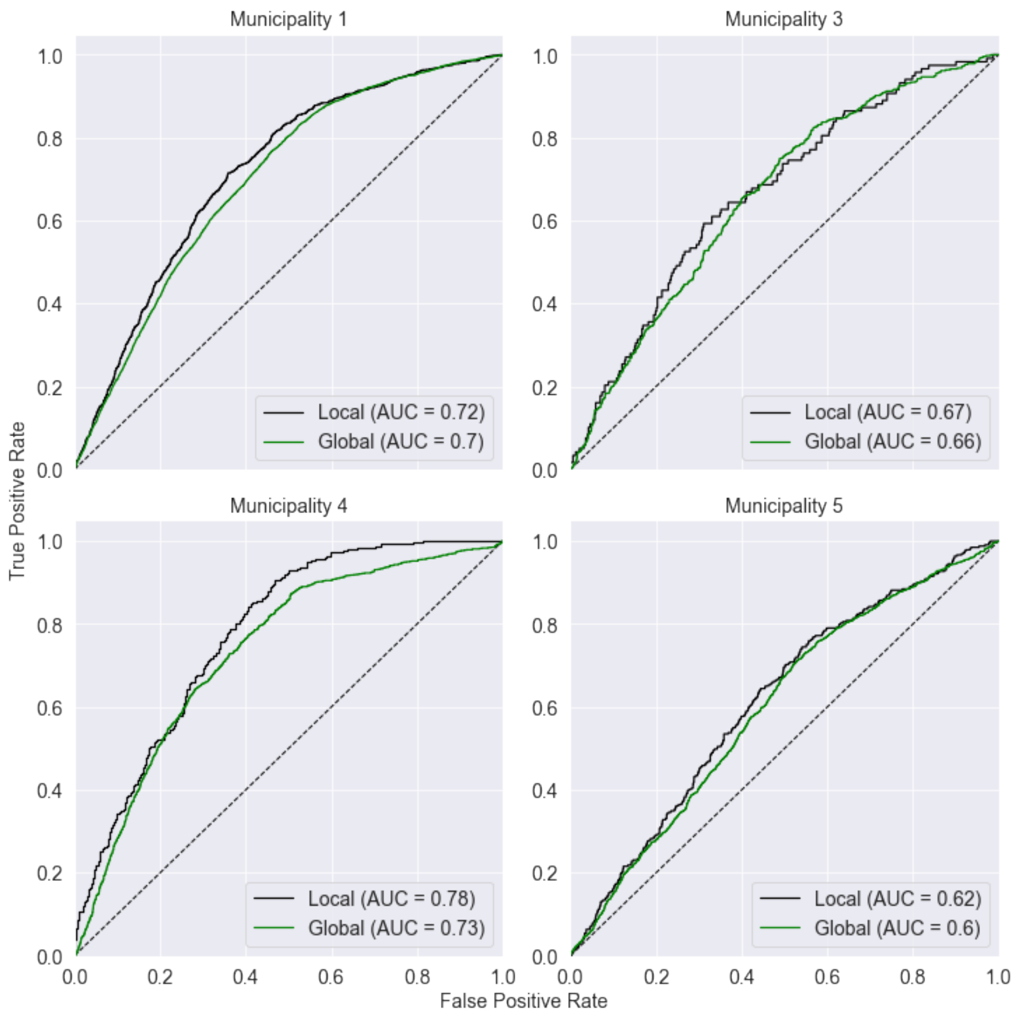

Figure 9: Comparison of micro averaged ROC curves for local and global model. Municipality 2 is left out as the as it lack pipes in condition class 4 and 5

Aggregated over each class, the models perform reasonably good in terms of the AUC score, where the local and global model for each case yields a similar score. Comparing it to the results from the RSF model (Figure 7), the trend in the different cases is similar. Municipality 4 performs best in terms of AUC, while municipality 5 performs worst. The trend is also that the local model is slightly better than the global one in all the cases.

The amount of classes used for classifying sewer pipes has shown to have a significant impact on the resulting performance of classifiers in previous studies (e.g. Mashford et al., 2011; Harvey and McBean, 2014b; Caradot et al., 2018). Mashford et al. (2011) showed that using five condition classes can give high accuracy, but this is often due to imbalanced dataset, typically a much higher amount of the good pipes than bad pipes. A more common approach is either a three class system, where pipes are grouped as good, medium or bad (e.g. Caradot et al., 2018; Hernández et al., 2021), or a binary system with good or bad pipes (e.g. Nguyen et al., 2022; Harvey and McBean, 2014b). To look into the effect of classifying with five classes or using binary classes, municipality 4 is used as a case study further on, as it is shown to yield the best results both for the SVM transferability, but also for the RSF transferability. In Table 7, the predictions are shown with the five class system, both with the local model and the global model.

Table 7: Five class confusion Matrix for SVM predictions by local (top) and global (bottom) model in municipality 4

	Predicted Condition (Local)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	37	19	5	6	6
	CC2	29	49	11	10	12
	CC3	3	15	10	8	4
	CC4	2	3	0	3	4
	CC5	0	1	2	3	9
	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	27	15	13	8	10
	CC2	30	30	27	11	13
	CC3	10	7	6	8	9
	CC4	4	6	1	1	0
	CC5	4	4	2	3	2

These predictions correspond to an accuracy by the local model of 43.0% and the by the global model of 26.3%. In general, the global models gets an accuracy between 20.0% and 26.0%, while the local models ranges between 24.0% to 43.0%. See Appendix C for the confusions matrices for all municipalities. By aggregating the predictions into binary classes, where good pipes are CC1-3 and bad pipes are CC4-5, the confusion matrix in Table 8 is obtained.

Table 8: Binary confusion matrix for SVM predictions by local (top) and global (bottom) model in municipality 4

	Predicted Condition (Local)		
		Good (CC1-3)	Bad (CC4-5)
Actual Condition	Good (CC1-3)	178	46
	Bad (CC4-5)	8	19
	Predicted Condition (Global)		
		Good (CC1-3)	Bad (CC4-5)
Actual Condition	Good (CC1-3)	165	59
	Bad (CC4-5)	21	6

Here the accuracy of the local model is 78.5% and the global model 68.1%. The study by Nguyen et al. (2022) used a binary classification system when comparing 17 machine learning algorithms in Ålesund, Norway. The data used is of similar size as the one used in municipality 4 in this study, and the accuracy of the models ranges from 67% to 78%, where the SVM model achieved a score of 74%. Similarly, Harvey and McBean (2014a) got accuracy of 58% and 89% with different probability cutoffs for a SVM model, while a Decision Tree model gave 76% and 89% with different cutoffs. Harvey and McBean (2014b) used a Random Forest model with different cutoffs, giving accuracy of 72%, 74% and 89%. In general, the mentioned studies utilize more explanatory variables than used in this study, showing that by using only the most basic pipe characteristics the predictive power of the model can still be reasonably good. Additionally, reducing the amount of condition classes from five to three or two, seems to be the most reasonable in terms of finding the most critical transition, namely between good or bad.

3.3 Model Comparability

To benchmark the predicted survival curves for the four transition states of the RSF model, a comparison between the RSF survival curves and GompitZ survival curves are conducted on municipality 4. Municipality 4 is chosen as it yielded the overall best predictions in the transferability study, overall replicating the global model best. As the amount of pipes in material cohort PVC and Other was sparse, the GompitZ calibration for these cohorts failed to converge, resulting in insignificant alpha and beta parameters (see Equation 15). Therefore the comparison is conducted only on the concrete survival curves from both models. Several studies has compared the output from machine learning models and statistical models, Caradot et al. (2018) and Rokstad and Ugarelli (2015) both compared a Random Forest classifier with the GompitZ model, while Laakso et al. (2019) compared survival curves from a RSF model, a Weibull model and a Kaplan-Meier curve.

The RSF model was trained and GompitZ was calibrated on the same 80% of data from municipality 4, while the remaining 20% was used for validation. The calibration process of GompitZ was done using the gompcal-module from Le Gat (2011) to estimate the α- and β parameters in Equation 15. Two cases are investigated, using a model without covariates, and a model with covariates. The covariates in GompitZ can be chosen to be time-dependent, time-independent or both. By testing different combinations of dependency of the covariates dimension, sewer type and length, a model where only length was used as a time-dependent covariate gave the best results in terms of parameter significance and log-likelihood value. From a physical and realistic point of view a time-dependent length does not make sense, but this is reviewed in detail in section 3.4. Laakso et al. (2019) used two Weibull curves, one without covariates and one with, where the former was used as an optimistic life span curve and the latter was used as a pessimistic life span curve. A similar approach is therefore used her. Further, Equation 15 to Equation 19 are written in Python to create survival curves for both cases for all transition states. As the curves without explanatory variables only depend on time, the survival probabilities were computed for the timespan given in Figure 10. With explanatory variables, curves for all pipes in the the validation set were computed by taking the average for all time-steps to obtain the final curve.

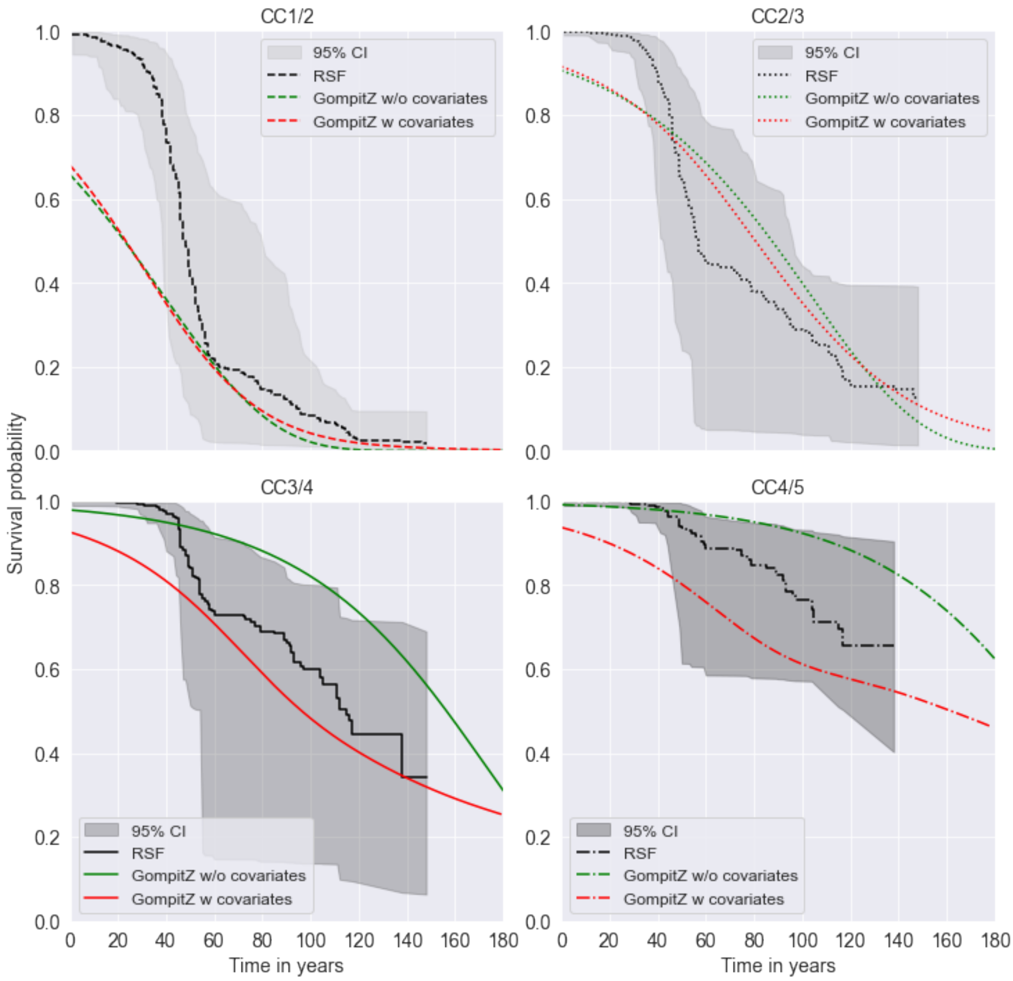

Figure 10: Comparison of survival curves generated from GompitZ and RSF for each transition state in municipality 4. The shaded area displays the 95% confidence interval for the RSF predictions. GompitZ uses covariates as notation for explanatory variables

The GompitZ survival curves tend to follow similar patterns as the survival curves obtained by the RSF model as seen in Figure 10. The tendency for the GompitZ curves is a higher deviation with increasing condition class between the optimistic and pessimistic curves, which probably is due to increased variable importance for increasing condition class. The major difference between the curves from GompitZ and the curves from RSF is the starting probability for the different transitions. For the CC1/2 transition, the probability of being in condition class 1 is approximately 65% by GompitZ, which increases up to between 90-100% for the higher class transitions depending on if it is the optimistic or pessimistic curve. RSF on the other hand always starts of at 100% certainty to be in class 1, 2, 3 or 4 for each transition curve. In practice the GompitZ is giving newly installed pipes, which usually are in condition class 1, a quite high probability of being in a higher condition class. Survival curves obtained on the sewer system of Dresden, Germany by Le Gat (2008) shows a probability of 50% that a newly constructed pipes is in the best condition class when it was constructed. Similarly, Caradot et al. (2018) survival curves from different pipes in different districts of Berlin, Germany shows a probability between 60-90% of being in the best condition class just as the pipe is constructed. As for the probabilities in Figure 10, a reason for the reduced starting probabilities could be that the data used contain several pipes inspected at a very young age to be in higher condition classes than CC1 at the time of inspection, and few pipes classified young in CC1. As the survival probabilities in the RSF model are computed by non-parametric methods (i.e. Nelson-Aalen estimator), the starting probability will always be 100% to be in CC1-4 depending on the transition state we look at, but the gradient of the curve will be affected by the young pipes being inspected in a higher condition class than CC1. One could argue that the curves from GompitZ and RSF would be more similar with less young pipes inspected in a higher condition class than CC1, and a possible solution could be reducing the number of condition classes, which could lead to reduced impact of young pipes inspected higher than CC1. Still, by looking at the curves from the transition CC3/4, i.e. the good and bad threshold, GompitZ still does not have a 100% probability of starting as a good pipe. Looking at the dataset used for modelling, very young pipes have been classified as being in bad condition (See Appendix. E), which will affect the starting probability here also.

The Cohort Survival Model (Baur & Herz, 2002), utilize the Herz distribution (Herz, 1996), where the survival probability is dependent on an ageing parameter, a transition parameter and a resistance time parameter. The resistance parameter can be thought of as a tool to reduce the impact on potentially high condition classes for young pipes, as it defines a timespan where the cohort should stay in a specific condition class. Still, the Cohort Survival model are dependent on extensive dataset, where each cohort should be small enough to be considered homogeneous, but also large enough to yield statistically significance (Kleiner et al., 2007). One could also argue that using parameters such as the resistance parameter to make the model more trustworthy is wrong, as the probable reason for a low starting probability could be young pipes in bad condition. The main advantage of GompitZ as a statistical model is similar as the findings by Laakso et al. (2019) that it is able to predict beyond the age of the oldest pipe, but the uncertainty of these predictions will most certainly become quite much higher than what is the case for the pipes who are inspected. The RSF model, who is purely data-driven, could be more reasonable if the data used is highly affected by young pipes being in condition classes above CC1, as the starting probabilities in theory are closer to the real situation than the predictions by GompitZ. Still. the GompitZ survival curves, both the optimistic and the pessimistic, are reasonably placed within the 95% confidence interval of the RSF curves, especially for the higher condition classes (i.e. CC3/4 and CC4/5). For the CC1/2 and CC2/3 transitions, GompitZ tend to follow a similar pattern as the RSF after the age of 50, but prior to this age the survival is highly affected by the reduced starting probability.

To study the prediction of RSF on the pipe level, the predicted condition on a binary scale (i.e. good or bad) is compared to the predictions obtained by the SVM model. The models trained on municipality 4 are used as a case also for this. As a measure of performance, the previously used K_pipe parameter is used to evaluate the predictions from the RSF model, as this was shown to yield more reasonable results than just maximizing accuracy, or minimizing the FNR or FPR measures. In Table 9 the predictions of the SVM model are displayed together with the predictions from the RSF model with two different probability cutoffs. The accuracy and K_pipe parameter for all three cases are displayed below each respective case.

Table 9: Confusion Matrices for SVM, and RSF with different probability cutoffs

SVM		Predicted Condition	
		Good	Bad
Actual Condition	Good	178	46
	Bad	8	19
Accuracy = 78.5%			
$K_{pipe} = 0.240$			

RSF with probability cutoff = 0.50		Predicted Condition	
		Good	Bad
Actual Condition	Good	221	3
	Bad	24	3
Accuracy = 89.2%			
$K_{pipe} = 0.513$			

RSF with probability cutoff = 0.88		Predicted Condition	
		Good	Bad
Actual Condition	Good	176	48
	Bad	6	21
Accuracy = 78.5%			
$K_{pipe} = 0.217$			

The RSF predictions are made by using the CC3/4 transition survival curve, as this denotes the threshold between good and bad pipes. The survival curve for each pipe is computed, and the survival probability at time equal to the age of the pipe are used to classify the pipe in a good or bad condition. This prediction is compared to the actual condition in the dataset, and based on this the prediction is denoted as a TP, TN, FP or FN. In the middle confusion matrix in Table 9 the predictions are shown with a probability cutoff of 0.50, meaning that if the survival probability at time equal pipe age is above 0.50, the pipe is predicted as good and below it is predicted as bad. Using a probability cutoff of 0.50 is thought of as intuitive, and are often the standard cutoff used in classifiers such as Random Forest or SVM (e.g. Harvey and McBean, 2014a; Harvey and McBean, 2014b). Nevertheless, this cutoff can be adjusted to maximize the predictive ability of the model, depending on the goal, such as maximizing accuracy, or minimizing FPR or FNR (Laakso et al., 2018). Adjusting the cutoff is also the basis for the ROC-curve, where TPR and FPR are computed with varying cutoff probabilities (Géron, 2019). Laakso et al. (2018) used three different cutoffs on a Random Forest model, where one was 0.50, one was 0.35 (by setting FNR to 0.20) and the last was 0.60. The highest accuracy was obtained by a cutoff of 0.50, but the lowest FNR was with 0.35 and lowest FPR was with 0.60. The trend was that lowering the cutoff resulted in more pipes being estimated in bad condition, decreasing the FNR. Nevertheless, this resulted in an increase of the FPR. The accuracy obtained by the RSF model with a cutoff of 0.50 gives a really good accuracy of 89.2%, but a significant higher K_pipe value than the SVM model. By looking at the confusion matrix for RSF with cutoff equal to 0.50, more of the good pipes are predicted to be good compared to the SVM predictions, but less of the bad pipes are predicted to be bad. Also, the bad pipes predicted to be good is higher for the RSF model. In general, the preferred ability of a deterioration model is to find the bad pipes, as rehabilitation and replacement are more urgent for those. The K_pipe metric is therefore a good metric as it tries to increase the amount of TP and TN, while reducing the amount of FP and FN. Therefore, a new threshold was obtained by minimizing the K_pipe value, displayed by the lower matrix in Table 9, resulting in a probability cutoff equal to 0.88.

By optimizing for K_pipe, the resulting confusion matrix from the RSF model is almost identical to the one by the SVM model. The accuracy is identical as the sum TP+TN is the same for both models, but K_pipe is lower for the RSF predictions, mainly due to an decrease in the FPR. A probability cutoff well above the initial 0.5 is shown to give better predictions in earlier studies. Harvey and McBean (2014a) tested a SVM and a Decision Tree model where the SVM had an optimal cutoff of 0.89 and the Decision Tree had a optimal cutoff of 0.95. The same numbers were obtained by Harvey and McBean (2014b) for a Random Forest model, where the optimal cutoff according to the ROC curve was 0.86. For the record, the mentioned studies uses the bad pipes as positive output and good pipes as negative output when computing, the opposite of what is done in this study. The results is therefore transferred to the same system as used in this study when comparing. In general, the results shows that increasing the cutoff decreases the TPR, as a higher cutoff predicts more of the good pipes to be bad. This results in a higher FNR as this is computed as $1 - TPR$, but the increasing cutoff results in lower FPR. A reduced FPR is preferred over a reduced FNR in most use cases of a deterioration model, as it can be argued that predicting a bad pipe as good is negative from a risk point of view. Even though a standard probability cutoff of 0.5 seems logic by intuition, the results shows that in an asset management use case increasing it up to above 0.8 have been shown to give better predictive performance in representative case studies.

3.4 Feature importance

The feature importance for both RSF and SVM are conducted for each global and local model. The resulting importance for RSF are shown in Figure 11, where each subplot denotes a transition state.

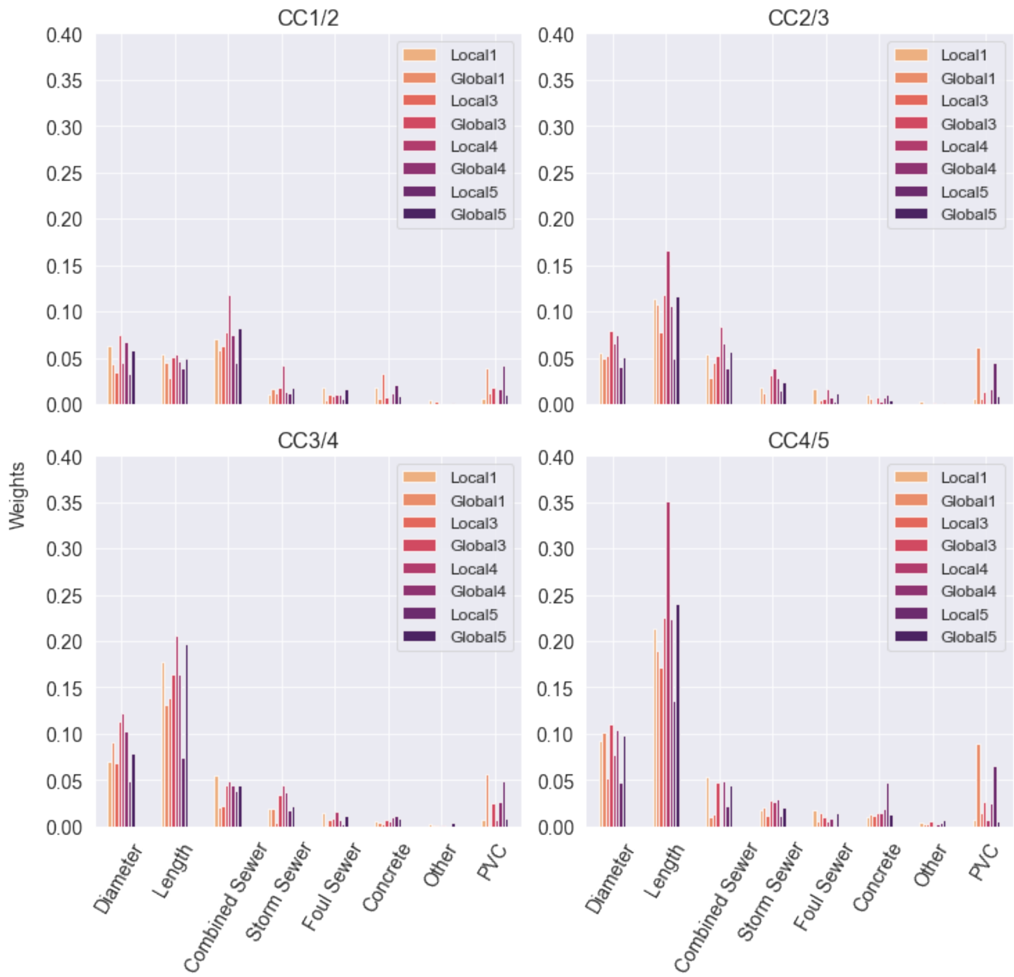

Figure 11: Feature importance for the Random Survival Forest models explanatory variables for the local- and global models for each condition class transition.

The most important features for all the models are shown to be the length of the pipes, getting more important for the high condition class sub-models. The tendency is a greater importance of the length as the condition class increases. This is reasonable, as the amount of data where an event has occurred gets more sparse as the condition class increases. Additionally, the local model for municipality 4 gets a significantly higher importance for the length of the pipes in the CC4/5 than what is the case for the for the global model. This could be an explanation of the significant variation in the C-index for the local model in municipality 4 (see Figure 7). On the other hand, the diameter shows quite similar importance for all transition states, while the sewer type tends to decrease as the condition class increases, especially for the combined sewer. For the record, age is a required parameter in the RSF models, as it models time-to-event, therefore this is not included in the importance. The importance of the features used in the SVM models are shown in Figure 12.

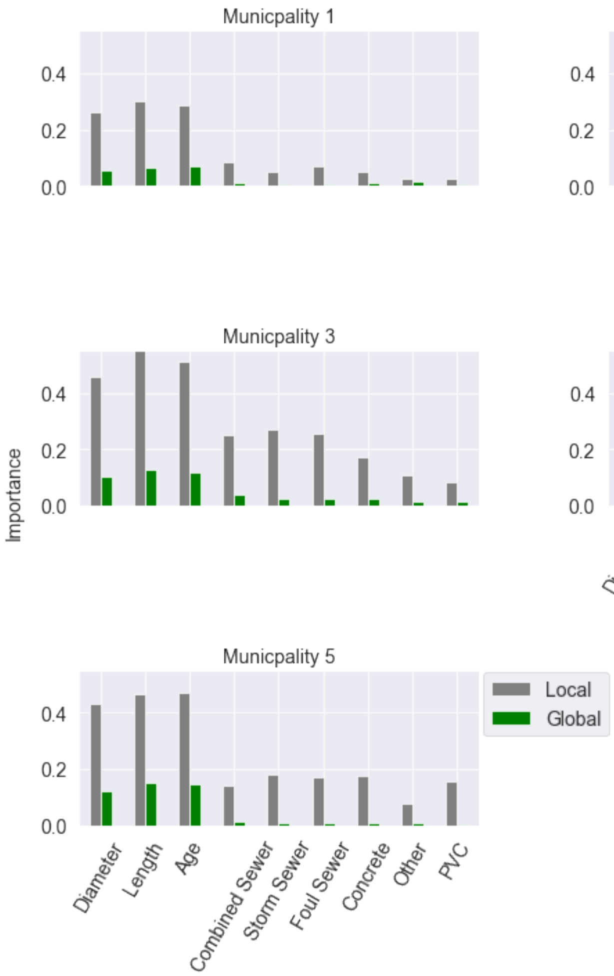

Figure 12: Feature Importance for the Support Vector Machine Models

There is a general tendency that dimension, length and pipe age are the explanatory variables that has the largest effect on the performance of the SVM. Also, the local models tend to have a larger importance of the variables than the global models, except in municipality 4. The pipe age is normally known as one of the most important variables explaining the condition of a sewer pipe (Najafi & Kulandaivel, 2005), but several studies have shown that age is not necessarily the most important factor in all cases (e.g. Nguyen et al., 2022; Laakso et al., 2018). For this study, the three variables mentioned (age, dimension, length) shows more or less similar importance in all of the cases. Several studies has looked into the effect dimension has on the deterioration process, and the results are contradictory, as in some cases increasing dimension results in decreasing deterioration rate and vice versa (Malek Mohammadi et al., 2020). For sewer length, studies have indicated that longer pipes have a tendency of having higher deterioration rate, as the probability of failure is higher in longer pipes. In the study of Nguyen et al. (2022), the material was the most important variable, ahead of the age. Both length and dimension was significantly less important than the other mentioned. In Laakso et al. (2018) the age and length was shown to have equal importance, but well above the importance of the diameter.

The result from the feature importance study of the RSF model (Figure 11) are showing that the sewer length is significantly more important than the other explanatory variables, while for the SVM model (Figure 12) the length shows similar importance as the dimension and age of the pipes, but the length is actually a bit higher for most case studies. A similar effect is seen under the calibration of the GompitZ model for municipality 4, where the set of explanatory variables giving the best results was obtained by using only the sewer length as a time-dependent variable. Still, a time-dependent length does not make sense from a realistic point of view, showing that there could be other reasons behind this effect. In the study of Ana et al. (2009) the significant factors affecting the deterioration of sewer pipes were shown to be age, material and sewer length. The length was marginally significant in this study and the results were showing an odds reduction of being in good condition of 0.60% for each 1 meter increase in length. Several studies have studied the effect of sewer length on the deterioration process, and the main findings in most studies is that longer pipes tends to increase the deterioration (Malek Mohammadi et al., 2020). Ana et al. (2009) argues that the main reason for increased deterioration with increasing length possibly is the presence of more joints in longer pipes. Joints between pipe segments are especially vulnerable to failure and defects, often due to lateral displacements (Malek Mohammadi et al., 2020). Ana et al. (2009) also describes higher vulnerability for blockages and deposition of sediments for longer pipes. The findings of Laakso et al. (2018) shows that pipes longer than 40 meters deteriorates faster than other pipes, which could be explained by structural factors such as higher potential of defects and bending stress. On the other hand, Baik et al. (2006) found that longer runs resulted in less deterioration than shorter ones. They argued that this could be due to fewer bends in the longer runs which results in less accumulation of debris. One could argue that this is case specific, as there is not necessarily less bends in longer pipes.

The coding system used in this study to calculate damage score from the CCTV inspections (Haugen, 2018), considers the length of the pipe as the registered damages are weighted, summed up and divided by the length. The reason is to make the scores relative over the pipe length, but one could argue that it has its effect on distributing the damages over the pipe length, resulting in worse condition of the pipe in total. Especially considering the difference in point damages, such as root intrusion or shifted joints, and longitudinal damages, such as cracks or corrosion, which in terms could have different effect on the total damage score. Another explanation of the importance by length is that it could work as a proxy for other variables that is not available. The explanatory variables used in this study is arguably the most basic variables to describe sewer pipes. In reality, factors regarding the physical properties of a sewer pipes (e.g. material, dimension, length) are not necessarily the most descriptive regarding the deterioration process of a pipe, but serves as proxies for other factors. This could be environmental factors such as groundwater level or soil type, or operational factors such as flow rate (Hawari et al., 2017). Still, studies have shown that such factors are not necessarily more important than the physical factors. Sewage flow was used as a variable by Laakso et al. (2018), showing quite high importance, but was still less important than the pipe slope. Nguyen et al. (2022) used, among others, environmental factors as rainfall, soil type and groundwater table, and even though they showed some importance, they were outcompeted by factors such as age, slope and material. Even though numerous studies using statistical sewer deterioration models or machine learning-based sewer deterioration models have been conducted over the years, the most important variables tend to be the physical factors of the pipes. Despite including far more variables than the five used in this study, the accuracy of the predictions does not get much better than 80% in the best cases, often due to very imbalanced datasets, favouring the good pipes. This is probably also an indication that there are factors that are still not known regarding the deterioration process.

4 Conclusion

In this study the potential of training machine learning-based sewer deterioration models on data from several Norwegian municipalities to address their transferability has been demonstrated. This has been done using two models, namely the machine learning-based survival analysis model RSF, and the classification algorithm SVM. As RSF has seldomly been used in sewer deterioration modelling, the outputs have been compared to other models on the network level and the pipe level. The network level comparison has been conducted using the statistical model GompitZ, while the pipe level comparison is done with the SVM model.

The study shows that a globally trained RSF model in some cases performs almost similar as a locally trained model, when the data being used is from representative municipalities. In general, the models perform reasonable with a C-index above 0.6 for most of the global models, and around 0.7 for the local models, meaning that the models perform better than randomly guessing. The estimated survival curves for each condition class transition of the global models are located reasonable within the 95% confidence interval of the same curves created with the local models. Still, in some of the cases the global model is more optimistic than the local model, while other cases the global model is more pessimistic. Nevertheless, the findings are showing the possibility of using survival curves from a globally trained model as a tool for planning rehabilitation and investments, especially in municipalities lacking data or competence to create their own deterioration models. Furthermore, the SVM models are shown to yield similar overall performance between the global- and the local model for the municipality, but the performance for each of the five condition classes varies significantly. The findings address the problem of low accuracy when using a five-class condition assessment system, but by combining them onto a binary scale, differentiating between good and bad pipes, yields accuracy around 70% in general, similar to previous studies.

Creating survival curves with the GompitZ model for the most representative municipality from the transferability study, gave reasonable results compared to the ones from the RSF model. For each transition state, the RSF survival curve was located between the pessimistic and optimistic curve from GompitZ. Still, one key difference between the models was the starting probability in each state in the GompitZ curves, which started out significantly lower than the ones from RSF. The findings indicate that both models has their advantages and disadvantages, so even with the increasing use of machine learning in deterioration modelling of water infrastructure one should not stop using statistical models, as they often are based on a deep knowledge of the deterioration process. For the predictive power of pipes being in good or bad condition from the RSF model, the study show that a probability threshold of 0.88 gives the best combination between TPR, FNR and FPR. A cutoff below this is shown to give better accuracy, but the ability to predict bad pipes as bad is reduced. These findings indicate that the RSF performs well both regarding network level predictions, and pipe level predictions.

For all models, the feature importance shows that the sewer length is the most important variable in the deterioration models. The increasing amount of pipe joints for longer pipe length, and greater bending stress is discussed as probable causes, together with the nature of the condition assessment system. Still, a probable cause based on the findings from the models applied in this study, is that the length serves as a proxy for some other variable that still haven't been figured out yet.

For the future, increasing the amount of data and including additional municipalities will be needed to further establish the transferability of the models. Additionally, grouping the municipalities based on their similarity, such as climate or geographical location, could be beneficial to increase the predictive power of a global model. Furthermore, reducing the number of condition classes will probably be beneficial when developing sewer deterioration models, as long as they still serve the desired objectives. Lastly, further research on understanding the deterioration process and the variables affecting it is needed to expand the model capability of predictions.

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *6*(4), 701–726. Retrieved March 22, 2023, from https://www.jstor.org/stable/2958850

Almheiri, Z., Meguid, M., & Zayed, T. (2021). Failure modeling of water distribution pipelines using meta-learning algorithms. *Water Research*, *205*. https://doi.org/10.1016/j.watres.2021.117680

Ana, E., Bauwens, W., Pessemier, M., Thoeye, C., Smolders, S., Boonen, I., & De Gueldre, G. (2009). An investigation of the factors influencing sewer structural deterioration [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15730620902810902]. *Urban Water Journal*, *6*(4), 303–312. https://doi.org/10.1080/15730620902810902

Ana, E. V., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15730620903447597]. *Urban Water Journal*, *7*(1), 47–59. https://doi.org/10.1080/15730620903447597

Atambo, D., Najafi, M., & Kaushal, V. (2022). Development and comparison of prediction models for sanitary sewer pipes condition assessment using multinomial logistic regression and artificial neural network. *Sustainability (Switzerland)*, *14*(9). https://doi.org/10.3390/su14095549

Baik, H.-S., Jeong, H., & Abraham, D. (2006). Estimating transition probabilities in markov chain-based deterioration models for management of wastewater systems. *Journal of Water Resources Planning and Management*, *132*(1), 15–24. https://doi.org/10.1061/(ASCE)0733-9496(2006)132:1(15)

Baioletti, M., Di Bari, G., Milani, A., & Poggioni, V. (2020). Differential evolution for neural networks optimization [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Mathematics*, *8*(1), 69. https://doi.org/10.3390/math8010069

Baur, R., & Herz, R. (2002). Selective inspection planning with ageing forecast for sewer types. *Water Science and Technology*, *46*(6), 389–396. https://doi.org/10.2166/wst.2002.0704

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests v3.1*. Retrieved June 4, 2023, from https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

Bruaset, S., Becker, M. A., Reksten, H., & Baade-Mathiesen, T. (2021). *Kommunalt investeringsbehov for vann og avløp 2021-2040*. Norsk Vann.

Butler, D., Digman, C. J., Makropoulos, C., & Davies, J. W. (2018). *Urban drainage* (Fourth). Taylor & Framcos.

Caradot, N., Riechel, M., Fesneau, M., Hernandez, N., Torres, A., Sonnenberg, H., Eckert, E., Lengemann, N., Waschnewski, J., & Rouault, P. (2018). Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in berlin, germany. *Journal of Hydroinformatics*, *20*(5), 1131–1147. https://doi.org/10.2166/hydro.2018.217

Caradot, N., Sonnenberg, H., Sampaio, P., Guilbert, A., Parez, V., & Dimova, V. (2021). Using deterioration modelling to simulate sewer rehabilitation strategy with low data availability. *Water Science and Technology*, *83*(3), 631–640. https://doi.org/10.2166/wst.2020.604

Chen, T.-J., Vladeanu, G., Yazdekhasti, S., & Daly, C. (2022). Performance evaluation of pipe break machine learning models using datasets from multiple utilities. *Journal of Infrastructure Systems*, *28*(2). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000683

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Daulat, S., Rokstad, M. M., Bruaset, S., & Tscheikner-Gratl, F. (2022). One model fits all - on the feneralizability of pipe deterioration modles between utilities. *EGU22-8046*. https://doi.org/https://doi.org/10.5194/egusphere-egu22-8046

Deepa, P., & Gunavathi, C. (2022). A systematic review on machine learning and deep learning techniques in cancer survival prediction. *Progress in Biophysics and Molecular Biology*, *174*, 62–71. https://doi.org/10.1016/j.pbiomolbio.2022.07.004

Dirksen, J., Clemens, F. H., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., Plihal, H., Müller, K., & Snaterse, C. T. (2013). The consistency of visual sewer inspection data [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15732479.2010.541265]. *Structure and Infrastructure Engineering*, *9*(3), 214–228. https://doi.org/10.1080/15732479.2010.541265

Dirksen, J., & Clemens, F. (2008). Probabilistic modeling of sewer deterioration using inspection data. *Water Science and Technology*, *57*(10), 1635–1641. https://doi.org/10.2166/wst.2008.308

Egger, C., Scheidegger, A., Reichert, P., & Maurer, M. (2013). Sewer deterioration modeling with condition data lacking historical records. *Water Research*, *47*(17), 6762–6779. https://doi.org/10.1016/j.watres.2013.09.010

EN 752. (2017). *Drain and sewer systems outside buildings.* (Standard). European Comittee for Standardization. Brussels, Belgium.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In F. Hutter, L. Kotthoff & J. Vanschoren (Eds.), *Automated machine learning: Methods, systems, challenges* (pp. 3–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1

Fugledalen, T., Rokstad, M., & Tscheikner-Gratl, F. (2021). On the influence of input data uncertainty on sewer deterioration models–a case study in norway. *Structure and Infrastructure Engineering*. https://doi.org/10.1080/15732479.2021.1998142

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow* (Second). O'Reilly Media.

Hahn, M., Palmer, R., Merrill, M., & Lukas, A. (2002). Expert system for prioritizing the inspection of sewers: Knowledge base formulation and evaluation. *Journal of Water Resources Planning and Management*, *128*(2), 121–129. https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(121)

Harrell, F., Jr, Califf, R., Pryor, D., Lee, K., & Rosati, R. (1982). Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, *247*(18), 2543–2546. https://doi.org/10.1001/jama.1982.03320430047030

Harvey, R., & McBean, E. (2014a). Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. *Journal of Hydroinformatics*, *16*(6), 1265–1279. https://doi.org/10.2166/hydro.2014.007

Harvey, R., & McBean, E. (2014b). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering*, *41*(4), 294–303. https://doi.org/10.1139/cjce-2013-0431

Haugen, H. J. (2018). *Dataflyt - klassifisering av avløpsledninger*. Norsk Vann.

Hawari, A., Alkadour, F., Elmasry, M., & Zayed, T. (2017). Simulation-based condition assessment model for sewer pipelines. *Journal of Performance of Constructed Facilities*, *31*(1). https://doi.org/10.1061/(ASCE)CF.1943-5509.0000914

Hawari, A., Alkadour, F., Elmasry, M., & Zayed, T. (2020). A state of the art review on condition assessment models developed for sewer pipelines. *Engineering Applications of Artificial Intelligence*, *93*, 103721. https://doi.org/10.1016/j.engappai.2020.103721

Herman, J., & Usher, W. (2017). SALib: An open-source python library for sensitivity analysis. *Journal of Open Source Software*, *2*(9), 97. https://doi.org/10.21105/joss.00097

Hernández, N., Caradot, N., Sonnenberg, H., Rouault, P., & Torres, A. (2021). Optimizing SVM models as predicting tools for sewer pipes conditions in the two main cities in colombia for different sewer asset management purposes. *Structure and Infrastructure Engineering*, *17*(2), 156–169. https://doi.org/10.1080/15732479.2020.1733029

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression.* Wiley Publishing.

Ishwaran, H., Kogalur, U., Blackstone, E., & Lauer, M. (2008). Random survival forests. *Annals of Applied Statistics*, *2*(3), 841–860. https://doi.org/10.1214/08-AOAS169

Jenkins, L., Gokhale, S., & McDonald, M. (2015). Comparison of pipeline failure prediction models for water distribution networks with uncertain and limited data. *Journal of Pipeline Systems Engineering and Practice*, *6*(2). https://doi.org/10.1061/(ASCE)PS.1949-1204.0000181

Jiang, X., Yuan, Y., & Liu, X. (2014). Multivariate probabilistic modelling for seepage risk assessment in tunnel segments. *International Journal of Reliability and Safety*, *8*, 228. https://doi.org/10.1504/IJRS.2014.069507

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/23789689.2019.1600960]. *Sustainable and Resilient Infrastructure*, *5*(6), 365–377. https://doi.org/10.1080/23789689.2019.1600960

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Kim, D., Lee, S., Kwon, S., Nam, W., Cha, I.-H., & Kim, H. (2019). Deep learning-based survival prediction of oral cancer patients. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-43372-7

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer. https://doi.org/10.1007/b97377

Kleinbaum, D. G., & Klein, M. (2012). *Survival analysis: A self.learning text* (Third). Springer.

Kleiner, Y., Rajani, B., & Wang, S. (2007). Consideration of static and dynamic effects to plan water main renewal.

Kley, G., Kropp, I., Schmidt, T., & Caradot, N. (2013). *Review of available technologies and methodologies for sewer condition evaluation* (No. D 1.1). Kompetenzzentrum Wasser Berlin. Berlin.

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kvamme, H., Borgan, O., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*.

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer condition prediction and analysis of explanatory factors. *Water (Switzerland)*, *10*(9). https://doi.org/10.3390/w10091239

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2019). Sewer life span prediction: Comparison of methods and assessment of the sample impact on the results. *Water (Switzerland)*, *11*(12). https://doi.org/10.3390/W11122657

Le Gat, Y. (2008). Modelling the deterioration process of drainage pipelines. *Urban Water Journal*, *5*(2), 97–106. https://doi.org/10.1080/15730620801939398

Le Gat, Y. (2011). Modelling the degradation of sewer pipelines in CaseS: The GompitZ tool user's guide - version 2.08.

Li, Y.-Y., Xu, Z.-M., Feng, C., & Jiang, Y. (2022). XGBoost-based survival analysis in business risk prediction, 320–324. https://doi.org/10.1109/HDIS56859.2022.9991448

Malek Mohammadi, M., Najafi, M., Kermanshachi, S., Kaushal, V., & Serajiantehrani, R. (2020). Factors influencing the condition of sewer pipes: State-of-the-art review [Publisher: American Society of Civil Engineers]. *Journal of Pipeline Systems Engineering and Practice*, *11*(4), 03120002. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000483

Mashford, J., Marlow, D., Tran, D., & May, R. (2011). Prediction of sewer condition grade using support vector machines. *Journal of Computing in Civil Engineering*, *25*(4), 283–290. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000089

Meijer, D., Scholten, L., Clemens, F., & Knobbe, A. (2019). A defect classification methodology for sewer image sets with convolutional neural networks. *Automation in Construction*, *104*, 281–298. https://doi.org/10.1016/j.autcon.2019.04.013

Mills, M. (2010, December 21). *Introducing survival and event history analysis* [Google-Books-ID: eL6ABljZpmMC]. SAGE.

Močkus, J. (1975). On bayesian methods for seeking the extremum. In G. I. Marchuk (Ed.), *Optimization techniques IFIP technical conference novosibirsk, july 1–7, 1974* (pp. 400–404). Springer. https://doi.org/10.1007/3-540-07165-2_55

Moncada-Torres, A., van Maaren, M., Hendriks, M., Siesling, S., & Geleijnse, G. (2021). Explainable machine learning can outperform cox regression predictions and provide insights in

breast cancer survival. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86327-7
Najafi, M., & Kulandaivel, G. (2005). Pipeline condition prediction using neural network models, 767–781. https://doi.org/10.1061/40800(180)61
Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945–966. https://doi.org/10.1080/00401706.1972.10488991
Nguyen, L. V., Bui, D. T., & Seidu, R. (2022). Comparison of machine learning techniques for condition assessment of sewer network [Conference Name: IEEE Access]. *IEEE Access*, *10*, 124238–124258. https://doi.org/10.1109/ACCESS.2022.3222823
Nikbakht, S., Anitescu, C., & Rabczuk, T. (2021). Optimizing the neural network hyperparameters utilizing genetic algorithm. *Journal of Zhejiang University-SCIENCE A*, *22*(6), 407–426. https://doi.org/10.1631/jzus.A2000384
Park, S., Kim, J., Newland, A., Kim, B., & Jun, H. (2008). Survival analysis of water distribution pipe failure data using the proportional hazards model. *316*. https://doi.org/10.1061/40976(316)500
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Pölsterl, S. (2020). Scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, *21*.
Price, K., Storn, R. M., & Lampinen, J. A. (2005). *Differential evolution: A practical approach to global optimization* (2005th edition). Springer.
Rajani, B., & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: Physically based models. *Urban Water*, *3*(3), 151–164. https://doi.org/10.1016/S1462-0758(01)00032-2
RIF. (2021). *State of the nation - norges tilstand: Vannforsyning- og avløpsanlegg*. Rådgivende Ingeniørers Forening. Oslo, Norway.
Roghani, B., Cherqui, F., Ahmadi, M., Le Gauffre, P., & Tabesh, M. (2019). Dealing with uncertainty in sewer condition assessment: Impact on inspection programs. *Automation in Construction*, *103*, 117–126. https://doi.org/10.1016/j.autcon.2019.03.012
Rokstad, M., Gat, Y. L., & Ugarelli, R. (2014). Assessment of the sensitivity of GompitZ. *International Conference on Hydroinformatics*. https://academicworks.cuny.edu/cc_conf_hic/44
Rokstad, M., & Ugarelli, R. (2015). Evaluating the role of deterioration models for condition assessment of sewers. *Journal of Hydroinformatics*, *17*(5), 789–804. https://doi.org/10.2166/hydro.2015.122
Røstum, J. (2000). *Statistical modelling of pipe failures in water networks* (Doctoral Thesis). Norwegian University of Science and Technology. Retrieved March 16, 2023, from https://core.ac.uk/reader/30813880
Saadati, M. (2022). Analyzing marriage tolerance by survival machine learning algorithms [ISSN: 0094-243X], *2662*. https://doi.org/10.1063/5.0111306
Sadrfaridpour, E., Razzaghi, T., & Safro, I. (2019). Engineering fast multilevel support vector machines. *Machine Learning*, *108*(11), 1879–1917. https://doi.org/10.1007/s10994-019-05800-7
Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1
Schmid, M., Wright, M., & Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, *63*, 450–459. https://doi.org/10.1016/j.eswa.2016.07.018
Senanayake, S., Barnett, A., Graves, N., Healy, H., Baboolal, K., & Kularatna, S. (2020). Using machine learning techniques to develop risk prediction models to predict graft failure following kidney transplantation: Protocol for a retrospective cohort study. *F1000Research*, *8*. https://doi.org/10.12688/f1000research.20661.2
Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682
Skjelde, J. (2022). *Machine learning-based sewer deterioration modelling - a review of different methods* (Project thesis). NTNU.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*. Retrieved April 22, 2023, from https://proceedings.neurips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, *55*(1), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

Song, H., Xi, L., Shu, Q., Wei, Z., & Qiu, S. (2023). Estimate forest aboveground biomass of mountain by ICESat-2/ATLAS data interacting cokriging [Number: 1 Publisher: Multidisciplinary Digital Publishing Institute]. *Forests*, *14*(1), 13. https://doi.org/10.3390/f14010013

Sousa, V., Matos, J., & Matias, N. (2014). Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Automation in Construction*, *44*, 84–91. https://doi.org/10.1016/j.autcon.2014.04.004

Statistics Norway. (2023). *Kommunalt avløp* [Ssb.no]. Retrieved February 14, 2023, from https://www.ssb.no/natur-og-miljo/vann-og-avlop/statistikk/utslipp-og-rensing-av-kommunalt-avlop

Storn, R., & Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359. https://doi.org/10.1023/A:1008202821328

Syachrani, S., Jeong, H. S. ", & Chung, C. S. (2013). Decision tree–based deterioration model for buried wastewater pipelines [Publisher: American Society of Civil Engineers]. *Journal of Performance of Constructed Facilities*, *27*(5), 633–645. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000349

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J., Ahmadi, M., Langeveld, J., Le Gat, Y., Scholten, L., Roghani, B., Rodríguez, J., Lepot, M., Stegeman, B., Heinrichsen, A., Kropp, I., Kerres, K., Almeida, M., Bach, P., Moy de Vitry, M., Sá Marques, A., ... Clemens, F. (2019). Sewer asset management–state of the art and research needs. *Urban Water Journal*, *16*(9), 662–675. https://doi.org/10.1080/1573062X.2020.1713382

Tscheikner-Gratl, F., Sitzenfrei, R., Rauch, W., & Kleidorfer, M. (2016). Enhancement of limited water supply network data for deterioration modelling and determination of rehabilitation rate [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15732479.2015.1017730]. *Structure and Infrastructure Engineering*, *12*(3), 366–380. https://doi.org/10.1080/15732479.2015.1017730

Tu, J. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9

Ugarelli, R., Venkatesh, G., Brattebø, H., Di Federico, V., & Sægrov, S. (2010). Asset management for urban wastewater pipeline networks. *Journal of Infrastructure Systems*, *16*(2), 112–121. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000011

Ugarelli, R. M.
bibinitperiod. (2008, May 20). *Asset management of wastewater networks* (Doctoral Thesis). Alma Mater Studiorum - Università di Bologna. Retrieved February 25, 2023, from http://amsdottorato.unibo.it/878/

United Nations. (2015). *UN SDGs* [Sdgs.un.org]. Retrieved February 14, 2023, from https://sdgs.un.org/goals/goal6

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, *4*(2), 65–85. https://doi.org/10.1007/BF00175354

Yang, J. (2004). Road crack condition performance modeling using recurrent markov chains and artificial neural networks. *USF Tampa Graduate Theses and Dissertations*. https://digitalcommons.usf.edu/etd/1310

Appendix

A Survival curves

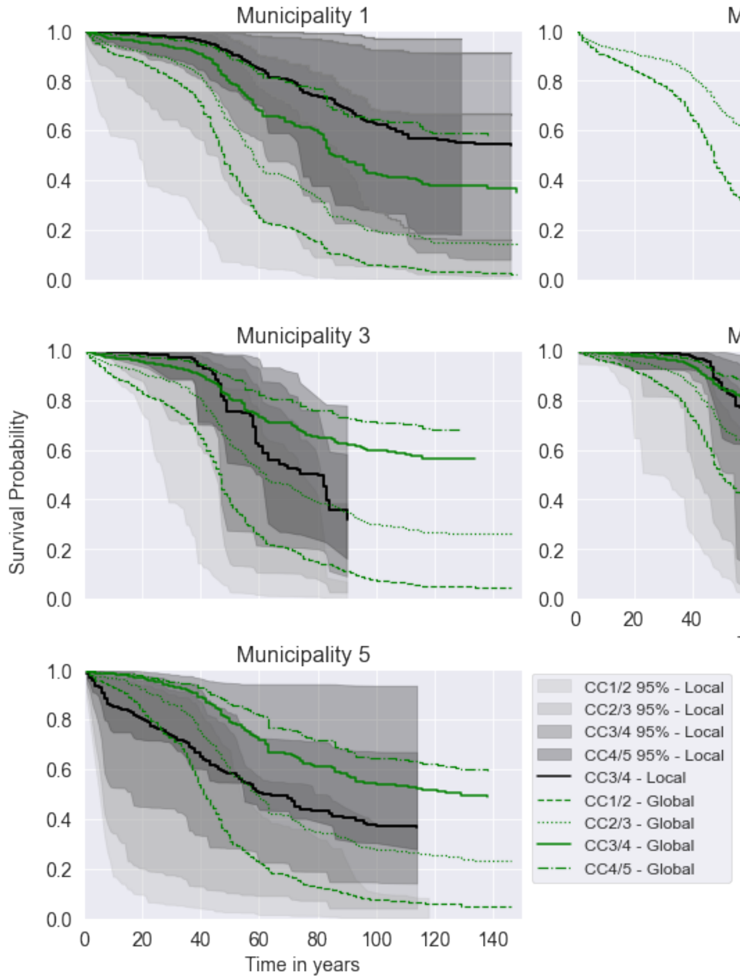

Figure A.1: Comparison of survival curves for the global models and the 95% confidence interval of the local models. The transition between good and bad pipes, i.e. CC3/4, are included from the local model, and are displayed with a thicker, solid line.

B ROC Curves

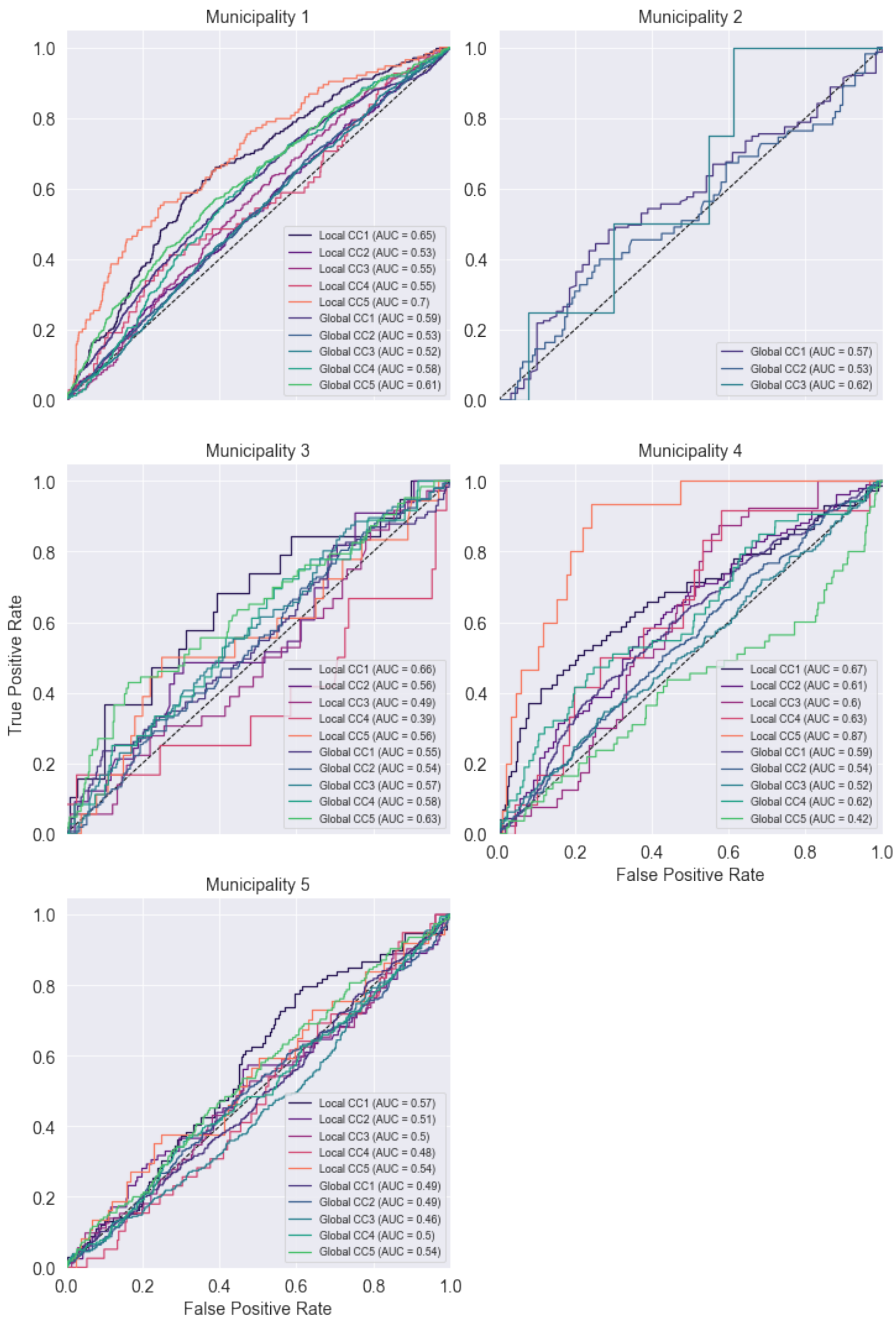

Figure B.1: Comparison of ROC curves for the local and global models for each municipality. The straight dashed line denotes a random model.

C Confusion matrices - Transferability

Table C.1: Five class confusion Matrix for SVM predictions by local (top) and global (bottom) model in municipality 1

	Predicted Condition (Local)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	133	88	70	24	21
	CC2	128	113	112	54	46
	CC3	63	66	94	47	46
	CC4	12	11	24	11	10
	CC5	9	25	27	15	38
	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	116	88	30	42	60
	CC2	135	118	32	58	110
	CC3	77	81	37	36	85
	CC4	19	13	3	9	24
	CC5	12	26	11	22	43
Accuracy local = 30.2%						
Accuracy global = 25.1%						

Table C.2: Five class confusion Matrix for SVM predictions by global model in municipality 2

	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	10	7	9	5	8
	CC2	6	3	1	2	1
	CC3	6	7	6	10	1
	CC4	0	0	0	0	0
	CC5	0	0	0	0	0
Accuracy global = 24.5%						

Table C.3: Five class confusion Matrix for SVM predictions by local (top) and global (bottom) model in municipality 3

	Predicted Condition (Local)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	9	8	2	0	0
	CC2	6	20	5	0	2
	CC3	4	13	12	2	5
	CC4	3	4	2	3	0
	CC5	4	5	1	1	7
	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	7	5	2	1	4
	CC2	6	7	5	10	5
	CC3	6	7	6	10	7
	CC4	1	1	2	5	3
	CC5	2	5	2	4	5
Accuracy local = 43.2%						
Accuracy global = 25.1%						

Table C.4: Five class confusion Matrix for SVM predictions by local (top) and global (bottom) model in municipality 4

	Predicted Condition (Local)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	37	19	5	6	6
	CC2	29	49	11	10	12
	CC3	3	15	10	8	4
	CC4	2	3	0	3	4
	CC5	0	1	2	3	9
	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition	CC1	27	15	13	8	10
	CC2	30	30	27	11	13
	CC3	10	7	6	8	9
	CC4	4	6	1	1	0
	CC5	4	4	2	3	2
Accuracy local = 43.0%						
Accuracy global = 26.3%						

Table C.5: Five class confusion Matrix for SVM predictions by local (top) and global (bottom) model in municipality 5

	Predicted Condition (Local)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition (Local)	CC1	34	32	17	12	4
	CC2	23	22	18	8	11
	CC3	12	26	15	10	9
	CC4	14	5	14	3	3
	CC5	7	10	9	5	6
	Predicted Condition (Global)					
		CC1	CC2	CC3	CC4	CC5
Actual Condition (Global)	CC1	32	16	22	20	9
	CC2	34	10	21	10	7
	CC3	14	15	20	12	11
	CC4	18	5	9	2	5
	CC5	9	9	11	5	3
Accuracy local = 24.3%						
Accuracy global = 20.4%						

D GompitZ Calibration

```
GompitZ v2.08 - Calibration results


*** Stratum BET ***
945 (total weight = 945.00) pipelines described in this stratum

945 (total weight = 945.00) inspections reported:
292 (total weight = 292.00) of which in condition C1
352 (total weight = 352.00) of which in condition C2
198 (total weight = 198.00) of which in condition C3
51 (total weight = 51.00) of which in condition C4
52 (total weight = 52.00) of which in condition C5

Convergence achieved in 36 iterations
Log-Likelihood = -1267.314676

Final Covariance Matrix =
{ +2.317e-002 , +2.970e-002 , +3.804e-002 , +4.311e-002 , -2.246e-002 , +3.081e-003 }
{ +2.970e-002 , +4.831e-002 , +6.679e-002 , +7.901e-002 , -3.052e-002 , -1.899e-003 }
{ +3.804e-002 , +6.679e-002 , +1.128e-001 , +1.355e-001 , -3.950e-002 , -1.076e-002 }
{ +4.311e-002 , +7.901e-002 , +1.355e-001 , +1.820e-001 , -4.482e-002 , -1.721e-002 }
{ -2.246e-002 , -3.052e-002 , -3.950e-002 , -4.482e-002 , +2.466e-002 , -3.777e-003 }
{ +3.081e-003 , -1.899e-003 , -1.076e-002 , -1.721e-002 , -3.777e-003 , +8.635e-003 }

Parameter estimates and Wald Chi2 tests
--------------------------------------------------------------------------------------------
Label                                          Estimate      Std. Error  DF  Chi2        Pr>Chi2
--------------------------------------------------------------------------------------------
Alpha(C1) (vs 0)                               -8.9366e-001 1.5222e-001 1   3.4467e+001 0.000000
Alpha(C2) (vs Alpha(C1))                       -2.3347e+000 2.1979e-001 1   1.7208e+002 0.000000
Alpha(C3) (vs Alpha(C2))                       -3.8695e+000 3.3590e-001 1   8.5492e+001 0.000000
Alpha(C4) (vs Alpha(C3))                       -4.7798e+000 4.2658e-001 1   3.4892e+001 0.000000
T (vs 0)                                       -3.8055e+000 1.5703e-001 1   5.8729e+002 0.000000
Sigma (vs 0)                                   +6.3698e-001 9.2925e-002 1   4.6988e+001 0.000000
--------------------------------------------------------------------------------------------
```

Figure D.1: GompitZ calibration results without covariates

```
GompitZ v2.08 - Calibration results


*** Stratum BET ***
945 (total weight = 945.00) pipelines described in this stratum

945 (total weight = 945.00) inspections reported:
292 (total weight = 292.00) of which in condition C1
352 (total weight = 352.00) of which in condition C2
198 (total weight = 198.00) of which in condition C3
51 (total weight = 51.00) of which in condition C4
52 (total weight = 52.00) of which in condition C5

Convergence achieved in 31 iterations
Log-Likelihood = -1243.254070

Final Covariance Matrix =
{ +2.259e-002 , +2.781e-002 , +3.411e-002 , +3.793e-002 , -1.195e-002 , -1.967e-004 , +2.621e-003 }
{ +2.781e-002 , +4.285e-002 , +5.518e-002 , +6.326e-002 , -1.696e-002 , -2.283e-004 , -2.724e-004 }
{ +3.411e-002 , +5.518e-002 , +8.801e-002 , +1.016e-001 , -2.259e-002 , -2.494e-004 , -5.134e-003 }
{ +3.793e-002 , +6.326e-002 , +1.016e-001 , +1.347e-001 , -2.584e-002 , -2.635e-004 , -8.601e-003 }
{ -1.195e-002 , -1.696e-002 , -2.259e-002 , -2.584e-002 , +1.661e-002 , -1.250e-004 , -1.184e-003 }
{ -1.967e-004 , -2.283e-004 , -2.494e-004 , -2.635e-004 , -1.250e-004 , +8.215e-006 , -3.761e-005 }
{ +2.621e-003 , -2.724e-004 , -5.134e-003 , -8.601e-003 , -1.184e-003 , -3.761e-005 , +5.057e-003 }

Parameter estimates and Wald Chi2 tests
--------------------------------------------------------------------------------------------
Label                                          Estimate      Std. Error  DF  Chi2        Pr>Chi2
--------------------------------------------------------------------------------------------
Alpha(C1) (vs 0)                               -9.7963e-001 1.5029e-001 1   4.2490e+001 0.000000
Alpha(C2) (vs Alpha(C1))                       -2.4478e+000 2.0701e-001 1   2.1958e+002 0.000000
Alpha(C3) (vs Alpha(C2))                       -3.9857e+000 2.9667e-001 1   1.1538e+002 0.000000
Alpha(C4) (vs Alpha(C3))                       -4.8804e+000 3.6700e-001 1   4.0919e+001 0.000000
T (vs 0)                                       -3.0212e+000 1.2890e-001 1   5.4938e+002 0.000000
T*LENGTH (vs 0)                                -1.7249e-002 2.8662e-003 1   3.6219e+001 0.000000
Sigma (vs 0)                                   +5.4691e-001 7.1111e-002 1   5.9151e+001 0.000000
--------------------------------------------------------------------------------------------
```

Figure D.2: GompitZ calibration results with covariates

E Damagescore and Pipe Age

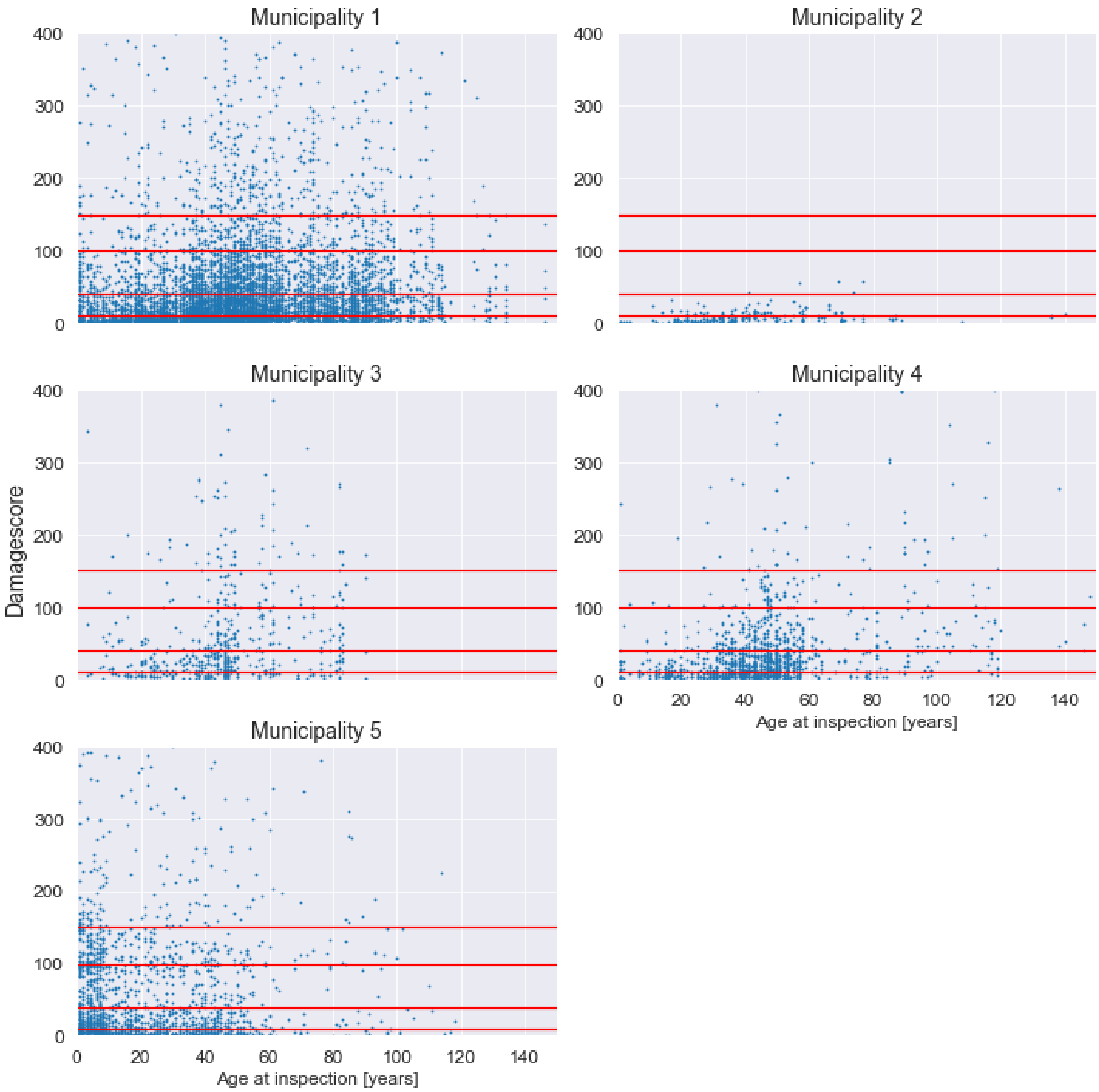

Figure E.1: Damagescore as a function of pipe age. Each red line denotes a condition class transition

F Python Scripts

The Python scripts developed and used in this study is attached as a zipped folder in Inspera. The folder contains a text-file, briefly describing the content of the different scripts.

G Project Thesis

As the methods used in this thesis are based on the work done in the authors project thesis the autumn of 2022, the thesis is to be found in the same folder as the Python scripts. This is to avoid plagiarism, as some of the sentences in the project thesis and the master thesis is quite similar. Also, the project thesis provides a further insight in the choice of the models used in this master thesis.

NTNU

Norwegian University of
Science and Technology