WDSA CCWI 2022

2nd International Joint Conference on Water Distribution
Systems Analysis & Computing and Control in the Water Industry

Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14755

IMPORTANT FACTORS FOR WATER MAIN BREAK PREDICTION ACROSS 13 CANADIAN SYSTEMS

Sadaf Gharaati[1] and Rebecca Dziedzic[2]

[1,2] Concordia University, Department of Building Civil and Environmental Engineering, Montreal, Quebec, Canada

[1] *gharaati.sadaf@gmail.com*, [2] *rebecca.dziedzic@concordia.ca*

Abstract

Water main breaks can jeopardize the safe delivery of clean water and incur significant costs. To mitigate these risks, water main breaks have been predicted through physical and statistical approaches. The latter are less complex and can provide satisfactory results with less data. While many factors can contribute to breaks, the factors applied in previous studies depended on local data availability. Because other studies have focused on a few systems at a time, a broad comparison of factor importance has not been possible. This limits the understanding of the impact of different factors on water main deterioration.

The present study identifies the most important factors driving water main breaks across 13 Canadian water systems. Twenty-eight factors describing physical, historical, protection, environmental and operational attributes were compiled and cleaned. Availability of each attribute differed by system. To evaluate the importance of both numerical and categorical attributes together, two approaches were tested, categorical principal component analysis (CATPCA) and recursive feature elimination with cross-validation (RFECV). The target variable in both cases was set as yearly break status, either broken or non-broken. While CATPCA provides the contribution of each attribute to the target, RFECV provides a tuned predictive model with selected attributes. The RFECV approach was applied with Random Forest and XGBoost models, both types of machine learning models which have been shown to produce accurate results in water main break prediction.

Results from both approaches showed that physical and historical attributes are generally important across all systems. Other types of data, i.e. protection and operational are less available. When protection data is available it was shown to be even more important than physical and historical attributes. Specifically, with CATPCA, lining age and lining material were found to have a higher contribution to break status than pipe age and lining status. With RFECV lining age and lining material were also included in the best models, in particular for systems with greater percentage of lined pipes. These results indicate the choice and timing of lining are key in extending the service life of water mains. Furthermore, this data should be collected if protection practices are in place, to more accurately predict deterioration and future costs.

The results also point to an opportunity to collect more operational data. Among attributes collected by only one utility, pipe pressure, roughness, and dead-end, were found to be important in RFECV. Thus, pipe dissipation and water stagnation could lead to greater pipe deterioration. Further studies are required to quantify the impacts of different pressure ranges and network designs on deterioration.

Keywords
Water main breaks, dimensionality reduction, machine learning, physical, historical, protection, environmental, operational.

1 INTRODUCTION

Water main deterioration is a global challenge that can jeopardize water systems' ability to deliver clean water safely. The failure of water mains can affect individuals, businesses, industries, and institutions. Water main breaks can directly disrupt the service provided by pipes [1]. According to the Canadian Infrastructure Report Card [2], the cost of upgrade and replacement of water and wastewater network in Canada is estimated to be more than CAD$ 80 billion. Hence, it is essential for water utilities to seek cost-effective rehabilitation and renewal strategies[3].

The factors that contribute to water main failure are diverse and collecting all data can be cumbersome. That is why statistical models have been preferred over physical models. However, the factors applied by previous studies differ, based on the availability of data for their given case studies. Accordingly, the present study seeks to identify the most contributing factors to water main breaks across 13 Canadian cities.

2 LITERATURE REVIEW

Physical pipe attributes such as diameter, length and age are widely collected and applied in predictive models. Diameter, length, age, soil type, previous failures, and failure type were consistently applied in historical studies[4]–[6]. Pipe length is identified as an important factor by many authors. However, there isn’t a consensus about whether breakage is positively[7] or negatively [8], [9] correlated with breaks [10]. Break rates also notably differ by material due to their structural resistance and vulnerability to corrosion[11], [12]. Unsurprisingly, protection of the pipe material can also extend service life[1], [12], [13]. While previous break prediction studies included data on lining status and material, the impact of lining age has not been explored.

Barton et.al. [13] note that operational and environmental factors such as sudden changes of temperature, pressure, and soil moisture level, can also increase probability of failure by increasing internal and external stress on the pipes. Martinez et.al [14] accounted for average pressure, in addition to diameter, install year and pipe depth. Snider and McBean [12] found varying break year pattern depended on various factors, most importantly weather. The impact of weather, specifically on soil movement, i.e. freeze-thaw cycles and ground swelling is confirmed by other studies[15]. It is also observed that, pipes are more likely to break once they have broken before [5], [16]–[18]. Previous breaks can be a proxy for local conditions such as soil type, weather conditions, traffic load, etc. However, the importance of these factors is not clear.

In order to identify the most important factors contributing to a target, dimensionality reduction approaches are commonly applied. However, they have not yet been applied to the analysis of water main break contributing factors. There are two general approaches for dimensionality reduction: feature elimination and feature extraction. The first reduces the number of variables by eliminating some, whereas the latter creates new independent variables from combinations of previous independent variables. A useful example of the first type is Recursive Feature Elimination with Cross-Validation (RFECV). RFECV finds the most important factors through a backward elimination process. This approach was initially introduced by Guyon et.al. [19] and is employed along with predictive models, either classification [20] or regression [21]. Previous studies have found a higher performance of RFECV with Random Forest [22], [23] and XGBoost [24]. One well-known feature extraction approach is Principal Component Analysis (PCA) [25]. This method however cannot handle categorical variables. Non-Linear PCA, also known as categorical PCA (CATPCA), is a dimensionality reduction method that, unlike PCA, can handle a non-linear relationship among variables. Categories of variables are replaced with numerical values through optimal scaling.

3 METHODOLOGY

The analysis of important factors driving water main breaks was divided into three key steps, data cleaning, data visualization and analysis, and dimensionality reduction. Each is explained in more detail in the following paragraphs.

3.1 Data preparation

Data from the utilities was provided as separate pipe inventory and historical break datasets. Thus, the first step for analyzing characteristics of broken and non-broken pipes was to merge the data. The datasets for each utility were merged based on unique IDs, identifying each pipe. Next, missing values were filled with three approaches, depending on data availability and type of attribute: 1. assumed value; 2. mirroring attribute; and 3. homogeneous groups. The first was applied for binary attributes with a clear common value. For example, anode status was only collected for pipes with anodes and all missing values were assumed to be related to pipes without protection. In the second method, missing values were replaced based on other attributes with equivalent and more detailed information. For instance, lining status (yes/no) was filled based on values of lining material. If lining material was "unlined", the lining status was set to "no", and "yes" for other actual lining materials. The third approach used clusters of similar pipes to replace missing values. For example, pipes with the same install year were assumed to generally be of the same material. After filling gaps, inconsistencies and outliers were detected and removed from the analysis. Lastly, categorical variables were converted to numerical through optimal scaling in R (optiscale package) for input to the correlation analysis and RFECV.

3.2 Data visualization and analysis

To better understand variations and correlations in the data, multiple graphs were generated and correlation analyses run. Because correlation reveals the relationship between numerical attributes, optimally scaled categorical variables were used. An initial analysis was performed between break status and common attributes across all cities (diameter, age, length and material). Then, a correlation analysis was run for all data for each city, and presented in a boxplot.

3.3 Dimensionality Reduction

Two dimensionality reduction methods were applied to identify the most important factors driving watermain failure, CATPCA and RFECV. The target variable in both cases was set as yearly break status (broken or non-broken). The CATPCA analysis was conducted in R (princals function, Gifi package). The number of PCs selected for each city was determined to account for around 78-85% of variance. Important factors were identified as those with a contribution above a cut-off, calculated as 100% divided by the number of attributes in each utility. The RFECV approach was conducted in python (Scikit-Learn library). Highly correlated predictors (correlation>0.8) were excluded from the analysis. Two types of estimators were employed, random forest and XGBoost. Hyperparameters were tuned for each estimator and each city. Overfitting was checked with 5-fold cross validation. Lastly, to evaluate the effectiveness of dimensionality reduction. the fit of the full data model and reduced data model were compared with F1 score and recall.

4 DATA DESCRIPTION

This study is part of the project "Best Practices for Predicting Water Main Breaks," a collaboration between the Canadian National Water and Wastewater Benchmarking Initiative (NWWBI) and the Concordia University research group "UrbanLinks". Thirteen utilities across Canada, in the provinces of Ontario, Nova Scotia, Newfoundland, Manitoba, Saskatchewan, and British Columbia, shared their water main inventories and historical records of main breaks as spreadsheets or GIS shapefiles. The inventory file contains information on the characteristics of existing pipes in the system, and the break file lists the failure records of broken pipes. The utilities are identified herein anonymously by the letters A through M. Overall, the data collected for this study can be categorized into five types of factors: physical, historical, protection, operational, and environmental. The attributes available in the datasets differ by utility, as shown in Table 1.

Table 1. Data available by utility (grey cells indicate available data, blank not available)

Attributes		A	B	C	D	E	F	G	H	I	J	K	L	M
Physical	**Joint type**								■					
	Diameter	■	■	■	■	■	■	■	■	■	■	■	■	■
	Material	■	■	■	■	■	■	■	■	■	■	■	■	■
	Length	■	■	■	■	■	■	■	■	■	■	■	■	■
	Restrained	■												
	Roughness									■				
	Dead-end		■											
Historical	**Failure Month**	■		■	■	■	■	■		■	■	■	■	■
	Install Month												■	
	Status	■		■				■	■					■
	Age	■	■	■	■	■	■	■	■	■	■	■	■	■
	Replaced Status								■					
Protection	**Casing Material**	■												
	Lining Material			■	■	■		■	■		■	■	■	
	Lining Status			■	■	■	■	■	■			■	■	
	Lining Age			■		■	■	■	■				■	
	Cathodic Protection Status	■		■			■							
	Cathodic Protection Age			■			■							
	Coating Material										■			■
Operational	**Service type**	■									■		■	
	Pressure											■		

It is clear that certain physical and historical attributes are collected consistently by all utilities: diameter, material, length and age. These attributes are not only among the easiest to collect as they are generally recorded at the time of design and installation, but are also the most commonly applied in predictive modelling. The majority of utilities also record information on lining. On the other hand, only one utility recorded pipe pressure in their inventory.

Pipe materials have evolved throughout the years. The most common pipe material installed in the early to mid-1900s was cast iron, as shown in Figure 1. This market was taken over by ductile iron pipes in the 70s. Soon after, with the advent of plastic pipes, these became the most popular, in particular PVC.

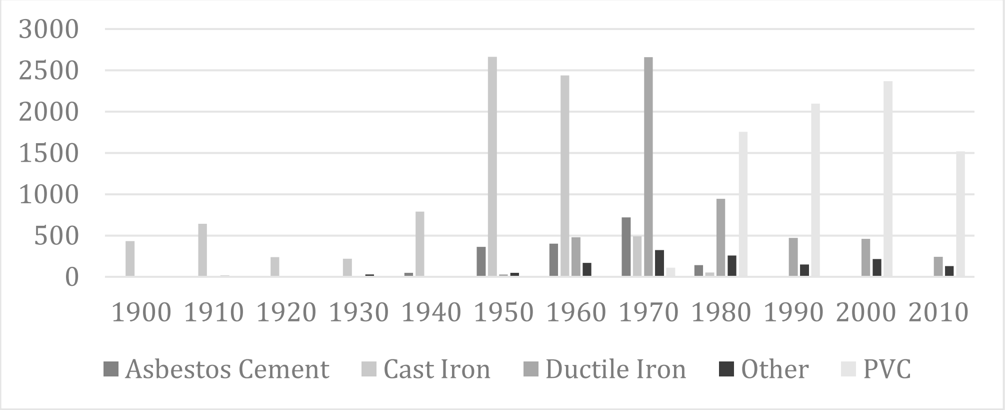

Figure 1. Total lengths of the pipes installed in each decade

Because material use trends changed over the years, the majority of pipes currently breaking are cast iron and ductile iron. This leads to a clear difference in material distribution for inventory and broken pipes, as illustrated in Figure 2. While almost 40% of pipes currently installed are PVC, more than 50%of the break records are for cast iron pipes.

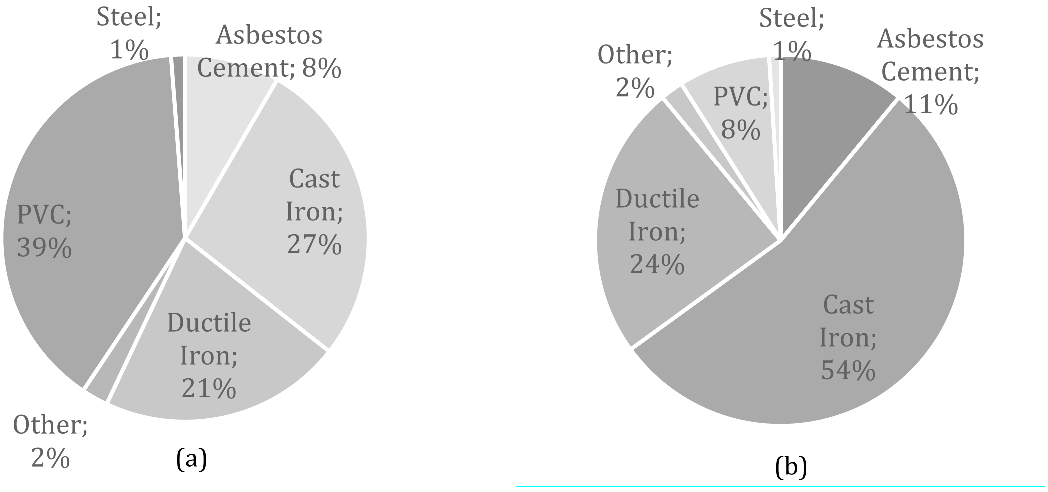

Figure 2. Breakdown of pipes by material for (a) all pipes and (b) broken pipes

It should be noted that he period of historical data collected by each utility differs. While the earliest data collecting utilities began in the 1950s (B, H and M), others only have the last two decades of data available (J and L). The size of the networks varies significantly as well. The largest is B with 6,811 km and the smallest is E with 12 km.

Table 2. General characteristics of pipes in each utility

Utility	Length (km)	Total breaks per KM	Break decades	% Cast Iron	% Ductile Iron	Average age	% Lined pipes	Average Lining age	% Protected pipes
A	897	15.1	1970-2010	16	25	27	-	-	4.3
B	6,811	15.1	1950-2010	21	23	30	-	-	-
C	3,183	13.3	1970-2010	17	20	31	13	32	13
D	2,710	14.9	1970-2010	44	48	39	44	-	-
E	12	956	1980-2010	24	35	34	0.3	018	-
F	1,501	15.7	1970-2010	9	14	23	11.2	33	27
G	392	10.2	1980-2010	10	32	32	25	26	-
H	1,363	27.4	1950-2010	19	0.2	34	0.9	5	-
I	694	15.4	1980-2010	43	44	40	-	-	-
J	1,577	25.5	1990-2010	43	54	43	47	-	-
K	351	9.3	1980-2010	49	37	56	5	-	-
L	481	13.4	1990-2010	31	15	36	11	14	-
M	4,862	33.2	1950-2010	25	1	38	-	-	-

5 RESULTS AND DISCUSSION

5.1 Correlation analysis

Analyzing all common utility attributes, material, diameter, length, age and their relation to break status does not reveal any high correlations. Figure 3 shows attributes are neither highly correlated with each other nor with the target break status. Among these common attributes, material and the length have the strongest association with the target.

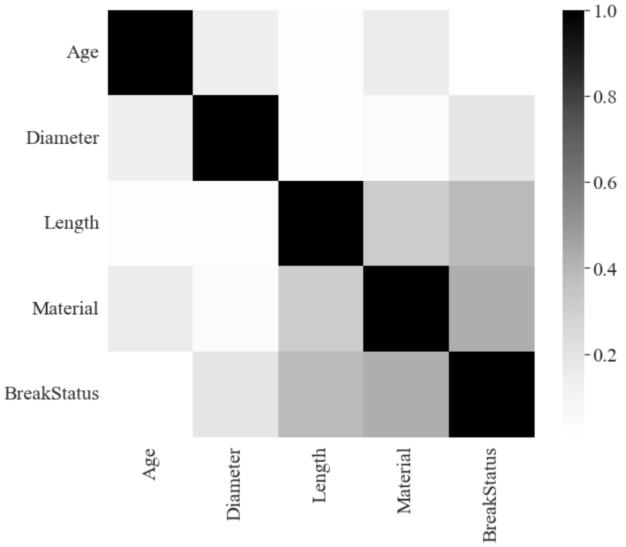

Figure 3. Correlation analysis between common attributes and break status

To further explore the correlations in each city and all attributes, separate correlation analyses were conducted. Results are summarized in Figure 4. The number of values in each box plot depends on how many utilities recorded that data. Attributes collected by only one utility are represented as a line. The most correlated attribute to break status is material. Nonetheless, results vary significantly by utility. This could be related to the variation of material within the utility. The lowest correlation (0.08) was found for utility J whose pipes are 97% either cast iron or ductile iron.

While previous studies commonly applied age in predicting watermain failure, results show cathodic protection age is more highly correlated with break status than age. This highlights the benefit of cathodic protection especially for largely metallic networks such as those analyzed herein.

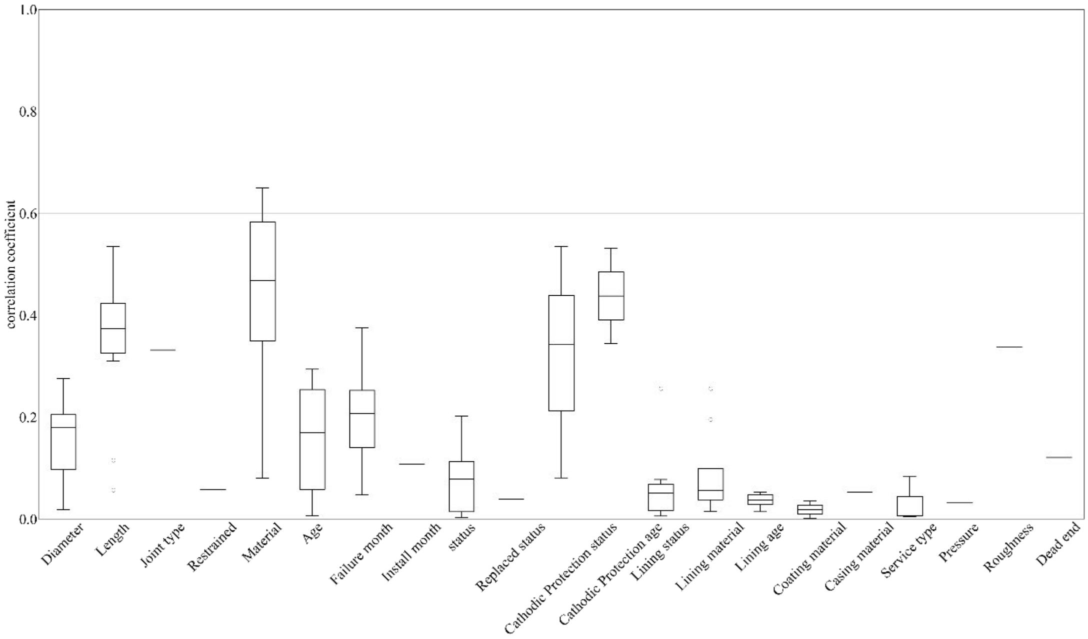

Figure 4 Correlation coefficients - Break status

5.2 Categorical Principal Component Analysis (CATPCA)

CATPCA results (Table 3) show the contribution of each attribute to break status. Overall results point to the importance of protection in general, i.e. lining, coating, and cathodic protection. In particular, the age of the protection, not only status is important. The type of lining material was often more important than the pipe material as well. Lining age was also found to be important for most utilities collecting this data, except utilities E and G, which have the lowest percentages of lined pipes, 0.3 and 0.9% respectively. Thus, collecting protection data can improve the selection and timing of protection activities, potentially reducing capital costs.

Table 3. CATPCA results - Break Status (Dark grey-important factors, light grey-available factors, blank cells-no available attribute)

Attributes		A	B	C	D	E	F	G	H	I	J	K	L	M
Physical	Joint type								8					
	Diameter	11	20	8	10	13	11	12	12	17	12	10	3	13
	Material	11	25	9	14	12	10	10	12	14	12	16	10	12
	Length	12	24	5	14	14	9	11	10	23	11	8	11	15
	Restrained	5												
	Roughness									12				
	Dead-end		10											
Historical	Failure Month	12		10	17	14	12	10		17	15	14	9	15
	Install Month												12	
	Status	11		9				11	4					16
	Age	9	21	8	14	11	9	11	12	16	11	10	10	15
	Replaced Status								10					
Protection	Casing Material	8												
	Lining Material			10	16	13		12	11		13	14	11	
	Lining Status			10	16	12	14	12	11			14	11	
	Lining Age			10		12	14	11	9				11	
	Cathodic Protection Status	12		10			11							
	Cathodic Protection Age			10			11							
	Coating Material										15			15
Operational	Service type	10									11		12	
	Pressure											14		
% Contribution cut-off level		**10**	**20**	**9**	**14**	**13**	**11**	**11**	**10**	**17**	**13**	**13**	**10**	**14**

The results also consistently identified the most collected physical factors, i.e., material, diameter, and length as important. This data was shown to be particularly important when other attributes such as protection factors were not available, as for utilities A and B. Less commonly collected data, such as install month and pressure were also found to be important. However, the range of their contribution requires further investigation as only one or two utilities collect this data.

Failure month was found to be important for most utilities. Graphing the distribution of breaks over months as shown in Figure 5 elucidates the relationship between time of year and breaks. Breaks are more likely in colder months, i.e. January and February. An increase in breaks can also be seen at the height of summer in July when weather is dryer, confirming previous finding (Bruaset and Saegrov, 2018).

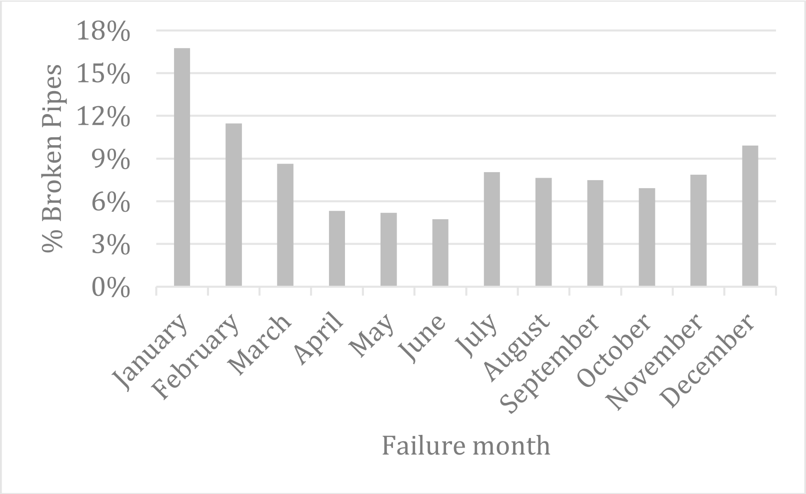

Figure 5. Total percentage of the failed pipes in each month

5.3 Random Forest Recursive Feature Elimination with Cross-Validation (RF-RFECV)

The RF-RFECV approach further reduces the number of selected features, compared with CATPCA. The reduced models, i.e. with fewer attributes, perform equally or slightly better than the full models, i.e. with all attributes, as shown in Table 4. Because the model predict the categorical target break status, they are evaluated according to F1 score and recall. Overall, the analysis rated physical and historical factors as the most important. Specifically length, age and material were consistently found to have the highest weights. Protection activities were rated less highly compared to CATPCA, but cathodic protection age and lining age were still found to be important. The results also selected joint type, pressure, roughness, and dead-end among the important features. This points to the opportunity to collect more operational data and explore the relation between operational decision and infrastructure service life. Pipe dissipation and water stagnation could lead to greater pipe deterioration.

To evaluate the performance and applicability of a model with even fewer attributes, models were developed with only common data (length, material, diameter and age). F1 and recall scores for these models are also provided in Table 4. The performance of these models is only slightly lower than the reduced models developed with RF-RFECV. Thus, for the purpose of predicting pipe deterioration for maintenance and capital planning, commonly available attributes should suffice. Nevertheless, in creating strategies for reducing maintenance and replacement costs, the relation between breaks and other adjustable factors such as pressure and protection should be explored.

Table 4. RF-RFECV weights and results - Break Status (Dark grey-important factors, light grey-available factors, blank cells-no available attribute)

Attributes		A	B	C	D	E	F	G	H	I	J	K	L	M
Physical	Joint type								0.10					
	Diameter	0.07	0.07	0.08	0.06	0.06	▨	0.10	0.06	0.06	0.05	0.07	0.04	0.03
	Material	0.35	0.34	0.32	0.33	0.21	0.40	0.08	0.19	0.11	0.03	0.09	0.14	0.40
	Length	0.22	0.26	0.19	0.34	0.38	0.17	0.35	0.28	0.27	0.43	0.31	0.34	0.30
	Restrained	▨												
	Roughness									0.11				
	Dead-end		0.01											
Historical	Failure month	0.12		0.10	0.10	0.14	▨	0.24		0.18	0.10	0.18	0.19	0.03
	Install month												▨	
	Status	0.03		▨				▨	0.003					0.003
	Age	0.19	0.32	0.18	0.17	0.22	0.18	0.23	0.37	0.26	0.36	0.29	0.23	0.24
	Replaced status								▨					
Protection	Casing material	0.02												
	Lining material			▨	▨	▨		▨	▨		0.04	0.01	▨	
	Lining status			▨	▨	▨	▨	▨	▨			▨	▨	
	Lining age			0.08		0.003	0.08	▨	0.005				0.06	
	Cathodic Protection status	0.01		▨			▨							
	Cathodic Protection age			0.08			0.17							
	Coating material										▨			
Operational	Service type	▨									▨		▨	
	Pressure											0.04		
Full F1		**97.5**	**97.5**	**97.5**	**95.3**	**97.2**	**98.9**	**98.6**	**97.4**	**95.2**	**99.4**	**92.9**	**97.3**	**96.7**
Reduced F1		**97.5**	**97.5**	**97.5**	**95.3**	**97.2**	**98.9**	**98.6**	**97.4**	**95.2**	**99.4**	**92.9**	**97.2**	**96.7**
Common F1		**96.8**	**97.4**	**96.3**	**95.2**	**97.1**	**98.0**	**98.0**	**97.1**	**94.4**	**99.4**	**92.3**	**96.8**	**96.7**
Full Recall		**98.4**	**98.3**	**98.6**	**96.7**	**98.8**	**99.2**	**100**	**98.6**	**97.6**	**99.9**	**96.8**	**98.5**	**97.2**
Reduced Recall		**98.6**	**98.3**	**98.6**	**96.4**	**98.8**	**99.4**	**99.9**	**98.6**	**97.6**	**99.9**	**96.8**	**98.5**	**97.2**
Common Recall		**97.7**	**98.3**	**98**	**96.3**	**98.7**	**99.2**	**99.7**	**98.5**	**96.8**	**100**	**96.2**	**97.3**	**97.2**

5.4 XGBOOST Recursive Feature Elimination with Cross-Validation (XGB-RFECV)

The XGBOOST-RFECV approach yields slightly underperforming models, compared to RF-RFECV, as shown in Table 5. The features selected, however are similar in both RFECV approaches. Physical and historical attributes are rated as the most important, specifically material, diameter, length, age, and failure month. Similar to RF-RFECV, material is the most important factor. Lining age is consistently important for utilities collecting this data except E, which is only 0.3% lined. Lining material was found to be important for utilities D and J, with more than 40% of lined pipes. Thus, the contribution of certain factors depends on local practices and conditions.

Table 5. XGB-RFECV weights and results - Break Status (Dark grey-important factors, light grey-available factors, blank cells-no available attribute)

Attributes		A	B	C	D	E	F	G	H	I	J	K	L	M
Physical	Joint type								0.10					
	Diameter	0.06	0.05	0.02	0.02	0.10	0.02	0.20	0.07	0.06	0.10	0.12	0.07	0.02
	Material	0.53	0.66	0.69	0.85	0.59	0.78	0.13	0.52	0.41	0.36	0.34	0.41	0.77
	Length	0.04	0.08	0.04	0.06	0.12	0.04	0.17	0.07	0.09	0.10	0.14	0.12	0.05
	Restrained	0.05												
	Roughness									0.16				
	Dead-end		0.01											
Historical	Failure month	0.04		0.03	0.02	0.11	0.02	0.11		0.15	0.12	0.24	0.16	0.01
	Install month												0.02	
	Status	0.07		▨				0.05	0.04					0.06
	Age	0.05	0.20	0.07	0.03	0.07	0.07	0.11	0.16	0.13	0.12	0.16	0.10	0.09
	Replaced status								▨					
Protection	Casing material	0.05												
	Lining material			▨	0.01	▨		▨	▨		0.14	▨	▨	
	Lining status			▨	▨	▨	▨	▨	▨			▨	▨	
	Lining age			0.12		▨	0.03	0.23	0.05				0.12	
	Cathodic Protection status	0.09		▨			▨							
	Cathodic Protection age			0.02			0.03							
	Coating material										0.05			▨
Operational	Service type	0.04									▨		▨	
	Pressure											▨		
Full F1		**86.5**	**91.8**	**89.7**	**90.9**	**79.6**	**90.3**	**61.3**	**88**	**77**	**35**	**69**	**74**	**86**
Reduced F1		**86.5**	**91.8**	**89.7**	**90.9**	**73**	**90.3**	**61.3**	**88**	**77**	**27**	**69**	**74**	**86**
Common F1		**85.9**	**85.9**	**88**	**91**	**78**	**90**	**58.8**	**86.2**	**72.5**	**24**	**68.5**	**66.5**	**85.6**
Full Recall		**84.8**	**89.3**	**85.9**	**89.1**	**78**	**86.6**	**60**	**84**	**71.5**	**27**	**62**	**68**	**83**
Reduced Recall		**84.8**	**89.3**	**85.6**	**89.1**	**72**	**86.6**	**60**	**84**	**71.5**	**24**	**62**	**69**	**84**
Common Recall		**84.8**	**84.8**	**84**	**89**	**71**	**87.5**	**52.6**	**81.8**	**66**	**17.5**	**61.6**	**62.6**	**83.5**

5.5 Discussion

Although all approaches applied in this study are of dimensionality reduction, they differ in nature. When data are numerical and linearly related to the target, correlation-based approaches are most appropriate. However, this is not the case for water main break prediction as a mix of numerical and categorical factors is available. CATPCA, also known as non-linear PCA, can handle a linear and non-linear relationship among variables and is recommended when mixed types of data in the analysis are not linearly related to the target.

RFECV can also handle different relations between predictors and targets through different estimators. The selection of an appropriate estimator based on the data structure is key to ensuring good results. In the present study XGBOOST and Random Forest estimators were selected and their hyperparameters tuned to maximize performance. In particular, the resulting random forest models were more accurate and, thus, provide more reliable feature selection results. However, the results are still largely dependent on the hyperparameter tuning.

The most important factors differed between approaches. While physical and historical factors were the most contributing factors in RFECV, CATPCA found protection activities to be the most important. CATPCA also selected a greater number of important factors compared to RFECV. Results also differed by utility, depending on the application of certain protection strategies and the variability of local practices, e.g. different types of installed materials. Thus, data collection strategies should be tailored to the factors impacting the most common materials and protection approaches in each utility.

6 CONCLUSION

Because data collection can be a time and cost intensive process, identifying the driving factors for water main breaks is a valuable endeavour. Based on the results of the dimensionality reduction approaches, a three-step data collection framework is proposed, summarized in Table 6. The first step represents the minimum level of data collection required to produce accurate water main break prediction models. Factors include material, diameter, length and age (calculated based on install date and failure date). This data is commonly collected across all utilities and was found to generate models with high F1 and recall scores, slightly below the optimal RFECV models.

Table 6. Three step data collection framework (1st step - dark grey, 2nd step- grey, 3rd step - light grey)

Physical	Historical	Protection activities	Operational
Material	Installation date	Cathodic Protection year	Pressure
Diameter	Failure date	Lining Material	Service Type
Length	Status	Lining Year	-
Join type	-	Coating Material	-
Roughness	-	Anode type	-
Dead-end	-	-	-
Restrained	-	-	-
Pipe Depth	-	-	-

The second step comprises factors that were found to be important when relevant, especially protection data. Collecting protection data can improve the selection and timing of protection activities, potentially reducing capital costs. More research is required on the extension of pipe service life caused by different types of protection and at different times. Lastly, the third step includes factors that were only collected by a few utilities and not identified as important in all approaches. Among attributes collected by only one utility, pipe pressure, roughness, and dead-end, were found to be important in RFECV. The results also point to an opportunity to collect more operational data and further research to quantify the impacts of different pressure ranges and network designs on deterioration.

7 REFERENCES

[1] I. American Water Works Service Co., "Deteriorating Buried Infrastructure Management Challenges and Strategies," Environ. Prot. Agency, pp. 1–33, 2002.

[2] CIRC, "Informing the Future," 2016. doi: 10.17226/11469.

[3] B. Rajani and Y. Kleiner, "Comprehensive review of structural deterioration of water mains: Statistical models," Urban Water, vol. 3, no. 3, pp. 151–164, 2001, doi: 10.1016/S1462-0758(01)00032-2.

[4] U. Shamir and C. D. D. Howard, "An analytic approach to scheduling pipe replacement.," Am. Water Work. Assoc. J., vol. 71, no. 5, May 1979, pp. 248–258, 1979, doi: 10.1002/j.1551-8833.1979.tb04345.x.

[5] T. M. Walski and A. Pelliccia, "Economic Analysis of Water Main Breaks.," J. / Am. Water Work. Assoc., vol. 74, no. 3, pp. 140–147, 1982, doi: 10.1002/j.1551-8833.1982.tb04874.x.

[6] A. J. Kettler and I. C. Goulter, "Analysis of Pipe Breakage in Urban Water Distribution Networks.," Can. J. Civ. Eng., vol. 12, no. 2, pp. 286–293, 1985, doi: 10.1139/l85-030.

[7] L. Berardi, O. Giustolisi, Z. Kapelan, and D. A. Savic, "Development of pipe deterioration models for water distribution systems using EPR," J. Hydroinformatics, vol. 10, no. 2, pp. 113–126, 2008, doi: 10.2166/hydro.2008.012.

[8] Y. Wang, T. Zayed, M. Asce, O. Moselhi, and F. Asce, "Prediction Models for Annual Break Rates of Water Mains," no. February, 2009.

[9] M. Nishiyama and Y. Filion, "Forecasting breaks in cast iron water mains in the city of Kingston with an artificial neural network model," Can. J. Civ. Eng., vol. 41, no. 10, pp. 918–923, 2014, doi: 10.1139/cjce-2014-0114.

[10] Y. Wang, O. Moselhi, F. Asce, T. Zayed, and M. Asce, "Study of the Suitability of Existing Deterioration Models for Water Mains," no. February, pp. 40–46, 2009.

[11] P. Rajeev, J. Kodikara, D. Robert, P. Zeman, and B. Rajani, "Factors contributing to large daimeter water pipe failure," Water Asset Manag. Int., vol. 10, no. 3, pp. 9–14, 2014.

[12] B. Snider and E. A. McBean, "State of watermain infrastructure: A canadian case study using historic pipe break datasets," Can. J. Civ. Eng., vol. 48, no. 10, pp. 1266–1273, 2021, doi: 10.1139/cjce-2020-0334.

[13] N. A. Barton, T. S. Farewell, S. H. Hallett, and T. F. Acland, "Improving pipe failure predictions: Factors effecting pipe failure in drinking water networks," Water Res., vol. 164, p. 114926, 2019, doi: 10.1016/j.watres.2019.114926.

[14] Martínez-Codina, M. Castillo, D. González-Zeas, and L. Garrote, "Pressure as a predictor of occurrence of pipe breaks in water distribution networks," Urban Water J., vol. 13, no. 7, pp. 676–686, 2016, doi: 10.1080/1573062X.2015.1024687.

[15] S. Bruaset and S. Sægrov, "An analysis of the potential impact of climate change on the structural reliability of drinking water pipes in cold climate regions," Water (Switzerland), vol. 10, no. 4, 2018, doi: 10.3390/w10040411.

[16] S. A. Andreou, D. H. Marks, and R. M. Clark, "A new methodology for modelling break failure patterns in deteriorating water distribution systems: Theory," Adv. Water Resour., vol. 10, no. 1, pp. 2–10, 1987, doi: 10.1016/0309-1708(87)90002-9.

[17] R. Jafar, I. Shahrour, and I. Juran, "Application of Artificial Neural Networks (ANN) to model the failure of urban water mains," Math. Comput. Model., vol. 51, no. 9–10, pp. 1170–1180, 2010, doi: 10.1016/j.mcm.2009.12.033.

[18] L. Scholten, A. Scheidegger, P. Reichert, M. Mauer, and J. Lienert, "Strategic rehabilitation planning of piped water networks using multi-criteria decision analysis," Water Res., vol. 49, pp. 124–143, 2014, doi: 10.1016/j.watres.2013.11.017.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using DCA,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5139 LNAI, pp. 62–72, 2008, doi: 10.1007/978-3-540-88192-6_8.

[20] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, “Decision variants for the automatic determination of optimal feature subset in RF-RFE,” Genes (Basel)., vol. 9, no. 6, 2018, doi: 10.3390/genes9060301.

[21] B. Butcher and B. J. Smith, “Feature Engineering and Selection: A Practical Approach for Predictive Models,” Am. Stat., vol. 74, no. 3, pp. 308–309, 2020, doi: 10.1080/00031305.2020.1790217.

[22] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” Chemom. Intell. Lab. Syst., vol. 83, no. 2, pp. 83–90, 2006, doi: 10.1016/j.chemolab.2006.01.007.

[23] S. Wang and S. Chen, “Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling,” J. Pet. Sci. Eng., vol. 174, no. November 2018, pp. 682–695, 2019, doi: 10.1016/j.petrol.2018.11.076.

[24] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, and S. Zhang, “A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data,” Diagnostics, vol. 9, no. 178, 2019.

[25] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.