Prediction of pipe failures in water supply networks using logistic regression and support vector classification

Alicia Robles-Velasco [a,b,c], Pablo Cortés [a,b], Jesús Muñuzuri [a], Luis Onieva [a]

[a] *Dpto. de Organización Industrial y Gestión de Empresas II. ETSI. Universidad de Sevilla. C/ Camino de los Descubrimientos S/N, 41092 Sevilla (Spain)*
[b] *Cátedra del Agua (EMASESA-Universidad de Sevilla)*
[c] *Corresponding author. E-mail address: arobles2@us.es*

Abstract

Companies in charge of water supply networks are making a huge effort to optimally plan the annual replacements of pipes. This would save costs, enable a higher quality of service and a sustainable management of infrastructure.

This study presents a methodology to predict pipe failures in water supply networks. Logistic regression and support vector classification are chosen as predictive systems. Both provide a failure probability associated with each sample which is increasingly required by companies that manage these infrastructures. Furthermore, several pre-processing techniques that seek to improve the accuracy of predictions are addressed.

The proposed methodology is illustrated with the real case of a Spanish city. This is an extensive water supply network whose recorded data contains 4,393 pipe failures. The results obtained state that the number of unexpected failures might be significantly reduced. Around 30% of failures could have been prevented by replacing only 3% of the network's pipes per year, which is a realistic and feasible option.

As a future line of research, the objective must be to develop a global tool that incorporates the failure probability and its consequence, generating the optimal pipe replacement plan.

Keywords: Water supply networks; Pipe failures; Support Vector Classification; Logistic Regression; Predictive algorithms

1. Introduction

Nowadays, ensuring access to drinking water is one of the most important challenges in the world. Water distribution networks are the infrastructure responsible for bringing this resource to population. Data collected in the 16th Drinking Water Supply and Sanitation Study in Spain [1] show that water distribution networks in the country traverse a total of 224,000km, which in average corresponds to 4.8 metres per person. The good maintenance of such networks is crucial to preserve the system quality, which has a strong impact on economy, environment and well-being of people. One of the symptoms of the intrinsic deterioration of a water supply system is the appearance of frequent breakages. The detection of these failures is difficult because of the buried nature of pipes. An unexpected failure in the network may lead to serious security risks as well as to the interruption of the supply. Figure 1 shows several examples of real dangerous situations derived from unexpected pipe breaks.

Companies in charge of water supply networks are making a huge effort to optimally plan the annual repairs of these utilities. This supposes cost savings, a higher quality of the service and a sustainable management of the infrastructure. Although the most common tools to predict pipe failures are rating indexes or lifetime curves [2], more robust and accurate techniques are necessary. For this reason, machine learning as predictive system is increasingly demanded by this sector.

Figure 1. *Real consequences of unexpected pipe failures (photographs taken by the authors)*

This paper demonstrates the ability of two binary classifiers to the challenge of predicting pipe failures in water supply networks. In the second section, an extensive literature review, including the main variables and approaches used to date, is presented. Section 3 describes the applied methodology and its three main blocks: (i) data pre-processing; (ii) predictive system: logistic regression (LR) and support vector classification (SVC); and (iii) the quality metrics used to measure and compare the obtained results. In the fourth section, the case of a Spanish city is used as illustration. Firstly, data are described and analysed. Secondly, the two predictive models are calibrated and some techniques are studied to improve their performance. Finally, results are shown and discussed from a quantitative and comparative point of view. Conclusions are presented in section 6.

2. Literature review

There are many studies aimed at determining the causes of pipe failures as well as the failure modes [3]–[6]. Nowadays, it is widely accepted that most breakages do not only occur in old pipes. Kleiner and Rajani [7] differentiated three phases in the pipe lifecycle: "burn-in", "in-usage" and "wear-out". They explained in their work that breakages in the first phase are often caused by manufacturing defects and storage, or by improper construction processes. A pipe "in-usage" can suffer a breakage because of inappropriate maintenance, natural hazard or external interference. While a "wear-out" pipe has more probability to break due to the intrinsic deterioration of the installation. Despite the major difficulty of predicting all the breakages in a water supply system, many authors have demonstrated that pipe failures follow certain patterns that can be extracted from historical data [8]–[13].

2.1. Variables considered in the literature to explain pipe failures

In most cases, it is impossible to obtain all the data that influence the operation and state of a network. Nevertheless, in the last years, available data in the industry have increased due to both the development of new technology and the growing interest in big data usefulness. This has enabled an in-depth study of the variables that influence pipe failures and has led to the application of different prediction models. The introduction of geographic information systems (GIS) tools for the storage, manipulation and access to the water network data suggested a new perspective in the field. In fact, case studies from many researches are based on data which have been extracted from this kind of tools [6], [8]–[11], [14]–[16].

Each water supply network has different physical, operational and environmental characteristics [10], [12], [17]. Physical features are those that characterise the layout and state of the network, for example, material, diameter or ageing of the pipes. Operational factors include parameters concerning network operation, as water properties, pressure or velocity, among others. Finally, environmental factors are external conditions that can affect the network performance, such as climate, soil corrosion, use of the land, etc.

The importance of physical factors in future breakages has been strongly demonstrated. Fares and Zayed [17], after consulting more than twenty experts, concluded that age has the highest effect on risk of failure, followed by material and failure rate. Meanwhile, Christodoulou *et al*. [9] considered the age of pipes as an output variable named "*LifeCycle*". Material is treated differently, some authors only study certain kind of materials [13], [16], [18], [19] while others consider all the diverse materials in the water network. Several studies stated that pipes with smaller diameters tended to suffer more failures [20], [21]. Regarding the pipe length, higher lengths broadly suppose more exposure to risk of failure [8]. An alternative to replacing a pipe might be to add some type of protection. In fact, protection methods have demonstrated to significantly extend the lifetime of pipes, specifically the cathodic protection in iron pipes [22].

Operational factors, such as water pressure or velocity, are more laborious to be obtained for the entire network. Among all of them, the number of previous failures (NOPF) is definitively the most used factor and its effect in the appearance of new pipe breaks has been widely demonstrated [3], [10]–[13], [17], [22]. Pipes which have already experienced a breakage are more prone to suffer a new one. Oliveira *et al.* [14] employed a density-based spatial clustering analysis to conclude that poor repairs of breaks might produce new breakages close to the previous one. Regarding water pressure, there are several points of view. On the one hand, Shirzad *et al*. [19] have argued that the performance of two predictive models, artificial neural networks and support vector regression, improved when the average of the hydraulic pressure was incorporated as input variable. On the other hand, Jafar *et al*. [10] have defended the major influence of the fluctuation of the pressure over its average.

Data referring to the environment of the pipes are less common and these factors are sometimes estimated per area under certain assumptions, for instance, using the clusterisation of pipes by location and historical breaks [12]. Debón et al. [11] included the traffic variable in their study, concluding that pipes under roadways with intense traffic are more likely to break than those under sidewalks or roadways with low traffic. In [23], the correlation between pipe failures and soil liquefaction is studied since the breaking behaviour changes if an earthquake occurs in the area. Fares and Zayed [17] defended that the best parameter to carry out an appropriate soil classification is corrosivity. Corrosivity is an electrochemical phenomenon between two materials in contact with each other that results in the deterioration of parent material [24]. Since it is difficult to obtain this parameter directly, the soil type is used to represent an approximation of its corrosivity. There are many factors that affect corrosivity, soil pH has been considered as a good indicator because corrosion occurs in a certain range of pH [25].

It is impossible to study the variability that some parameters experience along the year when the recorded data is annual. Authors, who have worked with data using shorter periods of time, defend the significance of seasonal changes on pipe breaks. One important cause of failure is the water renewal time inside pipes which typically increases during dry periods. The greater the renewal time is, the more breakages appear [3], [26], [27]. Moreover, pipes tend to suffer more failures during winter periods in places where heavy snowfalls are common due to the extra-loads pipes must support [21].

Table 1 summarises the main factors used by several studies in the last decade. The second column, type of study, clarifies whether it is a descriptive or a predictive study. Additionally, the utilised approaches are also emphasised. In addition to physical, operational and environmental features, the consequences of failures have also been considered in some studies and are also included in the table.

2.2. Approaches to predict pipe failures

This topic can be addressed from a descriptive or a predictive approach. In both cases, the use of high-quality historical data leads to well-founded conclusions. The descriptive or backward analysis helps to understand how the network works and which its most vulnerable points are [28]. Its objective is not to predict breakages but to analyse the characteristics and factors that promote them. Fuzzy logic (FL) and multicriteria optimisation (MCO) are techniques usually applied to this purpose [4], [5].

The classification of predictive studies according to [29] is: (i) statistical models; (ii) probability-based methods; and (iii) artificial intelligence. Statistical models are the most suitable to extract information about the variable interactions and also to predict the lifetime of pipes. Some survival models (SMs) have been employed to estimate the failure time of pipes [8], [11], [13], [30]. However, they cannot take into account the zeros or pipes that do not suffer any breaks.

Amongst probability-based methods, Bayesian Belief Networks (BBNs) receive a special attention in the scientific literature. BBNs are based on the Bayes' probability theorem. Its structure is represented by a directed acyclic graph where nodes represent parameters and arcs represent the probabilistic relationship between them. They allow performing prognostic and diagnostic reasoning [31]. Its main disadvantages are that the interpretation of results is not trivial and expert opinions are generally needed to generate conditional probabilities.

Artificial intelligence applications in this field have been mainly related to Artificial Neural Networks (ANN)[9], [10], [19], [32]. This methodology tries to emulate the functioning of human brains. Neurons are represented by nodes and the nerve impulses by a weighted sum of the input values in each neuron. The learning of networks is achieved by adjusting those weights, while its structure does not usually vary [33]. Although they have excellent pattern recognition and generalisation capabilities, ANNs are unable to explain the relation between parameters [9]. According to [19], support vector regression (SVR) algorithms are more appropriate to face this problem because of ANNs' lack of consistency with the physical behaviour observed. SVR allows predicting the failure rate [34] using a regression analysis after mapping the features into a high-dimensional space. The approaches considered in each study are also specified in table 1. The acronym GLM refers to generalised linear models, GP to genetic programing, NB to naïve Bayes classifier and OL to ordered lists, a heuristic process.

In this study, logistic regression and support vector classification are used to address the described problem. There are only two studies that have already used LR [8], [12]. While they only use scalar measurements such as mean squared error (MSE) and Akaike information criterion (AIC) to analyse the models' performance, we propose the use of specific quality metrics such as the confusion matrix and the ROC curves. Furthermore, our study contemplates all pipes present in the network and not only these which have previously suffered a failure.

To the best of our knowledge, support vector machine have only been utilised as a linear regression technique to estimate lifetime of pipes [19], [32] or their failure rate [34]. Nevertheless, no prior evidence has been found about its application to classify pipes regarding the prediction of future failures. We have applied this approach here because it has proved to be suitable working with unbalanced data [35] and has accomplished successful results in other fields of study [36]. Additionally, as such as LR, the output variable generated by SVC can be interpreted as a failure probability.

Table 1. *Summary of the main factors used by different authors and the resolution method applied by each one*

				Physical							**Operational**							**Environmental**							**Consequences**
	Reference	**Study type**	**Approach**	Age	Material	Diameter	Length	Thickness	Protection	Others*	NOF	Mean Press	Velocity	Water properties	Water age	Failure type	Others**	Traffic	Soil type	Soil corrosivity	Area type	Temperature	Freezing Index	Others***	
[8]	Yamijala *et al*., 2009	Pred	SMs; GLMs; LR	x	x	x	x				x	x					x		x	x	x	x		x	
[11]	Debón *et al*., 2010	Pred	SMs	x	x	x	x				x					x		x	x						
[10]	Jafar *et al*., 2010	Pred	ANNs	x	x	x	x	x			x						x		x		x				
[17], [37]	Fares and Zayed, 2009-10	Descr	FL; MCO	x	x	x			x		x			x				x	x		x				x
[9], [38]	Christodoulou *et al*., 2009-2010	Pred	FL; ANNs		x	x	x				x							x			x			x	
[18]	Xu *et al*., 2011	Pred	GP	x		x	x				x														
[12]	Kleiner and Rajani, 2012	Pred	OL; LR; NB; SM	x			x				x						x							x	
[6]	Royce *et al*., 2014	Descr	BBNs							x							x		x			x		x	x
[19]	Shirzad *et al*., 2014	Pred	SVR; ANNs	x		x	x			x		x													
[3]	Pietrucha, 2015	Descr	-	x	x	x	x			x	x						x								
[13]	Kabir *et al*., 2015a	Pred	SMs; NB	x		x	x				x									x	x		x	x	
[16]	Kabir *et al*., 2015b	Pred	BBNs	x		x	x	x				x	x	x	x			x		x	x		x		x
[5]	Li *et al*., 2015	Descr	MCO	x	x	x					x								x					x	
[21]	Sattar *et al*., 2016	Pred	GP;GA		x	x	x		x		x														
[4]	Al-Zahrani *et al.,* 2016	Descr	FL; MCO	x	x						x	x	x	x	x				x		x				x
[32]	Kutyłowska, 2016	Pred	SVR; ANN	x		x	x																		
[39]	Farmani *et al*., 2017	Pred	GP	x		x	x				x											x	x		
[22]	Sattar *et al*., 2017	Pred	ANN		x	x	x		x		x														
[34]	Kutyłowska, 2018	Pred	SVR				x				x														
[31]	Tang *et al*., 2019	Pred	BBNs	x	x	x	x			x	x					x				x				x	
[30]	Lin and Yuan, 2019	Pred	SMs	x		x	x																		

** Include network type, depth of installation or pipe lining.*

*** Include pressure fluctuation or time since last breakage.*

**** Include rainfall, soil resistivity or soil shrink swell among others.*

Consequences are mainly repair costs and damage to surrounding measures as population density.

3. Methodology

Figure 2 outlines the applied methodology in four main blocks. The three last blocks are explained in the following subsections. The first block is not addressed because this study assumes that the acquisition of data and the identification of factors have already been done by the company in charge of the network.

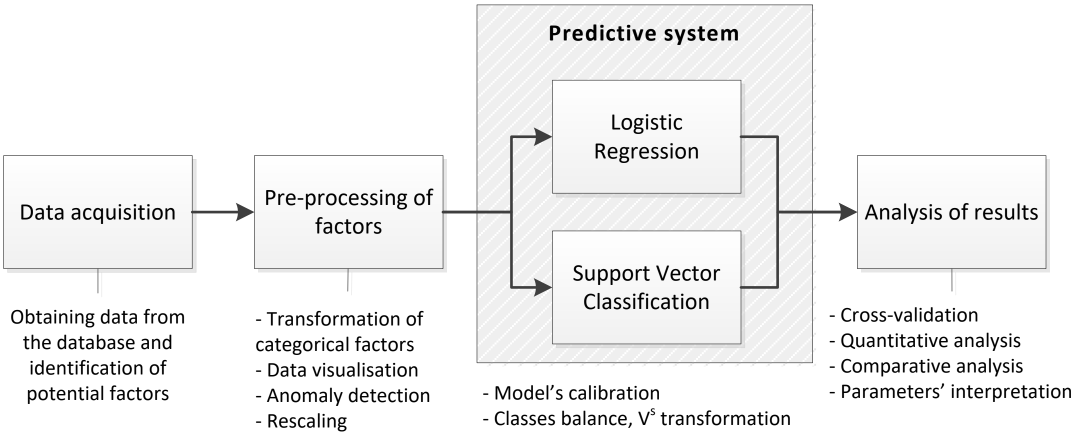

Predictive system
Data acquisition
Pre-processing of factors
Logistic Regression
Support Vector Classification
Analysis of results
Obtaining data from the database and identification of potential factors
- Transformation of categorical factors
- Data visualisation
- Anomaly detection
- Rescaling
- Model's calibration
- Classes balance, V^s transformation
- Cross-validation
- Quantitative analysis
- Comparative analysis
- Parameters' interpretation

Figure 2. *Main steps of the applied methodology*

3.1. Data pre-processing

Data pre-processing have an important role in all predictive algorithms due to its great influence on models' performances. Systems of the same nature usually have data with a similar structure. Therefore, the main steps to pre-process data from water supply networks are, in general, common to all of them.

Categorical data such as pipes material, pipes protection, type of network or traffic need to be transformed into numerical. In this study, it is proposed to assign an integer to each variable level depending on its failure rate per unit length (λ). Then all factors can be analysed by graphics and descriptive statistical measures to detect anomalies such as missing values or outliers. A sample which presents an anomaly could be removed, but relevant information about the rest of variables may be lost. Consequently, missing values and outliers are filled with the median of the factor. Finally, all factors are standardised in order to unify their scale.

3.2. Predictive system

Among all the explained predictive methodologies, logistic regression and support vector classification are selected to predict pipe failures in water supply networks because: (i) They generate an output between 0 and 1 that can be interpreted as a failure probability which is increasingly demanded by companies; (ii) Their capabilities to work with unbalanced data which allows including in the study those pipes which have not suffered any break, making the predictive system more realistic; and (iii) They work efficiently with small and medium size database, which is the case of many companies whose records are not too extensive.

In general, models need to be calibrated in order to fit some hyperparameters. Furthermore, water supply networks present a common characteristic: their data is unbalanced. The majority of pipes have never suffered a failure. This causes problems in the training phase. If data classes are not balanced, the model will not learn to predict sample of the minority class [35]. In this study, an under-sampling technique which removes samples of the majority class is applied to the training set. Other procedure to enhance prediction accuracies is the transformation of variables whose distribution or magnitude order is significantly different. For instance, the use of logarithms might avoid the expansion of data into high orders of magnitude.

3.2.1. Logistic regression

Logistic regression is a particular case of the generalised linear models that concerns the analysis of binary data, where the link function is the logit or logistic function [40].

$$p_i = \frac{1}{1 + e^{-(w^T x_i + b)}} \tag{1}$$

As stated in equation 1, the probability of occurrence of the success of interest (p_i) is a function of the vector of explanatory variables (x_i), their respective associated weights (w), which are common to the whole observations, and a constant term, *b*. The subscript *i* refers to each of the *N* observations forming the sample.

$$min_{w,b} \frac{\|w\|^2}{2} + C_{lr} \sum_{i=1}^{N} log(1 + e^{-y_i(w^T x_i + b)}) \tag{2}$$

In this study, weights are calculated by minimising the negative log-likelihood function (second term of equation 2)[41], [42]. Where C_{lr} is a previously fixed hyperparameter that controls the balance between the two terms of the objective function, being the first one a weight regularisation term according to L_2-norm. Once weights are estimated, predictions of new samples can be made by (eq. 3). The value of y_i, the output variable, depends on its associated probability and a fixed risk threshold whose value is usually *0.5*.

$$y_i = \begin{cases} -1 \; if \; p_i \leq threshold \\ 1 \; if \; p_i \; > threshold \end{cases} \tag{3}$$

3.2.2. Support vector classification

Support vector classification is a support vector machine algorithm based on the structured risk minimisation principle [43]. The explanatory variables are mapped through non-linear structures into a high dimensional space and then, a hyperplane that optimally separates both classes is generated. This hyperplane aims at minimising the classification errors while maximises the margins or distance sum from the hyperplane to the nearest training samples of each class [44].

The primal model is presented hereafter in equations (4-6) [45].

$$min_{w,b,\epsilon} \frac{\|w\|^2}{2} + C_{svc} \sum_{i=1}^{N} \in_i \tag{4}$$

Subject to:

$$y_i(w^T \cdot \emptyset(x_i) + b) \geq 1 - \in_i \qquad i = 1, \dots, N \tag{5}$$

$$\in_i \geq 0 \qquad i = 1, \dots, N \tag{6}$$

Where $\emptyset(x_i)$ is a non-linear function which maps each observation composed of its explanatory variables (x_i) into a higher-dimensional space. C_{svc} is again a regularisation parameter, *w* the weight vector associated to the explanatory variables in the new space, also named feature space, and *b* a bias term. Variables $\in_i$ are slack variables representing the distance between the observations *i* and the edge of the margin corresponding to their classes. Therefore, finding the optimal hyperplane (eq. 7) which maximises the margin (in the high-dimensional space) corresponds to minimising the vector weights' norm together with the number of misclassified instances (eq. 4). Finally, the labels or output variables, $y_i\{-1,1\}$, represent the sample class.

$$D(x_i) = w^T \cdot \emptyset(x_i) + b \qquad (7)$$

While the scale of the primal model depends on the dimensionality of the problem, its corresponding dual form depends on the number of samples. Consequently, when the dimensionality is high enough, it is more convenient to solve the dual model (eqs. 8-10).

$$max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (8)$$

Subject to:

$$\sum_{i=1}^{N} y_i \, \alpha_i = 0 \qquad (9)$$

$$0 \leq \alpha_i \leq C_{svc} \qquad i = 1, \ldots, N \qquad (10)$$

A Kernel function, $K(x_i, x_j)$, assigns to each pair of instances a corresponding in the feature space. There are many different Kernel functions, such as linear, polynomial, radial basis, sigmoidal, etc., whose only requirement is to be symmetric and positive semi-definite. Previous studies in this field have shown that the radial basis Kernel function (eq. 11) is the most suited to classification problems [46]. Thus, in our study, we utilise a radial basis Kernel function where γ is a hyperparameter that represents the inverse of the radius of influence of the instances selected by the model as support vectors [47].

$$K(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j) = exp(-\gamma \|x_j - x_i\|) \qquad (11)$$

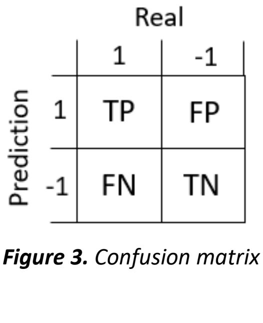
Predictions of new samples (eq. 12) are done once the weights and the bias term are estimated by solving the model.

$$\text{SVC} \quad y_i = \begin{cases} -1 \; if \; w^T \cdot \emptyset(x_i) + b \leq 0 \\ 1 \; if \; w^T \cdot \emptyset(x_i) + b > 0 \end{cases} \qquad (12)$$

3.3. Analysis of results: quantitative and comparative

In this study, the obtained results are analysed and compared using easily interpretable metrics such as confusion matrix and the ROC curves. These metrics are specific to measure the quality of classification approaches. Furthermore, cross-validation ensures that final results are independent of the partition between training and test data.

On the one hand, the confusion matrix is a tool to address the performance of binary classifiers. Once a model is estimated using training data, a prediction is made for each validation sample. Then, this matrix contains the real values against those predicted for the validation set (figure 3). There are four possible results for each sample: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). Each box would include the total number of observations of each type.



		Real	
		1	-1
Prediction	1	TP	FP
	-1	FN	TN

Figure 3. *Confusion matrix*

Some of the most frequent metrics derived from the confusion matrix are *accuracy* and *recall* [48].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

The *accuracy* (eq. 13) measures the total percentage of correct predictions; this metric gives the same importance to both classes, so in the case of an unbalanced sample it may cause misunderstanding. For this reason, the *recall* (eq. 14) is useful as it measures the percentage of right predictions made from class 1, in this study, pipes which suffer a failure. *Specificity* gives the percentage of correct classifications of class -1.

On the other hand, it is common to calculate the *Area Under the Curve* (AUC) as a metric which represents the ability of a classifier to avoid false classifications. The *Receiver Operating Curve* (ROC) depicts the *recall* (or true-positive rate) against 1-s*pecificity* (or false-negative rate) for different values of a risk threshold. The AUC is always between 0 and 1 [49]. A classifier whose AUC is 0.5 (red line of figure 4), will make random classifications, and closer to 1, better the performance of the classifier.

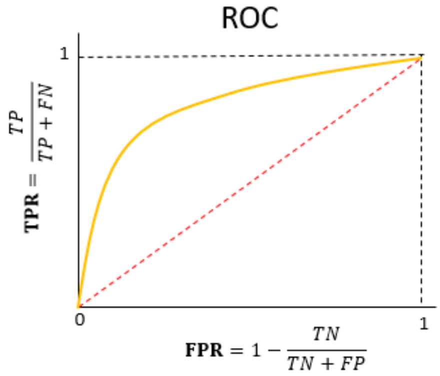


Figure 4. *A generic ROC curve*

4. Case study: City of Seville

Seville is a city located in the south of Spain with a warm Mediterranean climate. The network analysed in this study supplies fresh water to more than 1 million people. It covers a total area of 1,220km^2, including the city town and its metropolitan area, and it is composed of 3,800km of pipes.

4.1. Description and pre-processing of data

Table 2 presents a brief description of the factors used in this study.

Table 2. *Data description*

Factor	Description	Units	Count	Mean	Std	Min	Max
MAT	Material of which the pipe is made	---	88541	7.71	4.37	0	13
DIA	Nominal diameter of the section	mm*	88541	158.58	157.90	20	2000
AGE	Years since the installation	years	88541	25.60	19.66	0	118
LEN	Length of the section	m	88541	41.98	75.49	0.50	2522
CON	Number of connections of the section	---	88541	2.08	4.70	0	71
N_type	Net type	---	88541	0.91	0.28	0	1
ΔPRE	Pressure fluctuation	m**	87739	2.87	2.33	0	60.19
NOF	Total number of failures	---	88541	0.04	0.29	0	11

*1m = 1000mm **9,806.38Pa = 1m

There are fourteen different materials which have been grouped in cements (AMC), metal (FER) and plastics (PLA) following Jafar *et al*. [10] suggestion. As observed in table 3, cement pipes show the highest λ, 2.1 failures have occurred per kilometre during the seven years of study.

Table 3. *Pipe materials description*

Acronym	Material	% Length	$\boldsymbol{\lambda_i}$ (Breaks/u.L.)
FE	Metal	58.27	0.0005
PLA	Plastic	7.17	0.0006
AMC	Cement	34.56	0.0210

Available data only differs between two types of network, transport and secondary, whose percentage of length is 14.8% and 85.2% respectively. The number of recorded failures for the transport network is 0.42 breakages per kilometre, while the secondary network presents 1.14 breakages per kilometre.

Data visualisation

Histograms (figure 5) enable an overview of the frequency distributions of the variables. This allows observing the existence of possible outliers and the need for transformation of some variables in order to scale it. As shown in table 2, the average age of pipes is 25 years and its histogram shows that most of the sections are less than 50 years old. However, there is a substantial number of pipes that are 118 years old. The fluctuation of the pressure mainly varies between 0 and 10, existing atypical values of 60. Consequently, it will be impossible to extract patterns of pipes with these fluctuations of pressure values because of the lack of a significant sample. *DIA* and *LEN* cover large ranges of magnitude compared to the other of variables. Finally, the histogram of *NOF* shows one of the previously mentioned characteristic of this type of problems: the sample is totally unbalanced since most of the sections have not suffered any breakage during the seven years of study.

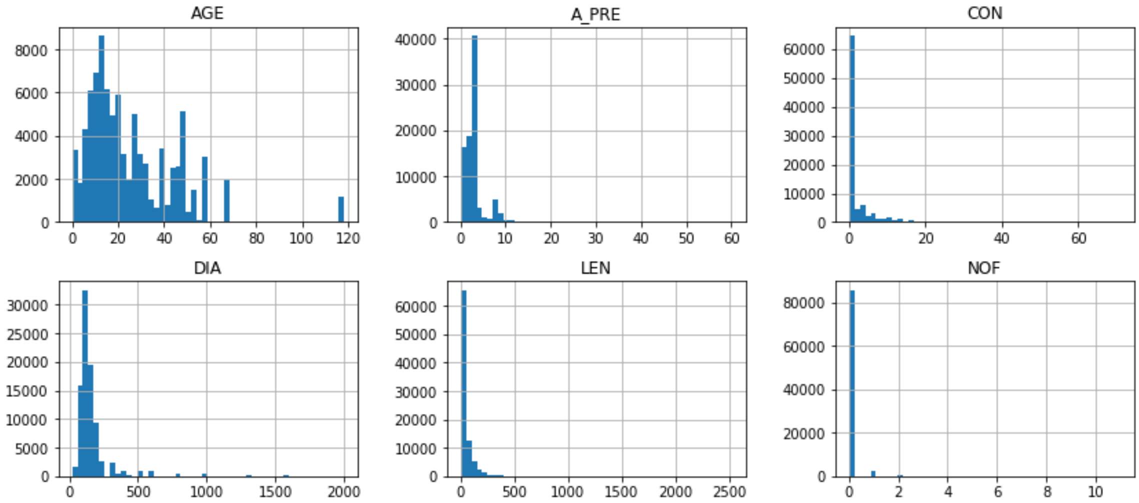


Figure 5. *Histograms of numerical variables*

Relationship between factors

The correlation coefficient is the covariance between two standardised variables. This parameter is only applied to a pair of quantitative variables and represents the linear relation between them. It is independent of the variables' scale and it moves from -1 to 1, 0 meaning non-relation and 1 (or -1) an intense linear relation.

Table 4 shows the matrix of correlation between variables. The highest relationship exists between *DIA* and *N_type* because, in general, transport network pipes have bigger diameters than secondary network ones. *MAT* and *AGE* also show a strong linkage because the higher numbers were assigned to the materials with the greatest failure rates. Therefore, it could imply a relationship between *AGE* and *NOF*. Variable *CON*, referring to the connections of a pipe, is changed by the number of connections per unit length, avoiding the possible dependence between *CON* and *LEN*.

Unfortunately, none linear relationship has been found between pressure and *NOF*. The reason could be the lack of precision of these data that have been recently added to the study.

Table 4. *Matrix of correlation between variables*

	MAT	DIA	AGE	LEN	CON	N_type	ΔPRE	NOF
MAT	1							
DIA	0.01	1						
AGE	0.51	0.01	1					
LEN	0.03	0.11	-0.03	1				
CON	0.04	-0.13	0.02	0.39	1			
N_type	-0.02	-0.73	-0.04	-0.12	0.13	1		
ΔPRE	0.05	-0.08	0.03	-0.01	0.02	0.05	1	
NOF	0.16	-0.02	0.12	0.11	0.11	0.01	-0.01	1

Analysis of pipe breaks

The total number of pipe breaks recorded during the period of study (2012-2018) is 4,393. Figure 6 shows the number of breaks of each material group per year. Since the percentage of cement pipes is only 34.56%, the existence of a problem concerning these pipes emerges as unquestionable. In addition, a progressive increase of AMC and PLA pipe breaks is appreciated.

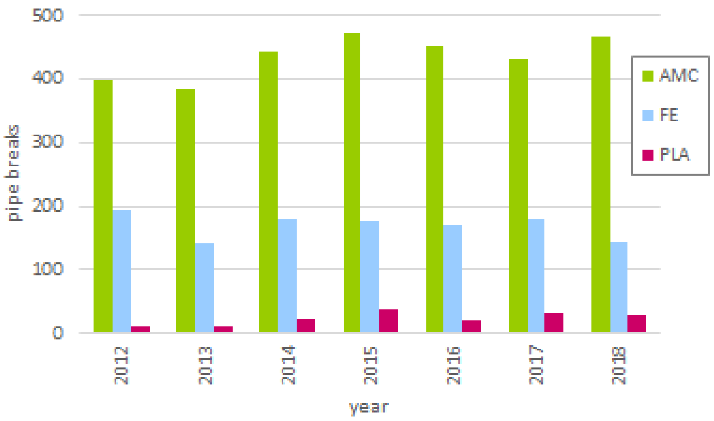


Figure 6. *Total number of pipe breaks per year*

The percentage of pipe breaks per range of diameter, length, age and pressure fluctuation is shown in figure 7. As stated in the literature, it can be noticed that more breakages appear in pipes with smaller diameters. In most cases, the breakage percentage of AMC pipes is higher than in the other materials. However, old metal pipes show large pipe failure percentage, so it should be more deeply studied. As previously mentioned, pipe length suggests more exposure to risk of failures, which is demonstrated in the graphic (b).

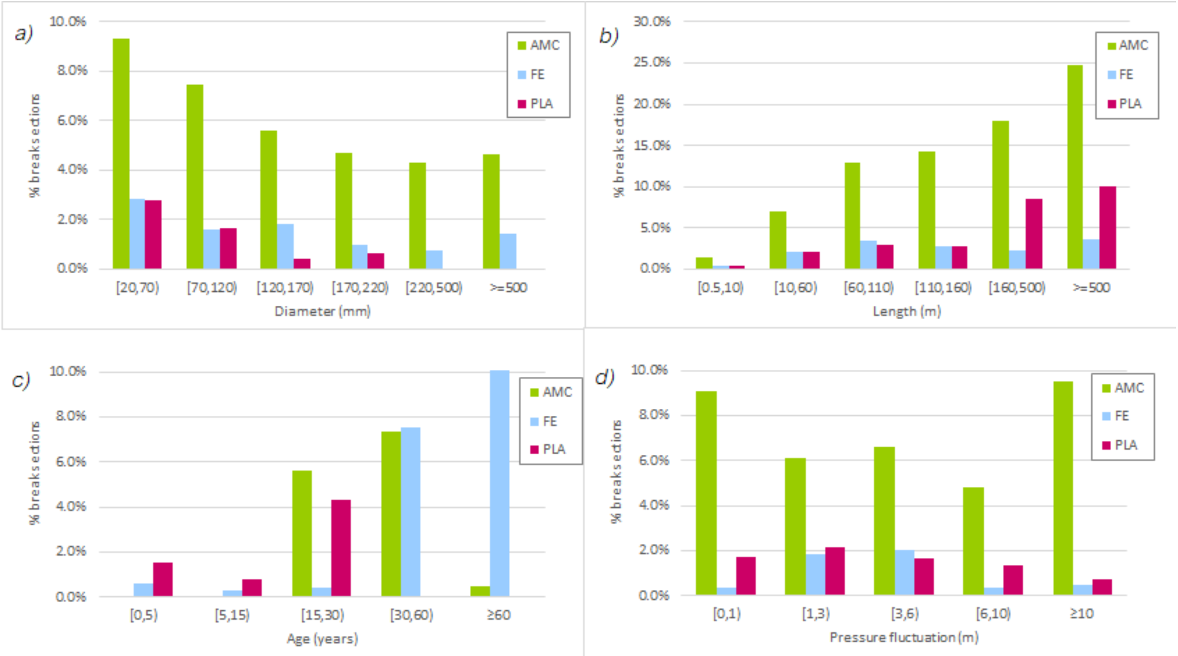


Figure 7. *Percentage (%) of pipe breaks per range of Diameter, Length, Age and Pressure fluctuation*

4.2. Predictive system

Both LR and SVC need label data to be trained because they are supervised methods. Four different models have been analysed. The global model which includes the entire network has eight input variables: MAT, DIA, AGE, LEN, CON, N_type, ΔPRE and NOPF, and in the three other models, which only include pipes of a kind of material (AMC, FE and PLA), MAT is not a variable. The output variable of every model is 1 if the pipe breaks in this year and -1 if not.

Calibration of the models

In the calibration phase, the first five years (2012-2016) are employed for training the models and the last two years (2017 and 2018) for their validations. All performance metrics of table 5 are calculated for the validation years and setting the hyperparameters to their default values (*C=1 and* γ*=1/8*). The accuracy represents the percentage of well-predicted pipes, while the recall shows the percentage of predicted failures. It can be appreciated that when the training set is not balanced, a minimal percentage of failures are predicted (simulations 1, 3, 5 and 7). The prediction of pipe failure is the main goal of the study, so the training set balance becomes a crucial procedure to obtain reasonable results. *Moreover, feature transformation, which includes the replacement of DIA and LEN variables by their logarithms, produces a significant increase of the accuracy.* Hence, it is demonstrated that both mechanisms entail a remarkable improvement in the methodologies' performance.

Table 5. *Results for the global model with and without feature transformation and balance of the training set. C=1 and* $\boldsymbol{\gamma}$*=1/8*

Approach	Sim. No.	Feature transformation	Training set balance	Resolution time (s)	Acc.	Recall	AUC
LR	1	No	No	5	0.993	0.036	0.750
	2	No	Yes	2	0.648	0.893	0.863
	3	Yes	No	3	0.994	0.009	0.773
	4	Yes	Yes	2	0.768	0.849	0.876
SVC	5	No	No	36,290	0.994	0.003	0.396
	6	No	Yes	89	0.589	0.849	0.798
	7	Yes	No	13,272	0.994	0.000	0.631
	8	Yes	Yes	21	0.735	0.886	0.855

Simulations have been done using Python 3.7 on a PC, 3 GHz dual core Intel 5 processor and 16.0 GB RAM and using Windows 10 as operating system. It should be noted that the resolution times (training and predicting) of SVC are greater than of LR, especially, when the training set is not balanced. Consequently, SVC is more suitable when the size of the database is small or medium.

A second set of simulations is made to estimate the optimal hyperparameters of the four models. These hyperparameters are the regularisation parameters, C_{lr} and C_{svc}, and γ, proper to the radial basis kernel function of SVC. The optimal configurations are marked in bold in tables 6 and 7. The criterion followed to decide which results are better is: (i) identifying the simulations whose mean between the accuracy and the recall are higher; (ii) if there is a difference lower than 1% between such means, the one with the highest AUC is chosen. In this case, feature transformation and balance of training set are always applied.

Table 6. *Estimation of the optimal hyperparameter* $\boldsymbol{C_{lr}}$ *of the LR algorithm*

C_{lr}	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC
0.1	Global	**0.766**	**0.853**	**0.876**	AMC	**0.660**	**0.761**	**0.784**	FE	0.868	0.775	0.871	PLA	**0.674**	**0.854**	**0.852**
1		0.768	0.849	0.876		0.662	0.757	0.784		0.871	0.772	0.871		0.697	0.792	0.842
10		0.768	0.849	0.876		0.662	0.755	0.784		**0.872**	**0.772**	**0.871**		0.698	0.771	0.836

Table 7. *Estimation of the optimal hyperparameters $\boldsymbol{C_{svc}}$ and γ of the SVC algorithm*

γ	C_{svc}	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC	Model	Acc.	Recall	AUC
0.01	0.1	Global	0.711	0.873	0.870	AMC	0.586	0.821	0.782	FE	0.893	0.738	0.861	PLA	0.985	0.333	0.835
0.01	1		**0.751**	**0.869**	**0.874**		**0.626**	**0.783**	**0.783**		0.891	0.755	0.864		**0.663**	**0.813**	**0.839**
0.01	10		0.736	0.880	0.870		0.659	0.759	0.782		0.895	0.769	0.870		0.719	0.771	0.816
0.125	0.1		0.738	0.880	0.859		0.623	0.777	0.776		0.887	0.762	0.876		0.612	0.792	0.780
0.125	1		0.735	0.886	0.855		0.630	0.785	0.778		0.889	0.769	0.879		0.691	0.792	0.784
0.125	10		0.741	0.890	0.851		0.644	0.755	0.771		**0.882**	**0.792**	**0.884**		0.744	0.708	0.760
1	0.1		0.704	0.894	0.828		0.545	0.840	0.739		0.912	0.520	0.869		0.754	0.458	0.735
1	1		0.744	0.883	0.830		0.643	0.771	0.745		0.877	0.785	0.875		0.675	0.792	0.752
1	10		0.752	0.828	0.820		0.655	0.705	0.715		0.836	0.812	0.855		0.738	0.563	0.763

4.3. Analysis of results

Quantitative and comparative analysis

The values presented in table 8 are the average of the quality metrics for seven generations *(7-fold cross-validation).* These values suggest that both methodologies have really good predictive abilities as well as generalisation capabilities. Firstly, it is observed that better results are reached by both approaches for the model that only includes FE pipes. This is due to the fact that these pipes have the lowest percentage of breakage. Secondly, the AMC model shows the worst results. In order to obtain moderately good recalls, the number of misclassified pipes considerably increases (low accuracies). This confirms that there is a difficulty on failure prediction of these pipes.

Table 8. *Summary of results attained by both methods for the optimal hyperparameters of each model*

Approach	Model	C	γ	Acc.	Recall	AUC
LR	Global	0.1	-	0.769	0.848	0.873
	AMC	0.1	-	0.666	0.727	0.774
	FE	10	-	0.873	0.807	0.888
	PLA	0.1	-	0.705	0.732	0.816
SVC	Global	1	0.01	0.750	0.866	0.872
	AMC	1	0.01	0.633	0.763	0.773
	FE	10	0.125	0.893	0.799	0.895
	PLA	1	0.01	0.717	0.724	0.816

A comparison between the global model performance and the performance of the three other models together (aggregated model) is presented in table 9. The accuracy of the aggregated model (eq. 16) is calculated as the weighted sum of the accuracy of each model (Acc_k) multiplied by its total number of pipes (NP_k). Meanwhile, for calculating the recall (eq. 17) only the number of pipes which suffer a failure (NP_k^f) is considered.

$$Accuracy_{agreg\_model} = \frac{\sum_k Acc_k \cdot NP_k}{\sum_k NP_k} \qquad k = AMC, FE, PLA \qquad (16)$$

$$Recall_{agreg\_model} = \frac{\sum_k Recall_k \cdot NP_k^f}{\sum_k NP_k^f} \qquad k = AMC, FE, PLA \qquad (17)$$

Although the accuracy of the global model is a bit lower, it achieves the prediction of a much greater percentage of breakages than the aggregated model. Since the corrective actions are more expensive than the preventive ones, it is concluded that the global model is the most suitable.

Table 9. *Comparison between the percentages of well-classified pipes and predicted failures using the global model and the aggregated model*

Approach	Model	Acc.	Recall
LR	Global	76.9%	84.8%
	Aggregated	79.4%	73.3%
SVC	Global	75.0%	86.6%
	Aggregated	79.6%	73.7%

A percentage of 76.9% of the pipes are well-predicted by using LR for the global model, avoiding 84.8% of pipe breaks. This represents, for instance, the possibility of predicting 541 of the 638 recorded failures during 2018. SVC gets to predict more failures, 86.6%, but at the cost of misclassifying more pipes, an accuracy of 75.0%. Since the annual replacement of the fourth part of the network is unfeasible both economically and physically, the replacement of those pipes, whose associated failure probability was the highest, is analysed. As previously mentioned, LR directly assigns a probability of failure to each pipe. In the case of SVC, this probability is also obtained based on the distance of the sample to the hyperplane which separates the classes. Once the pipes are ordered according to their probability of failure, the percentage of breakages that could be avoided by replacing only those with the greatest probability of breakage can be obtained. On the one hand, the use of the LR model for the year 2018 would have led to the prevention of 34.09% of breakages by replacing only 3.16% of the network's pipes (p_i> 0.85). On the other hand, SVC achieves avoiding 29.52% of breakages by replacing 3.84% of the pipes (pi> 0.85). These results show that costs could be significantly reduced.

The ROC curve is an appropriate tool to compare predictive models with the same response variable. Figure 8 depicts these curves for LR (red lines) and SVC (blue lines) and their corresponding AUC when the validation set is the year 2018. In all cases, the ability to avoid erroneous classifications is slightly higher using LR than SVC, except for the FE model. Both algorithms demonstrate to be very good classifiers with outstanding abilities to avoid false classification.

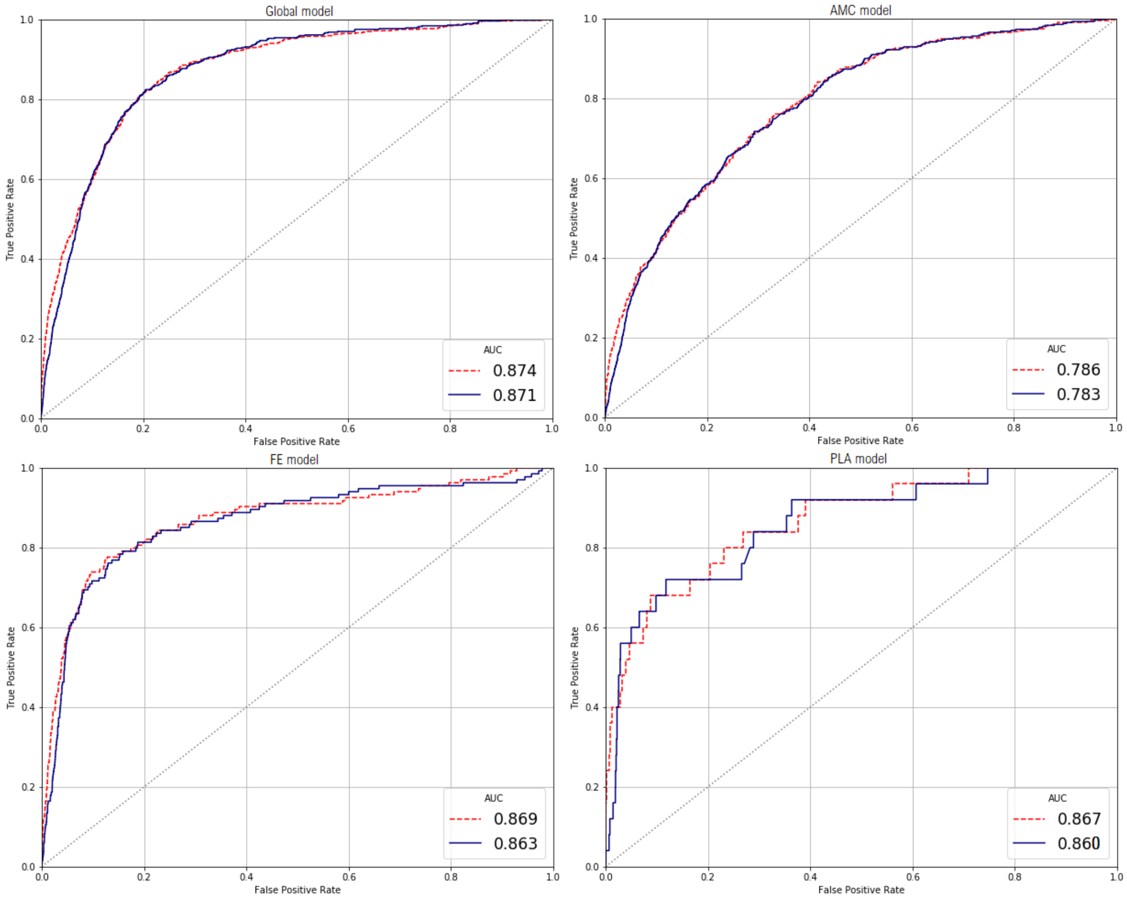


Figure 8. *ROC curves for the test data of year 2018*

Table 10 presents the AUCs obtained by various approaches used in the literature and those achieved in this work for the global model in the last recorded year. LR attains an AUC of 0.874 and SVC of 0.871, which represent a notable improvement with respect to the rest. Although the results might be influenced by the data quality because each method has been applied to different data, the applicability of the proposed methodologies to this particular problem is fully demonstrated. Moreover, it proves the importance of data pre-processing and the calibration of the models to enhance models' performance.

Table 10. *Comparison of AUCs for the problem of predicting pipe failures*

Authors	**Approach**	**AUC**
Debón *et al.* [11]	Cox model	0.769
	Poison GLM	0.828
Tang *et al.* [31]	BBNs automated	0.786
	BBNs guided	0.702
Our approaches	LR	0.874
	SVC	0.871

Physical interpretation of logistic regression parameters

LR allows extracting interesting information from the estimated weights (see table 11). Firstly, by substituting the intercept, b, in the equation $1/(1+e^{-b})$, it is obtained the probability of failure when all explanatory variables are equal to their means. Model assigns the highest breakage probability to PLA pipes, 0.40, followed by AMC 0.37.

***Table 11.** Estimated weights of LR model for the four studied models*

		Coefficients (w)								
		Intercept	Mat	log(DIA)	AGE	log(LEN)	CON	N_type	NOPF	APRE
Model	Global	-0.94	0.84	-0.13	0.27	0.78	0.01	-0.05	0.22	-0.09
	AMC	-0.52		-0.28	0.07	0.92	0.15	-0.07	0.27	-0.13
	FE	-1.33		-0.26	1.14	0.84	-0.02	0.07	0.19	-0.05
	PLA	-0.41		-0.56	0.10	0.67	-0.32	0.10	0.08	-0.04

According to the global model, the most influential factor is the pipe material followed by the pipe length, the age and the number of previous failures. The sign of the coefficients expresses if the effect of a unit change in an explanatory variable increases or decreases the probability of failure (e^{-w}). Therefore, it is confirmed that pipes with smaller diameters are more likely to break. Moreover, this factor becomes crucial to the appearance of failures in plastic pipes. The age is not relevant neither to cement nor plastic pipes, which means that other circumstances are causing these failures before the end of its lifetime. On the contrary, this factor does have a remarkable importance on the breakage of FE pipes as suggested in the initial historical data analysis. The length of the pipes exhibits a strong influence in all the models as well as, to a lesser extent, the number of previous failures. Regarding the number of connections per unit length, it only seems to be influential for cement pipes and, on an inverse sense, for plastic pipes. Plastic pipes with more connections are less prompt to break, perhaps, because the pressure inside these pipes is lower or more balanced.

5. Conclusions

This study presents a methodology to predict pipe failures in water supply networks. Firstly, data pre-processing is revised due to its great influence on models' performances. The visualisation and analysis of factors helps detect and prevent anomalies such as missing values and outliers. Secondly, logistic regression and support vector classification are utilised as a predictive system. Finally, results are analysed and compared using easily interpretable quality metrics such as confusion matrix and the ROC curves.

Among the explained predictive methodologies, logistic regression and support vector classification are chosen because they generate an output variable that can be interpreted as a failure probability which is increasingly demanded by companies. Additionally, both have key capabilities to deal with unbalanced class data which is the case of water supply networks. The major contributions of our study to the literature are the use of specific quality metrics to analyse and compare the results of the logistic regression model, and the use of support vector machines as a classifier, which has never been used before to address this problem.

Once models are chosen, these have to be calibrated to fit some hyperparameters. Moreover, mechanisms such as classes' balance and variables transformation demonstrate to greatly improve their performances.

The proposed methodology is illustrated with the real case of a Spanish city. This is an extensive water supply network, 3,800 kilometres, whose recorded data contains 4,393 pipe failures. It represents the largest water network analysed until now from a failure predictive point of view. Four different scenarios are analysed, one including the entire network and the other three in which pipes are grouped by kind of material (AMC, FE and PLA).

The results obtained are outstanding, showing LR a slightly better performance than SVC. The number of unexpected failures might be significantly reduced. Using the available historical data from the city of Seville, around 30% of failures could have been prevented by replacing only 3% of the network's pipes, which is a realistic and feasible option.

From a predictive point of view, the global model reaches better results than the three models separately. Moreover, AUCs attained for both algorithms are higher than those previously achieved in the literature. Accordingly, they have excellent abilities to avoid erroneous classifications. The coefficients derived from logistic regression show that material seems to be the most influential variable followed by pipe length, age and number of previous failures. It is also determined that pipes with smaller diameters are more prone to breakage.

Besides predicting the major number of breakages, the final goal of companies in charge is to avoid the highest priority ones. These are the ones that would cause the greatest problems as environmental and security risks, supply disruption of sensitive population or high reparation costs. Since the proposed methodology enables accurate predictions, the future line of research should be to study the consequences and costs derived from pipe failures. The objective must be to develop a global tool that incorporates the failure probability and its consequences, generating an optimal pipe replacement plan based on a budget and period of time.

Acknowledgements

The authors wish to acknowledge the financial support for the completion of this work by EMASESA (Empresa Metropolitana de Abastecimiento y Saneamiento de Aguas de Sevilla), through the Distinguished Chair in Water Network Management (Cátedra del Agua), a partnership programme between EMASESA and the Universidad de Sevilla (VI PPIT-US).

6. References

[1] AEAS, "XIV Estudio Nacional de Suministro de Agua Potable y Saneamiento en España," 2016. [Online]. Available: http://www.aeas.es/servlet/mgc?pg=ListNews&ret=next&news_id=1249&areaCode=publicarea&newsCategory=Noticias. [Accessed: 03-Jan-2019].

[2] A. M. St. Clair and S. Sinha, "State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models!," *Urban Water J.*, vol. 9, no. 2, pp. 85–112, 2012.

[3] K. Pietrucha-Urbanik, "Failure analysis and assessment on the exemplary water supply network," *Eng. Fail. Anal.*, vol. 57, pp. 137–142, 2015.

[4] M. Al-Zahrani, A. Abo-Monasar, and R. Sadiq, "Risk-based prioritization of water main failure using fuzzy synthetic evaluation technique," *J. Water Supply Res. Technol. - AQUA*, vol. 65, no. 2, pp. 145–161, 2016.

[5] S. Li, R. Wang, W. Wu, J. Sun, and Y. Jing, "Non-hydraulic factors analysis of pipe burst in water distribution systems," *Procedia Eng.*, vol. 119, no. 1, pp. 53–62, 2015.

[6] A. F. Royce, D. G. Seth, and L. Henneman, "Bayesian Belief Networks for predicting drinking water distribution system pipe breaks," *Reliab. Eng. Syst. Saf.*, vol. 130, pp. 1–11, 2014.

[7] Y. Kleiner and B. Rajani, "Comprehensive review of structural deterioration of water mains: physically based models," *Urban Water*, vol. 3, no. 3, pp. 151–164, 2001.

[8] S. Yamijala, S. D. Guikema, and K. Brumbelow, "Statistical models for the analysis of water distribution system pipe break data," *Reliab. Eng. Syst. Saf.*, vol. 94, no. 2, pp. 282–293, 2009.

[9] S. Christodoulou, A. Deligianni, P. Aslani, and A. Agathokleous, "Risk-based asset management of water piping networks using neurofuzzy systems," *Comput. Environ. Urban Syst.*, vol. 33, no. 2, pp. 138–149, 2009.

[10] 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.

[11] A. Debón, A. Carrión, E. Cabrera, and H. Solano, "Comparing risk of failure models in water supply networks using ROC curves," *Reliab. Eng. Syst. Saf.*, vol. 95, no. 1, pp. 43–48, 2010.

[12] Y. Kleiner and B. Rajani, "Comparison of four models to rank failure likelihood of individual pipes," *J. Hydroinformatics*, vol. 14, no. 3, pp. 659–681, 2012.

[13] G. Kabir, S. Tesfamariam, and R. Sadiq, "Predicting water main failures using Bayesian model averaging and survival modelling approach," *Reliab. Eng. Syst. Saf.*, vol. 142, pp. 498–514, 2015.

[14] D. P. De Oliveira, J. H. Garrett, and L. Soibelman, "A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage," *Adv. Eng. Informatics*, vol.

25, no. 2, pp. 380–389, 2011.

[15] M. Marzouk, S. A. Hamid, and M. El-Said, “A methodology for prioritizing water mains rehabilitation in Egypt,” *HBRC J.*, vol. 11, no. 1, pp. 114–128, 2014.

[16] G. Kabir, S. Tesfamariam, A. Francisque, and R. Sadiq, “Evaluating risk of water mains failure using a Bayesian belief network model,” *Eur. J. Oper. Res.*, vol. 240, no. 1, pp. 220–234, 2015.

[17] H. Fares and T. Zayed, “Hierarchical Fuzzy Expert System for Risk of Failure of Water Mains,” *J. Pipeline Syst. Eng. Pract.*, vol. 1, no. 1, pp. 53–62, 2010.

[18] Q. Xu, Q. Chen, W. Li, and J. Ma, “Pipe break prediction based on evolutionary data-driven methods with brief recorded data,” *Reliab. Eng. Syst. Saf.*, vol. 96, no. 8, pp. 942–948, 2011.

[19] A. Shirzad, M. Tabesh, and R. Farmani, “A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks,” *KSCE J. Civ. Eng.*, vol. 18, no. 4, pp. 941–948, 2014.

[20] A. J. Kettler and I. C. Goulter, “An analysis of pipe breakage in urban water distribution networks,” *Can. J. Civ. Eng.*, vol. 12, pp. 286–293, 1985.

[21] A. M. A. Sattar, B. Gharabaghi, and E. A. McBean, “Prediction of Timing of Watermain Failure Using Gene Expression Models,” *Water Resour. Manag.*, vol. 30, no. 5, pp. 1635–1651, 2016.

[22] A. M. A. Sattar, Ö. F. Ertuğrul, B. Gharabaghi, E. A. McBean, and J. Cao, “Extreme learning machine model for water network management,” *Neural Comput. Appl.*, vol. 31, no. 1, pp. 157–169, 2019.

[23] S. Toprak *et al.*, “Segmented pipeline damage predictions using liquefaction vulnerability parameters,” *Soil Dynamics and Earthquake Engineering*, vol. 125. p. 105758, 2019.

[24] M. O. Engelhardt, P. J. Skipworth, D. A. Savic, A. J. Saul, and G. A. Walters, “Rehabilitation strategies for water distribution networks: A literature review with a UK perspective,” *Urban Water*, vol. 2, no. 2, pp. 153–170, 2000.

[25] M. Najafi, *Trenchless Technology: Pipeline and Utility Design, Construction, and Renewal*, McGraw-Hil. 2005.

[26] P. F. Hudak, B. Sadler, and B. A. Hunter, “Analyzing underground water-pipe breaks in residual soils,” *Water Eng. Manag.*, vol. 145, no. 12, pp. 15–20, 1998.

[27] A. Baracos, W. D. Hurst, and R. F. Legget, “Effects of Physical Environment on Cast-Iron Pipe,” *J. Am. Water Work. Assoc.*, vol. 47, no. 12, pp. 195–206, 1955.

[28] F. Wang, X. zhong Zheng, N. Li, and X. Shen, “Systemic vulnerability assessment of urban water distribution networks considering failure scenario uncertainty,” *International Journal of Critical Infrastructure Protection*, vol. 26. 2019.

[29] N. M. Amaitik and C. D. Buckingham, “Developing a hierarchical fuzzy rule-based model with weighted linguistic rules: A case study of water pipes condition prediction,” *Proc. Comput. Conf. 2017*, vol. 2018-Janua, no. July, pp. 30–40, 2018.

[30] P. Lin and X. X. Yuan, “A two-time-scale point process model of water main breaks for infrastructure asset management,” *Water Research*. pp. 296–309, 2019.

[31] K. Tang, D. J. Parsons, and S. Jude, “Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system,” *Reliab. Eng. Syst. Saf.*, vol. 186, no. August 2017, pp. 24–36, 2019.

[32] M. Kutyłowska, “Prediction of Water Conduits Failure Rate – Comparison of Support Vector Machine and Neural Network,” *Ecol. Chem. Eng. A*, vol. 23, no. 2, pp. 147–160, 2016.

[33] V. Sze, Y. H. Chen, T. J. Yang, and J. S. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” in *Proceedings of the IEEE*, 2017, vol. 105, no. 12, pp. 2295–2329.

[34] M. Kutyłowska, “Forecasting failure rate of water pipes,” *Water Sci. Technol. Water Supply*, vol. 19, no. 1, pp. 264–273, 2018.

[35] J. Liu and E. Zio, “Integration of feature vector selection and support vector machine for classification of imbalanced data,” *Appl. Soft Comput. J.*, vol. 75, pp. 702–711, 2019.

[36] Y. Pu, D. B. Apel, and H. Xu, “Rockburst prediction in kimberlite with unsupervised learning method and support vector classifier,” *Tunn. Undergr. Sp. Technol.*, vol. 90, no. May 2018, pp. 12–18, 2019.

[37] H. Fares and T. Zayed, “Risk assessment for water mains using fuzzy approach,” *2009 Constr. Res. Congr.*, pp. 1125–1134, 2009.

[38] S. Christodoulou and A. Deligianni, “Neurofuzzy decision framework for the management of water distribution networks,” *Water Resour. Manag.*, vol. 24, no. 1, pp. 139–156, 2010.

[39] R. Farmani, K. Kakoudakis, K. Behzadian, and D. Butler, “Pipe Failure Prediction in Water

Distribution Systems Considering Static and Dynamic Factors," in *Procedia Engineering*, 2017, vol. 186, pp. 117–126.

[40] P. McCullagh and J. A. Nelder, *Generalized Linear Models*. Chapman and Hall Ltd, 1989.

[41] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[42] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region Newton methods for large-scale logistic regression," *J. Mach. Learn. Res.*, vol. 9, pp. 627–650, 2008.

[43] V. N. Vapnik, *Statistical learning theory*, John Wiley. 1998.

[44] S. Maldonado, J. Pérez, R. Weber, and M. Labbé, "Feature selection for Support Vector Machines via Mixed Integer Linear Programming," *Inf. Sci. (Ny).*, vol. 279, pp. 163–175, 2014.

[45] C. Chang, C. Lin, and T. Tieleman, "LIBSVM : A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 307, pp. 1–39, 2008.

[46] M. Aydogdu and M. Firat, "Estimation of Failure Rate in Water Distribution Network Using Fuzzy Clustering and LS-SVM Methods," *Water Resour. Manag.*, vol. 29, no. 5, pp. 1575–1590, 2015.

[47] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, Inc, 2017.

[48] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[49] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.